ADDALY · THE AI CLASSROOM

Making Things With AI

Images, video, voice and music — how they work, where they break, who owns them.

9 modules · 84 lessons · 29 figures · 9 case studies · 63,119 words · about 11 hours

Dr Mustafa Mun
Author and editor

ABOUT THIS BOOK

Making Things With AI, published by Addaly, 2026. Author and editor: **Dr Mustafa Mun**.

The manuscript was drafted with the assistance of a large language model and was edited and published under his name. That is stated plainly here rather than left to be discovered, because a reader is entitled to know how a book they are trusting was made.

This book is the printed form of the course of the same name at addaly.com/learn/making-things-with-ai. The website is the living version and is corrected first; where the two differ, the website is right.

Free to read, free to download, free to print, and free to teach from. There is no paid edition and there never will be. If you paid for this, somebody has sold you something that was given away.

Every diagram in it is drawn from data by code, not generated as a picture, so the numbers on the page are the numbers in the argument. Answers to every exercise, test and examination question are at the back, with a note on why each wrong option attracts people — those notes are the teaching, not the marking.

Contents

1. How the picture gets made

Trace one generated image from random noise to a saved file — naming the latent space, the text encoder, the denoising loop, the guidance term, the sampler and the seed — and say, with the published evidence, what a diffusion model does and does not retain from its training set

2. Steering it: prompts and what sits around them

Diagnose a failing prompt by identifying which part of the pipeline is ignoring it — encoder type, token limit, guidance, attribute binding or a missing reference — and rewrite it for the specific model in front of you rather than by adding adjectives

1. Writing a prompt that gets what you meant
2. Tag soup or sentences: why the advice reversed
3. What a negative prompt really does
4. Weighting, emphasis and where it breaks
5. The token limit that eats long prompts
6. When a reference image beats any sentence
7. Reading the metadata, keeping the record
8. The prompt you sent is not the prompt it got
9. A prompt is not an asset
10. Case study: The second batch did not match the first
11. Module test

3. Where it breaks, and why

Look at a failed generation and classify the failure by mechanism — binding, negation, resolution, distribution pull, physics or filtering — and state for each whether a prompt change, a structural control, a different model or a manual fix is the only thing that will work

1. What hands and text told us
2. Counting, and the objects that swap their attributes
3. Why "without" summons the thing
4. Left, right, behind: the relations it cannot hold
5. The pull toward the average
6. The bias, and what mitigating it actually does
7. Reflections, shadows and things that must agree
8. Faces in crowds and other small things
9. Refusals, filters and the false positives
10. The tells, and why they are disappearing
11. Case study: Forty cards, three wrong, and a term starting Monday
12. Module test

4. Control: getting the picture you actually need

Take a nearly-right generation to a finished deliverable using the appropriate mechanism at each stage — denoise strength, masked regeneration, a conditioning map, a small fine-tune, an upscaler and a manual composite — and say which of those a given defect requires

1. Editing what you already have
2. Denoise strength: the number that decides how much survives
3. Masks, feathering and the context window
4. Conditioning on structure: depth, edges and pose
5. Teaching it something it does not know
6. The same face twice
7. What an upscaler invents
8. Compositing is still the job
9. What runs on the machine you have
10. Case study: Sixty old product photographs and a trade fair in nine days
11. Module test

5. Time and space: video and three dimensions

Explain why generated video is short, expensive and inconsistent in terms of its underlying mechanism, design a fair test of a video model's claims, and say what a generated 3D asset is actually made of and where it can and cannot be used

1. Video, and what it still cannot do
2. Why a second of video costs what it does
3. What holds a moving picture together
4. Conditioning a clip: stills, frames and video in
5. Directing motion, and what actually lands
6. Why clips get worse as they get longer
7. Talking heads, lip sync and the consent line
8. Testing a claim that it understands physics
9. Testing a video model honestly
10. What synthetic video does to believing anything
11. Generated objects and the space around them
12. Case study: Thirty shots, one launch reel, and a bid due on Friday
13. Module test

6. Voice and speech

Describe the path from text to a waveform through tokens, an acoustic model and a vocoder, judge what a voice-cloning claim requires in consent and disclosure, and specify a verification control that defeats voice-impersonation fraud without relying on detection

1. Voice, and the line you do not cross
2. From text to a waveform
3. What a voice print actually is
4. Prosody, and why it still sounds slightly wrong
5. The languages it does badly, and who that is
6. The other half: turning speech into text
7. The dubbing chain, and where it breaks
8. Can you tell? Mostly not
9. The fraud that already happens
10. Case study: The voice that opened the account
11. Module test

7. Music and sound

Explain how a music model represents and generates audio, judge whether a generated track carries a real infringement risk and on what basis, and choose between generated music, licensed library music and a commissioned composer for a specific brief with reasons you could defend

1. Music, and whose music it learned from
2. How a machine writes a waveform
3. Why music is harder than speech
4. Pulling a mix apart
5. Sound effects, foley and where it is already good
6. Whose music it learned from
7. When a tune becomes somebody else's
8. What platforms do about it
9. When generated music is the right answer
10. Case study: The theme that sounded like something
11. Module test

8. Provenance, detection and disclosure

Assess an image or clip of unknown origin using provenance, source and corroboration rather than inspection, explain why a detector's positive result is a weak signal at realistic base rates, and write a disclosure that tells a reader what was synthetic and what was not

1. Why detection is the wrong hope
2. What a watermark can survive
3. Signed provenance, and where the chain breaks
4. Verify the claim, not the picture
5. When everything can be denied
6. Writing a disclosure somebody can use
7. What the platforms require now
8. A routine you can actually run
9. Case study: The strongest picture on the desk at eleven at night
10. Module test

9. Ownership, law and the parts nobody has settled

Identify which distinct legal question a given situation raises — training, output ownership, likeness, style, model licence, or disclosure duty — describe how the answer differs across jurisdictions, and draft the contractual terms that make a piece of AI-assisted client work defensible without relying on a resolution that does not exist

1. Who owns it, and who has to say it is AI
2. Was the training lawful?
3. Can anyone own the output?
4. Faces, voices and the rights in a person
5. Style is not owned, and other things are
6. Read the licence on the model
7. What an indemnity actually buys
8. Where labelling is already the law
9. What to write down before you start
10. A procedure for the cases nobody has answered
11. Case study: The mark the client could not own
12. Module test

Making Things With AI — final examination

The paper that opens the next course. Answers to everything follow it.

MODULE 1

How the picture gets made

Before any advice about prompts, the machinery. This block follows a single image from a request to a file: the noise it starts as, the small compressed space it is actually built in, how your words become a steering signal, what the guidance number does to the result, why the same prompt gives a different picture tomorrow, what the training run actually consisted of, and how much of the training images the finished model can be shown to hold.

> BY THE END OF THIS MODULE YOU CAN
>
> Trace one generated image from random noise to a saved file — naming the latent space, the text encoder, the denoising loop, the guidance term, the sampler and the seed — and say, with the published evidence, what a diffusion model does and does not retain from its training set

1 · 7 min

What the model is actually doing

THE ONE THING TO KEEP

An image model does not draw. It removes noise from static, over and over, steered by your words.

Start with a photograph you ruin

Take a photo. Add a little random speckle. Add a little more. Keep going for a thousand rounds and you have television static — the photo is gone. That

process is easy and needs no intelligence at all.

Here is the trick the whole field is built on: that ruining is reversible, if you can predict, for a slightly noisy image, exactly which speckle was added.

What training actually does

Training a diffusion model is boring in a good way. Take an image from the training set. Pick a random amount of noise. Add it. Show the model the noisy image and tell it how much noise is in there. The model guesses what noise was added. Compare with the real answer, nudge the weights, repeat a few hundred million times.

That is it. The model never learns "what a cat looks like" as a fact you could look up. It learns one narrow skill: given a mess, guess which part of it is mess.

Generating is that skill run backwards

Start with pure static — random numbers produced from a seed you can write down. Ask the model what noise it sees. Subtract a fraction of it. Ask again. After 20 to 50 rounds, static has become a picture, because at every step the model pushed the image slightly toward "looks like the training photos".

Two things follow immediately.

- The seed decides the starting static. The same seed with the same prompt gives you the same image every time. A different seed gives a different picture from identical words.
- The model has no canvas, no layers, no plan. Composition settles in the first handful of steps and gets refined afterwards. That is why changing one word can move the whole layout.

Where your words come in

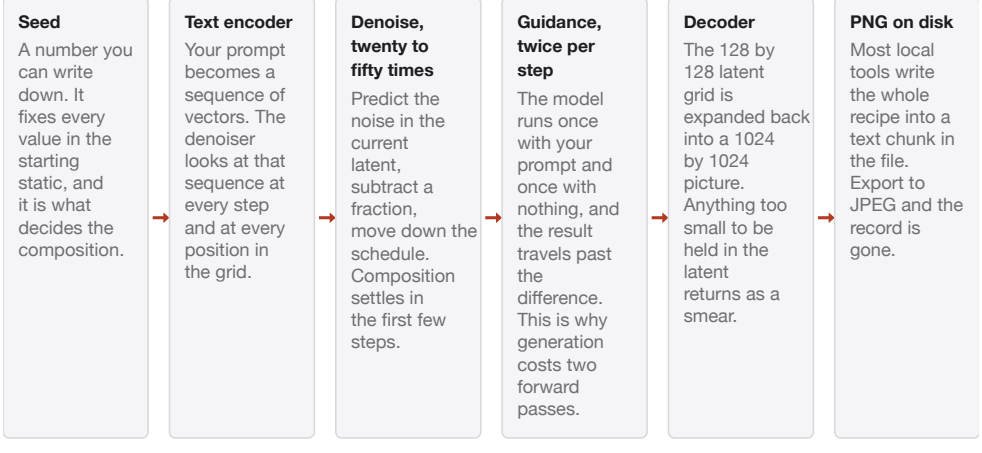
A text encoder turns your prompt into a list of numbers. Those numbers are fed into the denoiser at every step, so its guess becomes "what noise is here, given that this is supposed to be a wet street at night".

Then a technique called classifier-free guidance sharpens it. The model runs twice per step — once with your prompt, once with nothing — and the difference between the two is amplified. How much you amplify is the guidance scale, usually somewhere around 3 to 8. Set it low and the image drifts off your prompt. Set it high and you get the burnt, over-saturated, over-contrasted look people recognise as "an AI image". That look is not a style. It is a knob turned too far.

Doing it in a smaller room

Denoising a 1024x1024 image directly means handling over three million numbers on every step. So most models work in a compressed space. A separate autoencoder squashes the image roughly eight times in each direction — 1024x1024 becomes something like 128x128 with a few channels, about 48 times fewer numbers. All the denoising happens there, and a decoder expands the result back into pixels at the end.

One image, from a seed to a file



Nothing here is a plan or a canvas. That is why changing one word can move the whole layout, and why rerolling the seed is often the fix rather than another adjective.

That compression is why very fine detail comes back mushy: a face forty pixels tall, small lettering, thin wires. It was never represented finely enough to survive the round trip.

Not every model works this way

Newer image models swap the original network for a transformer, and swap noise prediction for flow matching, which learns a straighter path from noise to image and needs fewer steps. Some systems generate images the way a chat model generates text, one token at a time, which handles instructions and lettering better and behaves differently when you edit. The mental model above still gets you most of the way, but do not assume every tool has a seed field or a guidance slider.

Why bother knowing this

Because it changes what you do when something fails. If the model has no plan, then "be more specific" is often not the fix and rerolling the seed is. If detail was destroyed by compression, no number of adjectives recovers a tiny face, but regenerating that region larger will. Most prompt folklore is people mistaking mechanics for magic.

BEFORE YOU MOVE ON

You generate an image with a fixed seed and like everything except one background element. You change a single word in the prompt and generate again with the same seed. What should you expect?

A. Only the part of the picture described by the changed word moves, because the seed pins the rest of the image in place.

B. Nothing changes, because with the seed held constant the output is fully determined.

C. The whole image can shift noticeably, because the prompt steers every denoising step including the earliest ones, where composition is decided.

D. Only fine detail changes, because the early steps are governed by the seed and the prompt only influences the final steps.

2 · 9 min

Noise, steps and the schedule

THE ONE THING TO KEEP

Generation is a fixed number of small denoising steps down a noise schedule, which is why step count buys detail only up to a point and then stops.

Training is destruction, run backwards

A diffusion model is trained on a task that sounds pointless. Take a real photograph. Add a little random noise. Ask a network to predict the noise that was added. Repeat with more noise, and more, until the picture is indistinguishable from television static.

That forward process is fixed and requires no learning at all. It is arithmetic: at each level, keep a fraction of the image and add a fraction of Gaussian noise. The list of fractions — how much noise at each level — is the **noise schedule**. It is chosen by the people who train the model, written down, and shipped with it.

The network learns only the reverse: given a noisy image and a number saying how noisy it is, predict what the noise was. Subtract the prediction and you have a slightly cleaner image. Do that repeatedly, starting from pure static, and a picture appears that was never in the training set.

> *The model never learns "what a cat looks like" as a stored picture. It learns "given this smear and this noise level, which parts are noise". Run that from static and cats fall out, because cats were what the smears came from.*

What a step actually is

When your tool says **steps: 30**, that is the number of times the loop runs. Each step does the same three things:

1. Feed the current noisy latent, the timestep number and the text conditioning into the network.
2. Get back a prediction of the noise.
3. Remove a portion of it, sized by the schedule, and move to the next timestep.

The schedule decides how far each step travels. Early steps operate at high noise and settle the large structure — where the horizon sits, roughly where a body is. Late steps operate at low noise and settle texture — hair, fabric weave, the grain on a wall. This is why a generation that is going wrong is usually going wrong in the first five steps, and why waiting for step 28 to see whether the composition works is a waste of your electricity.

Why more steps stop helping

People assume steps behave like exposure time: more is better. They do not. Twenty steps of a modern sampler and eighty steps produce images that are close to identical, and the eighty-step version took four times as long.

The reason is that the loop is a numerical approximation of a smooth path. Each step is a guess at where the path goes next. Once the steps are small enough that the guess is accurate, halving them again buys nothing, because the error was already below what the eye — and the eight-bit output file — can hold. Modern samplers are accurate at 20 to 30 steps. Older ones needed 50 to 100 for the same result. Distilled models, trained specifically to take large steps, produce usable images in **one to four steps**.

A practical routine, which costs almost nothing:

```
seed 12345, 8 steps   -> is the composition right?
seed 12345, 20 steps  -> is the detail right?
seed 12345, 35 steps  -> is 35 visibly better than 20? usually
not.
```

Run that ladder once for each model you use and you will know its useful range for the rest of the year. Free tools make this easy: ComfyUI and

AUTOMATIC1111's WebUI both have a batch mode that varies one number and lays the results out side by side.

Where the schedule shows up in your results

Two models with the same architecture can behave differently because their schedules differ. A schedule that spends most of its steps at low noise gives crisp texture and weak composition; one that spends them at high noise gives strong composition and mushy texture. This is why a prompt that works beautifully on one checkpoint produces a flat, plasticky image on another that was fine-tuned with a different schedule.

There is a specific, documented flaw worth knowing. Several widely used schedules never quite reach pure noise at their top step — they leave a faint trace of the average brightness of the training set. The model therefore never learns to produce a genuinely very dark or very bright image, because at generation time it starts from true noise, which it never saw in training. That is the mechanism behind the complaint that certain models "cannot do night". It was a bug in the schedule, not a shortage of night photographs.

The honest limitation to carry forward: the schedule is a design decision made before you arrived, it is not exposed in most consumer interfaces, and it caps what your prompt can reach. When a model refuses to do something structural — very dark scenes, extreme aspect ratios — reach for a different checkpoint before you reach for another adjective.

BEFORE YOU MOVE ON

A generation at 20 steps and the same generation at 60 steps look nearly identical, though the 60-step run took three times as long. What is the best explanation?

A. The sampler's error is already below what the output file can represent at 20 steps, so extra steps refine a path that has already converged.
B. The model has a fixed internal resolution, so any extra computation is discarded before the image is written.
C. The prompt was too short to give the model enough to do with the additional steps, so it idled.
D. Step counts above roughly 25 are silently capped by most generation tools as a protection against runaway compute costs, so the additional steps were requested but never actually executed on the hardware.

3 · 9 min

The small space the picture is built in

THE ONE THING TO KEEP

Diffusion happens in a compressed latent grid roughly sixty-four times smaller than the image, and the autoencoder that compresses and restores it is the source of a specific family of defects.

A 512-pixel image is not built at 512 pixels

Running the denoising loop directly on pixels was tried, and it was ruinous. A 512 by 512 colour image is 786,432 numbers. Thirty passes of a large network over that many numbers, for every image, is why the first generation of these systems needed a data centre.

The fix that made image generation ordinary was to do the work somewhere smaller. Before the diffusion model ever sees your request, a separate network

called a **variational autoencoder** — the VAE — has been trained to squash images into a compact grid and to restore them afterwards. In the widely used Stable Diffusion family, the squash factor is eight in each direction, with four channels instead of three:

```
image   512 x 512 x 3  =  786,432 numbers
latent   64 x  64 x 4  =   16,384 numbers   (48x fewer)
```

The diffusion loop runs entirely on that 64 by 64 grid. Only at the very end does the VAE's decoder expand the finished latent back into pixels. Everything the model reasons about — composition, colour, the shape of a face — happens in a representation where one cell stands for an eight-by-eight block of the final picture.

Where the denoising actually happens

1024 by 1024 by 3 — the picture	3,145,728 numbers. Thirty passes of a large network over that many is why the first systems needed a data centre.
The autoencoder's encoder	Eight times smaller in each direction, and four channels instead of three.
128 by 128 by 4 — the latent grid	65,536 numbers, forty-eight times fewer. One cell stands for an eight-by-eight block of the final picture.
Every step, every mask, every conditioning map	All of it happens here. Composition, colour and the shape of a face are decided at this resolution.
The decoder, once, at the end	Small text, thin wires and distant faces were never represented finely enough to survive the trip back.

Push a photograph through the encoder and straight back out with no diffusion at all. What returns is the best possible output of that model: no prompt, fine-tune or sampler beats it, because everything leaves through this door.

What that buys, and what it costs

It buys the whole field. It is the difference between an image taking nine seconds on a laptop graphics card and ninety on a rented server.

The cost is that the VAE is lossy, and it is lossy in a patterned way rather than a random one. It was trained to reconstruct the *kinds* of images in its training

set with the least average error, which means it spends its capacity on what is common and starves what is rare. Three consequences you will meet:

- **Small text is destroyed even when the model got it right.** The diffusion model can place letter shapes correctly in the latent; if each letter is two latent cells wide, the decoder has no room to render it and returns a smear. Text at large sizes survives. Text on a signpost in the distance never will.
- **Fine regular patterns wobble.** Grille work, chain-link fence, printed circuit boards, distant window rows. The decoder reconstructs texture statistically, so a repeating grid comes back subtly irregular.
- **Faces at small scale degrade first.** A face occupying 40 pixels is five latent cells across. There is not enough there to hold two eyes, a nose and a mouth in the right relation, which is why crowd scenes look correct until you zoom in.

None of these are prompt problems. No adjective adds latent resolution. The fixes are structural: generate larger, or generate the small thing separately at full size and composite it, or use a model whose latent uses more channels. Newer architectures moved from four latent channels to sixteen precisely because four could not carry fine detail, and the improvement in small text and skin texture is visible without measuring anything.

The seam you can see

Because the decoder works over the whole latent at once but has a limited receptive field, generating an image much larger than the model's native size produces a specific artefact: repeated subjects. Ask a model trained at 512 for a 1024-wide image and you frequently get two heads, or a body with an extra torso. The model is not confused about anatomy. Each region of the oversized latent independently looks like a plausible start of a subject, and nothing coordinates them.

This is why the standard workflow is **generate at native size, then upscale**, rather than generating large in one pass. Native size is stated on every model card: 512 for the older Stable Diffusion checkpoints, 1024 for SDXL

and most current models. Working at native size and enlarging afterwards costs less and fails less.

Checking the ceiling for yourself

There is a five-minute experiment that teaches this better than any explanation. Take any photograph. Push it through the VAE's encoder and straight back out through the decoder, with no diffusion at all. Most local tools expose this; in ComfyUI it is a two-node graph, and it is free.

What comes back is the best possible output of that model. No prompt, no fine-tune, no sampler can beat it, because everything passes through this decoder on the way out. Compare the original and the round trip at 100% zoom and you will see exactly which details the pipeline cannot hold: the small type, the eyelashes, the moiré on a shirt. After that, you stop trying to fix those with words.

BEFORE YOU MOVE ON

A model generates a street scene at its native 1024 pixels. The shop sign in the middle distance is a convincing smear of letter-like marks rather than readable words. Which explanation fits the mechanism?

A. The text encoder truncated the prompt, so the sign's wording never reached the model.

B. The sign occupies too few latent cells for the decoder to reconstruct letterforms, so correct placement in the latent still decodes to a smear.

C. The model was not trained on photographs containing shop signs, so it has no letterforms to draw on.

D. Small text within a generated scene is deliberately removed by the pipeline's safety filter, which is designed to prevent the automated production of misleading or counterfeit signage.

4 · 9 min

How your words become a steering signal

THE ONE THING TO KEEP

A text encoder turns your prompt into a sequence of vectors that the network attends to at every step and every position, which is why prompts steer continuously rather than being read once.

The prompt is not read, it is attached

There is a mental model most people arrive with: the model reads the sentence, understands it, then draws. That is not the shape of the machinery, and the difference explains most prompt behaviour.

Your text goes first to a separate, already-trained **text encoder** — a language model that never generates anything. It converts the prompt into a sequence of vectors, one per token, each a few hundred to a few thousand numbers long. That sequence is then handed to the denoising network as a fixed side input. At every one of the thirty steps, and at every position in the latent grid, the network looks at that sequence and asks which parts of it are relevant *here*.

The mechanism for looking is **cross-attention**. For each cell of the latent, the network computes a similarity between what it is currently building there and each token vector, then blends the token vectors in proportion. A cell that is turning into fur attends strongly to "cat"; a cell in the background attends to "kitchen". Nothing assigns regions to words in advance. The assignment emerges, step by step, and it can go wrong.

Which encoder, and why it matters

The choice of text encoder is one of the largest differences between generations of image models, and it is not cosmetic.

- Early Stable Diffusion used **CLIP**, a model trained to match images with their captions. CLIP is excellent at the gist of a caption and poor at grammar. It was trained on alt text, where word order carries little meaning, and it inherits that.
- SDXL used two CLIP encoders together, concatenating their outputs, which improved reliability without changing the character.
- Current models add or replace these with **T5**, a general text model trained on prose. T5 tracks clauses, negation and relations far better, which is exactly why newer models can handle "the red mug is to the left of the blue one" when older ones could not.

This is the mechanism behind advice that seems contradictory across tools. On a CLIP-only model, comma-separated tags outperform sentences, because CLIP is a bag-of-concepts machine and the sentence structure is wasted. On a T5-based model, a plain sentence outperforms tags, because the encoder can use the syntax and the tag soup looks nothing like its training text. Neither piece of advice is wrong; each is about a different encoder.

The binding problem, at its source

Cross-attention explains the most common failure in prompting. Ask for **a red cube and a blue sphere** and you will sometimes get a blue cube and a red sphere, or two purple objects.

The token vectors for "red" and "cube" are separate. Nothing in the architecture ties an adjective to the noun that follows it. Both are available at every cell, and the cell picks by similarity. Regions building a cube attend to "cube" strongly and to both colour words weakly, so the colour that wins is close to arbitrary. Encoders that read syntax reduce the leak because "red" and "cube" end up with more similar representations; they do not eliminate it.

The practical consequence is that attribute binding gets more reliable when you separate the objects in the request — into different clauses, different regions specified explicitly, or different generation passes composited afterwards. It does not get more reliable when you shout.

The number nobody mentions

CLIP text encoders have a context of **77 tokens**, including two markers, so roughly 75 tokens of prompt. A token is about three-quarters of an English word. Past that limit, the rest is silently dropped or, in some tools, split into a second chunk and averaged with the first — which is not the same as being read.

Interfaces rarely warn you. A 200-word prompt on a CLIP-based model means about 55 words steer and the rest do nothing at all, and because the output still looks like the beginning of the prompt, people conclude their late clauses were ignored for stylistic reasons. They were not read. T5-based pipelines allow 256 or 512 tokens, which is why long prompts suddenly started working on newer models.

If your tool shows a token counter, watch it. If it does not, the test is simple: append something outrageous — "and a giraffe" — to the end of a long prompt. If no giraffe appears and no other change occurs, you are past the limit.

BEFORE YOU MOVE ON

A prompt asks for "a green bottle beside a brown loaf" and the model returns a brown bottle beside a green loaf. What does this reveal about the mechanism?

A. The model has learned an incorrect and persistent association between bottles and the colour brown from biased training captions, and is applying that learned association in place of the requested colour.

B. Attributes and nouns arrive as separate token vectors available at every latent cell, so nothing structurally binds "green" to "bottle".

C. The prompt exceeded the encoder's token limit, so the colour words were truncated.

D. The sampler introduced randomness late in the run, which shuffled colours between regions.

5 · 8 min

Guidance: the dial that makes it obey

THE ONE THING TO KEEP

Classifier-free guidance runs the model twice per step and extrapolates away from the unprompted prediction, so raising it increases obedience and saturation together, and past a point it destroys the image.

Two predictions, one exaggerated difference

Left alone, a diffusion model trained on captioned images follows a prompt loosely. The prompt is one influence among many, and early systems produced images that were plausible but only vaguely on topic.

The fix, **classifier-free guidance**, is a trick of arithmetic rather than training. At each step the model is run twice: once with your prompt, and once with an empty prompt. That gives two noise predictions — where the model would go with your words, and where it would go without them. The direction from the second to the first is "what your prompt is doing". Guidance amplifies that direction and follows it past where either prediction actually pointed.

```
guided = unconditional + scale * (conditional - unconditional)
```

At scale 1 the guided prediction equals the conditional one and the prompt is a gentle influence. At scale 7 you travel seven times the difference. At scale 20 you are far outside the range the model was trained to produce anything sensible in.

Two facts follow immediately, and they explain most of what people notice. First, guidance is why generation costs roughly twice the compute of a single forward pass — the model runs twice per step. Second, the "negative prompt" box in most interfaces is not a special feature. It replaces the empty prompt in

that second run. You are not telling the model to avoid something; you are moving the point it is being pushed away from.

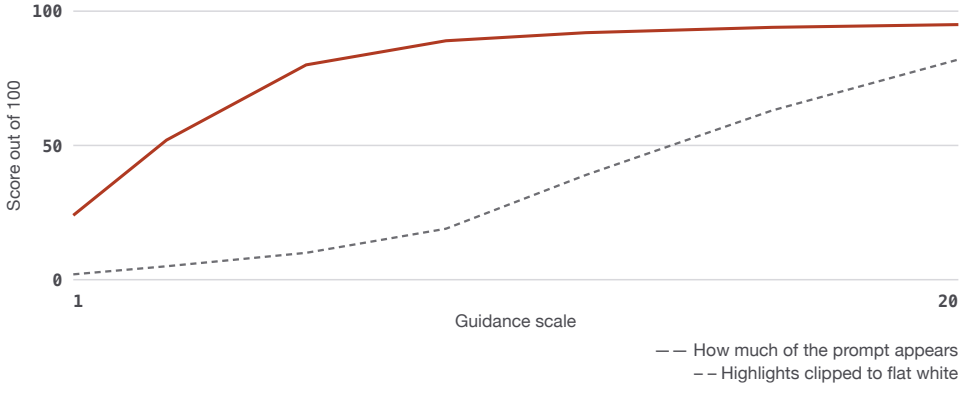
What each end of the dial looks like

Typical defaults sit at 6 to 8 for the Stable Diffusion family. Walking the range on a fixed seed teaches it in one minute:

- **1 to 3.** Loose, painterly, often ignores half the prompt. Colours are muted and natural. Occasionally the most photographic-looking result, because real photographs are not saturated.
- **6 to 9.** The usable band for most checkpoints. Prompt followed, contrast normal.
- **12 to 16.** Prompt followed hard. Colours push toward the edges of the range, skin goes waxy, highlights clip to flat white.
- **20 and above.** Burnt. Posterised colour, hard black outlines, sometimes a rainbow shimmer in flat areas.

The burning is a direct consequence of the arithmetic. Extrapolation drives the predicted latent outside the distribution the VAE was trained to decode, so the decoder is asked to render values it has never seen. Clipped, over-saturated output is what an autoencoder does when handed impossible inputs.

Walking the guidance dial on one fixed seed



Obedience is nearly all bought by nine and the burning is only starting. That is the whole trade, and it is why the usable band on this family is six to nine. Distilled and flow-matching models have moved the numbers, and on several of them the same burnt look arrives at seven.

The models where the dial has moved

Distilled and flow-matching models have changed the numbers, and advice written for one family is actively wrong for another. Several current models bake guidance in during training and expose a differently scaled parameter, where the useful band is 1 to 5 and setting it to 7 produces the burnt look that 15 produces elsewhere. Turbo and Lightning variants are trained to run at guidance 1 with three or four steps, and raising it destroys them.

There is no way to read the right number off the interface. It is on the model card, in one line, and reading that line is the difference between a model that seems broken and a model that works. This is a good habit to build early: before blaming a checkpoint, check the three numbers its author published — native resolution, step count, guidance scale.

There is also a scheduling trick worth knowing, because several free tools expose it. Guidance does not have to be constant across the run. Applying it strongly in the early steps, where composition is decided, and weakly in the late steps, where texture is laid down, gives prompt obedience without the burnt surfaces. Tools call this **guidance rescaling**, **dynamic thresholding** or **CFG scheduling** depending on who wrote them; ComfyUI ships several nodes for it and the WebUI has it behind an extension. If you find yourself stuck choosing between a picture that ignores you and a picture that looks like melted plastic, this is the setting that resolves the trade-off, and it is free.

When lowering it is the answer

Most people only ever raise guidance, because raising it makes the model obey. It is worth knowing the failure that lowering it fixes.

If your images look like stock illustration — plastic skin, impossible contrast, every surface glossy — the cause is usually guidance, not the prompt or the model. Photographs contain dull greys and blown-out corners. High guidance drives every region toward the most prototypical version of what it is, and the most prototypical mug is glossier than any real mug. Dropping from 9 to 4 and adding two steps often produces the image people were trying to reach with another twenty words of "photorealistic, 8k, award-winning". Those words are doing far less than the dial.

BEFORE YOU MOVE ON

Images from a checkpoint look plastic and over-saturated, with clipped white highlights, no matter how the prompt is rewritten. Which single change is most likely to fix it, and why?

A. Increase the step count, because more denoising steps recover the detail that saturation is hiding.
B. Add "natural lighting, muted colours" to the prompt, because the text conditioning is what sets tonal range.
C. Lower the guidance scale, because extrapolation is pushing latents outside the range the decoder was trained on.
D. Switch to a larger output resolution, because saturation artefacts are a symptom of generating below native size.

6 · 8 min

Seeds, samplers and reproducibility

THE ONE THING TO KEEP

The seed fixes the starting noise and the sampler decides the path away from it, so a result is only reproducible when both are recorded along with every other number.

Where the randomness actually is

Everything in the denoising loop is deterministic arithmetic. Given the same inputs, the network returns the same output every time. The variety in generated images comes from one place: the block of random numbers the loop starts from.

That block is produced by a pseudorandom generator, and a **seed** is the number that generator starts from. Seed 42 always produces the same static, so seed 42 with the same prompt, model, sampler, steps and guidance always

produces the same picture. Change the seed and every number in the starting static changes, which changes the composition entirely — not slightly.

This makes the seed the most useful control in the interface and the one beginners ignore. Fixing the seed converts generation from a slot machine into an experiment: change one thing, see what that thing did. Vary the seed and you are comparing two different pictures, which tells you nothing about the change you made.

Samplers: different routes, same destination

The sampler is the rule for turning the noise prediction into the next latent. They divide into two families, and the division matters more than the names.

Deterministic samplers — Euler, DDIM, DPM++ 2M, and the various "Karras" variants — follow a fixed path from the starting noise. Same seed, same result, every time. Increasing the steps refines the path toward the same picture.

Ancestral and stochastic samplers — anything with an `a` or `SDE` in the name, such as Euler a or DPM++ 2M SDE — inject fresh noise at each step. They often produce livelier textures, and they have two consequences people find baffling. The image keeps changing as you add steps rather than converging. And reproducing a result requires the same step count exactly, because the extra noise is drawn per step.

If you are exploring, ancestral samplers are pleasant. If you are working — comparing prompts, testing a fine-tune, building a batch that must match — use a deterministic sampler and you remove a variable you do not need.

What "the same seed" fails to preserve

This trips up everyone who tries to reproduce someone else's result and concludes their software is broken. Seed and prompt are not sufficient. All of the following change the picture:

- The model file, including a different quantisation of the same model.
- Step count, guidance scale, sampler, and the schedule variant.

- Output resolution, including a change of aspect ratio.
- The library version, which occasionally changes rounding.
- The hardware. Different GPUs, and CPU versus GPU, produce slightly different floating-point results that compound over thirty steps into a visibly different image.

The last one surprises people. Bit-identical reproduction across machines is not something these pipelines promise. What is reproducible is the *same machine, same versions* case, which is what you need for your own work.

Seed discipline, in practice

A workflow that costs nothing and saves hours:

1. Explore with random seeds until something is roughly right.
2. Note that seed. Every serious tool displays it after generation, and most write it into the PNG.
3. Fix it, then change exactly one thing at a time.
4. When you have the settings, sweep seeds with everything else locked to get variations of the same idea.

There is a habit worth adopting from research practice. Never judge a prompt change on one seed. A single generation tells you what happened to one sample; four fixed seeds tell you what happened to the prompt. Teams have shipped worse prompts than the ones they had because a lucky seed made the change look like an improvement.

Two final notes. Seed **-1** or **random** means the tool picks one for you and reports it afterwards — it is not a special mode, and the number it picked is recoverable. And seeds are not portable in any meaningful sense: seed 42 on one model has nothing in common with seed 42 on another, because the starting noise is the same numbers but everything that interprets it is different.

One further use of seeds is worth having. Because the starting noise determines composition, a seed you like is effectively a reusable layout. Keep a short list of seeds that gave you good framing for a given aspect ratio, and reuse them when you change the subject: the horizon, the placement of the fig-

ure and the balance of the frame carry over surprisingly often within a single model. Illustrators who work in sets — a series of covers, a run of cards — get consistency this way far more cheaply than by describing the composition in words. It stops working the moment you change model or resolution, which is a good reminder of what a seed actually is.

BEFORE YOU MOVE ON

A colleague sends you a prompt, a seed and a screenshot. You run it and get a clearly different image. What is the most likely cause, assuming neither of you made a mistake?

A. Seeds are model-specific, so the same seed on the same model still produces different noise on different runs.

B. The tool silently randomises the seed whenever the prompt contains uncommon words.

C. Your colleague used an ancestral sampler, and results from ancestral samplers cannot be reproduced under any circumstances because fresh noise is drawn on every single step of the run.

D. Something outside the prompt and seed differs — most likely the checkpoint, sampler, step count, guidance or resolution.

7 · 9 min

What the training run actually consisted of

THE ONE THING TO KEEP

These models were trained on billions of image-and-alt-text pairs scraped from the open web, so the caption vocabulary of that scrape sets the language your prompts have to speak.

Billions of pairs of picture and caption

The open image models of the past few years were trained on web-scale datasets built the same way: crawl the web, keep every image that has alt text, filter, and store the pairs. The best-documented is **LAION-5B**, released in 2022, with roughly 5.85 billion image-URL and caption pairs derived from Common Crawl. LAION itself stores no images — only links and text — which is part of how it was assembled at that scale and also why parts of it decay as pages disappear.

Two things about this deserve to be said plainly.

First, alt text is not description. It is what a webmaster typed to satisfy an accessibility checker or a search engine. A large share of it is a file name, a product code, a stock-photo watermark caption, or a keyword string. The model's idea of the English language is this corpus, not literature.

Second, the assembly was not curated in any meaningful sense. In December 2023, researchers at the Stanford Internet Observatory found that LAION-5B's index contained links to child sexual abuse material. The dataset was withdrawn, cleaned against known hash lists, and re-released as Re-LAION-5B in 2024. This is not a footnote. It is the clearest available evidence about what "scraped from the open web" means as a sourcing practice, and it is why later datasets from serious labs are filtered far harder and documented far less.

Why this dictates how you prompt

The strange conventions of image prompting are downstream of the caption corpus.

"Trending on ArtStation" worked because those words appeared in the alt text and surrounding metadata of a particular kind of polished digital illustration. "35mm, f/1.8, bokeh" works because photographers write their settings into captions, and those settings correlate with a shallow-focus look. "Unreal Engine" works for the same reason. None of these are instructions the model understands. They are strings that co-occurred with a visual style.

That also explains the failures. The model has weak coverage of anything the web does not caption well: specific tools, regional dress with names that vary by language, technical diagrams, anything photographed mainly by people who do not write English alt text. Ask for a *lungi* or a *sarpech* or a specific loom and you will get something adjacent, because the caption density is thin. Ask for a *wedding dress* and you will get white satin, because the caption density is enormous and Western.

> *When a prompt fails, the first question is not "how do I phrase this better" but "would anyone have written this caption on the web". If the answer is no, phrasing will not save you and you need a reference image or a fine-tune.*

Newer training, and what changed

The most visible recent improvement in prompt following came from **recaptioning**. Rather than train on the original alt text, labs pass every training image through a vision-language model that writes a long, literal description, and train on that. It is expensive and it works: models trained this way follow sentences with spatial relations, counts and multiple attributes far more reliably, because their captions during training actually contained those things.

It has a side effect worth knowing. Models trained on machine-written captions respond best to prompts that look like machine-written captions — full sentences, literal, describing what is in the frame. The old tag-soup style degrades on them. If a new model seems worse than an old one on your existing

prompts, this is usually why, and the fix is to rewrite the prompt as a description rather than a keyword list.

What this means for the ownership argument

Hold this lesson next to the legal module later in the course. Every serious dispute about generative images starts here: billions of copyrighted works were downloaded and used as training input, in most cases without a licence and without a workable way to opt out, on the argument that training is a transformative analysis rather than a reproduction. Courts in different countries have begun to answer that differently. Whatever you conclude, the factual base is not in dispute and is worth knowing precisely: web scrape, alt text, billions of pairs, no consent mechanism at collection time.

BEFORE YOU MOVE ON

A prompt for a specific regional garment returns a generic, roughly-similar item, while a prompt for a Western wedding dress is rendered accurately. What does this most directly demonstrate?

A. The model applies a bias correction that favours widely recognised subjects.

B. The regional garment is too visually complex for the latent resolution to represent.

C. Model output quality tracks the caption density of the training scrape, which is heavily weighted toward English-language web content.

D. The text encoder is unable to tokenise garment names that are not English words, so the term is discarded before conditioning and the model receives no signal about it at all.

8 · 8 min

Diffusion is not the only way

THE ONE THING TO KEEP

Generative images have been built with GANs, autoregressive token models and flow matching as well as diffusion, and the architecture explains what a given tool is fast at and bad at.

Four families, and what each is for

"AI image generator" covers several distinct machines. Knowing which you are using explains behaviour that otherwise looks arbitrary.

GANs. Two networks in a contest: a generator makes images, a discriminator tries to spot fakes, and both improve. GANs dominated until about 2021 and are still the fastest thing available — a single forward pass, milliseconds, no loop. StyleGAN's faces were startling in 2019 and still are. Their weakness is coverage: a GAN trained on faces makes superb faces and nothing else, and training one on everything tends to collapse onto a narrow slice of the data. They remain the right tool for narrow, high-volume jobs, and much free face-restoration and upscaling software is still GAN-based.

Autoregressive token models. Chop an image into a grid of discrete tokens, then predict them one at a time in sequence, exactly as a language model predicts words. This was DALL·E's original design, and it has returned: the image generation now built into several multimodal chat assistants works this way. Because the same network handles text and image tokens, it can use everything it knows about language, which is why these models are much better at rendering readable text in an image and at following long, fussy instructions. The cost is speed — one token at a time is a long sequence — and a different failure signature, where the picture drifts as it is written.

Diffusion. The subject of the last four lessons. Strong at texture and photographic realism, parallel over the whole image, controllable through a rich set

of conditioning methods because the loop offers thirty opportunities to intervene.

Flow matching. The current direction, and the basis of most 2024-onward open models. Instead of learning to remove noise at many discrete levels, the model learns a straight-line velocity field from noise to data. In practice it is a cleaner training objective that converges faster and needs fewer steps. From the outside it behaves like diffusion with different default numbers, which is why guidance values that suit older models are wrong for these.

Why the distinction reaches you

Three practical consequences.

Text in images. If you need a poster with correct spelling, an autoregressive or hybrid model will usually beat a pure diffusion model, because it generates the picture as a sequence with the language machinery attached. This is a real capability difference, not a matter of prompting harder.

Editing. Diffusion models offer masks, structural conditioning and strength dials because the loop can be interrupted. Sequential token models are typically edited by asking again in conversation. If your work is iterative retouching, the diffusion toolchain gives you more handles.

Running it yourself. Open weights are overwhelmingly diffusion and flow-matching. The strongest autoregressive image models are, at the time of writing, API-only. Anyone who needs offline work, private data or unlimited volume is choosing from the diffusion family whether they intended to or not.

The honest limitation

None of these families has solved the same set of problems. The autoregressive models write text well and still miscount objects. Diffusion models render skin and cloth better and still cannot spell. Both fail on the same negation prompts. This is evidence that the remaining failures are not artefacts of one architecture but something deeper about learning images from captions, which is worth remembering when a vendor claims their new architecture has solved understanding.

A useful habit: when you meet a new tool, find out which family it belongs to before you form an opinion about it. The model card, the research paper or a single search will tell you. Then test the three things that family is known to be weak at. You will know in ten minutes what would otherwise take a month of confused experimenting.

Increasingly the answer is "more than one". Several current systems are hybrids: a language model reads your request and writes a detailed internal description, which a diffusion or flow model then renders. That is why some assistants follow a complicated instruction well while still producing an image with diffusion's characteristic strengths and weaknesses — you are talking to one machine and being drawn for by another. It also explains a behaviour people find eerie, where the picture reflects something you implied rather than said. The intermediate description was written by a model that infers, and you never see it. When a result is oddly interpreted, asking the system to state the description it used is often available, and it is usually the fastest way to find where your meaning was lost.

BEFORE YOU MOVE ON

You need a printed flyer with a headline rendered in correct, readable English inside the generated image. Which choice is best supported by how these architectures work?

A. Reach for an autoregressive or hybrid model, whose sequential token generation carries the language machinery into the image.

B. Use a diffusion model at maximum guidance so the text tokens are followed exactly.

C. Use a GAN, since its single forward pass avoids the accumulated drift that corrupts letterforms.

D. Generate the flyer at four times the model's native resolution, so that each individual letterform has enough pixels available to come out sharp and legible in the printed piece.

9 · 9 min

What the model does and does not retain

THE ONE THING TO KEEP

A diffusion model stores a compressed statistical summary rather than a library of images, but images duplicated many times in the training data can be reproduced almost exactly, and the measured rate is small but not zero.

The size argument, and why it is not the whole answer

The most-repeated defence of these systems is arithmetic. Stable Diffusion 1.5 has roughly 860 million parameters, about 2 gigabytes on disk in half precision. It was trained on billions of images. If it stored them, it would be storing each in a fraction of a byte, which is impossible. Therefore it cannot contain the training images.

The arithmetic is right and the conclusion is too strong. Average compression across the whole set is not the same as compression of any particular item. A model can be an extreme lossy summary of nearly everything and still hold a handful of specific images almost exactly — precisely the ones it saw thousands of times.

What the measurements say

This has been tested rather than argued. The best-known study, by Carlini and colleagues in 2023, took Stable Diffusion, selected the training captions with the most duplicated images, and generated a very large number of samples — on the order of 175 million — checking each against the training set.

They recovered roughly a hundred near-exact copies. That is a small number against 175 million generations, and it is not zero. Every recovered image shared one property: it appeared many times in the training data, typically hundreds of times, usually because it was a stock photo, a film poster, a book cover or a widely reposted meme.

The mechanism is straightforward once you have the earlier lessons. Training reduces prediction error. An image seen once contributes a nudge that is averaged away among millions of others. An image seen a thousand times, with the same caption, becomes a reliable low-error target — the model can lower its loss by learning that specific mapping. Memorisation is not a bug in the architecture; it is what optimisation does with duplicates.

The consequences you should carry

Deduplication is the fix, and it is now standard. Later training runs remove near-duplicate images before training, which drops the memorisation rate substantially. When a lab says its dataset was deduplicated, this is the risk it is talking about.

Named artists and named works behave differently from styles. A prompt naming a very famous single painting has a real chance of producing something close to it, because that painting appears in the training data thousands of times. A prompt naming a living illustrator with a modest online presence usually produces a loose imitation of their style rather than a copy of any one work — which raises a different question, since style is not protected by copyright in most jurisdictions while a specific reproduction is.

"The model cannot copy" is not a safe thing to tell a client. The honest statement is that copying is rare, is concentrated in heavily duplicated images, and is your responsibility to check when the output will be published. The check is cheap: reverse image search the result. Google Lens, TinEye and Yandex are free, take fifteen seconds, and will catch the poster or stock photo that came back nearly intact.

Faces, and the sharper version of the problem

The same duplication effect applies to people. Individuals who appear in the training scrape thousands of times — actors, politicians, anyone with a heavy image presence — can be produced recognisably by name. Individuals who appear a few times generally cannot. This is why the practical risk of a model generating an identifiable private person by accident is low, and why the risk

of generating a public figure in a fabricated situation is high enough that most services block it by name.

The limitation nobody has solved: you cannot inspect a model to find out what it memorised. There is no index to query. The only method available is the one the researchers used — generate a great deal and compare against the training set — and that requires access to the training set, which most commercial models do not publish. So for a closed model, nobody outside the lab can tell you what it holds. That is worth saying out loud when someone offers you an indemnity based on a claim about training data you cannot verify.

One more distinction is worth keeping straight, because the two get argued as though they were one thing. Reproducing a specific image is a copyright question with a fairly clear shape in most countries. Reproducing a *style* is not — style as such is generally outside copyright, which is why a great deal of the anger about these systems does not map onto a cause of action, and why several artists' claims have been reframed around trademark, passing off or contract instead. That is covered properly in the final module. What the memorisation evidence settles is only the narrow question: near-exact copies are possible, they are rare, and they cluster on duplicates.

BEFORE YOU MOVE ON

A generated image of a famous album cover comes back near-identical to the original. Which explanation fits the published evidence?

A. The model contains a compressed copy of its entire training set, retrieved when the prompt matches closely enough.

B. The generation pipeline silently performs a web search and blends retrieved results into the output.

C. The album cover appeared many times in the training data, so learning that specific mapping reduced the training loss.

D. The prompt was long enough to exceed the encoder's token limit, and when conditioning is truncated the pipeline falls back on the nearest cached exemplar it holds for that subject.

CASE STUDY · MODULE 1

Forty thousand posters and a photograph nobody recognised

A district health office in Nashik, the evening before a dengue-awareness print run, with a grant that expires on the thirty-first

The Nashik district health office had a dengue season starting and a grant that lapsed at the end of the month. The communication officer, Sunita Rane, needed forty thousand posters in Marathi for panchayat noticeboards, primary health centres and school gates. The illustration budget was eleven thousand rupees, which in previous years had bought clip art.

This year an intern generated the artwork. A courtyard at mid-morning: an overturned bucket of standing water, a woman in a green sari tipping out a flowerpot, a child watching from a step. It was good. It was better than anything the office had printed in a decade, and everyone who saw the proof said so.

The press in Malegaon wanted the file by nine the next morning. At six in the evening a colleague looked at the proof for a long moment and said the woman looked familiar.

That is a thin thing to act on. Sunita acted on it anyway, because the check is free. She dropped the proof into three reverse image search services in turn. Two returned nothing useful. The third returned a stock photograph that had been on a library since 2015: the same courtyard geometry, the same pose, the same pot at the same angle, the same light falling in the same direction. It was not identical. The sari was a different colour, there was no child, and the wall behind was plain rather than tiled. But once she had the two side by side she could not unsee the relationship, and neither could anyone else in the room.

The intern's first response was arithmetic, and it is the argument everybody reaches for. The model file is two gigabytes. It was trained on billions of images. It cannot be storing them. Therefore it cannot have copied anything.

The arithmetic is right and the conclusion does not follow. Average compression across a whole training set says nothing about any particular item in it.

An image that appeared once contributes a nudge that is averaged away. An image that appeared a thousand times, with much the same caption each time, becomes a reliable low-error target that the model can lower its training loss by learning specifically. This stock photograph had been syndicated to health departments, NGOs and municipal websites across two decades. Its caption was, in effect, the intern's prompt.

So the choice in front of Sunita at half past six had a cost on both sides, and neither cost was hypothetical.

Holding the print run meant losing the press slot. The next slot was the following week, the grant lapsed on the thirty-first, and unspent grant money does not roll forward. Dengue cases in the district had already started climbing. A poster campaign that arrives after the peak is a filing exercise.

Running it meant publishing, under a government masthead, at a scale of forty thousand, an image that a stock library's automated matching could plausibly find. The library's remedy would be ordinary and manageable. The reputational problem would not be, because the whole product of a public health poster is that people believe it. A district office caught using somebody else's photograph badly disguised would be a story that outlived the dengue season.

There was a third thing she had to establish before she could decide anything, and it took four minutes. She asked the intern for the generation record. He had exported the final artwork as a JPEG for the press, which destroys the text chunk that local tools write into a PNG. But he had kept the original PNGs in a folder on his own laptop, out of habit rather than policy, and the prompt, the seed, the sampler, the step count and the guidance scale were all sitting in the file.

STOP HERE

Before turning the page: what would you have done, and what would it have cost if you were wrong? Write it down. The point of the next section is to compare.

WHAT HAPPENED

Sunita changed the seed.

That was the whole intervention, and it worked because of what a seed is. The seed fixes the block of random numbers the denoising loop starts from, and the starting noise is what decides composition. The prompt had steered the model toward a region of its training distribution where one heavily duplicated photograph sat as an attractor. A different seed starts the walk somewhere else entirely and arrives at a different arrangement of the same subject. Adding adjectives would have done the opposite: it would have kept her in the same caption neighbourhood and asked for more of it.

Eleven generations later, with the prompt untouched and only the seed varied, she had a courtyard that reverse-searched clean on all three services. The whole exercise cost ninety minutes. The press slot moved to eleven and the run went out on time. The grant survived.

The office wrote one page of rules the following week. Every image that will be published gets a reverse image search on at least two services before it leaves the building. The PNG is the master and the JPEG is the derived file, so the generation record survives. And the record for any published artwork — model, prompt, seed, date — goes in the job folder, not on an intern's laptop.

Six months later the same check caught a film poster that had come back nearly intact inside a nutrition leaflet. That one took two minutes to find and four to fix, and nobody argued about whether the check was worth the time.

A colleague said the woman looked familiar, and nobody could say why. Why did a reverse image search settle in four minutes what looking at the picture could not settle at all?

Inspection asks whether the image looks copied, which is a question about an impression. Reverse image search asks whether this composition already exists somewhere on the indexed web, which is a question about a record outside the file. The second question has an answer that does not depend on how good the generator was or how carefully anyone looked. It is also the check that catches the specific mechanism at work here: near-copies cluster on images that were duplicated many

times in training, and images duplicated many times in training are exactly the ones a search index has seen many times too.

The intern argued that a two-gigabyte model trained on billions of images cannot be storing any of them. Where does that argument break down?

It confuses average compression with per-item compression. A model can be an extreme lossy summary of almost everything and still hold a handful of specific images close to exactly. Training minimises prediction error, and an image that appears hundreds or thousands of times with a consistent caption is a target the optimiser can profitably learn as a specific mapping. The measured picture supports this: a well-known study generated on the order of 175 million samples from Stable Diffusion and recovered roughly a hundred near-exact copies, every one of them from a heavily duplicated training image. Rare, not zero, and concentrated on duplicates.

Why did rerolling the seed fix it, when adding more descriptive words to the prompt would probably not have?

The seed determines the starting noise, and composition is settled in the first few denoising steps out of that noise. A new seed produces a genuinely different arrangement from identical words. The prompt, by contrast, is what steered the model into the region of the distribution where the duplicated photograph sits. Adding adjectives refines a request inside that same region and can easily pull harder toward the very image being avoided. When the problem is that the output is too close to one specific thing the model saw repeatedly, the control that moves you is the one that changes where the walk begins.

MODULE 1 · TEST

Check what stuck

6 questions, one right answer each. This one is for you, not for a record: answers and the reason behind each are at the back. If two questions from the same module go wrong, re-read that module before the exam rather than after.

1 On a modern sampler, thirty steps and eighty steps produce images that are close to identical, and the eighty-step version takes nearly three times as long. What accounts for that?

A. The model has a fixed detail budget set by its parameter count, and it is exhausted after about thirty steps.

B. Guidance is applied only during the early steps, so the later ones have no prompt left to follow.

C. Once the steps are small enough, the sampler's remaining error is already below what the output file can hold.

D. The extra steps are spent on background regions of the latent, which have already settled.

2 A model places the letters of a shop sign in the right order in the latent, and the finished image still shows a smear. Which part of the pipeline destroyed it?

A. The autoencoder's decoder, because each letter occupied only a couple of latent cells.

B. The text encoder, because the prompt exceeded its token limit and the sign was cut off.

C. The sampler, because too few steps ran to resolve fine texture.

D. The guidance term, because high guidance clips small high-contrast detail.

3 Raising guidance from seven to sixteen on the same seed gives waxy skin, clipped white highlights and posterised colour. What is the mechanism?

A. The text encoder saturates and starts returning near-identical vectors for every token.

B. Higher guidance runs more denoising steps, so late-stage texture is over-refined.

C. The negative prompt gains weight and pushes the image toward the empty-prompt anchor.

D. The extrapolated prediction leaves the range of latents the decoder was trained on.

4 You send a colleague your prompt, seed, model file, sampler, step count, guidance and resolution. She runs it on her own machine and gets a visibly different picture. What is the most likely reason?

A. Her graphics card produces slightly different floating-point results, which compound over thirty steps.

B. Seeds are model-specific, so the same number means something different on her copy.

C. Her sampler injects fresh noise at every step, so results never repeat.

D. The prompt was re-tokenised on her machine and the token order changed.

5 Which prompt carries the highest chance of returning something close to an exact reproduction of a specific existing work?

A. A two-hundred-word prompt, because long prompts pull the model toward specific images.

B. A prompt naming a very famous single painting.

C. A prompt naming a living illustrator who has a modest online presence.

D. Any prompt run at guidance above twelve, because high guidance seeks the most typical example.

6 A prompt written as a long comma-separated keyword list gets good results on a 2022 checkpoint and vague, generic results on a current one. What changed?

A. The newer model has a shorter token limit, so most of the keyword list is discarded.

B. The newer model defaults to a much higher guidance scale, which washes out the keywords.

C. The newer model is autoregressive rather than diffusion, and ignores commas.

D. The newer model was trained on machine-written sentence captions rather than scraped alt text.

MODULE 2

Steering it: prompts and what sits around them

A prompt is a piece of text aimed at a specific encoder, and almost everything people believe about prompting is a description of one encoder's habits. This block covers what your words are competing against, what negative prompts and weighting actually do to the arithmetic, the token limit that silently eats long prompts, why the advice reversed when models changed encoder, when a reference image beats any sentence, and how to keep a prompt as a reproducible record rather than a lucky accident.

> BY THE END OF THIS MODULE YOU CAN
>
> Diagnose a failing prompt by identifying which part of the pipeline is ignoring it — encoder type, token limit, guidance, attribute binding or a missing reference — and rewrite it for the specific model in front of you rather than by adding adjectives

10 · 7 min

Writing a prompt that gets what you meant

THE ONE THING TO KEEP

Write the caption that would have sat under the picture you want, not a wish list of adjectives.

Your prompt is being matched, not understood

The model learned from images paired with text. Your prompt is pushed toward the region of that space where similar captions lived. So the useful question is never "how do I explain what I want". It is "what would the caption under this picture have said".

That one reframing fixes most beginner problems at once.

A rewrite

Weak: *a beautiful picture of a woman selling fruit, very detailed, 8k, masterpiece*

Nothing here locates an image. "Beautiful" sits under a billion different pictures. "8k" mostly appears in stock-photo spam. "Masterpiece" was genuinely useful in 2022, when some models were tuned on data tagged with aesthetic scores and the word pulled toward the higher-rated pile. In current models it does close to nothing, and people keep typing it out of habit.

Stronger: *A fruit seller in her fifties at a covered market in Kumasi, building a pyramid of oranges on a wooden crate. Morning light coming through a corrugated roof in stripes. Camera just above the fruit, looking slightly up at her. 35mm, shallow depth of field.*

Same idea. But now there is a who, doing what, where, in what light, from what angle, in what medium.

The five things worth naming

1. **Subject and action.** Verbs beat adjectives. "Building a pyramid of oranges" does more than "detailed".
2. **Setting**, specific enough to be a place rather than a category.
3. **Light.** The highest-leverage words you have. "Overcast noon", "a single bare bulb", "late sun through dust" give three completely different pictures of the same scene.
4. **Camera or medium.** 35mm, macro, ink line drawing, gouache, screen-printed poster. Name nothing and the model gives you the average of everything, which is a faintly plastic digital painting.
5. **Framing.** Close-up, wide, from below, over the shoulder.

Everything else is decoration.

Negation does not work the way you expect

"A street with no billboards" often returns billboards. The prompt is scored as a whole, the word "billboards" pulls billboard imagery in, and "no" is a weak signal pushing back. Models with larger text encoders handle this better than the 2022 generation, but better is not reliable.

The fix is always the same: say what should be there instead. "A street lined with plain shuttered storefronts and bare concrete walls." Some tools also give you a negative prompt field, which sets what guidance pushes away from. Use it for qualities — blurry, watermark, harsh flash — rather than for objects you are trying to keep out of a scene.

Counting and spatial relations are weak

"Five people" gives you four or six. "The red cup to the left of the blue one" swaps them. This is not a prompting mistake you can write your way out of; it is a real limit, and the next lesson explains where it comes from. Fix it in editing, or frame the shot so an exact count does not matter.

Iterating like an engineer

Most people change five things at once, get a worse image, then change five more. Instead:

- Lock the seed. Change one clause. Look. That is a controlled experiment, and it is the only way to learn what your words actually do.
- Generate four cheap, low-step images to find a composition, then rerun the one you want at full quality on the same seed.
- Keep a text file of prompts that worked. You will not remember, and prompts do not transfer cleanly between models — one tuned for a model trained on terse alt-text is mediocre in a model trained on long descriptive captions.

The honest ceiling

Prompt craft has a limit, and it is lower than the internet suggests. A prompt gets you a good starting frame. Getting a finished image — the right hands, the right sign, a client's product that actually looks like the product — happens in editing. That is the next lesson but one.

BEFORE YOU MOVE ON

Someone writes "a busy street with no advertising billboards" and keeps getting billboards. What is the best explanation?

A. The prompt is evaluated as a whole and "billboards" pulls billboard imagery in; the word "no" is a weak counterweight, not an instruction the model executes.

B. The model parsed the negation correctly, but nearly every street scene in the training data has billboards, so the prior overwhelms the instruction.

C. Negation is handled properly by models with the newer, larger text encoders, so the answer is simply to switch to one of those.

D. The prompt is too short; with more surrounding description the negation would have been respected.

11 · 8 min

Tag soup or sentences: why the advice reversed

THE ONE THING TO KEEP

Prompt style should match the captions a model was trained on, so keyword lists suit CLIP-era models and plain descriptive sentences suit models trained on machine-written captions.

Two pieces of contradictory advice, both correct

Search for prompting guidance and you will find two schools that flatly contradict each other. One says to write comma-separated keywords: `portrait, woman, red scarf, studio lighting, 85mm, shallow depth of field.` The other says to write a sentence: `A studio portrait of a woman wearing a red scarf, lit from the left, shot on an 85mm lens with the background thrown out of focus.`

Both are right, about different models, and the reason is the training captions.

A model trained on scraped alt text learned from text that looks like keyword lists, because that is what alt text is. Its encoder is good at recognising which concepts are present and poor at grammar. Feed it a sentence and the connective words consume tokens without adding steering.

A model trained on recaptioned data — where every training image was described in full sentences by a vision-language model — learned from prose. Feed it a keyword list and you have handed it text unlike anything in its training set. It will cope, but the spatial and relational vocabulary it was specifically trained on goes unused.

How to find out which you are holding, in two minutes

You do not need to read a paper. Run this test on a fixed seed:

```
A: red cube on the left, blue sphere on the right, white table
B: A red cube sits on the left side of a white table.
   A blue sphere sits on the right side of the same table.
```

If B places the objects correctly and A does not, you have a model that reads syntax and you should write sentences. If both come out the same and colours swap at random, you have a bag-of-concepts encoder and sentence structure is decoration.

The same two prompts, two encoders, eight fixed seeds

	Tag list	Full sentences
LIP-only encoder	**5 of 8** alt text looked like this	**4 of 8** the syntax is wasted
'5-based encoder	**5 of 8** unlike its training text	**8 of 8** the clause is actually read

The prompt is a red cube on the left, a blue sphere on the right, white table — written once as tags and once as two sentences. Both schools of advice are correct about different machines, and the test costs two generations.

A second test settles it further. Write a prompt with a clause that only means anything if grammar is being read: `a photograph of a dog that is not wearing a hat`. Bag-of-concepts encoders reliably produce a hat. Syntax-reading encoders often do not.

Style words are still tags, on every model

There is a part of the keyword tradition that survives the transition, and it is worth separating out. Words like `oil painting`, `linocut`, `Kodachrome`, `blueprint`, `medical illustration` are not describing the scene. They are naming a visual register that co-occurred with those words in captions. These work as tags on every architecture, because they are labels rather than relations.

So the practical shape of a good modern prompt is usually two parts: a sentence describing what is in the frame and where, followed by a short set of reg-

ister words. Mixing them is fine. Writing forty register words and no sentence is the old habit, and on a current model it produces a picture with a strong look and a vague subject.

The words that were never doing anything

A large amount of inherited prompt vocabulary is superstition. `8k`, `masterpiece`, `highly detailed`, `award-winning`, `trending on ArtStation`, `intricate` — these entered the culture because on 2022-era models they genuinely correlated with a certain sort of polished digital art in the caption corpus. On models trained since, most of them do very little, and some actively hurt by pulling toward that same over-rendered look.

The test is cheap and everyone should run it once. Fix the seed. Generate with the full prompt. Delete the trailing quality words. Generate again. Compare. On most current models you will find the two images are close, and on some the shorter prompt is better because it left the model free to interpret the subject rather than the aesthetic.

> *A prompt is not a spell. Every token you add competes for attention with every other token, and a word that contributes nothing is not free — it dilutes the words that were working.*

Where free tools help

You can do all of this without paying anyone. Free local interfaces — ComfyUI, AUTOMATIC1111's WebUI, Forge, InvokeAI, Fooocus — all support fixed seeds and batch comparison, and several hosted services offer enough free generations a day to run the tests above. On a phone, Draw Things on iOS runs small models locally for nothing. The point is not which tool. It is that any tool which lets you fix a seed and change one thing turns prompting from folklore into something you can check.

BEFORE YOU MOVE ON

A prompt that worked well for two years suddenly gives vague, over-rendered results on a newly released model. What is the most likely explanation?

A. The new model is larger, so it needs a correspondingly longer and more detailed prompt to have enough to work with.
B. The new model was trained on machine-written descriptive captions, so a keyword list no longer matches the text it learned from.
C. The new model applies a stricter safety filter that removes stylistic terms it associates with copyrighted artists' names before conditioning.
D. The prompt now exceeds the token limit, so the subject was truncated away.

12 · 8 min

What a negative prompt really does

THE ONE THING TO KEEP

A negative prompt replaces the empty prompt in the guidance calculation, so it defines what the image is pushed away from rather than issuing a prohibition, and it only exists on models that use classifier-free guidance.

Not a prohibition, a second anchor

The negative prompt box invites a wrong model of what is happening. People type `no extra fingers, no watermark, no blurry` and imagine a filter that removes those things.

Go back to the guidance arithmetic. Each step runs the model twice — once with your prompt, once with an empty one — and pushes the result along the difference:

```
guided = negative_prediction + scale * (positive_prediction -
negative_prediction)
```

The negative prompt simply fills in that second run. Instead of "where would the model go with no instruction", the second prediction becomes "where would the model go if asked for `blurry, watermark, low quality`". The image is then pushed *away from that point* and toward your prompt, by the guidance factor.

Three consequences follow, and each explains a common confusion.

Negative prompts have no effect at guidance 1. Look at the formula: at scale 1 the negative term cancels out entirely. On distilled models designed to run at guidance 1, or on models with no classifier-free guidance at all, the negative box does nothing whatsoever. Interfaces still show it, which is unhelpful.

A negative prompt costs you the same as a positive one. Both runs happen regardless, so a negative prompt is free in compute terms. But it is not free in effect — every token in it moves the anchor, including tokens you added carelessly.

Negation is not understood. Writing `no hands` in the negative box moves the anchor toward images that a caption-trained encoder associates with the word "hands". You are pushing away from hands, which is nearer to "hide the hands" than to "draw the hands correctly". This is why the fix for bad hands was never a negative prompt.

What actually belongs in there

Negative prompts work best when they name a *visual register you do not want*, because that is what encoders represent well.

Useful:

- `watermark, signature, text overlay, stock photo` — genuine visual patterns the model learned as a group.
- `cartoon, illustration, 3d render` when you want a photograph and the model keeps drifting toward illustration.

- `oversaturated, high contrast` as a partial counterweight to guidance burn.

Not useful:

- Counts. `two heads` in the negative does not enforce one head.
- Relations. `hand behind back` cannot be negated meaningfully.
- Anything specific to your image that has no visual signature of its own.

The enormous inherited negative prompts you see shared online — sixty terms including `deformed, mutated, poorly drawn, extra limbs, disfigured` — were tuned for one 2022 checkpoint and copied ever since. On most current models they measurably reduce quality by dragging the anchor into strange territory. Test yours by emptying it on a fixed seed. A surprising number of people find their images improve.

The related setting nobody explains

Some interfaces expose a separate control for how long the negative prompt applies — as a percentage of steps. This is more useful than it sounds. Composition is settled early; texture is settled late. A negative prompt aimed at texture problems, such as `plastic, waxy skin`, only needs to act in the last third of the run, and confining it there stops it from disturbing the composition.

If your tool offers it, the pattern is: composition-level negatives for the first 30% of steps, texture-level negatives for the last 40%, nothing in the middle. If it does not offer it, this is one of the reasons people move to node-based tools like ComfyUI, which expose scheduling for almost every parameter and cost nothing.

The honest limitation to hold onto: a negative prompt is a blunt instrument acting on the whole image at once. It cannot fix a local problem. When something is wrong in one region, the answer is a mask, which is the next module.

One last thing about the shared negative prompts you will meet. Many of them contain terms that are not visual at all — `bad anatomy`, `worst quality`, `ugly`. These entered circulation because one popular fine-tune was

trained on a booru-style dataset where those exact tags appeared as human quality ratings on the images. On that checkpoint they genuinely worked, since the model had learned the tag as a visual category. On a checkpoint trained on ordinary alt text they refer to nothing, and the anchor they move is essentially random. This is the general rule under all of it: a negative term only works if the model was trained on captions containing that term. When you inherit a prompt, the first question is what it was written for, and the second is whether that is what you are running.

BEFORE YOU MOVE ON

On a distilled four-step model that runs at guidance 1, a long negative prompt appears to have no effect at all. Why?

A. Distilled models discard the negative prompt because their reduced step count leaves no room to process a second conditioning input.

B. At guidance 1 the negative term cancels out of the guidance formula, so it has nothing to push away from.

C. The negative prompt is applied only after decoding, and distilled models decode with a different autoencoder that ignores it.

D. Four steps is too few for the negative anchor to accumulate any measurable influence on the latent.

13 · 8 min

Weighting, emphasis and where it breaks

THE ONE THING TO KEEP

Prompt weighting scales token embeddings before they reach the model, which is a blunt intervention on a representation the model never saw scaled during training.

