ADDALY · THE AI CLASSROOM

Video With AI

Plan the shots, generate ten takes, throw eight away, and cut what is left in a real editor.

---

7 modules · 66 lessons · 26 figures · 7 case studies · 51,153 words · about 10 hours

Dr Mustafa Mun
Author and editor

ABOUT THIS BOOK

Video 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/ai-video. The website is the living version and is corrected first; where the two differ, the website is right.

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

2. The first frame, and everything decided in it

Produce a first frame at the model's native ratio and resolution with a stated light direction, a hand-repaired subject and any real object already composited in, predict from that frame alone which part of the clip will fail, and store it with enough provenance to rebuild the shot six weeks later

1. The still frame is where the decisions belong
2. Compose for the frame you will end on
3. One light direction per scene, written down
4. Fix the frame yourself instead of re-rolling it
5. Put the real object in before you animate it
6. Deciding where the shot lands
7. A camera move with no video model at all
8. Predicting which second will break
9. Every still you keep is a part you do not have to make again
10. Where this is already the right tool
11. Case study: The label that was almost right
12. Module test

3. Directing, not describing

Write a shot instruction naming size, angle, lens character, one camera move with a speed, the subject's verb and what stays still; choose between language, a motion brush and an explicit conditioning signal for a given move; and run a fixed-seed experiment that determines which parts of your instruction the model obeyed

1. Cinematic, 8k, epic gets you the average of everything
2. The two words that decide most of the frame
3. Focal length, depth of field, and the words that carry them
4. Where people stand, and which way they face
5. Taking motion out of language
6. Depth, pose and flow as instructions
7. What a negative prompt can and cannot remove
8. An experiment protocol that fits on a card
9. The clauses that quietly do nothing
10. Case study: Eighteen students and one free allowance
11. Module test

4. Making separate shots belong to each other

Hold a character, a location and a look across a multi-shot sequence using a character sheet, reference conditioning or a trained model as appropriate, decide which of those a given job justifies, and state the compounding probability of a matching sequence from a measured per-shot hit rate

1. A cut breaks on the light before it breaks on the face
2. The document you build before you generate anything
3. What a reference image is actually doing
4. When a LoRA is worth the afternoon
5. Building a room that stays the same room
6. The blue that is not quite the same blue
7. Fixing a face after the clip exists
8. Deciding the look once, in writing
9. The arithmetic nobody puts in the demo reel
10. Six things it cannot do, and the reason for each
11. Case study: Six shots, one face, and the arithmetic
12. Module test

5. The half of video that is not picture

Build the audio half of a generated video — a speaking character whose consent you can document, a voice recorded or synthesised with a stated quality trade-off, room tone and spot effects under every shot, and music whose licence you could show a platform — and explain why sound repairs perceived physical errors that no regeneration would fix

1. Lip-sync is easy; the permission is not
2. What the model is actually reshaping
3. Acting is timing, and timing has to come from somewhere
4. Synthetic voice, and where it stops being good enough
5. A phone, a quiet room, and four minutes in Audacity
6. Why silence is the worst sound
7. Music you could defend if somebody asked
8. The viewer believes the sound
9. When the model makes the sound too
10. Case study: The retired doctor's voice
11. Module test

6. From generations to a finished piece

Take a folder of mixed-rate generations to a delivered file — conformed, assembled, cut on motion, colour-matched under one grade and one grain pass, titled, sound-finished and exported to a named platform specification — and name the free tool that performs each step on the hardware you actually have

1. The generator makes footage; the editor makes the video
2. Decide the frame rate before you import anything
3. Choosing takes before you start cutting
4. Cut on motion, and one frame before the failure
5. Thirty renders, one world
6. Putting back the texture the encoder removed
7. Real type, added afterwards
8. Slow motion is where fake physics goes to be caught
9. The last render, and what the platform does to it
10. The whole pipeline on a phone or an old laptop
11. Case study: The four-K admissions film
12. Module test

7. Money, clients and running the job

Plan, price and run a generative video job end to end — a shot list with durations and dependencies, an approved animatic before any final generation, a draft pass at the cheapest tier, a kill number per shot, a quote expressed in shots with included attempts, and a project record that makes a later revision cheap

1. An afternoon of curiosity can cost more than a day of work
2. The document everything else is built from
3. Get the yes before you spend the money
4. Test the idea at a tenth of the price
5. Queue everything, then judge it together
6. The arithmetic nobody does before subscribing
7. Price the thing that actually costs money
8. What to keep, and the day it saves you
9. Case study: The quote that said no to two shots
10. Module test

Video With AI — final examination

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

MODULE 1

What the machine is actually doing

Before the shot list, the machinery. A video model compresses your clip roughly fifty times, denoises the whole block of time at once, and attends across every patch of every frame simultaneously. This block walks that path end to end — encoder, schedule, attention, decoder — and turns each stage into something you can act on: which setting changes the picture, which only changes the bill, why the last second loosens, and what it takes to run any of this on hardware you own.

> **BY THE END OF THIS MODULE YOU CAN**
>
> Trace a prompt and a first frame through a latent video model — encoder, noise schedule, attention across time, decoder — and use that path to predict which settings change the picture, which change only the cost, why the final second degrades, and what VRAM a given model needs before you download it

1 · 8 min

Nobody is hiding the long version from you

THE ONE THING TO KEEP

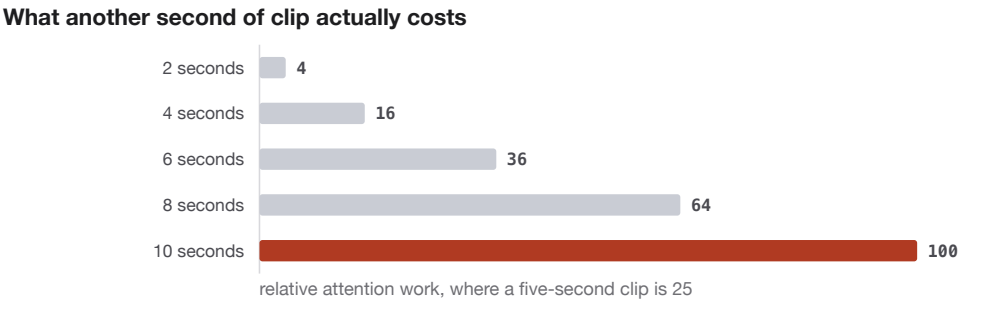
Clip length is a memory-and-attention limit, not a product tier, so plan in shots and cut before the drift rather than trying to generate through it.