The syntax, and what it is doing underneath

Most interfaces let you emphasise part of a prompt. The notations differ and the mechanism does not:

```
a (red:1.4) car          # common in WebUI, ComfyUI
a ((red)) car            # older shorthand, each bracket ~1.1x
a red car::2 sky::1      # Midjourney-style multi-prompt
weights
```

Underneath, the tool takes the vectors the text encoder produced for those tokens and multiplies them, or interpolates them toward or away from the vectors of an empty prompt. Then it hands the modified sequence to the diffusion model as if the encoder had produced it.

That last sentence is the whole lesson. The model was trained on encoder outputs, not on scaled encoder outputs. A weight of 1.2 produces a vector that is a little unusual and mostly behaves as you expect. A weight of 1.8 produces a vector unlike anything in training, and the model's response becomes unpredictable rather than merely stronger.

The observed behaviour

Push a weight up gradually and you see a consistent pattern:

- **1.0 to 1.3** — the concept becomes more present. This is the useful band.

- **1.3 to 1.6** — the concept starts to dominate the image, often spreading to places it does not belong. Weighting `red` this far will tint the whole frame.
- **Above 1.6** — texture breaks down. Colours posterise, edges harden, and sometimes the concept vanishes altogether, which surprises people who expected more of it.

Down-weighting is gentler and more reliable. `(background:0.6)` genuinely does calm a busy background, because scaling a vector toward the empty-prompt direction is closer to something the model has seen.

The practical rule most experienced users converge on: keep weights between 0.6 and 1.4, and if that is not enough, the problem is not emphasis. Either the concept is absent from the model, or it is being out-competed by something else in the prompt that you should delete instead.

Better tools than weighting

Weighting is popular because it is the only control many interfaces expose. Where the alternatives exist, they beat it.

Prompt editing over steps. Notation such as `[cat:dog:0.4]` swaps one token for another 40% of the way through the run. Since composition settles early and texture late, this gives you a dog-shaped animal with cat texture, or vice versa. It is far more precise than weighting and costs nothing extra.

Regional prompting. Assign different prompts to different areas of the canvas. The red cube genuinely goes on the left because the left region is conditioned only on the cube prompt. This solves attribute binding properly rather than compensating for it. Available free in ComfyUI and as extensions elsewhere.

Splitting into two generations. Generate the subject, generate the background, composite in GIMP, Krita or Photopea — all free. This is what a professional does when a job has to be right, and it is often quicker than twenty attempts at a prompt.

The failure that looks like a model problem

A specific trap: weighting interacts with guidance. Both are amplification, applied at different points. High guidance plus heavy weighting compounds, and the burnt, posterised result gets blamed on the checkpoint.

If your images look overcooked, take the weights out first, then adjust guidance, then put weights back only where they earned their place. Changing two amplifiers at once teaches you nothing about either.

There is also a difference between tools that matters if you follow instructions from elsewhere. Some implementations scale the token vector directly. Others interpolate between the prompt's embedding and the embedding of the prompt with that word removed, then renormalise the whole sequence so its overall magnitude is unchanged. The second approach is much better behaved at higher weights, because it never leaves the region the model was trained on. This is why the same `(red:1.6)` produces a mild change in one interface and a scorched image in another, and why weight values shared online are not comparable across tools. When you copy a prompt with weights in it, expect to retune the numbers, and do it on a fixed seed so you can see what each one did.

The honest limitation: weighting has no principled basis. It is a hack on an internal representation, discovered by users, adopted by tool authors, and never trained for. It works well enough within a narrow range and it will never be reliable outside it. Treating it as a precision instrument is the mistake.

BEFORE YOU MOVE ON

Raising a token's weight from 1.5 to 2.0 makes the concept less visible rather than more, and the image degrades. What does this show?

A. The interface caps prompt weights at 1.5 internally, and treats any value above that ceiling as an instruction to disable the token entirely rather than to emphasise it further.

B. The token was pushed past the encoder's vocabulary boundary, so it was replaced by an unknown-word placeholder.

C. Weighting scales an embedding into a region the model never saw during training, so its response stops being proportional.

D. The extra weight consumed the remaining token budget, truncating the rest of the prompt.

14 · 7 min

The token limit that eats long prompts

THE ONE THING TO KEEP

CLIP-based pipelines read about 75 tokens of prompt and silently discard or awkwardly average the rest, so a long prompt on an older model is mostly inert.

Seventy-five words that count

The CLIP text encoders used by the Stable Diffusion family accept exactly 77 token positions. Two are taken by start and end markers, leaving 75 for you. A token is roughly three-quarters of an English word, and less for unusual words, which get split into pieces.

So a prompt of about 55 to 60 ordinary English words fills the budget. Everything after that is, in the plain implementation, cut off and thrown away.

Nothing in the interface tells you. The generation succeeds. The image looks like your prompt, because the beginning of your prompt is the part that was read. People conclude the model ignored their later clauses out of preference. It did not receive them.

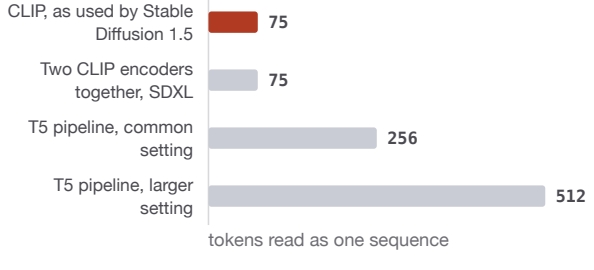
What tools do instead, which is not better

Most modern interfaces avoid the hard cut by chunking: they split the prompt into 75-token blocks, encode each separately, and concatenate or average the results. This is why you can paste a 300-word prompt into a WebUI and see it react to the end of it.

It is a workaround, not a fix, and it has a distinct signature. The chunks are encoded independently, so a clause split across a boundary loses its relation to the rest. `a woman holding a` in chunk one and `red umbrella` in chunk two do not compose. The tool usually shows the chunk count somewhere small — `75/150` means two chunks. When a long prompt behaves as if it half-understood you, look at that number.

Newer pipelines built on T5 or similar text models allow 256 or 512 tokens, and read them as one sequence. That is the real fix, and it is why long, careful, sentence-structured prompts became worth writing only recently.

How much prompt the encoder actually reads



A token is about three-quarters of an English word, so seventy-five tokens is fifty-five to sixty words. A two-hundred-word prompt on a CLIP model is mostly inert and nothing in the interface says so. The same words in Devanagari can exhaust the budget in fifteen.

How to find your limit in one minute

Append a token that would be unmistakable if read:

```
<your long prompt> , and a bright yellow rubber duck
```

Fix the seed. Generate. No duck means the tail is not reaching the model. Move the duck to the front and generate again to confirm that the model does render ducks. This test takes two generations and settles an argument people have for months.

Writing to a budget

Once you know you have 75 tokens, prompt writing becomes an editing problem, and editing is a skill you can actually improve at.

Cut, in this order:

1. **Quality incantations.** `masterpiece, best quality, 8k, highly detailed, award winning` is often 12 tokens doing almost nothing.
2. **Redundant synonyms.** `beautiful, gorgeous, stunning` are one concept in three costs.
3. **Articles and connectives**, but only on bag-of-concepts encoders. On a syntax-reading model these are load-bearing and cutting them is counterproductive.
4. **Anything you could achieve with a setting.** Aspect ratio, resolution and colour grade belong in parameters or post-processing, not in tokens.

What survives should be: the subject, its important attributes, the setting, the light, and the register. That fits comfortably.

The related limit on the other side

There is a mirror-image problem worth naming. Prompts that are too short leave the model to fill in everything, and what it fills in is the training-set average — the most typical face, the most typical lighting, the most typical composition. A three-word prompt does not give you a neutral image; it gives you the centre of the distribution, which is why every "photo of a woman" from an unmodified model looks like the same woman.

Between the two limits there is a band, roughly 20 to 60 tokens on a CLIP model and rather wider on a T5 one, where the prompt is specific enough to escape the average and short enough to be read. Almost all good prompts live there. It is worth counting your tokens once, on the model you use most, so you know where the edges are rather than guessing.

One consequence of the limit is worth stating for anyone working outside English. Tokenisers are trained on text corpora that are heavily English, so an English word is often a single token while the same word in Hindi, Bengali, Tamil or Arabic is split into five or six pieces. A prompt in Devanagari can exhaust a 75-token budget in fifteen words. This is a real disadvantage and it is not the writer's fault. The workable route today is to prompt in English for the scene and use references or a fine-tune for whatever the English caption vocabulary does not cover, which is an unsatisfying answer and an honest one. It will improve as models are trained on more multilingual captions, and it has improved already; it is not yet solved.

BEFORE YOU MOVE ON

A 200-word prompt on a CLIP-based pipeline produces an image matching only the first third of the request, but the interface shows no error. What is happening?

A. The generation timed out partway through conditioning and completed using whatever had been encoded so far.

B. The remaining tokens fell outside the encoder's 77-position context and were truncated or chunked without warning.

C. The guidance scale was too low for the later clauses to take effect on the image.

D. The later clauses described concepts the model has no training coverage for, so they produced no visible change in the output.

15 · 8 min

When a reference image beats any sentence

THE ONE THING TO KEEP

Conditioning on an image passes thousands of numbers of visual information where a prompt passes tens, so anything you can show is better shown than described.

The bandwidth argument

A 75-token prompt carries, at most, 75 vectors of a few hundred numbers each, and much of that capacity is spent on saying which objects exist. An image encoder converts a reference picture into a comparable set of vectors — but those vectors describe an actual arrangement of colour, light, texture and geometry rather than a verbal gesture at one.

This is why the single largest improvement most people can make to their results is not a better prompt. It is providing a picture.

There are several distinct ways to do it, and confusing them causes most of the frustration.

Four different things called "reference"

Style reference. An adapter encodes the reference image and injects it alongside the text conditioning, so the output takes on the reference's palette, texture and rendering without copying its content. This is what IP-Adapter and the various "style reference" or "sref" features do. Good for: matching an existing set of illustrations, keeping a brand look.

Structure reference. The reference is converted into an intermediate map — edges, depth, human pose, a rough segmentation — and the model is conditioned on that map at every step. This is ControlNet and its relatives. Good for: keeping a composition, redrawing a rough sketch, putting a new subject

into an existing layout. Content and style come from your prompt; only the geometry comes from the reference.

Image-to-image. The reference is encoded into a latent, noised part-way, and denoised from there. The output stays close to the original in proportion to how little noise was added. Good for: variations, a repaint, a change of medium.

Subject reference. The reference supplies a specific identity — a face, a product, a character — which the model then places in new scenes. This is the hardest of the four and still the least reliable.

Choosing the wrong one produces the classic complaint that "the reference did nothing". A style adapter will not preserve the composition. A depth map will not preserve the colour. Neither will preserve a face.

The practical routine

For most jobs the fastest path is not a better description but this:

1. Find or make an image with the *composition* you want. A phone photograph, a scribble in GIMP, a stock image, anything.
2. Extract a depth map or edge map from it. Free tools do this in one click and it takes under a second.
3. Prompt for the *content* you want, conditioned on that map.

The composition is now yours, exactly, and the prompt only has to carry subject and style — which is comfortably within 75 tokens.

Anyone who has spent an afternoon writing "the woman is on the left, slightly behind the table, looking away from the camera" and getting it wrong every time will find this a relief. Spatial language is exactly what these encoders are worst at, and it is exactly what an image conveys for free.

What you owe the source

A reference image is somebody's work unless you made it. Using a photograph as a depth map extracts geometry, which is a weaker taking than reproducing

the picture — but "weaker" is not "none", and the law here differs by country and is unsettled, which the final module covers properly. Two habits are worth adopting regardless of what any court eventually decides.

Keep a record of where every reference came from. If a client later asks, "is anything in this derived from someone else's image", you want to be able to answer precisely rather than from memory.

And prefer references you own or that are openly licensed. Your own photographs, images from Unsplash, Pexels or Wikimedia Commons under their stated terms, or a rough sketch you drew yourself in Krita. A sketch you drew badly works as well as a photograph for structure conditioning, because the model only reads the edges.

The honest limitation: structural conditioning constrains geometry, not meaning. A depth map of a person will get you a person-shaped thing in that pose. It will not stop the model producing four fingers, and it will not carry an expression. Reference images make composition solvable. They do not make everything solvable.

There is one more caution, about strength. Every reference method has a weight, and every one of them fails in the same two ways at the ends of its range. Too strong, and the output is the reference with a light coat of paint — recognisably the source, which is the case most likely to cause a rights problem as well as a boring image. Too weak, and you have paid for a conditioning step that did nothing. The useful setting is nearly always somewhere in the middle third, and the only way to find it for a given model is a short sweep with everything else fixed. Do that sweep once per method and write the number down.

BEFORE YOU MOVE ON

You want a generated illustration to sit in an exact composition you have sketched, but in a completely different visual style. Which conditioning method fits?

A. An image-to-image pass at high denoise strength, which preserves the layout while allowing the style to change freely.
B. A style-reference adapter fed with your sketch, so the model takes the arrangement from the reference.
C. A structural conditioning map — edges or depth — taken from the sketch, with style coming from the prompt.
D. A subject-reference method, since the sketch defines the subject that must appear in the output.

16 · 7 min

Reading the metadata, keeping the record

THE ONE THING TO KEEP

Generation parameters are usually written into the output file, so a result is recoverable and auditable if you keep the file and know where to look.

The parameters are in the file

Most local generation tools write the full set of parameters into the PNG itself, in a text chunk that survives copying and uploading to some services. Open a generated PNG in a text editor and you will often find something like:

```
a studio portrait of a woman wearing a red scarf
Negative prompt: watermark, text
Steps: 28, Sampler: DPM++ 2M Karras, CFG scale: 6,
Seed: 3819475210, Size: 1024x1024,
Model hash: 31e35c80fc, Model: sd_xl_base_1.0
```

That block is a complete recipe. With the same model file, it reproduces the image. Free tools read it for you: most WebUIs have a "send to" button that repopulates every field from a dropped image, and ComfyUI embeds the entire node graph, so dragging a PNG onto the canvas rebuilds the workflow that made it.

This is the single most useful habit in this module. When something works, you have the record already. You just have to not destroy it.

How the record gets destroyed

- **Saving as JPEG.** The text chunk is a PNG feature. Export to JPEG and it is gone.
- **Uploading to most social platforms.** They strip metadata on upload, partly for privacy, partly to reduce file size.
- **Editing in an image editor and re-saving.** Some preserve unknown chunks, many do not.
- **Screenshotting.** Obviously, but people do it.

So keep the original PNG. Treat exported versions as disposable and the generation output as the master, exactly as a photographer keeps the raw file.

What metadata does not prove

There is a temptation to treat the presence or absence of this metadata as evidence about whether an image is generated. It is not, in either direction.

Absence proves nothing: any conversion, upload or crop removes it, and the vast majority of generated images circulating online have none.

Presence proves very little: the chunk is plain text that anyone can write into any file. Adding a fake generation record to a photograph takes one command. It is not signed, not verified and not tamper-evident.

Provenance that resists tampering requires cryptographic signatures, which is a different system entirely and gets its own module later in the course. The generation metadata is a working convenience for you, not evidence for anybody else.

Keeping a record that survives you

For anything that matters — client work, anything published, anything you might have to explain — the parameters in the file are not enough on their own. Keep a short note alongside the deliverable with:

- Which model, at which version or hash. "Stable Diffusion" is not an answer; checkpoints differ enormously and are frequently updated in place.
- The prompt and seed, so the result can be regenerated.
- Every reference image used, with its source and licence.
- What you did afterwards by hand — the retouching, the compositing, the type.
- Whether the output was disclosed as AI-assisted, and where.

This takes two minutes per job and it answers, without argument, the questions that come up months later: was this generated, what went into it, can we make another one that matches, and can we prove what we did.

The version problem, stated plainly

There is a limitation here that no amount of discipline fixes. Hosted models change under you. A service can update its model with no version number and no announcement, and your recorded prompt and seed will then produce a different image. Nothing in your record is wrong; the machine it referred to no longer exists.

This is the strongest practical argument for keeping local copies of any model your work depends on. A 6-gigabyte checkpoint file on a disk is a version that cannot be revised without your consent, and open-weight models can be archived exactly like fonts or software. If a series of images has to match across two years, this is not a preference. It is the only way it works.

Two small habits make the rest of it painless. Name files so the record is in the name — `campaign_hero_seed3819475210_v3.png` tells you more at a glance than any folder structure. And keep the failures, at least for the duration of a project. The rejected variants are the evidence of what you tried, they are frequently the answer when a client changes their mind back, and on a job that

later gets questioned they show a process rather than a single lucky output. Disk is cheap; regenerating an image whose model has since been updated is not possible at any price.

BEFORE YOU MOVE ON

Someone shows you a JPEG with an embedded text block naming a prompt, seed and model, and argues this proves the image was AI-generated. What is wrong with the argument?

A. The text block is unsigned plain text that anyone can add to any file, so its presence is not evidence of anything.

B. Generation metadata is only ever written into PNG files, so a JPEG carrying such a block must have been fabricated deliberately by someone attempting to mislead.

C. The block would have been stripped during JPEG compression, so what is visible must be a rendering artefact.

D. Seeds are not portable between machines, so the recorded seed could not reproduce the image and the record is therefore invalid.

---

17 · 8 min

The prompt you sent is not the prompt it got

THE ONE THING TO KEEP

Hosted image services rewrite, expand and filter prompts before the model sees them, which explains results that do not match the words you typed and makes reproducibility a property of the service rather than of your record.

An invisible layer

When you type into a hosted image service, your text does not usually go straight to the model. Several things happen to it first, and none of them are

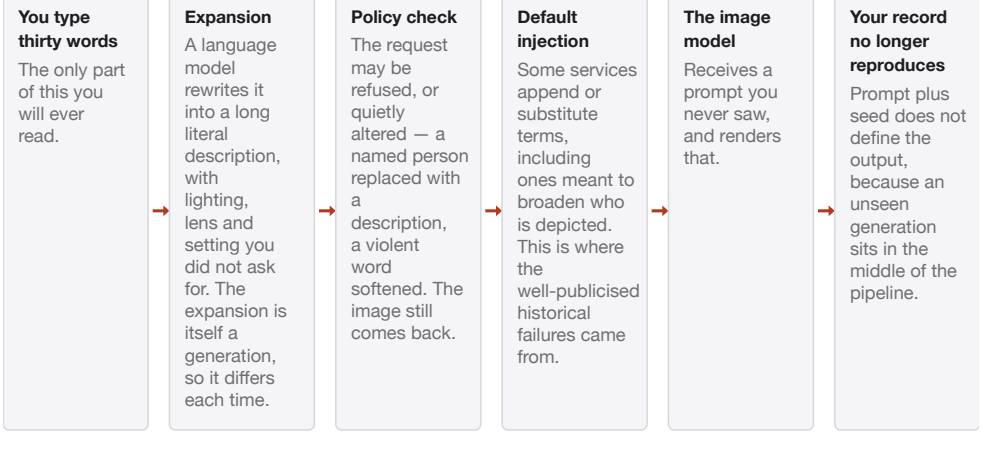
shown to you.

Expansion. A language model reads your short request and writes a longer, more detailed description for the image model to render. This is why a four-word prompt on some services returns a richly specified image with lighting, lens and setting you never asked for. It is also why the same four words give a different scene each time even at a fixed seed, if the service exposes seeds at all: the expansion is itself a generation.

Safety filtering and rewriting. Prompts are checked against policy. A prompt may be refused, or quietly modified — a named public figure replaced with a description, a violent word softened. The image comes back. Nothing says it answered a different question.

Diversity and default injection. Some services append or substitute terms intended to broaden the range of people depicted, because an unmodified model returns the training-set average. This is a defensible aim with a clumsy mechanism, and it has produced well-publicised failures where the injected terms were applied to requests where they made no sense, including historical ones. The lesson is not that the aim was wrong. It is that the intervention happened at the prompt, invisibly, where the user could neither see it nor correct it.

What a hosted service does to your prompt before the model sees it



None of this is visible and all of it is detectable: ask for an empty room with a single grey box and nothing else. If furniture, plants and a window come back, something expanded your request.

How to detect it

You cannot read the rewritten prompt on most services. You can detect that rewriting is happening:

- **Ask for something structurally odd** — an empty room with a single grey box, nothing else. If the result contains furniture, plants and a window, something expanded your request.
- **Repeat the identical prompt several times.** Wide variation in scene content, as opposed to variation in the same scene, points at an upstream generator.
- **Give a prompt that contradicts itself mildly.** Expansion layers tend to resolve contradictions into something coherent; a raw model tends to produce the contradiction.

Some services will tell you the final prompt if you ask in the same conversation, and some show it in a details panel. Where it is available, look at it once. It is usually instructive about how much of your result was your idea.

Why this matters beyond curiosity

Three practical consequences.

Reproducibility is not yours. If the pipeline includes a rewriting step you cannot see or pin, then your prompt plus seed does not define the output. For work that must be repeatable — a series, a brand, anything a client will ask you to match — this is a real limitation of hosted services, and it is the main reason production studios keep local models for the shots that must be consistent.

Attribution of failure goes wrong. People conclude that a model "cannot" do something when it was the filter that refused, or that a model is "biased toward" something when it was the injection. Diagnosing anything requires knowing which layer you are talking to.

Disclosure gets complicated. If you are telling a client what went into a piece of work, "I wrote this prompt" is not quite true when an unseen model

wrote most of the operative description. It rarely matters. It occasionally does, and it is better to know.

The free alternative, and its price

Running a model locally removes every one of these layers. Your prompt is the prompt. The seed reproduces. Nothing is injected, nothing is silently refused, and the record you keep is complete. ComfyUI, the various WebUIs, InvokeAI, Fooocus and Draw Things are all free, and small models run on modest hardware.

The price is honest to state. You maintain it. You choose the model, the sampler, the guidance and the upscaler yourself, and nothing helps you when a result is bad. The removal of the safety layer is also the removal of the safety layer: what you generate is entirely your responsibility, legally and ethically, and the law does not care that no filter warned you.

Most people end up using both — hosted services for speed and for the things the big models do best, local models for anything that must be reproducible, private or consistent. Knowing which one you are on, and what it is doing to your words, is the point of this lesson.

BEFORE YOU MOVE ON

A hosted service returns wildly different scenes for the identical short prompt on repeated runs, while a local model with a fixed seed returns the same image every time. What does this most likely indicate?

A. The hosted service uses ancestral sampling, which cannot converge on a single result across runs.

B. The hosted model is far larger, and larger models are inherently less deterministic in their output.

C. An upstream layer is expanding your short prompt into a different detailed description on each request.

D. The hosted service caches results and returns previously generated images from other users matching the same prompt.

18 · 7 min

A prompt is not an asset

THE ONE THING TO KEEP

Prompts are tightly coupled to one model version and transfer badly, so the durable asset is the reference material, the workflow and the record, not the words.

The marketplace problem

There is a small industry selling prompts. Packs of a hundred prompts for a fee, prompt marketplaces, prompt libraries as a subscription. It is worth understanding why almost all of it is worthless, because the reasoning also tells you what to invest your own effort in.

A prompt only means something in combination with a specific model, at a specific version, with specific sampler and guidance settings. Move any of those and the output changes, often completely. The models change every few months. A prompt pack is therefore a set of instructions for a machine that will not exist next year, sold without the machine.

There is a second problem. The good prompts in any pack are the obvious ones — describe the subject, name the register, state the light — and you can derive them from the last five lessons in an afternoon. The rest is the quality-incantation vocabulary that stopped working two model generations ago.

What does transfer

Some things survive a model change, and these are worth building deliberately.

Reference images. A depth map, a pose, a colour reference, a sketch. These are inputs to a process, not instructions to a particular model, and they work on whatever comes next.

Workflows. The sequence — generate at native size, inpaint the face, upscale, colour-correct in a free editor, composite the type — is a method. Node graphs in ComfyUI are files you can keep, share and adapt when the model underneath changes.

Fine-tunes you trained. A small LoRA trained on your own material encodes something the prompt cannot. It is tied to a base model, so it also expires, but it holds information no sentence can carry.

Judgement about what is wrong. This is the real asset and the reason for this whole module. Someone who can look at a failed image and say "that is guidance, not prompt" or "that is the latent resolution, no wording will fix it" is not dependent on any particular tool.

Version your own prompts anyway

Within a project, prompts are worth keeping properly. Not as a mystical library, but the way you would keep any working configuration.

A plain text file next to the project, one block per shot, each with model, seed and parameters, is enough. If you use version control for anything else, put it there; a prompt is text and diffs well. When a client asks for "the same but with a blue background six months later", the difference between a two-minute job and an afternoon of guessing is entirely whether this file exists.

The uncomfortable part

There is a related claim worth being honest about, because people building a business on prompting need to hear it: prompting is not, on its own, a durable skill. Every model release makes prompts easier to write and less necessary. Systems that rewrite your request internally have already removed much of the craft, and that direction is not reversing.

What does not get automated away is everything around the prompt. Knowing what image the job actually needs. Composition and typography. Retouching and compositing. Judging when a result is nearly right and what specifically is wrong. Managing rights and disclosure. Those are the reasons a client pays

somebody, and they are all covered elsewhere in this track — design, image editing, vector work, storytelling, and the business of it.

> *The prompt is the cheapest part of the process and the part most likely to be obsolete. Put your effort into the parts that were already valuable before any of this existed.*

That is not a discouraging conclusion. It means the ceiling on this work is set by ordinary craft rather than by access to a particular tool, and ordinary craft is learnable by anyone with time, which is the only genuinely fair distribution of ability there has ever been.

There is one narrow exception worth naming, because it is where prompt knowledge does hold value. Inside an organisation, a prompt that has been tested against a house style, checked for the failure modes that matter to that business, and recorded with its model version is a piece of operational documentation. It is valuable for the same reason a tested configuration file is valuable: somebody did the work of finding out what breaks. That value is local, it decays with the next model, and it belongs in the project's repository rather than for sale. Notice the difference from a prompt pack. What is worth something is the testing, not the sentence.

BEFORE YOU MOVE ON

Why does a prompt bought in a prompt pack usually fail to reproduce the sample image shown with it?

A. Prompt packs deliberately withhold key terms so buyers return for more, which is why the sample cannot be matched from the text supplied to them.

B. A prompt's effect is defined by a particular model version and settings, none of which travel with the text.

C. The sample images are rendered at higher resolution than a normal account is permitted to generate.

D. Prompt text loses meaning when copied between interfaces because each tool tokenises punctuation differently.

CASE STUDY · MODULE 2

The second batch did not match the first

A two-person knitwear export business in Tiruppur, six months after a set of lookbook images that can no longer be reproduced

Kavitha Anandh and her brother run a knitwear unit in Tiruppur with nine machines and one important customer: a German buyer who takes about seventy per cent of what they make. In March the buyer asked for lifestyle images for a lookbook. Not product shots on white — garments worn in a room, the same room every time, the same light, so that the set read as one place across a season.

Neither of them had a studio or a budget for one. Kavitha built the room on a hosted image service over four evenings and produced thirty images. The buyer was pleased enough to ask for thirty more in September, for the second half of the range.

In September she opened the text file where she had saved every prompt, pasted the first one in exactly as written, and got a different apartment. Warmer light. Different furniture. A window on the wrong wall. She tried the next one, and the next. Nothing matched.

Her brother's diagnosis was that the service had got worse, which is the diagnosis everybody reaches first. Kavitha's was that she did not actually know what the service had been doing in March, and she ran two tests to find out.

The first: she re-ran one March prompt six times without changing a character. She did not get six variations of one apartment. She got six different apartments, with different furniture and different architecture. Variation within a scene is what a seed change produces. Variation of the scene itself points at something upstream deciding what the scene is.

The second test was the one that settled it. She asked for an empty room, bare walls, a single grey box on the floor, nothing else. She got a styled living room with a plant, a rug and afternoon light through a window. Nothing she had typed asked for any of that. A language model was reading her short request, writing a longer and more detailed description, and handing that to the image

model. Her words were an input to that expansion, not the instruction the image model received. And the expansion was itself a generation, which is why identical text produced different rooms.

That left three ways forward, with the buyer's deadline three weeks out.

The first was to rent a flat in Coimbatore, hire a photographer and models, and shoot the second thirty properly. Roughly a lakh and a half, three days, and it does not solve the problem — the March images would still not match the September ones, so the set would be split down the middle whichever way she went.

The second was to accept the mismatch and present it to the buyer as a new season look. Cheapest, fastest, and it asks a customer who is seventy per cent of the business to absorb an inconsistency they did not ask for.

The third was to rebuild locally. A second-hand machine with twelve gigabytes of video memory, a free node-based interface, an open checkpoint archived onto a disk she controlled, and then redo all sixty images so the whole set came from one place. That meant spending money on hardware, losing days of a three-week window to setting it up, and accepting that a mid-sized open model would not be quite as polished out of the box as the hosted service had been.

She chose the third, and her reasoning was not about this season. It was that the buyer would ask again in March, and in September after that, and a pipeline that cannot be re-run is not a pipeline.

STOP HERE

Before turning the page: what would you have done, and what would it have cost if you were wrong? Write it down. The point of the next section is to compare.

WHAT HAPPENED

The machine cost forty-two thousand rupees. The evening she had budgeted for installation became two.

Her first local results were worse than the hosted ones, and badly so: waxy skin, hard black outlines, colours pushed to the edge of the range. She assumed the open checkpoint was simply inferior. It was not. She had carried her habits across, including a guidance scale of seven, and the model card for the checkpoint specified a useful band of three to four because the model bakes guidance in during training. Everything she had generated had been extrapolated far outside the range the decoder could render. Dropping the number to three and a half fixed it in one generation.

The second thing she carried across was her prompt style — long comma-separated keyword strings, written for a service whose behaviour she had never actually diagnosed. The local checkpoint used a text encoder trained on written-out descriptions, and it responded much better to plain sentences describing what was in the frame. Her prompts got shorter and the results got more accurate.

Then she did the thing that made the rest work. She fixed one seed per room, saved the node graph as a file, kept the checkpoint on two disks, and photographed the corner of her own office on her phone to use as a depth map for the composition. Sixty images took four days.

The buyer noticed that the set was more consistent than the spring one, not less. The following season Kavitha reused the same room seeds and the same depth map and changed only the garments.

What she keeps now is the checkpoint file, the node graph, the phone photographs and a text file of parameters with a date against each. The prompts themselves turned out to be the least durable thing in the folder.

She pasted the March prompt in character for character and got a different room. What does that tell you about where reproducibility actually lives on a hosted service?

It tells her the prompt is not the input to the image model. A hosted service can expand, rewrite and filter a request before the model sees it, and the expansion is itself a generative step, so identical text produces a different operative description each time. On a pipeline like that, the prompt and the seed together do not define the output, and neither does any record she keeps. Reproducibility is a property the service either provides or does not, and it can be withdrawn silently by a model

update. That is the argument for keeping a local copy of any model a body of work depends on.

Why was 'an empty room, bare walls, a single grey box, nothing else' a good test prompt, where 'a woman in a red coat' would have taught her nothing?

Because it is structurally sparse and the expansion is visible against it. A detailed subject prompt returns a detailed image whether or not anything was added, so there is nothing to compare. A deliberately under-specified, slightly odd request has an obvious correct answer — an almost empty frame — and every plant, rug and window that comes back is something a layer she cannot see decided to add. The test works by leaving a gap and seeing who fills it.

Her first local images looked burnt and waxy and she blamed the checkpoint. What was the actual cause, and what is the general habit that would have caught it faster?

Guidance. She was running at a scale of seven on a model whose card specified three to four, so the guided prediction was being extrapolated outside the range the decoder was trained to render, which is what produces clipped highlights, posterised colour and plastic skin. The general habit is to read the three numbers the model's author published — native resolution, step count and guidance scale — before forming any opinion of a checkpoint. Guidance values are not comparable across model families, and advice written for one is actively wrong for another.

MODULE 2 · TEST

Check what stuck

6 questions, one right answer each. This one is for you, not for a record: answers and the reason behind each are at the back. If two questions from the same module go wrong, re-read that module before the exam rather than after.

1 A colleague fills the negative prompt box on a distilled model that runs at guidance 1 and reports that it makes no difference at all. Why not?

A. Distilled models strip the text encoder, so neither prompt box is read.

B. The negative term cancels out of the guidance formula when the scale is one.

C. The negative prompt is only applied during the final third of the steps, and there are too few steps.

D. Negative prompts require a separate conditioning path that distilled models remove during training.

2 You append 'and a bright yellow rubber duck' to the end of a long prompt, fix the seed and generate. No duck appears and nothing else changes. What have you established?

A. That guidance is too low for the model to follow late clauses.

B. That the model has no coverage for rubber ducks.

C. That the tail of the prompt is not reaching the model.

D. That the seed is dominating the composition and overriding the prompt.

3 Weighting a colour at 1.2 makes it more present. At 1.8 the colour sometimes disappears entirely. What is going on?

A. Weights above 1.5 are clamped by most interfaces and then inverted.

B. The token is dropped once its vector exceeds the encoder's normalisation range.

C. High weights consume proportionally more of the token budget, pushing the colour out of the prompt.

D. A heavily scaled vector is unlike anything the model saw in training, so its response is unpredictable rather than stronger.

4 Why does a depth map extracted from a rough sketch place two objects more reliably than the sentence 'the red mug is to the left of the blue one'?

A. The map is spatially aligned with the latent, so each cell is told directly what belongs there.

B. The map is encoded into far more tokens than a sentence, so it dominates the conditioning.

C. The map raises the guidance scale locally in the regions it covers.

D. The map is processed by a larger text encoder that can represent spatial relations.

5 You suspect a hosted service is rewriting your prompt before the model sees it. Which test would show it most clearly?

A. Generate the same prompt at three different guidance scales and compare saturation.

B. Open the returned PNG in a text editor and look for the parameter block.

C. Ask for an empty room with a single grey box and nothing else.

D. Generate twice with the same seed and check whether the two images are identical.

6 Your studio is moving from one base checkpoint to a newer one. Which part of your existing work is most likely to still do its job unchanged?

A. The list of seeds that gave good framing on the old model.

B. The sixty-term negative prompt block copied from a forum.

C. The depth maps and pose references built for each shot.

D. The weighting values tuned over several months of work.

MODULE 3

Where it breaks, and why

Every generative failure has a mechanism, and knowing the mechanism tells you whether a prompt can fix it. This block works through the durable ones: counting and attribute binding, negation, spatial relations, the pull toward the training-set average, the bias that pull carries, physics and reflections, detail at small scale, refusals and false positives from safety filters, and the visible tells that used to identify a generated image and are now disappearing.

> BY THE END OF THIS MODULE YOU CAN
>
> Look at a failed generation and classify the failure by mechanism — binding, negation, resolution, distribution pull, physics or filtering — and state for each whether a prompt change, a structural control, a different model or a manual fix is the only thing that will work

19 · 7 min

What hands and text told us

> THE ONE THING TO KEEP
>
> These models nail local texture and fail global constraints — anything you must count, spell, or match across a picture.

Two famous failures

In 2022, hands. Six fingers, thumbs branching off thumbs, a knuckle where a nail should be. And shop signs reading RESTAVRANT or BAKREY — letter-

shaped marks arranged with total confidence into nothing.

Both are largely fixed. Understanding *why* they broke is worth much more than the fix, because it tells you what else will break.

Hands

A hand is small in the frame, has roughly 27 degrees of freedom, appears in thousands of configurations, and is constantly blurred, cropped or half-hidden in training photos. And no caption ever says "with five fingers", because no human writes that.

So what could the model learn? Finger-ish texture. Skin, knuckle, nail, the taper of a fingertip. What it could not learn from that training objective is a *count*, because a count is never locally visible. Standing at any single patch of the image, four fingers and six fingers look equally plausible. Local plausibility is all the objective ever rewarded.

Text

Spelling is worse, because it is an exact, ordered, discrete constraint stretched across a distance. The older text encoders made it close to impossible: they chop your prompt into subword tokens and hand the image model a semantic vector. There is no character-level information in there. The model was asked to render BAKERY while genuinely not being told which letters that word contains. Letter-shaped marks were the honest best guess.

What actually fixed them

Not a clever trick. Four unglamorous things:

- Bigger, character-aware text encoders that can see the actual string.
- Higher working resolution, so a hand occupies enough of the compressed space to be represented at all.
- Recaptioning the training data with vision models, so captions describe what is in the picture rather than whatever alt-text a website happened to have.

- Scale.

The pattern worth keeping

These models are extremely good at **locally plausible texture** and weak wherever correctness is a **global, discrete, exact constraint** — something you can only verify by taking in the whole image and counting or comparing.

That predicts failures you have not hit yet:

- Exact quantities. Nine coins, three windows, a hand of five cards.
- Reflections and shadows that must agree with the scene rather than merely look like reflections.
- Clock faces, calendars, dice, sheet music, chessboards, wiring diagrams.
- Anything repeated identically: a logo appearing twice in one image, a tiled pattern, the same character in two separate pictures.
- Maps, and writing in scripts thin on the ground in training data. Devanagari, Arabic and Thai still fail in models that spell English perfectly, for exactly the reason English used to fail.

What to do about it

Stop trying to prompt your way through a global constraint.

- Put real text on afterwards, in a real font, in a design tool. Nobody in production generates a poster headline.
- Generate the countable thing separately, or crop so it sits out of frame.
- Use masked editing and structural conditioning for the parts that must be exact. That is the next lesson.
- For a character who must recur, use the model's reference or consistency features rather than describing the person again and hoping.

One more use for this

The same reasoning applies to spotting synthetic images. As models improve, the tells stop being fingers and become global consistency: a reflection that

does not match the room, shadows at two different angles, earrings that change between two shots of the same person, background text that reads fine from a distance and dissolves up close.

Fingers were never the point. Counting was.

BEFORE YOU MOVE ON

A new image model spells shop signs perfectly and draws flawless hands. Which task should you still expect it to struggle with?

A. A macro shot of tree bark under unusual light.

B. A hand gripping a mug at an awkward angle.

C. A photorealistic portrait of a person who does not exist.

D. A two-panel illustration where the same nine coins appear in both panels, arranged differently.

---

20 · 8 min

Counting, and the objects that swap their attributes

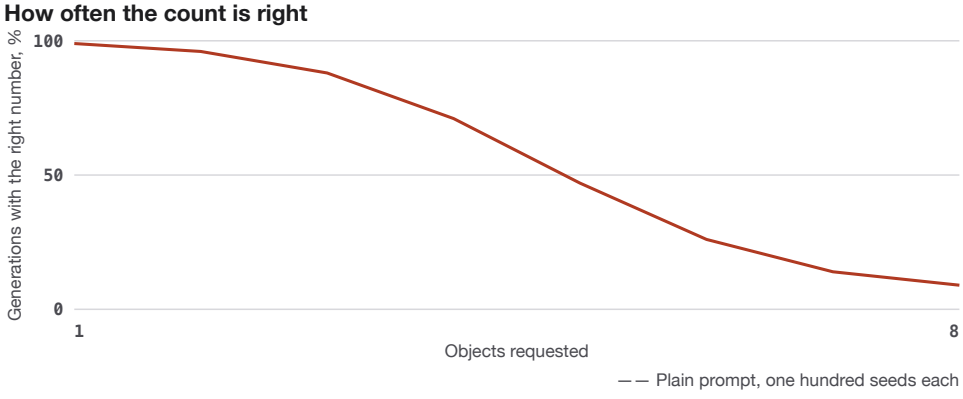
THE ONE THING TO KEEP

Nothing in the architecture ties an adjective to a noun or enforces a quantity, so counts and attribute assignments are emergent rather than guaranteed and degrade sharply above three or four objects.

Ask for five and get four, or six

Request `five apples on a plate` and count what comes back. One, two and three are usually right. Four is usually right. Five is a coin toss. Beyond six, the model produces "a lot of apples" and the exact number is essentially random.

This is not a shortage of training images of five apples. It is that nothing in the generation process counts.



Nothing in the pipeline counts. Each latent cell settles into something plausible for its neighbourhood and no register anywhere holds three placed, two to go. Past four the number is close to whatever the local dynamics produced, and repeating it in the prompt does not touch it.

Recall how the image forms. Each cell of the latent grid attends to the prompt and settles into something plausible for its neighbourhood. A cell surrounded by apple-ish cells becomes more apple. There is no global register holding "we have placed three so far, two to go". The number that emerges is whatever the local dynamics produced, and the prompt influences it only as a general pressure toward apple-density.

Models trained on machine-written captions do better, because their captions actually said "five apples" while alt text rarely did. Better is not solved. Published evaluations of current models still show accuracy falling off sharply past four objects, and the same model that renders skin convincingly will hand you six fingers on a hand it drew perfectly otherwise.

The same mechanism, different symptom

Attribute binding is the counting failure wearing another coat. `a woman in a green coat next to a man in a red coat` returns, with real frequency, both coats red, both green, or the colours swapped.

The token vectors for `green`, `coat`, `red` and `man` are all available everywhere. A latent region turning into a coat attends to `coat` strongly, and then

must pick a colour from two candidates neither of which is bound to it. Whichever has slightly higher affinity in that region wins, and the winner is not stable across seeds.

Failure gets worse as the number of attributed objects rises, for the same reason counting does, and it gets worse when the objects are visually similar. Two coats compete far more than a coat and a teapot.

What actually fixes it

In rough order of effort:

1. **Separate clauses and full stops.** On a syntax-reading model this genuinely helps: `A man wears a red coat. A woman wears a green coat.` gives the encoder a structure to bind within.
2. **Regional prompting.** Condition the left region on one prompt and the right region on another. Now nothing competes, because the colour words are not present in the other region's conditioning at all. Free in ComfyUI and available as extensions in most WebUIs.
3. **Generate separately and composite.** Two images, one subject each, combined in GIMP, Krita or Photopea. For a count, this is often the only reliable route: generate one apple, duplicate it five times, vary the rotation and lighting slightly.
4. **Inpaint the count.** Generate four apples, mask an empty area, inpaint a fifth. The model only has to add one object, which it can do.

What does not work: repeating the number, weighting the number, writing `exactly five`, `5 apples (five:1.5)`, or putting `four apples, six apples` in the negative prompt. All of these are attempts to argue with a machine that has no counter, and every one of them is a thing people spend an afternoon trying.

The professional consequence

If a job specifies a count — three product variants in a row, a family of four, five icons — do not plan to prompt your way there. Plan to generate the elements and assemble them. This is not a workaround forced on you by an im-

mature technology; it is how commercial illustration has always been produced. The generated element is a component, and the composition is your work.

There is an honest ceiling here worth stating. Some models have improved counting markedly by training on synthetic images with programmatically correct captions. It is a genuine gain and it is bounded: the improvement is on the kinds of arrangement the synthetic data contained. A count in an unusual arrangement — five people around a table, three specific tools in a row — still fails, because the fix was more training data of a certain shape rather than a mechanism for counting. Watch for this pattern generally. A capability that arrives through targeted data works exactly as far as the data went.

BEFORE YOU MOVE ON

A prompt asking for six identical bottles reliably returns five or seven, on every seed and every model tried. Which response follows from the mechanism?

A. Generate one bottle and assemble the row in an image editor, because nothing in the generation process maintains a count.

B. Add "exactly six, six bottles, 6" and put "five bottles, seven bottles" in the negative prompt to constrain the number from both sides of the guidance calculation.

C. Increase the resolution so each bottle occupies enough latent cells to be individually resolved.

D. Switch to a model with a larger text encoder, which will parse the numeral correctly.

21 · 8 min

Why "without" summons the thing

THE ONE THING TO KEEP

A caption-trained encoder represents which concepts are present far better than logical operators over them, so naming something in order to exclude it tends to make it more likely.

The classic demonstration

`A room with absolutely no elephant in it.` Run it. On most models, some of the time, there is an elephant.

The reason is in the training data, not in any perversity. Captions on the web overwhelmingly describe what is in a picture. Almost nobody writes "a photograph of a street with no elephant", so the encoder had little opportunity to learn what "no" does to a following noun. What it learned instead is that the token `elephant` appears in captions of pictures containing elephants. The word is in your prompt. The elephant becomes more probable.

Models with a proper language encoder do considerably better. Ask a T5-based model for a dog not wearing a hat and you will usually get a bare-headed dog. The improvement is real and it is uneven: negation of a whole object works reasonably; negation of an attribute (`a cup that is not red`) works less well; negation of a relation (`a man not looking at the camera`) still fails often.

Positive descriptions of the desired state

The reliable technique is to describe what you want rather than what you do not.

- Not `a street with no cars` but `an empty street at dawn, bare tarmac` .

- Not `a portrait without glasses` but `a portrait of a woman, bare eyes, clear view of both eyes`.
- Not `a landscape with no people` but `an uninhabited valley, wilderness, no trace of settlement` — and note that even here, `no trace of settlement` is doing less work than `wilderness`.

This is a general habit worth acquiring, because it applies far beyond images. The system responds to the concepts present in your text. Every unwanted thing you name is a concept you have made present.

Where the negative prompt does and does not help

The obvious question is whether the negative prompt solves this. Partly, and less than people assume.

Putting `elephant` in the negative box moves the guidance anchor toward elephant-containing images, and pushes the result away from them. For a whole distinctive object, this often works and is the right tool.

It fails in the same places prompt negation fails. `two heads` in the negative does not enforce one head, because there is no visual signature for "two heads" distinct from "heads". `not smiling` fails because the anchor moves toward smiling faces and pushing away from them lands somewhere unpredictable — often an unpleasant expression rather than a neutral one. When you want neutral, ask for neutral.

The absence problem underneath

There is a deeper limitation worth understanding, because it explains a family of failures at once.

These models are trained to produce images that match captions. Absence is not a visual feature. A picture of a street without cars and a picture of a street where the cars happen to be out of frame are the same image. The model has no representation of "the thing that is not there", because nothing in the pixels distinguishes the cases.

So requests that hinge on absence are structurally hard: an empty shelf that is specifically empty rather than merely bare, a person with no shadow, a room stripped of a particular item. In each case the fix is to find a positive visual description of the same situation. An empty shelf is `bare wooden shelf, dust marks where objects stood`. That is an image the model can match a caption to.

The practical rule, which will save you a great deal of time: if you cannot describe what the picture looks like without using the word "no" or "without", you have not finished deciding what you want. Work out what is in the frame instead. Nine times out of ten the prompt improves for reasons beyond negation, because you have replaced a vague exclusion with a concrete scene.

There is one exception where naming the unwanted thing is right, and it is worth separating out so the rule does not become superstition. When the unwanted element is a **medium artefact** rather than a subject — a watermark, a signature, a border, a caption bar, a stock-photo overlay — the negative prompt is the correct tool, because those things have a consistent visual signature the encoder learned as a group. `watermark` in the negative genuinely reduces watermarks. `elephant` in the negative genuinely reduces elephants, too. The distinction is that a watermark is not something you would otherwise have to describe positively, whereas an empty street is. Use exclusion for the furniture of the image format; use description for the content of the picture.

BEFORE YOU MOVE ON

Why does asking for "a landscape with no buildings" often produce a building?

A. The safety layer rewrites prompts containing negations to avoid ambiguous instructions, and the rewrite drops the exclusion.

B. The negative prompt field is the only place negation is processed, so negations in the positive prompt are ignored and stripped.

C. Captions rarely describe absences, so the encoder learned that the word marks presence rather than exclusion.

D. The model interprets "no" as a stylistic modifier because it appears frequently in art-related captions.

22 · 8 min

Left, right, behind: the relations it cannot hold

THE ONE THING TO KEEP

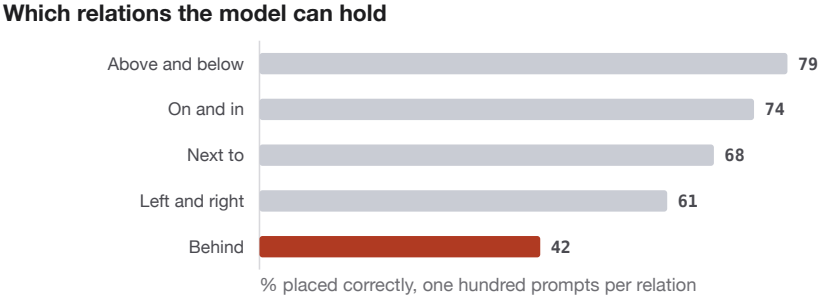
Spatial relations are learned as weak statistical associations rather than represented geometrically, so relational prompts fail unpredictably and are better solved with a conditioning map than with words.

The measurement, not the impression

Benchmarks built specifically for this — sets of prompts of the form "A to the left of B" with automatic checking — find that image models score far below what people assume from their general quality. Older models sat close to chance on left-and-right. Current ones are substantially better and still make errors at rates you would not tolerate in any other tool: a relation that fails one time in four is not a relation you can build a layout on.

The asymmetry is informative. Models handle **above** and **below** better than **left** and **right**. They handle **on** and **in** better than **behind**. And they handle relations between dissimilar objects better than between similar ones.

Each of those follows from the training captions. Vertical relations are described more often and more consistently, because gravity makes them stable and worth mentioning. Left and right depend on the viewer's frame and are frequently written from the subject's perspective instead, so the training signal is genuinely contradictory. `Behind` requires depth, which is not directly visible in a flat caption-image pair.



Vertical relations are described more often and more consistently in captions, so they land more often. Left and right are frequently written from the subject's point of view rather than the viewer's, so the training signal contradicts itself. A relation that fails one time in three is not something to build a layout on.

What "learned as association" means here

There is no coordinate system anywhere in the model. When it renders `the cat to the left of the dog`, it is not placing objects on a plane. Latent cells on the left side settle into cat-like content because the conditioning made cat slightly more likely in regions that co-occurred with left-ness in training. That is a soft statistical pressure competing with everything else in the image, and it loses whenever another pressure is stronger — for instance, the compositional habit of putting the larger subject in the centre.

This is the same machinery as attribute binding, and it fails for the same reason: no explicit structure holds the relation.

Solve it with geometry, not language

The reliable approach is to stop asking and start showing. Three routes, all free:

A conditioning map. Sketch two blobs in the right places in GIMP, Krita or any drawing app — twenty seconds, no skill required. Feed it as a scribble or depth control. The layout is then fixed by the map, and the prompt only has to say what the blobs are.

Regional prompting. Divide the canvas explicitly and condition each region. This solves relation and binding at once, since the left region's prompt contains only the cat.

Composite. Generate the cat, generate the dog, place them. This is how nearly all commercial work is done anyway, and it also solves scale, which relational prompts never handle: `a cat next to a dog` gives you no control over their relative size, and the model's default is frequently wrong.

The related failure with viewpoint

A neighbouring problem worth naming, since it wastes as much time. Prompts specifying camera position — `low angle`, `from behind`, `three-quarter view`, `bird's eye` — work only to the degree those phrases appeared in captions with consistent meaning. `Low angle` and `bird's eye` are photographic vocabulary and work reasonably. `From behind and slightly to the left` is not a caption anyone writes and behaves accordingly.

For anything where viewpoint matters, the cheap professional trick is to build the scene roughly in free 3D software — Blender, or even a simple posing tool — screenshot it, and use the screenshot as a depth or edge map. You get exact camera control from a program designed for exact camera control, and the generative model does what it is good at, which is surfaces.

The honest limitation: none of these give you a model that understands space. They give you ways to supply the spatial information yourself. That distinction matters when someone claims a new model has solved composition. Ask how it does on relations between two similar objects, and on "behind". Those are the cases where the association is weakest and the claim gets tested.

A last practical note, because it explains a lot of wasted time. Relation failures are not stable across seeds, which means the first thing everybody does — regenerate and hope — sometimes works. That is the worst possible feedback. Getting the relation right one time in three teaches you that the prompt nearly works and encourages another twenty attempts, when the actual state of affairs is a coin weighted slightly in your favour. The test for whether a relational prompt is reliable is four fixed seeds, not one lucky one. If it fails on two of four, it will fail on the shot the client picks.

BEFORE YOU MOVE ON

Which prompt pair would you expect a current image model to handle most and least reliably?

A. Most reliable: a lamp above a table. Least reliable: a red chair behind a blue chair.
B. Most reliable: a red chair behind a blue chair. Least reliable: a lamp above a table.
C. Both are handled equally well, since modern encoders parse prepositions correctly and pass the parsed relation through to the layout stage of generation.
D. Most reliable: the one with fewer words, since shorter prompts leave less room for the encoder to lose the relation.

23 · 8 min

The pull toward the average

THE ONE THING TO KEEP

Guidance and short prompts both push output toward the most typical example in the training distribution, which is why unspecified images converge on a narrow, glossy sameness.

Everyone has seen this and few people name it

Generate `a photo of a woman` twenty times on an unmodified model and look at the set. The faces will differ, and they will differ around a centre: a similar age, a similar symmetry, similar skin, similar lighting, a similar pleasant expression. The same happens with `a house`, `a meal`, `an office`.

This is the distribution pull, and it has two separate causes that compound.

The prompt underspecifies. Everything you do not say is filled in by what is most probable. Most probable is not neutral. It is the mode of a corpus that

heavily over-represents commercial photography, stock imagery and social media, all of which are already selected for a narrow idea of what looks good.

Guidance amplifies the typical. Recall the arithmetic: guidance pushes away from the unconditional prediction and toward the conditional one, past where either pointed. That extrapolation drives every region toward the most prototypical version of what it is. High guidance is, quite literally, a dial for how stereotypical the output is.

The visible signature

Once you know what to look for, the pull is easy to spot:

- Skin without pores, blemishes or asymmetry.
- Teeth uniformly white and even.
- Lighting that is always flattering and always from a plausible key with fill.
- Objects that are new, clean and undamaged.
- Food that is styled.
- Interiors that are tidy.

None of this is a rendering defect. Every element is a plausible thing. It is the *combination* that never occurs in an unposed photograph, and it is why generated images read as advertising even when the subject is mundane.

Working against it

Three approaches, in increasing order of effectiveness.

Specify the atypical. Say the things you would not have thought to say: the time of day, the state of the objects, the imperfection. `A cluttered kitchen at 7am, unwashed pan on the hob, low winter light through a dirty window, someone's coat over a chair.` Every clause you add takes a decision away from the mode.

Lower the guidance. Dropping from 8 to 4 does more for realism than most prompt vocabulary. It costs you prompt obedience, which is a real trade and worth making when the subject is simple.

Condition on a real photograph. A depth or edge map taken from an ordinary snapshot imports the composition of a real, unposed scene — which is the very thing the model will not produce on its own.

Why this matters more than an aesthetic complaint

The pull toward the average is the mechanism behind the bias covered in the next lesson, and it is worth seeing them as one thing. When a model produces the same kind of face for `a doctor` every time, it is not applying a rule about doctors. It is doing what it does for kitchens and lamps: returning the centre of a distribution. The distribution has a demographic shape because the corpus does.

Understanding it as one mechanism has a practical benefit: the same interventions work. Specification, lower guidance and reference conditioning all reduce demographic collapse for the same reason they reduce aesthetic collapse.

There is a limitation nobody has removed. You can push away from the mode, but you cannot get to a region the training data barely covers. Ask for a kind of face, a kind of room or a kind of tool that is rare in the corpus and specification will not conjure it; you will get the nearest well-covered thing with your adjectives applied on top. That boundary — between "underspecified" and "not there" — is the single most useful thing to be able to tell apart, and the test is simple: add the specification and see whether the image moves toward what you meant or merely acquires a label of it.

It is worth noticing that the pull is also why generated images age so visibly. The mode of the corpus at the time of training becomes the look of everything the model makes, so a model trained in one year produces a recognisable house style that dates the way a decade's stock photography dates. Anyone building a body of work on a single checkpoint is committing to that look. That is an argument for a light touch — generate elements, finish them yourself, keep your own grade — rather than for shipping the model's default and calling it a style.

BEFORE YOU MOVE ON

Twenty generations of "a photo of a kitchen" all return tidy, evenly lit, magazine-like rooms. Which single change most directly counteracts the mechanism?

A. Raise the guidance scale so the model follows the word "photo" more strictly and produces a result that reads as documentary rather than as commercial photography.
B. Increase the number of steps so the model has more opportunity to introduce variation across the run.
C. Generate at a higher resolution so more of the room's incidental detail can be rendered.
D. Specify the unglamorous particulars — clutter, time of day, worn surfaces — so fewer decisions are left to the mode.

24 · 9 min

The bias, and what mitigating it actually does

THE ONE THING TO KEEP

Demographic skew in generated images is the distribution pull applied to people, measured and reproducible, and every available mitigation trades one problem for another rather than removing it.

What has been measured

This is one of the better-documented properties of image models, because it is easy to test at scale: generate a few hundred images for an occupation prompt and classify what comes back.

The consistent findings across several published audits:

- Occupation prompts return images skewed well beyond real-world workforce statistics, in the direction of the stereotype. Prompts for high-status professions return predominantly light-skinned men; prompts for care and service work return predominantly women.
- Skin-tone distributions are compressed toward the lighter end of standard scales, including for prompts that specify a country where that is not representative.
- Prompts naming a country return the most touristically photographed version of it. `An Indian street` returns a crowded market with saturated colour; `a German street` returns clean cobbles. Neither is false; both are a magazine's idea of the place.
- Adding an attribute like `poor` or `criminal` shifts the demographic distribution markedly, which is the most troubling finding of the group because it shows the association is available and reachable.

None of this requires malice anywhere in the pipeline. It is the corpus, and the mode-seeking behaviour of the last lesson, working exactly as described.

What the mitigations do

Three families are in use, and each has a cost worth knowing.

Prompt injection. The service silently adds or substitutes demographic terms. It is cheap, it works on the immediate complaint, and it is unaware of context. This is the mechanism behind well-publicised incidents where diversity terms were applied to historical prompts and produced images that were absurd or offensive. The failure was not the goal; it was applying a blunt textual rule to every request without knowing which requests it made sense for.

Retraining and rebalancing. Filter or reweight the training data. This is the principled route and it is expensive, slow, and only as good as the labels used to rebalance — which means somebody has decided which categories exist and what the target proportions are. Those are contested choices, and they are usually made without publication.

User-level specification. Push the decision back to the person prompting. Honest, effective and unevenly distributed: it works if you know to do it,

which puts the burden on the people most likely to be misrepresented.

What this means for your work

Two practical positions, and they are not in tension.

If you are producing images of people for anything public, specify. Leaving it to the default means shipping the corpus's assumptions under your name, and "the model chose it" is not a defence anybody accepts. This is the same discipline as casting: someone always decides, and pretending otherwise just hides the decision.

If you are evaluating a tool or a vendor's claim, ask what was measured and how. "We have addressed bias" is not a claim you can check. "Here is the distribution over skin tone for these fifty occupation prompts, before and after" is. Insist on the second form. The methods for measuring this are public and cheap; a supplier who has not run them has not looked.

The part that stays unresolved

There is no agreed answer to what the right output distribution is, and it is worth being clear that this is a genuine disagreement rather than a gap waiting to be filled.

Should `a doctor` reflect the real demographics of doctors in the user's country, which vary enormously? The global figures? An equal split, on the grounds that a generated image is not a statistic? The demographics of the people likely to see it? Each has been argued seriously, each implies different outputs, and no principle chooses between them.

What can be said without controversy is narrower and still useful: the unmitigated default is more skewed than any of the candidate answers, the skew is measurable, mitigation by silent prompt injection has produced visible failures, and the choice is currently being made by vendors without publication. Anyone who tells you this is a solved problem is describing their preferred answer as though it were the only one.

One further note for anyone working in a country the corpus under-represents. You will find that the model's idea of your own surroundings is a foreigner's, and that specifying harder produces a caricature rather than a correction, because the specific vocabulary has thin coverage. The route that works is the one from the previous module: photograph your own references and condition on them, or train a small fine-tune on a few dozen of your own images. Twenty photographs taken on a phone are enough to move a model's idea of a particular kind of street, shop or garment further than any prompt will. It is more work than it should be, and it is work you can actually do.

BEFORE YOU MOVE ON

A vendor states that its image model's demographic bias has been "addressed". What is the most useful follow-up question?

A. Which countries' workforce statistics were used as the target distribution for each occupation prompt in the mitigation?

B. Was the mitigation applied at the training stage rather than by prompt injection?

C. Which measured distributions changed, over which prompts, and by how much before and after?

D. Does the model still permit prompts containing occupational terms?

25 · 8 min

Reflections, shadows and things that must agree

THE ONE THING TO KEEP

The model matches local texture statistics without enforcing any global constraint, so anything that must agree across the picture — a reflection, a shadow, a continuing line — fails in a characteristic way.

The class of failure

Look closely at a generated image of a room with a mirror. The mirror contains a plausible reflection. It is frequently not a reflection *of that room*: the furniture is different, the angle is impossible, a figure appears who is not in the scene.

The same shape of failure recurs:

- **Shadows** that fall in different directions from different objects, or that are missing under one item and present under its neighbour.
- **Water and glass** that refract convincingly in each patch and inconsistently across the whole.
- **Continuing lines** — a railing, a cable, a road marking, the edge of a table — that go behind an object and re-emerge at a different height.
- **Symmetry** — a pair of earrings, two shoes, a jacket's buttons and buttonholes.
- **Repeated instances of one thing**, such as the same building at two scales, or a character's clothing changing between one side of the frame and the other.

Why this is one mechanism and not five

The denoising network operates over the latent grid with a limited receptive field per layer. Deeper layers see wider context, so the model does have some global reach — enough to keep a composition coherent. What it does not have is any *constraint solver*. Nothing checks that the reflection is consistent with the room, because consistency is not something the training objective ever asked for.

The objective was: given noisy input, predict the noise. A picture where the mirror shows the wrong room is, locally, exactly as easy to denoise as one where it shows the right room. Both are mirror-shaped regions containing mirror-like content. The training signal that would have distinguished them never existed.

This generalises into the most useful heuristic in the module:

Local and global, and which one the objective ever rewarded

Decided inside one patch — reliable	Must agree across the picture — unreliable
The weave of a fabric	A reflection that matches the room
How a metal surface takes light	Shadows from a single light source
Depth of field falling off	A railing continuing behind a post
Skin texture and grain	A pair of earrings, or two shoes
The colour of a shadow	The same logo drawn twice
	A word with the right letters in it

A picture where the mirror shows the wrong room is, locally, exactly as easy to denoise as one where it shows the right room. The training signal that would have distinguished them never existed. Hands and spelling were the two famous cases of the same rule.

> *These models are excellent at anything decided locally and unreliable at anything that must agree across the picture. Texture, material and light quality are local. Counting, spelling, reflection, symmetry and continuity are global.*

Hands and text — the reused lesson opening this module — are the two most famous instances of the same rule. A hand must have a globally consistent number of fingers. A word must have globally consistent letters. Neither is a local texture problem.

What to do about it

Choose scenes that avoid the trap. This is the professional answer and it is not defeatist. Cinematographers avoid mirrors too, for their own reasons. If a reflection is not the point of the picture, do not put one in the frame.

Fix it in post. A wrong reflection is a two-minute job in any layered editor: select the mirror, paste a flipped copy of the actual scene, blur and reduce opacity. GIMP, Krita and Photopea all do this for nothing. This is far faster than re-rolling for a model to get lucky.

Inpaint the disagreeing region with the correct content visible as context. Masked regeneration sees the surrounding pixels, so a shadow inpainted with the light source in view often comes back correct.

Use a 3D pass for the geometry. For anything where physical consistency is the whole point — a product on a reflective surface, an architectural view — build the block-out in Blender, which is free and physically correct by construction, and use the render as a structural reference.

The claim to be sceptical of

Every generation of models is announced with improved "understanding of physics". Read this precisely. What improves is the statistical fidelity of local appearance and, in video, short-run plausibility of motion. What has not been demonstrated is a model that enforces a constraint.

The test that separates them is easy to run and worth running on any model that makes the claim: put a mirror or a still water surface in a scene with two distinctive objects, and check whether the reflection contains those objects in the right arrangement. If the answer is sometimes, the model is still doing statistics.

There is a version of this test for video too, and it is the one to reach for when a video model is described as a world simulator. Show it a scene where something is occluded and then revealed — a ball rolling behind a chair, a hand passing in front of a face. Local plausibility carries the shot until the object comes back out, and what comes back is often a different object, or nothing.

Object permanence is a global constraint over time, which is the same category of thing as a reflection over space, and it fails for the same reason.

BEFORE YOU MOVE ON

A generated interior shows a mirror whose reflection contains furniture that is not in the room. Which description of the cause is correct?

A. The mirror region was generated in a separate pass and then composited into the finished picture without any reference to what the main image actually contained.

B. The model has no mechanism enforcing agreement between regions, and a plausible-looking reflection is as easy to denoise as a correct one.

C. The reflection is drawn from a memorised training image of a different room.

D. The latent resolution is too low in the mirror region for the reflected furniture to be reconstructed accurately by the decoder.

---

26 · 7 min

Faces in crowds and other small things

THE ONE THING TO KEEP

Detail quality is set by how many latent cells a subject occupies, so anything small in frame degrades regardless of prompt, model quality or output resolution.

The arithmetic of a face in a crowd

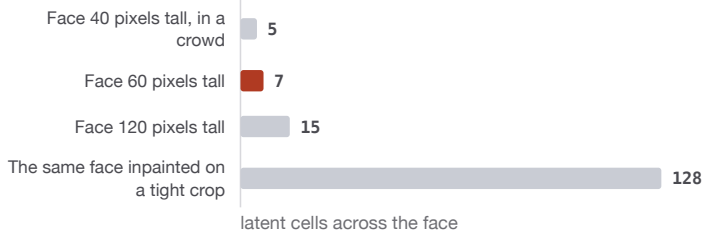
A 1024-pixel image on a model with an eight-times-downscaling autoencoder has a 128 by 128 latent grid. A face that occupies 60 pixels of the final image is about seven latent cells across.

Seven cells must hold two eyes, a nose, a mouth, the shape of a jaw and hair. It cannot. What comes back is a face-shaped region with facial statistics — right

colour, right soft transitions, wrong or absent features. The characteristic result is the melted background face that everyone recognises in crowd scenes.

The same arithmetic governs everything small: a wristwatch, a distant window, a logo on a shirt, a bird against the sky, the teeth in a smile at middle distance.

Latent cells across a face, on a 1024-pixel model



The grid is 128 by 128, so a sixty-pixel face gets seven cells to hold two eyes, a nose, a mouth and a jaw. Going from 1024 to 1536 takes it to about ten, which is why raising the output size is not the fix. Detail is a budget, and the budget is cells.

Why raising the output resolution does not fix it

The intuition is to generate larger. It does not work as expected, for two reasons.

First, if you exceed the model's native size in a single pass, you get the repeated-subject artefact from the first module — two heads, duplicated bodies — because nothing coordinates distant regions of an oversized latent.

Second, and less obvious: even when the resolution increase works, the face gets more cells only in proportion. Going from 1024 to 1536 gives your seven-cell face about ten cells. That is a marginal improvement for a substantial cost, and the failure mode does not change in kind.

What actually works

Generate the small thing large, separately. Make the face its own 1024-pixel image. Scale it down. Composite. It will hold up under any zoom, because it was made with the cells it needed.

Detail passes with masked regeneration. The standard professional workflow: generate the wide shot, then mask each face and regenerate that region *at full resolution*, with the mask's contents upscaled before denoising and scaled back afterwards. Free implementations of this exist under names like `ADetailer`, face-detailer nodes in ComfyUI, and similar. It is the single biggest quality gain available to anyone generating scenes with people in them, and most people never turn it on.

Accept the limit and design around it. Photographers do exactly this. Depth of field exists partly because distant faces are not renderable in detail on film either. A generated crowd with a genuinely shallow focus is both more convincing and cheaper than one where every face was fixed.

The related case: thin structures

A neighbouring failure with the same root. Hair strands, wire fences, rigging, spider webs, whiskers, text serifs — anything one or two pixels wide in the final image is a fraction of a latent cell.

The decoder handles these by producing the statistical impression of the structure rather than the structure. Hair looks like hair at a glance and dissolves under a zoom. Fences develop breaks and reconnections. This is why generated images often survive at Instagram size and fall apart at print size, and it is worth knowing before you promise a client a billboard.

The general principle to carry: **detail is a budget, and the budget is latent cells, not pixels.** Before you generate anything, ask how many cells the important part of the picture will get. If the answer is under twenty across, plan a separate pass for it. This one habit removes most of the frustration people have with fine detail, and it costs nothing but a little sequencing.

Why the crop test is the honest one

There is a review habit worth adopting from print production. Never judge a generated image at the size it appears on your screen. Zoom to 100% and look at four crops: the main subject's face or focal detail, one background region,

one edge, and any area containing type or thin structure. This takes thirty seconds and it is where every problem in this lesson shows itself.

The reason to make it a routine rather than an occasional check is that the failure is invisible at fit-to-window and unmissable at full size, so an image can pass every casual look and fail at the printer. Designers who work in print already have this habit. Anyone whose work has only ever been seen on a phone will not, and the first billboard is a bad place to acquire it.

Newer models with sixteen-channel latents and larger native resolutions have improved the numbers, and the arithmetic has not changed shape. Doubling the native size doubles the cells across a small subject; it does not remove the ceiling. The question is always the same: how many cells does this part of the picture get?

BEFORE YOU MOVE ON

A generated group photograph looks good overall, but every face beyond the front row is malformed. What is the most effective fix?

A. Regenerate the image at four times the resolution in a single pass so the distant faces receive more detail.

B. Add "detailed faces, sharp features, high quality faces" to the prompt so the model allocates more attention to them.

C. Mask each face and regenerate it at full resolution, then scale the result back into place.

D. Switch to a model with a stronger text encoder, which will bind the face description to the correct regions.

27 · 8 min

Refusals, filters and the false positives

THE ONE THING TO KEEP

Safety filtering happens at several separate points with different failure characteristics, and the classifier-based stages produce predictable false positives on medical, artistic and non-Western content.

Four places a request can be stopped

Understanding which one stopped you determines whether there is anything to be done.

Prompt blocklist. A string match or classifier on your text, before anything runs. Fast, crude, and the source of the most obviously silly refusals — a prompt refused for a word with an innocent meaning in context.

Prompt classifier. A model judges the request's intent. Better at context and worse at explaining itself, since there is no term you can point to and remove.

Output classifier. The image is generated and then checked. This is where the "black image" or "content not available" results come from — you paid for the generation and the result was withheld. Output classifiers are typically tuned for high recall on nudity and violence, which means a lot of false positives.

Named-entity blocking. Specific people, usually public figures, blocked by name. Increasingly common and, for elections and public figures, increasingly required by law or platform policy.

The false positives, and who gets them

These are not random, and the pattern is consistent enough to be worth stating.

Medical and anatomical content. Breast examination diagrams, wound care, dermatology, childbirth. The classifier sees skin and body parts and cannot distinguish clinical from sexual context. This is a genuine harm to educational and medical work, and it is rarely acknowledged as a cost.

Art history and figurative art. Classical sculpture, life drawing, anatomical study. The same mechanism.

Non-Western clothing and context. Filters trained predominantly on Western imagery misjudge what is exposed and what is ordinary. Traditional dress from several regions triggers nudity classifiers at elevated rates; images of religious gatherings have been flagged as crowds in distress.

Violence in news and history. Historical photographs, conflict reporting, memorials.

The common thread: classifiers are trained on labelled examples, and the labels come from a labelling operation with its own cultural defaults. A false positive is not a bug being fixed on a schedule; it is the classifier working as trained on a case its training under-represented.

What you can actually do

Rephrase toward the clinical or the compositional. `A medical illustration of` and `an anatomical diagram, line art` frequently pass where a photographic phrasing does not. This is not gaming the system; it is describing the thing you actually wanted.

Split the request. Generate the setting and the subject separately when the combination is what triggers.

Use a local model when the work is legitimate and keeps being refused. Open-weight models on your own machine have no service filter. This is the honest answer for medical educators, art teachers and researchers, and it comes with the honest caveat: removing the filter removes it for everything, and the responsibility is then entirely yours.

Report it. Most services have a route for false positives. It is slow, and it is the only mechanism that improves the classifier for the next person.

The part worth being clear-eyed about

Two things are true at once and people usually pick one.

Filtering is doing real work. Sexual imagery of children, non-consensual sexual imagery of adults, and targeted harassment material are the categories these systems exist to prevent, and the first two are illegal in most of the world regardless of how they were produced. Weakening that is not a neutral act.

And the same filters silently restrict medical education, art, journalism and cultural material, disproportionately outside the West, with no appeal that works at speed and no published measurement of the error rate. Both of these are true. A vendor that reports only its catch rate and never its false-positive rate is telling you half of what you need, and asking for the other half is a reasonable thing to do.

One practical warning about the workaround culture. There is a large amount of advice online about phrasings that get past filters, and much of it is aimed at producing exactly the material the filters exist to stop. Following that advice puts you on the wrong side of both a service's terms and, for some categories, the criminal law of most countries — and "the model made it" has never been a defence anywhere. The legitimate version of the same move is narrow and worth stating precisely: describing your actual subject accurately in clinical or compositional terms is not evasion, it is a better prompt. If the rephrasing you are considering would embarrass you to explain, it is the other thing.

BEFORE YOU MOVE ON

A dermatology teaching resource keeps having its illustrations refused by a hosted image service. Which explanation and response fit together best?

A. The prompts contain specialist medical terminology that the service's blocklist treats as explicit content, so systematically replacing those terms with everyday language will let the requests through.

B. The service has a policy against medical content, so the only route is a licensing agreement with the vendor.

C. An output classifier tuned for high recall on skin is flagging clinical images, so clinical or diagrammatic phrasing, or a local model, is the practical route.

D. The named-entity filter is blocking the request because dermatological conditions are associated with identifiable patients in the training data.

28 · 8 min

The tells, and why they are disappearing

THE ONE THING TO KEEP

The visible artefacts people use to identify generated images are transient consequences of specific architectures, so a detection habit built on them expires and confident identification by eye is not reliable.

The tells, as they stood

There has been a folk practice of spotting generated images, and the list is worth writing down precisely — partly because some of it still works, and mostly to see how it has decayed.

- **Hands.** Six fingers, fused digits, thumbs on the wrong side. Largely fixed on current models; still fails in complex interactions such as two people holding hands.

- **Text.** Letter-like marks that spell nothing. Substantially fixed on large models, and still reliable on small local ones.
- **Teeth and ears.** Too many teeth, ears of different shapes. Mostly fixed.
- **Backgrounds.** Melted faces in crowds, structures that do not connect, patterns that shift. Still common, because this is the resolution limit from the earlier lesson rather than a training gap.
- **Jewellery and eyewear.** Asymmetric earrings, spectacle arms that vanish. Still frequent, because it is the symmetry-and-continuity failure.
- **The surface.** Skin without pores, uniform lighting, a general glossiness. Still the strongest single tell, and it is the distribution pull rather than any defect.

Notice how the list divides. The failures that were about training coverage got fixed by more data. The failures that are structural — global consistency, latent resolution, mode-seeking — have not been fixed, because more data does not address them.

Why "I can always tell" is not true

Two facts make confident visual identification unreliable, and both are uncomfortable.

Selection. You see the generated images that were noticed. Every convincing one passed by without registering. This is survivorship bias operating on your entire sense of how good these systems are, and there is no way to correct for it by looking harder.

The failures that remain are removable. A skilled retoucher fixes hands, adds grain, breaks the lighting, and photographs a screen to add real optical noise. Every tell on the list above is a fifteen-minute job for someone who intends to deceive. The tells identify *casual* generation, which is a different and much easier problem than identifying deliberate deception.

The consequence is that a false accusation is now a real risk. Photographers have had genuine work called AI-generated on the basis of smooth skin, which is what studio lighting and ordinary retouching produce. The costs land on individuals with no route to prove a negative.

What to use instead

The next-to-last module covers provenance properly. The short version, which is worth having now:

Verify the **source and the claim**, not the pixels. Where did this file come from, who published it first, does anyone with a name stand behind it, does a reverse image search show it appearing earlier somewhere else. Free tools — Google Lens, TinEye, Yandex — answer the last question in seconds and answer it far more reliably than any inspection of fingers.

Treat an image with no traceable origin as unverified, whether or not it looks generated. That is the same standard journalism used before any of this existed, and it survives every improvement in the models.

The habit worth building

When you catch yourself thinking "that looks AI", finish the thought properly: *which mechanism would have produced that artefact?* Melted background faces mean latent resolution. Wrong reflections mean no global constraint. Glossy uniformity means distribution pull and high guidance.

If you can name the mechanism, you have a real observation. If you cannot, you have a feeling — and feelings about images are exactly what everybody's is being trained on at the moment, by both the generators and the people who benefit from doubt.

What to say when someone asks

You will be asked, because everybody with any reputation for knowing about this gets asked. A good answer has three parts and takes fifteen seconds.

Say what you can see and what mechanism it suggests, if anything. Say that visual inspection cannot establish origin either way, and that people have been wrongly accused on exactly this basis. Then move to the question that can be answered: where did the file come from, and who is standing behind it.

That answer is unsatisfying to somebody who wanted a verdict, and it is the truthful one. Declining to give a confident verdict on an image is not a failure

of expertise. At the moment it is what expertise consists of.

BEFORE YOU MOVE ON

Why has the "count the fingers" heuristic become unreliable, while melted faces in crowd backgrounds remain common?

A. Hand rendering was a training-coverage problem that more data fixed, whereas small background faces are limited by latent resolution, which data does not change.

B. Hands are simpler geometrically than faces, so the models mastered them first and will master faces next.

C. Modern hosted services run a dedicated hand-correction pass after the image has been generated, while background regions are deliberately excluded from that pass in order to save compute on every request.

D. Users learned to write better prompts for hands, and no equivalent prompt vocabulary exists for background faces.

CASE STUDY · MODULE 3

Forty cards, three wrong, and a term starting Monday

A small educational publisher in Patna, proofing a Class 2 counting set eleven days after an intern started generating it

Saraswati Prakashan prints workbooks and teaching aids for low-fee private schools across three districts. In June they took an order for a Class 2 numeracy set: forty flashcards covering the numbers one to ten, two cards for each number, each card showing that many of something familiar. Mangoes. Kites. Bangles. Lassi glasses.

An intern generated the cards over eleven days. He was careful. He regenerated each card until the count looked right, sometimes fifteen or twenty times, and he checked every one before handing them over.

The proofreader was Ramesh Jha, a retired primary teacher who had been reading proofs for the firm for nine years. He printed the set at full size, laid it out on the table, and counted.

Seven kites had eight kites. Nine bangles had eleven. And five mangoes had five mangoes and, behind a leaf at the left edge, most of a sixth.

The intern's defence was that he had checked them, and he had. He had checked them on a laptop screen at the size four cards fit side by side, which is the size at which a sixth mango behind a leaf is a leaf.

Anjali Prasad, who runs the firm, had to decide by that evening. The distributor's van to the districts left Monday morning. The next van was in three weeks, which was after the counting chapter would have been taught.

She had three options and none of them was free.

She could ship thirty-seven cards. But a counting set missing five, seven and nine is not a counting set, and the schools had ordered it because the counting chapter starts in week two.

She could have Ramesh count all forty at full size and have the intern regenerate the three that failed. This was the option everyone in the room preferred, and it was the one Anjali distrusted, for a reason worth stating precisely. The intern's method had produced three failures in forty. Applying the same method to the three failures produces failures at the same rate. It also assumes Ramesh caught everything, and two people in the office had already disagreed about the mango, which meant the error count was not three but somewhere between three and five.

Or she could stop prompting for counts altogether. Generate one good mango, cut it out, duplicate it five times, vary the rotation and the shadow, and place them by hand in a free editor. Forty cards at roughly twenty-five minutes each is about sixteen hours. Two people, two days. Monday becomes Wednesday, which means the van is missed and a private courier is paid.

There was a second thing on the table, and it had eaten two of the intern's eleven days. The lassi cards kept coming back as milkshakes in tall American glasses with straws. He had tried thirty variations of the wording. He had tried weighting the word. He had put milkshake in the negative prompt. Nothing moved it, and he had concluded that he was bad at prompting.

He was not. Those are two different failures with two different fixes, and the eleven days had gone on treating them as one.

STOP HERE

Before turning the page: what would you have done, and what would it have cost if you were wrong? Write it down. The point of the next section is to compare.

WHAT HAPPENED

Anjali chose the composite. Two people, two days, and a private courier on Wednesday that cost six thousand rupees against a job worth about ninety.

All forty cards were correct, because a person placed each object and a person can count. The generative model did what it is reliably good at — producing one convincing mango, one convincing kite, one convincing bangle, each gen-

erated large and therefore with enough latent cells to hold real detail — and the arrangement, which is the part with a global constraint in it, was done by hand.

The lassi problem took eleven minutes. They put six steel glasses on a table in the office kitchen, photographed them on a phone, extracted a depth map and generated from that. The result was a glass anybody in Patna would recognise, because the geometry came from a real glass rather than from a caption corpus that has very little to say about lassi and a great deal to say about milkshakes.

Anjali wrote two rules afterwards and put them on the wall.

If a specification contains a number, nobody prompts for the number. The elements are generated and the arrangement is assembled.

And every proof is read at full print size, not on a screen at a size that flatters it.

The second rule was not new. It is what Ramesh had been doing for nine years and what everybody else in the building had quietly stopped doing when the work moved onto screens.

The intern regenerated 'seven kites' until it looked right, and shipped eight. What is wrong with 'regenerate until it looks right' as a method for a count?

It selects on the reviewer's inspection rather than on the property being specified, so its error rate is the error rate of the inspection. Nothing in the generation process counts — each region of the latent settles into something plausible for its neighbourhood, with no global register tracking how many objects have been placed — so whether a card has seven or eight kites is close to random above about four objects. Regenerating until it passes a glance produces a set that passes glances, and a card that fails at full size on a table was always going to. The count has to be established by something that actually counts, which means a person placing objects.

Would 'exactly seven kites, (seven:1.5), eight kites in the negative prompt' have worked? Say why in terms of the mechanism.

No. All three are attempts to argue with a machine that has no counter. Repeating or weighting the number scales a token vector that the model was never trained to see scaled, which mostly makes the concept spread rather than making it precise. A negative prompt moves the point the guidance pushes away from, and there is no distinct visual signature for 'eight kites' separate from 'kites', so pushing away from it pushes away from kites. A count is a global, discrete, exact constraint, and none of the available prompt controls acts on global constraints at all.

The lassi glasses came back as milkshakes and no amount of rewording moved them. Why is that a different failure from the wrong count, and what fixed it?

The count failure is a mechanism problem: the model has the concept and no way to enforce a quantity. The lassi failure is a coverage problem: the caption corpus the model learned from has thin and mostly wrong material for that object, so the nearest well-covered thing is a milkshake, and specifying harder only adds adjectives on top of the wrong object. The distinction is between 'underspecified' and 'not there'. Coverage gaps are fixed by supplying the thing rather than a word for it — a reference image, a depth map, or a small fine-tune. Six glasses on a kitchen table and a phone took eleven minutes and moved the model further than two days of wording had.

MODULE 3 · TEST

Check what stuck

6 questions, one right answer each. This one is for you, not for a record: answers and the reason behind each are at the back. If two questions from the same module go wrong, re-read that module before the exam rather than after.

1 Requests for one, two or three objects return the right number almost always; five is a coin toss and eight is random. What explains the cliff?

A. There are far fewer training images containing five objects than containing two.

B. Above four objects the prompt exceeds the encoder's effective attention span.

C. Each latent cell settles by local plausibility, and nothing holds a running count.

D. Guidance drives the image toward the most typical arrangement, which is two or three objects.

2 'A room with absolutely no elephant in it' returns an elephant a noticeable share of the time. Why?

A. The encoder represents which concepts are present far better than operations over them.

B. The word 'absolutely' acts as an emphasis weight on the noun that follows it.

C. Negation is handled by the negative prompt path, which is empty in this case.

D. The tokeniser splits 'no elephant' into pieces that the encoder recombines as 'elephant'.

3 Image models handle 'above' and 'below' more reliably than 'left' and 'right'. What is the most likely reason?

A. Vertical relations are easier to render because the latent grid is scanned row by row.
B. Left and right depend on the viewer's frame, and captions are written from both, so the training signal is contradictory.
C. Horizontal placement is decided in the last few steps, when the prompt has least influence.
D. Most models are trained on portrait-orientation images, which under-represent horizontal arrangements.

4 A crowd scene generated at 1024 pixels has melted faces in the background. You regenerate at 1536 and the faces are still wrong. Why did the larger size not fix it?

A. The upscaling happened after generation, so the extra pixels carry no new information.
B. Background regions receive less guidance than the subject, so they are always weaker.
C. The faces gained cells only in proportion, and are still far too small to hold features.
D. The model's face-rendering capability is trained only at its native resolution.

5 A generated room contains a mirror, and the mirror shows a plausible room that is not the room in the picture. Which statement identifies the mechanism?

A. Reflections are rendered by a separate pass that does not see the main image.
B. Mirrors are rare in training data, so the model has little coverage for them.
C. The latent grid has too few cells in the mirror region to hold the reflected scene.
D. Nothing in the training objective ever asked for agreement across the picture.

6 A relational prompt places two objects correctly about one time in three, so you keep regenerating. Why is that the worst possible kind of feedback?

A. Repeated generation drifts the output toward the model's distribution average.

B. Occasional success suggests the prompt nearly works, encouraging attempts at a coin weighted slightly in your favour.

C. Each regeneration changes the seed, which changes the prompt's effective weighting.

D. The model learns from your rejections within a session and narrows its output.

MODULE 4

Control: getting the picture you actually need

Prompting produces a starting frame; control produces a deliverable. This block covers the mechanisms that let you decide rather than hope — the denoise strength that governs how much of an existing image survives, masked regeneration and its context window, structural conditioning maps, small fine-tunes trained on your own material, the honest state of character consistency, what an upscaler invents, why compositing is still the job, and the free toolchain that runs all of it on ordinary hardware.

> BY THE END OF THIS MODULE YOU CAN
>
> Take a nearly-right generation to a finished deliverable using the appropriate mechanism at each stage — denoise strength, masked regeneration, a conditioning map, a small fine-tune, an upscaler and a manual composite — and say which of those a given defect requires

29 · 8 min

Editing what you already have

THE ONE THING TO KEEP

Prompting gets you a starting frame; masks, denoise strength and structural inputs get you a finished image.

Nobody good ships a first generation

The workflow people actually use looks like this: generate until you have a base with the right composition and light, fix regions one at a time, extend the frame if you need to, upscale last. The prompt got you sixty percent. The rest is masking.

Inpainting, and the setting that changes everything

Inpainting means painting a mask over a region, describing what should be there, and regenerating only that region while the model reads the surrounding pixels for context.

The setting that separates a clean result from a smear is usually called "inpaint at full resolution" or "inpaint only masked area". Without it, the entire image is squeezed down to the model's working size, so a face occupying five percent of a wide shot gets maybe forty pixels and comes back as mush. With it, the tool crops around your mask, generates that crop at the model's native resolution, and blends it back. Same model, same prompt, completely different outcome.

Three practical rules:

- Mask a little wider than the problem. Tight masks leave visible seams.
- Change the prompt to describe the region, not the scene. If you are fixing a hand, prompt the hand.
- Fix one thing per pass, and save between passes.

Denoising strength is the dial you actually turn

Whether you are inpainting or running image-to-image, one number governs how far the result may travel from the input. It sets how much noise is added before denoising begins.

- **0.2 to 0.35** — texture and finish only. Shapes survive intact.
- **0.4 to 0.6** — the same objects with different details. Faces will change. Most real work happens here.
- **0.7 and up** — the input is a loose suggestion. Composition survives, content does not.

If your edit is not taking, raise it. If your subject keeps turning into a different person, lower it and do two passes.

Keeping the structure, changing everything else

Structural conditioning feeds the model a second input alongside your prompt: a depth map, an edge trace, a pose skeleton, a rough scribble. The model must respect that structure while the prompt decides everything else.

This is how you get the exact same room at night in the rain, or a pencil sketch turned into a finished illustration that keeps your composition. It is the single biggest jump in control that most people have never tried.

Instruction editing, and its quiet problem

A newer class of model edits from plain instructions with no mask at all. "Make it evening." "Remove the man on the left." "Put her in a green jacket." These are genuinely good and have become the default for casual work.

Their weakness is worth knowing. Most of them re-render the whole image, not just the part you named. One edit is invisible. After five sequential edits, the face has drifted, the sign on the wall has changed its wording, and fine texture has softened toward the model's house style. It is a photocopy of a photocopy.

So keep the original. Do the biggest change first. Where you can, run independent edits from the same original and composite them, rather than stacking edits on edits.

Upscaling is not detail recovery

An upscaler invents plausible detail. It does not recover detail that was never captured. A good one turns a soft image sharp and convincing, and it will also confidently invent the wrong lettering, the wrong eyelashes, and a fabric pattern that was not in the shot. Upscale last, look closely at anything that carries meaning, and expect one more round of inpainting afterwards.

The mindset shift

Treat generation as photography, not as ordering. You take a lot of frames, choose one, then do the darkroom work. The people producing consistently good AI images are not writing better prompts than you. They are doing four rounds of masking you never see.

BEFORE YOU MOVE ON

Someone inpaints a small, distant face in a wide landscape shot. The face comes back soft and mushy, while the same tool fixes faces beautifully in close-up portraits. What is the most likely cause?

A. The denoising strength was set too low, so the model preserved the blur it started from.

B. The masked region is being generated at a small fraction of the model's native resolution, so the face is only a few dozen pixels wide when the model works on it.

C. The prompt described the whole scene instead of just the face, so the model's attention was divided across everything.

D. The model has a weak prior for distant faces, because training photos of faces are overwhelmingly close-ups.

30 · 8 min

Denoise strength: the number that decides how much survives

THE ONE THING TO KEEP

Image-to-image adds noise to an existing picture and denoises from there, so the strength value sets how far back along the schedule you start and therefore how much of the original can survive.

One number, and what it means

Every tool that transforms an existing image has a slider called strength, denoise, or image weight. It is the most misunderstood control in the toolkit, and it has an exact mechanical meaning.

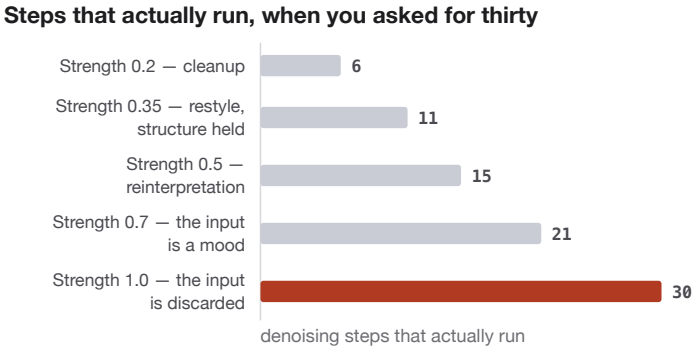
Recall that generation walks down a noise schedule from pure static to a clean image. Image-to-image does not start at the top. It takes your input image, encodes it to a latent, adds noise corresponding to some point partway down that schedule, and runs the loop from there.

Strength is where on the schedule you start.

```
strength 0.2  -> start near the bottom, 20% of the steps run
strength 0.5  -> start halfway
strength 0.9  -> start near the top, almost pure noise
strength 1.0  -> the input is entirely destroyed; this is ordi-
nary generation
```

Two consequences follow immediately, and both surprise people.

Steps and strength multiply. At 30 steps and strength 0.4, only about 12 steps actually run. This is why an image-to-image pass finishes faster than a fresh generation, and why setting a low strength with few steps produces a barely-changed image no matter what the prompt says.



Strength is where on the noise schedule the loop starts, so it multiplies with the step count. At 0.2 with thirty requested, six run, and no wording makes a difference the model has no steps left to make. At 1.0 the input is gone and this is ordinary generation.

Strength 1.0 ignores your input completely. People set it high because they want a big change, and at the top of the range they have simply thrown the image away. The band where interesting things happen is narrower than the slider suggests.

What each band is for

Ranges vary by model, and these are close enough to start from:

- **0.15 to 0.30.** Cleanup. Removes JPEG artefacts, unifies lighting, tidies edges. Composition and identity fully preserved. This is the pass that makes a rough composite look like one image.
- **0.30 to 0.50.** Restyling with structure held. A photograph becomes a painting, a sketch becomes a rendering. Faces survive at the lower end and start drifting toward the upper.
- **0.50 to 0.70.** Reinterpretation. The composition is a suggestion; the content is redecided. Useful for turning a rough block-out into a real image.
- **0.70 and above.** The input is a mood, not a plan.

A practical routine, which takes a minute and saves hours: fix the seed, generate the same image-to-image pass at 0.25, 0.4, 0.55 and 0.7, and look at all four. You will develop an intuition for the model's particular scale far faster than by reading anyone's recommended numbers.

The trap that catches everybody

Iterating image-to-image repeatedly degrades the picture in a specific way. Each pass adds noise and denoises, and each denoising pulls the result slightly toward the model's mode. Ten passes at 0.3 do not equal one pass at 3.0; they produce an image that has drifted toward the distribution average, with contrast climbing and detail smoothing on every round.

This is the same mechanism as the distribution pull from the previous module, applied repeatedly. It is also, incidentally, what makes the "keep re-uploading the same image" experiments circulating online converge on a similar glossy face regardless of where they started.

If you need several changes, make them in one pass at a suitable strength, or make them with masks so untouched regions are genuinely untouched.

Where it belongs in a workflow

The professional shape is nearly always the same:

1. Generate or photograph a base.
2. Fix the layout by hand — crop, move elements, paint over what is wrong, however crudely. A rough paint-over in GIMP or Krita costs two minutes and does not need to look good.
3. Run image-to-image at 0.3 to 0.45 to unify it.
4. Mask and inpaint the specific regions that are still wrong.
5. Upscale and finish.

Step 2 is the one people skip, and it is the one that makes the rest work. The model is much better at making a bad painting look like a photograph than it is at inventing a composition you have not shown it. A five-year-old's standard of paint-over is genuinely enough, because at strength 0.4 the model is reading shapes and colours, not craftsmanship.

The honest limitation: strength is a single global number. It cannot preserve the face and restyle the background, because it applies to the whole latent. When you find yourself wanting two different strengths in one image, you want masks, which is the next lesson.

BEFORE YOU MOVE ON

An image-to-image pass at strength 0.85 returns something almost unrelated to the input picture, though the prompt describes the input accurately. Why?

A. The prompt overrode the image conditioning because text conditioning always takes precedence over image conditioning in the pipeline.
B. The input image was encoded at a different resolution from the output, so the latent was resampled and its structure lost.
C. At 0.85 the run starts near the top of the noise schedule, so almost none of the original latent survives to influence the result.
D. The model applied the maximum guidance scale automatically at high strength values, which drove the output toward the prototypical version of the prompt.

31 · 8 min

Masks, feathering and the context window

THE ONE THING TO KEEP

Masked regeneration replaces only the masked latent while conditioning on the surrounding pixels, so what the model can see around the mask determines whether the repair matches.

What a mask actually does

Inpainting is the same denoising loop with one addition. At every step, the parts of the latent outside your mask are overwritten with the correctly-noised version of the original image. Only the inside of the mask is free to change.

That means the region being regenerated is surrounded, at every step, by the real picture. The model is not repairing in isolation; it is completing a scene it can see. This is why inpainting produces results that match lighting and

colour without being told to, and why it is a different and far more reliable tool than regenerating the whole image and hoping.

The three settings that decide the result

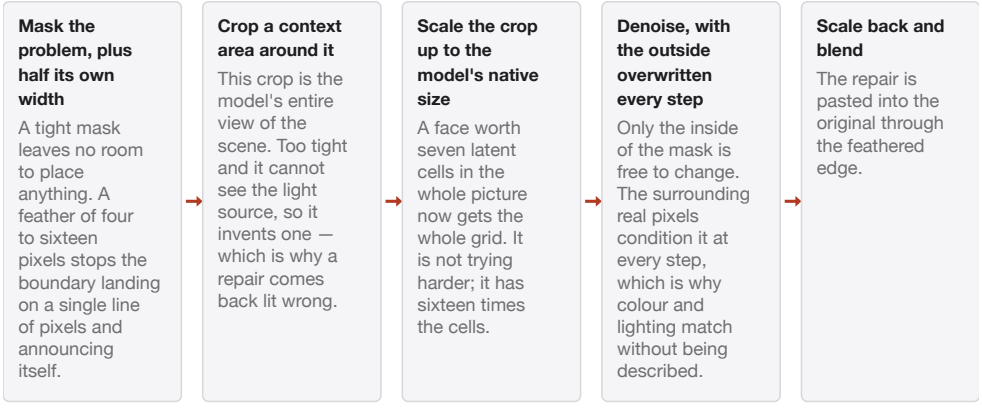
Mask size and shape. Too tight and the model has no room to place anything; the fix is squeezed into an area the wrong shape. Too loose and you have thrown away good pixels. A useful default is to mask the problem plus roughly half its own width of surrounding area.

Feather or blur. A hard-edged mask leaves a visible seam because the boundary between generated and original falls on a single pixel line. A feather of 4 to 16 pixels blends the transition. This one setting is the difference between a repair that reads and one that announces itself.

Context area — the setting nobody explains. Most implementations do not process the whole image for an inpaint. They crop a region around the mask, scale it to the model's native size, generate, and paste back. The size of that crop is the model's entire view of the scene.

This last one explains the most common inpainting failure. Mask a face in a large image with a tight context crop, and the model sees a face-sized rectangle with no shoulders, no horizon and no light source. It has no idea which way the light falls, so it invents one. Widen the context and the same mask produces a correct result. In WebUI this appears as "only masked padding"; in ComfyUI it is the crop size on the inpaint node. When a repair comes back lit wrong, this is nearly always why.

One inpaint, and where the extra detail comes from



Outpainting is the same machinery pointed outwards: pad the canvas, mask the new area, and extend in steps of 128 or 256 pixels so each step still has context to hold on to.

Why the crop is also a resolution gift

There is a bonus in the crop mechanism worth exploiting deliberately. Because the cropped region is scaled up to the model's native size before generation, a small masked area is being generated at *full model resolution* and then scaled back down.

A face that occupies seven latent cells in the full image occupies the whole latent grid when inpainted with a tight crop. This is the mechanism behind face-detailer workflows, and it is why inpainting a small element gives dramatically more detail than the same element in the original generation. It is not that the model tries harder. It has sixteen times the cells.

Outpainting is the same thing pointed outwards

Extending an image beyond its borders uses identical machinery: pad the canvas, mask the new area, inpaint with the existing image as context. The failure modes are the same. Extend too far in one pass and the far edge has no context and drifts; extend in steps of 128 or 256 pixels and each step stays anchored.

A practical note that saves a lot of confusion: many models are markedly worse at outpainting downward than in other directions, because photographs

are cropped at the bottom far more consistently than at the top, so the training data contains fewer examples of what continues below a frame.

Where inpainting stops being the answer

Inpainting is the right tool for a wrong region. It is the wrong tool for a wrong picture. If the composition is not working, no amount of masked repair will fix it, and you will spend an hour discovering that.

There is also a class of problem inpainting handles badly: anything requiring agreement with a region you are not regenerating. Fixing one earring to match the other means the model must reproduce the other, which it can only see as context. It often gets close and rarely gets exact. For symmetry, the reliable fix is mechanical — copy the good one, flip it, place it, blend — in GIMP, Krita or Photopea, all free.

The general rule that emerges across this module: use the generative tool for texture and content, and use ordinary image editing for geometry and symmetry. Each is far better than the other at its own job, and most of the frustration people report comes from asking one to do the other's work.

BEFORE YOU MOVE ON

An inpainted face comes back correctly rendered but lit from the wrong side. Which setting is the first thing to change?

A. The context crop around the mask, since the model can only match lighting it can actually see.

B. The mask feather, since a hard edge causes the lighting mismatch to appear at the boundary between regenerated and original pixels.

C. The denoise strength, since a lower value would retain more of the original lighting inside the masked region.

D. The guidance scale, since higher guidance would make the model follow a lighting description in the prompt more closely.

32 · 9 min

Conditioning on structure: depth, edges and pose

THE ONE THING TO KEEP

A control network injects a spatial map into the denoising loop at every step, fixing geometry while leaving content and style to the prompt, and each map type constrains a different thing.

The mechanism, briefly

A structural control takes an image-shaped map — an edge drawing, a depth field, a stick-figure skeleton — and feeds it into the denoising network alongside the text conditioning, at every step and at matching spatial positions.

The crucial difference from a text prompt is that the map is *spatially aligned* with the latent. Cell by cell, the network is told what should be there. There is no attention competition, no binding problem, no reliance on the caption corpus containing a word for the arrangement you want. This is why structural conditioning solves the problems the previous module said prompting cannot.

What a depth map pins, and what it leaves alone

Fixed by the map	Still entirely the prompt's business
• Where every surface sits in space	• What the objects are made of
• Camera position and perspective	• Colour, light and time of day
• The silhouette of each object	• The medium — photograph, gouache, linocut
• Relative scale between objects	• How many fingers are on the hand
• The composition you sketched in twenty seconds	• Whether a sign spells anything

The map is spatially aligned with the latent, cell by cell, so there is no attention competition and no reliance on the caption corpus having a word for your arrangement. That is why conditioning solves what prompting cannot, and why it fixes none of the failures that were never spatial.

The best-known implementation is ControlNet, which trains a copy of part of the base model to accept the map. Others exist under different names — T2I-Adapter, and various built-in equivalents in hosted tools. The behaviour is

close enough that the choice of map matters more than the choice of implementation.

Which map for which job

Canny edges. A hard black-and-white line drawing produced by an edge detector. Constrains outlines exactly. Use for: redrawing an existing image in a new style while keeping every contour, turning a line drawing into a rendered image, architectural work. Limitation: it is *strict*, so any noise in the edge detection becomes a line the model dutifully draws.

Depth. A greyscale map where brightness is distance. Constrains three-dimensional arrangement while leaving surface detail free. Use for: keeping a composition and camera while changing everything about the content. This is the most generally useful map and the best first choice.

Pose. A skeleton of joint positions. Constrains only the figure's arrangement, not its size, clothing or surroundings. Use for: putting a specific gesture or stance into an otherwise free image. Limitation: it carries no depth, so a figure can come back facing the wrong way, and hands are only a few points.

Soft edge and scribble. Loose line maps that constrain suggestion rather than outline. Use for: turning a rough sketch into an image without the model copying your bad lines.

Segmentation. A map of coloured regions, each colour meaning a category. Use for: laying out a scene by region — sky here, building there.

Normal maps and line art cover more specialised cases, and the same principle applies throughout: pick the map that constrains what you care about and nothing else.

The strength and timing settings

Two controls decide how hard the map is applied.

Conditioning strength. At 1.0 the map is followed closely; at 0.4 it is a suggestion. Very high strength on a noisy map produces artefacts, because the model is being forced to render detection noise.

Start and end steps. Applying the map only for the first 60% of steps fixes the composition and then releases the model to render freely. This is a genuinely useful trick: it gives you the layout without the map's stiffness, and it usually improves texture.

Where the maps come from

You do not need a photograph. All of these are useful sources:

- **A rough sketch,** drawn in Krita, GIMP or on paper and photographed. Scribble and soft-edge maps are designed for exactly this.
- **A 3D block-out** in Blender, which is free and gives you an exact depth map and exact camera control. For product shots and architecture this is the professional route and it is not difficult at the block level.
- **A photograph you took,** run through a depth estimator. Most tools include one.
- **A pose you built** in a free posing tool or extracted from any reference photograph.

The general point: you supply the geometry from something that is good at geometry, and the model supplies the surfaces, which is what it is good at.

The honest limits

Structural conditioning constrains where things are. It does not constrain what they are, and it does not fix any of the failures from the previous module that are not spatial. A depth map of a hand does not produce five fingers. A pose does not produce a consistent face. An edge map does not spell.

It also introduces its own failure: **over-constraint**. A model given a strong edge map of a photograph and asked for an oil painting will produce a photograph with paint texture, because every contour is pinned. If your restyling looks like a filter rather than a reinterpretation, lower the strength or switch from canny to depth. The fix is almost always to constrain less.

BEFORE YOU MOVE ON

You want to reinterpret a photograph as a woodcut print, keeping the composition but not the photographic detail. Which conditioning choice fits best?

A. Canny edges at full strength, so every contour of the original is preserved exactly in the print.

B. A pose map, since the arrangement of the subject is what needs to carry over.

C. An image-to-image pass at strength 0.9, using the original photograph as the input image and putting "woodcut print, bold carved lines" in the prompt to drive the change of medium.

D. A depth map at moderate strength, ending partway through the run, so the arrangement is fixed and the surfaces are free.

33 · 9 min

Teaching it something it does not know

THE ONE THING TO KEEP

A small fine-tune such as a LoRA adjusts a low-rank correction to the model's weights from a few dozen images, which is the only reliable way to add a subject, style or object the base model has no coverage for.

When a prompt cannot get there

The previous module drew a line between "underspecified" and "not there". Everything on the not-there side of that line needs training, not wording: your face, your product, a specific character, a regional garment the caption corpus never described, a house illustration style.

Training a whole model is out of reach. Training a small correction to one is not, and this is the single largest capability most people are unaware they have.

What a LoRA is

Full fine-tuning updates every weight in the model — billions of numbers, enormous memory, and a separate multi-gigabyte file per subject.

A **LoRA** — low-rank adaptation — instead learns a small pair of matrices whose product is added to the original weights. The mathematics is that most useful adjustments are low-rank: they can be expressed as a small correction rather than a full rewrite. In practice the trained file is 10 to 200 megabytes rather than 6 gigabytes, several can be loaded at once, and each can be applied at an adjustable strength.

The related methods are worth distinguishing:

- **Textual inversion.** Learns a new *word* rather than new weights — a few kilobytes, very fast, limited to concepts the model can already almost express.
- **DreamBooth.** Full fine-tuning on a small set with a preservation term. Higher fidelity, large files, more compute.
- **LoRA.** The middle, and where nearly everyone lands.

What training actually requires

The numbers, so you can judge feasibility:

- **A subject** (a face, a product, an object): 15 to 30 images.
- **A style**: 30 to 100 images.
- **Time**: 10 to 40 minutes on a mid-range consumer GPU; a free Colab session is enough for a small one.
- **Cost**: nothing, if you have a suitable computer or use a free tier.

The images matter far more than the settings, and this is where almost every failed attempt goes wrong.

Vary everything except the subject. Different backgrounds, lighting, distances, angles, clothing. If every photograph was taken in the same room, the model learns the room. This is the classic failure: a LoRA that reproduces a face beautifully and always with the same bookshelf behind it.

Crop and resize to the training resolution — square or the model's bucket sizes. Sloppy crops teach sloppy framing.

Caption honestly and describe the variable parts. Caption what changes between images (the background, the pose, the clothing) and *not* the thing you are teaching. Anything you leave uncaptioned is absorbed into the trigger word. Caption "a photograph of sks person in a garden" and the garden is described, so the garden stays variable; leave the garden out and the model may bind it to the subject.

Twenty good images beat two hundred mediocre ones. Duplicates and near-duplicates cause the memorisation effect from module one, at small scale and quickly.

Over-training, and how to see it

The most common outcome of a first attempt is a LoRA that has learned too hard. Symptoms: the subject appears correctly and everything else in the image degrades; prompts stop working; every output has the same composition; the style bleeds onto objects it should not touch.

The fix is to save checkpoints during training — most trainers do this by default, every few hundred steps — and test several. The best one is usually not the last. Generate the same three prompts with each checkpoint at the same seed and pick by eye. This takes ten minutes and is the difference between a usable LoRA and a broken one.

Applying strength below 1.0 at generation time is the other half of the control. A slightly over-trained LoRA at 0.7 is often perfect.

The part that is not technical

Training on somebody's face requires their consent, and the same standard applies as for voice, which the voice module covers in full: informed, specific, written, and revocable. Training on a living artist's work to reproduce their style is legal in some places, contested in others, and the subject of the final module. What is not contested anywhere is that publishing the result as though it were theirs is passing off.

Training on your own photographs, your own products, your own drawings and openly licensed material is uncomplicated, and for most commercial work it is also what you actually need. A LoRA of your client's product photographed from twenty angles solves a problem no prompt can, and nobody has any claim on it.

BEFORE YOU MOVE ON

A LoRA trained on 25 photographs reproduces the subject accurately but places them in nearly the same setting every time. What went wrong?

A. The training resolution was too low, so the model could not separate the subject from its surroundings during training.

B. The LoRA rank was set too high for a set this small, giving it enough capacity to memorise whole training images rather than isolating the subject that was supposed to be learned.

C. The training images shared a background and the captions did not describe it, so the setting was absorbed into the learned concept.

D. The LoRA was applied at strength 1.0 at generation time, which forces every learned feature to appear.

34 · 8 min

The same face twice

THE ONE THING TO KEEP

Character consistency has no exact solution because nothing in the pipeline stores an identity, so the working methods all trade fidelity, effort and flexibility against each other.

Why this is genuinely hard

Every generation starts from noise and is steered by conditioning. There is no persistent object anywhere that means "this person". A face that appeared in one image exists only as pixels in that image.

So consistency is always reconstructed, never retrieved, and every method is an approximation. Anyone who tells you a tool has solved it is describing an approximation they have not stressed.

The demand is real, though. Comics, children's books, explainer videos, brand mascots, storyboards, product ranges — all of them need the same thing to appear repeatedly and be recognisably itself.

The methods, honestly rated

Fixed seed and identical prompt. Free, instant, and only works while nothing else changes. Change the pose or the setting and the face changes with it. Useful for variations on one shot, useless for a sequence.

Describe the character in detail and reuse the description. A very specific description — age, face shape, hair, eye colour, a distinctive feature — narrows the distribution enough that the outputs are cousins. They are not the same person. Adequate for background characters, not for a protagonist a reader will look at forty times.

A trained LoRA. The strongest general method. Twenty to thirty images of the character in varied settings, trained as in the previous lesson. The problem is circular for an invented character: you need images of someone who does not exist. The standard route out is to generate a set, fix the best ones by hand until they genuinely match, train on those, then regenerate a better set with the LoRA and retrain. Two rounds usually suffice.

Face-swapping and identity adapters. A method that extracts a face embedding from one image and injects it during generation. Fast, no training, and it holds the face while leaving everything else free. The limitations are specific: it carries the face and not the body, hair or clothing; quality drops at unusual angles; and it makes generating a convincing image of a real person from one photograph very easy, which is why several services restrict it.

Reference-image conditioning in hosted tools. Convenient, closed, and version-dependent — it can change under you between releases, which is the reproducibility problem again.

Draw or photograph the character once and use structural conditioning. For sequential work this is often the fastest honest route. A puppet built in Blender, or a real photographed model, gives you exact pose and identity, and the generative pass supplies the rendering.

What professionals actually do

The workflow used in production is a combination, and it is worth stating plainly because it removes the expectation of a single button:

1. Establish the character with a LoRA or a strong reference set.
2. Generate each shot with the character's conditioning applied.
3. Fix the face in every shot with a masked detail pass.
4. Colour-match all the shots to each other afterwards in a free editor, which does more for perceived consistency than most people expect.

Step 4 matters more than it sounds. A large part of what reads as "the same character" is consistent colour and light, not exact geometry. Two images with matched grade and different face proportions read as more consistent than the reverse.

What still does not work

Clothing and props remain the weakest link. Faces are now reasonably solvable; a specific jacket with a specific number of buttons, across twelve shots, is not. The reliable answer is to make the costume simple and distinctive — one strong silhouette, one colour, one accessory — which is exactly what character designers have always done for animation, and for the same reason: consistency is cheaper when there is less to be consistent about.

Hands holding specific objects, characters interacting, and any shot where two of your consistent characters appear together are the remaining hard cases. Budget for hand fixing, and expect to composite two single-character generations for the two-shot rather than to generate it.

The honest summary: consistency is achievable to a standard that survives a reader's attention, with effort, using several methods together. It is not achievable by prompting, and it is not yet a solved feature in any tool regardless of what the marketing page says.

BEFORE YOU MOVE ON

Why does an identity adapter that holds a face across shots still fail to keep a character consistent in a comic?

A. Identity adapters degrade with each successive generation, so consistency decays across a long sequence of panels.

B. It carries facial identity only, leaving hair, clothing and props to be re-decided in every panel.

C. It requires a fixed seed, and a comic needs different seeds for different panels.

D. It operates after generation as a post-process, so it cannot influence anything about the panel other than the pixels of the face region itself.

35 · 8 min

What an upscaler invents

THE ONE THING TO KEEP

Upscaling adds plausible detail rather than recovering real detail, so it improves photographs and fabricates evidence, and the distinction matters most in exactly the cases people reach for it.

Two different operations with one name

Interpolation — bicubic, Lanczos — estimates intermediate pixels from neighbours. It cannot add information. A blurred photograph enlarged this way is a larger blurred photograph. Every image editor has done this for thirty years.

Generative upscaling does something categorically different. A model trained on pairs of small and large images predicts what the large version would look like. Where the small image has a grey smudge, the model produces the texture that most often appears where such a smudge appeared in training: brick, or fabric, or an eye.

The output looks sharper because it *is* sharper. The detail is real detail of a plausible image. It is not detail of your image, because that information was never in the file.

Why this matters more than an aesthetic preference

Two consequences.

For creative work, it is straightforwardly good. If you generated the picture, invented detail is no less legitimate than the rest of it. Upscaling a generated image to print size is the normal final step and there is nothing to worry about.

For anything evidential, it is dangerous. A CCTV frame upscaled produces a face. The face is what a model thinks belongs in a smudge of that shape. It is not the person's face, and it can be a different person's face entirely. Published demonstrations have shown upscalers turning low-resolution images of well-known faces into confidently wrong different people, and the failure is not a bug — it is the model doing exactly what it was trained to do.

The rule worth carrying: **an upscaler is a plausibility engine, not a recovery engine.** Never present an upscaled image as evidence of what was in the original, and be sceptical when anyone does.

The families, and when to use each

Convolutional upscalers — the ESRGAN family and its many community variants, all free. Fast, deterministic, faithful to the input's structure. They add texture without changing content. Specialised variants exist for faces, anime, text and photographs, and choosing the right one matters more than choosing the newest.

Diffusion upscalers. A generation pass conditioned on the small image, usually at low denoise. Much stronger detail invention, much more drift. At denoise above about 0.4 they cheerfully add objects that were not there.

Tiled upscaling. Cut the image into overlapping tiles, upscale each, blend. This is how a 6-gigabyte-of-memory card produces a 6000-pixel image. The characteristic failure is tile seams and repeated content — a face appearing in each tile, because each tile is independently a plausible place for a face. Overlap and a low denoise value control this.

Face restoration — CodeFormer, GFPGAN and similar, all free. Purpose-built for faces and very effective. They have a documented tendency to regularise a face toward the average, so an unusual face comes back more conventionally attractive and slightly less itself. There is a fidelity slider; the default is usually too aggressive.

A workable ladder

For most work:

1. Generate at native resolution.
2. Masked detail pass on faces and any small critical element.
3. Convolutional upscale to twice the size.
4. Optional low-denoise diffusion pass at 0.2 to 0.3 to unify texture.
5. Sharpen and grade in a free editor, last.

Doing the detail pass *before* the upscale is the ordering people get wrong. Upscaling a wrong face gives you a large wrong face, and the upscaler has no idea it was wrong.

The limitation nobody removes

Every upscaler has a training distribution, and it invents what that distribution contained. An upscaler trained on photographs will render text as text-like texture. One trained on anime will smooth a photograph's grain into flat colour. A face upscaler applied to a painting will make it look like a photograph of a face.

So the practical discipline is to match the upscaler to the material, and to look at the result at 100% rather than trusting that bigger is better. A great deal of the "AI slop" look in circulation is not the generator at all. It is an aggressive upscaler applied by default, adding invented micro-detail everywhere, and turning it off is often the single biggest improvement available.

One habit worth borrowing from print. Work out the size you actually need before upscaling anything. A 300 dots-per-inch A4 page is 2480 by 3508 pixels; a full-page magazine image is not much more; most screen use never exceeds 2000 pixels on the long edge. People routinely upscale to 8000 pixels for a piece that will be seen at 1200, which quadruples the invented detail, the file size and the processing time for no visible gain. Decide the output dimension first and stop there.

BEFORE YOU MOVE ON

A journalist upscales a low-resolution security image and obtains a clear face. What is the correct assessment?

A. The face is a plausible reconstruction generated from training data, not recovered information, and cannot identify anyone.
B. The face is reliable if the upscaler was trained specifically on facial imagery, since such models are optimised to recover facial structure accurately from limited pixel information.
C. The face is reliable at low upscaling factors and unreliable above four times.
D. The face is reliable provided the upscale was run at a low denoise strength.

36 · 8 min

Compositing is still the job

THE ONE THING TO KEEP

Generated elements assembled by hand in an ordinary editor beat single-pass generation for anything that must be exact, which is why the finishing skill matters more than the prompting skill.

The professional's actual workflow

There is a widespread assumption that using these tools well means getting a finished image out in one go. Almost nobody working commercially does that. The output of a generative model is treated as **stock** — raw material, gathered and then assembled.

A typical commercial piece is:

- A generated background, upscaled and graded.
- A subject generated separately, cut out.
- Two or three small elements generated or sourced individually.

- Type set in a real typesetting tool.
- Colour, grain and grade applied over the whole thing.

Every one of these steps happens in an ordinary image editor. None of them is a prompting problem.

Why assembly beats generation for exactness

The reasons are the mechanisms from the previous module, all at once. Counting works when you place the objects. Symmetry works when you flip a copy. Relations work when you position things. Text works when you set it. Colour matching works when you sample it.

Every failure that comes from the absence of a global constraint disappears when a human supplies the constraint. That is not a workaround. It is the correct division of labour: the model produces surfaces and textures that would take hours to paint, and you produce the structure it cannot hold.

The free tools, named

Everything here can be done for nothing, and the free tools are genuinely capable rather than a consolation:

- **GIMP** — layers, masks, selections, colour. Free, offline, runs on modest hardware.
- **Krita** — better for painting and drawing over generations, excellent brush engine, free, and its AI Diffusion plugin runs local generation inside the canvas.
- **Photopea** — runs in a browser, opens PSD files, and behaves like the industry tool closely enough that skills transfer in both directions. Free with advertisements.
- **Inkscape** for vector work, **Scribus** for print layout, **Blender** for 3D and its surprisingly good compositor.

The paid names — Photoshop, Illustrator, InDesign, After Effects — matter because job advertisements name them, and it is worth being able to say you

know what they do. The operations are the same operations. Someone who can composite in GIMP can composite in Photoshop within a day.

The five techniques that carry most of the work

If you learn only a handful, learn these:

1. **Layer masks.** Never erase; hide. Everything stays editable, which means every decision is reversible.
2. **Cutting out with a decent edge.** Selection, refine, a small feather, and attention to hair. The single most common tell of an amateur composite is a hard cut edge.
3. **Matching light and colour.** Sample the background's shadow and highlight colour and grade the subject toward it. Most "it looks pasted on" problems are colour, not geometry.
4. **Adding a unifying pass.** A shared grain, a shared slight blur, a shared grade over everything. This is what makes disparate sources read as one photograph.
5. **A final low-denoise generative pass** at 0.15 to 0.25 over the assembled image, which smooths the remaining inconsistencies. This is the one step that did not exist before, and it is the reason composites are now much faster than they used to be.

The argument for learning this properly

The generative part of this work is the part being automated hardest. Each model release makes the prompt matter less. What does not get easier is knowing that the shadow is falling wrong, that the edge is too hard, that the type is set badly, that the colour is off by a degree.

Those judgements are ordinary craft, they are the same judgements a retoucher made in 1995, and they are what the person paying you is actually buying. The rest of this track — design foundations, image editing, vector work, print — is exactly this material, and it is worth more than another year of prompt vocabulary.

The honest limitation of compositing: it is slower per image and it does not scale to volume. If you need four hundred variations, you need generation and automation, not assembly. Know which job you are on. Assembly for the hero image; generation for the long tail.

There is a second limitation, and it is about records rather than craft. A composite has several sources, and each of them carries its own rights position — one generated, one licensed stock, one photographed by the client, one taken from a reference the client supplied without saying where it came from. When a question arises months later it arrives about the finished piece, and the only way to answer it is a layered file with named layers and a note saying where each element came from. Flattening a composite and keeping only the export destroys that record permanently. Keep the layered file.

BEFORE YOU MOVE ON

Why is a hand-assembled composite usually more reliable than a single-pass generation for a commercial hero image?

A. Composites use higher-resolution source material, so the final file exceeds what a generative model can produce in one pass.

B. The human supplies the global constraints — count, position, symmetry, type — that the generation process has no mechanism to enforce.

C. Editors apply corrections that generative models cannot, because image editors work directly on pixels rather than on a latent representation.

D. Composites avoid the model's training distribution entirely, so they do not carry its aesthetic bias.

37 · 8 min

What runs on the machine you have

THE ONE THING TO KEEP

Local generation is possible on modest hardware through quantisation, smaller models and tiled processing, so the choice between hosted and local is a real one rather than a matter of budget alone.

The question everyone asks first

Can I run this on my laptop? The honest answer is: some of it, and more than you would expect, and the constraint is video memory rather than anything else.

Rough figures, for the diffusion family:

- **4 GB of video memory** runs the older 512-pixel models comfortably, with community optimisations.
- **6 to 8 GB** runs 1024-pixel models such as SDXL and its many fine-tunes.
- **12 GB** runs most current open models in quantised form.
- **24 GB** runs the largest open image models at full precision and makes video generation possible.

Apple Silicon Macs use shared memory and work well from 16 GB upward. A machine with no discrete graphics card can still generate on the processor alone, at roughly one to three minutes per image rather than seconds — slow, and entirely usable for someone producing a few images a day.

The techniques that lower the requirement

Quantisation. Model weights are stored at reduced precision — 8-bit or 4-bit instead of 16. A 12-billion-parameter model that needs 24 GB at half precision fits in 8 to 12 GB quantised, with a quality loss that is measurable and, at

8-bit, usually invisible. Community-distributed quantised files are standard and free.

Sequential offloading. Parts of the model are moved between system memory and video memory as needed. Slower, and it lets a large model run on a small card at all.

Tiling. The decoder is the memory spike at high resolutions. Processing it in tiles removes the spike at the cost of a little time.

Smaller and distilled models. A well-chosen 2-billion-parameter model produces better results than a badly-driven 12-billion one, and runs four times faster.

On a phone

This is worth stating because much of this readership has a phone and no laptop. Local generation on a phone is real: apps exist that run small diffusion models entirely on the device, offline, for free, taking tens of seconds per image on a recent handset. Quality is behind the large hosted models and ahead of what was available on a workstation three years ago.

For everything else, hosted free tiers are the route, and the practical advice is to treat free generation quota as a scarce resource: plan the image, use a fixed seed, and do the assembly and finishing in Photopea, which runs in a phone browser.

Choosing between local and hosted

The trade is not really about money, since both have free options. It is about four things:

Reproducibility. Local models do not change under you. Hosted ones do.

Privacy. A client's unreleased product photograph uploaded to a hosted service has left your control, and many contracts prohibit exactly that. Local generation never sends the image anywhere.

Control. Local tools expose every parameter, allow arbitrary conditioning combinations and custom fine-tunes, and impose no content filter — which is both the advantage and the responsibility.

Quality and convenience. The largest hosted models are ahead of the best open ones on several axes, and they require no maintenance.

Most working people use both deliberately: hosted for exploration and for the things the big models do best, local for client-confidential work, for anything that must be reproducible, and for volume.

Getting started without spending anything

A concrete route, in order:

1. Install **ComfyUI** or a WebUI. Both are free and both have one-click installers now.
2. Download one well-regarded open checkpoint suited to your memory.
3. Learn the five settings from this module: steps, guidance, seed, denoise strength, and one conditioning map.
4. Add **Krita** with its diffusion plugin if you prefer painting to node graphs.
5. Use free Colab or Kaggle notebook time for the occasional job too large for your machine, and for training a LoRA.

The honest caution: local setup consumes an evening the first time, and dependency problems are real. The payoff is a toolchain that costs nothing per image, works offline, keeps your work private, and does not change without your permission. For anyone doing this regularly, that is worth an evening.

One more consideration people rarely weigh, and should. Generation uses electricity, and a graphics card under load draws a few hundred watts. A single image is a trivial amount — comparable to running a desk lamp for a few minutes — and a habit of generating two hundred images to pick one is not trivial, either in cost or in the emissions behind it. The discipline that improves your results is the same one that reduces the waste: think before generating, fix the seed, change one thing, and stop when the picture is right rather than when the batch finishes. Careful work happens to be the cheap kind.

BEFORE YOU MOVE ON

Why does quantising a model to 8-bit allow it to run on a smaller graphics card?

A. It reduces the number of denoising steps required, which lowers peak memory during the run.
B. It reduces the output resolution the model produces, so the latent and decoder need less memory.
C. It stores each weight in fewer bits, so the same parameters occupy substantially less video memory.
D. It removes the least useful parameters from the model, shrinking the file and the memory footprint together.

CASE STUDY · MODULE 4

Sixty old product photographs and a trade fair in nine days

A knitwear exporter in Ludhiana, deciding whether to upscale a 2011 photo archive or reshoot it

Harbans Knitwear has been selling cardigans and pullovers to domestic wholesalers for thirty years. Gurpreet Sandhu, who runs it now, had nine days until a trade fair in Delhi and no catalogue.

What he had was the website his cousin built in 2011: sixty product photographs, six hundred pixels on the long edge, shot against a bedsheet. On a phone they look acceptable. At A4, which is about two and a half thousand pixels wide at print resolution, they are unusable.

Reshooting meant pulling sixty garments out of the warehouse, two days of a photographer at Ludhiana rates, and the packing staff working around it in their busiest month. It also meant discovering that eleven of the sixty styles had been discontinued and no longer existed in the building or anywhere else.

The design studio he hired proposed something cheaper. They ran the six-hundred-pixel files through a generative upscaler and sent back sixty images at print size, sharp, clean and convincing. The quote for the whole catalogue dropped by about two-thirds.

Meena Kaur, who has run production for eleven years, opened one at full size and said the stitch was wrong.

The original photograph showed a cable-knit cardigan at six hundred pixels, where the cable is a grey suggestion of texture rather than a countable pattern. The upscaled version had a clean, confident, entirely plausible cable — with a different repeat, a different rope width and a different crossing rhythm from the cable the factory actually knits. The button had become a different button. The cuff rib had become a deeper rib than the garment has.

Gurpreet's first instinct was that this did not matter much, and it is worth sitting with why, because it is the instinct most people have. The picture looked

better. It looked more like a good cardigan than the original did. Nothing in it was ugly or obviously wrong.

Meena's objection was commercial rather than aesthetic. A buyer at a trade fair points at a photograph and orders four hundred pieces against a style number. If the photograph shows a cable the factory does not knit, the shipment that arrives is not the one that was ordered, and a buyer who has to explain that to their own customer does not come back.

So the decision had a real cost on both sides, nine days out.

Shipping the upscaled catalogue meant a book that looked better than anything Harbans had ever printed, delivered on time and within budget, carrying sixty photographs of garments that were slightly not the garments. Reshooting meant thirty-five thousand rupees, a day and a half the packing bay could not spare, eleven styles with no photograph at all, and a catalogue that would look plainer than the competition's on the next table.

Meena proposed a test before anyone decided. It took forty minutes.

She took one cardigan that still existed, photographed it properly on the shop floor, and downscaled her own photograph to six hundred pixels. Then she ran that through the same upscaler and laid the result beside the real photograph at full size.

Eight visible differences. Four of them were on things a buyer specifies in an order: the stitch repeat, the rib depth at the cuff, the button, and the woven label, which had come back with confident lettering that was not the company's name.

That settled it, and it settled it with evidence rather than with an argument about whether AI is trustworthy.

STOP HERE

Before turning the page: what would you have done, and what would it have cost if you were wrong? Write it down. The point of the next section is to compare.

WHAT HAPPENED

They reshot the forty-nine styles that still existed, in a day and a half, in the packing bay. A grey paper sweep, two windows, a bedsheet clipped over one of them as a diffuser and a tripod. The photographer they had quoted was not needed for a garment flat-lay, and Meena took most of them herself.

The eleven discontinued styles went into the catalogue at the size the old photographs honestly supported — a seventy-millimetre square on the page, captioned as archive images of styles no longer in production — rather than enlarged to fill a page with detail nobody could order.

Generative tools did not leave the job. They moved to where they invent nothing a buyer can specify. Each new photograph went through an image-to-image pass at a denoise strength of about zero point two, which unified the light across two days of changing weather without moving a single edge. Each button got a masked detail pass, generated at the model's native resolution inside a tight crop and scaled back, which is where the detail the original frame could not hold actually came from. And the backgrounds were cleaned with masked regeneration rather than by hand.

The difference between those passes and the upscale is not the technology. It is that a buyer cannot order a background, and can order a cable.

The catalogue went to the fair. Gurpreet has since had the same conversation with two other suppliers who were offered the cheap version, and the round-trip test — photograph it properly, shrink it, upscale it, compare — is what he shows them, because it takes forty minutes and ends the argument.

Upscaling a generated landscape is uncontroversial and upscaling the cardigan was not. What is the distinction, stated precisely?

An upscaler is a plausibility engine, not a recovery engine: it produces the detail that most often appeared where a smudge of that shape appeared in its training data. For a generated landscape, invented detail is no less legitimate than the rest of the picture, because nothing in it is a claim about a real object. The cardigan photograph is a representation of a specific garment that a buyer will order against a style number, so invented detail is a false statement about a product. The same

operation is finishing in one case and fabrication in the other, and the difference is whether anything outside the file is supposed to match it.

Why was Meena's round-trip test more useful than the studio and the factory arguing about whether the upscaled images looked right?

Because it produced a ground truth. Looking at an upscaled image tells you whether the invented detail is plausible, which it always is — plausibility is what the model optimises. Shrinking a known photograph and upscaling it back gives you the real garment and the model's guess side by side at the same size, so every difference is a measured error rather than a matter of taste. It is the same move as running an image through an autoencoder and straight back out to see the pipeline's ceiling: construct a case where you already know the answer, then see what the tool says.

They ran a generative pass over the new photographs anyway, at a denoise strength of about 0.2, and a masked detail pass on every button. Why is that acceptable when the upscale was not?

Because of what each operation is allowed to change. A denoise strength around 0.2 starts the loop near the bottom of the schedule, so shapes and identity survive intact and only surface finish and lighting move — which is exactly the inconsistency two days of changing daylight produced. A masked detail pass regenerates a small region at the model's native resolution while conditioning on the surrounding real pixels, so the button is rendered with the cells it needs from a photograph that actually contains that button. Neither invents a specification. The upscale, by contrast, was asked to produce detail that was never captured at all, and it obliged.

MODULE 4 · TEST

Check what stuck

6 questions, one right answer each. This one is for you, not for a record: answers and the reason behind each are at the back. If two questions from the same module go wrong, re-read that module before the exam rather than after.

1 You set an image-to-image pass to 30 steps and a strength of 0.4, and it finishes much faster than a fresh generation at 30 steps. Why?

A. The input image is cached, so the first twelve steps are skipped.
B. Strength reduces the guidance scale, which halves the work per step.
C. Strength decides where on the noise schedule you start, so only about twelve steps run.
D. Image-to-image skips the decoder, which is the slowest part of the pipeline.

2 You inpaint a face in a large photograph and it comes back lit from the wrong side. What is the most likely cause?

A. The context crop was too tight, so the model could not see the light source.
B. The mask feather was too wide and blended the lighting from the wrong region.
C. The denoise strength was too low for the model to change the lighting.
D. The prompt described the scene rather than the region being repaired.

3 Inpainting a small face in a wide shot produces dramatically more detail than the same face did in the original generation. What makes the difference?

A. Inpainting runs at a higher guidance scale by default in most tools.

B. The cropped region is scaled up to the model's native size, so the face gets many more latent cells.

C. The mask restricts the loss to that region, so the model concentrates its capacity there.

D. Inpainting models are separately trained on close-up photographs of faces.

4 You condition on a canny edge map of a photograph and prompt for an oil painting. The result is a photograph with paint texture over it. What should you change?

A. Raise the guidance scale so the prompt overpowers the map.

B. Add 'photograph, photorealistic' to the negative prompt.

C. Increase the conditioning strength so the map is applied more decisively.

D. Lower the conditioning strength, or switch from edges to depth.

5 A LoRA trained on twenty-five photographs of a person reproduces the face well and always puts the same bookshelf behind them. What went wrong in the training set?

A. The learning rate was too high, so the model over-fitted to high-contrast regions.

B. There were too few images; a subject needs at least a hundred.

C. The photographs were taken in one room, and the captions did not describe the background.

D. The images were not cropped to the training resolution, so the framing was learned.

6 A low-resolution security camera frame is run through a generative upscaler and a recognisable face appears. What has the upscaler produced?

A. The texture that most often appeared where such a smudge appeared in its training data.

B. Detail recovered from information that was present in the file but below the display threshold.

C. An averaged reconstruction of all faces consistent with the camera's optics.

D. A sharpened version of the original pixels with no new content added.

MODULE 5

Time and space: video and three dimensions

Adding time to a generative model changes the arithmetic and adds a whole class of failure. This block covers why a second of video costs what it costs, the mechanisms that hold a moving picture together and the ones that let it flicker, how motion and camera are conditioned, why clips drift as they lengthen, how a talking head is made and what consent it requires, how to test a claim that a model understands physics, how to evaluate a video model honestly, what synthetic video means for believing anything, and the separate world of generated three-dimensional objects.

> BY THE END OF THIS MODULE YOU CAN
>
> Explain why generated video is short, expensive and inconsistent in terms of its underlying mechanism, design a fair test of a video model's claims, and say what a generated 3D asset is actually made of and where it can and cannot be used

38 · 8 min

Video, and what it still cannot do

THE ONE THING TO KEEP

Video models produce shots, not films; plan the edit and budget for the takes you throw away.

The extra dimension costs more than you think

An image model has to make one frame plausible. A video model has to make 150 frames plausible *and* mutually consistent. It cannot make them one at a time, because frame 80 has to know what frame 12 did. So it compresses time as well as space — a video autoencoder might squash eight times in each spatial direction and four times in time — and a transformer attends across that whole compressed block at once.

Attention cost grows steeply with the size of the block. That is the real reason clips are short. Five to ten seconds is where quality, price and attention length currently meet. It is a physics-of-the-method limit, not an arbitrary product decision.

What genuinely works now

Single shots up to about ten seconds, with a clear subject and a simple camera move, look good. Image-to-video — where you generate or photograph a still first, then animate it — is far more controllable than text-to-video, and it is what most working people use. Several models now produce synchronised dialogue and ambience along with the picture, which was science fiction two years earlier.

What still fails

Physics at the point of contact. Anything touching anything else. Hands gripping, feet meeting ground, liquid pouring, cloth folding, a knife going

through a tomato. The model learned what these look like, not what they are, and contact is where the difference surfaces.

Object permanence. When something leaves frame and comes back, it comes back subtly wrong. Crowds are the worst case; background faces churn continuously.

Identity across shots. Your character in shot three is a cousin of your character in shot one. Locked keyframes and reference features help. They do not solve it.

Timed action. "She looks up on the fourth beat" is not currently something you can ask for. You get an approximation and cut around it.

Text. Signage in video is roughly where image models were in 2022.

Length. You do not get a film. You get shots.

The number that matters is cost per usable second

List prices in 2026 sit roughly between ten cents and seventy-five cents per generated second, depending on model, resolution and tier. That number misleads, because what you actually spend is cost per second you can use. At a one-in-eight hit rate — normal once a shot has a specific requirement — an eight-second clip listed at two dollars really cost sixteen, plus an hour of your attention.

Budget for the takes you throw away. Teams that budget for the keeper run out of money at shot four of thirty.

A workflow that survives a deadline

1. Write the shot list first, as shots, with durations. Treat it like a real shoot, because it is one.
2. Make a still for every shot and get the still right. Image tools give you far more control than video tools do.
3. Animate one shot at a time, several takes each.

4. Assemble in an ordinary editor. Cuts hide an enormous number of sins. A long continuous move hides none.
5. Do sound separately unless the model's native audio happens to land.

Where the honest line sits

For mood pieces, abstract backgrounds, product motion and B-roll nobody studies frame by frame, this is production-ready and cheap. For anything with a named character doing specific things in a specific order, you are fighting the tool, and a small live shoot or 2D animation is often faster, cheaper and better.

Say that in week one rather than week four. Being the person who knows where the tool stops is more valuable than being the person who can drive it.

BEFORE YOU MOVE ON

A team plans a three-minute explainer as one continuous camera move through a factory, generated as a single piece. What is the strongest reason to change the plan?

A. No model accepts a prompt long enough to describe three minutes of action in adequate detail.
B. Three minutes of generation costs far more than thirty six-second clips would.
C. Consistency degrades over duration, because the model has no persistent representation of the objects it must keep identical from start to finish.
D. A single unbroken shot removes the cuts an editor would use to hide errors, so any mistake stays on screen.