The limit is arithmetic, not marketing

Most people assume the few-second cap on generated clips is a pricing tactic — that a longer version exists behind a higher tier. It does not. The length falls out of how these models are built, and understanding why changes how you plan everything else.

A video model does not make frames one at a time. Frame 90 has to agree with frame 12 about where the wall is, so the model generates the whole clip as a single object. To make that affordable it compresses first: a video autoencoder squeezes the picture by roughly eight times in each spatial direction and by four times or more along time, turning a few hundred frames into a much smaller block of latent tokens. A transformer then denoises that entire block, attending across all of it.

Attention cost grows with roughly the square of the token count. Double the duration, roughly double the tokens, roughly quadruple the attention work — and the whole block has to sit in GPU memory at once while it is denoised. A ten-second clip is not twice the price of a five-second one. It is several times worse, and past a point it does not fit on the card at all.



Attention work grows with roughly the square of the token count, and tokens grow with duration. Doubling the length does not double the bill, it quadruples the attention work — and the whole block has to sit in GPU memory at once while it is denoised. Five to ten seconds is where quality, price and memory currently meet.

That is the entire explanation. Five to ten seconds is where quality, price and memory currently meet.

Why the last second is the worst second

In image-to-video the first frame is given to the model. Every later frame is inferred, and each one is anchored by the frames before it rather than by anything fixed. Error accumulates with distance from the anchor.

You will see this in a consistent pattern:

- The first quarter-second is a settle. Motion starts slightly wrong and corrects.
- The middle is the good part.
- The last second is where the face loosens, the background churns, the colour creeps warm, and any object that left frame comes back subtly changed.

So generate long and deliver short. If you need six seconds on screen, generate ten. Treat the head and tail as offcuts, the way a printer treats bleed.

Extending compounds the damage

Every tool offers an extend button. What it does mechanically: takes the last frame of the previous generation — which is already a decoded, lossy output —

and uses it as the conditioning image for a fresh generation. That frame gets re-encoded into latent space, generated from, decoded again.

This is a photocopy of a photocopy. Contrast climbs. Saturation creeps. Fine texture smooths into a painted look. Three extends in, the room is a different colour from where it started, and no prompt fixes it because nothing in the prompt changed.

If you must extend, do one of two things. Colour-match the sections back together in the edit afterwards. Or better, treat each extension as a separate shot and cut between them — a mismatch that reads as a cut is invisible, while the same mismatch inside a continuous move is glaring.

Plan in shots, and put the cut where the model is weakest

A three-minute explainer is thirty shots. Write it as thirty lines with a duration next to each one before you generate anything. This is not an AI workflow, it is a shot list, and it is the same document a crew would work from.

Then place your cuts deliberately. If the hand goes wrong at six seconds, the shot is five seconds long. You are not compromising; you are doing what editors have always done, which is end the shot before it stops working.

You can write this list on a phone in a notes app. It costs nothing and it is the highest-value thing in the whole pipeline.

When longer clips arrive, less changes than you expect

Models will keep stretching the window. It matters less than it sounds. Commercial editing runs at two to four seconds a shot; a sixty-second unbroken take is not something most projects want. The constraint you are fighting is already close to normal film grammar.

The thing a longer window does not fix is scenes. Two people talking across a cut, matching eyelines, the same jacket, the same key light — that is the unsolved problem, and it is about consistency between generations, not length within one. A later lesson deals with it honestly.

Today: take whatever you actually want to make, write it as a shot list with a duration on every line, and count the lines. That number, times your cost per shot, is your project.

BEFORE YOU MOVE ON

A team generates an eight-second clip and then presses extend three times to reach thirty seconds. The final section is noticeably more contrasty, more saturated and softer in detail than the opening. What happened?

A. Prompt adherence decays over long durations, so the later sections followed the description less closely than the first.

B. Each extension conditions on the last decoded frame of the previous generation, so encoding and decoding error compounds like a copy of a copy.

C. The clip passed the model's trained duration, so the tool fell back to frame interpolation to fill the remaining time.

D. Later generations were served from a lower-cost tier once the session's high-quality credits ran down.

---

2 · 10 min

Your clip is compressed fifty times before anything happens

THE ONE THING TO KEEP

A video autoencoder throws away roughly ninety-eight per cent of the pixel data before generation begins, so any detail finer than an eight-pixel block is being reinvented by the decoder rather than preserved.

The number that explains most of the artefacts

Take a five-second clip at 1080p and 24 frames per second. That is 120 frames, each 1920 by 1080 with three colour channels: about 746 million

numbers.

No transformer attends over 746 million numbers. So the first thing a video model does is hand your frames to an autoencoder — usually called a 3D VAE or a causal video VAE — which squeezes them. A typical configuration compresses by 8× in width, 8× in height, 4× along time, and expands the channel count from 3 to 16.

Run the arithmetic. 120 frames become 30 latent frames. 1920×1080 becomes 240×135. With 16 channels that is $30 \times 240 \times 135 \times 16 \approx 15.5$ million numbers.

746 million in, 15.5 million out. Roughly forty-eight times smaller. Generation happens entirely in that small space. The decoder then expands it back to pixels at the end.

Where the 746 million numbers go

Pixels: 746 million numbers	A five-second 1080p clip at 24 frames per second. 120 frames, each 1920 by 1080, in three colour channels. No transformer attends over this.
Encoder: 8 by 8 by 4	Eight times narrower, eight times shorter, four times fewer frames, with channels rising from 3 to 16. Lossy, and lossy in a predictable direction.
Latent: 15.5 million numbers	30 latent frames at 240 by 135 by 16 channels. Roughly forty-eight times smaller. Every step of generation happens in here and nowhere else.
Decoder: back to pixels	Anything finer than about an eight-pixel block was never carried, so the decoder invents a plausible replacement: small type, fabric weave, a thin wire, an eyelash.

The squeeze is the reason the attention step is affordable at all, and it is also the reason generated footage has a slightly waxy surface and melting text. Neither was preserved and then damaged. Neither was ever in the latent.

What survives the squeeze, and what does not

The encoder is not a zip file. It is lossy, and it is lossy in a specific, predictable way: it keeps what it was trained to consider important and discards the rest.

What survives well:

- Large shapes, silhouettes, overall colour, the direction of light.

- Smooth gradients and broad textures.
- Motion of large objects.

What does not survive:

- Anything smaller than about an 8×8 pixel block at output resolution. That is a letter in small type, the weave of a fabric, a thin wire, an eyelash.
- High-frequency noise, including real camera grain.
- Fine specular highlights — the tiny bright points on wet surfaces, metal edges, eyes.

When you see those things in the output, they were not preserved. The decoder invented plausible replacements. That is why generated footage has a slightly waxy, denoised surface even when nothing has obviously gone wrong. It is also why text melts: the letterform was never in the latent to begin with.

The round-trip test you can run for free

This is the single most convincing demonstration in the subject, and it costs nothing because it involves no generation at all.

In ComfyUI, build a three-node graph: load a real video, encode it with the model's VAE, decode it straight back out. No diffusion, no prompt, no sampling. Compare input and output.

```
LoadVideo -> VAEEncode -> VAEDecode -> SaveVideo
```

What comes back is the absolute ceiling on quality for that model. Nothing generated through that VAE will ever be sharper than your round-tripped real footage. Do it once with a clip containing a sign, a patterned shirt and a face, and you will stop blaming the sampler for things the encoder did.

If you cannot run ComfyUI, several Hugging Face Spaces expose the same encode-decode path in a browser.

Causal encoders and why the first frame is different

Most current video VAEs are causal: latent frame *n* is computed from pixel frames at or before it, never from later ones. The practical consequence is that the very first frame is encoded on its own, without temporal compression, while every later group of four frames is squeezed together.

That asymmetry is part of why image-to-video works as well as it does. Your supplied still enters the latent space with less loss than any frame that follows it. It is the sharpest frame in the clip before generation even starts.

It is also why the first quarter-second often looks subtly different from the rest — the settle described in the previous lesson has an encoder component as well as an attention component.

What you do differently knowing this

1. **Do not put load-bearing fine detail in the still.** If the logo has to be legible, it goes on in the editor as real text over the top, not in the frame you animate.
2. **Do not send a 4000-pixel still into a 720p model.** It is downscaled on the way in. You paid for pixels that were deleted before the model saw them.
3. **Do not expect an upscaler to recover what the VAE removed.** Upscaling invents new detail; it cannot restore the specific detail that was discarded. It is a finishing step, not a repair.
4. **Add grain at the end, deliberately.** The encoder removed real grain. A uniform grain pass over the finished timeline replaces a texture the pipeline is structurally unable to keep, and it is why generated footage that has been grained reads as film while the raw export reads as CGI.
5. **Judge a model's VAE separately from its sampler.** Two models with the same sampler quality can differ enormously in how much the decoder smears. The round-trip test isolates it.

The honest limitation

Bigger latents would help. Some newer models compress less aggressively along time, or use more channels, and they look visibly crisper. They also cost proportionally more to generate, because everything downstream scales with the size of the latent. The waxiness is not a bug somebody forgot to fix — it is the price paid to make the attention step affordable at all, and it moves only when somebody accepts a bigger bill.

Today: run the encode-decode round trip on ten seconds of your own phone footage. Whatever comes back is the best this model will ever give you.

BEFORE YOU MOVE ON

You generate a shot containing a shop sign. The letters shimmer and reshape across the clip. You raise the sampler steps from 20 to 50 and regenerate. The letters still shimmer. Why?

A. Fifty steps is still too few for text; letterforms typically need 80 to 120 steps before they stabilise.

B. The letters are below the resolution the autoencoder preserves, so the decoder reinvents them each frame however well the latent was denoised.

C. The prompt did not specify the text clearly enough, so the model had no target to converge on.

D. The temporal attention window is shorter than the clip, so the letters lose their anchor to the first frame partway through and begin drifting away from it.

3 · 10 min

The whole clip is made at once, not frame by frame

THE ONE THING TO KEEP

A video model denoises every frame of the clip simultaneously across a fixed number of steps, which is why nothing can be streamed out early, why step count hits diminishing returns fast, and why distilled turbo models trade motion range for speed.

Starting from noise, all at once

Generation begins with the entire latent block filled with random noise — all 30 latent frames of it. The model then makes a prediction about that noise, subtracts part of it, and repeats. After a fixed number of steps the block is clean and gets decoded.

The important word is *simultaneously*. At step 7, frame 1 and frame 120 are both 7 steps along. They are being refined together, seeing each other, agreeing with each other about where the wall is.

This is the opposite of how most people imagine it, and it explains several things that otherwise look arbitrary:

- **You cannot watch the first second while the rest renders.** There is no first second yet. Progress bars that show a preview are decoding a partially denoised block, which is why early previews look like coloured fog.
- **You cannot generate a bit more onto the end.** The block size was fixed before step 1. Extending means a new run conditioned on a decoded frame, with all the photocopy damage that implies.
- **Cost is decided at the start.** Duration, resolution and step count are all fixed before any picture exists. There is no way to spend more only on the interesting part.