39 · 8 min

Why a second of video costs what it does

THE ONE THING TO KEEP

Video generation is not image generation repeated, because the model must attend across frames as well as within them, so cost rises faster than the frame count and memory becomes the binding limit.

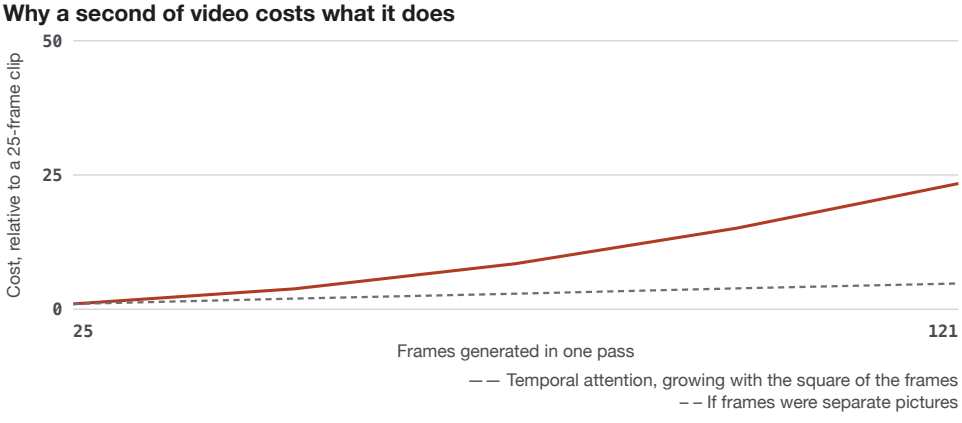
The naive arithmetic, and why it is wrong

A five-second clip at 24 frames per second is 120 frames. If video were images repeated, it would cost 120 times an image. That would be expensive and manageable.

It costs considerably more than that, and the reason is the part that makes video work at all.

A model that generated each frame independently would produce 120 unrelated pictures. To make frame 47 continue frame 46, the model must see other frames while generating each one. That means attention *across time* as well as across space, and attention is quadratic: doubling the number of frames quadruples the cost of the temporal attention layers.

Current video models handle a fixed window of frames jointly. Within that window, every latent position can attend to every other, across both space and time. A 121-frame clip at a 60 by 104 latent resolution has on the order of 750,000 positions attending to each other. That is the number that decides what hardware you need.



Every latent position attends to every other, across space and time, so doubling the frames roughly quadruples that part of the bill. The odd frame counts — 49, 81, 121 — come from the video autoencoder compressing four input frames into one latent frame, plus one.

Where the memory goes

The practical consequence is that video generation is memory-bound rather than time-bound. You can wait longer for a slow generation. You cannot wait your way past running out of video memory.

Rough current figures for open video models: a small one generating a few seconds at low resolution fits in 8 to 12 GB with quantisation; the larger open models want 24 GB and gratefully accept more. This is why video generation went to hosted services for most people first, while image generation went local.

The mitigations are the same family as for images, applied harder:

- **Quantisation** of the weights.
- **Tiled decoding**, because the video autoencoder's decode step is the single largest memory spike.
- **Fewer frames per pass**, generating in chunks and joining them.
- **Lower resolution**, then upscaling each frame afterwards — which works, and introduces its own flicker if the upscaler is not temporally aware.

The video autoencoder

Images use a spatial autoencoder that compresses by eight in each direction. Video autoencoders compress in time as well, typically by four: four input frames become one latent frame. A 121-frame clip becomes about 31 latent frames.

This is why video models produce frame counts in odd-looking numbers — 49, 81, 121 — rather than round ones. The counts are set by the temporal compression factor plus one.

It also introduces a specific artefact. Because four frames share one latent, fast motion within those four frames has to be reconstructed by the decoder from a single compressed representation. Rapid movement therefore comes back smeared or stuttering in a way that is not present in the latent itself. If your clip has a fast pan or a quick gesture and the result looks like a badly encoded video file, this is the mechanism, and generating at a higher frame rate does not fix it because the compression factor is fixed.

What this means for how you work

Three practical consequences follow directly from the arithmetic.

Generate short. The cost curve punishes length superlinearly, and the drift problem covered later in this module punishes it again. Five seconds is not an arbitrary limit; it is where several curves cross.

Generate small, upscale after. Resolution multiplies into the same attention cost. A 480-pixel generation upscaled to 1080 costs a fraction of a native 1080 generation and is frequently indistinguishable after grading.

Expect to throw work away. Because each generation is expensive, the temptation is to accept a nearly-right clip. The professional habit is the opposite: decide in advance what "usable" means for the shot, and discard anything that is not, because a flawed clip costs you far more in the edit than a regeneration costs in credits.

The honest limitation to carry: none of this is a temporary inefficiency waiting for better engineering. The quadratic term is in the architecture. Approaches

that reduce it — attending only to nearby frames, or compressing time harder — reduce the cost and give up exactly the long-range consistency that the attention was buying. That trade is the subject of the next lesson.

It is worth grounding the cost in something concrete, because "expensive" is not a number. On hosted services, generated video is currently priced in the region of a few cents to a few tens of cents per second of output, depending on model and resolution. Against a hit rate of one usable shot in six or eight attempts — which is a realistic figure for text-conditioned work and improves considerably with a first frame — a usable five-second shot lands somewhere between the price of a coffee and rather more. Multiply by the thirty shots a two-minute piece needs and you have the real budget. Doing that arithmetic before starting is the difference between a project that finishes and one that stops halfway.

BEFORE YOU MOVE ON

Why do open video models list frame counts like 49 and 81 rather than 48 and 80?

A. Those values are chosen to match the common video frame rates used in broadcast and film, so that generated clips divide evenly into whole seconds when placed on a timeline.

B. They are the maximum counts that fit in typical consumer video memory at the model's resolution.

C. The temporal compression factor of the video autoencoder produces counts of the form four times a number, plus one.

D. Odd frame counts guarantee a central keyframe, which the model uses as an anchor for temporal attention.

40 · 8 min

What holds a moving picture together

THE ONE THING TO KEEP

Consistency across frames comes from temporal attention over a limited window, so anything that must persist longer than that window has no mechanism keeping it stable.

The problem, stated exactly

Take an image model and generate 120 frames with the same prompt and 120 different seeds. You get a slideshow of unrelated pictures. Use the same seed for all of them and you get 120 copies of the same picture, which is a still.

Video needs the middle case: frames that are related but progressing. Producing that requires the model to know, while generating frame 47, what frames 40 to 60 look like.

Three architectural approaches have been used, and each one's characteristic failure is worth recognising.

Temporal layers bolted onto an image model. The earliest working approach: take a trained image model, insert layers that attend across frames, and train only those. Cheap, and it inherits the image model's quality. Its signature failure is *texture flicker* — the overall shapes hold while surface detail boils and shimmers, because the spatial layers were never trained to keep texture stable.

Joint space-time attention. The whole clip is one tensor and attention runs over space and time together, in a single trained model. Much more coherent, much more expensive, and this is what current strong models do. Its signature failure is at the window boundary rather than within it.

Causal or autoregressive video. Each chunk is generated conditioned on the previous one, in order. This allows arbitrary length and streams in real

time, at the cost of drift, which the extension lesson covers.

The window, and why it decides everything

Whichever architecture, the model has a finite attention window — the number of frames it can consider at once. Inside that window, consistency is enforced by the model's training. Outside it, nothing enforces anything.

This single fact predicts most of what people observe:

- A jacket keeps its buttons for four seconds and changes at five.
- A character walks out of frame and returns as someone else.
- A pan across a room shows a door that is no longer there when the camera comes back.
- Two clips generated from the same prompt do not match, and no setting makes them match.

Object permanence is the clearest case. Something occluded and then revealed must be reconstructed from information that has fallen outside the window. What comes back is a plausible thing of that kind, not the thing that went in.

What improves it, and what does not

Improves it: a stronger base model, generating within a single window rather than joining clips, simple scenes with few objects, and shots where nothing is occluded.

Does not improve it: prompting for consistency. `consistent lighting, same character throughout` addresses nothing mechanical. It is the same category of mistake as writing `exactly five` for a count.

Partly improves it: conditioning every chunk on a shared reference image, which anchors identity without enforcing continuity of everything else.

The flicker you can fix in the edit

Some temporal artefacts are cheaper to remove afterwards than to prevent.

High-frequency texture boil responds well to a light temporal denoise in a video editor — DaVinci Resolve's free edition has one, and so do free plugins for Kdenlive and Shotcut. It works because the flicker is uncorrelated between frames while the real image is correlated, so averaging across a few frames removes one and keeps the other.

Brightness flicker between frames is fixed by a deflicker filter, which normalises frame luminance. This is a solved problem borrowed from film restoration.

What cannot be fixed afterwards is content changing — a different jacket, a vanished object. That is a reshoot, which in this medium means a regeneration.

How to find a model's window

You are rarely told the number, and you can measure it. Generate the longest clip the model allows, of a subject with one distinctive, arbitrary detail — a patterned scarf, a coloured mug, a sticker on a laptop. Step through the output frame by frame in any free editor and note where the detail changes.

Do this three times and take the earliest failure rather than the average. That frame number, divided by the frame rate, is the honest length of a shot you can plan on that model. It is usually shorter than the maximum the interface offers, and knowing it converts a vague frustration into a production constraint you can design around.

The honest summary: consistency is bounded by the attention window, the window is set by the model and the memory available, and no amount of prompting extends it. Plan shots that fit inside it, exactly as a director plans shots that fit inside a magazine of film.

BEFORE YOU MOVE ON

A generated shot holds together for four seconds, then a character's clothing changes. What does this indicate?

A. A cut in the underlying model's output, at the point where two separately generated segments were joined together into a single continuous file.
B. The prompt did not specify the clothing precisely enough for the model to maintain it.
C. The temporal denoiser in the pipeline averaged across too many frames and lost the detail.
D. The earlier frames have passed outside the model's attention window, so nothing constrains the clothing any more.

41 · 8 min

Conditioning a clip: stills, frames and video in

THE ONE THING TO KEEP

A video model can be anchored by a first frame, a last frame, a full input video or a reference image, and each conditioning route constrains a different thing and fails differently.

Text alone is the weakest way to make a clip

Prompting a video model with text only asks it to decide the subject, the composition, the lighting, the camera and the motion, all at once, from a sentence. Every uncertainty in the image case is present, plus motion.

The result is the familiar experience: attractive clips that are not what you asked for, at a hit rate that makes any planned sequence impractical.

Every other conditioning route is better, and knowing what each one pins is the whole skill.

Image to video

Supply a still as the first frame. The model continues it.

This is the dominant professional workflow and the reason is simple: it moves the entire composition problem into the image domain, where you have masks, structural conditioning, fine-tunes, compositing and unlimited cheap iterations. You settle the frame using everything in the previous module, then ask only for motion.

What it pins: composition, subject appearance, colour, lighting, style at frame one.

What it does not pin: what happens next. The subject can still turn into someone else by second four. Motion is still decided by the model.

A practical detail people miss: the input still is often re-encoded, so the first frame of the output is not pixel-identical to your input. For a shot that must cut cleanly from a still, check this and, if needed, replace frame one in the edit.

First and last frame

Supply both ends. The model interpolates a plausible path between them.

This is the strongest control available for a specific action, and it is underused. It turns "make him pick up the cup" into "here he is with an empty hand, here he is holding the cup", which is a much better-posed question. It is also how you make a clip loop: use the same image for both ends.

Its failure is characteristic and worth expecting: given two ends that cannot be connected by a plausible continuous motion, the model produces a morph rather than a movement. Objects melt into each other. When you see morphing, the two frames were too far apart, and the answer is an intermediate frame rather than a better prompt.

Video to video

Supply a clip; get a restyled clip. Underneath this is usually per-frame structural conditioning — depth or edges extracted from each input frame — plus

the model's temporal layers.

What it pins: motion and geometry, exactly, because they come from the source.

What it does not pin: appearance stability, which is why early attempts at this flickered so badly. Current models are much better, and the remaining tell is texture that boils on surfaces the source video kept flat.

This is the route for anyone who can shoot. A phone video of the actual motion, restyled, beats generated motion for anything requiring specific action. It also solves the rights position for the performance, which the talking-heads lesson takes up.

Reference conditioning

Supply an image not as a frame but as a description of who or what should appear. Identity and style come from the reference; the shot is generated freshly.

Useful for keeping a character across shots that cannot be one continuous take. Weaker than a first frame, more flexible.

Choosing, in one line each

- Need a specific composition: **first frame**.
- Need a specific action: **first and last frame**, or **video to video**.
- Need the same character across several shots: **reference conditioning**, plus a face pass afterwards.
- Need a specific camera move: **video to video** from a phone clip or a Blender render.
- Need something to look nice and do not much mind what happens: **text only**.

There is one more route worth knowing, because it is free and often overlooked: **animating a still without a video model at all**. A slow push, a parallax move on a layered image, a subtle drift of clouds, a flicker of light — all of these are keyframe work in an ordinary editor, and they cost nothing,

never flicker, never drift, and are perfectly reproducible. A great deal of what people generate as video would be better made as a still with a move on it. Kdenlive, Shotcut and Resolve's free edition all do this; so does Blender. Ask whether the shot actually needs generated motion before spending generated motion on it.

The honest limitation across all of them: conditioning constrains the frames you supply and the frames near them. Nothing constrains the middle of a long clip. This is the same window limit as the previous lesson, and it is why the answer to almost every video problem in practice is "make the shot shorter and control both ends".

BEFORE YOU MOVE ON

A first-and-last-frame generation returns a smooth morph rather than a movement, with objects melting into one another. What is the likely cause?

A. The interpolation strength was set too low, so the model blended the two frames instead of animating between them.

B. The two supplied frames are too far apart for a plausible continuous motion to connect them.

C. The two frames were encoded at different resolutions, so the latent representations could not be aligned.

D. The model's temporal attention window was shorter than the clip, so the middle frames were generated without reference to either end.

42 · 8 min

Directing motion, and what actually lands

THE ONE THING TO KEEP

Camera and motion instructions work only where the training captions contained consistent film vocabulary, so standard shot language lands and invented descriptions do not, and explicit trajectory conditioning beats both.

Why some camera words work

The mechanism is the same as for image prompts. Video training data is captioned, increasingly by machine, and those captions describe motion using whatever vocabulary the captioner used. Terms that appear consistently in that corpus behave like reliable controls. Terms that do not are decoration.

In practice this means **standard film vocabulary works**, because it is what any competent captioner uses and what stock video libraries have always tagged with:

- `slow push in`, `dolly out`, `pan left`, `tilt up`, `crane down`
- `handheld`, `locked off`, `tracking shot`, `orbit`
- `rack focus`, `shallow depth of field`
- `slow motion`, `time-lapse`

And **invented or compound descriptions do not**: `the camera swoops dramatically around the subject and then rises to reveal the city` is one caption nobody wrote, and the model will pick whichever fragment it recognises.

The practical instruction is to learn a small amount of real film vocabulary and use it plainly, one move per shot. This is a genuine reason to read a cinematography primer: the words are the interface.

Subject motion versus camera motion

These are separate and models frequently confuse them. `moving left` is ambiguous — the subject, or the frame? Written captions resolve this by convention and the conventions are inconsistent.

Say which. `The camera pans left while the woman stands still` is unambiguous and lands far more often than `panning left`.

A useful diagnostic: if you cannot tell from your own prompt whether the background should move, neither can the model.

Explicit trajectory conditioning

Several systems now accept camera motion as data rather than words — a path, a set of positions per frame, or a drag gesture indicating where an element should travel. Where this exists it is categorically better than prompting, for the same reason a depth map beats "on the left": it is spatially explicit and does not depend on caption vocabulary.

Where it does not exist, the free substitute is video-to-video from a source clip with the camera move you want. That source can be a phone video, or a Blender render of a moving camera over grey blocks, which takes ten minutes to set up and gives exact, repeatable moves.

The motion-strength setting

Most image-to-video systems expose an amount-of-motion control. It is worth understanding what it trades.

Low motion gives you stable, coherent, nearly-still clips. It is also where models look best, and where the demonstration reels quietly live: a slow push on a static subject hides almost every temporal weakness.

High motion gives you movement and buys the whole class of failures — morphing, limbs multiplying, backgrounds reorganising. Fast motion is where the temporal compression from the first lesson smears, and where the attention window is stretched hardest.

The professional consequence: **design shots that need little motion**. This sounds like a constraint and it is also good practice — a held shot with a small camera move and one thing happening is a better shot than a busy one in any medium. Most people arrive wanting the model to do something spectacular and get much better results the moment they ask it for something restrained.

What no model does yet

Two things worth knowing before promising them.

Continuity of action across a cut. Two shots of the same action from different angles, matching. This is basic film grammar and there is no reliable route to it in generation. The workaround is to generate one wide take and crop into it for the closer shots, which is how a lot of low-budget live action is cut anyway.

Precise timing. Making something happen on a specific frame — a hit on a beat, a door closing on a word — is not controllable. You get the action somewhere in the clip and you place the cut in the edit. Budget for it.

There is also a cheap substitute for a camera move that nobody should be embarrassed to use. Generate a wider, static shot at higher resolution, then create the move in the edit by animating a crop across it. A push in, a slow pan, a reframe — all of these are a keyframed crop, they are exact, they are repeatable, and they introduce no temporal artefacts at all because nothing in the picture is changing. Documentary editors have made a living from this technique on still photographs for decades. Against a generated camera move that lands one time in five, it is frequently the better shot as well as the cheaper one.

The honest summary: camera control has improved from "unusable" to "usable within film vocabulary", and precise directing is still done by conditioning on real footage or a 3D render rather than by asking.

BEFORE YOU MOVE ON

Which prompt is most likely to produce the intended shot on a current video model?

A. A cinematic sequence where the camera glides elegantly through the doorway and sweeps upward to reveal the whole room in a single dramatic movement.

B. Dynamic camera work, high energy, lots of movement, epic feel.

C. The camera does a slow dolly in. The woman stands still, facing the window.

D. Camera moves left to right across the scene at a steady pace, while at the same time the subject in the foreground moves in the opposite direction toward the door.

43 · 8 min

Why clips get worse as they get longer

THE ONE THING TO KEEP

Extending a clip by conditioning each chunk on the previous one accumulates error, because each chunk's small imperfections become the ground truth for the next.

The compounding problem

The obvious way to make a long video from a short-window model is to generate a chunk, take its last frames, and use them to condition the next chunk. Repeat.

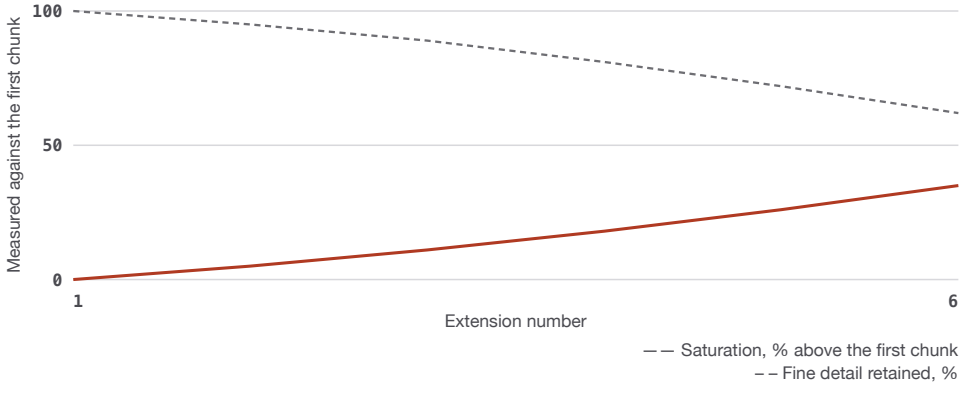
It works, and it degrades in a predictable way. Each generation introduces small errors — a slight colour shift, softening, a face marginally off. Those errors are then treated as the truth for the next chunk, which adds its own on top. After five or six extensions the result has drifted somewhere the first chunk would not recognise.

The characteristic progression, which you can watch happen:

1. Contrast and saturation creep upward, because each pass pulls slightly toward the model's mode.
2. Detail softens, because each pass through the autoencoder loses a little.
3. Faces regularise toward the average face.
4. The scene reorganises, because nothing outside the current window constrains it.

This is the same mechanism as repeated image-to-image from the control module, running automatically, several times, inside one apparently simple feature.

What six chained extensions do to a clip



Each chunk is conditioned on the last chunk's errors, so the errors become the ground truth. Correcting the anchor frame by hand at extension two means regenerating one chunk; noticing at six means regenerating four. Put the first and last frames side by side after every extension.

Why long-video demonstrations are usually cut

When you see a minutes-long generated video, look at the cutting. Almost always it is many short generations edited together, with cuts placed exactly where continuity would have broken. That is legitimate filmmaking and it is worth naming, because it is frequently presented as a single capability.

The exception is the current class of causal streaming models, which generate frame by frame and can run indefinitely. They face the same drift and manage it with explicit mechanisms — periodically re-anchoring on a reference frame,

or keeping a compressed memory of earlier content. These reduce drift; nothing yet removes it.

Working with it rather than against it

Re-anchor deliberately. Rather than chaining chunk to chunk, condition every chunk on the *original* reference image as well as the previous frames. Drift is then measured against a fixed point rather than against the last error.

Regenerate the anchor by hand. If chunk three's last frame has drifted, fix that frame as an image — inpaint the face, correct the colour — and use the corrected version to condition chunk four. Ten minutes of image work resets the accumulation.

Grade at the end, not per chunk. Colour-correcting each chunk to look good individually makes the drift harder to see and no smaller. Assemble everything, then grade the whole sequence to one reference. Free tools do this well; Resolve's free edition is the obvious one.

Cut on motion. A cut during movement hides a mismatch that would be obvious across a held frame. This is standard editing practice and it is doubly useful here.

The arithmetic worth doing before you start

Before committing to a piece, work out how many independent shots it needs and how long each is. A two-minute explainer is not two minutes of generation; it is perhaps thirty shots of three to six seconds, most of which will need several attempts.

That number tells you whether the project is feasible on your budget and your time, and it is far better to know at the start. It also usually reveals that half the shots do not need generation at all — a title card, a screen recording, a still with a slow move on it, a piece of stock footage. Mixing sources is not a compromise. It is how the medium is actually produced.

A drift check that takes ten seconds

Put the first frame and the last frame of your extended clip side by side and look at them as a pair, not as the ends of a sequence. Watching the video hides drift, because your eye follows the change; seeing the two frames together makes it obvious.

Do this at every extension rather than at the end. Drift is far cheaper to correct at chunk two than at chunk six, because correcting it at chunk two means regenerating one chunk and correcting it at chunk six means regenerating four.

The same trick works for a sequence of separate shots that are supposed to match. Lay the first frame of every shot in a row in any image editor. Mismatched grade, drifting colour and inconsistent contrast are visible instantly in that strip and invisible when the shots play in order.

The honest limitation: there is no setting that makes a long generated take hold together. Length is bought with cuts, anchors and manual correction. Anyone promising a single continuous generated scene of any duration is either using a very forgiving subject or has not looked closely at the result.

BEFORE YOU MOVE ON

A five-chunk extended clip ends up brighter, glossier and softer than it began. What is the mechanism?

A. The upscaler applied to each individual chunk accumulated sharpening and contrast artefacts that compounded across the length of the finished sequence.

B. The video autoencoder compresses time by a factor of four, so later chunks contain fewer distinct frames.

C. Each chunk was conditioned on the previous chunk's output, so each pass's small errors became the next pass's ground truth.

D. The attention window shortened as the total clip length grew, leaving later chunks with less context.

44 · 9 min

Talking heads, lip sync and the consent line

THE ONE THING TO KEEP

A talking-head system drives face and mouth motion from an audio track, which makes convincing video of a person saying words they never said cheap, and the only real control on it is consent recorded before the fact.

How the mechanism works

There are two broad approaches and both are now cheap enough to run on ordinary hardware.

Audio-driven reenactment. A model takes a still photograph or short clip of a face and an audio track. It predicts, per frame, the mouth shape and head motion that would produce that audio, and renders the face accordingly. Underneath, the audio is converted to a sequence of features that correlate with mouth position, and the face model is conditioned on them frame by frame. The rest of the image is held.

Full generation with audio conditioning. A video model generates the whole shot with the audio as a conditioning input, producing head, body and background together.

The first is more stable and more obviously an edit. The second is more convincing and drifts more.

The quality determinant most people miss is the **audio**, not the face. Clean, close-miked speech produces accurate mouth shapes; noisy or compressed audio produces mush, because the features the model reads are degraded before it starts.

What it is genuinely good for

This is not a purely dangerous technology and it is worth being specific about the useful cases, because a blanket warning teaches nothing.

- **Dubbing with matched lips.** A film translated into another language, with the mouths adjusted to the new dialogue. This is a real improvement in accessibility and the licensing arrangements for it are ordinary.
- **Presenter video at scale** for training material, where a real presenter has consented to a synthetic version of themselves.
- **Restoring a damaged performance** in post-production — a line re-recorded and re-synced rather than reshot.
- **Accessibility**, including giving a synthetic presence to someone who cannot appear on camera.

Every one of these has the same structure: the person depicted agreed, specifically, in advance.

The consent standard

The voice module in this course sets out the standard in full and it applies identically to faces. In short, consent must be:

- **Informed** — the person knows what will be produced and in what contexts.
- **Specific** — to this use, not to any future use anyone thinks of.
- **Written** — because the dispute will be about what was agreed.
- **Revocable** — with a stated process and a stated timescale.
- **Compensated**, where the use is commercial.

And separately from consent, the audience needs to know. Someone who consented to a synthetic presenter has not thereby consented on behalf of the viewer to be deceived. Disclosure is a second duty, and in several jurisdictions it is now a legal one, which the final module sets out.

The part that is not solvable by good practice

Every technique described here works from a photograph and an audio sample. Both are publicly available for most people who have ever appeared online. That means the capability to produce convincing video of an arbitrary person is widely distributed and cannot be recalled.

The harms are documented and heavily skewed. The overwhelming majority of synthetic video circulating is non-consensual sexual imagery, and the overwhelming majority of its targets are women. Fraud is the second category — a video call from a colleague's face authorising a payment. Political fabrication is real and, by volume, third.

The realistic responses are institutional rather than technical:

- **Verification out of band.** A second channel for any instruction that moves money or grants access. This defeats the fraud case entirely and costs nothing to adopt.
- **Provenance**, covered in a later module — signed capture rather than detection after the fact.
- **Law.** Several jurisdictions have created specific offences for non-consensual synthetic sexual imagery and for fraudulent impersonation; the coverage is uneven and expanding.

What does not work is detection by eye. Assume that any video of a person can be fabricated and that the question worth asking is where the file came from, not whether the mouth looks right.

One practical note for anyone commissioning this work. A performer's agreement to appear on camera is not an agreement to a synthetic version of themselves, and a standard release form written before any of this existed does not cover it. Several performers' unions negotiated specific terms on exactly this point, and the shape they arrived at is a reasonable template even outside a union context: separate consent for the creation of a digital replica, separate consent for each use, a stated term after which it expires, and payment for use rather than only for the capture session. If you are the performer, ask for those four things. If you are commissioning, offering them unprompted is cheap and it is the difference between a working relationship and a dispute.

For your own work the line is simple enough to state in a sentence: do not make a recognisable person say or do anything without their specific written agreement, and tell the audience when what they are watching is synthetic. Both parts are required. Neither is difficult.

BEFORE YOU MOVE ON

An organisation wants to use a synthetic version of a consenting presenter for internal training videos. What else is required beyond their consent?

A. A visible watermark applied to every frame of the video file, on the grounds that the presenter's consent alone does not satisfy the provenance requirements that apply to synthetic media.

B. A separate agreement with the audio model provider covering the presenter's voice as a distinct work.

C. Nothing further, provided the consent is written, specific and revocable.

D. Disclosure to the viewers, because the presenter's consent does not cover the audience's interest in knowing.

---

45 · 8 min

Testing a claim that it understands physics

THE ONE THING TO KEEP

Claims that a video model has learned physics should be tested with occlusion, conservation and causal-order cases, because local plausibility over a few seconds is achievable without any physical model at all.

The claim, and why it is attractive

Video models are increasingly described as world models or world simulators — systems that have learned how things behave, not merely how they look.

The claim matters commercially, because a model that understands physical consequence is a different product from a model that generates pretty footage.

It is also, in a weak form, plausible. To predict the next frames of a falling object you must produce acceleration that looks right. Something about the dynamics has been captured.

The question is what "captured" means, and it is answerable by test rather than by argument.

Four tests that separate the claims

Occlusion and permanence. Show a distinctive object passing behind a larger one and re-emerging. A model with an internal object representation returns the same object. A model doing local plausibility returns *an* object of roughly that kind, or a different one, or none. This is the single most informative test and it is easy to run.

Conservation. Pour liquid from one vessel into another and watch the volumes. Cut an object and count the pieces. Break a plate and see whether the fragments could reassemble. Quantity is a global constraint, and the counting lesson tells you what to expect.

Causal order. Ask for the consequence of an event rather than the event: a glass that has already fallen, a candle after it was blown out, a footprint after the walker has passed. Models trained on captioned clips have seen far more of the event than of its aftermath.

Counterfactuals and unusual physics. A ball bouncing on a surface at an unusual angle, an object with the wrong weight, a sport played with the wrong equipment. If the model has learned dynamics it can extrapolate; if it has learned the appearance of familiar clips, it snaps back to the familiar version.

Run these on any model whose marketing makes the claim. It takes twenty minutes and it is more informative than any published benchmark score, because these are exactly the cases benchmarks under-represent.

What the tests currently show

Honestly reported: current strong models do well on short, simple, common dynamics — falling, splashing, cloth, smoke, walking. They fail with regularity on occlusion and permanence, on conservation of quantity, on causal aftermath, and on anything counterfactual.

That pattern is consistent with a model that has learned a very good statistical prior over how footage of the world looks over short spans, and inconsistent with one that maintains a representation of objects and their properties.

This is not a criticism of the models. A very good visual prior is enormously useful and is what most production work needs. It is a criticism of the word "understands", which does specific work in a sales conversation.

Why it matters beyond argument

Two practical reasons to care.

It predicts your failures. If the model has no object representation, you should expect exactly the problems this module has described, and you should plan shots that avoid them rather than hoping the next release fixes them. It has not fixed them for three releases.

It matters for downstream claims. World-model language is being used to argue that these systems can serve as simulators for robotics, driving and planning — domains where being wrong about occlusion has consequences. The claim may eventually be true. The tests above are the honest way to find out, and they are the ones a buyer should ask for.

> *A model that predicts what footage looks like is not the same as a model that predicts what happens. The two agree most of the time, which is exactly why the disagreements are worth finding on purpose.*

There is a related test worth adding, because it separates memorised footage from learned dynamics more sharply than the physics cases. Ask for something that is physically ordinary and visually rare: a cricket ball rolling across a snooker table, a person walking backwards up stairs, a candle burning in a strong wind. Each is easy to imagine and unlikely to be well represented in

any video corpus. A model with dynamics extrapolates; a model with a strong appearance prior returns the familiar version — the ball on a cricket pitch, the person walking forwards — or produces motion that reads as reversed footage rather than as the action requested.

The unsettled part, stated plainly: there is a genuine research disagreement about whether enough video training induces a physical model as a by-product, or whether an explicit structure is required. Serious people hold both positions. What is not in dispute is that current systems fail the tests above, and anyone claiming otherwise should be asked to run them.

BEFORE YOU MOVE ON

Which test most directly distinguishes a model with an internal object representation from one producing locally plausible footage?

A. Generating a long clip and checking whether the style stays consistent from beginning to end.

B. Passing a distinctive object behind an obstacle and checking whether the same object re-emerges.

C. Generating fast motion and checking whether it smears, since smearing indicates the absence of a physical model.

D. Comparing the model's output against real footage of the same scene using an automated similarity score across every frame of both sequences.

46 · 8 min

Testing a video model honestly

THE ONE THING TO KEEP

A demonstration reel is selected output, so judging a video model requires a fixed prompt battery, a stated number of attempts per prompt, and a usable-shot rate rather than an impression.

Why reels tell you nothing

Every video model is launched with a reel. The reel is real output and it is also the survivors of an unknown number of generations, chosen by people whose job is to choose well, on prompts selected because they worked.

Nothing about it is dishonest and nothing about it is informative. The number you need is the one the reel cannot contain: how many attempts produced how many usable shots, on prompts you did not choose.

A battery you can run in an hour

Fix a set of prompts before you test anything, and reuse it for every model. Ten is enough. A serviceable set:

1. A person walking toward the camera, medium shot.
2. Two people talking, one gesturing.
3. Hands performing a task — tying a lace, pouring tea.
4. A slow dolly in on a static object.
5. An animal moving quickly.
6. Water or fire.
7. A vehicle passing through frame.
8. Text on a sign, held in shot.
9. An object passing behind another and re-emerging.

10. A specific action with a first and last frame supplied.

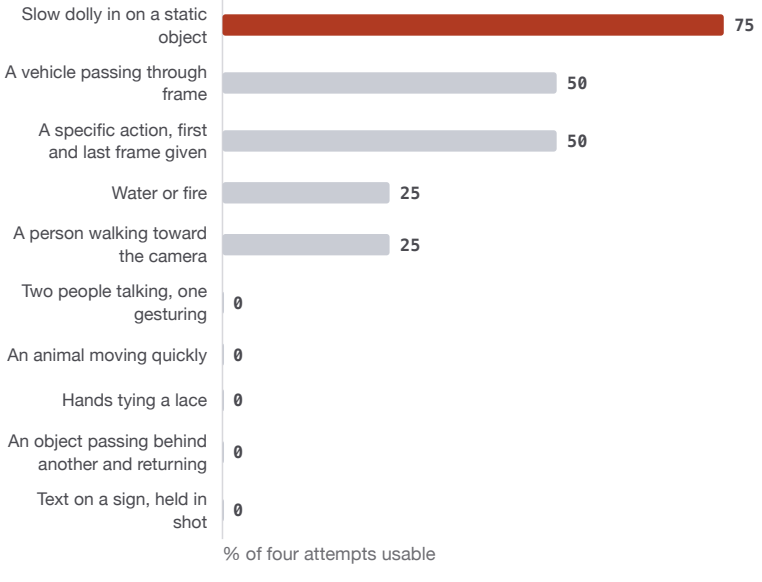
Generate four attempts for each on every model. Forty clips is an hour and a modest cost, and it gives you something no review will.

Scoring without deceiving yourself

Decide what usable means *before* you look, and write it down. A workable definition: a shot you would place in a paying client's edit without an apology. Then score each clip as usable or not, with no middle category — the middle category is where self-deception lives.

Report the result as a rate: "6 of 40 usable, 15%". That single number is comparable across models, across months, and across the claims anyone makes to you.

Forty clips, one battery: where a model actually fails



Nine of forty usable, twenty-three per cent. The average is the least useful part of it. The four prompts that never produced anything are the model's shape, and they are what you plan shots around.

Record two more things while you are there:

- **Which prompts never produced anything usable.** This is the model's shape, and it is more useful than the average.

- **Time and cost per usable shot**, not per generation. A model with a 30% hit rate at twice the price is cheaper than one at 10%.

The failure taxonomy worth keeping

When a clip fails, record why, using the categories from this module: temporal drift, morphing, anatomy, physics, camera did not follow, subject did not follow, texture flicker, resolution. After forty clips you have a profile rather than an impression, and profiles are what let you choose the right model per shot instead of adopting one for everything.

This is exactly the method the evaluation course in this catalogue teaches in general form: a fixed set, a stated definition of success, a rate rather than an anecdote. It applies here because video is the domain where impressions are least reliable — the outputs are seductive, individually memorable, and heavily selected before you ever see one.

What to re-test, and when

Models change. Hosted ones change without announcement. Re-run the battery when a version number changes, when your hit rate drops noticeably, and before any commitment that depends on the result.

Keep the clips. A folder of the same forty prompts across four models and two years is a genuinely valuable thing to own, it costs nothing but disk, and it is the only way to answer "is this actually better" rather than "does this feel better", which it always does at launch.

Judging without knowing which model made it

One refinement is worth the small extra effort, and it comes straight from ordinary experimental practice. Rename the clips so you cannot tell which model produced which, shuffle them, and score them in that state. Then reveal the labels.

People who have done this are usually surprised. Expectation does a great deal of work in judging generative output — a clip from the model you have just read about looks better than the identical clip from the one you have not

— and a blind pass removes it for nothing but a minute of file renaming. If you are choosing a tool your team will pay for, this is the difference between a decision and a preference.

The honest limitation of the method: a ten-prompt battery covers what you thought to include. It will miss the failure that ruins your particular project. So add prompts drawn from your real work as you meet them, and treat the battery as a growing regression suite rather than a fixed exam. Each shot that failed on a real job becomes prompt eleven, and the set gets more useful with every disappointment.

BEFORE YOU MOVE ON

A vendor's reel is flawless and your own testing gives a 12% usable rate. What is the most likely reconciliation?

A. The vendor's model is served at a higher quality tier than the public endpoint, which is why identical prompts give different results.

B. Your prompts are less well written, so the difference reflects prompting skill rather than the model.

C. The reel shows selected survivors of many generations on prompts chosen because they worked.

D. The reel was produced with first-and-last-frame conditioning while your test used text prompts, so the comparison is between two different capabilities of the same system.

47 · 8 min

What synthetic video does to believing anything

THE ONE THING TO KEEP

The main effect of convincing synthetic video is not that fakes are believed but that real recordings can be denied, so the useful response is provenance and verification rather than detection.

Two harms, and the larger one is not the obvious one

The obvious harm is fabrication: a video of someone doing something they did not do. It is real, it is documented, and its distribution is heavily skewed toward non-consensual sexual imagery of women and toward financial fraud.

The less obvious harm is the mirror of it. Once everyone knows video can be fabricated, any genuine recording can be dismissed as a fake. This is sometimes called the liar's dividend, and it does not require anybody to actually make a fake. It requires only that fakes are known to be possible.

This is a corrosion of a specific kind. Photographic and video evidence has, for about a century, functioned as a shared reference point that survived disagreement about everything else. What replaces it is not nothing, and it is slower, more institutional and less available to individuals: chains of custody, corroborating sources, and cryptographic provenance.

Why detection will not save it

The provenance module goes into the mechanisms. The short reason detection is the wrong hope:

- Detectors are trained on the outputs of known models and degrade sharply on models they have not seen, which includes every model released after the detector.

- Any lossy step — a re-encode, a screenshot, a platform upload — removes much of the statistical signal they rely on.
- The base rate problem is severe: apply a 95%-accurate detector to a population where 1 in 1000 videos is fabricated and most of your positives are wrong.
- And a detector that is public is a target that adversaries can optimise against.

The consequence: a detector's output is a weak signal, not a verdict, and using one to make a public accusation is how innocent people get harmed.

What actually works

Provenance at capture. Sign the recording when it is made, in the device, and carry the signature with the file. This proves origin rather than detecting fakery, and it is the only approach that improves as the generators improve. Its problem is adoption, not principle.

Corroboration. Multiple independent recordings of the same event, from angles that agree. This is how newsrooms have verified footage for years and it does not care how good the generators get.

Source over content. Who published this first, do they have a name, does the account have a history, does a reverse search find it earlier elsewhere. The free tools do this in seconds.

Out-of-band verification for anything consequential. Any instruction arriving by video or voice that moves money, grants access, or changes a payment detail gets confirmed on a different channel. Organisations that adopted this rule stopped losing money to this class of fraud immediately. It is the cheapest control in the whole field.

What to do as somebody who makes things

Two responsibilities, both small and both worth taking seriously.

Label your synthetic work. Not because you are ashamed of it, but because an unlabelled synthetic video that later circulates as real is a harm you

contributed to. The final module covers what the law requires in several countries; the practice runs ahead of the law and should.

Do not build the demonstration. There is a persistent genre of "look how easy this is" content — a fabricated clip of a public figure, made to raise awareness. It raises awareness and it also adds a fabricated clip to the world, permanently, with your caption stripped by the second share. Show the workflow on yourself or a consenting colleague. The point lands identically.

There is one asymmetry worth naming, because it decides who bears the cost. Fabricating a convincing clip now takes minutes and a few pounds. Refuting one takes a newsroom, a forensics team and several days, and the refutation reaches a fraction of the audience the original did. That imbalance is not fixed by better tools on the defending side, because the effort ratio is the problem rather than the accuracy. It is the strongest practical argument for provenance at capture, which shifts the work to the moment of recording, where it is cheap, and away from the moment of dispute, where it is not.

The unsettled part, stated as such: nobody knows where this settles. Some argue that societies adapted to photographic manipulation before and will adapt again, with institutions and habits rather than technology doing the work. Others argue that the speed, volume and personalisation are different in kind. Both positions are held by serious people, the evidence is a few years old, and anyone confident in either direction is guessing. What can be said now is what to do about it, and the list above is not controversial.

BEFORE YOU MOVE ON

Why is a detector's positive result a weak basis for a public accusation?

A. Detectors are trained on limited data, so their accuracy figures are estimates with wide error bars around the reported value.

B. Detectors report a probability rather than a category, so the result cannot be interpreted without a threshold.

C. Detectors work only on uncompressed files, and any published video has been compressed.

D. Detectors degrade on models they were not trained on, and at realistic base rates most of their positives are false.

48 · 8 min

Generated objects and the space around them

THE ONE THING TO KEEP

Generated 3D comes in two incompatible forms — meshes you can edit and animate, and radiance representations you can only view — and confusing them wastes a great deal of time.

Two different things, both called 3D

Meshes. Vertices, edges and faces, with a texture map. This is what every game engine, animation package and 3D printer consumes. It can be rigged, animated, cut, measured and manufactured.

Radiance representations — neural radiance fields and, more usefully now, **Gaussian splatting**. These store a scene as a cloud of coloured, semi-transparent blobs optimised to reproduce a set of photographs from any viewpoint. They render beautifully and photorealistically, in real time, and they are not geometry. There is no surface to attach anything to, no clean silhouette, nothing to 3D print.

The single most common disappointment in this area is someone capturing a splat of an object, finding it looks perfect, and discovering it cannot be imported into anything that expects a mesh. Conversion exists and loses much of what made the splat good.

Two different things, both called 3D

Mesh — vertices, edges, faces	Gaussian splat — a cloud of coloured blobs
• Opens in any game engine or animation package	• Renders photorealistically, in real time
• Can be rigged and animated	• Reproduces a captured scene from any viewpoint
• Can be measured and manufactured	• Has no surface to attach anything to
• Arrives as dense irregular triangles	• Cannot be 3D printed
• Needs retopology before anything moves	• Converting it to a mesh loses what made it good

Choose by what happens next. Viewing, walkthroughs and virtual tours: splats. Anything to be edited, animated, simulated or manufactured: meshes. The most common disappointment in this area is discovering the difference after the capture.

Choose by what happens next. Viewing, walkthroughs, product presentation, virtual tours: splats. Anything to be edited, animated, simulated or manufactured: meshes.

How a mesh gets generated

Three approaches are in use.

Multi-view then reconstruct. Generate several consistent views of the object with an image model, then reconstruct geometry from them. Cheap and it inherits the consistency problems of this whole module — the back of the object frequently disagrees with the front.

Native 3D generation. A diffusion model trained directly on 3D representations. Better geometry, limited by the scarcity of 3D training data, which is orders of magnitude smaller than image data. This scarcity is the field's binding constraint and it is why 3D generation lags image generation by several years.

Photogrammetry from real photographs. Not generative at all: dozens of overlapping photographs reconstructed into geometry. Free software does

this well. For a real object you can photograph, this remains the highest-quality route and it is often forgotten in the enthusiasm for generation.

What generated meshes are actually like

Honest description of the current state, because the promotional images do not show the wireframe:

- **Topology is poor.** Generated meshes are dense, irregular triangle soup rather than clean quad flow. They render acceptably and they deform badly, which means anything intended to be animated needs retopology.
- **Scale and orientation are arbitrary.** Everything arrives needing to be scaled and rotated to fit a scene.
- **Textures are baked and low resolution**, frequently with the lighting of the generated views burned into the colour map, which makes them look wrong under different lighting.
- **Hollow interiors and non-manifold geometry** are common, which matters enormously for 3D printing and not at all for a background prop.

The realistic use today is as **greyboxing and background assets**: blocking out a scene, filling a shelf, prototyping a shape before modelling it properly. As a hero asset for animation, it is a starting point that a modeller cleans up.

Where it fits with everything else in this course

There is a workflow worth knowing that runs the other way, and it is the most practical use of 3D for most people here.

Build a rough scene in Blender — free, on any platform — with grey blocks in the right positions and the camera where you want it. Render a depth map. Use that as structural conditioning for an image or video generation. You now have exact composition, exact camera, exact perspective and repeatable framing, supplied by software designed for precisely that, while the generative model does the surfaces.

This is the answer to nearly every spatial problem in this course. It requires learning a small amount of Blender — moving objects, placing a camera, ren-

dering — which is a weekend rather than a career, and it converts an unreliable prompt into a reliable pipeline.

A note on the rights position, which differs from images in a way people miss. A generated mesh of a recognisable product, vehicle or building carries the same trademark and design-right questions as a photograph of it would, and in some cases more, because a 3D model can be manufactured. Printing a generated model of a branded object is not the same act as generating a picture of it, and the relevant law is design right and trademark rather than copyright in an image. The final module covers the shape of these questions; the practical note here is that "the model generated it" changes nothing about the object you then make from it.

The honest limitation: 3D generation is the least mature part of this field, the training-data scarcity is structural rather than temporary, and the gap between a viewable result and a usable asset is wider than in any other medium covered here. Expect it to improve and do not plan a production pipeline around it yet.

BEFORE YOU MOVE ON

A team captures a product as a Gaussian splat, then finds they cannot use it in their animation. What did they get wrong?

A. The capture resolution was too low to reconstruct usable geometry from the source photographs.

B. Splats store a scene as view-dependent blobs rather than surfaces, so there is no mesh to rig or animate.

C. Splats can only be viewed in a browser, so the file cannot be opened by desktop software at all.

D. The splat was captured without a turntable, so the object's underside is missing and the model is incomplete.

CASE STUDY · MODULE 5

Thirty shots, one launch reel, and a bid due on Friday

A three-person video house in Ahmedabad, deciding whether to bid a co-operative bank film as generated work

Brightbox is Nilesh Trivedi, an editor and a junior. They make corporate films for banks, hospitals and two engineering firms, mostly two to four minutes, mostly shot.

A district co-operative bank invited them to bid on a two-minute film explaining a new micro-loan product, to run on branch televisions and to be forwarded on WhatsApp. Budget two lakh forty thousand rupees. Done the usual way that is a two-day shoot, a small crew, two hired actors and a location fee, and about a lakh eighty of the budget is committed before anybody opens an edit timeline.

Riya, the junior, had watched a video model's launch reel the week before and proposed generating the whole film. Twenty thousand rupees of credits, no crew, no location, no actors. The margin was the entire argument and it was a good one.

Nilesh's objection was not that it would not work. It was that nobody in the room knew the hit rate, and a launch reel cannot contain that number. Every clip in a reel is real output. It is also the survivor of an unknown number of attempts, chosen by people whose job is to choose well, on prompts selected because they produced something.

It was Tuesday. The bid was due Friday.

Option one was to bid on the strength of the reel. Two lakh forty, a healthy margin, and find out during production whether the shots the bank actually needs are shots this model can make.

Option two was to spend Wednesday and about nine thousand rupees of credits measuring it, and bid on Thursday with a number. The cost of that is specific: two of the three days before the bid go into a test rather than into the

treatment and the storyboard, and a competing production house on the second floor of the same building was bidding on the same job.

The shot list decided it, or rather the shot list made the question answerable. The bank's film needed a farmer signing a form, which is hands on paper. Two people talking across a counter. A branch board with the bank's name legible. A scooter arriving at a gate. And the same woman appearing in four separate shots across the film, because she is the borrower the story follows from application to first instalment.

Every one of those is a known weak spot, and Nilesh knew that in general. Hands at the point of contact with an object are where the model's learned appearance of a gesture stops matching what a gesture is. Legible text in video is roughly where image models were in 2022. A character who must be recognisably the same person in shot one and shot nine has no mechanism keeping her the same, because nothing in the pipeline stores an identity. He could have recited all of that on Tuesday afternoon.

What he could not recite was the rate, and the rate is the only thing a bid needs. Knowing that hands are hard tells you to expect trouble. Knowing that hands come back usable one time in twelve tells you whether to put them in the quote, and knowing that they come back usable zero times in twelve tells you the shot does not exist at any budget.

There was also a number underneath the whole argument that Riya had not included in her twenty thousand rupees. List prices are quoted per generated second, and what a production actually spends is the cost per second it can use. At a one-in-eight hit rate — which is ordinary once a shot has a specific requirement attached to it rather than being pretty — an eight-second clip listed at two hundred rupees has really cost sixteen hundred, plus the hour of somebody's attention spent watching seven failures. Teams that budget for the keeper run out of money somewhere around shot four of thirty.

They ran the test.

STOP HERE

Before turning the page: what would you have done, and what would it have cost if you were wrong? Write it down. The point of the next section is to compare.

WHAT HAPPENED

They fixed ten prompts before generating anything: a person walking toward camera, two people talking with one gesturing, hands performing a task, a slow dolly on a static object, an animal moving quickly, water, a vehicle crossing frame, legible text on a sign held in shot, an object passing behind another and re-emerging, and a specific action with a first and last frame supplied. Four attempts each. Forty clips, a little over nine thousand rupees, about five hours including the waiting.

Before they looked at anything, Riya wrote the scoring rule on a sheet of paper: usable means a shot I would put in a paying client's edit without an apology. Usable or not usable. No middle category, because the middle category is where people talk themselves into a clip.

Six of forty.

The average was not the useful part. The distribution was. Slow dolly on a static object: four of four. Water: three of four. Vehicle crossing frame: two of four. Hands on paper: zero of four, every attempt. Legible signage: zero of four. The same person across separate shots: the face changed every time, so the count was zero before they discussed it.

That is three of the five things the film needed, at a rate of nothing, and no amount of budget for retries changes a zero.

They bid two lakh twenty for a hybrid, and the treatment named which shots were which. A half-day shoot at a real branch with two bank staff who volunteered covered the seven shots involving hands, faces and signage, which cost about forty thousand including the day. Twelve atmosphere shots were generated: a field at dawn, a market, traffic, a shutter going up, none of them containing a face anyone had to recognise or a word anyone had to read. Six more

shots were stills with a keyframed crop moving across them, which cost nothing, never drift and are exact.

They won. The bank's marketing officer said afterwards that the competing bid was three pages of adjectives and theirs was a shot list with a method against each line.

The forty clips are still on a disk. When the model released a new version in November, Riya ran the same ten prompts again in an afternoon and the number moved from six to eleven, which is the only way either of them would believe it had improved.

Every clip in the launch reel was genuine output from the model. Why is the reel still uninformative for a bid?

Because it is selected on two axes the viewer cannot see: how many attempts produced each clip, and which prompts were tried at all. A model with a one-in-twenty hit rate and a model with a one-in-two hit rate produce identical reels, and a reel will not contain the prompt that never worked. The number a bid needs is a usable-shot rate on prompts chosen by the buyer, and no reel can carry it. That is not dishonesty on the vendor's part; it is a structural property of showing chosen output.

Riya wrote the definition of 'usable' down before generating anything, and allowed no middle category. Why do both of those matter?

Writing it first stops the standard moving to fit the results, which it does silently once effort and money have been spent. Refusing a middle category forces each clip to be counted as something, and the middle category is where a clip that would need an apology in a client edit gets recorded as nearly fine. Together they turn an impression into a rate that is comparable across models, across months and against anyone's claim. It is the same discipline as fixing a test set before looking at it.

Hands on paper and legible signage scored zero of four. Should they have spent a day writing better prompts for those shots?

No. Both are global, exact constraints — a hand has a correct number of fingers in a correct arrangement, and a sign has a correct sequence of letters — and the model is strong on local plausibility and weak on constraints that can only be checked

by taking in the whole frame. Prompt wording acts on what concepts are present, not on whether a global constraint holds. A rate of zero over four attempts is a signal to change the method rather than the words: shoot those shots, or supply the structure another way. They shot them, in half a day, for about forty thousand rupees.

MODULE 5 · TEST

Check what stuck

6 questions, one right answer each. This one is for you, not for a record: answers and the reason behind each are at the back. If two questions from the same module go wrong, re-read that module before the exam rather than after.

1 A five-second clip at 24 frames per second is 120 frames, and it costs far more than 120 images. Where does the extra cost come from?

A. Each frame is generated at a higher resolution and then downscaled for playback.

B. The model must attend across frames as well as within them, and attention grows quadratically.

C. Video models run many more denoising steps per frame to suppress flicker.

D. The video autoencoder is larger than an image autoencoder and is run once per frame.

2 Video models offer frame counts like 49, 81 and 121 rather than round numbers. What sets those figures?

A. They are the counts that divide evenly into common frame rates.

B. They are limits imposed by the maximum tensor size the attention layers accept.

C. They follow the autoencoder's temporal compression factor, plus one.

D. They are chosen so the clip length is a whole number of seconds at 24 frames per second.

3 A character's jacket keeps its buttons for four seconds and changes at five. Which explanation fits?

A. The prompt's influence decays over the length of the clip.

B. Colour drift accumulates and crosses a threshold at around five seconds.

C. Beyond the model's attention window, nothing enforces consistency.

D. The temporal layers are trained only on four-second clips and extrapolate poorly.

4 You supply a first and a last frame and the model produces a morph in which the objects melt into one another rather than a movement. What does that tell you?

A. The two frames cannot be connected by a plausible continuous motion.
B. The motion-strength setting is too low for the model to move anything.
C. The two frames were encoded at different resolutions.
D. The model does not support last-frame conditioning and ignored the second image.

5 You build a thirty-second clip by generating six chunks, each conditioned on the last frames of the previous one. By chunk five the contrast has crept up, faces have regularised and the scene has reorganised. What is the best correction?

A. Raise the resolution of the later chunks so detail is not lost.
B. Colour-correct each chunk as it is made so the drift is not visible.
C. Generate all six chunks from the same text prompt with no image conditioning.
D. Condition every chunk on the original reference image as well as the previous frames.

6 A client wants a generated 3D asset of a product that will be animated and eventually 3D printed. Why is a Gaussian splat capture the wrong answer however good it looks?

A. Splats are stored at a fixed viewpoint and cannot be rotated freely.
B. Splats have no surface geometry, so there is nothing to rig, animate or manufacture.
C. Splats cannot represent colour accurately enough for product work.
D. Splats require far more memory to render than a mesh of the same object.

MODULE 6

Voice and speech

Speech is the medium where synthesis became convincing first and where the consequences are sharpest. This block covers how a synthetic voice is actually produced, what a voice print is and why a few seconds of audio is enough, why synthetic speech still sounds flat and what fixes it, how uneven the coverage of languages and accents is, the recognition half of the pipeline and where it fails, what breaks in a dubbing chain, why detecting a synthetic voice is harder than people assume, and the fraud pattern that every organisation should already have a control for.

> BY THE END OF THIS MODULE YOU CAN
>
> Describe the path from text to a waveform through tokens, an acoustic model and a vocoder, judge what a voice-cloning claim requires in consent and disclosure, and specify a verification control that defeats voice-impersonation fraud without relying on detection

49 · 9 min

Voice, and the line you do not cross

THE ONE THING TO KEEP

Clone a voice only with the owner's informed, written, revocable consent — and tell the audience before they believe it.

Cloning is fast now

Modern voice models do not retrain on your voice. They extract a speaker embedding — a compact numerical description of timbre and delivery — from a short reference clip, then condition a speech model on it. Three to thirty seconds is usually enough.

Quality of reference beats quantity. One clean, dry, mono recording with no music, no reverb and no second speaker will beat ten minutes of a noisy phone call. Match the delivery too: a clone built from calm reading sounds wrong shouting.

What this is genuinely good for

Audiobook narration in an author's own voice. Dubbing a course into eight languages while keeping the teacher's voice. Voice banking for people losing speech to ALS or throat cancer, which on its own justifies the field. Fixing one mispronounced word in an hour of finished narration without recalling the talent.

Every one of those shares a feature: the person whose voice it is agreed.

The line

You may clone a voice when the person it belongs to has agreed, knows what it will be used for, and can withdraw. That is the whole rule.

It is not softened by any of the following, all of which people say with real confidence:

- It is satire.
- He is a public figure.
- It is my father and he would have found it funny.
- I paid for a dataset of their podcast.
- I put a disclaimer in the description.

What ignoring it enables

Family-emergency fraud. A cloned voice of someone's child on a deliberately bad line, asking for money urgently. It works because fear disables judgement before scepticism arrives. The defence that actually works is a family passphrase agreed in advance, offline, and never sent over anything.

Payment fraud, using an executive's voice lifted from a conference recording, aimed at a finance team under time pressure.

Elections. In January 2024, voters in New Hampshire received a robocall in a synthetic voice imitating the US president, telling them not to vote in the primary. The US telecoms regulator then confirmed that AI-generated voices in robocalls fall under existing anti-robocall law, and proposed a multi-million-dollar fine against the operative behind it.

The law is arriving, unevenly

Tennessee's ELVIS Act, in force since 2024, added voice explicitly to the state's right of publicity — a music state protecting its industry.

India has built protection through personality-rights injunctions rather than a statute. The Delhi High Court protected Anil Kapoor's name, image, voice and mannerisms in 2023, and in 2024 the Bombay High Court granted the singer Arijit Singh an order dealing directly with AI voice cloning.

Denmark moved to give individuals a copyright-like right in their own likeness and voice.

The EU AI Act requires anyone deploying a deepfake to disclose it.

A US federal statute on voice and likeness has been introduced more than once and, as far as I know, has not been enacted. Check the current position rather than trusting that sentence.

The direction is unambiguous even where the text is not final. This is becoming a right people hold in themselves.

Consent that actually holds

If you commission a voice, get a written agreement naming:

- whose voice it is, and which recordings the model was built from
- what it may say, as categories rather than vibes
- where it may appear, and for how long
- what it may never do: political speech, endorsements, sexual content, anything defamatory
- how the person withdraws, and what happens to the model files when they do
- what they are paid, including for reuse

Keep the recording of the consent conversation with the model files. And say the sentence out loud once, in the room: we are making a copy of your voice that can say things you never said. If that makes the conversation awkward, the awkwardness is information.

Dead people

Estates can license, and often do. The harder question is not legal. In a 2021 documentary about the chef Anthony Bourdain, the director had a few lines of Bourdain's own written words performed in a synthetic version of his voice, and did not tell the audience. The words were his. The performance was not. Viewers felt deceived, and the fallout has shaped practice since.

The lesson holds generally: disclosure has to reach the audience before the belief does. Not in the credits.

BEFORE YOU MOVE ON

A documentary team clones the voice of a subject who has died, to read letters the subject genuinely wrote. The family has agreed and been paid. A note appears in the closing credits. What is the sharpest criticism?

A. The consent that matters is the family's and it is present; the remaining failure is that the disclosure arrives long after the audience has already heard the voice as real.

B. No disclosure is needed at all, because the words are the subject's own writing and nothing has been put in his mouth.

C. It is unacceptable in any form, because a dead person cannot consent and a family cannot consent on his behalf.

D. It is fine once the estate is paid a licence fee, which makes it an ordinary rights transaction like any archive clearance.

50 · 8 min

From text to a waveform

THE ONE THING TO KEEP

Modern speech synthesis converts text into discrete audio tokens with a language model and turns those tokens back into sound with a neural codec, which is why it now sounds natural and why it hallucinates.

The old pipeline, and why it sounded like that

For decades, text-to-speech had three stages. Convert text to phonemes with pronunciation rules and a dictionary. Predict a spectrogram — a picture of the sound's frequency content over time — from those phonemes. Convert the spectrogram to a waveform with a vocoder.

Each stage was separately engineered, and the joins were audible. Concatenative systems stitched together recorded fragments, which is why an-

nouncements at railway stations have that particular seam between words. Parametric systems produced the flat, buzzy voice everyone associates with a computer.

The quality ceiling came from the middle stage. Predicting a spectrogram from phonemes is an averaging problem: many real deliveries of a sentence are correct, and a model trained to minimise error over all of them produces the average of them, which is the flat one.

The current pipeline

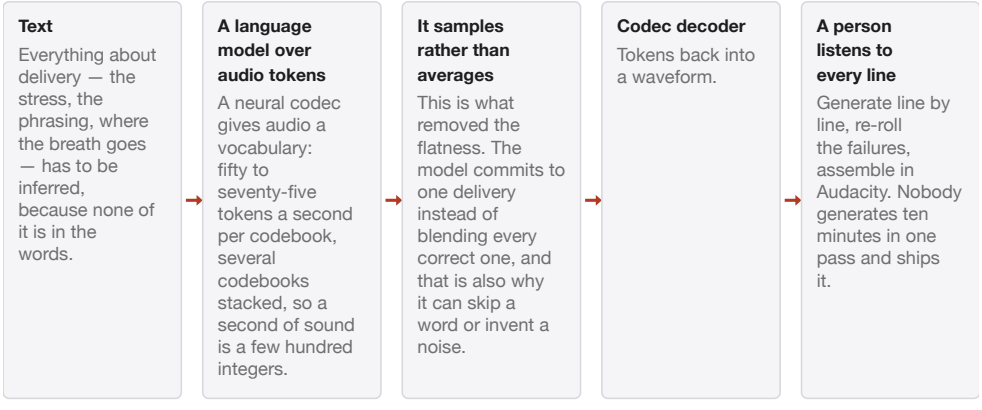
The shift that changed everything was to treat audio as a sequence of discrete tokens, exactly as a language model treats text.

A **neural audio codec** — trained to compress audio and reconstruct it — produces a small vocabulary of tokens, typically at something like 50 to 75 tokens per second per codebook, with several codebooks stacked to capture successive levels of detail. A one-second clip becomes a few hundred integers.

Now the problem is familiar: given text, predict a sequence of audio tokens. A transformer does this, and it does it the way language models do — by sampling from a distribution rather than by averaging. That single change removes the flatness, because the model commits to *one* delivery rather than blending all of them.

The codec's decoder then turns the predicted tokens back into a waveform.

From text to a waveform, as it is done now



The old pipeline predicted a spectrogram from phonemes, which is an averaging problem: many deliveries of a sentence are correct and the average of them is the flat one. Sampling is the whole difference, and it brings the whole failure family with it.

```
text -> (language model) -> audio tokens -> (codec decoder) ->
waveform
```

What follows from the architecture

Three consequences that explain nearly everything users notice.

It hallucinates, because it samples. A generative sequence model can produce tokens that do not correspond to your text. In practice this appears as repeated syllables, a word skipped, an invented noise, or — the most disconcerting failure — the voice trailing off into unrelated babble. This is the same failure family as a language model repeating itself, and it is why any synthetic voice used in production must be listened to, every time, before it ships.

Temperature and sampling settings exist. Where a system exposes them, they behave as they do for text: low values give consistent, slightly duller delivery; high values give expressive readings with more failures. There is a genuine trade here, and for long-form narration the low setting plus a re-roll of the bad lines is usually the right one.

Long inputs degrade. Attention over a long token sequence is the same constraint as everywhere else in this course. Most systems handle a sentence or a paragraph well and drift over several minutes. Generating line by line and

assembling in an editor is the standard practice, and it also gives you the ability to re-roll one bad line rather than a whole chapter.

The free tools, and what they can do

This is an area where open tools are genuinely competitive, which is not true everywhere in this course.

- **Piper** — small, fast, runs on a Raspberry Pi, dozens of languages, good for narration where perfect naturalness is not required.
- Several open codec-based systems produce voices close to commercial quality, run on consumer hardware, and support cloning from a short sample.
- **Audacity** — free — for assembling, trimming and normalising the output, which you will need to do regardless of the generator.

The practical shape of a narration job is: generate line by line, listen to every line, re-roll the failures, assemble in Audacity, apply gentle compression, export. That is perhaps an hour for a ten-minute script, and most of the hour is listening.

The limitation worth remembering

The naturalness of current speech synthesis is real and it makes the failures more dangerous, not less. A flat robotic voice announced itself. A voice that is indistinguishable from a person, that occasionally inserts a word that was not in the script, and that is being used for something consequential — a medical instruction, a legal notice, an emergency announcement — is a different risk profile entirely.

The control is not technical. It is that synthetic speech in any consequential setting gets listened to by a person against the script before it is used. Every organisation deploying this at scale eventually learns that, usually after an incident.

BEFORE YOU MOVE ON

Why does current speech synthesis sound natural where the older spectrogram-prediction approach sounded flat?

A. It samples a single delivery from a distribution over audio tokens rather than predicting the average of all valid deliveries.
B. It uses a much higher audio sample rate, which preserves the high-frequency detail that carries vocal expressiveness in human speech recordings.
C. It applies a separate prosody model after synthesis to add emphasis and intonation to an otherwise flat rendering.
D. It concatenates recorded fragments from a much larger library of human speech, so more natural transitions are available.

51 · 8 min

What a voice print actually is

THE ONE THING TO KEEP

A speaker embedding is a few hundred numbers summarising vocal identity, extracted from seconds of audio and used as conditioning, which is why cloning is fast, cheap and effectively impossible to prevent.

A few hundred numbers

A speaker encoder is a network trained on a task that sounds unrelated to synthesis: given two audio clips, say whether they are the same person. Trained on many thousands of speakers, it learns to produce a vector — typically 192 to 512 numbers — that is close for two recordings of one person and far apart for two people.

That vector is the voice print. It captures the things that make a voice recognisable: fundamental pitch range, the resonances of the vocal tract, the characteristic timing and attack.

To clone a voice, you compute that vector from a sample and use it as conditioning while generating. No training run, no fine-tuning, nothing stored beyond the vector. This is called **zero-shot** cloning and it is why the whole thing takes seconds.

How little audio is enough

The numbers matter here because people's intuitions are badly calibrated.

- **Three seconds** is enough for a recognisable clone with early research systems, and current ones do better.
- **Ten to thirty seconds** of clean speech produces a clone most listeners cannot distinguish from the source in a short clip.
- **Several minutes**, used for actual fine-tuning rather than zero-shot conditioning, produces a clone that holds up across long-form narration and unusual phrasing.

Quality of the sample matters far more than quantity. Clean, close, single-speaker, no music, no reverb, one consistent emotional register. A five-minute sample from a noisy video call produces a worse clone than fifteen clean seconds.

The uncomfortable consequence: everyone who has ever posted a video with their voice in it, left a voicemail greeting, appeared on a podcast, or spoken in a recorded meeting has published enough material. There is no meaningful way to withhold it, because the amount needed is below the threshold of ordinary social participation.

Fine-tuning versus zero-shot

Worth distinguishing, because the consent and security implications differ.

Zero-shot conditions on a vector at generation time. Nothing persists. The sample can be discarded. This is what a service does when you upload a clip and get speech back immediately.

Fine-tuning trains model weights on a person's recordings, producing a durable artefact — a file that is, in a practical sense, a copy of a voice. It

sounds better and it is a thing that exists, can be copied, shared, sold and stolen. Any agreement about a fine-tuned voice needs to say who holds the artefact, where it is stored, who may use it, and what happens to it when consent is withdrawn. "We deleted the recordings" is not the same as "we deleted the model", and a contract that only mentions the recordings has missed the point.

Why watermarking the output is not a control

Several services embed an inaudible watermark in generated speech. This is worth doing and it does not solve the problem, for reasons the provenance module covers in general: the watermark identifies audio that came from that service, and open models with no watermark exist and run on a laptop. A control that only binds the compliant is a useful record and not a defence.

What to do about your own voice

Three things, none of them heroic.

Assume it is clonable and act accordingly. Do not treat voice as an authentication factor for anything of yours, and ask your bank whether they do. Some do, and you can usually opt out.

Agree a verification method with the people who matter. A word, a fact, a callback number. This is the control that actually works, and it is covered in full at the end of this module.

Get the terms right when you consent. If you record for a client, read what the contract says about synthetic reproduction. A release written before 2022 almost certainly does not address it, and silence is not protection — it is an argument waiting to happen.

The honest limitation of all advice here: none of it prevents anyone from cloning your voice. The technology is distributed, the samples are public, and the barrier is not technical. What the advice does is remove the value of doing it, by ensuring that nobody who matters to you will act on a voice alone. That is a smaller claim than people want and it is the true one.

BEFORE YOU MOVE ON

Why does deleting the original recordings not fully satisfy a withdrawal of consent for a fine-tuned voice?