What a step actually buys

Each step is one full forward pass of the transformer over the whole latent. Twenty steps means twenty passes; fifty means fifty. Cost is close to linear in steps.

Quality is not. The curve is steep at the start and flat after a point that is lower than most people expect. On a typical modern video model:

Cost is close to linear in steps and quality is not. The shape is what transfers, not the numbers — run the same shot at four step counts with the seed fixed and find your own model's elbow. Note also where the turbo curve stops: it arrives fast and it stops lower, which is why a draft answers a different question from a final.

- 4 to 8 steps: shapes and colour are right, motion is mushy, fine structure is unresolved.
- 15 to 25 steps: most of the quality. This is where the majority of production work sits.
- 30 to 50 steps: small improvements in fine detail and edge stability.
- Beyond 50: you are usually paying for a different random walk, not a better one.

Run this yourself with a fixed seed. Generate the same shot at 10, 20, 30 and 50 steps and put the four side by side. Most people are surprised at how early the returns stop, and it is worth knowing your own model's number rather than a general one.

Flow matching, and why older advice does not transfer

Early diffusion models used a noise schedule with hundreds of possible steps and samplers with names like DDIM and Euler-a. Most current video models are trained with **flow matching** instead, which learns a straighter path from noise to data and converges in far fewer steps.

This matters practically because a lot of advice written for 2022-era image models is wrong here. Step counts of 100, high guidance values, and sampler names that no longer exist in the interface all come from that lineage. If a tutorial tells you to use 80 steps on a flow-matching video model, it was written for different machinery.

Flow-matching models also expose a **shift** or **sigma shift** parameter, which redistributes where the steps are spent along the path. Higher shift spends more effort in the early, structural part of the denoise. Video models generally want a higher shift than image models, and the tool usually sets a sensible default. Change it last, if at all.

Distilled and turbo models

A distilled model is trained to imitate the output of a many-step model in very few steps — often 4 to 8. They are genuinely useful and they are how draft passes become affordable.

What they cost you is real and worth naming:

- **Reduced motion range.** Distillation tends to pull toward the average trajectory. Turbo outputs are noticeably calmer, which sometimes flatters a shot and sometimes kills it.
- **Weaker prompt adherence on unusual requests.** Common instructions survive distillation. Odd ones get smoothed toward the common case.
- **Guidance behaves differently.** Most distilled models want a guidance scale near 1. Feeding them the value you use on the full model produces burnt contrast and frozen motion.

The workflow that follows from this: draft on the turbo variant to test whether a motion idea works at all, then do the final on the full model with the same seed and still. The draft is not a preview of the final — it is a cheap answer to a different question.

Why the same settings give different pictures

Two runs with everything identical except the seed produce different clips because they start from different noise. That is expected. But two runs that you believe are identical can also differ, because the effective inputs include model version, precision, batch size and the specific GPU kernels used. The next lesson deals with that properly.

The free path

Every point above can be tested on a free Hugging Face Space or a Colab session, because they are all about settings rather than about scale. Fix the still and the seed, move one number, compare. An afternoon of that teaches you more about your chosen model than any comparison video, and it costs a queue rather than a balance.

Today: fix a seed and a still, generate at four step counts, and find the point on your own model where the improvement stops being visible. That number is now your production default.

BEFORE YOU MOVE ON

A studio wants to show a client the first two seconds of a ten-second generation while the rest finishes rendering, to save waiting time. Why is this not possible?

A. The API does not expose partial results, though the model itself produces frames sequentially and a self-hosted setup could stream them out as they finish.

B. Early frames are rendered at lower quality first and only refined at the end, so a partial result would misrepresent the shot.

C. The decoder can only run on the complete latent block, although the latent frames themselves finish at different times.

D. Every frame is at the same stage of denoising throughout, so before the last step the first two seconds are as unfinished as the rest.

4 · 10 min

Why a chair stays a chair and a face does not

THE ONE THING TO KEEP

Attention lets every patch of every frame see every other, which is what holds a scene together, but it costs the square of the token count and carries no object identity — so coherence is an emergent statistical effect, not a guarantee about any particular thing.

What attention is doing here

After the encoder, the latent is cut into patches — typically 2×2 in space, sometimes across a pair of latent frames as well. Each patch becomes a token. For our five-second 1080p clip: 30 latent frames, 240×135 latent pixels, patched 2×2, gives 30 × 120 × 68 ≈ 244,800 tokens.

In **full 3D attention**, every one of those tokens computes a relationship with every other one. That is $244{,}800^2 \approx 60$ billion pairs, per attention layer, per

step. Modern kernels do not materialise that matrix, but the work still scales that way.

This is the expensive thing in the whole pipeline, and it is also the thing that makes video models work. A patch on frame 118 can look directly at a patch on frame 3. That is how the wall stays where it was.

Full, factorised, and windowed

Three arrangements, with a clear trade-off.

- **Full 3D attention.** Every token sees every token. Best coherence, highest cost. Used by the models with the best physical consistency, and the reason they are expensive.
- **Factorised attention.** Attend across space within each frame, then across time at each spatial position, in alternating layers. Cost drops enormously. Objects that move diagonally across the frame are handled worse, because a token only attends along its own row through time.
- **Windowed or sparse attention.** Tokens attend within a local neighbourhood in space and time, with occasional global layers. Cheapest. Long-range agreement weakens, which shows up as slow drift over a clip rather than sudden breakage.

You cannot usually pick. It is baked into the model. But you can read the symptom: if a model handles fast diagonal motion badly while holding static scenes well, you are probably looking at factorisation. If everything holds for three seconds and then slowly slides, suspect a window.

Attention has no idea what an object is

This is the part that matters most, and it is routinely misunderstood.

There is no variable anywhere in the model holding "the red mug". There are tokens whose values happen to correlate strongly across frames because they were trained on real footage where mugs persist. Persistence is a statistical habit, not a stored fact.

Three consequences follow directly.