A. The fine-tuned model retains the voice as trained weights, which is a separate artefact from the recordings.
B. The recordings will have been backed up automatically, so deletion is rarely complete in practice.
C. A speaker embedding computed from the recordings persists in the service's cache and can be used to regenerate them.
D. The audio watermark embedded in previously generated output allows the voice to be reconstructed from any published clip that used it.

52 · 8 min

Prosody, and why it still sounds slightly wrong

THE ONE THING TO KEEP

Prosody carries meaning that the text does not, so a system given only text has to guess the interpretation, and its guesses are wrong in ways that are subtle and cumulative.

The information that is not in the words

Say "I never said she took the money" seven times, stressing a different word each time. Seven different meanings, one sentence. The stress carries information the text does not contain.

A text-to-speech system receives only the text. Everything about delivery — which word carries the emphasis, where the phrase breaks, whether the sentence rises at the end, how fast, how much energy — has to be inferred. The model does this from what usually happens with sentences of that shape,

which is right on average and wrong on the specific sentence often enough to notice.

This is why synthetic narration passes at the sentence level and fails at the paragraph level. Each line is plausible; the sequence has no argument running through it. Human readers place emphasis according to what they are *doing* with the passage — contrasting, conceding, listing, concluding — and no such intention exists in the model.

The specific failures to listen for

Once you know the list, you will hear them:

- **Contrastive stress missed.** "It was not the red one, it was the blue one" delivered with even weight, so the contrast disappears.
- **Lists flattened.** Every item read identically, with no rise-rise-fall shape, so the listener cannot tell when the list ends.
- **Questions that are not questions.** Rising intonation on a statement, or a flat delivery on a genuine question.
- **Wrong phrase breaks.** A pause inside a noun phrase rather than at the clause boundary. This is the one that most reliably reveals a synthetic reader.
- **Emotional monotony over length.** Three minutes at exactly the same energy, which no person produces.
- **Homographs.** "Read", "lead", "live", "bow", "record", "tear". Systems guess from context and get it wrong on the harder cases, and in some languages the equivalent problem is far worse.

What actually improves it

Reference audio for style. The strongest control available. Supply a short clip in the delivery you want — energetic, conversational, sombre — and the model conditions on the style as well as the identity. This works far better than any adjective, for the reasons the reference-image lesson gave: you are supplying the thing rather than a word for it.

Rewrite the text for the ear. Shorter sentences. Explicit connectives. Punctuation that marks the breath. Spelling out numbers and abbreviations the way you want them read. This is real work and it improves human narration too. Anyone who has written for radio already knows it.

Markup where it is supported. Some systems accept tags for emphasis, breaks and rate. The dialect varies; where it exists it is precise and worth learning for the twenty lines in a script that matter.

Split and re-roll. Generate line by line, keep the deliveries that land, regenerate the ones that do not. Assembling in Audacity is free and this is how professional-sounding synthetic narration is actually made. Nobody generates ten minutes in one pass and ships it.

Direct it like a performance. Where a system accepts an instruction alongside the text — "read this as if explaining to a friend" — use it, and be specific about intent rather than emotion. "Sceptical, then convinced by the end of the paragraph" outperforms "happy".

The part that remains

There is a category of prosodic meaning that no current system produces reliably: the delivery that depends on understanding what the passage is arguing. Irony, dramatic timing, a pause that lands because of what came three sentences earlier, the shift in register when a speaker changes their mind.

This is not a data problem in an obvious way, and it is a reasonable place to expect slow progress. It is also the reason that synthetic narration is now entirely adequate for instructional material, announcements, audio descriptions and drafts, and still noticeably inferior for anything where the reading is an interpretation — audiobook fiction, poetry, advertising with a joke in it, drama.

Knowing which of those two categories your job is in saves an enormous amount of argument. For the first, use synthesis and put the effort into the script. For the second, hire someone, or record it yourself.

One more practical note, because it is the fix most people never try. If a line will not come out right after four attempts, the problem is almost always the

writing rather than the model. Read the line aloud yourself. If you have to think about how to deliver it, so does the system, and it has less to go on. Splitting the sentence, moving the important word to the end, or replacing a subordinate clause with a full stop will usually produce a clean read on the first attempt. Writing for the ear is a real skill, it is learnable in a week, and it makes both synthetic and human narration better.

BEFORE YOU MOVE ON

Why does synthetic narration often pass at the sentence level and fail over a paragraph?

A. Long inputs exceed the model's attention window, so later sentences are generated without reference to earlier ones.

B. Audio token drift accumulates across a long generation, degrading quality progressively toward the end of the passage.

C. Prosody encodes the speaker's intention across the passage, and the model has no intention to encode.

D. The sampling temperature is fixed for the whole passage, so no variation in energy is possible between sentences.

53 · 8 min

The languages it does badly, and who that is

THE ONE THING TO KEEP

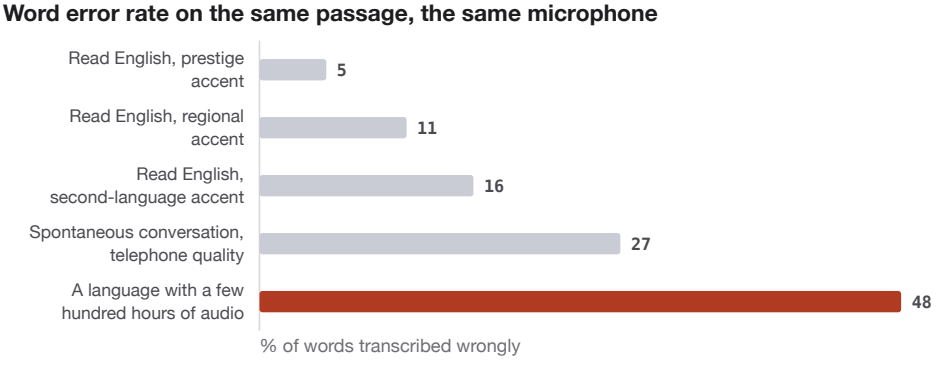
Speech systems inherit the distribution of their training audio, so quality varies by orders of magnitude across languages and accents, and the gap falls on exactly the people least able to route around it.

The unevenness, stated plainly

English has thousands of hours of clean, transcribed, permissively available speech. So do a handful of others — Mandarin, Spanish, French, German, Japanese. Below that, availability falls off a cliff.

There are roughly seven thousand living languages. Fewer than a hundred have speech systems anyone would call good. Several hundred million people speak languages with no usable speech recognition at all.

Within a well-served language, the same unevenness repeats across accents. Recognition error rates for a language's prestige accent and for a regional or second-language accent of the same language routinely differ by a factor of two or three. Published audits have found large disparities in commercial systems between speakers of different varieties of English, in the same recording conditions, saying the same things.



The figure a vendor quotes is the first bar. The material people actually have is the fourth. The gap is the training distribution rather than the speaker, and it falls on exactly the people least able to route around it.

This is not a design choice. It is the training distribution, exactly as the caption corpus was for images. But the consequence is not aesthetic here — it is a person being unable to use a voice interface, an automatic transcript of their evidence being wrong, or a subtitle system garbling their name.

What breaks specifically

Code-switching. Speakers of many languages mix in English or another language mid-sentence as a matter of ordinary speech. Most systems are built around one language per utterance and handle a switch by mangling both sides of it. Hinglish, Taglish, Spanglish and dozens of comparable everyday registers are, from the system's point of view, an error condition.

Names. Personal and place names outside the training distribution are transcribed phonetically and wrongly, which is a small daily indignity and occasionally a serious problem in medical or legal transcription.

Tone and length distinctions that English does not make are frequently flattened by both recognition and synthesis, changing meaning rather than merely accent.

Scripts with ambiguous vowel marking — Arabic, Hebrew, and abjads generally — require the system to infer pronunciation that the text does not carry. Diacritics resolve it and are usually absent in ordinary writing.

What is being done, and what you can do

The serious work here is community-driven and worth knowing about because you can join it.

Open, community-collected speech corpora exist that anyone can contribute to by reading sentences aloud, and anyone can use to train. Several languages have crossed the threshold from unusable to usable purely on this basis, contributed by their own speakers.

The practical routes if your language is poorly served:

- **Fine-tune an open recognition model** on a few hours of your language or accent. This is genuinely achievable — free notebook time is enough — and the improvement on a specific accent from even a small, well-matched set is substantial.
- **Use a language-specific model** rather than a multilingual one where it exists. The multilingual giants are convenient and frequently worse than a small model trained on one language.
- **Contribute recordings** to an open corpus. This is slow, collective and the only thing that fixes it durably.
- **Check the numbers before you promise anything.** Published error rates are usually reported on clean read speech in the best-served accent. Test on your own audio before building a product on it.

Why to be sceptical of "supports 100 languages"

That claim is now standard and it means very little. Support ranges from "trained on thousands of hours" to "trained on a few hundred, mostly religious texts read aloud". The word covers both.

The question that separates them: what is the measured error rate on this language, on spontaneous conversational speech, in ordinary recording conditions? Vendors that have the number will give it. Vendors that quote a single headline figure across all languages are quoting the best one.

The unresolved part, and it is worth naming as unresolved: nobody has a route to good coverage of the long tail that does not depend on the speakers of

those languages doing the data collection themselves, largely unpaid. That is a real problem of fairness with no agreed answer, and the current arrangement — that the world's least-served speakers must volunteer to build their own tools — is defended by some as community ownership and criticised by others as free labour. Both readings are available and both are honest.

BEFORE YOU MOVE ON

A vendor advertises support for 100 languages. Which question most usefully tests the claim?

A. How many hours of training audio were used in total across all supported languages combined?

B. Which of those languages are supported for synthesis as well as recognition?

C. Does the system handle code-switching between any two of the supported languages?

D. What is the measured error rate for this language on spontaneous speech?

---

54 · 8 min

The other half: turning speech into text

THE ONE THING TO KEEP

Recognition models are trained to produce plausible transcripts, so they fail by writing fluent text that was not said, which is more dangerous than failing by writing nothing.

Why this belongs in a course about making things

Every workflow in this module runs both directions. Subtitles, dubbing, transcripts, searchable video, editing by editing the text — all of it depends on recognition, and it fails in ways that are easy to miss precisely because the output reads well.

How it works and what that predicts

Current systems are sequence models. Audio is encoded into features; a decoder generates text tokens conditioned on them, one at a time, exactly like a language model.

That last part is the important one. The decoder is a language model. It produces text that is *fluent* whether or not it is *accurate*, because fluency is what the objective rewarded.

The characteristic failures follow:

- **Hallucinated text during silence or noise.** A recording with a long pause can produce an invented sentence, frequently something that appeared in the training data — a subtitle credit, a stock phrase, a repeated line. Studies of medical and general transcription have documented invented content, including invented clinical statements, at rates that are small per segment and significant across a corpus.
- **Plausible substitutions.** An unfamiliar name or term is replaced with a common word that fits the sentence. The transcript reads perfectly and says something else.
- **Repetition loops**, where a phrase repeats for many seconds.
- **Silent language switching** on multilingual audio, translating rather than transcribing.

Compare this with the older approach, which produced obvious rubbish when it failed. Obvious rubbish is safer. A confident wrong transcript is trusted.

The numbers, honestly

Word error rate on clean, read English is in the low single digits for the best models, which is genuinely at or near human parity for that material. On the material people actually have — overlapping speakers, accents, background noise, telephone bandwidth, domain vocabulary — it is commonly 10 to 25%, and worse for under-served accents.

Do not plan on the headline number. Test on twenty minutes of your own real audio and count the errors yourself. It takes an hour and it is the only figure

that means anything for your work.

Working with it properly

Improve the audio first. This is the highest-return action by a wide margin. A close microphone, a quiet room and separate tracks per speaker will do more than any change of model. Free noise reduction in Audacity is worth a pass; aggressive processing is not, because it removes the detail the model needs.

Supply context where the system allows it. Many systems accept an initial prompt or a vocabulary list. Feeding in the names, jargon and product terms that appear in the recording measurably reduces substitutions.

Segment on silence rather than on fixed lengths. Cutting mid-word produces errors at every boundary.

Never ship an unreviewed transcript of anything consequential. Legal, medical, journalistic, or anything a person's account of themselves depends on. The failure mode is fluent invention, and only a human comparing against the audio catches it.

Keep the audio. The transcript is a derived artefact and it is wrong sometimes. The recording is the record.

The free path

Whisper and its many optimised community implementations are free, run offline, and are strong across a wide range of languages. Local processing also means confidential audio never leaves the machine, which matters for legal, medical and journalistic work and is frequently a contractual requirement.

Free editors — Kdenlive, Shotcut, DaVinci Resolve's free edition — have subtitle workflows that import a transcript and let you correct it against the picture, which is by far the fastest way to review one.

There is one setting worth knowing about because it changes the failure profile. Many recognition systems will accept a prompt or an instruction alongside the audio, and some accept a request to transcribe verbatim including

hesitations and repetitions. Turning that on makes the output less pleasant to read and considerably more faithful, because the model is no longer tidying speech into prose. For interviews, evidence and anything where what was said matters more than how it reads, verbatim is the right mode, and the tidy default is quietly editorialising.

The limitation to hold onto: recognition quality is not a single number, it varies enormously with the speaker and the conditions, and the errors it makes are the fluent kind. The discipline that follows is simple and rarely observed — measure it on your own material, and read the transcript against the audio whenever it matters.

BEFORE YOU MOVE ON

Why is a modern recognition system's failure mode more dangerous than an older system's?

A. It fails silently by dropping segments, so missing content leaves no visible trace in the transcript.

B. It produces fluent, plausible text that was never spoken, so errors do not announce themselves.

C. It reports a confidence score that is poorly calibrated, so users cannot tell which segments to check.

D. It processes audio in fixed-length windows, so errors cluster at the boundaries where sentences are cut in half.

55 · 8 min

The dubbing chain, and where it breaks

THE ONE THING TO KEEP

A dubbing pipeline is four error-prone models in series, so errors compound and the only reliable control is a human check at each stage rather than at the end.

Four stages, four failure modes

Automatic dubbing chains together:

1. **Recognition** — speech to text in the source language.
2. **Translation** — source text to target text.
3. **Synthesis** — target text to speech, often in the original speaker's cloned voice.
4. **Alignment** — fitting the new audio to the timing, sometimes with lip adjustment.

Each stage takes the previous stage's output as truth. A name misheard at stage one is translated confidently at stage two, spoken clearly at stage three, and lip-synced convincingly at stage four. The pipeline's output is fluent, well-timed and wrong, with nothing anywhere signalling a problem.

This compounding is the central thing to understand. If each stage is 95% right, the chain is not 95% right.

The specific breakages

Recognition errors on names and terms propagate all the way through, and they are precisely the words a viewer will notice.

Translation without register. Machine translation chooses a register and often chooses wrong: formal where the speaker was casual, or the wrong sec-

ond-person form in languages that distinguish them. In several languages this makes a speaker sound rude or absurdly deferential, which is a serious failure that no automatic check catches.

Length mismatch. Languages differ in how long the same content takes to say — a translation can be 30% longer than the original. The pipeline then either speeds up the speech, which sounds rushed, or lets it run over the shot, which breaks the edit. Good human dubbing solves this by rewriting for length, which is a craft skill with a name — adaptation — and it is the part automatic systems handle worst.

Cloned prosody carried across languages. A voice print captures the speaker's identity; the delivery patterns of the source language do not necessarily suit the target. The result can sound like a foreigner reading the target language even though the accent is correct.

Cultural content. Idioms, jokes, references and units. A joke translated literally is not a joke, and nothing in the chain knows a joke was there.

Where automatic dubbing is genuinely good

Being fair about this matters, because the failure list makes it sound useless and it is not.

It works well for **informational content with a controlled script**: training material, product explanations, lectures, documentation. The register is neutral, the terms can be supplied in advance, the timing is flexible, and there are no jokes.

It works well as a **draft** for a human adapter, who then fixes register, length and the twenty lines that matter. This is the shape most professional workflows have converged on, and it makes translation affordable for material that would never have been dubbed at all.

It works well for **accessibility**, where a rough version now beats a perfect version never.

The controls that make it safe

Check the transcript before translating. Ten minutes of review at stage one removes most of the compounding. This is the single highest-value intervention in the chain.

Have a speaker of the target language review the translation. Not a checker of grammar, a checker of register and meaning.

Listen to the final audio against the original. Not read it — listen.

Disclose it. A viewer watching a dubbed video in a cloned voice should know. Several jurisdictions now require this in some form, which the final module covers, and the ethical case does not depend on the legal one.

Get consent for the voice, from the speaker, specifically for synthetic reproduction in other languages. A performer's agreement to be filmed is not an agreement to speak Portuguese.

There is a cheaper alternative that deserves more consideration than it gets: **subtitles**. They are far more reliable, they cost a fraction as much, they preserve the original performance entirely, and a large part of the world's audience already prefers them. Dubbing is the right answer for young children, for people who cannot read the subtitles comfortably, and for material watched while doing something else. For everything else, the choice between a good subtitle track and a mediocre automatic dub is not close, and the dub is frequently chosen because it sounds more impressive rather than because it serves the viewer better.

The honest limitation of the whole area: this is a chain of statistical systems, none of which knows what it is talking about, producing output that is confident at every stage. It is enormously useful and it is not a fire-and-forget pipeline. Every organisation that has deployed it at scale has learned this by publishing something embarrassing, and the ones who put a human at stage one learned it more cheaply.

BEFORE YOU MOVE ON

Why is checking the transcript the highest-value intervention in an automatic dubbing chain?

A. Recognition is the least accurate stage of the four, so it contributes most of the total error in the finished dub.
B. Transcript errors are the only errors a reviewer without knowledge of the target language can detect.
C. It is the first stage, so an error there is treated as truth by translation, synthesis and alignment in turn.
D. Transcripts are the only artefact in the chain that can be corrected without regenerating the audio.

56 · 8 min

Can you tell? Mostly not

THE ONE THING TO KEEP

Synthetic-speech detection relies on artefacts that compression and re-recording destroy and that new models do not produce, so detection is a weak signal and the practical defence is verification instead.

What detectors look for

Three broad families, and each has a specific weakness.

Artefact detectors look for traces of the generation process: characteristics of the neural codec, unnatural regularity in the noise floor, missing very-high-frequency content, absent breath and mouth noise. These work well on raw output from known systems.

Their weakness is that the artefacts live in exactly the parts of the signal that compression discards. Audio sent through a phone call, a messaging app or a

social platform has been compressed heavily, and much of the evidence goes with it. Re-recording — playing the audio and capturing it with a microphone — removes nearly all of it.

Learned classifiers are trained on large sets of real and synthetic speech. They perform impressively on their test sets and drop sharply on generators they were not trained on, which is every generator released after them.

Watermark detectors look for a deliberately embedded signal. These are reliable when the signal is present and say nothing when it is absent, which covers every open model and every deliberate misuse.

The numbers to hold in your head

Published detection accuracies of 95% or more are common and are measured under favourable conditions: known generators, clean audio, no compression. Under realistic conditions — unseen generator, telephone-quality audio — reported figures fall substantially, and there is an active research literature on exactly how far.

Then apply the base rate. Suppose one call in a thousand to your organisation uses a cloned voice, and your detector is 95% accurate in both directions. Out of a thousand calls you flag the one real fake and about fifty genuine callers. Fifty false accusations for one catch is not an operational control; it is a way of insulting your customers.

This base-rate arithmetic is the same one that governs medical screening and fraud detection generally, and it is the single most important idea for anyone deploying a detector of any kind. The evaluation course in this catalogue works through it properly.

What humans can hear, and for how much longer

People perform close to chance on short clips of current systems, and better on longer ones, where prosodic monotony and the absence of breath give it away. Trained listeners do better than untrained ones. Everybody does worse over a telephone.

The tells that remain, for what they are worth: unnaturally consistent pacing, no breath before long sentences, background that is either perfectly silent or a loop, and emotional flatness under stress — a caller who claims to be in an emergency but whose voice does not have the physiological signature of one.

Do not build anything on these. They are being closed, deliberately, because naturalness is the product.

What to do instead

The whole weight should sit on verification rather than detection, because verification does not care how good the generator is.

- **A shared secret** with family: a word, a question with an answer only they know. Agree it now, in person.
- **Call back on a known number.** Not the number that called; the number in your address book.
- **A second channel** for anything consequential — a message, an email, a different app.
- **Institutional process** for money and access: no payment instruction acted on from a single voice or video contact, ever, regardless of how certain the recipient is.

These are unglamorous and they work completely, because they make the attacker's capability irrelevant rather than trying to out-run it.

There is one setting where detection does earn its place, and it is worth naming so the lesson does not read as blanket dismissal. As a **triage signal inside a system that already has a verification step**, a detector adds value: flag the call for a callback rather than for refusal, route the flagged case to a slower path rather than a rejection. The false positives then cost a few seconds of extra checking instead of an accusation. The rule that follows is a good one generally — a weak signal is useful when it routes and harmful when it decides.

The honest summary: treat any detector's output as a weak signal to investigate, never as a verdict and never as grounds for a public accusation. The literature on this is clear, the failure of the same approach for generated text is in-

structive — one widely-publicised text detector was withdrawn by its own developer for low accuracy — and the same forces apply here.

BEFORE YOU MOVE ON

A support desk deploys a synthetic-voice detector reported at 95% accuracy. Roughly one caller in a thousand is an impersonator. What should they expect?

A. Around half of impersonators will be caught, since accuracy figures are typically halved under telephone conditions.

B. The detector will be reliable, because 95% accuracy on a binary decision is far above chance.

C. Detection accuracy will improve over time as the system accumulates examples from its own call traffic.

D. Far more genuine callers flagged than impersonators caught, because at that base rate most positives are false.

57 · 8 min

The fraud that already happens

THE ONE THING TO KEEP

Voice-impersonation fraud follows a fixed script — urgency, authority, secrecy, an irreversible payment — and the control that defeats it is an out-of-band check, which works regardless of how convincing the voice is.

Two patterns, both documented

The family emergency. A call from a relative's voice: an accident, an arrest, a hospital, an urgent need for money and a request not to tell anyone else. The voice is right. The details are right, because they are on social media. The caller is under time pressure and asks for a transfer, a gift card code, or a courier.

The corporate instruction. A call or video call from a senior colleague's voice authorising an urgent payment or a change of bank details. The largest reported cases of this kind have involved multi-million-pound transfers, including a widely reported incident in which an employee joined a video call populated entirely by synthetic colleagues and made a series of transfers.

Neither pattern is new. Both are decades-old confidence tricks. What has changed is the cost of making the voice convincing, which has fallen to nearly nothing, and the availability of samples, which is total.

The script, and why naming it helps

Every version of this contains four elements:

1. **Urgency.** A deadline measured in minutes.
2. **Authority.** Someone whose instruction you would not normally question.
3. **Secrecy.** A reason not to check with anyone else — a confidential deal, an embarrassment, a legal restriction.
4. **An irreversible transfer.** Bank transfer, cryptocurrency, gift cards, cash to a courier.

Teach those four to anyone you care about. Recognising the *shape* is more robust than recognising the voice, because the shape does not improve with the technology. Any request containing all four is fraudulent, whoever appears to be making it, and there is no legitimate business or family situation in which all four genuinely apply.

The four elements, and the control that defeats all of them

Urgency	A deadline measured in minutes, so there is no room to think.
Authority	A voice whose instruction you would not normally question: a child, a chief executive.
Secrecy	A reason not to check with anyone else — a confidential deal, an embarrassment, a legal restriction.
An irreversible transfer	Bank transfer, cryptocurrency, gift card codes, cash to a courier.
Call back on a number you already had	Not the number that called. This works whatever the voice sounds like, which is the point: it makes the attacker's capability irrelevant instead of trying to out-run it.

Any request carrying all four is fraudulent, whoever appears to be making it. There is no legitimate family or business situation in which all four genuinely apply, and recognising the shape does not expire the way recognising a voice does.

The controls, in order of value

A family safe word. Agreed in person, never shared electronically, never used except in this situation. It costs a two-minute conversation and it ends this attack completely for the people you love. Do it this week. Choose something a stranger could not guess and a relative under stress could still remember.

Call back on a known number. The single most effective control and the most frequently skipped, because hanging up on a distressed relative feels wrong. Say plainly: "I am going to call you straight back on your usual number." A genuine caller will wait. This is worth rehearsing as a sentence, because in the moment people find it hard to say.

Dual authorisation for payments. No single person can move money or change payee details on one contact. This is standard financial control and it defeats the corporate case outright.

A written policy that says voice and video are not authentication. Explicit, so that a junior member of staff refusing a chief executive's instruction is following policy rather than risking their job. This is the part organisations get wrong: the technical control exists and the social pressure defeats it. Make refusal the safe option.

Reduce the sample surface, marginally. Voicemail greetings in a generic voice, care with public video. This helps a little and it is not a defence, for the reasons the voice-print lesson gave. Do not mistake it for one.

For the people most at risk

Older relatives are targeted disproportionately and the reason is not gullibility. It is that the attack exploits exactly the impulse to help a grandchild in trouble, and that is not a defect.

The conversation that works is not "be careful of scams". It is concrete: agree the safe word, agree that you will always be able to call back, and agree in advance that hanging up to check is a normal thing to do and will never cause offence. Removing the social cost of checking is the whole intervention.

What to do if it happens

Stop the transfer immediately with the bank — recovery is sometimes possible in the first hours and rarely afterwards. Report it to the national fraud reporting body. Keep the call details and any recording. Tell other people, because the same attacker works through a contact list, and because the shame that stops people reporting is precisely what makes the pattern profitable.

The honest note to end on: none of this depends on anything about how these systems work. It is old advice, made newly urgent by a capability that is now cheap and cannot be recalled. The technical chapters of this course explain why detection will not save anyone. This lesson is the part that actually protects people, and it costs nothing to put in place.

BEFORE YOU MOVE ON

Why does an out-of-band callback defeat voice-impersonation fraud regardless of how convincing the clone is?

A. Because it introduces a delay, and fraudulent transfers can usually be reversed within that window.

B. Because network carriers can trace a returned call to its true origin and flag the impersonator.

C. Because voice clones cannot maintain a conversation for the length of a second call, so the impersonation collapses.

D. Because it verifies the channel rather than the content, and the attacker does not control the number you dial.

CASE STUDY · MODULE 6

The voice that opened the account

A co-operative credit society in Kolhapur, after a member's voice from a public meeting was used on its telephone line

The Sahakari Patsanstha in Kolhapur has about nine thousand members. A large number of them are farmers and retired mill workers, and a meaningful number have no smartphone at all.

In 2023 the society added voice verification to its telephone line. A member calls, speaks a short passphrase, a speaker-verification system compares it against an enrolled sample, and a clerk can then read a balance, discuss a loan instalment or accept a change to a standing instruction. Complaints about the phone line fell by roughly two-thirds inside a year. The manager, Prakash Kulkarni, considered it the best thing the society had done for its older members in a decade.

In March a member in Ichalkaranji discovered that his balance had been read out to someone, his registered mobile number had been changed, and two transfers totalling one lakh forty thousand rupees had left the account over the following week.

The member had not called. His son had not called. What had happened was simpler than anyone expected. The member had spoken for about ninety seconds at a co-operative federation's annual meeting the previous year, and the video was on a Facebook page with four hundred followers.

A speaker embedding — a few hundred numbers summarising vocal identity — can be extracted from a few seconds of clean speech and used to condition a speech model directly, with no training run at all. Ninety seconds of a man speaking clearly into a microphone at a public meeting is not a marginal sample. It is a good one.

The society recovered about forty thousand rupees and wrote off the rest.

At the board meeting in April there were two proposals.

The first was to remove voice verification entirely and go back to one-time passwords. This is the safe-sounding answer and it has a real cost. The OTP flow is what produced the complaints in the first place. Around six hundred members have no smartphone, and for a farmer eleven kilometres out during the monsoon, being told to come to the branch to change a mobile number is not a minor inconvenience.

The second came from a younger director who had been approached by a vendor. A synthetic-speech detector, one lakh ten thousand rupees a year, quoted at ninety-six per cent accuracy. It preserved the service and it addressed the attack directly.

Prakash's nephew, who works in software in Pune, was at the meeting because he had driven his uncle there. He asked for the whiteboard and did an arithmetic problem that took four minutes.

Suppose one call in a thousand uses a cloned voice, which is generous. The society takes about ten thousand calls a month. That is ten fraudulent calls, of which a ninety-six per cent detector catches about ten. It also examines nine thousand nine hundred and ninety genuine calls and flags four per cent of them, which is about four hundred.

Four hundred members a month, most of them elderly, most of them people for whom the telephone line is the only route into their own money, treated as suspected impostors, so that ten fraudulent calls are caught.

The room went quiet, the director said the vendor had not mentioned that, and the nephew said the vendor would not have, because the ninety-six per cent is a true number and the four hundred is what the true number does at this base rate.

STOP HERE

Before turning the page: what would you have done, and what would it have cost if you were wrong? Write it down. The point of the next section is to compare.

WHAT HAPPENED

The board took neither proposal.

Voice stopped being authentication and became convenience. A member may still speak the passphrase and be greeted by name, and the clerk may still discuss a balance. But nothing that moves money and nothing that changes a contact detail happens on the strength of that call. A payment instruction or a standing-instruction change is confirmed by the clerk calling back on the number already held on file — not the number that called — and a change to the registered number itself is done at the branch, in person, with the passbook.

The rule went onto a laminated card at every clerk's desk. The sentence Prakash considers most important on that card is the last one: a clerk who declines to act on a telephone instruction is following society policy and will not be questioned for it. The March incident had been processed by a clerk who thought the caller sounded impatient and did not want a complaint about him.

They also sent a printed note to every member about the other kind of call — the grandchild in hospital, the urgent transfer, the request not to tell anyone — suggesting a family word agreed in person, never sent by message, and making the point that hanging up to call back is a normal and reasonable thing to do. Eleven months later two members reported exactly that call. Both hung up. Neither lost anything.

The detector was not bought. The director who had proposed it now does the base-rate arithmetic himself before buying anything, and has said so more than once, with some feeling, about a different vendor.

Ninety seconds of a man speaking at a public meeting was enough. What does that imply about advising members to protect their voices?

That the advice cannot work as a defence. Cloning is zero-shot: a speaker encoder produces a vector of a few hundred numbers from a short clean sample, and that vector is used as conditioning at generation time with no training run. Three to thirty seconds is the working range, and quality of the sample matters far more

than length. The amount required is below the threshold of ordinary social participation — a voicemail greeting, a video, a recorded meeting — so nobody can withhold it. Advice about reducing the sample surface is marginal. The control that works removes the value of the clone by ensuring nobody acts on a voice alone.

The detector was quoted at ninety-six per cent accuracy, and the nephew's arithmetic produced four hundred flagged members a month. Explain where those four hundred come from.

From the base rate. Accuracy is a rate per call, and the number of errors it produces depends on how many calls of each kind there are. With one fraudulent call in a thousand, ten thousand monthly calls contain about ten fakes and about nine thousand nine hundred and ninety genuine calls. A four per cent false-positive rate applied to the large group is about four hundred people, against about ten correct catches from the small one. The detector is working exactly as specified; the rarity of the event swamps a good error rate. This is why any screening system for a rare event needs a confirmatory step rather than a decision.

Why does the card say that a clerk who declines will not be questioned, and why is that sentence load-bearing rather than decorative?

Because the attack works on social pressure, not on technology. The fraud script is urgency, authority, secrecy and an irreversible transfer, and the authority element is aimed precisely at making refusal feel costly to the person answering the phone. A policy that a clerk is technically allowed to invoke but will be blamed for invoking is not a control. Naming refusal as the safe option converts declining from an act of individual courage into compliance, which is the only version that survives a caller who sounds impatient and senior.

MODULE 6 · TEST

Check what stuck

6 questions, one right answer each. This one is for you, not for a record: answers and the reason behind each are at the back. If two questions from the same module go wrong, re-read that module before the exam rather than after.

1 A synthetic narration inserts a word that was not in the script, and occasionally trails off into unrelated babble. Which property of the architecture predicts that?

A. The vocoder reconstructs phase information and sometimes fills gaps with noise.

B. The pronunciation dictionary substitutes an entry when a word is not recognised.

C. The system predicts a sequence of audio tokens by sampling from a distribution.

D. The speaker embedding drifts across a long passage and pulls the delivery off script.

2 Cloning a voice from a short sample takes seconds and produces no file that is recognisably a copy of anyone. What is actually happening?

A. The model is fine-tuned on the sample using a very high learning rate.

B. A speaker encoder produces a vector of a few hundred numbers, used as conditioning at generation time.

C. The sample is stored and fragments of it are concatenated to form new words.

D. The sample is converted into a pronunciation profile that adjusts the vocoder's output.

3 An automatic transcript of a long interview reads perfectly and contains a sentence nobody said. Why is that failure more dangerous than the older kind?

A. It occurs more often than the errors older systems produced.

B. It is concentrated in the opening minutes, which reviewers read most carefully.

C. It is fluent, so nothing in the text signals that anything went wrong.

D. It cannot be corrected without re-running the whole recording.

4 An automatic dubbing pipeline delivers a fluent, well-timed Portuguese track in which a speaker's name is wrong throughout. Where is the highest-value place to intervene?

A. At synthesis, by supplying a pronunciation hint for the name.

B. At alignment, by re-timing the affected lines against the picture.

C. At translation, by giving the translator a glossary of proper nouns.

D. At recognition, by reviewing the source transcript before anything else runs.

5 A helpline takes 10,000 calls a month, about ten of which use a cloned voice. A detector accurate 95 per cent of the time in both directions is deployed. Roughly what happens?

A. About ten fakes are caught and about five hundred genuine callers are flagged.

B. About ten fakes are caught and fewer than twenty genuine callers are flagged.

C. About half the fakes are caught, with a negligible number of false flags.

D. The detector's accuracy improves over time as it sees more of the helpline's calls.

6 Which single control defeats voice-impersonation fraud regardless of how convincing the clone becomes?

A. A synthetic-speech detector on the incoming line.

B. Keeping recordings of your voice off public platforms.

C. Training staff to listen for absent breath and flat emotion under stress.

D. Confirming any consequential instruction on a separate, already-known channel.

MODULE 7

Music and sound

Music is the medium where the technical problem and the legal problem are hardest at the same time. This block covers how audio is generated at all, why music is harder than speech, what source separation makes possible, where generated sound effects are genuinely good, the state of the training-data dispute, when a melody becomes infringement, what platforms do about generated tracks, and where generated music is the right answer rather than the cheap one.

> BY THE END OF THIS MODULE YOU CAN
>
> Explain how a music model represents and generates audio, judge whether a generated track carries a real infringement risk and on what basis, and choose between generated music, licensed library music and a commissioned composer for a specific brief with reasons you could defend

---

58 · 8 min

Music, and whose music it learned from

> THE ONE THING TO KEEP
>
> Generated music can still infringe a real melody, and "a model wrote it" is not a defence.

How a generated song is made

Most commercial music generators are language models over audio tokens. A neural codec compresses a waveform into a stream of discrete tokens — in-

stead of 44,100 numbers per second, perhaps 50 to 150 tokens per second. A transformer predicts the next tokens, conditioned on your style text and your lyrics. A decoder turns tokens back into sound. Some systems instead run diffusion over an audio latent, which is closer to how an image model works.

The token framing explains the artefacts you can hear. Transients go gluey, because the attack of a snare or a plucked string is exactly the fast detail a codec discards. Cymbals smear. Consonants in vocals dissolve into the mix. And structure drifts after two or three minutes, because the model has no score, only a sequence it is extending.

What it is good for

Background beds under video. Temp tracks so a client can hear pacing before you commission a composer. Mood pieces, meditation audio, loops for a game menu. Sketching an arrangement in ninety seconds to find out whether the idea works at all. Jingles nobody will hear twice.

What it is not good for

Mixing properly. Stem separation exists on most platforms, but it separates a finished mix after the fact. You do not get the tracked parts, and the separated stems carry artefacts.

Precise control. You cannot say "drop the second guitar in bar 17". You reroll and hope.

A hook that survives thirty listens. Generated music is usually pleasant and rarely memorable. Fine for a bed. Fatal for a single.

Anything where a musician's identity is the point. That is not a technical limitation and no model release will fix it.

The training-data question, answered commercially

In June 2024 the major record labels sued the two largest music generation companies, alleging their models were trained on copyrighted recordings at

scale. It was set up to be the case that decided whether training on recordings without a licence is fair use.

It did not decide it. Through late 2025 those disputes moved toward settlements bundled with licensing deals: money changed hands, the companies agreed to licensed models, artist opt-ins and tighter download rules, and the underlying legal question went unanswered by any court. Separate claims from music publishers over compositions and lyrics have run on their own track.

That is worth sitting with. The largest open question in generative audio was resolved by commercial negotiation rather than by law. Which means the answer can move again with the next negotiation, and it gives you no precedent to rely on.

"In the style of" is three separate problems

Style is generally not protected by copyright anywhere. You cannot own "sounds like reggaeton", or even a particular producer's approach to drums.

A voice is protected in more and more places, as a personality or publicity right — see the previous lesson. A generated vocal that sounds like a specific living singer is the risky part, not the arrangement.

A melody is protected, and this is where people get caught. Substantial similarity is judged on the output. If the generated hook lands close enough to an existing hook, it infringes, and "a model made it" is not a defence — the model was trained on that catalogue, which makes the usual claim of independent creation hard to run.

Most services block artist names in prompts. Read that as a signal about where the companies think the risk sits, not as an assurance that you are clear.

Before you use a track commercially

Check which tier you are on. Most platforms grant commercial rights only on paid plans, and some retain rights in what free accounts generate. Check whether there is any indemnity; usually there is not. And whether you own the

output at all is the next lesson, where the answer is less comfortable than the terms of service imply.

BEFORE YOU MOVE ON

A studio uses a generated instrumental under an advertisement. A listener points out the hook is very close to a hit from the 1990s. What is the soundest analysis?

A. The platform's terms say the studio owns the output, so any liability sits with the platform, not the studio.

B. A model generates statistically rather than copying files, so a resemblance cannot amount to copying.

C. A short instrumental bed is a brief excerpt, and brief excerpts are treated as too small to matter.

D. Similarity is judged on the finished output, so how it was made does not change whether the hook infringes.

59 · 8 min

How a machine writes a waveform

THE ONE THING TO KEEP

Music models generate stacks of discrete audio tokens produced by a neural codec, so the codec's bitrate sets the ceiling on quality and the token rate sets the cost of length.

Audio is a lot of numbers

A stereo track at CD quality is 44,100 samples per second per channel. Three minutes is about 16 million numbers. Predicting those directly, one at a time, was tried and is impractical for anything beyond a few seconds.

The solution is the same one the voice module described, applied at higher quality. A **neural audio codec** learns to compress audio into a small stream of discrete tokens and to reconstruct it. A common configuration produces tokens at 50 per second, with four to eight stacked codebooks, so one second of music is a few hundred integers rather than 88,000 samples.

The stacking matters and is worth understanding. The first codebook carries the coarse structure — roughly, the shape of the sound. Each subsequent codebook encodes the residual error left by the previous ones, adding detail. This is **residual vector quantisation**, and it means you can trade quality for speed by generating fewer codebooks, which is exactly what fast preview modes do.

Given that representation, generation is sequence prediction: a transformer predicts the next tokens conditioned on a text description, a melody, or the tokens so far. The codec decoder turns the result back into sound.

The other approach, and why both survive

Some systems work on **spectrograms** instead — the time-frequency picture of a sound — treating them as images and applying diffusion. A vocoder converts the spectrogram back to a waveform.

This inherits the strengths of image diffusion: strong texture, parallel generation, good conditioning tools. It also inherits the vocoder problem, because a spectrogram discards phase information, and reconstructing phase is where the characteristic metallic or watery artefacts in some generated music come from.

Current strong systems are frequently hybrids. Which one you are using is usually stated in the model card, and it predicts the artefacts: token-based systems tend toward abrupt glitches and structural drift, spectrogram-based ones toward smearing and phasiness.

The numbers that govern quality

Codec bitrate is the ceiling. No amount of model capacity produces audio better than the codec can reconstruct. The test from the image module works

identically here: run a real track through the codec's encoder and straight back out, with no generation at all. What comes back is the best the system can ever produce. Do this once and you will stop expecting master-quality output from a model whose codec cannot deliver it.

Sample rate. Many open music models generate at 32 kHz rather than 44.1, which means nothing above 16 kHz. On a phone this is inaudible. On good headphones the top end sounds slightly closed, and it matters if the track will be mastered alongside recorded material.

Token rate times duration is the cost, which is why generated tracks are typically 30 seconds to a few minutes and why extending is done in chunks with the same drift problem as video.

What free tools exist

The open end of this field is genuinely usable:

- Open music-generation models run on consumer hardware and produce instrumental tracks of reasonable quality. Licences vary considerably — several are non-commercial only, which is a detail people miss until it matters. Read the licence before using output in paid work.
- Some open models are trained specifically on permissively licensed and public-domain audio, which materially changes the rights position and is worth seeking out.
- **Audacity** for editing, **LMMS** and **Ardour** for full production, all free.

Where the quality actually is

Being specific rather than dismissive: current generated music is good at texture, atmosphere and short loops. A four-bar bed, an ambient wash, a percussion loop, a sting — these are frequently indistinguishable from library music and take seconds.

It is weak at everything that depends on structure over time, which is the next lesson, and it is weak at anything with a lead vocal that has to be intelligible, expressive and in tune across a whole song.

There is one more setting worth finding in whatever tool you use: the number of codebooks or the quality tier. Fast preview modes generate fewer residual codebooks, which is why the draft sounds thin and grainy and the final render does not. People frequently judge a model on its preview output and conclude it is poor. Generate one full-quality render before forming an opinion, and use the preview for what it is good for, which is deciding whether the idea is right.

The limitation to carry: audio quality has a hard ceiling set by a component most users never look at, and structure has a soft ceiling set by the same attention-window problem as every other medium in this course. Neither is a matter of prompting better.

BEFORE YOU MOVE ON

A generated track sounds slightly closed at the top end compared with commercial recordings. What is the most likely cause?

A. The model generates at a sample rate that carries no content above about 16 kHz.

B. The generation used too few denoising steps, so the high-frequency detail was never resolved.

C. The prompt did not describe the desired brightness, so the model produced a darker mix by default.

D. The stereo field was collapsed to mono during generation, which removes the perception of high-frequency air in the recording.

60 · 8 min

Why music is harder than speech

THE ONE THING TO KEEP

Music requires structure over minutes and simultaneous independent voices, both of which fall outside a model's attention window and neither of which local plausibility supplies.

Two problems speech does not have

Long-range structure. A song has form. A verse, a chorus that returns and is recognisably the same chorus, a bridge that departs and resolves, a harmonic plan that creates and discharges tension over three minutes. Almost none of that is local. A listener who hears the chorus return is comparing it with something they heard ninety seconds earlier.

That comparison happens outside any current model's attention window. The result is the most common complaint about generated music: it sounds convincing for fifteen seconds and then wanders. Sections appear and do not return. The energy stays flat. A key change happens for no reason and is not resolved.

This is the same window limitation as everywhere else in this course, and music is the medium where it is most audible, because form *is* the content.

What the window holds, and what a song needs

Inside a few seconds — reliable	Over three minutes — nothing is tracking it
• Timbre and the texture of an arrangement	• A chorus recognisable as the same chorus
• A groove that sits right	• Tension created and then discharged
• One chord change into the next	• A cadence that ends rather than fades
• A four-bar bed, a loop, a sting	• An energy shape across the whole piece
• Atmosphere	• Two independent lines that keep agreeing

Speech reached convincing quality first because its structure comes from the text you supply. Nothing supplies a song's form, so the model must invent it over a duration far longer than anything it can attend to. A longer generation is six minutes of wandering rather than three.

Polyphony. Speech is one voice. Music is several simultaneous independent lines that must be consistent with each other harmonically and rhythmically while remaining distinguishable. The bass and the melody have to agree about the chord and disagree about the note.

A model generating a single token stream that decodes to a mixture is not tracking parts. It is producing something that sounds like a mixture, which works locally and produces the characteristic mush when the arrangement gets busy.

The failures to listen for

- **Structure that does not return.** No recognisable repeat of a section.
- **Endings that do not end.** The track fades or stops rather than resolving, because a cadence is a structural event.
- **Rhythmic drift** over long durations, particularly at joins between generated chunks.
- **Harmonic wandering** with no destination.
- **Lyrics that dissolve** into vowel sounds, because intelligible sung words are the hardest case in the whole field: pitch, rhythm and phonetics simultaneously.
- **Instruments that morph** — a guitar becoming a synth mid-phrase, the same failure as an object changing in a video.

What improves it

Conditioning on structure. Systems that accept a melody, a chord chart or a reference track produce far more coherent results than text alone, for the same reason a depth map beats a description of composition. If your tool accepts a hummed melody or a MIDI input, that is the strongest control it has.

Generating sections separately and arranging them. Generate a verse and a chorus, then build the arrangement in a free digital audio workstation. Structure supplied by you, texture supplied by the model. This is the same division of labour as compositing in the image module, and it is what people who get good results actually do.

Short loops. Anything under about fifteen seconds is inside the reliable zone. A great deal of practical use — beds, stings, loops, transitions — lives there.

Explicit form in the prompt, where supported: "intro, verse, chorus, verse, chorus, outro" helps some systems and is ignored by others. Test yours rather than assuming.

The comparison worth making

It is instructive that speech synthesis reached convincing quality before music did, given that speech carries meaning and music does not.

The reason is exactly this. Speech is one voice, and its structure is supplied by the text you provide. The model handles perhaps a sentence of context at a time and that is enough, because the coherence of a paragraph comes from the writing.

Music has no equivalent external source of structure. Nothing supplies the form. The model has to invent it, over a duration far longer than its window, with no scaffold.

That is why the strongest results come from putting the structure in yourself, and why the systems that improve fastest are the ones that accept structural conditioning rather than the ones that generate longer.

A test that shows you the ceiling

Generate a two-minute track and listen to it twice: once from the beginning, and once starting at the ninety-second mark. If the second listen tells you nothing about where you are in the piece — no sense that this is the end rather than the middle — the model has produced texture rather than form.

Do this deliberately rather than waiting to notice it. It takes four minutes, it is the clearest demonstration of the limitation in this whole module, and after you have heard it once you will stop expecting the next release to fix it by generating longer.

The honest limitation: this is not close to solved, and length is not the fix. A model that generates six minutes instead of three produces six minutes of wandering. What would fix it is an explicit representation of musical form, and that is a research direction rather than a product feature.

BEFORE YOU MOVE ON

Why does generated music tend to wander after fifteen seconds while generated speech holds together?

A. Music contains several simultaneous instruments, and these consume the model's available attention window far faster than a single speaking voice does.

B. Music models are trained on far less data than speech models, so they have seen fewer examples of long-form structure.

C. Speech has its structure supplied externally by the text, whereas musical form must be invented over durations no window covers.

D. Speech is generated at a lower sample rate, so a given attention window covers more seconds of audio.

---

61 · 8 min

Pulling a mix apart

THE ONE THING TO KEEP

Source separation splits a finished mix into instrument stems with free tools at usable quality, which enables real work and does not change the rights position of the recording at all.

What it does

Give a separation model a stereo mix and it returns several tracks: vocals, drums, bass, and everything else. Better models split further — guitar, piano,

strings.

The mechanism is a network trained on pairs of mixes and their known stems, learning to predict a mask over the time-frequency representation for each source. It is not unmixing in a mathematical sense; the information genuinely overlaps and cannot be perfectly recovered. It is a very good estimate.

Demucs is the best-known free implementation, released under a permissive licence, and it runs on ordinary hardware. Several free web tools wrap it. Quality on modern well-produced music is high enough that separated stems are used in professional workflows.

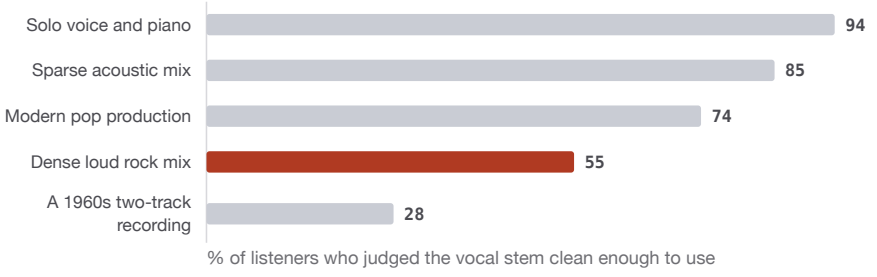
Where the artefacts are

Knowing them tells you when separation is fit for purpose.

- **Reverb tails** belong to whichever source they came from and get split badly, so a vocal stem carries a ghost of the room and the instrumental carries a hole where the reverb was.
- **Overlapping frequency content** — a snare and a vocal consonant in the same band — produces momentary bleed in both directions.
- **Cymbals** smear, because their content is broadband noise, which is exactly what the model cannot attribute confidently.
- **Dense mixes** separate worse than sparse ones. A solo voice with piano comes out nearly perfect; a loud rock mix does not.
- **Older recordings** with everything on two tracks separate poorly, because the sources were never independent.

The practical test is to separate, then sum the stems back together and compare with the original. What you hear missing or added is the error, and it tells you exactly how much you can trust the stems for your purpose.

How cleanly a mix comes apart



Separation is a very good estimate, not an unmixing: the information genuinely overlaps and cannot be perfectly recovered. Sources that were never independent cannot be made independent. Sum the stems back and compare with the original, and what you hear missing is the error.

What it makes possible

Remixing and mashups — the obvious use, and the one with the rights problem below.

Practice and study. Removing the guitar to play along with, isolating the bass line to learn it, slowing a solo down. This is transformative for anyone learning an instrument and it is free.

Restoration and repair. Removing a cough, fixing a bad note, rebalancing a live recording that was mixed badly.

Cleaning dialogue. Separating speech from background music in a recorded interview where music was playing — a common problem in documentary work.

Karaoke and rehearsal tracks, made in seconds rather than sourced.

Retiming and re-editing music to picture, which is far easier with stems than with a mix.

The part people get wrong

Separating a track does not give you rights to it. This is worth saying plainly because the technical ease creates a strong intuition otherwise.

A vocal stem extracted from a commercial recording is a derivative of that recording. Using it in something you publish involves the recording copyright,

the composition copyright, and usually the performer's rights, exactly as sampling the original would. That the extraction was automatic changes nothing.

The uses that are uncomplicated: your own recordings, practice at home, material you have licensed, public-domain recordings, and anything where the rights holder has granted a remix licence. The uses that are not: publishing a remix of a commercial track without a licence, whatever the platform culture suggests, and building a cloned voice from a separated vocal, which adds a personality-rights problem to the copyright one.

What separation is doing to the field

Two consequences worth noticing.

Stems have become an expected deliverable. Clients increasingly ask for them because they can be re-edited, and the assumption that a mix is final has weakened.

And it has made unauthorised voice cloning of singers dramatically easier, since a clean isolated vocal is exactly what a cloning system wants. This is the single largest driver of the artist-voice disputes covered in the next lessons, and it followed directly from a tool that was built for entirely benign reasons and released freely. That pattern — a useful capability whose main effect turns out to be somewhere its authors did not look — is worth remembering generally.

One practical tip for the uses that are uncomplicated. Separation quality improves markedly if you feed the model the best available source. A lossless file separates better than a compressed one, and a compressed file separates better than audio captured from a video platform, because each lossy step has already discarded exactly the fine detail the model uses to attribute sound to sources. If a stem comes back watery, check what you gave it before blaming the tool.

BEFORE YOU MOVE ON

A separated vocal stem from a commercial track sounds clean. What is the correct rights position?

A. It is a new recording created by the separation model, so a fresh copyright arises in the extracted stem.
B. It is a derivative of the original recording, so publishing it engages the same rights as sampling would.
C. It is fair use, because separation is a transformative technical process applied to the recording.
D. The vocal is a performance rather than a composition, so only the performer's consent is needed and not the label's.

62 · 8 min

Sound effects, foley and where it is already good

THE ONE THING TO KEEP

Generated sound effects are at production quality for many categories because a sound effect is short and unstructured, which is exactly the regime these models handle well.

The one place generation is unambiguously ready

Everything in this module so far has been qualified. Sound effects are the exception.

A sound effect is typically under five seconds, has no long-range structure, no polyphony to keep independent, and no requirement to be a specific recognisable thing. It is texture and envelope. That is precisely the regime where generative audio is strong.

The practical consequence: a text-to-sound model produces usable effects for a large fraction of what a video needs — impacts, whooshes, ambiences, mechanical noises, footsteps on surfaces, weather, crowd tone, interface sounds. In seconds, at any length, tuned by re-rolling.

Why this matters more than it sounds

Sound is half of video, and it is the half amateurs neglect. A well-shot clip with no sound design reads as amateur; an average clip with good sound reads as professional. This is not an opinion, it is a reliable finding of anyone who has run the comparison.

Historically, the barrier was libraries. A decent effects library cost money, and the free ones were small, badly labelled and inconsistent. Generation removes that barrier for a large category of sounds.

There is also a real advantage over libraries, and it is worth naming: an effect can be generated to length. A three-second whoosh that must be 2.4 seconds is a fight with a library file and a re-roll with a generator.

Where it still fails

Specific real objects. A particular model of car, a named machine, a specific bird. The model produces something in the category, and anyone who knows the object will hear that it is wrong.

Speech-adjacent sounds. Crowds where words are almost intelligible, laughter, a baby crying. These sit close enough to speech to trigger the model's speech machinery and come out uncanny.

Musical instruments in isolation. Better handled by a sample library or a real instrument.

Anything that must sync tightly to picture. Generation produces a sound, not a sound with a hit point at 1.32 seconds. You place it in the edit, which is normal practice anyway.

The free path, which is strong here

This is a category where free sources are genuinely excellent and deserve to be tried first:

- **Freesound** — a large community library, mostly Creative Commons, with licences stated per file. Check each one; they differ, and some require attribution.
- **The BBC sound effects archive**, available for personal, educational and research use under its own terms.
- Several national archives and museum collections with open licences.
- **Recording it yourself.** A phone in a quiet room records better foley than most people expect. Footsteps, cloth, doors, water, paper — all of it is available in your own house, it is free of any rights question, and it is more specific than anything generated.

That last one deserves emphasis because it is chronically undervalued. Foley artists have always made sounds with objects rather than recording the real thing, because the real thing usually sounds wrong. This craft is available to anyone with a phone and a quiet hour, and it produces the most convincing results of any option here.

The workflow

The professional shape is a layered one, and it applies whatever the sources:

1. **Ambience** — a continuous bed establishing the space. Generated ambiences are excellent for this.
2. **Spot effects** — the specific events. Generated, library, or recorded.
3. **Detail** — the small sounds that sell it: cloth, breath, a distant car. This layer is what separates good sound from adequate sound and it is almost always missing from amateur work.

Mix these in any free editor. The rule that matters more than the sources: quiet detail under a scene does more than loud effects on it.

One technique is worth knowing because it costs nothing and improves almost any sound: **layering**. A single generated impact rarely sounds convincing on

its own. The same impact plus a low thump underneath and a short bright crack on top does, because that is how real sounds are built — a body, a low end and a transient. Generate three variations, stack them, adjust the relative levels and trim the starts to line up. This one habit converts adequate generated effects into good ones and is the reason a sound designer's library of individually unremarkable files produces remarkable results.

The honest limitation: generated effects are unspecific by nature. For a documentary about a particular factory, generated machine noise is a fabrication, and using it raises the same disclosure question as any other synthetic element in factual work. For drama, atmosphere and explainer video, it is simply a tool, and a good one.

BEFORE YOU MOVE ON

Why are generated sound effects at production quality when generated music is not?

A. Sound effects are short and unstructured, which avoids the long-range coherence problem entirely.

B. Sound effects are monophonic, so the model does not have to keep several simultaneous voices consistent with each other across the duration of the sound.

C. Sound effects models are trained on much larger datasets than music models.

D. Sound effects are generated at a lower sample rate, so the model has fewer tokens to predict correctly.

63 · 9 min

Whose music it learned from

THE ONE THING TO KEEP

The music industry's dispute over training data is further advanced than in any other medium, and it has produced both litigation and licensing, which makes the shape of the disagreement unusually visible.

Why music moved first

Music has three features that made this the sharpest front.

Concentrated ownership. Three major labels control a large share of recorded music, which means a small number of parties can act together and afford to litigate.

Existing licensing machinery. The industry already has collecting societies, mechanical licences and sync licensing. There is an established answer to "how do you pay for using this", so "we could not have licensed it" is a weaker argument here than elsewhere.

Precedent for tight infringement standards. Music litigation has long treated short similarities seriously, which the next lesson covers.

What has actually happened

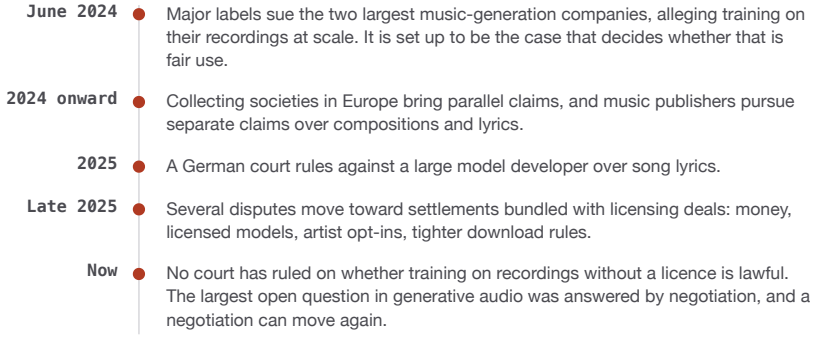
The pattern, described without picking a side.

Major labels brought infringement claims against leading music-generation companies in 2024, alleging that training on their recordings without licence was infringement at scale. Collecting societies in Europe brought parallel claims, and a German court ruled against a large model developer over song lyrics in 2025.

During 2025 several of those disputes moved toward licensing rather than judgment: label groups announced arrangements with generation companies covering training and, in some cases, artist participation. Other proceedings continued.

That combination — litigate, then license — is a familiar shape in music, and it happened before with sampling and with streaming. It is not a ruling on whether training is infringement. It is a commercial settlement of a question the courts did not have to reach, which leaves the legal position genuinely unresolved while the market moves on.

Litigate, then license: how the music dispute has gone



This shape is familiar in music; it happened with sampling and with streaming. It leaves you no precedent to rely on, which is why the practical advice is about licences, records and melody checks rather than about who was right.

What "licensed" means, and what to check

"Trained on licensed data" is now a marketing claim, and it covers several different things:

- **Fully licensed catalogue**, with rights cleared for training and for output, and payment flowing to rights holders.
- **Public domain and openly licensed material** only, which is a real and verifiable position and constrains the sound.
- **Licensed from a stock library** whose own contributor terms may or may not have permitted training — a gap that has caught several providers.

- **Licensed for training but silent about output**, which leaves you exposed on the thing you actually care about.

The questions worth asking a vendor: what exactly was licensed, from whom, does the licence cover the output as well as the training, do you indemnify commercial use, and what are the conditions on that indemnity. The final module covers what an indemnity is actually worth.

The artists' position, stated fairly

Musicians' objections are not all the same objection, and conflating them makes the argument unproductive.

Some object to **uncompensated use** of their recordings as training input. This is a licensing dispute and it is the one that settles.

Some object to **style imitation** — a model producing tracks in a recognisable manner. Style is not protected by copyright almost anywhere, so this is a moral objection with no legal instrument behind it, which is exactly why it feels unanswerable to those making it.

Some object to **voice cloning**, which is a personality-rights question rather than a copyright one, and which several jurisdictions have legislated on specifically.

Some object to **displacement** — that generated library music takes the work that paid the rent, regardless of any rights question.

These need different answers and only the first is on a path to one.

Where this leaves you

If you use generated music commercially, the practical position today:

- Read the licence on the model and on the output. Several capable open models are non-commercial only.
- Prefer providers who state their training sources and indemnify output, and read the conditions.

- Keep a record of what you generated, with what, and when. If a question arises in two years it will be about a specific track.
- For anything high-value or long-lived, consider licensed library music or a commissioned composer instead. The cost is not as different as people assume, and the rights position is settled.

The unresolved part, plainly: whether training on copyrighted recordings without licence is lawful has not been decided consistently anywhere, courts in different countries have begun to answer adjacent questions differently, and the commercial settlements have removed several of the cases that would have produced an answer. Anybody who tells you this is settled is describing their jurisdiction, their preference, or their sales position.

BEFORE YOU MOVE ON

Why does the wave of settlements between labels and music-generation companies leave the legal question unresolved?

A. Settlements are confidential, so their terms cannot be cited as precedent by other parties.

B. Settlements apply only to the specific recordings named in the claim, leaving the rest of each catalogue untouched.

C. Settlements bind only the parties and produce no judgment, so no court decided whether training was infringement.

D. Settlements were reached in Europe rather than the United States, so they have no bearing on the American doctrine that matters most.

64 · 8 min

When a tune becomes somebody else's

THE ONE THING TO KEEP

Music infringement can turn on a few seconds of melody, the standards are inconsistent and unpredictable, and "a model generated it" is not a defence anywhere.

Two rights in every song

Any recorded song carries at least two copyrights: the **composition** (melody, harmony, lyrics) and the **sound recording** (the specific captured performance).

A generated track cannot infringe a sound recording unless it actually reproduces the recorded audio, which is rare. It can infringe a composition, because a melody is exactly the kind of thing a model trained on melodies might reproduce.

The standard, and why it is unpredictable

Infringement of a composition generally requires access to the original and substantial similarity to it. Neither is measured in a way that produces reliable predictions.

There is no number of notes below which you are safe. Cases have turned on a handful of notes and been lost on longer passages. Prominent litigation of the last decade produced outcomes that are hard to reconcile: one high-profile case found infringement on the basis of feel and production style rather than a copied melodic line, another found no infringement on a widely-noticed similarity in a famous opening, and a substantial verdict in a third was overturned on the ground that the shared material was a commonplace musical building block.

The honest summary is that outcomes are inconsistent, expert testimony carries enormous weight, and juries respond to what sounds similar rather than to a technical analysis. This is not a criticism of the courts so much as an observation that music similarity resists formalisation.

Why a model might reproduce a melody

The memorisation mechanism from the first module applies here with unusual force. Popular songs appear in training data many times: the original, covers, live versions, remixes, karaoke tracks, tutorials, background in videos. That is exactly the duplication pattern that produces near-exact reproduction.

So the risk is not spread evenly across all music. It concentrates on the most famous melodies in the world, which are also the ones most likely to be recognised and most vigorously defended.

Some systems block prompts naming artists or songs. This reduces deliberate copying and does nothing about the accidental case, which is the one that will catch you.

What to actually do

Listen critically. Play the generated melody and ask whether it reminds you of anything. Your ear is a better first filter than any tool.

Ask other people. Melodic similarity is heard more reliably by a fresh listener than by the person who has been iterating on it for an hour.

Search it. Melody-search services and query-by-humming features in ordinary music apps will identify a close match. This takes a minute for a hook you are about to build a campaign on.

Change it if it is close. Move an interval, change the rhythm, alter the contour. Cheap now, expensive later.

Prefer generated material that has no strong hook for background use. Ambient and textural music carries almost no melodic infringement risk, because there is no melody to be substantially similar to. A great deal of what people need music for is exactly this.

Keep the record. Prompt, model, date, and the generated file. If a claim ever arrives, the ability to show what you did and when is worth a great deal.

The defence that does not exist

Say it plainly: "the model generated it" is not a defence in any jurisdiction. Liability for publishing an infringing work falls on the publisher. Whether the model's developer is also liable is a separate question that the courts are working through, and their answer will not remove your exposure for what you published.

Nor is a vendor's indemnity a substitute for care. Indemnities have conditions — you must have used the service as intended, not modified the output, notified promptly, allowed the vendor to control the defence — and the final module goes through what they typically cover and what they do not.

The unsettled part: what happens when a model produces a melody that is substantially similar to an existing one without any human intending it. The composition was not copied by a person, and it was learned from a corpus that included the original. Nobody has a settled answer to how the traditional access-and-similarity test applies to that, and the first cases to test it directly will be worth watching.

BEFORE YOU MOVE ON

A generated instrumental hook sounds strongly reminiscent of a well-known song. Which response is best supported by how the risk arises?

A. Publish it, since the vendor's terms assign all rights in the output to the user and therefore transfer any liability with them.

B. Publish it, since a melody generated by a model was not copied by any person and so cannot meet the access requirement.

C. Alter the melodic contour and rhythm, and check the result against a melody-search service before using it.

D. Keep it but credit the original song in the description, which converts the use into permitted quotation.

65 · 8 min

What platforms do about it

THE ONE THING TO KEEP

Distribution platforms enforce their own rules on generated music through matching systems and upload policies, and those rules bite faster and harder than any court would.

The rules that actually reach you

For most people, the operative constraint on generated music is not copyright law. It is the terms of the platform where the music will be heard, enforced automatically, at speed, with limited appeal.

Content matching on video platforms. An audio fingerprinting system compares uploads against a database of registered recordings. If it matches, the rights holder's chosen policy applies — monetisation redirected, the video blocked in some territories, or muted.

Two consequences specific to generated music. First, a generated track that is substantially similar to a registered recording can be matched, and you will be arguing with an automated system. Second, and stranger: if you upload your own generated track to a distribution service that registers it in the matching database, and someone else generates something similar, they will get claimed — and so, occasionally, will you, by your own registration through a different channel.

Streaming service policies. Services have moved from silence to explicit rules. Several now require disclosure of AI involvement in submitted metadata, remove tracks that impersonate an artist's voice, and act aggressively against bulk uploads of generated material used for streaming fraud. Reported figures during 2025 put the share of daily uploads to some services that were fully generated at a substantial and rising fraction, and the response has been filtering rather than acceptance.

Stock and library platforms have their own positions, ranging from outright prohibition to acceptance with disclosure. Read before submitting; a rejected batch is a wasted week.

Streaming fraud, and why it makes life harder for everyone

There is a large-scale abuse pattern worth understanding, because it drives the policies you will encounter.

Generate thousands of tracks. Upload them across many artist names. Stream them with automated accounts. Collect per-stream royalties from a pot that is divided among all rights holders. Prosecutions have followed, involving very large numbers of tracks and substantial sums.

The response has been tighter upload controls, identity checks, minimum thresholds and generated-content flags. If you are a legitimate small artist using generation as part of your process, you are inside the blast radius of a policy written for someone else. That is unfair and it is the situation.

The practical consequence: expect more friction, disclose accurately, and do not distribute in bulk. A hundred generated tracks uploaded in a week looks exactly like the fraud pattern regardless of your intentions.

Disclosure in metadata

Several distribution standards now carry fields for indicating AI involvement, at varying levels of detail — none, in composition, in performance, in production.

Fill them in accurately. Three reasons: some services require it and reject on discovery, a false declaration is a contractual breach that can end a distribution relationship, and the honest declaration costs you nothing while the dishonest one is a liability that never expires.

The practical checklist

Before distributing anything with generated music:

1. Check the model's licence permits commercial use and distribution.

2. Check what the platform requires disclosed, and disclose it.
3. Check the melody against a search service if there is a hook.
4. Keep the generation record.
5. If you registered the track with a matching system, know which system and under what identity, because you will need that when a claim arrives.

The honest position

Platform rules are moving faster than law, they differ between services, and they change without much notice. Anything written here about specific policies will be out of date; the structure will not be.

What to do when a claim lands

It will happen eventually, on something you have every right to use, because these systems produce false matches. The routine that works:

Do not delete the upload; a deletion removes your evidence and sometimes your appeal. Read the claim to see exactly which recording is asserted and at which timestamps. Compare that section against the asserted work yourself. If it is a genuine match, take the track down and replace it. If it is not, dispute with the specific facts — model, date, generation record, licence — rather than with an assertion that you made it yourself.

The reason the record-keeping advice keeps appearing in this course is this moment. A dispute is answered with dates and files, and a dispute answered without them usually fails regardless of the merits.

That structure is: automated matching decides first and asks later, disclosure obligations are real and enforced contractually, bulk generated uploads are treated as presumptively fraudulent, and appeals are slow. Plan for the automated system rather than for the legal position, because the automated system is what you will actually meet.

BEFORE YOU MOVE ON

Why is platform policy usually a more immediate constraint on generated music than copyright law?

A. Platform terms of service override national copyright law for any material distributed through their services, which is why they take precedence in practice.
B. Copyright law does not apply to generated works, leaving platform terms as the only applicable rule.
C. Platforms enforce automatically and immediately, while a legal claim is slow, expensive and rarely brought against a small user.
D. Platforms are legally required to remove generated content, so their action precedes any rights holder's decision to act.