Occlusion is dangerous. When something passes behind something else, the tokens that carried it stop existing for a while. When it reappears, nothing retrieves the original; the model re-samples something consistent with context. A person who walks behind a pillar can come back with a different shirt, and no setting prevents it.

Leaving frame is worse than occlusion. An object that exits the edge and returns has been gone from every token for many frames. Expect it to come back changed.

Count is not conserved. Three birds can become four. There is nothing counting.

The production response is to design around it. Do not write shots where a specific object leaves and returns. Cut instead — the cut is free and the re-entry is not.

Why faces break first

The model is no worse at faces than at chairs. You are far better at faces.

Human vision has dedicated machinery for facial identity, tuned to differences of a few per cent in the geometry between features. A chair whose back changes by thirty per cent still reads as the same chair. A face whose inter-ocular distance shifts by three per cent reads as a different person.

So the same drift rate produces a shrug on furniture and a rejected take on a face. That asymmetry is in your visual cortex, not in the model, and knowing it tells you where to spend: shorter shots on faces, longer shots on everything else.

Text conditioning is a small part of this

Your prompt enters through **cross-attention**: text tokens from an encoder such as T5 or a CLIP variant, attended to by the video tokens. There are a few hundred text tokens against a quarter of a million video tokens.

That ratio is worth sitting with. The prompt is a light steering influence over an enormous amount of structure that is mostly determined by the first frame

and the model's priors. It explains why adding a fourth clause to a prompt often changes nothing, and why the still frame overwhelms the text every time the two disagree.

What you can act on

1. Shots where an object leaves and returns are not shots. Split them.
2. Put the face in fewer, shorter shots than your instinct says.
3. Prefer motion that stays roughly in the same region of frame — a push-in holds better than a fast lateral whip, on factorised models especially.
4. When a model drifts slowly rather than breaking suddenly, do not fight it with prompts. Shorten the clip.
5. Expect the prompt to lose any argument with the first frame.

Today: generate one shot where a person walks behind a column and out the other side. Watch what comes back. That single clip teaches the whole lesson better than the arithmetic does.

BEFORE YOU MOVE ON

In a generated shot, a woman carrying a blue bag walks behind a pillar and emerges with a green bag. What does this reveal about the model?

A. Nothing stores the bag, so once its tokens are occluded the model re-samples something merely consistent with the surroundings.

B. The temporal attention window ended at the pillar, so the frames after it had no access to the frames before.

C. The prompt described the bag only once, and prompt influence decays across the length of a clip so later frames were effectively unconditioned.

D. The autoencoder discarded the bag's colour because it fell below the preserved spatial resolution, and the decoder chose a new one.

5 · 9 min

The prompt vocabulary was written by whoever captioned the training clips

THE ONE THING TO KEEP

Video models learned from automatically captioned clips filtered for aesthetics and motion, so the words that work are the words those captions used, and the model's defaults are the defaults of stock footage.

Where the training data came from

A video model is trained on tens or hundreds of millions of clips, each paired with a text description. Almost none of those descriptions were written by a person.

The pipeline, in the order it runs:

1. **Scrape and segment.** Long videos are cut at shot boundaries by a scene-detection tool, so the training unit is a single continuous shot — usually a few seconds. This alone shapes the model: it has essentially never seen a cut.
2. **Filter.** Clips are scored and discarded. Aesthetic scorers remove ugly ones. Optical-flow scorers remove clips that are too static to teach motion and clips that are too chaotic to learn from. Text detectors often remove clips with heavy on-screen graphics. Watermark detectors remove the obvious ones and miss the rest.
3. **Caption.** A vision-language model writes a description of each surviving clip. Sometimes two, at different lengths.
4. **Train.** The model learns the mapping from those captions to those clips.

Everything you experience as the model's taste was set in steps 2 and 3.

Why mood words fail and shot words work

An auto-captioner describes what it sees. It writes things like "a slow dolly-in on a woman sitting at a desk, warm window light from the left, shallow depth of field". It does not write "epic" or "masterpiece" or "8k", because those are not observations.

So those words appear in training text only where a human-written description leaked in, scattered across wildly unrelated footage. Ask for them and you get the average of that scatter.

Meanwhile the captioner's own vocabulary — shot sizes, camera moves, light direction, lens character — appears tens of millions of times, attached precisely to the thing it names. That is the vocabulary that steers.

The practical rule: **write the caption you would want a machine to produce for the clip you want.** That is literally the function the model learned.

The stock-footage prior

The filtering stage has consequences that show up in every default the model has.

- **It loves a slow push-in.** Static clips were filtered out for having too little motion; frantic ones for having too much. What survives the middle of that distribution is a great deal of gentle, drifting, tripod-and-slider stock footage. That is why an under-specified prompt returns a slow dolly.
- **It centres the subject.** Stock footage composes safely.
- **It is landscape.** Most of the corpus is 16:9. Vertical performance on many models is visibly weaker, not because 9:16 is harder but because there was less of it, and much of what existed was phone footage of lower quality.
- **It grades teal and orange.** That look dominates the commercial footage that scores well aesthetically.
- **It has a narrow idea of who is in a shot.** Stock libraries over-represent some people and settings and under-represent others. Ask for a spe-

cific profession or setting and check what you are handed, because the average is somebody else's average.

None of this is a conspiracy. It is a filter applied at scale, and filters have shapes.

Why the model cannot cut

Worth stating on its own. Scene detection meant every training clip was one unbroken shot. The model has no representation of a cut, which is why asking for "a montage" or "then it cuts to" produces either a morph or a camera move rather than an edit.

Cuts are your job, in the editor. That is not a limitation to work around; it is the correct division of labour, and it is why the edit module exists.

Recaptioning, and why prompts stopped needing to be weird

Several labs now recaption their entire corpus with a better VLM before training, in a consistent style and at a consistent length. The effect on users is direct: prompts that read like ordinary descriptive English work better than keyword soup, because the training text was ordinary descriptive English.

If a model's documentation shows long, prose-style example prompts, that is a recaptioned model and you should write prose. If the examples are comma-separated keyword lists, write keyword lists. The examples are not marketing — they are a sample of the training distribution, and matching it is worth more than any prompt trick.

The free path to finding a model's vocabulary

You do not need to guess. Take one still and generate it six times with six different camera-move phrasings, one word changed each time. Keep the ones that produced a distinct, correct move. That list of six or ten phrases is your working vocabulary for that model, and it will be more reliable than any prompt guide written for a different one.

This costs a free daily allowance and about twenty minutes, and it is the highest-value use of a free tier there is.

Today: read the example prompts in your model's own documentation, and count how many are mood adjectives versus observations. Write your next prompt in whichever style dominates.

BEFORE YOU MOVE ON

A team asks a video model for "a montage of three locations" and consistently gets a single continuous camera move through a morphing space. What in the training pipeline explains this?

A. Montages are rare in the aesthetic-filtered corpus, so the model saw too few of them to learn the pattern reliably.

B. The caption model described montages as camera moves, so the words 'montage' and 'dolly' became entangled in the text encoder.

C. Attention across frames enforces continuity, and a hard cut would violate the temporal coherence that those attention layers were trained to produce.

D. Clips were segmented at shot boundaries before training, so every example was one unbroken shot and the model has no notion of a cut.

6 · 8 min

Same settings, different clip, and why that is not a bug

THE ONE THING TO KEEP

A seed fixes the starting noise, but reproducibility also requires the same weights, precision, batch shape and kernels — which is why local runs repeat exactly and hosted APIs usually will not promise it.

What a seed actually fixes

Generation starts from random noise. A seed is the number that initialises the random number generator that produces that noise. Same seed, same starting noise.

That is all it fixes. Everything else in the path has to match too:

- The model weights, down to the exact checkpoint version.
- The numerical precision — fp32, bf16, fp8 all give different answers.
- The step count, guidance value, scheduler and shift.
- The text encoder and how the prompt was tokenised.
- The batch shape, because some kernels reduce in a different order at different batch sizes.
- The GPU and driver, because attention and convolution kernels are chosen by hardware and some are non-deterministic by design.

Lock all of those and a local run repeats exactly. Miss one and it does not.

Why hosted services will not guarantee it

This is not evasiveness. A hosted endpoint batches your request with other people's, may route it to a different GPU generation, and may silently update

the model — a point release, a safety-filter change, a new default scheduler. Providers usually reserve the right to do all three.

The practical consequence is the one that costs money: **a seed that produced a perfect take last month may not reproduce it today.** People discover this when a client asks for one small change to a shot from a finished project.

So the record you keep has to be stronger than a seed. For every shot that ships, save:

- The first frame as an actual image file.
- The exact prompt text.
- The model name *and version string*.
- Every numeric setting.
- The seed.
- **The output file itself**, at the highest quality you received.

That last one is not a fallback, it is the primary record. The others let you get close; the file is the only thing that is certain.

Using a seed as an instrument

The real value of a seed is not reproducibility. It is isolation.

If you want to know what a prompt change does, you must hold the noise constant. Otherwise you are comparing two different random draws and reading meaning into the difference. This is the single most common error in prompt experimentation, and it wastes entire afternoons.

The protocol:

1. Fix the seed and the still.
2. Change exactly one thing.
3. Generate.
4. Compare against the previous output, not against your memory of it.

With a fixed seed, a change that produces no visible difference genuinely produced no difference. Without one, you cannot tell.

When the tool gives you no seed

Several consumer products hide it. You then have a different, cruder method: generate four takes of A, four of B, and compare the *distributions* rather than individual clips. If three of four A-takes hold the face and one of four B-takes does, that is a signal. One clip each is not.

This costs four times as much, which is a real argument for tools that expose a seed, and a real argument for doing prompt experimentation on an open-weights model even if you deliver on a hosted one.

Seed ranges and the myth of good seeds

You will see people share "good seeds". A seed is meaningful only in combination with the exact prompt, model and settings it was found with. Transplanted anywhere else it is an arbitrary number. There is no such thing as a seed that produces good faces in general.

What is true is that for a *fixed* prompt and still, some seeds land better than others, and scanning ten seeds is often a better use of money than rewriting the prompt ten times. When a shot is nearly right, roll the seed. When it is structurally wrong, the seed will not save it.

Making a local run deterministic

If you are running the model yourself and want bit-exact repeats, the usual measures are worth knowing:

```
import torch
torch.manual_seed(0)
torch.use_deterministic_algorithms(True)
```

Plus a fixed `CUBLAS_WORKSPACE_CONFIG`, a fixed batch size, and no autocast changes between runs. Determinism costs some speed, because it forbids the

fastest kernels. Turn it on while you are experimenting and off when you are producing.

ComfyUI is generally reproducible run to run on the same machine with the same graph, which is one of the strongest practical reasons to prototype there even if you finish elsewhere.

Today: pick a shot you have already generated and try to reproduce it from your notes alone. Whatever you could not reproduce is the field missing from your record sheet.

BEFORE YOU MOVE ON

A freelancer reruns a six-week-old hosted generation with the identical prompt, still, settings and seed, and gets a visibly different clip. What is the most likely explanation?

A. Seeds on hosted platforms are per-session identifiers rather than RNG seeds, so the same number means nothing across sessions.

B. Video models are inherently stochastic even at a fixed seed, because the temporal attention introduces fresh randomness at every step.

C. The provider changed the model or serving stack meanwhile, and a seed fixes only the starting noise, not what is done to it.

D. The still was re-encoded by the upload pipeline at a different compression level, which changed the conditioning image enough to alter the result.

7 · 9 min

The four numbers, and which of them you should touch

THE ONE THING TO KEEP

Guidance scale controls how hard the model is pushed toward your prompt and away from its unconditional prediction, so raising it past a narrow band burns contrast and freezes motion rather than improving adherence.