66 · 8 min

When generated music is the right answer

THE ONE THING TO KEEP

Generated music competes with library music rather than with composition, so the honest choice depends on whether the music needs to be specific, and saying which is which protects both your work and your relationships.

The comparison nobody makes explicitly

The argument about generated music usually gets framed as machines against composers. That is not the real comparison for most projects.

The alternative to generated music, in the overwhelming majority of cases, is **library music** — pre-recorded tracks licensed for a fee, chosen from a catalogue, used unchanged. A composer writing to picture is a different and much more expensive product that most projects were never going to buy.

So the honest question is: does this project need music that is *specific to it*, or music that is *appropriate to it*?

Where generated music genuinely fits

- **Background beds** for explainer videos, tutorials, corporate material, podcasts. Nobody is listening to the music; it is there so the room is not silent.
- **Temp tracks** for editing, where the final music will be licensed or composed. This is a real workflow improvement: cutting to a temp track that fits exactly beats cutting to something approximate.
- **Loops and stings** for interfaces, transitions and short-form content.
- **Personal and hobby projects** with no budget, where the alternative is silence.
- **Rapid iteration** on tone before commissioning — generating six candidates to show a client what "warmer" means is far more useful than a conversation about adjectives.
- **Accessible music-making** for people learning, exploring or with disabilities that make instruments difficult. This is a real and under-discussed benefit and it deserves to be counted.

Where it does not fit

- **When the music is the point.** A film's score, a song, a piece somebody will listen to for its own sake.
- **When it must be memorable.** Generated hooks are weak precisely where memorability lives, and a brand theme that nobody remembers has failed at its only job.
- **When it must be defensible.** High-value, long-lived campaigns where an infringement claim would be expensive. Licensed library music has a clear chain and an indemnity behind it.
- **When there is a relationship to protect.** If you work with musicians, quietly replacing paid work with generation will be noticed, and the damage is not recoverable by explaining it well.

- **In documentary and journalism**, where fabricated sound has the same disclosure problem as any other fabricated element.

Saying it out loud

Two conversations worth having plainly.

With clients. "The bed is generated, the licence permits commercial use, here is what it cost." Some clients care and most do not. The ones who care would care far more about discovering it later, and being the person who said so first costs nothing.

With collaborators. If you brought in generated music on a project where a musician expected work, say so before they hear it. This is professional courtesy and it is also self-interest, because the industry is small.

The wider judgement

There is a real cost to the working musicians whose income came from library and commission work, and it is being felt now rather than in some hypothetical future. Pretending otherwise, or framing it purely as democratisation, is dishonest.

There is also a real benefit to the enormous number of people who could never afford any music at all, and for whom this is the difference between a piece of work existing and not existing. Dismissing that as unimportant is dishonest in the other direction.

Both are true and no framing dissolves the tension. What an individual can do is narrow and worth doing anyway: use generation where the music does not need to be specific, pay a person where it does, disclose either way, and do not pretend the second category is empty because the first one is cheap.

That is not a rule anyone can enforce on you. It is a standard you can hold, and holding it is what makes the difference between a practice you can explain and one you have to avoid explaining.

There is a third route that gets forgotten between the two extremes, and it is often the best answer. A great deal of excellent music is already available un-

der Creative Commons licences and in the public domain — community archives, national collections, artists who release deliberately, and every recording old enough to be out of copyright. It costs nothing, the licence is explicit, attribution is usually the only condition, and a real musician made it. For a project with no budget that still wants music somebody cared about, this beats both generation and silence, and it takes twenty minutes to search properly.

BEFORE YOU MOVE ON

What is the most useful question when deciding whether generated music suits a project?

A. Whether the budget can support a commissioned composer, since that determines what is actually available.

B. Whether the platform requires disclosure of AI involvement in the audio track.

C. Whether the audience is likely to notice that the music was generated rather than recorded.

D. Whether the music needs to be specific to this project or merely appropriate to it.

CASE STUDY · MODULE 7

The theme that sounded like something

A Malayalam cooking channel in Kochi, four days before a twelve-episode season goes live

Ruchi Vibes has three hundred and forty thousand subscribers and a house style: a home kitchen, one cook, no studio lighting. Twelve episodes of the new season were shot, cut and graded, with a launch on Monday.

The title theme was eight seconds under the card, with a longer version used as a bed under the recipe segments. Jomon, a musician in Thrippunithura who had written the previous two seasons, quoted twenty-eight thousand rupees for a theme and three stings. The producer, Deepa Menon, said no, and the editor, Fahad, generated a theme in an afternoon from a text description.

It was good. It was catchy in the specific way a title theme has to be, which is that you can hum it after hearing it twice.

On Thursday, four days out, Deepa's mother heard it from the next room and said it was from a film. She could not name the film. Fahad said he had never heard it before in his life, which was true and not relevant. Two people in the office said it reminded them of something and two said it did not.

The decision had to be made that day, because Fahad was colour-grading for the rest of the week.

Launching meant going out with a hook that four people had split evenly on. If it turned out to be close to an existing melody, the consequence would not arrive as a lawyer's letter. It would arrive as an automated content match on the video platform, on every episode at once, with the revenue redirected to whoever registered the recording while the channel argued with a queue.

Changing it meant re-rendering twelve episode openers and thirty-six segment transitions, about a day and a half of the only editor's time in a week that had no day and a half in it, plus either paying a rush rate to Jomon with four days' notice or generating something else and having the identical conversation about that.

Fahad's position, argued with some heat, was that a model had made the melody, that he had never knowingly heard the film song anybody was vaguely gesturing at, and that a generated tune cannot be somebody else's tune because nobody wrote it.

Deepa did not know enough to answer that properly and she knew she did not. What she did know, from a year of running a channel on a platform, was that the question of who wrote it and the question of what would happen on Monday were not the same question and might not have the same answer.

There is a reason the risk concentrates on exactly this eight seconds. Popular film melodies appear in training data many times over — the original recording, covers, live versions, karaoke tracks, reaction videos, tutorials, and hundreds of hours of other people's content with the song playing underneath. That is the duplication pattern that produces something close to reproduction rather than something merely in the same idiom. A hook is also short, which is the length at which music infringement is actually litigated, and memorable, which is the property a title theme is chosen for. The channel had asked for the most dangerous possible thing and had been delighted when it arrived.

There was a third cost that did not appear on any spreadsheet. Jomon had written the last two seasons and had not been told the channel was generating this one. He would hear the theme on Monday like everybody else, in a city where the people who do this work all know each other.

STOP HERE

Before turning the page: what would you have done, and what would it have cost if you were wrong? Write it down. The point of the next section is to compare.

WHAT HAPPENED

Deepa spent an hour on it and did three separate things.

She hummed the hook into the query-by-humming feature of an ordinary music app. It returned nothing conclusive, which she correctly treated as no evidence rather than as a clearance.

She sent the eight seconds to four people who had not heard it, with no context beyond asking what it reminded them of. Three named the same 2007 film song. When she found that song, the first four notes of the phrase and its rhythm lined up. Not the whole melody, and enough.

Then she checked the thing that actually decides what happens to a channel. The film's label had the recording registered in the platform's audio matching system. A sufficiently close match on twelve simultaneous uploads would be handled by software, applied immediately, and appealed into a queue, and the season's opening week is where a season's revenue is.

She changed it.

Fahad generated new material for the segment beds — ambient and textural, no hook, nothing anyone could hum — because that is the regime where generated audio is genuinely strong and where there is no melody to be substantially similar to anything. Those took an afternoon and carried no risk worth the word.

For the eight-second title theme she called Jomon and paid him twelve thousand rupees for the theme alone, without the stings. He wrote it in two days. She told him, unprompted, that the beds were generated, what the licence permitted and what it had cost. He said he would much rather be told than find out, and that library music had been taking that work from him for fifteen years before any of this.

The openers re-rendered overnight. The launch did not move.

What Deepa took from it, and now applies as a rule, is a distinction rather than a position. Where the music has to be memorable, it is worth paying a person, because memorability is exactly where generated music is weakest and where infringement risk is concentrated. Where the music only has to be appropriate, generation is competing with library music rather than with a composer, and it wins on cost and on fitting the cut exactly.

Why did Deepa treat the eight-second title theme as risky and the segment beds as not, when both were generated by the same tool on the same afternoon?

Because infringement in music turns on the composition, and a composition claim needs a melody to be substantially similar to. A hook is a melodic figure short enough and distinctive enough to be compared with another one, and popular hooks are exactly the material that appears many times over in training data — the original, covers, karaoke versions, background in other videos — which is the duplication pattern that produces near-reproduction. Ambient, textural material has no melodic line to compare, so the risk is close to nil. The risk is not evenly spread across generated audio; it concentrates precisely where memorability lives.

Fahad had genuinely never heard the 2007 song. Why does that not help?

Because liability for publishing an infringing work falls on the publisher, and the usual defence of independent creation is hard to run when the model was trained on the catalogue containing the original. 'A model generated it' is not a defence in any jurisdiction. Access and substantial similarity are assessed on the output and the training corpus, not on what the operator personally remembers hearing. Fahad's innocence is real and it is about his state of mind, which is not the question a claim asks.

She checked the platform's matching position separately from the question of whether it actually infringed. Why is that a distinct question and the more urgent one?

Because the platform's system decides first and asks later. Audio fingerprint matching is automatic, applies the rights holder's chosen policy immediately across every upload at once, and is appealed into a queue rather than to a person. A track can be matched without a court ever agreeing it infringes, and a track can infringe without ever being matched. For a channel whose revenue is concentrated in a launch week, the operative constraint is the contractual and automated layer, not the legal one, and it bites far faster.

MODULE 7 · TEST

Check what stuck

6 questions, one right answer each. This one is for you, not for a record: answers and the reason behind each are at the back. If two questions from the same module go wrong, re-read that module before the exam rather than after.

1 You want to know the best audio quality a music model can ever produce, before judging any of its output. What do you do?

A. Generate the same prompt twenty times and keep the best result.

B. Run a real track through the codec's encoder and straight back out with no generation.

C. Generate at the highest available sample rate and compare with a reference recording.

D. Compare a fast preview render against a full-quality render of the same prompt.

2 Speech synthesis became convincing before music did, even though speech carries meaning and music does not. What best explains the order?

A. Speech has a narrower frequency range, so the codec reconstructs it more accurately.

B. There is far more transcribed speech than licensed music available for training.

C. Speech is one voice and its structure is supplied by the text you provide.

D. Music models are trained on shorter clips, so they never learn long-form behaviour.

3 You separate a commercial recording into stems with free software and use the isolated vocal in a track you publish. Which statement is accurate?

A. Separation creates a new work, so the stem carries its own copyright.
B. The stem is a derivative of the recording and the original rights all still apply.
C. Automatic extraction falls under the same exception as private study, so publication is permitted.
D. Only the recording copyright is engaged, because the composition was not copied.

4 Generated sound effects are production-ready for many categories while generated songs are not. What property of a sound effect accounts for that?

A. Sound effects are monophonic, so the codec has fewer sources to reconstruct.
B. Sound effects appear more often in training data than complete songs do.
C. Sound effects are short, unstructured texture and envelope, with no long-range form to hold.
D. Sound effects are generated at a lower sample rate, which hides the codec's artefacts.

5 Major labels sued the largest music-generation companies in 2024 over training data, and by late 2025 the disputes were moving toward licensing arrangements. What does that leave you with?

A. No precedent, because a commercial settlement is not a ruling on whether training was lawful.
B. A clear rule that training on recordings requires a licence, now established by agreement.
C. A finding that training is fair use, since the claims were not pressed to judgment.
D. An answer that applies to recordings but not to compositions or lyrics.

6 You publish twelve videos carrying a generated theme that turns out to resemble a registered recording. What happens first, in practice?

A. A rights holder's lawyer writes to you asserting infringement.

B. The platform requests a licence declaration before the videos go live.

C. An automated fingerprint match applies the rights holder's chosen policy to all twelve.

D. Nothing, until a viewer reports the similarity to the platform.

MODULE 8

Provenance, detection and disclosure

If you cannot tell by looking, the question becomes what can be established and how. This block covers why detection is structurally weak, what a watermark can and cannot survive, how cryptographic content credentials work and where the chain breaks, how to verify a claim rather than an image, the second-order harm of universal doubt, how to write a disclosure that actually informs somebody, what the platforms currently require, and a verification routine borrowed from newsrooms that anybody can run.

> BY THE END OF THIS MODULE YOU CAN
>
> Assess an image or clip of unknown origin using provenance, source and corroboration rather than inspection, explain why a detector's positive result is a weak signal at realistic base rates, and write a disclosure that tells a reader what was synthetic and what was not

67 · 8 min

Why detection is the wrong hope

THE ONE THING TO KEEP

Detectors learn the fingerprints of the generators they were trained on, so they degrade on new models, on compressed files and at realistic base rates, which makes their output a signal to investigate rather than a verdict.

Four independent reasons, any one of which would be enough

Generalisation. A detector is a classifier trained on real and generated examples. It learns what the generators in its training set leave behind — characteristic frequency signatures, upsampling traces, statistical regularities in noise. Those fingerprints are specific to an architecture and to a version. A model released after the detector produces different traces, and measured performance on unseen generators falls sharply, sometimes to near chance. The detector is always chasing.

Fragility. The evidence lives in fine detail. Re-encoding a JPEG, resizing, cropping, applying a filter, uploading to a platform that recompresses, screenshotting, or photographing a screen — each of these removes a substantial part of the signal. By the time an image has passed through two social platforms, most detectors are guessing.

Adversarial pressure. A public detector is a target. Anyone who wants to evade it can test against it and adjust until it passes. This is not hypothetical; it is a standard research result across every detection domain.

Base rates. This is the one that matters operationally and the one people skip.

The arithmetic worth doing once

Suppose a detector is 95% accurate in both directions, which is better than most are in the field. Suppose one image in a thousand in some feed is generated.

Out of 100,000 images: 100 are generated, of which the detector flags 95. The other 99,900 are real, and it wrongly flags 5% of them — about 4,995.

So of roughly 5,090 flagged images, about 95 are actually generated. Fewer than two in a hundred flags are correct.

A 95% detector on 100,000 images, one in a thousand generated

	Detector flags it	Detector clears it
ıted (100 of them)	**95** the correct catch	**5** missed
l (99,900 of them)	**4,995** a false accusation	**94,905** correct

About 5,090 images are flagged and 95 of them are generated — fewer than two flags in a hundred are right. Nothing is wrong with the detector; the rarity of the thing swamps a good error rate. A flag is a reason to look, never a reason to accuse.

Nothing is wrong with the detector. The rarity of the thing being detected swamps a good error rate. Any screening system for a rare event has this property, and it is why medical screening programmes are designed around confirmatory testing rather than around the screen alone.

The consequence is that a positive result from a detector is a reason to look more carefully, and never a reason to accuse.

The precedent worth knowing

Text detection went through this cycle publicly and quickly. A widely-publicised detector for AI-written text was withdrawn by its own developer within seven months, on the stated grounds of low accuracy — it correctly identified a minority of generated text while wrongly flagging some genuine human writing.

Meanwhile students were being accused of cheating on detector output, and writing in a plain, structured style — which is what non-native English writers and careful technical writers produce — was flagged disproportionately. The harm landed on people with the least ability to contest it.

Image and video detection is subject to the same forces with worse fragility, because pixels get recompressed far more aggressively than text does.

What detectors are actually good for

They are not useless, and it is worth being precise about the fit.

Triage inside a system with a second step. Flag for human review, route to a slower path, request additional verification. The false positives then cost a check rather than an accusation.

Aggregate measurement. "Roughly what proportion of uploads this month were generated" tolerates a poor per-item error rate because the errors partly cancel.

Known-generator settings. A closed platform checking its own outputs against its own detector, where the generator is known and the file is unmodified.

What they are not good for: deciding whether the picture in front of you is real, in public, about a named person.

The direction that works instead

Detection asks "does this look generated". Provenance asks "where did this come from". The second question has an answer that improves as the generators improve, because it depends on the record rather than on the artefact.

That is the subject of the next three lessons. The important shift is one of framing: stop trying to prove a negative about the pixels and start establishing a positive about the origin.

One consequence is worth naming for anyone in a position of authority over other people — a teacher, an editor, a manager, a moderator. Do not act on a

detector's output alone against an individual. Not because the tool is worthless, but because the arithmetic above means most of the people you flag will be innocent, and the cost falls entirely on them while the cost of your error falls on nobody. If a flag is the beginning of a conversation in which the person can show their working, the tool is doing something useful. If it is the end of the conversation, it is producing false accusations at a rate you would never accept if the number were printed on the screen.

BEFORE YOU MOVE ON

Why is a 95%-accurate detector unhelpful for flagging generated images in a feed where one image in a thousand is generated?

A. Because 95% accuracy is measured on clean files and real feeds contain compressed ones.

B. Because the detector was trained on generators that predate the images it now sees.

C. Because accuracy is the wrong metric and precision should be reported instead of accuracy in every detection context.

D. Because the rare positive class is swamped by false positives, so most flags are wrong.

68 · 8 min

What a watermark can survive

THE ONE THING TO KEEP

An invisible watermark embeds a signal robust to ordinary handling, which makes it a useful record of compliant generation and not a defence against anyone who does not want to be marked.

Two kinds, and only one is interesting

Visible watermarks — a logo, a label burned into the corner. Honest, immediately informative, and removed by a crop.

Invisible watermarks embed a signal into the content itself, distributed across the image or audio in a way designed to survive handling. The best-known implementations from large labs mark generated images, audio, video and text, and provide a detector that reads the mark back.

The interesting question about any of these is not whether they work, but what they survive.

Where the signal is put

For images, a watermark is embedded in a representation robust to common operations — frequency components rather than raw pixels, or, in the strongest designs, applied during generation itself so it is woven through the content rather than added afterwards.

Typical published claims: survives JPEG compression, resizing, cropping, colour adjustment, moderate filtering, and screenshots. This is genuinely impressive engineering and it covers the ordinary path an image takes through the world.

For audio, the mark is placed in a perceptually masked part of the spectrum and survives compression and re-encoding. For text, the mark is a statistical

bias in token selection, which survives light editing and disappears under paraphrase.

What removes it

Being honest about the limits, because vendors state the robustness and rarely state the ceiling:

- **Heavy transformation.** Strong crops, large rotations, aggressive re-compression, significant resizing.
- **Regeneration.** Passing the image through another generative model — an image-to-image pass at moderate strength, or an upscaler — substantially disrupts embedded signals. Published research has demonstrated this against several watermarking schemes.
- **Deliberate attack.** There is an active literature on watermark removal, and it works.
- **Not being there.** The decisive one. Open-weight models run locally produce no watermark at all, and anyone intending to deceive will use one.

What it is for, then

That last point makes the whole thing sound pointless, and it is not. The right way to think about a watermark is as **a record of compliant generation, not a barrier to malicious generation**.

Its real uses:

Platform labelling at scale. A platform can detect the mark and label the post automatically, which handles the enormous volume of ordinary, non-malicious generated content that nobody would otherwise label.

Keeping training data clean. Model developers use it to avoid training on their own output, which matters more than it sounds — a corpus increasingly filled with generated material degrades the next generation of models, and marking is the only practical way to filter it.

Internal accountability. An organisation can tell whether an asset came from its own approved pipeline.

Establishing intent. Removing a watermark is a deliberate act, and in some jurisdictions is itself becoming an offence. A stripped mark is evidence of something.

None of those is "proving an image is fake", and a vendor who implies otherwise is overselling.

The absence problem, stated clearly

The most important property of every watermarking scheme: **the absence of a watermark proves nothing**. Most images in the world are unmarked, including nearly all real photographs and all output from open models.

So a watermark detector can say "this came from that system" with reasonable confidence and can never say "this is real". This asymmetry is the whole design, and it means watermarking and provenance are complements rather than alternatives: one marks generated content where the generator cooperates, the other attests to captured content where the camera cooperates.

There is a practical point for anyone publishing generated work. If your tool embeds a mark, leave it in. Removing it gains you nothing legitimate, and in an increasing number of places it is becoming an offence in its own right, quite separate from anything about the content. If you are asked by a client to strip one, that request is worth a conversation rather than a keystroke, because the only reason to want it gone is to make the origin harder to establish.

The unsettled part: whether marking should be mandatory. Several jurisdictions are legislating in that direction, which the final module covers. The case for is that it makes automatic labelling possible at scale. The case against is that a mandate binds only compliant actors while imposing a cost on everyone, and that a mark identifying which system produced a file is itself a privacy consideration for people who make things they do not want traced. Both arguments are serious.

BEFORE YOU MOVE ON

What does the absence of a watermark on an image tell you?

A. Effectively nothing, since most images carry no watermark, including nearly all real photographs.
B. That the image was probably captured rather than generated, since the major generators now mark their output by default as a matter of policy.
C. That the image has been edited, since editing is the most common cause of a watermark being lost.
D. That the watermark detector used is incompatible with the scheme that produced the file.

69 · 9 min

Signed provenance, and where the chain breaks

THE ONE THING TO KEEP

Content Credentials attach a cryptographically signed record of how a file was made and edited, which establishes origin rather than detecting fakery, and the chain breaks wherever a tool or platform does not carry it.

The idea

Instead of examining a file for signs of generation, attach a signed statement about where it came from and what was done to it.

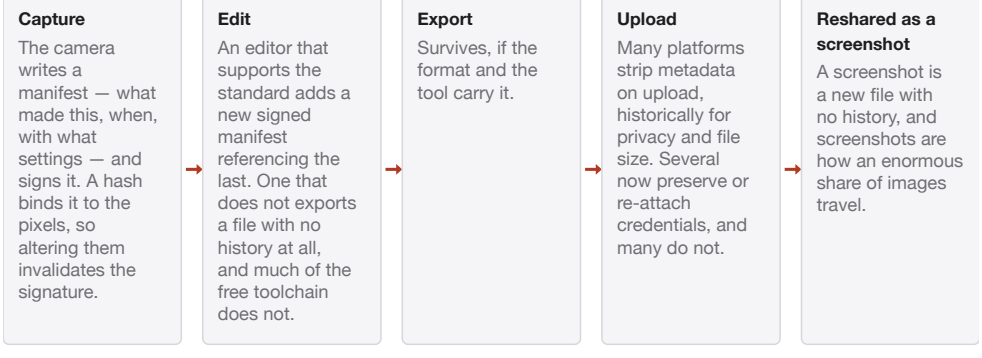
The main standard for this is the C2PA specification, whose user-facing name is **Content Credentials**. It was built by a coalition of camera manufacturers, software companies, news organisations and chip makers, and it is moving into formal international standardisation.

The mechanism is ordinary public-key cryptography applied to a manifest:

1. A capture device or a piece of software creates a **manifest** — a structured record: what made this, when, with what settings, and what actions were applied.
2. The manifest is **signed** with a certificate belonging to the device or software vendor.
3. The manifest travels with the file, and a hash binds it to the content, so altering the pixels invalidates the signature.
4. An editor that supports the standard adds a new signed manifest referencing the previous one, building a chain.

Anyone can then verify the chain: this was captured by that model of camera at that time, opened in that editor, cropped and colour-corrected, and exported. Or: this was generated by that model, then edited.

A signed chain, and the places it breaks



Provenance attests to a process and never to truth: a manifest saying a camera captured this says nothing about whether the scene was staged. And absence proves nothing, which today describes almost every file you will meet.

What it does and does not claim

This is the part people get wrong in both directions.

It attests to a process, not to truth. A signed manifest saying "captured by a camera" does not mean the scene was not staged, that the caption is accurate, or that the photograph was not taken somewhere else entirely. Provenance answers where a file came from and nothing about what it means.

It does not detect anything. There is no analysis of the pixels. If a file has no credentials, the system has nothing to say, which is the same absence prob-

lem as watermarking.

A valid signature is only as good as the certificate. Trust flows from a list of accepted signers. A compromised or improperly issued certificate signs false manifests that verify correctly.

Where the chain breaks in practice

This is the honest part, and it is where the technology currently stands.

Most software does not participate. Open a credentialed file in an editor without support and export it, and the credentials are gone. Much of the free toolchain this course recommends does not yet carry them.

Most platforms strip metadata on upload, historically for privacy and file size. Several major platforms now preserve or re-attach credentials, and many do not.

Screenshots break everything. A screenshot is a new file with no history, and screenshots are how an enormous share of images travel.

Cameras are only starting to ship it. Support arrived first in a small number of professional bodies and by firmware update on some others. The overwhelming majority of photographs are taken on phones, where deployment is partial and recent.

The consequence: today, a file with valid credentials tells you something useful, and a file without them tells you nothing, which describes almost everything you will encounter.

Why it is still the right direction

Two reasons it deserves support despite the gaps.

It improves as generators improve. Detection gets harder every year. A signature does not care how good the generator is.

It handles the harder half of the problem. The urgent question is increasingly not "is this fake" but "is this real" — a genuine recording that some-

body wants to dismiss. Only provenance can answer that, and no detector ever will.

What to do now

If you make things: turn credentials on where your tools support them. Recording that a piece is generated, and what was done to it, is the same discipline as keeping the layered file, and it costs a setting.

If you assess things: check for credentials as a first step, treat their absence as uninformative, and remember that a valid manifest tells you about a process rather than about a claim.

There is a privacy consideration that deserves stating, because it is the strongest objection to the whole approach. A provenance record can carry device identifiers, timestamps and edit history, and a photographer working somewhere dangerous may need none of that attached to their file. The specification allows redaction and selective disclosure for exactly this reason, and whether those controls are used well in practice is a live question. Anyone advocating provenance as a default should be able to answer what it means for a source photographing a protest, and "you can turn it off" is only an answer if turning it off does not itself become evidence of something.

And keep expectations calibrated. This is infrastructure being built while it is needed, with real gaps between the specification and what actually happens to a file on its way through the world. Saying so is not scepticism about the project. It is the difference between using it correctly and trusting it wrongly.

BEFORE YOU MOVE ON

A news photograph carries valid Content Credentials showing capture by a specific camera at a specific time. What has been established?

A. That the file came from that device and has not been altered since.

B. That the photograph is an accurate and unstaged record of the event described in its caption.

C. That the image contains no generated elements, since a generative edit would have invalidated the signature.

D. That the photographer holds copyright in the image, since the signing certificate is issued to the rights holder.

70 · 8 min

Verify the claim, not the picture

THE ONE THING TO KEEP

Almost every question about a suspicious image is answered by finding its earliest appearance and its source, which is free, fast and unaffected by how good the generator was.

The reframe

The instinct on seeing a doubtful image is to study it. That is the least productive available action, for every reason the earlier lessons gave.

The productive question is not "is this generated" but "what is being claimed, and what supports it". Almost always the claim is bundled with the image — that this shows a specific event, at a specific place, involving specific people. Those are verifiable independently of the pixels.

This is not new technique. It is what newsrooms have done with user-submitted photographs for fifteen years, and it survived every previous improvement

in image manipulation.

The routine

Find the earliest appearance. Reverse image search across several services, because they index differently — Google Lens, TinEye, Yandex and Bing all return different results, and all are free. TinEye in particular sorts by oldest, which is exactly what you want. If the image appears three years before the event it supposedly shows, you are finished.

Identify the source account. How old is it, what has it posted before, does it have a name attached, does anybody with a reputation stand behind the claim. An image with no source is unverified regardless of how it looks.

Check the place. If the image claims a location, compare against satellite and street-level imagery. Building shapes, road layouts, signage, the side of the road traffic drives on, the language on shopfronts, vegetation. This is slower and it is decisive, because a generated or misattributed image rarely matches a real place in detail.

Check the time. Shadows give a rough time of day and, with a known location, a rough date. Weather records are public. If the image shows heavy rain and the archive says clear skies, that is a fact rather than an impression.

Look for corroboration. Real events of consequence generate multiple independent recordings from different angles. A single image with no other coverage is a reason for caution, whether or not it is generated.

Read the metadata, knowing it proves little. Its presence can be informative; its absence is normal.

Check provenance for credentials, knowing most files have none.

Why this beats inspection every time

Every step above depends on facts outside the file. The generator's quality does not affect a satellite photograph, an archived weather record, or the date an image was first indexed.

That is what makes the approach durable. Inspection-based methods expire with each model release. This one does not.

The discipline of not concluding

The hardest part is being willing to end at "unverified".

There are three possible outcomes, and people reliably collapse them into two. Verified: origin and claim check out. Debunked: found earlier elsewhere, or the place is wrong. **Unverified**: nothing established either way, which is the most common result and a legitimate place to stop.

An unverified image should not be shared as true and should not be denounced as fake. Both mistakes are common, and the second one is now doing serious damage to people whose genuine work gets dismissed.

Practical habits

Slow down at the moment of highest emotion. The images engineered to spread are the ones that make you angry. That reaction is the trigger to check, not to share.

Screenshot before checking, keeping the original file and its address. Content disappears while you are looking at it.

Search the claim in words, not just the image. If something significant happened, credible outlets will have it; if only one account has it, that is informative.

Write down what you established and what you did not. For anything you publish, this record is both your defence and your reason for confidence.

A note on what to do with an old image being reused. This is the single most common form of misleading media and it involves no generation at all: a real photograph of a real event, presented as a different event. Reverse search catches it in seconds, and it accounts for a large share of what circulates during any crisis. Anyone who learns only the first three steps of this routine will catch far more misinformation than someone who becomes expert at spotting generated pixels, because the recycled photograph is both more common and

more persuasive — it survives every kind of inspection, since nothing about it is fake except the caption.

The limitation of this method, stated plainly: it works on images making a factual claim about the world, which is where the harm concentrates. It does nothing for an image where the claim is about a private person and there is no public record to check against — non-consensual imagery, harassment, fabricated evidence in a personal dispute. Those cases need law and platform action, not verification technique, and they are where the current gap is widest.

BEFORE YOU MOVE ON

Why does source-and-claim verification remain effective as generators improve, when visual inspection does not?

A. Because generators cannot yet reproduce the metadata patterns that verification tools examine.

B. Because verification tools are updated more frequently than detection models are.

C. Because it relies on facts outside the file — earliest appearance, geography, weather, corroboration — which model quality does not affect.

D. Because reverse image search indexes every generated image at the moment it is created, allowing any generated file to be traced to its origin.

71 · 8 min

When everything can be denied

THE ONE THING TO KEEP

Universal awareness that media can be fabricated lets genuine evidence be dismissed, so the second-order harm arrives without anyone needing to make a fake at all.

The shape of the problem

The first-order harm of synthetic media is obvious: false things are believed.

The second-order harm is that true things can be disbelieved. Someone caught on camera can now say the recording was fabricated, and enough of the audience will accept it. This has a name in the academic literature — the **liar's dividend** — and it was described before the technology was good, precisely because the mechanism does not depend on the technology being good. It depends only on the audience knowing that fabrication is possible.

The dividend is paid to whoever benefits from doubt. It costs nothing, requires no technical capability, and cannot be countered by improving detection, because the claim is not that a specific analysis is wrong. It is that nothing can be established at all.

Documented and predictable consequences

Politicians and public figures have dismissed authentic recordings as AI-generated. Defendants have challenged audio and video evidence on the basis that fabrication is possible, and courts in several jurisdictions have begun to work out how to handle that argument.

Journalists report that sources are more reluctant to provide recordings, since a recording no longer settles anything. Fact-checking organisations report

that debunking one item now prompts a general "you cannot trust anything" response rather than acceptance of the correction.

The corrosive part is not that people believe fakes. It is that the shared reference point disappears, and disputes that used to end with a recording now do not end.

Why this is harder than the first problem

Fabrication has technical countermeasures, however imperfect: provenance, watermarking, verification routines. The dividend has none, because it is a claim about epistemology rather than about a file.

Worse, the countermeasures partly feed it. Every article explaining how convincing generated video has become is also an article establishing that any video might be generated. This is a genuine dilemma for anyone doing public education on the subject, this course included, and there is no clean way out of it. The honest response is to pair every statement about capability with a statement about what can still be established — which is what the previous lesson is for.

What actually helps

Provenance at capture, again. It is the only mechanism that supports a positive claim of authenticity, which is precisely what the dividend attacks. A recording that can be shown to have come from a specific device at a specific time is much harder to wave away.

Institutional custody. Evidence handled through a documented chain — police, courts, established newsrooms — resists dismissal in a way an anonymous upload does not. This is not new technology; it is the reason chains of custody exist.

Multiple independent sources. Three recordings from three unconnected people are far harder to dismiss than one, and the argument required to do so becomes visibly absurd.

Naming the move. When someone dismisses evidence as fabricated without giving a reason, saying so plainly is worth something: "that is an assertion, not an analysis; what specifically indicates fabrication". Making the burden visible does not always work and it costs nothing.

Calibrated public language. Reporting that says what was established and how, rather than "experts say it is real", gives an audience something to hold. Vague authority is exactly what the dividend erodes.

The honest position

Nobody knows how this resolves. One view holds that societies absorbed photographic manipulation, staged photographs and edited audio before, and that new norms and institutions will form again — slower, more mediated, and functional. Another holds that the volume, speed and personalisation are different in kind, and that the shared factual ground genuinely does not reconstitute.

Both are held by serious people. The evidence available is a few years old and mostly about elections and public figures, which is a narrow slice of where the harm lands.

There is also a personal version of this that is easier to act on than the civic one. If you record something that might matter — a conversation with a landlord, an incident at work, evidence of a repair that was never done — the dividend applies to you. What protects that recording is the same set of things: capture it on a device that records provenance if you have one, keep the original file rather than a re-shared copy, note the time and place immediately, and tell someone contemporaneously. None of that is technical and all of it is what makes a recording hold up when somebody says it was fabricated.

What can be said with confidence is narrow: the dividend is real, it is being used, it does not require any fake to exist, and the only countermeasures that touch it are about establishing origin rather than about detecting fabrication. That is a reason to support provenance infrastructure even while being clear-eyed about how partial it currently is.

BEFORE YOU MOVE ON

Why can improved detection technology not counter the liar's dividend?

A. Because detectors are proprietary, so their results cannot be independently verified by the public.

B. Because detection only works on recent generators, and a denial can always cite a newer one.

C. Because the denial asserts that nothing can be established, rather than disputing any particular analysis.

D. Because detectors report probabilities rather than certainties, which a denial can characterise as inconclusive.

72 · 8 min

Writing a disclosure somebody can use

THE ONE THING TO KEEP

A useful disclosure says what was synthetic and what was not, at the point of viewing, in the viewer's terms, which is a different thing from a legal label that protects the publisher.

Two kinds of label

A compliance label exists to satisfy a rule. "This content may contain AI-generated material." It is placed where a rule requires, worded to cover everything, and tells a reader nothing.

An informative disclosure tells somebody what they are looking at. "The presenter's voice is synthetic; the footage is real. The background of the third image was extended using generative fill."

Only the second changes what a viewer understands, and it costs about the same to write.

What makes a disclosure useful

Specific about what. "AI was used" covers a spectrum from spellchecking to fabricating the entire subject. Say which element.

Specific about what was not. Frequently the more valuable half. In factual work, "the photographs are unaltered; the diagrams are generated" tells a reader exactly how much weight to put on each.

Placed where the content is. A disclosure in the terms of service is not a disclosure. It belongs adjacent to the thing, visible at the moment of viewing. For video, in the video or in a persistent overlay, because descriptions do not travel when a clip is reshared.

Written for the reader. Not "generative adversarial synthesis was employed in post-production" but "this voice is a computer, not the presenter".

Proportionate. A stock background image on a corporate page does not need a paragraph. A synthetic voice in a news report does. The test is whether a reasonable person would change how they interpret the piece if they knew.

A usable rule

The threshold worth adopting, because it is both defensible and easy to apply:

> *Disclose when the synthetic element affects how the audience should interpret what they are seeing.*

Under this rule:

- **Disclose:** a synthetic voice presented as a person; a photograph of an event that did not occur; a person depicted doing something they did not do; a testimonial from someone who does not exist; sound design in a documentary presented as recorded on location.
- **Usually do not need to:** removing a distracting object from a background; upscaling; colour correction; generated texture in an obviously illustrative graphic; a stock-style illustration used decoratively.
- **Judgement:** an illustration of a real event; a composite of several real photographs; generated background music in a factual piece.

Notice this is the standard journalism already used for photographic manipulation. Dodging and burning was always acceptable; adding or removing a person was not. The line is whether the change alters the meaning.

Where over-labelling does harm

There is a failure mode in the other direction worth naming, because good-faith organisations are falling into it.

Labelling everything trains the audience to ignore the label. If a "may contain AI" notice appears on every page, it carries no information and the genuinely important disclosure is invisible inside the noise.

Some platforms' automatic labelling has already produced this: photographs that were merely opened in an editor with generative features have been labelled as AI, which is technically defensible and misleads the reader about what happened. The photographers affected reasonably objected, because the label said something untrue about their work.

Accuracy in both directions is the standard. A label that overstates is a false statement about your own work.

The practical shape

For most working people this comes down to two habits.

Keep the record — model, prompt, what was generated, what was captured, what was done by hand. The layered file and the note from earlier modules.

Write two sentences at delivery. What is synthetic, what is not. Put them where they will be seen.

That is the whole practice. It takes a minute, it is honest, it becomes an ordinary professional habit rather than an admission, and it is far cheaper than the conversation that happens when a client discovers it later.

The tone question

A disclosure is often written apologetically, and it does not need to be. There is a difference between confessing and informing. "Some elements of this were created with AI" reads as a hedge; "the landscape is a photograph, the creature is generated" reads as a caption.

The second is better writing and it is also better positioning. Work that states plainly what it is invites judgement on what it is, which is the position you want to be in. Work that hedges invites the reader to wonder what else is being hedged. Museums, science publishers and good documentary producers have handled this for a long time with reconstructions and illustrations, and their convention is worth copying: name the thing, say how it was made, and move on.

BEFORE YOU MOVE ON

Which disclosure best meets the standard of informing rather than merely complying?

A. This content may contain material generated or modified using artificial intelligence technologies.

B. Made with AI.

C. The voice is synthetic. The footage was filmed on location and is unedited.

D. Certain visual and audio elements in this production were created with the assistance of generative machine learning systems, in accordance with our editorial policy on synthetic media.

73 · 8 min

What the platforms require now

THE ONE THING TO KEEP

Platform disclosure rules are contractual, automatic and inconsistent between services, so the practical obligation on a publisher is set by where the work is distributed rather than by where they live.

The layer that reaches you first

Legal duties are covered in the final module. This lesson is about the rules that actually bite: the ones written by the services where your work appears, enforced by their systems, with account-level consequences and slow appeals.

The pattern across major services, described structurally because the details change:

Self-declaration at upload. A checkbox or field asking whether the content is synthetic or has been meaningfully altered. Video platforms introduced requirements of this kind for realistic altered content, with a stated obligation to disclose when a viewer might mistake it for real.

Automatic detection and labelling. Services read embedded watermarks and provenance metadata and apply their own label. This is why watermarking exists commercially: it makes labelling possible at scale without asking anyone.

Category-specific rules. Stricter requirements for political and election content, for health claims, and for anything depicting a real person. Some services prohibit synthetic political advertising outright; others require prominent disclosure.

Prohibitions rather than labels for the worst categories: synthetic sexual content of real people, fabricated statements by public figures presented as real, impersonation for fraud. These are removals, not disclosures.

What "meaningfully altered" tends to mean

The threshold most services have converged on is close to the interpretive standard from the previous lesson: disclosure is required when a viewer could reasonably mistake synthetic content for a real recording of a real event or person.

So generally not required: colour correction, stylistic filters, obviously unreal or fantastical imagery, background cleanup, production assistance such as scripting or editing help.

Generally required: a realistic depiction of a real person saying or doing something they did not; a realistic-looking event that did not happen; a real place altered to look materially different; synthetic voice presented as a real person's.

The frictions to expect

Automatic labels you did not ask for. A photograph edited in software with generative features can pick up provenance metadata that a platform reads as AI, producing a label on a real photograph. Contest it, and expect it to take time.

Labels that do not survive. A label applied by one platform disappears when the file is downloaded and reposted elsewhere. Anything you rely on being disclosed should be disclosed *in* the content, not around it.

Inconsistency. The same piece can carry a label on one service and none on another. There is no shared threshold, and there is no sign of one arriving soon.

Advertising review. Advertising policies are stricter and enforced harder than organic content rules, with faster consequences. If the work is an advertisement, read that policy specifically, because it is a different document.

What to do

1. **Read the actual policy** for each service you publish on, once, and reread when you hear it has changed. They are short.

2. **Declare accurately at upload.** A false declaration is a terms breach that can cost the account, and the upside of a false one is nothing.
3. **Put the disclosure in the content** as well as in the platform's field, so it survives redistribution.
4. **Keep the generation record**, which is what you will need if a label is applied wrongly.
5. **Assume the strictest rule** among the platforms you publish to, and make one version.

The honest assessment

These rules are new, unevenly enforced, and largely dependent on self-declaration and watermark reading — both of which the previous lessons showed can be avoided by anyone who wants to. They are therefore doing very little about deliberate misuse.

What they are doing is establishing a norm for the vastly larger volume of ordinary, non-malicious synthetic content, and creating a contractual hook that lets platforms act. That is worth something and it is not what it is often described as.

One more thing to know about how these rules are enforced, because it changes how you should respond. Enforcement is overwhelmingly complaint-driven and automated. Nobody is reviewing your upload; a system checks a watermark, a rights holder's matching system checks a fingerprint, or another user reports it. That means the decisions arrive without context and are appealed into a queue rather than to a person. It also means the single most effective protection is to be boring: accurate declarations, a clear in-content disclosure, no bulk uploading, and a record you can produce quickly. Accounts that never generate a signal are not the ones that get caught in the machinery.

For you, the practical summary is unromantic: the rules that constrain your work are contractual, they vary by service, they change without notice, and complying accurately costs you a minute per upload. Do it, keep the record, and put the disclosure where it travels with the file.

BEFORE YOU MOVE ON

Why should a disclosure be placed inside the content rather than only in a platform's disclosure field?

A. Because platform fields are not visible to viewers, only to the platform's moderation systems.
B. Because platform labels do not survive when the file is downloaded and reposted on another service.
C. Because platform fields are optional, whereas an in-content disclosure satisfies the legal requirement in every jurisdiction.
D. Because in-content disclosure exempts the publisher from the platform's own declaration requirement.

74 · 8 min

A routine you can actually run

THE ONE THING TO KEEP

A short, written, repeatable checklist beats expertise applied inconsistently, because the failure in verification is almost always skipping a step under time pressure.

Why a written routine

Everything in this module is knowable and most of it is known by people who nonetheless get it wrong. The failure is not ignorance. It is that verification happens under pressure — a story is breaking, a client is waiting, a post is spreading — and under pressure people skip steps and trust their eyes, which is exactly what does not work.

A checklist removes the judgement about whether this case needs checking. Aviation, surgery and newsrooms all reached the same conclusion for the same reason.

The routine

Before anything: preserve. Save the file. Record the URL, the account, the time you found it, and the claim being made. Screenshot the surrounding context. Content vanishes.

Step 1 — What is the claim? Write it in one sentence. "This shows X happening at Y on Z date." If you cannot write it, there is nothing to verify and you are reacting to a feeling.

Step 2 — Where has this appeared before? Reverse image search on at least two services. Sort by oldest where available. An earlier appearance ends the process.

Step 3 — Who is the source? Named or anonymous, account age, posting history, whether anyone accountable stands behind it. Find the original poster, not the account you saw it on.

Step 4 — Does the place check out? Satellite and street imagery against the visible geography. Signage, language, vehicles, side of the road, architecture, vegetation.

Step 5 — Does the time check out? Shadows, weather records, seasonal cues, anything dated in frame.

Step 6 — Is there corroboration? Independent recordings, credible reporting, official statements. Absence of corroboration for a major claim is itself a finding.

Step 7 — Provenance and metadata. Check for content credentials. Check metadata. Remember both are informative when present and meaningless when absent.

Step 8 — Only now, look at the image. Any artefacts consistent with the mechanisms in module three. Treat this as the weakest evidence you have collected, because it is.

Step 9 — Record the verdict as one of three. Verified, debunked, or unverified — with the reasons written down.

The routine, ordered by how much each step can settle

Preserve first	Save the file, the address, the account and the time you found it. Content disappears while you are looking at it.
Write the claim in one sentence	This shows X happening at Y on Z date. If you cannot write it, there is nothing to verify and you are reacting to a feeling.
Earliest appearance — often decisive	Reverse search on at least two services, sorted by oldest where possible. An earlier appearance ends the process in seconds.
The source, the place, the time	Who stands behind it; satellite and street imagery against the visible geography; shadows and archived weather records.
Corroboration	Events of consequence generate several independent recordings. Their absence is itself a finding.
Credentials and metadata	Informative when present, meaningless when absent, which is the usual case.
Only now, look at the image	Treat this as the weakest evidence you have collected, because it is.
Verified, debunked or unverified	Unverified is the most common result and a legitimate place to stop. Do not upgrade it to either certainty because a decision is needed.

Every step above the last depends on facts outside the file, which is why this method does not expire when a better model is released. For a group chat, the first three take two minutes and catch most of what circulates.

The rules that go with it

Unverified is a legitimate stopping point and the most common one. Do not upgrade it to either certainty because a decision is needed.

Never accuse in public on inspection alone. The cost of a false accusation falls on a real person and is not recoverable.

Time-box it. Decide in advance how long you will spend. Thirty minutes settles most cases; the ones it does not settle are rarely settled by three hours.

Say what you did. When you publish a conclusion, publish the method: "the image first appears in 2019 in an archive of a different event". This is far more persuasive than authority, and it lets others check you.

Adapting it

For personal use, steps 1 to 3 catch the overwhelming majority of misleading material circulating in group chats, and take under two minutes. That short version is the one worth teaching to family, and it is more useful than any amount of explanation about how the models work.

For a newsroom or an organisation, the full routine with a named person responsible and a written log is the professional standard, and the log matters as much as the checks — it is what lets you defend a decision later.

For a moderation or trust-and-safety context, the same routine plus scale: automated triage that routes to this process rather than deciding, for exactly the base-rate reasons from the first lesson.

The last word on this module

Nothing here requires you to identify a generated image, and that is the point. The methods that work are about origin, source and corroboration; the methods that expire are about pixels.

That is a more optimistic conclusion than most writing on this subject reaches. The tools for establishing where something came from are free, they are the same ones journalists have used for years, and they do not become obsolete when a better model is released. What they require is the willingness to spend two minutes and to accept "unverified" as an answer, which is a matter of discipline rather than of technology.

BEFORE YOU MOVE ON

Why does looking closely at the image come last in the routine rather than first?

A. Because artefact analysis requires specialist knowledge that most people applying the checklist will not have.

B. Because examining the image first would bias the interpretation of the later external checks.

C. Because platform compression removes the artefacts, so inspection can only work on original files.

D. Because inspection is the weakest evidence and the external checks usually settle it first.

CASE STUDY · MODULE 8

The strongest picture on the desk at eleven at night

The digital desk of a Marathi daily in Chiplun, during a night of flooding in the Konkan

Late July, heavy rain across the Konkan for a third day. Asmita Joshi runs the digital desk of a Marathi daily with a print edition in three districts and most of its readership on a phone.

At twenty to eleven a photograph arrived on the paper's tipline: a bus submerged to the windows, a man standing on the roof holding a child, a temple spire in the background. It was the strongest image anyone had of the night. Three competing outlets were running flood coverage with nothing like it.

The sender was a number with no name. The messaging app labelled the image as forwarded many times.

The desk publishes at eleven. On a night like that one, the traffic is the month's traffic, and the paper's own district is exactly where its readers expect it to be first.

Running the picture with a line saying the paper could not independently verify it is the ordinary compromise, and on that night it was not a small thing. People in low-lying parts of Chiplun were deciding whether to move to a relative's house. An image of a bus underwater on a named road is information they would act on. If it was fabricated or misplaced, the paper would be the outlet that laundered it into the record, and a correction at nine the next morning reaches a fraction of the audience the picture does.

Holding it meant a competitor running it within the hour, the paper looking slow on its own ground, and — if the picture was real — failing to show readers something true on the night it mattered.

A reporter, Nikhil, had put the file through a detector he found by searching. It reported eighty-seven per cent AI, and he wanted to lead with that.

Asmita asked him four questions. Eighty-seven per cent of what. Trained on which generators, and is this image from one of them. What happens to that number after a photograph has been through two rounds of messaging-app compression. And at a base rate where the great majority of pictures sent to a tipline on a flood night are real photographs taken by real people, what proportion of the images this tool flags are actually generated.

She was not being clever at Nikhil's expense. She had been accused twice of running a fake and once, more damagingly, of calling a real photograph fake.

Then she ran the routine, and it took twenty minutes.

Preserve first: save the file, screenshot the message and the number, note the time it arrived and the claim being made. Write the claim in one sentence — this shows a bus submerged tonight on the Chiplun road.

Reverse image search on two services. One returned nothing. The other returned a visually similar photograph from the 2021 Konkan floods, which was a different bus on a different road, so it ruled nothing in or out.

Source: a number with no name, forwarded many times, no original poster findable. A dead end, and dead ends are findings.

Place: the temple spire was distinctive. Satellite imagery showed a temple of that shape on a road near Chiplun. Street-level imagery from 2022 showed the building beside it as a single-storey shop. In the photograph it was two storeys with a blue plastic water tank on the roof.

Time: that was the one that ended it, and it had been in front of everybody for twenty minutes. The photograph is in daylight. It had been sent at twenty to eleven at night, as a picture of that night.

STOP HERE

Before turning the page: what would you have done, and what would it have cost if you were wrong? Write it down. The point of the next section is to compare.

WHAT HAPPENED

The picture was from the 2021 flood on the same road. The building had been rebuilt in the years since and had gained a storey and a water tank, which is why the street imagery disagreed. The bus was a real bus and the water was real water, and the only false thing about it was the date it was being given.

This is the most common form of misleading media and it involves no generation at all. It survives every kind of inspection, because nothing in the pixels is fake. The detector's eighty-seven per cent was pointing at an image that no model had touched.

Asmita did not run it. The paper ran the story at ten past eleven with two photographs from its own stringer in Khed — a flooded approach road and a rescue boat, both less dramatic — and a short note at the foot explaining how images sent to the tipline are checked, with the method in it rather than an assertion of care. A competitor ran the bus picture at twenty past eleven and took it down at nine the next morning.

Three days later the other half of the problem arrived. A reader sent a genuine photograph of a collapsed retaining wall behind a new municipal building, and within a few hours supporters of a local politician were saying online that it was AI-generated. On another day that would have been unanswerable. Because the desk had the stringer's own capture with its time and place, a second photograph of the same wall from a different angle taken by a different person an hour later, and a written note of what had been established, it could publish the basis rather than the conclusion.

Asmita now keeps the routine printed and taped inside a cupboard door, and the version she teaches to reporters on their phones is only the first three steps, because those three catch most of what circulates in a group chat and take under two minutes.

The detector reported eighty-seven per cent AI on a photograph that no model had touched. What does that tell you about the tool, and what does it tell you about this case?

About the tool: nothing surprising. Detectors learn the fingerprints of the generators in their training set, degrade sharply on anything else, and lose much of the

signal they rely on to compression — and this file had been through repeated messaging-app re-encoding. At a base rate where nearly every tipline image is a real photograph, most of a detector's positives are wrong however good its stated accuracy. About the case: nothing at all. It pointed at a real photograph and would have pointed the desk away from the actual problem, which was the date rather than the pixels.

Which step actually settled it, and why is that kind of step durable when inspection is not?

The time check — daylight in a photograph presented as tonight — supported by the place check, where street-level imagery showed a single-storey building where the photograph had two storeys and a water tank. Both depend on facts outside the file: an archive of street imagery, the position of the sun, a building that was rebuilt. None of them is affected by how good any generator becomes, so the method does not expire with a model release. Inspection-based methods are calibrated against the artefacts of current systems and are obsolete by design.

Three days later a genuine photograph was publicly called AI-generated. Why did detection have nothing to offer there, and what did?

Because the claim being made was that a real recording was fabricated, and no detector can establish that something is real — the absence of generation artefacts, like the absence of a watermark, proves nothing. That is the liar's dividend: it costs the accuser nothing, needs no fake to exist, and cannot be answered by better detection. What answered it was provenance and corroboration: the paper's own stringer capturing the image, a recorded time and place, and a second photograph of the same wall from a different angle by a different person. Positive evidence about origin is the only thing that touches this, which is why provenance at capture matters more as generators improve.

MODULE 8 · TEST

Check what stuck

6 questions, one right answer each. This one is for you, not for a record: answers and the reason behind each are at the back. If two questions from the same module go wrong, re-read that module before the exam rather than after.

1 A detector that is 95 per cent accurate in both directions is run over 100,000 images, of which about 100 are generated. Roughly what fraction of its positive results are correct?

A. About ninety-five in a hundred.
B. Fewer than two in a hundred.
C. About half.
D. About one in five.

2 An image carries no invisible watermark from any known generator. What does that establish?

A. Nothing, because most images in the world are unmarked.
B. That it is more likely than not to be a photograph.
C. That it was produced before watermarking was introduced.
D. That it has been re-encoded at least once since it was created.

3 A photograph arrives with valid Content Credentials showing capture by a named camera model at a stated time. What have you established?

A. That the image has not been edited since capture, and that the scene is as described.
B. That the accompanying caption is accurate, since the manifest binds to the content.
C. That the file came from that device by that process, and nothing about what it means.
D. That the image is not generated, because generated files cannot carry credentials.

4 During a crisis, the single most common form of misleading imagery in circulation is not generated at all. What is it, and why does inspection miss it?

A. Heavily filtered photographs, because the filter conceals the original content.

B. Composites of two real photographs, because the seam is hidden by compression.

C. Staged photographs, because nothing in the file distinguishes staging.

D. A real photograph of an older event with a new caption, because nothing in the pixels is false.

5 Why can better detection not address the liar's dividend?

A. Detectors are proprietary, so their results cannot be presented publicly.

B. The claim is that nothing can be established, not that a particular analysis is wrong.

C. Detectors cannot run on video at the resolutions news footage is shot at.

D. The dividend applies only to audio, where detection is weakest.

6 An organisation adds 'this content may contain AI-generated material' to every page it publishes. What is the main cost of that policy?

A. It exposes the organisation to liability for content that is not in fact synthetic.

B. It violates the disclosure rules of most major publishing platforms.

C. It trains the audience to ignore the label, so the important disclosure disappears into noise.

D. It prevents the organisation from claiming copyright in its own work.

MODULE 9

Ownership, law and the parts nobody has settled

The legal questions around generative media are genuinely unresolved, differ by country, and are moving. This block sets out the shape of each disagreement rather than pretending to resolve it: whether training on copyrighted work is lawful, whether the output can be owned and by whom, what protects a person's face and voice, why style sits outside copyright, what the licence on a model actually permits, what a vendor's indemnity is worth, the disclosure duties that are already law in several places, and what to write into a contract when the law will not tell you.

BY THE END OF THIS MODULE YOU CAN

Identify which distinct legal question a given situation raises — training, output ownership, likeness, style, model licence, or disclosure duty — describe how the answer differs across jurisdictions, and draft the contractual terms that make a piece of AI-assisted client work defensible without relying on a resolution that does not exist

75 · 12 min

Who owns it, and who has to say it is AI

THE ONE THING TO KEEP

Ownership of AI output differs by country and is genuinely unsettled; the duty to label it is not.

Part one: who owns it

The United States: no human author, no copyright

The US Copyright Office has been consistent. In 2023 it partly cancelled the registration of *Zarya of the Dawn*, a comic illustrated with Midjourney: the writer's text and her arrangement of panels stayed protected, the images themselves did not. A prize-winning image, *Théâtre D'opéra Spatial*, was refused. In 2025 a federal appeals court confirmed that a machine cannot be an author.

The Office's 2025 report addressed the question everyone actually cares about. Prompts alone do not make you an author, however long, however original, however many iterations you ran. But your own expressive contributions can be protected: a drawing you fed in, edits you made afterwards, the selection and arrangement of generated pieces into a larger work.

The UK and its family: an author by legal fiction

UK law has an unusual provision from 1988. Where a literary, dramatic, musical or artistic work is computer-generated with no human author, the author is deemed to be the person who made the arrangements necessary for its creation, with a fifty-year term. Ireland, New Zealand, Hong Kong, South Africa and India have similar provisions. The UK has consulted more than once on removing it.

India's own experience shows how unresolved this is. In 2020 the Indian Copyright Office registered an artwork listing an AI as co-author, then issued

a withdrawal notice. The registration and the retraction are both on the record.

China: the opposite conclusion

In November 2023 the Beijing Internet Court held that an AI-generated image *was* protected, because the user's prompts, parameter choices and repeated refinement showed personalised intellectual investment. Similar facts, opposite result.

So

The same picture may be unprotectable in Washington, protected in Beijing, and owned by legal fiction for fifty years in London. Anyone telling you this is settled is selling something.

What to do while it stays unsettled

- Treat the prompt as worthless as property. It is not a moat, and in the US it is not authorship.
- If you need exclusivity, add human authorship you can point to, and keep the working files: your own photograph underneath, your edits, your compositing, your arrangement.
- Read platform terms as contract, not copyright. "You own your outputs" means the company will not claim them. It cannot manufacture a right that does not exist, and it usually does not stop another user generating something near-identical.
- For logos and brand marks, route around the problem. Trademark protects a mark through use and registration regardless of how it was drawn.
- The bigger commercial risk is rarely "do I own this". It is "does this infringe something": a trademarked character, a recognisable person, a style so specific it names a living artist.

Part two: who has to say it is AI

Here the law is moving faster, and in one direction.

China went first and hardest. Labelling rules in force since 1 September 2025 require generated content to carry both an explicit label a person can see or hear and an implicit label inside the file's metadata, and require platforms to check declarations and label what they detect.

The European Union. The AI Act's transparency article requires providers to mark generated output in machine-readable form, and requires anyone deploying a deepfake to disclose it. It was scheduled to apply from August 2026, and that timetable has been amended more than once — so confirm the current position rather than trusting any date, including this one.

India amended its IT Rules for synthetically generated information after a late-2025 consultation. The striking part is that it prescribes *visibility*: a label covering a defined portion of the frame and the opening portion of audio, plus a duty on platforms to verify user declarations. It is an answer to a real problem. A label nobody sees is not a label.

The United States has no general federal labelling law. It has the 2025 TAKE IT DOWN Act, requiring platforms to remove non-consensual intimate imagery including synthetic imagery, and a patchwork of state election-deepfake laws, several challenged on free-speech grounds.

Denmark moved to give people a right in their own likeness and voice.

The technical layer, and why it is not enough

Content Credentials, built on the C2PA standard, attach a cryptographically signed manifest to a file recording what created it and what edited it. Some cameras and most major generators support it. Invisible watermarking — patterns embedded in the pixels or the audio — is a second layer.

Both are real and both are fragile. A screenshot destroys them. Many re-encodes and social uploads strip them. And detectors claiming to identify AI content from pixels alone are unreliable in both directions: they clear synthetic images and they accuse real photographers, which has already cost people work and reputation.

The practical consequence: provenance from the source beats detection after the fact, and a visible label beats both, because it is the only part that survives

a screenshot.

The rule that outlives the statutes

If a reasonable viewer could mistake your output for a record of something that happened, say that it is not — in the frame, at the start, where they will see it before they believe it. Not in the credits. Not in the alt text. Not in a caption they will scroll past.

BEFORE YOU MOVE ON

A designer generates a logo entirely from prompts for a client in the United States, and the client wants exclusive rights. What is the most accurate advice?

A. Writing a long, highly original prompt establishes authorship, since every creative choice in it belongs to the designer.

B. The image is unlikely to attract US copyright, so exclusivity has to come from trademark registration and contract rather than from copyright.

C. The generator's terms transfer ownership of outputs to the user, so the designer already holds an exclusive right to assign onward.

D. Registering the logo with the Copyright Office secures exclusivity, since registration is what creates the right.

76 · 9 min

Was the training lawful?

THE ONE THING TO KEEP

Whether training on copyrighted work without permission is lawful is being answered differently in different countries and differently for different facts, so there is no single answer and any confident one is a jurisdiction or a sales position.

What everyone agrees on

The factual base is not in dispute. Large models were trained on very large quantities of copyrighted material — images, text, recordings — collected by web scraping or from existing datasets, in most cases without individual permission and without a workable route for creators to object beforehand.

The disagreement is entirely about whether that is lawful, and it takes different forms depending on where you are.

The argument for lawfulness

Training is analysis, not reproduction of expression. The model extracts statistical regularities. What it retains is not the works but abstractions from them, and copyright protects expression rather than facts, ideas or style.

Copies made during training are incidental to a transformative process, in the same family as the intermediate copying that search engines and text-and-data-mining research have long relied on.

The output is generally not substantially similar to any particular input, and where it is, that specific output is the infringement rather than the training.

The alternative is unworkable at scale. Licensing billions of works individually is not possible, so a rule requiring it forbids the technology rather

than pricing it.

The argument against

Copying happened, at enormous scale. Every training run required reproductions of the works, whatever was done with them afterwards.

Commercial substitution. A model that produces illustration in a style competes with the illustrators whose work trained it. That market effect is precisely what several fair-dealing tests weigh most heavily.

Consent was never sought even where it could have been, and opt-out schemes arrived after the training.

Scale is not an excuse. "We could not have licensed it all" is an argument about business models, not about rights, and other industries with licensing problems solved them by building licensing systems.

Where courts have got to

Enough has happened to describe the terrain, without treating any of it as settled.

Courts in the United States have reached different conclusions on different facts. A ruling found no fair use where a system reproduced editorial content to compete directly in the same market. Another found that training a language model on lawfully acquired books was transformative fair use, while holding that building a library from pirated copies was a separate and serious matter, ultimately resolved by a very large settlement. Another granted judgment to a developer on the record before it while explicitly noting that a better-evidenced case on market harm might come out differently. The pattern is that *how the material was acquired* and *what market effect is proved* are doing most of the work.

In the United Kingdom, a major image case ended with the copyright claims largely falling away — partly on evidence about where training took place — and a limited trademark finding. The UK's text-and-data-mining exception covers non-commercial research only, and a government consultation on

widening it, with an opt-out for rights holders, was contested by creators and remains unresolved.

The European Union has TDM exceptions with a machine-readable opt-out for commercial use, and the AI Act requires general-purpose model providers to publish a summary of training content and to respect those reservations.

Japan has one of the broadest permissions for machine learning, with limits where the use unreasonably prejudices the rights holder.

Other jurisdictions, India among them, have not tested the question in a decided case, and the applicable exceptions were written long before any of this.

What this means for you

Three practical positions that hold regardless of how it resolves.

Your exposure is on output, not training. You did not train the model. The risk you carry is publishing something that infringes, which is the melody and memorisation problem from earlier modules.

Prefer providers who say what they trained on, and read the answer carefully rather than accepting the word "licensed".

Do not repeat either side's confident version. The honest sentence is: it depends on the country, on how the material was obtained, on what market harm can be shown, and it has not been resolved.

The reason to know the argument rather than just the conclusion is that you will be asked, by clients, by collaborators, and by people who are angry about it for reasons that are entirely legitimate. Being able to set out both cases accurately is more useful, and more respectful to the people affected, than picking one.

BEFORE YOU MOVE ON

Why is "training on copyrighted work is fair use" an unsafe thing to state as settled?

A. Because fair use has been rejected in every case decided so far, making the opposite the safer statement.
B. Because fair use exists in few jurisdictions, and outcomes have varied with the facts.
C. Because fair use applies only to non-commercial use, and model training is commercial.
D. Because the question concerns the output rather than the training, so fair use is not the relevant test at all.

77 · 9 min

Can anyone own the output?

THE ONE THING TO KEEP

Most copyright systems require a human author, so purely prompted output is unprotected in several major jurisdictions while human contribution to selection, arrangement and editing can be, and a few countries have a specific rule for computer-generated works.

The question, split properly

"Who owns AI output" collapses two different questions that need separating.

Is there copyright at all? A work with no protection is not owned by anybody; it is free for anyone to use, including your client's competitor.

If there is, who holds it? That is answered by contract and by employment law, and it is the easier half.

Where the human-authorship requirement bites

In the United States, the Copyright Office has stated its position across a series of decisions and a formal report: copyright requires human authorship, prompts alone do not supply it however detailed, and material generated by a machine in response to a prompt is not protectable. Where a human selects, arranges, or modifies generated material with sufficient creativity, that human contribution can be protected — but the protection covers the contribution, not the generated elements underneath.

A well-known registration involving a graphic novel with generated images ended exactly there: the text and the arrangement of the pages were protected, the individual images were not. A separate line of cases confirmed that a work with no human author cannot be registered at all.

Where the rule is different

The United Kingdom has an unusual provision for computer-generated works with no human author, giving them a term of fifty years with the author deemed to be the person who made the arrangements necessary for creation. It has been on the books since 1988, it has barely been tested, and the government has consulted on whether to keep it.

India has a similar mechanism: for a computer-generated work, the author is the person who causes the work to be created. There was a much-discussed registration involving an AI system named as co-author, which the office subsequently sought to withdraw, leaving the position ambiguous.

The European Union has no equivalent provision, and its case law emphasises the author's own intellectual creation and free creative choices, which points toward requiring human authorship.

China has produced decisions finding copyright in AI-generated images where the court was satisfied the user made sufficient intellectual investment through prompts, parameters and selection — a notably different emphasis from the United States on very similar facts.

The same image can therefore be protected in one country and free in another. This is not a temporary anomaly awaiting harmonisation; it follows from gen-

uinely different foundations in each system.

China has found copyright on facts close to the top-left cell, reasoning that prompts, parameters and repeated refinement showed intellectual investment. The same picture, protected in one country and free in another. Anyone who calls this settled is describing one office's current guidance.

What this means commercially

A client wanting exclusivity should be told the truth. If a logo, a character or a campaign image is purely generated, it may not be protectable in some markets, which means a competitor could reproduce it. That is a real commercial fact and hiding it is the kind of omission that ends relationships.

Human contribution is the answer, and it is not a trick. Substantial editing, compositing, arrangement, typography, and creative selection produce a work with real human authorship — and, not incidentally, a better piece of work. This is another reason the craft chapters of this track matter more than the prompting.

Document the human contribution. Layered files, working versions, notes on what was decided and why. If protection is ever asserted, this is the evidence.

Other rights may still apply. A logo can be a registered trademark whether or not there is copyright in the artwork, which is frequently the protection that actually matters for a brand mark.

The unresolved part

Nobody has drawn the line between "prompting" and "authorship" in a way that survives contact with real practice. Is a hundred iterations with masking,

structural conditioning and a custom fine-tune authorship? Most people's intuition says yes and no office has said where the threshold is. Is a single prompt with a careful choice among four hundred outputs authorship? Photography's history suggests selection can matter enormously, and the current guidance suggests it does not.

Anyone who gives you a confident threshold is describing their jurisdiction's current guidance, which is guidance rather than legislation and which has already changed once.

None of this is legal advice, and a decision that actually matters commercially is worth an hour of a qualified lawyer's time in the relevant country. What this lesson gives you is the ability to know that the question exists, which is what stops people promising a client something nobody can deliver.

BEFORE YOU MOVE ON

A client asks for exclusive rights in a purely generated brand image. What is the accurate response?

A. Exclusivity can be granted by contract, which is sufficient because contractual assignment creates the underlying right.

B. Copyright arises automatically on creation, so the image is protected and the rights can simply be assigned.

C. Purely generated material may not be protectable in some countries, so exclusivity may need a registered trademark or substantial human authorship.

D. The rights belong to the model provider under its terms of service, so securing exclusivity for the client would require negotiating a separate licence directly with that provider.

78 · 9 min

Faces, voices and the rights in a person

THE ONE THING TO KEEP

Rights in a person's face and voice sit outside copyright, vary enormously by country, and are the area moving fastest in legislation, so this is where a generated work is most likely to be actionable.

A separate body of law

Copyright protects works. It does not protect people. A photograph of you is owned by the photographer; the fact that it is you is governed by something else entirely.

That something else has different names and different shapes:

- **Right of publicity or personality rights** — a right to control commercial use of your identity. Strong in parts of the United States, recognised in various forms elsewhere.
- **Image rights and passing off** in jurisdictions with no dedicated right, where a false endorsement claim does similar work.
- **Data protection.** In the European Union and the United Kingdom, a person's face and voice are personal data and, when used to identify them, biometric data with heightened protection. This is frequently the strongest and most overlooked instrument.
- **Defamation**, where the depiction says something false and damaging.
- **Specific criminal offences** for non-consensual sexual imagery and for fraudulent impersonation, which more jurisdictions add each year.

Where the law has moved recently

This is the fastest-moving area in the whole field, and a few concrete markers are worth knowing.

Several US states have enacted statutes aimed specifically at synthetic replicas of voice and likeness, including one framed around musicians' voices, and further federal proposals have been introduced without yet becoming law. Coverage of deceased persons varies by state and is one of the sharper differences.

Indian courts have granted a series of personality-rights orders protecting named public figures against unauthorised use of their name, image, voice and mannerisms, including at least one specifically addressing AI voice cloning of a singer. These are injunctions in individual cases rather than a statute, and they have built quickly into a recognisable line.

Denmark has proposed giving people a copyright-like right over their own likeness, which would be a significant departure if enacted, and at the time of writing it remains a proposal.

The European Union's AI Act adds transparency duties for deepfakes on top of the existing data-protection position, with those obligations applying from 2026.

The direction is consistent even though the instruments differ: more protection, more specifically aimed at synthetic replicas, arriving faster than most legal change.

The practical standard

Because the law is uneven and moving, the workable rule is not a legal test. It is a practice standard that satisfies most of them:

Consent that is informed, specific, written, time-limited and revocable, with payment where the use is commercial.

Read those five conditions again, because each one fails a common shortcut. Verbal agreement fails "written". A blanket release fails "specific". A perpetual grant fails "time-limited". "You agreed once" fails "revocable".

And separately, **disclosure to the audience**, which the person depicted cannot waive on the audience's behalf.

The cases that are simply off limits

Some things need no legal analysis:

- Sexual imagery of any real person without consent. Criminal in a growing number of jurisdictions and wrong everywhere.
- Anything depicting a child.
- Fabricated statements attributed to a real person, presented as real.
- Impersonation for fraud.
- Using a dead person's likeness where the family objects. The law varies; the judgement does not have to.

The awkward middle

Satire and comment have protection in many systems and it is not unlimited, and a photorealistic fabrication is treated differently from an obvious caricature precisely because one can be mistaken for a record.

A person who resembles someone. Prompting for a type rather than a name can still produce a recognisable individual, and recognisability is generally the test rather than intent.

Historical figures. Generally lower risk, and reconstructions of real historical people presented without labelling raise a different and serious problem about the record.

Your own likeness, licensed to a client. Increasingly common and worth negotiating deliberately, because a synthetic version of you is an asset that outlives the campaign.

The unresolved part: nobody has settled how far these rights extend to a style of performance rather than an identity — a voice that sounds like a genre rather than a person, a face assembled from features. Cases are beginning to test it. Until then, recognisability is the practical line, and if people recognise the person, treat it as their right regardless of how the image was assembled.

BEFORE YOU MOVE ON

Which factor most reliably indicates that a generated depiction engages someone's personality rights?

A. That the person's name was used in the prompt, since the right attaches to use of the name.
B. That the depiction is photorealistic rather than illustrated, since only realistic depictions can be mistaken for a record.
C. That the use is commercial, since personality rights protect only commercial exploitation of identity.
D. That an ordinary viewer would recognise the individual, however the depiction was produced.

79 · 8 min

Style is not owned, and other things are

THE ONE THING TO KEEP

Copyright does not protect artistic style, so the strongest objections to style imitation have no copyright remedy, while trademark, passing off and false endorsement can reach the same conduct from another direction.

The rule that surprises people

Copyright protects the expression of a work, not the manner of making works. A painter's palette, brushwork, subject matter and recurring motifs are not, as such, protected. Anyone may paint in the style of anyone. This has always been true, it is why art movements exist, and it is not a loophole introduced by generative models.

What is protected is a specific work. Reproducing a particular painting infringes; painting a new picture in the same manner does not.

So the most emotionally powerful objection to these systems — that a model produces work in a living artist's style, trained on that artist's work, competing with them — has, in most jurisdictions, no copyright answer. This is worth understanding precisely, because it explains why litigation on behalf of artists has been reframed around other theories.

What can reach it

Trademark. An artist's name can be a trademark for goods and services. Using a name to promote generated images can be trademark use and can mislead as to origin. In one prominent case, a claim about generated images bearing a stock library's watermark succeeded in trademark where the copyright claims did not.

Passing off and false endorsement. Presenting work in a way that suggests an artist made it or approved it is actionable in many systems without any copyright question arising.

Trade dress and design rights. The distinctive look of a product or packaging can be protected separately from any artwork.

Contract. A commission agreement, a platform's terms, a licence on a training set — these can restrict what copyright does not.

Moral rights, in systems that have them, cover attribution and derogatory treatment of specific works rather than style.

The practical line

For anyone making things, the distinction that matters:

Generally acceptable: learning from work you admire, producing in a genre or movement, using descriptive style vocabulary — impressionist, art deco, low-poly, mid-century.

Risky: naming a living artist in a prompt and publishing the result; presenting work as being by or approved by someone; training a fine-tune on one living artist's work and marketing it under their name; anything that could mislead about origin.

Beyond the legal question: naming a living artist to imitate them is, whatever any court eventually says, taking something from a person who is trying to make a living. Many artists have said so plainly and there is no reason to disbelieve them. A great deal of the anger directed at this technology comes from people who found their name used as a prompt keyword and had no way to object.

You do not need a legal instrument to decline to do that. "It is not illegal" is a poor foundation for a professional practice, and the reputational cost inside a creative industry is real and long-lasting.

The style-transfer question in fine-tunes

Training a small fine-tune deserves its own note because the module on control encouraged it.

On your own work, your client's assets, openly licensed material or public-domain art: uncomplicated.

On a living artist's work, to reproduce their style: legally uncertain, ethically contested, and commercially dangerous. Several artists have pursued people who did this publicly, and platforms hosting fine-tunes have removed them on request.

On a deceased artist whose work is in the public domain: legally clear and still worth thinking about, particularly where an estate or a cultural community has a stake.

What is genuinely unresolved

Whether style should be protected at all is a live policy argument rather than a settled question. The case for is that automated, scaled imitation is different from a human learning from an influence, and that the existing rule assumed a scale that no longer holds. The case against is that protecting style would be a drastic expansion of copyright that would restrict every artist who ever learned from another, and that the boundaries would be unworkable.

What an artist can actually do

Because the question comes up and "nothing" is not the true answer, the available steps are worth listing.

Register trademarks for a name and any distinctive mark used in trade. Put explicit terms on a portfolio site and in commission agreements about training and machine reading. Use the machine-readable reservations that the EU exception and several platforms recognise, imperfect as adoption is. Contact platforms hosting fine-tunes trained on your work, since several remove them on request. Watermark and register works so a specific reproduction can be proved rather than argued.

None of it stops style imitation, because nothing does. What it does is protect the things that are protectable, which is a smaller and real amount.

Both arguments above are serious. Nothing about the current law was designed with this in mind, and the argument that it needs revisiting is not the same as an argument about what the law currently is. Keeping those two apart is most of what it takes to discuss this well.

BEFORE YOU MOVE ON

Why have several artist-led claims been framed around trademark rather than copyright?

A. Because trademark claims tend to carry higher damages awards, which makes them the commercially preferable route for a claimant with limited resources.

B. Because copyright does not protect style, so imitation without reproduction of a specific work has no copyright remedy.

C. Because trademark registration is automatic for artists' names, making the claim easier to bring.

D. Because copyright claims cannot be brought until a work has been formally registered in the relevant jurisdiction.

80 · 8 min

Read the licence on the model

THE ONE THING TO KEEP

Model weights come with licences that restrict use, and outputs come with terms of service that assign or condition rights, so what you may do with a generated file is set by two documents most people never open.

Two separate documents

The model licence governs the weights. It says whether you may use the model commercially, whether you may modify or redistribute it, and what uses are forbidden.

The terms of service of a hosted product govern the output. They say who owns what comes out, whether you may use it commercially, whether the provider may use your inputs for training, and what happens on a dispute.

These are independent. A permissively licensed model used through a restrictive service leaves you bound by the service's terms.

What model licences actually say

The range is wider than people expect, and the differences matter.

Fully permissive — Apache 2.0, MIT. Commercial use, modification and redistribution allowed. Several capable open models are here.

Non-commercial only. Some prominent, high-quality open models are released for research and non-commercial use, with commercial use requiring a separate paid licence. Using one for client work without that licence is a breach, and this catches people constantly because the model is freely downloadable and works perfectly.

Community or bespoke licences with conditions: revenue thresholds above which a commercial licence is required, attribution requirements, restrictions on training other models on the output, or clauses that terminate the licence if you bring certain legal claims.

Responsible-AI licences that permit broad use but forbid enumerated purposes — surveillance, discrimination, generating certain content. Their enforceability is untested and they express intent clearly.

Derived weights inherit conditions. A fine-tune of a non-commercial base model is non-commercial. A merge of several models carries every restriction in the set. Community model repositories are full of merges whose licence position is genuinely unclear, and "it was on the internet" is not a defence.

What service terms typically say about output

Most major services now assign whatever rights they may have in outputs to the user, and this is worded carefully — it does not create copyright where none exists, and it does not warrant that the output infringes nothing.

Then read the conditions:

- **Commercial use** may depend on your subscription tier, with free-tier output restricted.
- **Training on your inputs** — whether your uploaded images and prompts are used to improve the service. This is often on by default on consumer tiers and off on business ones. For client-confidential material, this is a contractual problem before it is a privacy one.
- **Attribution** may be required on some tiers.
- **Retention and deletion.** How long inputs are kept, and whether you can require deletion.
- **Content restrictions** beyond the law, which can be broader than you expect.

The checks worth running before a job

Five minutes, once per tool:

1. Find the model licence. If it says non-commercial, do not use it for paid work.
2. Find the output terms. Confirm commercial use at your tier.
3. Check whether your inputs are used for training, and turn it off if the material is confidential.
4. Check for attribution requirements and comply with them.
5. Save a copy of both documents with the date. They change, and what governed your use is the version in force at the time.

That last point is not pedantry. Terms are revised, sometimes materially, and a screenshot with a date has resolved more than one dispute.

Where this collides with client contracts

A client contract that warrants you have all necessary rights, combined with a non-commercial model licence, puts you in breach of one or the other. This is a common and entirely avoidable failure.

Equally, a client contract requiring confidentiality is inconsistent with a service that trains on your uploads. Read both sides before signing either.

One more provision worth looking for, because it catches people building on top of these tools: **restrictions on using the output to train another model**. Several services forbid it outright. If your plan involves generating a dataset and fine-tuning something on it — which is a common and otherwise sensible technique for building a house style — that clause decides whether the plan is available to you. Open models with permissive licences generally allow it; hosted services frequently do not.

The limitation worth stating: none of these documents tells you whether the output infringes somebody else's rights. They allocate rights between you and the provider. Everything about third parties — the melody, the memorised image, the recognisable face — sits outside them entirely, which is what the next lesson is about.

BEFORE YOU MOVE ON

Why can a freely downloadable, high-quality open model be unusable for client work?

A. Because open models produce output that cannot be copyrighted, so a client cannot receive exclusive rights.
B. Because open model weights carry no licence at all, leaving the legal position undefined.
C. Because its licence may permit only non-commercial use, requiring a separate paid licence for client work.
D. Because locally run models produce no watermark, and platforms require watermarked output for commercial distribution.

81 · 8 min

What an indemnity actually buys

THE ONE THING TO KEEP

A vendor indemnity is a conditional contractual promise with exclusions and caps, so it shifts some risk on some facts and never removes the obligation to check your own work.

What is being offered

Several providers now offer to defend and cover customers against third-party copyright claims arising from output. This is a genuine commercial commitment and it is worth having. It is also narrower than the headline.

An indemnity is a contract term: if a third party brings a claim of a specified kind, the vendor will take on the defence and pay defined amounts. Everything turns on the specification.

The conditions to read

Every indemnity worth the name has conditions. The usual set:

Tier. Frequently available only on paid business or enterprise plans, not free or consumer tiers.

Unmodified output. Some require that you used the output as generated. If you edited it, composited it, or ran it through another model, cover may lapse — which is exactly the workflow the rest of this course recommends, so read this one carefully.

Safety features left on. If you disabled filters or circumvented protections, cover usually goes.

No knowledge. If you knew or should have known the output was infringing, you are outside it. Prompting for a named artist and then claiming surprise sits here.

Prompt conduct. You must notify within a stated period, and failing to notify promptly can void the whole thing.

Vendor control of the defence. They choose the lawyers and the strategy, and may settle in a way you would not have chosen.

Caps. Often limited to what you have paid, or to a stated figure, and consequential losses — your lost business, your client's — are typically excluded.

What is usually not covered

Personality and publicity rights. Most indemnities are drafted around copyright. A claim from a person about their face or voice is often outside them, and that is one of the most likely claims to arise.

Trademark, sometimes.

Defamation.

Regulatory penalties, including for failing to disclose where disclosure is required by law.

Your client's losses, which is what will actually be claimed against you.

That last one deserves emphasis. If a campaign is pulled, the client's loss is commercial — the media spend, the reprint, the delay. A vendor indemnity aimed at a rights holder's copyright claim frequently does nothing about it, and the exposure sits with you under your own contract.

What an indemnity buys, and what is claimed against you

Usually inside it	Usually outside it
• A third party's copyright claim about the output	• A claim about someone's face or voice
• The cost of a defence the vendor controls	• Trademark, sometimes
• Damages up to a stated cap	• Defamation
• On a paid tier, with safety features left on	• Regulatory penalties for failing to disclose
• If you notified within the stated period	• Your client's losses — media spend, reprint, delay

Read the exclusions first, because that is where the drafting effort went. The rights most likely to be asserted against you personally are frequently outside it, and the loss your client will actually claim is almost always outside it.

How to use one properly

Read it before relying on it. It is a page. Read the exclusions first; that is where the content is.

Keep the evidence of compliance. Which tier, which settings, what you generated and when. If cover depends on conditions, you have to be able to show you met them.

Do not let it replace checking. Reverse image search on anything that will be published. Melody search on a hook. Recognisability check on any face. These take minutes and they prevent the claim rather than paying for it.

Match your client contract to it. If you warrant to a client something broader than your vendor covers, the gap is yours. Say what you can honestly say, and no more.

What to tell a client

An honest formulation, and clients respect it:

> *The generated elements come from a service that indemnifies commercial use on our plan, within its stated conditions. We have checked the output against reverse image search and it does not match any known work. We cannot guarantee that no claim will ever be made, because nobody can. Here is what we have done and here is the record.*

Compare that with "it is all fine, they indemnify us". The first is defensible. The second is a promise you cannot keep, and it is the one that ends badly.

It is also worth knowing why these offers exist. They appeared as a competitive response when enterprise buyers began refusing to adopt generative tools without one, and they are priced into business tiers. That is a perfectly ordinary commercial arrangement and it tells you something useful: the indemnity is a sales instrument as much as a legal one, which is why the headline is generous and the exclusions are where the drafting effort went. Read them in that spirit rather than as a reassurance.

The honest summary: an indemnity is a useful risk transfer for a specific class of claim on specific facts, offered by a party with far deeper pockets than you. It is not a licence to skip the checks, it usually excludes the rights most likely to be asserted against you personally, and it never protects your client from their own losses. Treat it as one control among several rather than as the answer.

BEFORE YOU MOVE ON

Which claim is a vendor's output indemnity least likely to cover?

A. A copyright claim from a rights holder whose work resembles the generated image.

B. A claim about the resemblance of a generated character to a copyrighted fictional character.

C. A claim that the output reproduced a substantial part of a protected photograph.

D. A claim from an individual about the unauthorised use of their face or voice.

82 · 9 min

Where labelling is already the law

THE ONE THING TO KEEP

Several jurisdictions now impose legal duties to label synthetic content, with different triggers, timings and penalties, so the applicable rule depends on where the audience is rather than where you are.

The shift from ethics to obligation

Earlier modules argued for disclosure as good practice. In several places it is now a legal duty, and the trend is one-directional.

The rules differ in a way that matters operationally: they attach to *audience location* as much as to publisher location, so a small studio publishing internationally can be subject to several at once.

The European Union

The AI Act creates transparency obligations for synthetic content. Providers of systems generating synthetic audio, image, video or text must mark output in a machine-readable way so it can be detected as artificially generated. Deployers who generate or manipulate content constituting a deepfake must disclose that it has been artificially generated or manipulated, with carve-outs including artistic and satirical work where the disclosure must not spoil the display of the work, and adjustments for law enforcement.

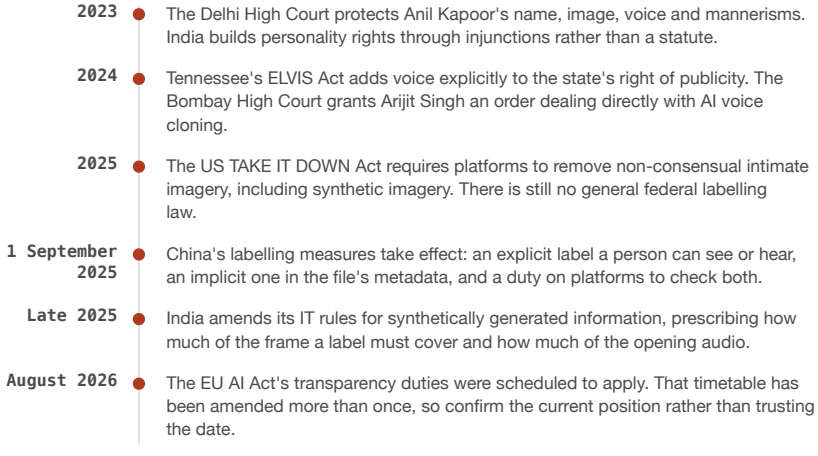
These transparency provisions apply from August 2026, with the Act's other obligations phased across earlier and later dates. Penalties are set by member states within ranges the Act specifies, and they are substantial.

China

Measures on labelling AI-generated synthetic content took effect on 1 September 2025. They require both an **explicit** label that users can see or hear and an **implicit** label embedded in file metadata. Obligations fall on generation service providers, on distribution platforms, and on users who publish. Platforms must check for the implicit label and act where content is unlabelled or falsely labelled.

This is the most comprehensive labelling regime currently in force anywhere, and it is worth knowing about even if you never publish there, because it is the closest thing to a working implementation of the idea.

Disclosure duties, as they have arrived



Three features have held across every regime so far: the duty attaches to where the audience is rather than where you are, machine-readable marking sits alongside a human-visible label rather than replacing it, and the trigger is whether a reasonable person could take the content for a real record.

The United States

No general federal statute at the time of writing. Instead:

- **State laws** on synthetic media in elections, on non-consensual sexual imagery, and on voice and likeness replicas, differing considerably between states.
- **Federal Trade Commission** enforcement on deceptive advertising, which reaches fabricated testimonials and endorsements without any AI-

specific statute.

- **Proposed federal bills** on likeness and on election content, introduced and not enacted.

Elsewhere

Several countries have election-specific rules requiring disclosure of synthetic content in political advertising within defined periods. India's information technology rules and advisories have addressed synthetic media and labelling, with proposals to tighten them. Broadcast regulators in several countries have applied existing accuracy rules to synthetic material without new legislation.

What this means practically

Identify the strictest rule that applies to your audience and build to that. For anyone publishing internationally, that currently means a visible disclosure plus preserved provenance metadata, which is a reasonable specification to adopt generally.

Political and election content is the sharpest edge. Rules here are stricter, timed to election periods, and enforced. If you take political work, read the specific rules for that election, in that country, for that period.

Advertising has its own regime. Consumer-protection law reaches deceptive synthetic testimonials and endorsements everywhere, whether or not there is a labelling statute.

Machine-readable marking is becoming a compliance requirement, not just good practice. Keep the credentials and watermarks your tools produce, and choose tools that produce them.

The gap between the law and reality

Two honest observations.

These rules bind compliant actors. The material they are aimed at — fraud, non-consensual imagery, election manipulation — is produced by people who will not comply, using open models that mark nothing. The realistic effect is to

establish norms and enable platform enforcement for the vast ordinary volume, not to stop the deliberate misuse.

And enforcement is barely begun. Most of these regimes are new, the authorities are staffing up, and how the thresholds work in practice will be established case by case over years.

A note on why the details in this lesson will date and the structure will not. Specific dates, thresholds and penalty figures change; three features have been stable across every regime introduced so far. Duties attach to the audience rather than the publisher. Machine-readable marking sits alongside human-visible labelling rather than replacing it. And the trigger is whether a reasonable person could take the content for a real record. If you build to those three, a change in the detail is an adjustment rather than a rebuild.

None of this is legal advice, and the applicable rules for your specific work in your specific country are worth checking directly, with a professional where the stakes justify it. What this lesson gives you is the map: there are duties, they differ, they attach to your audience, and they are getting stricter rather than looser. Building an honest disclosure habit now is much cheaper than retrofitting one later.

BEFORE YOU MOVE ON

A studio publishes a synthetic-voice advertisement to audiences in several countries. What follows from how these rules are structured?

A. Only the studio's home country's rules apply to the campaign, since transparency obligations of this kind attach to the publisher's place of establishment rather than to the audience.

B. The strictest applicable rule effectively sets the specification, because obligations attach to where the audience is.

C. No rules apply until enforcement authorities publish detailed guidance for synthetic advertising.

D. Advertising is exempt from synthetic-media labelling because it is already covered by consumer-protection law.

83 · 8 min

What to write down before you start

THE ONE THING TO KEEP

Because the law is unsettled, the contract is what actually allocates risk in AI-assisted client work, and the terms that matter are short, specific and better raised at the start than discovered at the end.

Why the contract carries the weight

Everything in this module has ended at "unsettled" or "depends on the country". That is not a reason for paralysis. It is the ordinary condition of commercial work, and the ordinary answer is to write down what the parties have agreed rather than to rely on a default nobody can state.

The terms below are short, and raising them makes you look like somebody who has done this before.

The clauses worth having

Disclosure of use. State that AI tools may be used and in what capacity. Some clients have policies — many public sector, pharmaceutical, financial and news organisations restrict or prohibit it — and finding that out at delivery is the expensive way.

What you warrant, and what you do not. Warrant what you can: that you have the necessary licences for the tools used, that you have carried out stated checks, that you are not aware of any infringement. Do not warrant that the work is free of any third-party right, because nobody can establish that about generated material.

Ownership and its limits. Assign whatever rights you have. Say plainly that purely generated elements may not attract copyright in some jurisdictions, and that exclusivity in those elements cannot be guaranteed. If exclusiv-

ity matters, agree the route — substantial human authorship, trademark registration, or commissioning original artwork.

Third-party material. List it: stock, references, fonts, music, generated elements, and the licence for each. This is the record that answers every later question.

Confidentiality of inputs. If the client's material will pass through a hosted service, say so, and confirm whether that service trains on inputs. For sensitive work, commit to local processing and mean it.

Likeness and voice. If any real person is depicted or heard, attach their consent. Say who obtained it and what it covers.

Disclosure obligations. Who is responsible for labelling at publication, and in what form. This is increasingly a legal duty and it is frequently nobody's job by default.

Liability cap and what is excluded. Standard commercial practice and doubly worth having here.

What happens on a claim. Who notifies whom, within what period, and who controls the response. Your vendor indemnity probably has notification conditions; make sure your client contract does not prevent you meeting them.

The conversation to have at the start

Most of this fits into one exchange at the briefing, and it takes five minutes:

- Does the client have a policy on AI use?
- Does the work need to be protectable, and in which markets?
- Will any real person be depicted?
- Where will it be published, and who is labelling it?
- Is any of the input material confidential?

The answers determine the whole approach. A client who needs a protectable exclusive mark for three markets is a different job from one who needs forty

social images by Friday, and the difference is decided before you open any tool.

The record to keep

Repeated through this course because it answers everything: model and version, prompts and seeds, reference images with sources and licences, what was done by hand, layered files, the licences and terms in force at the time, consents, and the disclosure as published.

A folder per job. Ten minutes across the whole project. It is the difference between answering a question in five minutes and reconstructing a year of work from memory.

Pricing, briefly

One practical note because it comes up immediately. Clients who know generation is involved will ask for a lower price. The honest answer is that the generation is a fraction of the work, and that what they are buying — the judgement, the composition, the finishing, the rights position, the accountability — is unchanged.

Pricing by time spent invites exactly this argument and loses it. Pricing by the value of the deliverable does not. The freelancing course in this track goes through this properly; the connection to this module is that the rights and disclosure work described here is real professional labour, it is what makes the deliverable safe to use, and it belongs in the price rather than being absorbed silently.

None of this is legal advice and none of it substitutes for a professional where the amounts justify one. What it is, is the list of things that cause disputes, written down before they do.

BEFORE YOU MOVE ON

Why should a supplier avoid warranting that AI-assisted work infringes no third-party rights?

A. Because such warranties are unenforceable in most jurisdictions and would be struck out in any event.
B. Because vendor indemnities already provide the client with an equivalent warranty directly.
C. Because it is impossible to establish, so the warranty creates an obligation the supplier cannot meet.
D. Because copyright does not subsist in generated material, so there are no third-party rights capable of being infringed.

84 · 8 min

A procedure for the cases nobody has answered

THE ONE THING TO KEEP

When the law is unsettled, a written decision procedure based on recognisability, substitution, deception and reversibility produces defensible choices without needing a resolution that does not exist.

The situation this course leaves you in

You now know the mechanisms, the failure modes, the disclosure practice and the shape of every major legal disagreement. What you do not have — because it does not exist — is an authority to consult that will tell you whether a particular thing is allowed.

That is uncomfortable and it is the actual condition of the work. Professionals in every field operate under uncertainty using procedures rather than certainty, and this is what one looks like here.

Four questions, in order

1. Is a real person recognisable?

If yes, stop and get consent — informed, specific, written, time-limited, revocable, paid where commercial — and plan the disclosure. If you cannot get consent, do not make it. This question resolves the largest category of serious harm and it resolves it first because the answer is not a balance.

2. Does this substitute for a specific person's work?

Not "does it resemble a style" but: was somebody going to be paid for this, and did you replace them by imitating them specifically? Naming a living artist, training on one person's portfolio, reproducing a distinctive identifiable manner. If yes, the legal position may be open and the professional position is not. Commission them, or do something else.

3. Would the audience change their view if they knew?

If yes, disclose. This is the test from the labelling lesson and it covers factual work, testimonials, journalism, anything presented as a record. If disclosure would defeat the purpose of the piece, that is a strong signal the piece is deceptive rather than a reason to skip the disclosure.

4. What happens if this is wrong?

Reversible and cheap — a social post, a draft, a personal project — proceed and fix it if needed. Irreversible or expensive — a printed run, a broadcast campaign, a registered mark, anything about a real person — slow down, check properly, and get advice where the amounts justify it.

Then write it down

Whatever you decide, record the decision and the reasoning, in three lines, in the job folder.

This is the step people skip and it does more work than the rest. A written contemporaneous note showing that you identified the question, considered it, and decided on stated grounds is the difference between a mistake and neg-

ligence — in a dispute, in a client conversation, and in your own judgement six months later when you cannot remember what you were thinking.

What to do when somebody objects

You will occasionally be challenged by someone who believes this technology should not be used at all. Some of them are the people whose work trained the models, and their anger is not unreasonable.

The response that works is not a defence of the technology. It is to say what you actually did: what was generated, what was not, who was paid, what was disclosed, what you declined to do. Specifics are answerable; generalities are not, and a practice you can describe in detail is one you have already thought about.

If you cannot describe it comfortably, that is information about the practice rather than about the person asking.

Where to look things up

The habit worth building is knowing what kind of question you have, because that determines who answers it.

- Copyright in your jurisdiction: the national copyright office, which publishes guidance in plain language.
- Personality and likeness: a media or entertainment lawyer, because it is not a copyright question.
- Disclosure duties: the relevant regulator, and the platform's own policy, which will usually be stricter.
- Model and service terms: the documents themselves, saved with a date.
- Anything with real money attached: a qualified professional in the right country. An hour of proper advice is cheaper than the alternative by an enormous margin.

The last word

Nothing in this course resolves the underlying disagreements, because they are not resolved. What it gives you is the ability to tell which question you are facing, to say accurately what is known and what is not, and to make a decision you can explain.

That is a lower standard than certainty and a higher one than most practice currently meets. It is also, in a field changing this fast, the only standard that will still be worth anything in two years.

BEFORE YOU MOVE ON

What is the decisive value of writing down the reasoning behind an uncertain decision?

A. It converts an unsettled legal question into a documented position that binds the other party.

B. It transfers liability to the client, who is deemed to have accepted the stated reasoning on delivery.

C. It shows you identified and considered the question, which distinguishes a judgement call from negligence.

D. It satisfies the record-keeping requirements imposed by the transparency provisions of current AI legislation.

CASE STUDY · MODULE 9

The mark the client could not own

A four-person design studio in Delhi, in the meeting where a six-lakh identity job is being agreed

Studio Sundial is Ira Bhatt and three others, working out of two rooms in Lado Sarai. The largest job they had been offered came from a two-year-old Gurugram company making a water-quality sensor for housing societies and small water plants. The company was closing a funding round and preparing to sell in the United Kingdom and the United States. They wanted a full identity: a mark, packaging, and a campaign of about forty images. The fee was six lakh forty thousand rupees, which was more than the studio had billed in the previous five months together.

The designer on the job had built the mark by generating around three hundred candidates on a hosted service, choosing one, and tidying it. The campaign images had been generated on the same service, on the studio's consumer subscription, which was what the studio had because it was what the studio could afford.

In the second meeting the founder asked the question every client asks, in the form every client asks it. We will own all of this, obviously, exclusively, everywhere.

The accurate answer is four paragraphs long and contains the words unsettled and depends on the country.

In the United States, the Copyright Office has been consistent across a series of decisions and a formal report: copyright requires a human author, prompts alone do not supply authorship however detailed or however many iterations were run, and material generated by a machine in response to a prompt is not protectable — though a human's own expressive contributions, including selection, arrangement and subsequent editing, can be. A registration for a graphic novel illustrated with generated images ended precisely there: the text and the page arrangement protected, the individual images not.

In the United Kingdom and in India there is an older provision deeming an author for a computer-generated work with no human author, being the person who made the arrangements necessary for its creation. It has barely been tested in either country, the UK has consulted on removing it, and the Indian position was left ambiguous by a registration naming an AI as co-author which the office then sought to withdraw.

In China, a court has found copyright in a generated image on the basis that the user's prompts, parameters and repeated refinement showed sufficient intellectual investment. Similar facts, opposite conclusion.

So the same mark might be unprotectable in one of the client's three markets, protected on different reasoning in another, and owned by legal fiction in the third.

Ira had to decide whether to say that, in the meeting where the fee was being agreed.

Saying yes was what every other studio the founder had spoken to had said. It is what a client expects to hear, and a long careful answer in a pitch sounds like a studio that is not sure of itself.

The cost of saying yes was not immediate and it was large. If the client registered the mark, built a brand on it and a competitor in the United States reproduced the artwork, there would be no copyright to stop them. And the founder's lawyers would read Studio Sundial's contract during funding diligence, where a warranty that the work was free of any third-party right and exclusively owned everywhere is a warranty the studio could not support.

The cost of saying the accurate thing was six lakh forty thousand rupees going to the studio that said yes.

STOP HERE

Before turning the page: what would you have done, and what would it have cost if you were wrong? Write it down. The point of the next section is to compare.

WHAT HAPPENED

She said it, in three sentences at the meeting and a page sent afterwards.

The three sentences were the jurisdictional position, stated without hedging in either direction: purely generated artwork may attract no copyright at all in the United States, is protected on different reasoning in China, and in India and the United Kingdom rests on an old provision that has hardly been tested.

The page was what to do about it, and that was the part that mattered. The mark would be redrawn as vector by a person, in Inkscape, using the generated candidate as a reference and nothing more, with the working files kept as evidence of the human work. The protection that actually matters for a brand mark is a registered trademark, which does not care how the artwork was drawn, and she costed registration in three jurisdictions. The campaign images would carry substantial human contribution — compositing, retouching, typography, arrangement — with layered files retained, because that contribution is the part that can be owned and the layered file is the proof it happened.

Two things changed in the studio's own position at the same time.

They moved to the service's business tier. The consumer tier's terms allowed inputs to be used for training, and the client's unreleased product photographs were going through it, which was a contractual problem before it was a privacy one. The tier also carried the vendor's indemnity, which did not exist on the plan they had been using — although Ira read the exclusions first, and found that it was drafted around copyright, said nothing about personality or publicity rights, required the output to be used substantially as generated, and was capped in a way that would not reach a client's own losses if a campaign had to be pulled.

And she rewrote her warranty. She warranted that she held the licences for the tools used, that she had carried out named checks — a reverse image search on every published image and a recognisability check on every face — and that she was not aware of any infringement. She struck the clause warranting that the work was free of any third-party right, because nobody can establish that about generated material and a warranty you cannot support is a liability rather than a reassurance.

She got the job. The founder told her afterwards that the page was why. Three other studios had said yes without qualification, and none of them had raised the question at all, which made him wonder what else they had not raised.

Why is trademark the right route for the mark, when copyright is the right the client asked about?

Because trademark protects a sign used in trade as an indicator of origin, and it is acquired through use and registration regardless of how the artwork was made or whether any copyright subsists in it. It gives the client what they actually wanted — the ability to stop a competitor using a confusingly similar mark in their markets — without depending on an authorship question that different countries answer differently. Copyright in the artwork would be a second, weaker and jurisdictionally unreliable layer. Routing around the unsettled question is usually better than waiting for it to settle.

Moving to the business tier bought an indemnity. What does that change, and what does it leave exactly where it was?

It transfers some risk on some facts: if a third party brings a copyright claim of the specified kind and the conditions are met, the vendor defends and pays within a cap. It leaves almost everything else. Most indemnities are drafted around copyright and exclude personality and publicity rights, which is among the most likely claims where real people are depicted. Many require the output to be used substantially as generated, which sits badly with a workflow built on compositing and retouching. Caps commonly stop at what was paid and exclude consequential loss, so a client's own losses from a pulled campaign — the media spend, the reprint — are not covered, and those are what gets claimed against the studio under its own contract.

Striking the warranty that the work is free of any third-party right sounds like a studio weakening its own offer. Why is it the stronger position?

Because it is a warranty nobody can honestly give about generated material, and giving it converts an unknowable risk into a certain contractual breach the moment anything is asserted. Warranting what can be established instead — that the tool licences are held, that named checks were carried out, that no infringement is known — is enforceable, verifiable from the job folder, and true. It also shifts the conversation to what the studio actually did, which is a defensible record, rather

than to a promise about the world. A client's lawyer reading it in diligence finds a supplier who understood the question.

MODULE 9 · TEST

Check what stuck

6 questions, one right answer each. This one is for you, not for a record: answers and the reason behind each are at the back. If two questions from the same module go wrong, re-read that module before the exam rather than after.

1 An illustrator objects that a model produces work in her recognisable style. Which route could actually reach that conduct in most jurisdictions?

A. Copyright, because style is the expression of a body of work.
B. Moral rights, because style imitation is derogatory treatment.
C. Database rights, because her portfolio was collected as a set.
D. Trademark and passing off, where a name or an endorsement is implied.

2 Under the United States position, which part of a graphic novel illustrated with generated images is protectable?

A. The written text and the author's arrangement of the pages.
B. The individual images, because the prompts were long and heavily iterated.
C. Everything, because the author made the arrangements necessary for creation.
D. Nothing, because the presence of generated material taints the whole work.

3 You fine-tune a non-commercially licensed open model and use the result for a paid client job, through a hosted interface whose terms allow commercial use. What is your position?

A. Covered, because the hosted service's terms govern the output.
B. Covered, because a fine-tune is a new work with its own licence.
C. In breach, because derived weights inherit the base model's conditions.
D. Uncertain, because non-commercial clauses are unenforceable against end users.

4 Your tool vendor offers an indemnity on its business tier. Which claim is most likely to fall outside it, and is also among the likeliest to arise?

A. A stock library's copyright claim over a near-reproduced photograph.

B. A person's claim about the use of their face or voice.

C. A claim that the output reproduced a registered typeface.

D. A rights holder's claim over a memorised training image.

5 A one-person studio in Pune publishes a synthetic presenter video aimed at viewers in the European Union and China. Whose disclosure rules apply?

A. Only India's, since that is where the work was produced.

B. None, since none of these regimes reaches an individual publisher.

C. Whichever regime the hosting platform is incorporated under.

D. Those of each audience, so the studio should build to the strictest.

6 Working through the decision procedure for an unsettled case, which question is answered first, and why is it not a balancing exercise?

A. Whether a real person is recognisable, because consent is required rather than weighed.

B. Whether the work will be irreversible or expensive, because that sets how much care is due.

C. Whether the audience would change their view if they knew, because disclosure governs everything else.

D. Whether the output substitutes for a specific person's paid work, because that is the commercial risk.

EXAMINATION

Making Things With AI — final examination

This paper covers the whole course: how a diffusion model builds a picture, how prompts and references steer it, the failures that come from its architecture rather than your wording, the editing controls that turn a generation into a deliverable, what video and three-dimensional generation can and cannot do, the speech and music pipelines and their consent and rights positions, provenance and disclosure, and the law on training, ownership, likeness and labelling. Twenty-five questions are drawn from a larger pool, so a retake is a different paper. The pass mark is 70 per cent. There is no timer: many readers are working in a second or third language, and a clock measures panic alongside understanding. If you do not pass, you may retake after thirty minutes with fresh questions, as many times as you need. Failing locks nothing: a teacher can open the next course for any learner at any time, recorded as an override and never disguised as a pass. Every question tests a mechanism rather than a definition, and the explanation shown after you submit is part of the teaching.

On the website a sitting is 25 questions drawn from this pool and the pass mark is 70%. There is no time limit, there and here. Passing opens the next course in the track.

1 `1. How the picture gets made`

Training a diffusion model consists of adding a random amount of noise to a training image and asking the network to predict the noise that was added. Which consequence follows directly from that objective?

A. The model learns to recognise which parts of a mess are mess, rather than storing what objects look like.

B. The model builds an internal library of object shapes that can be queried by name.

C. The model learns a mapping from captions to compositions, which is why prompts control layout so precisely.

D. The model learns to evaluate whether an image is realistic, which is what guidance later amplifies.

2 1. How the picture gets made

A checkpoint is said to be incapable of producing genuinely dark night scenes. The documented cause is a flaw in its noise schedule. What is the flaw?

A. The schedule spends too few steps at low noise, so dark texture is never resolved.

B. The schedule never quite reaches pure noise, leaving a trace of the training set's average brightness.

C. The schedule applies guidance more strongly at high noise, which lifts the exposure.

D. The schedule was calibrated on a latent with four channels, which cannot represent deep shadow.

3 1. How the picture gets made

Asking a model trained at 512 pixels for a 1024-wide image in a single pass frequently returns two heads or an extra torso. What is happening?

A. The model is confused about anatomy because it has seen few full-body training images.

B. The decoder tiles the latent and each tile is decoded independently.

C. Each region of the oversized latent independently looks like a plausible start of a subject, and nothing coordinates them.

D. Guidance is applied per region, so distant parts of the frame receive contradictory conditioning.

4 1. How the picture gets made

'A red cube and a blue sphere' sometimes returns a blue cube and a red sphere. Which part of the machinery makes that possible?

A. The tokeniser splits colour words into pieces that are recombined in the wrong order.

B. The sampler visits latent regions in an order that does not match the prompt's word order.

C. Guidance amplifies colour terms more than noun terms, so colours become mobile.

D. Cross-attention lets every latent cell draw on every token, and nothing ties an adjective to its noun.

5 1. How the picture gets made

You want to compare two prompts fairly, and you want to be able to reproduce whichever wins. Which sampler family should you use, and why?

A. A deterministic sampler, because it follows a fixed path from the starting noise.

B. An ancestral sampler, because the extra noise averages out over many steps.

C. Either, provided you record the step count, since both converge to the same image.

D. An ancestral sampler, because livelier texture makes differences easier to see.

6 1. How the picture gets made

You need a poster with correctly spelled text in the image. Which architecture is most likely to manage it, and on what grounds?

A. A GAN, because a single forward pass avoids the drift that accumulates over a denoising loop.

B. A flow-matching model, because a straighter path from noise to data preserves fine structure.

C. Any diffusion model at a high enough guidance scale, because guidance enforces prompt fidelity.

D. An autoregressive token model, because it generates the picture as a sequence with language machinery attached.

7 1. How the picture gets made

Why did phrases like 'trending on ArtStation' and '35mm, f/1.8, bokeh' ever work, and what does that predict about prompting for a regional garment?

A. They were instructions the model was trained to follow, so an equally precise garment name will work.

B. They co-occurred in captions with a visual style, so an object with thin caption coverage will fail.

C. They were aesthetic-score tags injected by the training pipeline, which garment names also carry.

D. They increased the effective guidance scale, which a specific garment name will also do.

8 2. Steering it: prompts and what sits around them

A generated PNG contains a text chunk with the prompt, seed, sampler and model hash. Somebody asks whether that proves the image was generated. What is the accurate answer?

A. Yes, because only generation tools write that chunk and it cannot be forged.

B. Yes for the presence of the chunk, and its absence proves the image is a photograph.

C. Very little, because the chunk is plain text that anyone can write into any file.

D. Only if the model hash matches a published checkpoint, in which case it is conclusive.

9 2. Steering it: prompts and what sits around them

Which entry belongs in a negative prompt and will do something useful?

A. 'two heads', to enforce a single head.

B. 'hand behind back', to keep the hand in view.

C. 'watermark, signature, stock photo', to suppress medium artefacts.

D. 'five apples', to prevent a sixth apple appearing.

10 2. Steering it: prompts and what sits around them

Every 'photo of a woman' from an unmodified model looks like roughly the same woman. Which explanation is correct?

A. A three-word prompt leaves everything to be filled in, and what fills it is the mode of the corpus.

B. The model has a small number of face templates that it selects between by seed.

C. Short prompts are padded to the token limit with the training set's most common words.

D. The autoencoder compresses faces more aggressively than other content, converging them.

11 2. Steering it: prompts and what sits around them

A writer prompting in Devanagari finds that fifteen words exhaust a 75-token budget. What is the cause, and what is the honest workaround today?

A. The model has no Devanagari coverage, so the workaround is transliteration into Latin script.

B. Tokenisers are trained on heavily English corpora, so the workaround is to prompt the scene in English and supply the rest as references or a fine-tune.

C. Non-Latin scripts are encoded twice for compatibility, so the workaround is to disable the second pass.

D. The encoder reserves half its positions for a translation buffer, so the workaround is a shorter prompt.

12 2. Steering it: prompts and what sits around them

You supply a reference image to keep a composition and complain that 'the reference did nothing' — the layout changed completely while the palette carried over. Which conditioning route did you use?

A. A depth map, which constrains arrangement and not colour.

B. Image-to-image at low strength, which preserves the original closely.

C. A subject reference, which carries identity into new scenes.

D. A style adapter, which injects palette and texture alongside the text conditioning.

13 2. Steering it: prompts and what sits around them

Why is the top of a reference method's strength range the setting most likely to cause a rights problem as well as a dull image?

A. The output becomes the reference with a light coat of paint, and is recognisably the source.
B. High strength disables the model's output safety classifier.
C. High strength writes the reference image into the output file's metadata.
D. High strength causes the model to memorise the reference for future generations.

14 2. Steering it: prompts and what sits around them

A series of images must match across two years. Which measure actually delivers that, and why?

A. Recording the prompt, seed and parameters for every image in a version-controlled text file.
B. Using the same hosted service throughout and keeping the same subscription tier.
C. Keeping local copies of the checkpoint files the work depends on.
D. Exporting every image as PNG so the generation record survives.

15 3. Where it breaks, and why

Hands and misspelled signage were largely fixed by better encoders, higher working resolution, recaptioning and scale. What does the pattern of what was fixed predict about what has not been?

A. Failures that were about training coverage got fixed; failures that are structural did not.
B. Failures in the foreground got fixed; failures in the background need higher resolution still.
C. Failures involving people got fixed; failures involving objects were never prioritised.
D. Failures at small scale got fixed; failures at large scale require larger latents.

16 3. Where it breaks, and why

Generated images read as advertising even when the subject is mundane. Two mechanisms compound to produce that. Which pair?

A. Upscaler artefacts and the autoencoder's loss of fine texture.

B. High step counts and deterministic samplers, which both converge on a single answer.

C. Under-specified prompts filling in with the mode, and guidance driving regions toward the prototypical.

D. Recaptioning, which standardised descriptions, and deduplication, which removed unusual images.

17 3. Where it breaks, and why

A service silently appends demographic terms to prompts and produced well-publicised absurdities on historical requests. What was the failure?

A. The terms were drawn from an out-of-date list of categories.

B. A blunt textual rule was applied to every request without knowing which requests it suited.

C. The injection happened after the text encoder, so the terms were never properly conditioned.

D. The aim of broadening the range of people depicted was itself misconceived.

18 3. Where it breaks, and why

You wait for a generation, pay for it, and receive a black frame or a 'content not available' message. Which filtering stage did that?

A. The prompt blocklist, which matched a string before anything ran.

B. Named-entity blocking, which refused a public figure by name.

C. The output classifier, which checked the image after it was generated.

D. The prompt classifier, which judged the request's intent.

19 3. Where it breaks, and why

Safety classifiers produce false positives at elevated rates on medical illustration, life drawing and traditional dress from several regions. What is the common cause?

A. These categories are rare in training data, so the classifier defaults to refusing.

B. These categories contain high-contrast skin tones that trigger the detector's colour heuristics.

C. These categories are deliberately restricted by policy in most services.

D. The classifiers were trained on labelled examples whose labels carry a labelling operation's cultural defaults.

20 3. Where it breaks, and why

Why is 'I can always tell whether an image is generated' unreliable, even for someone who has looked at thousands?

A. The convincing ones never register, and the remaining tells are a short job for anyone intending to deceive.

B. Screen calibration differs between devices, so artefacts are not consistently visible.

C. The tells vary by model, so accuracy depends on knowing which model produced the image.

D. Human visual memory for texture decays quickly, so comparisons across time are unreliable.

21 3. Where it breaks, and why

Why does a generated image often survive at social-media size and fall apart at print size?

A. Print reproduction applies a colour profile that exposes compression artefacts.

B. Structures one or two pixels wide are a fraction of a latent cell, so the decoder returns their statistical impression.

C. Upscaling to print size introduces tiling seams that are invisible on screen.

D. Print requires a higher bit depth than the generation pipeline produces.

22 4. Control: getting the picture you actually need

Five sequential instruction edits on one image leave the face subtly changed, a sign reworded and fine texture softened. Why?

A. The instructions are concatenated, so later ones are conditioned on earlier text.

B. Each edit increases the denoise strength applied to the whole frame.

C. Most instruction editors re-render the whole image, so each pass is a copy of a copy.

D. The editor stores only the most recent version, discarding detail from earlier saves.

23 4. Control: getting the picture you actually need

Ten image-to-image passes at strength 0.3 do not equal one pass at a much higher strength. What do they produce instead?

A. An image with contrast climbing and detail smoothing, drifted toward the distribution average.

B. An image identical to the input, because low strength changes nothing.

C. An image with accumulated tiling seams from the repeated decode steps.

D. An image that converges on the prompt, because each pass adds prompt influence.

24 4. Control: getting the picture you actually need

Many models outpaint upward and sideways better than downward. What is the most likely explanation?

A. The latent grid is processed top to bottom, so the lower rows have less context.

B. Photographs are cropped at the bottom more consistently than at the top, so training data is thin on what continues below a frame.

C. Gravity cues make downward extension physically harder to render plausibly.

D. Mask feathering behaves differently along the bottom edge in most implementations.

25 4. Control: getting the picture you actually need

One earring must match the other exactly. Inpainting gets close and rarely gets exact. What is the reliable fix?

A. Raise the conditioning strength so the model reproduces the visible earring precisely.
B. Inpaint both earrings in one masked pass so the model sees them together.
C. Copy the good earring, flip it, place it and blend, in an ordinary image editor.
D. Train a small fine-tune on photographs of that earring and regenerate the region.

26 4. Control: getting the picture you actually need

In a production character workflow, the last step is to colour-match every shot to one reference. Why does that step matter more than people expect?

A. Colour matching normalises the latent statistics, which reduces face drift on subsequent passes.
B. A large part of what reads as 'the same character' is consistent colour and light rather than exact geometry.
C. Consistent grading is required for the face-detection stage of the detail pass to work.
D. Matched grading is what allows shots to be composited without visible edges.

27 4. Control: getting the picture you actually need

A 12-billion-parameter model needs about 24 GB of video memory at half precision. Which measure most directly brings it within reach of an 8 to 12 GB card?

A. Tiled decoding, which removes the autoencoder's memory spike at high resolution.
B. Sequential offloading, which moves parts of the model between system and video memory.
C. Reducing the step count, which shortens the time the weights are resident.
D. Quantisation, which stores the weights at 8-bit or 4-bit instead of 16.

28 4. Control: getting the picture you actually need

In a finishing ladder, why does the masked detail pass on faces come before the upscale rather than after?

A. Upscaling a wrong face gives you a large wrong face, and the upscaler cannot know it was wrong.
B. Upscalers overwrite mask boundaries, so any later detail pass would leave a seam.
C. A detail pass at higher resolution would exceed the model's native size and duplicate the subject.
D. Face restoration models only accept inputs below a certain resolution.

29 5. Time and space: video and three dimensions

Which camera instruction is most likely to be followed accurately by a video model, and why?

A. 'The camera swoops around the subject and rises to reveal the city', because it is descriptive and unambiguous.
B. 'Cinematic camera movement', because it names the register the model was trained on.
C. 'Slow push in', because standard film vocabulary is what captioners and stock libraries actually use.
D. 'Move left', because short instructions leave less room for misinterpretation.

30 5. Time and space: video and three dimensions

Demonstration reels for video models are full of slow pushes on static subjects. What does the motion-strength control actually trade?

A. Low motion hides temporal weaknesses; high motion buys movement and morphing, multiplying limbs and reorganised backgrounds.
B. Low motion reduces the frame count, so clips are shorter but more stable.
C. High motion increases the attention window, which improves long-range consistency.
D. Low motion applies the prompt for fewer steps, so the subject is rendered more accurately.

31 5. Time and space: video and three dimensions

Image-to-video is the dominant professional route. What does supplying a first frame pin, and what does it leave open?

A. It pins the motion and leaves the composition to the model.

B. It pins the whole clip, since every later frame is conditioned on it.

C. It pins composition, appearance, colour and lighting at frame one, and leaves what happens next open.

D. It pins the subject's identity across the clip and leaves the camera free.

32 5. Time and space: video and three dimensions

You need a specific camera move and a specific action, repeatably. Which route gives you both, and what does it not give you?

A. Text prompting with film vocabulary, which gives motion but not appearance stability.

B. Video-to-video from a phone clip or a Blender render, which pins motion and geometry but not appearance stability.

C. Reference conditioning on a still, which pins identity but not the camera.

D. First-and-last-frame conditioning, which pins the ends but not the camera path.

33 5. Time and space: video and three dimensions

A vendor describes its video model as a world simulator. Which single test is most informative about that claim?

A. A close-up of water or fire, since fluid dynamics are the hardest thing to fake.

B. A long clip of a walking figure, since gait reveals whether the model understands bodies.

C. A rapid camera move through a complex scene, since that stresses the attention window.

D. An object passing behind a larger one and re-emerging.

34 5. Time and space: video and three dimensions

You are choosing between two video models for a team that will pay for one. Why rename and shuffle the clips before scoring them?

A. To remove expectation effects, which do a great deal of work in judging generative output.
B. To prevent the scorer from counting how many attempts each model needed.
C. To ensure the clips are scored in the same order they were generated.
D. To stop file metadata from revealing the generation parameters.

35 5. Time and space: video and three dimensions

A client asks for a synthetic presenter using a staff member who signed a standard on-camera release in 2019. What is the correct position?

A. The release covers it, since it grants use of the recorded footage in any medium.
B. It covers creation but not distribution, so a second release is needed at publication.
C. No release is needed at all, since the output is generated rather than recorded.
D. The release predates the capability and does not cover a digital replica, so separate consent is required.

36 6. Voice and speech

Which prosodic failure most reliably reveals that a passage was read by a machine?

A. A pause placed inside a noun phrase rather than at the clause boundary.
B. A slightly faster reading rate than a human narrator would use.
C. Consonants that are clipped at the start of words.
D. A voice that sounds younger than the character it is reading for.

37 6. Voice and speech

You want a synthetic narration delivered conversationally rather than as a formal reading. Which control works best?

A. Raising the sampling temperature until the delivery loosens.
B. Prefixing the text with adjectives describing the desired tone.
C. Supplying a short reference clip in the delivery you want.
D. Generating several takes and choosing the most expressive.

38 6. Voice and speech

A voice agreement says the recordings will be deleted if the speaker withdraws consent. Why is that insufficient for a fine-tuned voice?

A. Deletion cannot be verified, so the clause is unenforceable in practice.
B. Fine-tuning produces a durable model artefact that exists separately from the recordings.
C. Recordings must be retained for a statutory period, so the clause cannot be honoured.
D. The speaker embedding is derived at generation time and is not covered by any deletion clause.

39 6. Voice and speech

A vendor advertises support for a hundred languages. Which question separates real support from a headline?

A. How many hours of audio were used in total across all languages?
B. Which languages were included in the most recent training run?
C. Is the model multilingual or a set of language-specific models?
D. What is the measured error rate on this language, on spontaneous speech, in ordinary conditions?

40 6. Voice and speech

A recognition system garbles both halves of every sentence in which a speaker mixes Hindi and English. Why is that the expected behaviour?

A. Most systems are built around one language per utterance, so a switch is an error condition.

B. Hinglish has no standard orthography, so the decoder cannot choose a spelling.

C. The acoustic model normalises pitch differently for the two languages.

D. The language identification step runs once and cannot be re-run mid-utterance.

41 6. Voice and speech

A service embeds an inaudible watermark in the speech it generates. What does that achieve?

A. It prevents the generated audio from being used to impersonate anyone.

B. It allows a listener to verify that a recording is of a real person.

C. It records that audio came from that service, and binds nobody who does not want to be marked.

D. It survives re-recording through a microphone, which is the main evasion route.

42 6. Voice and speech

Consent for a cloned voice must be informed, specific, written, time-limited and revocable. Which shortcut fails which condition?

A. A perpetual grant fails 'time-limited', and a blanket release fails 'specific'.

B. A verbal agreement fails 'informed', and a paid agreement fails 'revocable'.

C. A recorded conversation fails 'written', and a narrow grant fails 'informed'.

D. A signed release fails 'revocable', and a dated agreement fails 'specific'.

43 7. Music and sound

A generated track has a metallic, watery quality on sustained sounds. Which generation approach does that point to?

A. Token-based generation, where structural drift produces a smeared sustain.

B. Spectrogram diffusion, where phase must be reconstructed by a vocoder.

C. Token-based generation at a reduced number of residual codebooks.

D. Spectrogram diffusion at a sample rate above 32 kHz.

44 7. Music and sound

How can you demonstrate to yourself in four minutes that a music model produces texture rather than form?

A. Generate two tracks from the same prompt and check whether they differ.

B. Generate a track and count how many distinct instruments appear.

C. Generate a two-minute track and listen once from the start and once from ninety seconds in.

D. Generate the same track at two quality tiers and compare the transients.

45 7. Music and sound

A separated vocal stem comes back watery and full of holes. Before blaming the separation model, what should you check?

A. Whether the source was a lossless file or audio captured from a video platform.

B. Whether the model was run at its highest number of output stems.

C. Whether the track was mixed in mono, which prevents spatial separation.

D. Whether the vocal was recorded with reverb, which the model cannot process.

46 7. Music and sound

A single generated impact sounds thin on its own. Which technique most reliably makes it convincing?

A. Raising the generation quality tier so more residual codebooks are produced.
B. Applying compression and a short reverb to place it in the scene.
C. Layering a low thump underneath and a short bright transient on top.
D. Generating a longer version and trimming it to the required length.

47 7. Music and sound

A vendor says its music model was 'trained on licensed data'. Which follow-up question most usefully narrows what that means for you?

A. How many hours of audio were licensed?
B. Does the licence cover the output as well as the training?
C. Which collecting societies were involved in the arrangement?
D. Were any artists able to opt out of the training set?

48 7. Music and sound

A legitimate small artist uploads a hundred generated tracks in one week and hits identity checks, thresholds and generated-content flags. Why?

A. Distribution platforms cap uploads to protect catalogue quality.
B. Generated audio is watermarked, and bulk marks trigger automatic review.
C. The pattern is indistinguishable from the streaming-fraud pattern the policies were written for.
D. Metadata fields for AI involvement cannot be filled in for batches.

49 7. Music and sound

For most projects, what is generated music actually competing with, and what question decides whether it fits?

A. A commissioned composer, and the question is whether the budget allows one.

B. Silence, and the question is whether music is needed at all.

C. Library music, and the question is whether the music must be specific to this project or merely appropriate to it.

D. A staff musician, and the question is whether the organisation employs one.

50 8. Provenance, detection and disclosure

Four independent reasons make detection unreliable. Which one applies specifically because the detector is available to the public?

A. Adversarial pressure, since anyone can test against it and adjust until they pass.

B. Fragility, since public detectors are run on re-compressed uploads.

C. Generalisation, since public detectors are trained on public models only.

D. Base rates, since public detectors are applied to much larger populations.

51 8. Provenance, detection and disclosure

Why do model developers care about watermarking their own output, quite apart from labelling?

A. It lets them demonstrate compliance with disclosure law in every jurisdiction.

B. It lets them prove authorship of generated images in a dispute.

C. It lets them charge differently for marked and unmarked output.

D. It lets them keep their own output out of the next training corpus.

52 8. Provenance, detection and disclosure

Which of these most commonly destroys a Content Credentials chain in ordinary use?

A. Cropping the image, because the hash no longer matches the content.
B. A screenshot, because it is a new file with no history at all.
C. Colour correction, because it alters the pixels the manifest binds to.
D. Uploading to a platform that preserves credentials but re-encodes the file.

53 8. Provenance, detection and disclosure

A verification routine ends with three possible verdicts. Which set is correct, and which is the most common?

A. Verified, debunked and unverified, with unverified the most common.
B. Verified, likely fake and confirmed fake, with likely fake the most common.
C. Authentic, manipulated and generated, with manipulated the most common.
D. Verified and debunked, with debunked the most common.

54 8. Provenance, detection and disclosure

In a written verification routine, looking at the image for artefacts is the last step rather than the first. Why is the ordering deliberate?

A. Artefacts are easier to see once the file has been preserved and copied.
B. Inspection primes the reviewer, which biases the later searches.
C. Inspection is the weakest evidence available and the ordering stops it anchoring the conclusion.
D. Most artefacts are only interpretable once the source and place are known.

55 8. Provenance, detection and disclosure

A platform label marks your video as containing synthetic content. Why is that insufficient if disclosure matters to you?

A. Platform labels are applied automatically and are frequently wrong.
B. Platform labels do not satisfy any legal disclosure duty.
C. Platform labels are visible only to logged-in viewers.
D. The label disappears when the file is downloaded and reposted elsewhere.

56 8. Provenance, detection and disclosure

Applying the interpretive threshold for disclosure, which of these clearly requires it?

A. A distracting bin removed from the background of a corporate photograph.
B. An upscaled illustration used decoratively on a product page.
C. Sound design in a documentary presented as recorded on location.
D. Colour correction applied across a set of event photographs.

57 9. Ownership, law and the parts nobody has settled

Courts in the United States have reached different conclusions on training claims. What has been doing most of the work in distinguishing them?

A. The size of the model and the number of works in the training set.
B. Whether the developer is a commercial or a research organisation.
C. How the material was acquired, and what market effect was proved.
D. Whether the model can reproduce any training work on demand.

58 9. Ownership, law and the parts nobody has settled

The United Kingdom's provision for computer-generated works deems an author. Who is it, and what is notable about the provision?

A. The person who made the arrangements necessary for creation, with a fifty-year term, barely tested.
B. The operator of the software, with the ordinary life-plus-seventy term, applied routinely.
C. The developer of the model, with a twenty-five-year term, recently enacted.
D. The commissioner of the work, with a fifty-year term, harmonised across the European Union.

59 9. Ownership, law and the parts nobody has settled

A Chinese court found copyright in an AI-generated image on facts that would likely fail in the United States. What was the court's reasoning?

A. That the model developer had assigned its rights to the user by contract.
B. That the image was registered before publication, which creates a presumption of authorship.
C. That computer-generated works are protected by a specific statutory provision.
D. That the user's prompts, parameter choices and repeated refinement showed personalised intellectual investment.

60 9. Ownership, law and the parts nobody has settled

In the European Union and the United Kingdom, which instrument protecting a person's face and voice is most often overlooked?

A. Moral rights, which attach to the depiction as a work.
B. Data protection, under which a face or voice used to identify someone is biometric data.
C. Database rights, where a set of images of one person has been compiled.
D. Performers' rights, which extend to any depiction of a performer.

61 9. Ownership, law and the parts nobody has settled

A client contract requires confidentiality of all materials supplied. You are using a consumer-tier service whose terms allow inputs to be used for training. What is the position?

A. Acceptable, provided the material is deleted from the service after the job.

B. Acceptable, because training on inputs does not reproduce them.

C. Acceptable if the client is told after delivery which tools were used.

D. Inconsistent: the two documents cannot both be honoured.

62 9. Ownership, law and the parts nobody has settled

Which warranty can a studio honestly give a client on AI-assisted work?

A. That the work is free of any third-party right in every market of publication.

B. That the work is original and exclusively owned by the client on delivery.

C. That the tool licences are held, stated checks were carried out, and no infringement is known.

D. That the vendor's indemnity will cover any claim arising from the generated elements.

63 9. Ownership, law and the parts nobody has settled

China's labelling measures require two labels on generated content. What are they, and why two?

A. An explicit label a person can see or hear, and an implicit label in the file's metadata.

B. A label at the start and a label at the end, so it survives partial reposting.

C. A label from the generating service and a second from the distributing platform.

D. A visible label and a registered declaration filed with the regulator.

Answers

Each entry gives the correct option, then what each wrong one reveals about how somebody was thinking. The second part is worth more than the first: a wrong answer here is a named misreading, not a mistake.

Lesson checks

1. What the model is actually doing — C

A — Treats the seed as fixing the content of the image rather than only the random starting noise.

B — Assumes the seed alone determines the result, forgetting the prompt steers every step.

D — Imagines the noise already contains a layout that the prompt merely decorates at the end.

2. Noise, steps and the schedule — A

B — Invents a hard internal limit; the loop genuinely runs sixty times, and the results are near-identical because the numerical path has converged, not because work was thrown away.

C — Treats steps as time the model fills with prompt content; step count controls how finely the denoising path is approximated and is independent of prompt length.

D — A plausible-sounding safety story with no basis; the steps really run, which is exactly why the slower version cost three times as much.

3. The small space the picture is built in — B

A — Confuses two different limits; truncation would drop the request entirely, whereas here the model clearly attempted a sign and produced letter-shaped marks.

C — Contradicts the evidence in the image; letter-like marks appear precisely because the model has seen a great many signs.

D — Attributes an ordinary resolution artefact to moderation; safety filters block or blur whole outputs, they do not selectively degrade small type.

4. How your words become a steering signal — B

A — Would predict a consistent bias toward brown bottles across seeds, whereas attribute swapping varies run to run and affects both objects symmetrically.

C — Truncation drops content entirely; here both colours appeared, merely on the wrong objects.

D — Misplaces the cause in time; the colour assignment is settled in the early high-noise steps by attention, not shuffled by the sampler at the end.

5. Guidance: the dial that makes it obey — C

A — Steps refine an already-converged path; they do not pull latent values back inside the range the decoder can render.

B — Text nudges content, but the clipping comes from extrapolation in the guidance arithmetic, which prompt words cannot undo.

D — Generating below native size causes composition problems and softness, not clipped highlights and waxy skin.

6. Seeds, samplers and reproducibility — D

A — Half right and half wrong; seeds are indeed model-specific in effect, but the same seed on the same model and machine reproduces the noise exactly.

B — No such behaviour exists; the seed is used verbatim, and the reported seed after generation is the one that was used.

C — Overstates it; ancestral samplers do reproduce when seed, sampler and step count all match, so this alone would not explain the difference.

7. What the training run actually consisted of — C

A — No such correction exists; the disparity comes from the raw caption distribution, not from an added mechanism.

B — Complexity is not the constraint; the model renders comparably intricate Western garments well, so the difference lies in training coverage.

D — Tokenisers split unfamiliar words into sub-pieces rather than discarding them; the word reaches the model, but with little learned meaning attached.

8. Diffusion is not the only way — A

B — Guidance amplifies the prompt's overall direction; it has no mechanism for enforcing spelling, and high values degrade the image.

C — GANs are fast but narrow; a face or texture GAN has no general text-rendering capability at all.

D — Resolution helps a model that already forms correct letters; it does not give a model the ability to spell.

9. What the model does and does not retain — C

A — Contradicts the parameter arithmetic; near-exact recovery is confined to a small set of heavily duplicated images, not available for the corpus at large.

B — No such retrieval step exists in a standard diffusion pipeline; the image is produced from noise and weights alone.

D — Invents a cache; truncation drops tokens and changes what is generated, but there is nothing to fall back on.

10. Writing a prompt that gets what you meant — A

B — Credits the model with understanding negation and blames the data instead of the mechanism.

C — Assumes a known weakness has been fully solved by newer architectures rather than merely reduced.

D — Believes adding detail changes how the model handles semantics rather than just what it draws.

11. Tag soup or sentences: why the advice reversed — B

A — Size does not create a need for more words; the mismatch is in the form of the text, and adding more keywords usually makes this case worse rather than better.

C — Safety filters block or refuse prompts outright; they do not silently strip style words and leave a working generation behind.

D — Truncation drops the end of a prompt; here the subject appears but reads as vague, which points at prompt form rather than length.

12. What a negative prompt really does — B

A — Sounds mechanical but is wrong about the cause; the step count is irrelevant, and the negative conditioning would be processed normally at any guidance above 1.

C — Places the negative prompt in the wrong stage entirely; it acts during the denoising loop, never after decoding.

D — Influence per step scales with guidance, not with the number of steps; at scale 4 a four-step model shows a strong negative-prompt effect.

13. Weighting, emphasis and where it breaks — C

A — Invents a cap; the scaled vector is genuinely passed through, which is precisely why the output degrades rather than reverting to the unweighted result.

B — Weighting scales an already-computed embedding; it cannot change which token was looked up or push it out of the vocabulary.

D — Weights are numbers applied to existing tokens; they consume no additional positions in the sequence and cannot cause truncation.

14. The token limit that eats long prompts — B

A — No such partial-conditioning failure exists; the encoder runs to completion in milliseconds and either truncates or chunks the input.

C — Low guidance weakens the whole prompt evenly; it does not produce a clean cut-off after the first third.

D — Would give vague or approximate renderings of those concepts rather than their complete absence, and would not fall so neatly at a length boundary.

15. When a reference image beats any sentence — C

A — High denoise strength is precisely the setting that discards the original layout; low strength would keep the layout but also keep the style.

B — Confuses the two roles; a style adapter transfers palette and texture, and would import your sketch's rough look rather than its geometry.

D — Subject reference carries identity, such as a particular face or product, and has no mechanism for enforcing a layout.

16. Reading the metadata, keeping the record — A

B — Overstates it in the other direction; tools can and do write parameters into other formats, and the real weakness is the absence of any signature rather than the file type.

C — Compression affects pixel data, not text segments; the block can survive a deliberate export.

D — True but beside the point; irreproducibility across hardware says nothing about whether the metadata was authentic in the first place.

17. The prompt you sent is not the prompt it got — C

A — Ancestral sampling varies texture within the same composition; it does not produce entirely different scenes from the same conditioning.

B — Model size does not affect determinism; a large model with fixed conditioning and a fixed seed is exactly as reproducible as a small one.

D — No mainstream service serves other users' generations as your result, and cached delivery would produce repetition rather than the variation described.

18. A prompt is not an asset — B

A — A cynical explanation that is unnecessary; even a complete, honestly supplied prompt fails once the model, sampler or guidance differ.

C — Resolution differences change sharpness, not subject, composition or style, which is what usually differs.

D — Tokenisation differences are minor; the dominant factor is the model and settings behind the prompt.

19. What hands and text told us — D

A — Assumes visual complexity is what makes a task hard, when dense texture is exactly the model's strength.

B — Treats hands as a permanent special weakness rather than one instance of a broader category that got fixed.

C — Assumes faces are hard because they are semantically loaded and viewers are sensitive to them.