The controls, in order of how often they should move

Most interfaces expose four things: steps, guidance, sampler, and shift. In practice you should touch them in almost the reverse of the order people usually do.

- **Steps.** Set once per model from your own step-count test. Then leave it.
- **Guidance.** The one worth understanding and occasionally adjusting.
- **Sampler.** Pick the documented default. Differences are small on video.
- **Shift.** Leave alone unless the model's own guide tells you otherwise.

Everything else — prompt, still, motion strength, duration — matters more than all four combined.

What guidance actually does

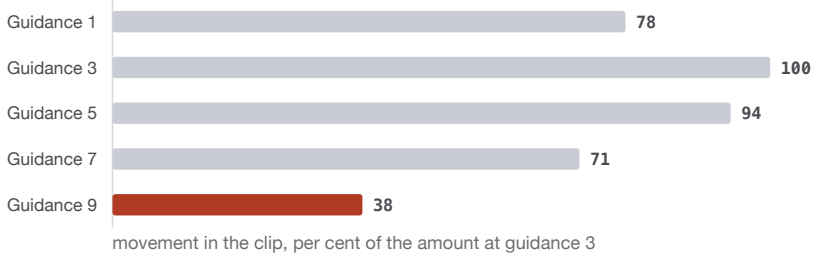
At each step the model makes two predictions: one conditioned on your prompt, one unconditioned. Classifier-free guidance takes the difference and extrapolates along it:

```
prediction = uncond + scale * (cond - uncond)
```

At scale 1 you get the conditioned prediction unchanged. Above 1 you are pushing *past* it — deliberately overshooting in the direction the prompt pulls.

That overshoot is why guidance works and why it breaks. Push harder and the prompt is obeyed more literally. Push too hard and you leave the region where the model's predictions are valid. The symptoms are specific and recognisable:

Raising guidance does not buy adherence, it buys stillness



High guidance drives every frame toward the strongest single interpretation of the prompt, and the strongest interpretation does not change over time — so the clip stiffens toward a photograph while the colours cook. A full model usually wants 3 to 7. A distilled one wants about 1, and feeding it 7 produces exactly the burnt, frozen output on the right.

- Contrast and saturation climb until highlights clip to white.
- Edges get a hard, posterised outline.
- **Motion freezes.** This one is particular to video and catches people out. High guidance drives every frame toward the strongest interpretation of the prompt, and the strongest interpretation does not change over time, so the clip stiffens into a near-still.

If your motion looks locked and your colours look cooked, lower the guidance before you touch anything else.

The usable ranges

Video models generally want lower guidance than image models. Typical bands, which you should verify on your own model rather than trust here:

- Full (non-distilled) video models: roughly 3 to 7, with 5 a common default.
- Distilled or turbo variants: near 1. These were trained to need no guidance, and feeding them 7 produces exactly the burnt, frozen output described above.
- Some models expose separate guidance for the image condition and the text condition. Raising image guidance tightens adherence to your first

frame at the cost of motion; lowering it lets the model drift from your still.

The diagnostic: generate the same shot at guidance 2, 5 and 9 with a fixed seed. You will see the motion die and the colour cook, and you will know your model's ceiling for the rest of your career with it.

Samplers, briefly and honestly

On image models sampler choice was a genuine subject. On flow-matching video models it is mostly not. The differences between Euler, DPM variants and the rest are small compared with the differences between prompts, stills and guidance values.

Use whatever the model card recommends. If you want to test it, test it once, with a fixed seed, and then stop thinking about it. Time spent on samplers is time not spent on the shot list, and the shot list is where the quality is.

Shift, and why it exists for video

Flow-matching models spend their steps along a path from noise to data. **Shift** biases where along that path the effort goes. Higher shift concentrates steps early, where large-scale structure and motion are decided; lower shift concentrates them late, where fine detail resolves.

Video wants more early effort than images do, because getting the motion trajectory right matters more than getting a texture crisp. That is why video models ship with a higher default shift, and why raising resolution sometimes calls for raising shift too — a bigger latent has more structure to settle.

It is a real control and a rarely useful one. If motion is consistently mushy at good guidance and adequate steps, try raising it. Otherwise leave it.

What people reach for instead, and should not

Three habits that waste money:

1. **Raising steps to fix a bad shot.** Steps resolve detail. They do not change what the model decided to depict. A shot that is wrong at 20 steps is the same shot, sharper, at 50.

2. **Raising guidance to force an ignored instruction.** If the model ignored "she turns left", guidance 9 gives you a stiffer clip in which she still does not turn left. The fix is the still or the phrasing.
3. **Changing four things at once and keeping the result.** You now have a setting you cannot explain and cannot repeat.

The free path

All of this is testable for nothing on a Hugging Face Space or a Colab notebook, because none of it needs scale — it needs one still, one seed, and patience with a queue. Build a small grid: three guidance values across two step counts, six generations, one page of notes. That page is worth more than any settings guide, including this one, because it is about your model.

Today: generate one shot at guidance 2, 5 and 9 with the seed fixed, and watch the motion disappear as the colour cooks.

BEFORE YOU MOVE ON

A shot generated at guidance 9 comes back with clipped highlights and almost no movement, although the prompt asked for a walking figure. What is happening mechanically?

A. Guidance extrapolates past the conditioned prediction, and at a high scale every frame is pushed toward the same fixed interpretation.

B. The prompt is being followed so precisely that the model has no freedom left to invent motion between the frames.

C. High guidance requires proportionally more steps, and at the default step count the sampler has not converged yet.

D. The unconditional branch is skipped above a threshold scale, so the model loses the motion prior that this branch normally carries into each step.

8 · 9 min

Generated at sixteen, delivered at twenty-four

THE ONE THING TO KEEP

Many models generate at a low native frame rate and are interpolated up afterwards, which is cheap and usually invisible — except at occlusions and fast motion, where interpolation produces a specific smeared artefact worth recognising.