20. Counting, and the objects that swap their attributes — A

B — Repeats the number in a system with no counter; the negative anchor also cannot express a quantity, so this changes the image without addressing the count.

C — Resolution affects how well each bottle is rendered, not how many of them the process settles on.

D — A better encoder represents the numeral more faithfully, but there is still no mechanism downstream that enforces a quantity.

21. Why "without" summons the thing — C

A — Attributes a general architectural limitation to a rewriting layer; the effect appears identically on local models with no such layer.

B — Nothing strips them; the tokens are encoded and do influence the result, which is precisely the problem, since they make the concept more present.

D — Invents a specific misreading; the mechanism is the general one that concepts named become more probable, not a special meaning attached to the word.

22. Left, right, behind: the relations it cannot hold — A

B — Reverses the pattern; vertical relations between dissimilar objects are the best-supported case and depth relations between similar objects the worst.

C — Encoders do parse better than they used to, but parsing the preposition is not the same as the image process enforcing the geometry.

D — Prompt length is not the operative variable here; the type of relation and the similarity of the objects are.

23. The pull toward the average — D

A — Backwards; higher guidance strengthens the pull toward the prototypical, which is the cause of the effect rather than a cure for it.

B — Steps refine a converging path; they do not move the result away from the centre of the distribution.

C — Adds detail to the same prototypical scene; the tidiness and the lighting are set by the distribution, not by the pixel budget.

24. The bias, and what mitigating it actually does — C

A — Presumes one particular answer to the contested question of what the target should be, and a vendor could answer it while having measured nothing.

B — A reasonable question, but a training-stage fix can still be unmeasured and ineffective; the method matters less than the evidence.

D — Confuses mitigation with refusal; blocking the prompt is not a way of addressing what the model produces for it.

25. Reflections, shadows and things that must agree — B

A — No such separate pass exists; the whole latent is denoised together, which is why the failure is one of consistency rather than of assembly.

C — Memorisation is rare and applies to heavily duplicated whole images; ordinary inconsistent reflections are generated, not retrieved.

D — Resolution governs how finely the reflection is rendered; it does not determine whether the reflection depicts the correct objects.

26. Faces in crowds and other small things — C

A — Exceeding native size in one pass typically introduces duplicated subjects, and the cell budget per face rises only in proportion anyway.

B — Text conditioning cannot change how many latent cells a region occupies, which is what determines whether features can be represented at all.

D — Better binding puts the right content in the right place; it does not create the spatial capacity that small regions lack.

27. Refusals, filters and the false positives — C

A — Gets the stage wrong; clinical terminology is rarely on a blocklist, and everyday language for the same subject is more likely to trigger a classifier, not less.

B — Assumes a categorical policy where the usual cause is a classifier's error rate on skin; most services permit medical illustration in principle.

D — Named-entity blocking applies to specific people by name and has no bearing on generic clinical illustration.

28. The tells, and why they are disappearing — A

B — Hands are not geometrically simpler; the difference is between a coverage limit and a spatial-capacity limit, not a difficulty ordering.

C — Detail passes on faces and hands do exist as optional workflows, but the improvement in hands came from training, and no service excludes backgrounds by policy.

D — Prompt vocabulary never fixed hands; the improvement arrived with new model releases regardless of how the prompt was written.

29. Editing what you already have — B

A — Names a real cause of soft inpainting results, but one that would affect close-ups equally.

C — Imagines the model splits a fixed budget of effort across the things named in the prompt.

D — Reaches for a training-data explanation when the cause is a mechanical property of how the crop is processed.

30. Denoise strength: the number that decides how much survives — C

A — No such precedence rule exists; the two are combined, and the balance is set by where on the schedule the run begins.

B — Resampling softens detail; it does not discard the composition, which survives fine at lower strengths with the same resolution mismatch.

D — Strength and guidance are independent settings; raising one does not change the other.

31. Masks, feathering and the context window — A

B — Feathering controls how visible the seam is; it has no bearing on which direction the model believes the light comes from.

C — Lower strength retains the original pixels, which is the thing you were trying to replace; the lighting error would return with the face.

D — Prompting the light direction is a weak and unreliable substitute for letting the model see the surrounding scene.

32. Conditioning on structure: depth, edges and pose — D

A — Pins every contour including photographic noise, producing a photograph with woodcut texture rather than a woodcut.

B — Pose carries only a figure's joints; it says nothing about the rest of the composition and does not apply to a scene without people.

C — At that strength almost nothing of the input survives, so the composition you wanted to keep is discarded.

33. Teaching it something it does not know — C

A — Resolution affects fidelity of detail, not whether the background was learned as part of the concept.

B — Rank affects capacity and can contribute to over-fitting, but the specific symptom here points at the images and captions rather than the rank.

D — Lowering the strength would weaken the effect, but the background was bound to the concept during training and would still be present.

34. The same face twice — B

A — There is no accumulation across generations; each panel is conditioned independently from the same source image.

C — Identity adapters work across different seeds; that is the point of them.

D — Adapters inject conditioning during the denoising loop rather than swapping pixels afterwards, which is why the result is lit and posed correctly.

35. What an upscaler invents — A

B — Face-specific training makes the invention more convincing, not more truthful; there is no additional information in the source to recover.

C — The factor changes how much is invented, but even a modest factor invents rather than recovers, so no threshold makes the result evidential.

D — Low denoise limits drift from the input; it does not turn generated detail into recorded detail.

36. Compositing is still the job — B

A — Resolution can be reached by upscaling either way; the decisive advantage is control over structure, not pixel count.

C — Working on pixels is not inherently more capable; a low-denoise generative pass over a composite is a standard and useful final step.

D — The generated elements carry exactly that bias; the grade and assembly reduce how much it dominates, but they do not avoid it.

37. What runs on the machine you have — C

A — Quantisation does not change the step count; steps affect time, not the memory needed to hold the weights.

B — Output resolution is a separate setting; a quantised model generates at the same resolutions as the original.

D — That describes pruning, which is a different technique; quantisation keeps every parameter and stores each one less precisely.

38. Video, and what it still cannot do — C

A — Locates the limit in how much instruction the model can take in, rather than in what it can hold together.

B — Assumes the binding constraint is price, when roughly the same number of frames gets generated either way.

D — A true point about editing craft that skips the prior question of why the footage would be wrong at all.

39. Why a second of video costs what it does — C

A — Neither 49 nor 81 divides evenly by 24, 25 or 30, so this cannot be the reason.

B — Memory sets an upper bound but would not produce these particular values, which appear regardless of the hardware available.

D — No such central-anchor mechanism exists; attention operates across the whole window rather than around a designated middle frame.

40. What holds a moving picture together — D

A — Would produce an abrupt discontinuity in everything at once, whereas this is a single element changing while the shot continues.

B — Prompt specificity influences what appears initially; it provides no mechanism for holding an element stable over time.

C — Temporal denoising smooths high-frequency flicker; it does not substitute one garment for another.

41. Conditioning a clip: stills, frames and video in — B

A — No blending setting is involved; the model generates intermediate frames, and morphing is what it produces when no plausible motion connects the ends.

C — A resolution mismatch is rejected or resampled by the pipeline; it does not produce a smooth morph between recognisable states.

D — Would produce unrelated content in the middle rather than a continuous melt between the two supplied states.

42. Directing motion, and what actually lands — C

A — A compound, invented description; the model will latch onto fragments of it, and the several moves compete with one another.

B — Register words with no specific motion vocabulary; this sets a mood and leaves the actual movement entirely undetermined.

D — Two simultaneous, opposed motions in one shot is exactly the case where the model tends to apply one and drop the other.

43. Why clips get worse as they get longer — C

A — Sharpening would increase apparent detail rather than soften it, and the colour shift would not follow.

B — Temporal compression is constant across every chunk and does not vary with position in the sequence.

D — The window is a fixed property of the model; it does not shrink as more chunks are generated.

44. Talking heads, lip sync and the consent line — D

A — A useful measure and not the missing piece; the gap is what the audience is told, not what is embedded in the file.

B — A licensing detail that may or may not apply; it is not the ethical or legal requirement that consent leaves unaddressed.

C — Treats consent as covering the audience as well; the presenter cannot agree on the viewer's behalf to be misled about what they are watching.

45. Testing a claim that it understands physics — B

A — Style consistency is a temporal-attention property and is maintained by models with no object representation whatsoever.

C — Smearing comes from temporal compression in the autoencoder and says nothing about object representation.

D — Similarity scores reward appearance, which is precisely what a locally plausible model is good at; the disagreements it would miss are the informative ones.

46. Testing a video model honestly — C

A — Possible in principle and rarely the explanation; a strong reel is consistent with a low hit rate without any difference in the model.

B — Prompt quality matters, and a battery of ordinary professional prompts is the relevant test, since that is what the work consists of.

D — Worth checking, and it would not account for the gap on the conditioned prompts in the battery, which are tested the same way.

47. What synthetic video does to believing anything — D

A — True of most measurements and not the specific problem; the decisive issues are generalisation and base rates.

B — Thresholds are a routine part of using any classifier; setting one does not resolve the underlying difficulty.

C — Compression weakens the signal substantially rather than eliminating it, so this overstates a real problem.

48. Generated objects and the space around them — B

A — Resolution affects the quality of the splat; no capture resolution turns a radiance representation into geometry.

C — Splat files open in several desktop applications; the limitation is what they contain, not where they can be loaded.

D — A real capture problem with a real fix, and it would leave a hole rather than making the asset unusable for animation.

49. Voice, and the line you do not cross — A

B — Treats authenticity of the words as removing the need to label a synthetic performance.

C — Applies consent absolutely in a way that would also forbid ordinary archival and estate licensing practice.

D — Treats payment as what makes the use legitimate, and forgets the audience is a separate party with an interest.

50. From text to a waveform — A

B — Sample rate affects fidelity and was never the cause of the flatness; older systems at high sample rates still sounded averaged.

C — Post-hoc prosody adjustment was tried in the older pipelines and is not what changed; the improvement came from the generative objective.

D — That describes concatenative synthesis, the older approach whose seams are exactly what people found unnatural.

51. What a voice print actually is — A

B — A real operational concern about any deletion, but not the specific gap that fine-tuning creates.

C — Describes the zero-shot case; with fine-tuning the identity lives in the model weights rather than in a cached vector.

D — Watermarks mark provenance; they carry no information from which a voice could be rebuilt.

52. Prosody, and why it still sounds slightly wrong — C

A — A real constraint on very long inputs, and the paragraph-level failure appears even in short passages generated together.

B — Drift affects audio quality over long runs; the paragraph problem is about meaning and emphasis, and it appears from the first sentences.

D — Temperature governs variability of delivery, and even with variation the emphasis would still not track the argument.

53. The languages it does badly, and who that is — D

A — A total tells you nothing about distribution; almost all of it will belong to a handful of the languages.

B — A reasonable scoping question that still does not distinguish a well-trained language from a barely-trained one.

C — A good question about a real weakness, and it addresses one failure rather than the overall quality claim.

54. The other half: turning speech into text — B

A — Dropped segments do occur and leave gaps, which are noticeable; the harder problem is content that is present and wrong.

C — Calibration is a genuine weakness, and many systems expose no confidence at all, so this is not the primary risk.

D — Segmentation errors are real, predictable and easy to mitigate by cutting on silence, which makes them far less dangerous.

55. The dubbing chain, and where it breaks — C

A — Recognition is often not the least accurate stage; its significance comes from its position rather than its rate.

B — A practical convenience and not the mechanism; a target-language reviewer is also needed, at a later stage.

D — Translations can be corrected without regenerating audio too, and correction cost is not what makes stage one decisive.

56. Can you tell? Mostly not — D

A — Plausible about degradation and it addresses the wrong problem; the dominant issue is the volume of false positives among genuine callers.

B — Accuracy above chance is not sufficient when the positive class is rare; the base rate determines what a flag actually means.

C — Assumes labelled feedback that a support desk does not have; it rarely learns which flagged calls were genuinely synthetic.

57. The fraud that already happens — D

A — Reversal is unreliable and often impossible; the callback works by verifying identity, not by buying recovery time.

B — Tracing is not what protects the caller, and it plays no part in the moment the decision is made.

C — Assumes a technical limit that does not exist; a clone can sustain a long conversation perfectly well.

58. Music, and whose music it learned from — D

A — Reads a terms-of-service ownership clause as an indemnity against third-party claims.

B — Assumes the internal mechanism decides the legal question, when the comparison is made between the two outputs.

C — Relies on a widely repeated 'a few seconds is fine' rule that does not exist in copyright law.

59. How a machine writes a waveform — A

B — Step counts affect detail in diffusion pipelines generally; a systematic absence of the top octave points at the sample rate instead.

C — A prompt can shift the tonal balance within what the format carries; it cannot add frequencies the sample rate excludes.

D — Mono collapse changes the sense of space rather than the frequency content, and would be audible as narrowness rather than as a closed top end.

60. Why music is harder than speech — C

A — Polyphony is a genuine difficulty, and the window is measured in time rather than being consumed faster by more instruments.

B — Data volume differs, and the decisive problem is structural: form operates over durations no window covers, regardless of training volume.

D — Speech is often generated at a lower rate, but the window is defined over tokens representing time, so this does not explain the difference in coherence.

61. Pulling a mix apart — B

A — The model does not create new expression; it isolates existing recorded material, which remains the original recording.

C — Transformativeness concerns the use made of a work, not the technical process; extraction is not itself a defence, and fair use is not a global doctrine.

D — Splits the rights incorrectly; a commercial recording engages the recording copyright and usually the composition as well.

62. Sound effects, foley and where it is already good — A

B — Many effects are broadband and complex rather than monophonic; the decisive factor is duration and absence of form, not the number of voices.

C — Effects corpora are not obviously larger, and the difference in outcome would not follow from data volume alone.

D — Effects benefit from a high sample rate for brightness, and fewer tokens would not confer quality in any case.

63. Whose music it learned from — C

A — Confidentiality limits what others can learn from the terms; the deeper point is that a settlement produces no ruling at all.

B — Licensing deals are typically catalogue-wide; the scope of the deal is not what leaves the law unresolved.

D — Proceedings ran in several jurisdictions, and in any case a settlement anywhere produces no ruling.

64. When a tune becomes somebody else's — C

A — Assigning ownership of output does not transfer liability to third parties whose rights may have been infringed.

B — Nobody has established that; the material entered the model from the original, and the point is untested rather than resolved in your favour.

D — Attribution is not a licence; crediting a work you have reproduced without permission does not make the reproduction lawful.

65. What platforms do about it — C

A — Terms of service are contractual and cannot displace copyright law; they simply operate alongside it.

B — Copyright certainly applies to any pre-existing works that a generated track resembles, which is the whole risk.

D — No such general requirement exists; platform action follows their own policies and rights holders' registered preferences.

66. When generated music is the right answer — D

A — Budget decides what you can buy and not what the project needs, which is the judgement that should come first.

B — A necessary compliance step at the end rather than the question that decides the approach.

C — Reduces the decision to whether you will be caught, which is neither a reliable test nor a defensible standard.

67. Why detection is the wrong hope — D

A — A genuine problem that compounds the difficulty, and the arithmetic fails even if accuracy held at 95% on every file.

B — Also true and also insufficient; even a detector that generalised perfectly would be swamped by the base rate.

C — Names the right metric without giving the reason; precision is low here specifically because the positive class is rare.

68. What a watermark can survive — A

B — Ignores open models, which produce no mark, and every transformation that removes one.

C — Editing can remove a mark, and absence is equally consistent with the file never having carried one.

D — Possible in a specific case, and it does not license any general inference from absence.

69. Signed provenance, and where the chain breaks — A

B — Provenance attests to the file's history, not to the truthfulness of the scene or the caption attached to it.

C — A conforming editor records a generative edit in the chain rather than invalidating it, so the chain must actually be read.

D — Certificates are issued to devices and software vendors and carry no statement about ownership of the work.

70. Verify the claim, not the picture — C

A — Metadata is trivially forged and usually absent; the method does not depend on it.

B — Update frequency is not the reason; the method relies on external records rather than on any model at all.

D — No such index exists; reverse search finds published copies, not the fact of generation.

71. When everything can be denied — C

A — Transparency would help their credibility, and the dividend attacks the possibility of establishing anything rather than the openness of any particular method.

B — A real weakness of detection, and the denial does not depend on citing any generator at all.

D — Probabilistic output is a rhetorical opening and not the mechanism; the dividend works even against a confident result.

72. Writing a disclosure somebody can use — C

A — Covers everything and therefore says nothing; a reader learns nothing about what to trust in the piece.

B — Brief and uninformative in the same way; it does not distinguish a colour correction from a fabricated scene.

D — Long, formal and still unspecific about which elements; length is not the same as informativeness.

73. What the platforms require now — B

A — Many platforms do render a visible label from that field; the weakness is what happens to the file elsewhere.

C — Declaration fields are frequently mandatory, and no single approach satisfies every jurisdiction's differing legal requirements.

D — It does not; the platform's field must still be completed accurately, and a false declaration is a breach regardless.

74. A routine you can actually run — D

A — The module has already supplied that knowledge; the ordering reflects the strength of the evidence, not the skill required.

B — Anchoring is a genuine concern, and the primary reason for the ordering is that inspection is the weakest evidence available.

C — Compression does degrade artefacts, which is one reason inspection is weak rather than the reason it is placed last.

75. Who owns it, and who has to say it is AI — B

A — Treats creative effort in the instruction as legally equivalent to creative effort in the expression.

C — Confuses a contractual promise not to claim the work with the existence of an underlying property right.

D — Assumes registration constitutes copyright rather than recording a right that must already exist.

76. Was the training lawful? — B

A — Not accurate; outcomes have gone both ways depending on acquisition and market evidence.

C — Commerciality is one factor among several rather than a bar, and non-commercial uses have also been found infringing.

D — Output similarity is a separate question; the lawfulness of the training copies is genuinely at issue in these cases.

77. Can anyone own the output? — C

A — A contract can only transfer a right that exists; it cannot manufacture copyright in unprotected material.

B — Automatic protection depends on there being an author; in several jurisdictions purely generated output has none.

D — Most providers assign whatever rights they have in outputs to the user; the difficulty is whether any copyright exists to assign.

78. Faces, voices and the rights in a person — D

A — Naming is strong evidence of intent, and a depiction produced without the name can still be recognisable and actionable.

B — Realism affects the seriousness of the harm; personality rights can be engaged by caricature and illustration too.

C — Commercial use is central in some systems, and defamation, data protection and criminal provisions apply to non-commercial uses as well.

79. Style is not owned, and other things are — B

A — Remedies vary case by case and are not the reason; the choice follows from what each right actually covers.

C — Trademark rights arise from registration or established use in trade, not automatically from having a name.

D — Registration is a precondition to suit in some countries only, and it is not what drove the reframing.

80. Read the licence on the model — C

A — A separate and real issue about output protection, and it is not what makes the model itself unusable.

B — Open releases carry explicit licences; the problem is what those licences say rather than their absence.

D — Platforms label content; none require a watermark as a condition of commercial distribution.

81. What an indemnity actually buys — D

A — This is the central case such indemnities are drafted to cover.

B — Still a copyright claim, and squarely within the type of dispute these terms address.

C — A straightforward copyright claim of exactly the kind covered.

82. Where labelling is already the law — B

A — Several of these regimes attach to where the content is made available, not only to where the publisher is established.

C — Obligations take effect on their stated dates; absence of detailed guidance does not suspend them.

D — Consumer-protection law adds to the obligations rather than displacing labelling duties, and advertising is generally treated more strictly.

83. What to write down before you start — C

A — They are ordinarily enforceable; the problem is that you cannot meet the obligation, not that it is void.

B — Vendor indemnities run to their customer under conditions and do not create any promise to your client.

D — Confuses two questions; the output's own protection says nothing about pre-existing works it might resemble.

84. A procedure for the cases nobody has answered — C

A — A record of your own reasoning binds nobody; its value lies elsewhere.

B — Liability is allocated by the contract, not by a note in your own file.

D — Those provisions concern labelling and training-data summaries rather than a practitioner's internal reasoning.

Module test — 1. How the picture gets made

1. C

The denoising loop is a numerical approximation of a smooth path from noise to image, and each step is a guess at where that path goes next. Once the steps are small enough for the guess to be accurate, halving them again reduces an error that was already smaller than an eight-bit image can represent. Modern samplers reach that point at twenty to thirty steps; older ones needed fifty to a hundred, and distilled models get there in one to four.

2. A

Diffusion happens in a compressed latent grid, typically eight times smaller in each direction, so one latent cell stands for an eight-by-eight block of the final picture. A letter two cells wide has no room in the decoder to become a letter shape, and what comes back is the statistical impression of small type. No prompt adds latent resolution. The fixes are structural: generate larger, generate the sign separately, or set the type afterwards in a real font.

3. D

Guidance takes the difference between the prompted and unprompted predictions and travels several times that distance past either of them. At high scale the resulting latent sits outside the distribution the autoencoder ever learned to decode, so the decoder is asked to render values it has never seen. Clipping and posterisation are what an autoencoder does with impossible input. The step count does not change, and the burnt look is a dial turned too far rather than a model defect.

4. A

Everything in the loop is deterministic arithmetic, but not bit-identical arithmetic across different hardware. Small floating-point differences at step one are amplified by twenty-nine further steps into a visibly different image. Bit-identical reproduction across machines is not something these pipelines promise; what is reproducible is the same machine with the same versions. Seeds do behave differently across models, but she is using the same model file, and a stochastic sampler would also break repeatability on her own machine, which is not what is described.

5. B

Memorisation is driven by duplication in the training data, not by prompt length or guidance. An image that appeared once is averaged away; one that appeared hundreds of times with a consistent caption becomes a target the optimiser can profitably learn exactly. Famous paintings, film posters, book covers and stock photographs are the duplicated cases. A living illustrator with a thin online presence usually yields a loose style imitation rather than a copy of any one work, which raises a different question, since style is generally outside copyright while a reproduction is not.

6. D

Prompt style should match the captions a model was trained on. Alt text reads like a keyword list, so a model trained on it is a bag-of-concepts machine and keyword lists suit it. Recaptioning — passing every training image through a vision-language model that writes a long literal description — means the newer model learned from prose, and a tag list is unlike anything in its training set. It will cope, and the relational and spatial vocabulary it was specifically trained on goes unused. Token limits on recaptioned models are generally longer, not shorter.

Module test — 2. Steering it: prompts and what sits around them

1. B

The guided prediction is the negative prediction plus the scale times the difference between the positive and negative predictions. At a scale of one that reduces to the positive prediction alone, so whatever is in the negative box has no influence. The box is still displayed, which is unhelpful. This is the same arithmetic that explains what a negative prompt is doing in the first place: it replaces the empty prompt in the second run, moving the point the result is pushed away from rather than issuing a prohibition.

2. C

CLIP text encoders accept 77 token positions, two of them markers, so roughly 75 tokens of prompt — about 55 to 60 ordinary English words. Past that the remainder is silently discarded or split into a separate chunk and averaged. Nothing in the interface warns you, and the image still looks like the beginning of your prompt. A deliberately unmistakable token at the end is a two-generation test. Moving the duck to the front and seeing it appear confirms the model can draw ducks, which separates a limit from a coverage gap.

3. D

Weighting multiplies or interpolates the vectors the text encoder produced and hands the result to the diffusion model as if the encoder had made them. The model was trained on encoder outputs, not on scaled encoder outputs. A small adjustment stays near familiar territory and behaves as expected; a large one produces a vector outside anything in training, and outside that range more weight does not mean more of the concept. Keeping weights between about 0.6 and 1.4 is the rule most experienced users converge on.

4. A

A structural control feeds an image-shaped map into the denoising network at every step, at matching spatial positions. There is no attention competition, no binding problem, and no dependence on the caption corpus containing a phrase for the arrangement you want. A prompt, by contrast, is a sequence of vectors available everywhere in the grid, and a relation between two objects is learned only as a weak statistical association that loses to any stronger compositional pressure.

5. C

A deliberately sparse, slightly odd request has an obvious correct answer — an almost empty frame — so anything added is visible as an addition. If a plant, a rug and a window come back, a language model expanded your request before the image model saw it. A detailed prompt returns a detailed image either way, teaching you nothing. Hosted services rarely show the rewritten prompt; the metadata block is a local-tool feature and would not be present; and a seed comparison confuses an expansion step with ordinary sampling variation.

6. C

References are inputs to a process rather than instructions to a particular model, so they transfer. Seeds do not: the starting noise is the same numbers, and everything that interprets them has changed. Inherited negative prompts were tuned for one checkpoint and frequently reduce quality elsewhere, because a negative term only works if the model was trained on captions containing it. Weight values are not even comparable between interfaces, let alone between models, because implementations scale token vectors differently.

Module test — 3. Where it breaks, and why

1. C

A count is never locally visible. Standing at any one patch of the image, four objects and six objects look equally plausible, and local plausibility is the only thing the training objective ever rewarded. Small counts survive because the arrangement is simple enough to settle correctly by accident. Models trained on machine-written captions do better, because their captions actually said 'five apples', and published evaluations still show accuracy falling sharply past four. Better is not a mechanism.

2. A

Captions on the web overwhelmingly describe what is in a picture, so the encoder had little chance to learn what 'no' does to the noun that follows. What it learned is that the token 'elephant' appears in captions of pictures with elephants in them. Putting the word in your prompt makes the elephant more probable. Models with a proper language encoder do considerably better on whole-object negation and still fail on negated attributes and relations. The reliable technique is to describe the state you want rather than the thing you do not.

3. B

There is no coordinate system in the model; relations are learned as statistical associations from captions. Gravity makes vertical relations stable and consistently described. Left and right are frequently written from the subject's perspective rather than the viewer's, so the corpus genuinely disagrees with itself. The same reasoning explains why 'behind' is worse than 'on', and why relations between two similar objects fail more than between dissimilar ones. The fix is to supply the geometry with a conditioning map rather than to phrase it better.

4. C

Detail is budgeted in latent cells, not pixels. On an eight-times autoencoder, a sixty-pixel face is about seven cells across, which cannot hold two eyes, a nose, a mouth and a jaw in the right relation. Going to 1536 makes it about ten cells: a marginal gain for a substantial cost, and the failure does not change in kind. Exceeding the model's native size in one pass also risks the repeated-subject artefact. What works is generating the face separately at full size, or a masked detail pass that regenerates the region at native resolution.

5. D

The objective was to predict the noise in a noisy input. A picture whose mirror shows the wrong room is, locally, exactly as easy to denoise as one where it shows the right room: both are mirror-shaped regions containing mirror-like content. There is no constraint solver anywhere. This is one mechanism, not five, and it also produces disagreeing shadows, asymmetric earrings and lines that re-emerge at the wrong height. The heuristic that follows: these models are excellent at anything decided locally and unreliable at anything that must agree across the whole frame.

6. B

Relation failures are not stable across seeds, so regenerating sometimes works. That is exactly what makes it misleading: it reads as a prompt that is close, when the real state of affairs is a weighted coin. The test for whether a relational prompt is reliable is four fixed seeds, not one lucky generation. If it fails on two of four, it will fail on the shot the client picks. Nothing is learned between generations, and drift toward the average comes from repeated image-to-image, not from regenerating from noise.

Module test — 4. Control: getting the picture you actually need

1. C

Image-to-image encodes your input to a latent, adds the noise corresponding to some point partway down the schedule, and runs the loop from there. Strength is that starting point. Steps and strength multiply, so 30 steps at 0.4 runs about twelve. Two consequences follow: a low strength with few steps produces a barely-changed image however you word the prompt, and a strength of 1.0 destroys the input entirely, which is ordinary generation. The useful band is narrower than the slider suggests.

2. A

Most implementations do not process the whole image for an inpaint. They crop a region around the mask, scale it to the model's native size, generate, and paste back. That crop is the model's entire view of the scene. A face-sized rectangle with no shoulders, no horizon and no window gives the model no idea which way the light falls, so it invents one. Widening the context — 'only masked padding' in some interfaces, the crop size on the node in others — usually fixes it with the same mask and the same prompt.

3. B

It is not that the model tries harder; it has more cells. A face occupying seven latent cells in the full image occupies most of the grid when the crop around it is scaled to native resolution before generation and scaled back afterwards. This is the mechanism behind face-detailer workflows, and it is the single largest quality gain available to anyone generating scenes with people in them. There is no per-region loss at generation time, and the model is the same model.

4. D

This is over-constraint. A canny map pins every contour, and a painting differs from a photograph mainly in how contours are made — a painter simplifies, merges and omits edges. With all of them held, the prompt can only change surface. The fix is almost always to constrain less: lower the strength, apply the map for only the first part of the run, or use a depth map, which fixes three-dimensional arrangement while leaving contour and surface free. Raising guidance or adding negatives fights the wrong control.

5. C

Anything constant across the training images and left uncaptioned is absorbed into whatever the trigger word means. Two rules follow. Vary everything except the subject — backgrounds, lighting, distance, angle, clothing — so the constant is the person and nothing else. And caption what changes between images, including the background, so it stays a variable rather than binding to the subject. Fifteen to thirty images is the right range for a subject; the problem here is their sameness, not their number.

6. A

A generative upscaler is a plausibility engine, not a recovery engine. It was trained on pairs of small and large images and predicts what the large version would look like, so a grey smudge becomes whatever most often occupies a smudge of that shape — brick, fabric, or an eye. The result is real detail of a plausible image and not detail of your image, because that information was never captured. Published demonstrations have turned low-resolution pictures of well-known faces into confidently wrong different people. Never present an upscaled image as evidence of what was in the original.

Module test — 5. Time and space: video and three dimensions

1. B

A model generating frames independently would produce 120 unrelated pictures. For frame 47 to continue frame 46, the model must see other frames while generating each one, which means attention across time as well as space. Doubling the frame count quadruples the cost of the temporal attention layers. A 121-frame clip at a modest latent resolution has on the order of three quarters of a million positions attending to each other, and that number decides the hardware. The quadratic term is in the architecture, not in the engineering.

2. C

A video autoencoder compresses in time as well as space, typically turning four input frames into one latent frame. The odd counts are multiples of that factor plus one. The same compression produces a specific artefact: because four frames share one latent, fast motion inside those four frames must be reconstructed by the decoder from a single compressed representation, so a quick pan or gesture comes back smeared or stuttering. Generating at a higher frame rate does not help, because the compression factor is fixed.

3. C

Whatever the architecture, the model considers a finite number of frames at once. Inside that window, consistency is enforced by training. Outside it, nothing enforces anything, so a detail that has fallen out of the window is reconstructed as a plausible thing of that kind rather than as the thing that went in. The same fact predicts characters returning from off-screen as someone else and doors that vanish when a pan comes back. Prompting for consistency addresses nothing mechanical. Measure the window with a distinctive arbitrary detail and plan shots that fit inside it.

4. A

First-and-last-frame conditioning asks the model to interpolate a plausible path between two states. Where no continuous motion connects them in the number of frames available, what it produces instead is a blend. Morphing is therefore diagnostic: the ends are too far apart. The answer is an intermediate frame rather than a better prompt — split the action into two shorter generations. This conditioning route is the strongest control available for a specific action, and it is also how a looping clip is made, by using the same image at both ends.

5. D

Each generation introduces small errors, and chaining makes those errors the ground truth for the next chunk, so they compound. Re-anchoring on a fixed original means drift is measured against a fixed point rather than against the last error. Grading per chunk hides the drift without reducing it; the grade belongs at the end, applied to the whole sequence. Dropping image conditioning removes continuity altogether. The other practical control is to fix a drifted anchor frame by hand as an image before using it.

6. B

Radiance representations store a scene as a cloud of coloured semi-transparent blobs optimised to reproduce photographs from any viewpoint. They render beautifully in real time and they are not geometry: no surface to attach anything to, no clean silhouette, nothing to print. Meshes — vertices, edges and faces — are what engines, animation packages and printers consume. Choose by what happens next: viewing, walkthroughs and virtual tours favour splats, anything to be edited, animated, simulated or manufactured needs a mesh. Conversion exists and loses much of what made the splat good.

Module test — 6. Voice and speech

1. C

Modern speech synthesis converts text into discrete audio tokens with a language model and turns those tokens back into sound with a neural codec. Because it samples rather than averages, it commits to one delivery — which is exactly why it no longer sounds flat — and it can produce tokens that do not correspond to your text. Repeated syllables, a skipped word and trailing babble are the same failure family as a language model repeating itself. The control is not technical: anything consequential is listened to against the script before it ships.

2. B

A speaker encoder is trained to say whether two clips are the same person, and in doing so learns to produce a vector that is close for two recordings of one voice and far apart for two people. Cloning computes that vector from a sample and conditions generation on it, with no training run. This is why three to thirty seconds is enough, why quality of the sample matters far more than length, and why it cannot be prevented: the amount of audio required is below the threshold of ordinary social participation. Fine-tuning is the other route and produces a durable artefact, which has different consent and security implications.

3. C

The decoder in a recognition model is a language model, trained to produce text that is fluent whether or not it is accurate. So the characteristic failures are invented sentences during silence or noise, and plausible substitutions where an unfamiliar name is replaced by a common word that fits. Older systems produced obvious rubbish when they failed, and obvious rubbish is safer, because a confident wrong transcript is trusted. The discipline that follows: measure the error rate on your own audio, and read the transcript against the recording whenever it matters.

4. D

Each stage treats the previous stage's output as truth. A name misheard at recognition is translated confidently, spoken clearly, and lip-synced convincingly, with nothing anywhere signalling a problem. Ten minutes of review on the source transcript removes most of the compounding, and it is the only intervention that acts before the error has been laundered through three further systems. If each stage is ninety-five per cent right, the chain is not ninety-five per cent right — which is the central fact about any pipeline of statistical systems in series.

5. A

Accuracy is a rate per call, so the number of errors depends on how many calls of each kind there are. Ten fakes yield about nine or ten catches. The 9,990 genuine calls yield about five per cent false positives, which is roughly five hundred people treated as suspected impostors. Nothing is wrong with the detector; the rarity of the event swamps a good error rate. This is why a detector belongs inside a system that has a confirmatory step — routing a call to a callback rather than to a refusal — and never as the decision itself.

6. D

Out-of-band verification makes the attacker's capability irrelevant rather than trying to out-run it: calling back on the number in your own records, or confirming by a second channel, works whether the voice was perfect or terrible. Detection is a weak signal at realistic base rates. Reducing the sample surface helps marginally and is not a defence, because seconds of ordinary public audio suffice. The audible tells are being closed deliberately, since naturalness is the product. The other half of this control is organisational: refusal must be the safe option for the person answering the phone.

Module test — 7. Music and sound

1. B

The codec is the ceiling. Everything the model produces is decoded by it, so a round trip with no generation at all shows you the best the system can do. No amount of model capacity beats it. This is the same test as pushing an image through an autoencoder and back to see the pipeline's limit. The preview comparison is worth doing for a different reason — fast modes generate fewer residual codebooks, which is why drafts sound thin — but it tells you about the tier, not about the ceiling.

2. C

A speech model needs about a sentence of context, because the coherence of a paragraph comes from the writing you hand it. Music has no external source of form: nothing supplies the verse, the returning chorus, the harmonic plan, and those operate over durations far longer than any attention window. Music also demands polyphony — simultaneous independent lines that must agree harmonically and stay distinguishable — while a single token stream produces something that sounds like a mixture. Length is not the fix: a six-minute model produces six minutes of wandering.

3. B

The technical ease creates a strong intuition that something has changed, and nothing has. A vocal extracted from a commercial recording is a derivative of that recording: the recording copyright, the composition copyright and usually the performer's rights are all engaged, exactly as sampling the original would engage them. That the extraction was automatic is irrelevant. Uncomplicated uses are your own recordings, licensed material, public-domain recordings, practice at home, and anything where a remix licence has been granted.

4. C

A sound effect is typically under five seconds, has no returning sections, no independent simultaneous voices, and no requirement to be a specific recognisable object. That is precisely the regime where generative audio is strong, and it is why impacts, whooshes, ambiences, weather and interface sounds are usable immediately. Generation also produces an effect to length, which a library file cannot. Where it still fails is specific real objects, speech-adjacent sounds such as crowds and laughter, and anything that must sync tightly, which you place in the edit anyway.

5. A

The largest open question in generative audio was resolved by negotiation rather than by a court, which means no precedent exists to rely on and the answer can move again with the next negotiation. Litigate-then-license is a familiar shape in music, seen before with sampling and with streaming. Claims over compositions and lyrics ran on their own track and a German court did rule against a developer over lyrics, so the position is patchy rather than silent. Anyone describing this as settled is describing their jurisdiction, their preference or their sales position.

6. C

For most people the operative constraint is not copyright law but the platform's audio matching system, which compares uploads against a database of registered recordings and applies the rights holder's policy automatically — monetisation redirected, territory blocks or muting — across every upload at once. It decides first and is appealed into a queue rather than to a person. A track can be matched without a court agreeing it infringes, and can infringe without being matched. Plan for the automated system, because that is what you will actually meet.

Module test — 8. Provenance, detection and disclosure

1. B

It flags 95 of the 100 generated images and about 5 per cent of the 99,900 real ones, which is roughly 4,995. Out of some 5,090 flags, about 95 are right. Nothing is wrong with the detector: the rarity of the thing being detected swamps a good error rate, which is why every screening system for a rare event is designed around a confirmatory step. The consequence is that a positive is a reason to look more carefully and never a reason to accuse, particularly for anyone with authority over the person being flagged.

2. A

Absence proves nothing. Nearly all real photographs are unmarked, all output from open models run locally is unmarked, and heavy transformation, regeneration through another model, or deliberate attack will remove a mark that was there. A watermark detector can say with reasonable confidence that a file came from a particular system; it can never say a file is real. That asymmetry is the whole design, and it is why watermarking is best understood as a record of compliant generation rather than as a barrier to malicious generation.

3. C

Provenance attests to a process, not to truth. A signed manifest saying a camera captured this does not mean the scene was not staged, that the caption is accurate, or that it was taken where it claims. Editing does not break the chain either, because a supporting editor adds a new signed manifest referencing the previous one. Generators also sign their output, so credentials appear on generated files too. And a valid signature is only as good as the certificate behind it. Provenance answers where a file came from, which is a narrower and more durable thing than detection.

4. D

The recycled photograph is both more common and more persuasive than a fabrication, and it survives every kind of inspection because only the caption is wrong. Reverse image search across two or three services, sorted by oldest where the service allows it, catches it in seconds. Anyone who learns only the first three steps of a verification routine — preserve, write the claim, search for the earliest appearance — will catch far more misinformation than someone who becomes expert at spotting generated pixels.

5. B

The dividend is paid to whoever benefits from doubt, and it requires no fake to exist — only that fabrication is known to be possible. It is a claim about what can be known rather than about a file, so no per-item analysis touches it, and an absence of generation artefacts cannot establish that something is real. What helps is positive evidence about origin: provenance at capture, institutional custody, multiple independent recordings, and naming the move when an assertion is offered in place of an analysis.

6. C

A label that appears everywhere carries no information. Worse, it misstates the work in both directions: photographs that were merely opened in an editor with generative features have been labelled as AI, which is technically defensible and tells the reader something untrue. A useful disclosure is specific about what was synthetic and what was not, placed where the content is, written in the reader's terms, and proportionate. The threshold worth adopting is whether the synthetic element changes how the audience should interpret what they are seeing.

Module test — 9. Ownership, law and the parts nobody has settled

1. D

Copyright protects the expression of a particular work, not the manner of making works, so painting in someone's style has never infringed. That is why claims on behalf of artists have been reframed around other theories: an artist's name can function as a trademark, and presenting work as being by or approved by someone is passing off or false endorsement without any copyright question arising. Moral rights attach to specific works rather than to style. Whether style should be protected at all is a live policy argument, and it is separate from what the law currently is.

2. A

Copyright there requires a human author. Prompts alone do not supply authorship however detailed or however many iterations were run, so the generated images are not protected — but the author's own expressive contributions are, including text, selection and arrangement. A registration for exactly such a work ended there. 'The person who made the arrangements necessary' is the United Kingdom's computer-generated-works provision, which India also has in similar form and which has barely been tested in either country. The same image can be unprotected in one market and protected in another.

3. C

A fine-tune of a non-commercial base model is non-commercial, and a merge of several models carries every restriction in the set. The model licence governs the weights; a service's terms govern the output of that service. They are independent documents and a permissive one does not cure a restrictive one. This catches people constantly, because the restricted model is freely downloadable and works perfectly. Community repositories are full of merges whose licence position is genuinely unclear, and 'it was on the internet' is not a defence.

4. B

Most indemnities are drafted around copyright and say nothing about personality and publicity rights, which is exactly the area moving fastest in legislation and one of the likeliest claims where real people are depicted. Read the exclusions first. Cover is also typically conditional on the paid tier, on using the output substantially as generated — which sits badly with a compositing workflow — on safety features being left on, on prompt notification, and on the vendor controlling the defence. Caps commonly exclude your client's own losses, which is what gets claimed against you.

5. D

Disclosure duties attach to audience location as much as to publisher location, so a small studio publishing internationally can be subject to several at once. Three features have been stable across every regime introduced so far: duties follow the audience, machine-readable marking sits alongside human-visible labelling rather than replacing it, and the trigger is whether a reasonable person could take the content for a real record. Building to the strictest applicable rule — a visible disclosure plus preserved provenance metadata — makes a change in the detail an adjustment rather than a rebuild.

6. A

Recognisability comes first because it resolves the largest category of serious harm and the answer is not a balance: if a real person is identifiable, you obtain consent that is informed, specific, written, time-limited, revocable and paid where commercial, and plan the disclosure — or you do not make it. The other three questions follow in order and are genuine judgements. Whatever is decided, the decision and its reasoning go into the job folder in three lines, which is the step people skip and the one that does the most work later.

Making Things With AI — final examination

1. A ***1. How the picture gets made***

The objective rewards one narrow skill: given a noisy input and a stated noise level, say which part is noise. Nothing in it asks the model to represent objects as retrievable facts, to plan a layout, or to judge realism. Generation is that skill run backwards from static. Two things follow immediately — there is no canvas and no plan, so composition settles in the first few steps and one changed word can move the whole frame; and the seed, which fixes the starting static, is the control that changes composition.

2. B ***1. How the picture gets made***

Several widely used schedules leave a faint residue of average brightness at their top step, so the model never learned to produce a genuinely very dark or very bright image. At generation time it starts from true noise, which it never saw in training. The complaint that a model 'cannot do night' is a bug in a design decision made before you arrived, not a shortage of night photographs, and it is not exposed in most consumer interfaces. When a model refuses something structural, change checkpoint rather than adding adjectives.

3. C ***1. How the picture gets made***

The denoising network has a limited receptive field per layer, so in a latent much larger than the model's native size there is nothing holding distant regions in agreement. Each independently settles into a plausible subject. This is why the standard workflow is to generate at native size and upscale afterwards rather than to generate large in one pass: it costs less and fails less. Native size is stated on every model card, and reading that line saves a great deal of confusion.

4. D ***1. How the picture gets made***

The text encoder produces one vector per token, and at every step every cell of the latent computes a similarity against all of them and blends in proportion. A region turning into a cube attends strongly to 'cube' and weakly to both colour words, so which colour wins is close to arbitrary and unstable across seeds. Nothing in the architecture binds an adjective to the noun that follows. Encoders that read syntax reduce the leak and do not remove it; separating the objects into clauses, regions or passes is what actually helps.

5. A ***1. How the picture gets made***

Deterministic samplers — Euler, DDIM, DPM++ 2M and their Karras variants — give the same result from the same seed every time, and adding steps refines toward the same picture. Ancestral and SDE samplers inject fresh noise at each step, so the image keeps changing as steps increase rather than converging, and reproducing a result requires the identical step count. If you are exploring, ancestral samplers are pleasant. If you are comparing, use a deterministic one and remove a variable you do not need.

6. D ***1. How the picture gets made***

Autoregressive image models predict image tokens one at a time using the same network that handles text, so character-level information is available to them in a way it is not to a pure diffusion pipeline handed a semantic vector. This is a real capability difference rather than a matter of prompting harder. The trade is speed and a different failure signature, where the picture drifts as it is written. Knowing which family a tool belongs to, and testing what that family is known to be weak at, takes ten minutes and saves a month.

7. B ***1. How the picture gets made***

The model's language is the caption corpus, which is largely alt text: file names, product codes, photographers' settings and keyword strings. Those phrases are strings that co-occurred with a look, not instructions anybody taught it. The same reasoning predicts the failures: anything the web does not caption well has thin coverage, so a lungi, a sarpech or a specific loom returns something adjacent, while a wedding dress returns white satin. When a prompt fails, ask whether anyone would have written that caption. If not, phrasing will not save you and you need a reference or a fine-tune.

8. C ***2. Steering it: prompts and what sits around them***

The chunk is not signed, not verified and not tamper-evident, so adding a fake generation record to a photograph takes one command. Its absence proves nothing either: any conversion, crop or platform upload removes it, and most generated images circulating online have none. The record is a working convenience for you, not evidence for anybody else. Provenance that resists tampering requires cryptographic signatures, which is a different system with different properties.

9. C ***2. Steering it: prompts and what sits around them***

A negative prompt moves the point guidance pushes away from, so it works on things the encoder represents as a coherent visual group. Watermarks, signatures, caption bars and stock overlays are exactly that: the furniture of the image format, with a consistent look the model learned as a category. Counts have no visual signature distinct from the object, relations cannot be negated meaningfully, and pushing away from a number pushes away from the object. The general rule is to use exclusion for format artefacts and description for the content of the picture.

10. A ***2. Steering it: prompts and what sits around them***

A short prompt does not give you a neutral image; it gives you the centre of a distribution that heavily over-represents commercial photography, stock imagery and social media, all already selected for a narrow idea of what looks good. Guidance compounds it by driving every region toward the most prototypical version of what it is. Between a prompt too short to escape the mode and one too long to be read there is a usable band, roughly twenty to sixty tokens on a CLIP model. Specification, lower guidance and reference conditioning all push away from the centre.

11. B ***2. Steering it: prompts and what sits around them***

An English word is often a single token while the same word in Hindi, Bengali, Tamil or Arabic is split into five or six pieces, so the same meaning costs several times as many positions. This is a real disadvantage and not the writer's fault. The workable route is to prompt the scene in English and use references or a small fine-tune for whatever the English caption vocabulary does not cover. That is an unsatisfying answer and an honest one; it has improved as models are trained on more multilingual captions and it is not solved.

12. D ***2. Steering it: prompts and what sits around them***

Four distinct things are called 'reference' and choosing the wrong one produces exactly this complaint. A style adapter encodes the reference and injects it beside the text conditioning, so palette, texture and rendering carry over while content and composition do not. A structure reference — edges, depth, pose, segmentation — is what pins a layout. Image-to-image keeps the original in proportion to how little noise was added. A subject reference carries an identity. A style adapter will not preserve a composition and a depth map will not preserve colour.

13. A ***2. Steering it: prompts and what sits around them***

Every reference method fails the same two ways at the ends of its range. Too weak and you have paid for a conditioning step that did nothing. Too strong and the result is recognisably the source material, which is both the least interesting image and the one most likely to raise a claim, since the objection to reproduction is about recognisability rather than method. The useful setting is nearly always in the middle third, and the only way to find it for a given model is a short sweep with everything else fixed, done once per method and written down.

14. C ***2. Steering it: prompts and what sits around them***

Hosted models change under you, with no version number and no announcement, and when they do your recorded prompt and seed produce a different image. Nothing in the record is wrong; the machine it referred to no longer exists. A checkpoint file on a disk is a version that cannot be revised without your consent, and open-weight models can be archived like fonts or software. The record, the PNG masters and the version control are all necessary and none of them is sufficient without the model that the record refers to.

15. A ***3. Where it breaks, and why***

More data addressed the gaps that were gaps. It did not address the global consistency problem, the latent resolution ceiling or mode-seeking, because those are properties of the method rather than of the corpus. So the predictable remaining failures are exact quantities, reflections and shadows that must agree with a scene, clock faces and sheet music, anything repeated identically, and scripts thin in the training data — Devanagari, Arabic and Thai still fail in models that spell English perfectly, for exactly the reason English used to.

16. C ***3. Where it breaks, and why***

Everything you do not say is filled in by what is most probable, and most probable is not neutral — it is the mode of a corpus already weighted toward commercial photography. Guidance then pushes every region past the conditional prediction toward the most prototypical version of what it is, so it is literally a dial for how stereotypical the output is. Nothing in the picture is a rendering defect; it is the combination — poreless skin, even teeth, flattering light, undamaged objects, styled food, tidy rooms — that never occurs in an unposed photograph.

17. B ***3. Where it breaks, and why***

Prompt injection is cheap, addresses the immediate complaint, and is context-blind. The failure was applying it uniformly and invisibly, where the user could neither see it nor correct it. The alternatives have their own costs: retraining and rebalancing is principled, expensive and requires somebody to decide which categories exist and in what proportions, usually without publication; user-level specification is honest and effective and puts the burden on the people most likely to be misrepresented. No principle chooses between the candidate target distributions.

18. C ***3. Where it breaks, and why***

Knowing which stage stopped you determines whether anything can be done. A blocklist or a prompt classifier refuses before generation, so you have not paid for anything. Paying and receiving nothing is the signature of an output classifier, which are typically tuned for high recall on nudity and violence and therefore produce a lot of false positives. The practical responses differ by stage: rephrasing toward the clinical or compositional, splitting the request, using a local model for legitimate work that keeps being refused, and reporting the false positive.

19. D ***3. Where it breaks, and why***

A false positive here is not a bug on a fix schedule; it is the classifier working as trained on a case its training under-represented. The same mechanism produces clinical content read as sexual, figurative art read as nudity, and religious gatherings read as crowds in distress. Two things are true at once: filtering is doing real work on categories that are illegal in most of the world, and the same filters silently restrict medical education, art and journalism, disproportionately outside the West, with no published error rate. A vendor reporting only its catch rate is telling you half.

20. A ***3. Where it breaks, and why***

Two facts undo it. Selection: you only notice the generated images you noticed, which is survivorship bias operating on your entire sense of how good these systems are, and looking harder cannot correct it. Removability: fixing hands, adding grain, breaking the lighting and photographing a screen is fifteen minutes of work. The tells identify casual generation, which is a different and much easier problem than identifying deliberate deception. Photographers have had genuine work called AI on the basis of smooth skin, and the cost falls on individuals with no way to prove a negative.

21. B ***3. Where it breaks, and why***

Hair strands, wire fences, rigging, whiskers and text serifs are below the latent's resolution, so the decoder produces something that looks like hair at a glance and dissolves under a zoom, and fences that develop breaks and reconnections. The habit that catches it is from print production: never judge at fit-to-window. Zoom to full size and look at four crops — the focal detail, one background region, one edge, and anything with type or thin structure. Thirty seconds, and the first billboard is a bad place to learn this.

22. C ***4. Control: getting the picture you actually need***

One edit is invisible. The weakness is cumulative: most of these systems re-render everything rather than only the part you named, so five passes compound five small re-interpretations and a drift toward the model's house style. The working habits that follow are to keep the original, do the biggest change first, and where possible run independent edits from the same original and composite them rather than stacking edits on edits.

23. A ***4. Control: getting the picture you actually need***

Each pass adds noise and denoises, and each denoising pulls the result slightly toward the model's mode. Repeated, that is the distribution pull applied over and over: contrast rises, detail smooths, faces regularise. It is also what makes the 'keep re-uploading the same image' experiments converge on a similar glossy face whatever they started from. If you need several changes, make them in one pass at a suitable strength, or use masks so untouched regions are genuinely untouched.

24. B ***4. Control: getting the picture you actually need***

Outpainting is inpainting pointed outwards: pad the canvas, mask the new area, generate with the existing image as context. It inherits its behaviour from the training data, and photographers crop the bottom of a frame far more consistently than the top, so the corpus contains fewer examples of what continues below. The general controls are the same as for inpainting: extend in steps of a hundred and twenty-eight or two hundred and fifty-six pixels so each step stays anchored, rather than extending far in one pass and letting the far edge drift.

25. C ***4. Control: getting the picture you actually need***

Symmetry is a global constraint, and inpainting can only see the other earring as context, so it approximates. The general rule this module arrives at is a division of labour: use the generative tool for texture and content, and ordinary image editing for geometry and symmetry. Each is far better than the other at its own job, and most reported frustration comes from asking one to do the other's work. A copy, a flip, a placement and a blend takes a minute in GIMP, Krita or Photopea, all free.

26. B ***4. Control: getting the picture you actually need***

Two images with matched grade and slightly different face proportions read as more consistent than two with identical proportions and different grades. The full workflow is a combination — establish the character with a fine-tune or a strong reference set, generate each shot with that conditioning, fix the face in every shot with a masked detail pass, then colour-match everything — and the grading step is the cheapest of the four and the one most often skipped. Consistency is achievable with effort and several methods together, and it is not achievable by prompting.

27. D ***4. Control: getting the picture you actually need***

Quantisation attacks the weights themselves, which is where the bulk of the requirement is, with a quality loss that is measurable and at 8-bit usually invisible. Community-distributed quantised files are standard and free. Offloading and tiling also help and address different parts of the problem — offloading trades speed for residency, tiling removes the decoder's spike at high resolution. Step count has no bearing on how much memory the weights occupy. A well-chosen smaller model often beats a badly-driven large one anyway.

28. A ***4. Control: getting the picture you actually need***

This is the ordering people get wrong. An upscaler is a plausibility engine working from what is in front of it; it has no idea the face was malformed and will happily produce a sharper version of the error. Fixing the face first means the upscale is enlarging something correct. The workable ladder is: generate at native resolution, masked detail pass on faces and small critical elements, convolutional upscale, an optional low-denoise pass at 0.2 to 0.3 to unify texture, then sharpen and grade last.

29. C ***5. Time and space: video and three dimensions***

Terms that appear consistently in the training captions behave like reliable controls, and standard film vocabulary — push in, dolly out, pan, tilt, crane, handheld, locked off, tracking, orbit, rack focus — is what any competent captioner uses and what stock libraries have always tagged with. Invented compound descriptions are captions nobody wrote, and the model takes whichever fragment it recognises. 'Move left' fails for a different reason: it does not say whether the subject or the frame moves, and if you cannot tell from your own prompt, neither can the model.

30. A ***5. Time and space: video and three dimensions***

Low motion is where these models look best, and where demonstration reels quietly live. High motion is where the temporal compression smears and the attention window is stretched hardest, which is where morphing and reorganisation come from. The professional consequence is to design shots that need little motion — a held frame with a small camera move and one thing happening is a better shot in any medium, and most people get much better results the moment they ask for something restrained.

31. C ***5. Time and space: video and three dimensions***

It moves the entire composition problem into the image domain, where you have masks, structural conditioning, fine-tunes, compositing and unlimited cheap iterations — so you settle the frame with everything from the control module and then ask only for motion. What it does not pin is what happens after frame one: the subject can still become someone else by second four, and motion is still decided by the model. A practical detail people miss is that the input still is usually re-encoded, so output frame one is not pixel-identical to your input.

32. B ***5. Time and space: video and three dimensions***

Video-to-video applies per-frame structural conditioning — depth or edges extracted from each source frame — so motion and geometry come from the source exactly. What it does not pin is appearance stability, which is why early attempts flickered badly and why the remaining tell is texture boiling on surfaces the source kept flat. It is also the route for anyone who can shoot, and it settles the rights position for the performance. A Blender render of a moving camera over grey blocks takes ten minutes and gives exact repeatable moves.

33. D ***5. Time and space: video and three dimensions***

Occlusion and permanence is the sharpest and cheapest test. A model with an internal object representation returns the same object; a model doing local plausibility returns an object of roughly that kind, or a different one, or none. Water and fire are exactly where a strong appearance prior succeeds, so they test nothing. The other informative cases are conservation of quantity, causal aftermath rather than the event, and counterfactuals — and an action that is physically ordinary but visually rare, such as a person walking backwards up stairs.

34. A ***5. Time and space: video and three dimensions***

A clip from the model you have just read about looks better than the identical clip from the one you have not. Blind scoring removes that for the cost of a minute of file renaming, and people who have done it are usually surprised. It sits alongside the rest of the method: a fixed prompt battery reused across models, a written definition of 'usable' decided before looking, no middle category, and a rate reported as a number. Together they are the difference between a decision and a preference.

35. D ***5. Time and space: video and three dimensions***

A performer's agreement to appear on camera is not an agreement to a synthetic version of themselves, and a form written before any of this existed does not cover it. The shape several performers' unions negotiated is a reasonable template outside a union context: separate consent for creating a digital replica, separate consent for each use, a stated term after which it expires, and payment for use rather than only for the capture session. Disclosure to the audience is a second and separate duty, which the person depicted cannot waive on the audience's behalf.

36. A ***6. Voice and speech***

Phrase breaks are where the model's lack of an intention shows. It infers delivery from what usually happens with sentences of that shape, which is right on average and wrong on the specific sentence often enough to notice. The other reliable tells are missed contrastive stress, flattened lists, statements delivered with rising intonation, emotional monotony over several minutes, and homographs like read, lead, live, bow, record and tear. Each line is plausible; the sequence has no argument running through it.

37. C ***6. Voice and speech***

Reference audio for style is the strongest control available, for the same reason a reference image beats a sentence: you are supplying the thing rather than a word for it, and the model conditions on the style as well as the identity. Temperature trades expressiveness against failures and is a blunt instrument. Adjectives underperform an instruction about intent where a system accepts one — 'sceptical, then convinced by the end of the paragraph' beats 'happy'. And if a line will not come out right after four attempts, the problem is usually the writing.

38. B ***6. Voice and speech***

Zero-shot cloning conditions on a vector at generation time and nothing persists, so discarding the sample genuinely ends it. Fine-tuning trains model weights on a person's recordings and produces a file that is, in a practical sense, a copy of a voice — a thing that can be copied, shared, sold and stolen. Any agreement about a fine-tuned voice has to say who holds the artefact, where it is stored, who may use it and what happens to it on withdrawal. 'We deleted the recordings' is not 'we deleted the model'.

39. D ***6. Voice and speech***

Support ranges from thousands of hours to a few hundred, mostly religious texts read aloud, and the word covers both. Published error rates are usually measured on clean read speech in the best-served accent, and vendors quoting a single headline figure across all languages are quoting the best one. Vendors who have the per-language number on realistic material will give it. There is also a practical corollary: a small model trained on one language frequently beats a convenient multilingual giant on that language.

40. A ***6. Voice and speech***

Code-switching is ordinary speech for a very large number of people, and most systems treat one utterance as one language and mangle both sides of a switch. Hinglish, Taglish, Spanglish and dozens of comparable registers are, from the system's point of view, a fault. The other characteristic failures are names outside the training distribution transcribed phonetically and wrongly, tone and length distinctions flattened, and scripts with ambiguous vowel marking. Fine-tuning an open model on a few hours of your own accent is genuinely achievable and helps substantially.

41. C ***6. Voice and speech***

A watermark is a record of compliant generation, not a barrier to malicious generation. Open models with no watermark run on a laptop, and anyone intending to deceive will use one. Its real uses are automatic platform labelling at scale, keeping training corpora free of a developer's own output, internal accountability, and establishing that removal was a deliberate act. It also cannot say that unmarked audio is real, because almost nothing in the world is marked. Re-recording destroys most of the evidence detectors rely on.

42. A ***6. Voice and speech***

Read the five conditions against the common shortcuts and each one fails a different clause: a verbal agreement fails 'written', a blanket release fails 'specific', a perpetual grant fails 'time-limited', and 'you agreed once' fails 'revocable'. Payment is a separate requirement where the use is commercial, not a substitute for any of them. And disclosure to the audience is a second duty entirely, because somebody who consented to a synthetic presenter has not consented on the viewer's behalf to be deceived.

43. B ***7. Music and sound***

Spectrogram-based systems treat the time-frequency picture as an image and apply diffusion, which inherits image diffusion's strong texture and good conditioning — and the vocoder problem, because a spectrogram discards phase and reconstructing it produces the characteristic metallic or watery artefacts. Token-based systems tend the other way: abrupt glitches and structural drift. Reduced codebooks give a thin, grainy draft rather than phasiness. Which family a tool belongs to is usually on the model card, and it predicts the artefacts.

44. C ***7. Music and sound***

If the second listen tells you nothing about where you are in the piece — no sense that this is the end rather than the middle — the model has produced texture rather than form. Form is what a listener recognises by comparing the present with something ninety seconds earlier, and that comparison falls outside any current attention window. Doing this deliberately once is the clearest demonstration of the limitation in the module, and it stops you expecting a longer-generating model to fix it. A six-minute model produces six minutes of wandering.

45. A ***7. Music and sound***

Every lossy step has already discarded exactly the fine detail the model uses to attribute sound to sources, so a lossless file separates better than a compressed one and a compressed one better than audio pulled off a video. The artefacts to expect regardless are reverb tails split badly, momentary bleed where sources share a frequency band, smeared cymbals, and worse results on dense mixes and older recordings. The practical test is to separate, sum the stems back together, and compare with the original: what is missing or added is your error budget.

46. C ***7. Music and sound***

Real sounds are built from a body, a low end and a transient, and layering three generated variations with their starts lined up reproduces that structure. This one habit converts adequate generated effects into good ones, and it is why a sound designer's library of individually unremarkable files produces remarkable results. The wider workflow is layered too: a continuous ambience establishing the space, spot effects for the events, and a detail layer of small sounds that is almost always what separates good sound from adequate sound.

47. B ***7. Music and sound***

The claim covers several different positions: a fully cleared catalogue with payment flowing; public-domain and openly licensed material only, which is verifiable and constrains the sound; stock library material whose own contributor terms may not have permitted training, a gap that has caught several providers; and licensed for training but silent about output, which leaves you exposed on the thing you actually care about. The full set of questions is what was licensed, from whom, whether output is covered, whether commercial use is indemnified, and on what conditions.

48. C ***7. Music and sound***

The abuse pattern is to generate thousands of tracks, spread them across many artist names, and stream them with automated accounts to collect per-stream royalties from a shared pot. Prosecutions have followed, involving very large numbers of tracks. The response has been tighter upload controls, identity checks and flags, so a legitimate small artist is inside the blast radius of a policy written for somebody else. That is unfair and it is the situation: expect friction, declare accurately, and do not distribute in bulk.

49. C ***7. Music and sound***

A composer writing to picture is a different and much more expensive product that most projects were never going to buy; the realistic alternative is a pre-recorded track licensed from a catalogue. So the honest question is specificity. Beds, temp tracks, loops, stings and rapid tonal iteration are all cases where appropriate is enough. A score, a song, a brand theme that has to be remembered, high-value work where a claim would be expensive, and anything in documentary or journalism are not. A third route is often forgotten: Creative Commons and public-domain music, made by real musicians, free and explicitly licensed.

50. A ***8. Provenance, detection and disclosure***

A public detector is a target, and optimising against a fixed classifier is a standard result across every detection domain. The other three are independent and any one would be enough: generalisation, because fingerprints are specific to an architecture and version so anything released after the detector defeats it; fragility, because the evidence lives in fine detail that re-encoding, resizing and screenshots destroy; and base rates, which is the one people skip and the one that decides what a positive result is worth in practice.

51. D ***8. Provenance, detection and disclosure***

A corpus increasingly filled with generated material degrades the next generation of models, and marking is the only practical way to filter it. The other real uses are automatic platform labelling at scale, which handles the enormous volume of ordinary non-malicious generated content nobody would otherwise label; internal accountability, so an organisation can tell whether an asset came from its approved pipeline; and establishing intent, since removing a mark is a deliberate act and in some places an offence in itself.

52. B ***8. Provenance, detection and disclosure***

An editor that supports the standard records a crop or a colour correction as a new signed manifest referencing the previous one, so ordinary editing extends the chain rather than breaking it. A screenshot has no history to extend, and screenshots are how an enormous share of images travel. The other real gaps are software that does not participate at all, platforms that strip metadata on upload, and cameras where deployment is partial and recent. Today a file with valid credentials tells you something useful and a file without them tells you nothing.

53. A ***8. Provenance, detection and disclosure***

People reliably collapse three outcomes into two, and unverified — nothing established either way — is both a legitimate stopping point and the usual result. An unverified image should not be shared as true and should not be denounced as fake; the second mistake is now doing serious damage to people whose genuine work gets dismissed. The discipline of not concluding is the hardest part of the routine, and the practical supports are a time box decided in advance and writing down what you established and what you did not.

54. C ***8. Provenance, detection and disclosure***

The routine runs preserve, state the claim, find the earliest appearance, identify the source, check the place, check the time, look for corroboration, check provenance and metadata, and only then inspect. Inspection is placed last because it is the least reliable evidence and because under pressure people trust their eyes, which is precisely what does not work. A checklist removes the judgement about whether this case needs checking, which is the failure mode aviation, surgery and newsrooms all reached the same conclusion about.

55. D ***8. Provenance, detection and disclosure***

A label applied around the content does not travel with the file, and neither does a description, which is why anything you rely on being disclosed belongs in the content itself — in the video or in a persistent overlay. The other frictions to expect are automatic labels you did not ask for, since a photograph edited in software with generative features can pick up metadata a platform reads as AI; inconsistency between services with no shared threshold; and advertising policies, which are a separate and stricter document.

56. C ***8. Provenance, detection and disclosure***

The rule is to disclose when the synthetic element affects how the audience should interpret what they are seeing. Fabricated location sound in factual work is presented as a record of a real place and changes what the viewer believes about it. Background cleanup, upscaling, decorative illustration and colour correction do not. This is the standard journalism already used for photographic manipulation: dodging and burning was always acceptable, adding or removing a person was not, and the line is whether the change alters the meaning.

57. C ***9. Ownership, law and the parts nobody has settled***

A ruling found no fair use where a system reproduced editorial content to compete directly in the same market. Another found training a language model on lawfully acquired books transformative, while holding that building a library from pirated copies was a separate and serious matter that was ultimately settled at very large cost. Another granted judgment to a developer on the record before it while noting explicitly that a better-evidenced case on market harm might come out differently. Acquisition and proved market effect are the variables.

58. A ***9. Ownership, law and the parts nobody has settled***

It has been on the books since 1988, applies where there is no human author, and runs for fifty years. Ireland, New Zealand, Hong Kong, South Africa and India have similar provisions, and the UK has consulted more than once on removing it. India's position is further muddied by a registration that listed an AI as co-author and a subsequent withdrawal notice, both on the record. The European Union has no equivalent and its case law emphasises the author's own free creative choices, which points toward requiring a human.

59. D ***9. Ownership, law and the parts nobody has settled***

The Beijing Internet Court's emphasis was on intellectual investment through prompting, parameter selection and iteration — which is close to what the United States Copyright Office has said does not amount to authorship. Similar facts, opposite result, on genuinely different foundations rather than a temporary anomaly awaiting harmonisation. The practical consequence is that a purely generated mark may be unprotectable in one of a client's markets and protected in another, which is a commercial fact worth telling them.

60. B ***9. Ownership, law and the parts nobody has settled***

Copyright protects works, not people: a photograph of you is owned by the photographer, and the fact that it is you is governed by something else. In the EU and UK that something else is frequently data protection, with heightened protection for biometric data, and it is the strongest and most commonly forgotten instrument. The others in the family are publicity and personality rights, passing off and false endorsement, defamation, and a growing set of specific criminal offences for non-consensual sexual imagery and fraudulent impersonation.

61. D ***9. Ownership, law and the parts nobody has settled***

This is a common and entirely avoidable failure, and it is a contractual problem before it is a privacy one. Training on inputs is often on by default on consumer tiers and off on business ones, so the fix is usually a setting or a tier. The mirror-image version is a client contract warranting that you hold all necessary rights combined with a non-commercial model licence, which puts you in breach of one or the other. Read both sides before signing either, and save a dated copy of the terms, because they change.

62. C ***9. Ownership, law and the parts nobody has settled***

Warrant what can be established and verified from a job folder. Nobody can establish that generated material is free of any third-party right, so warranting it converts an unknowable risk into a certain breach the moment anything is asserted. Exclusive ownership cannot be promised where purely generated elements may attract no copyright at all in some jurisdictions. And a vendor indemnity is a contract between you and the vendor with its own conditions and caps; it is not a warranty you can pass on.

63. A ***9. Ownership, law and the parts nobody has settled***

The explicit label serves the viewer, who has to know before they believe; the implicit one serves the platforms, which must check for it and act where content is unlabelled or falsely labelled. It is the most comprehensive regime currently in force and worth knowing even if you never publish there. India's amended rules go further on visibility, prescribing a label covering a defined portion of the frame and the opening portion of audio — an answer to a real problem, since a label nobody sees is not a label.