The frame rates you are actually dealing with

A video model has a native frame rate baked into training. Common values are 16, 24 and 30. A model trained at 16 fps that hands you a 24 fps file did not generate 24 frames per second — something interpolated the extra ones.

This is a good trade and worth understanding rather than resenting. Generating more frames costs quadratically through attention. Interpolating them costs a small, separate network run at a tiny fraction of the price. For most footage the result is indistinguishable.

The interpolators you will meet: **RIFE** (fast, free, open, the default in most pipelines), **FILM** (Google's, handles large motion better, slower), and the optical-flow retimer inside DaVinci Resolve. All three estimate motion between two real frames and warp pixels along it.

Where interpolation gives itself away

Motion estimation fails in three specific places, and once you can name them you will see them in other people's work constantly.

1. **Occlusion boundaries.** Where one object passes in front of another, some pixels exist in frame A and not in frame B. There is nothing to warp, so the interpolator smears or invents. You see a soft halo hugging the moving edge.

2. **Fast motion.** If something moves more than the estimator's search range between frames, the match is wrong and the intermediate frame tears. A hand crossing the frame quickly gets a ghost.
3. **Anything appearing or disappearing.** A flash, a spark, a blink, a hard cut inside the clip. Interpolating across an event produces a half-existing frame.

The practical rule: interpolate calm footage freely, and be suspicious of interpolated fast action. If a clip has one difficult second, interpolate the rest and let that second run at native rate rather than fixing it everywhere.

Motion blur, which generated footage often lacks

A real camera exposes each frame for a fraction of the frame interval — the classic 180-degree shutter means the sensor is open for half of it. Anything moving smears across that exposure. Your eye expects it.

Generated frames are frequently too clean. Every frame is crisp, so motion reads as a sequence of sharp positions rather than as movement. The effect is subtle and it is one of the things that makes footage feel synthetic without anyone being able to say why.

Two fixes, both cheap:

- Add a small amount of motion blur in the editor. Resolve, Kdenlive and most compositors have it. Keep it light; overdone blur is worse than none.
- Interpolate *up* and then conform *down*, blending frames. Generating at 16, interpolating to 48, then rendering to 24 with frame blending gives a natural smear for free.

Conform the project rate before you import a single clip

This is the mistake that costs an afternoon.

Mixed frame rates on one timeline judder, because frames are being duplicated or dropped to fit the project rate. A 30 fps clip on a 24 fps timeline drops one frame in five, and on a slow pan that reads as a stutter you cannot un-see.

Decide the project rate first:

- **24** for anything meant to read as film.
- **25** if you are delivering for broadcast in a PAL region.
- **30** for most social platforms, which is also where 60 shows up for fast content.

Set it before importing. In DaVinci Resolve, changing project frame rate after media is on the timeline ranges from awkward to blocked depending on version, and re-conforming a finished cut is a genuine nuisance.

If a clip arrives at the wrong rate, retime it properly with optical flow rather than letting the timeline drop frames.

Odd durations

Models hand back clips at 5.1 or 4.8 seconds, because the latent frame count divided by the frame rate does not land on a round number. Nothing is wrong. Trim to what you need; you were going to trim the head and tail anyway.

The free path

Everything here is free. RIFE ships inside ComfyUI and inside **Flowframes** on Windows. **FFmpeg** will interpolate, conform and blend from the command line on any machine, including a very old one:

```
# conform 30 fps source to 24 fps with frame blending
ffmpeg -i in.mp4 -filter:v "minterpolate=fps=24:mi_mode=blend"
out.mp4
```

Resolve's free tier includes optical-flow retiming. Kdenlive and Shotcut handle conforming. On a phone, CapCut will change frame rate on export, though with less control.

Today: find out your model's native frame rate from its documentation, then look at one of your own clips frame by frame at an occlusion edge and see

whether the halo is there.

BEFORE YOU MOVE ON

A clip generated at 16 fps and interpolated to 24 looks clean except for a soft halo hugging a hand as it passes in front of a torso. What produced the halo?

A. The autoencoder compressed the hand region more aggressively than the torso because it carried more motion.

B. At the occlusion boundary some pixels exist in one real frame and not the next, so the interpolator has nothing to warp and fills the gap by guessing.

C. Interpolation to a non-integer multiple of the native rate forces the estimator to blend three frames instead of two, softening edges.

D. The hand moved faster than the attention window could track, so the model produced inconsistent hand pixels that the interpolator then averaged together.

9 · 8 min

The ranking will be wrong by the time you read this

THE ONE THING TO KEEP

Benchmark models on one hard shot from your own list, because public rankings measure which clip looks nicer, not whether it did what you asked.

What is out there, in classes rather than in order

The specific league table changes every few months. The categories have been stable for a while, and knowing the categories is what transfers.

Hosted flagships. Google's Veo line, notable for generating synchronised dialogue and ambience along with the picture. OpenAI's Sora line, app-first,

with a consent-gated likeness feature built into the product. Runway, which pairs its generation models with director-style controls — camera moves, a motion brush, performance transfer. Kling from Kuaishou, strong on physical motion and start-and-end frames. Hailuo from MiniMax, cheap and lively. Luma's Dream Machine, with keyframes. Pika, effects-led.

Aggregators and director layers. Higgsfield is the clearest consumer example: it resells access to several underlying models and adds camera-motion presets, character tools and a queue on top. fal.ai and Replicate do the same for developers, with an API instead of a UI. What you gain is one balance, one interface and the ability to switch models without opening new accounts. What you pay is a margin on top of the underlying price, and one more step of distance from the model's own controls.

Open weights you can run. Wan from Alibaba, HunyuanVideo from Tencent, LTX-Video from Lightricks, CogVideoX, Mochi. Usually driven through ComfyUI. These are behind the hosted flagships on raw quality and they are catching up faster than most people expect.

Names and versions here are as they stood in 2026. Check before you commit money.

Read a leaderboard for what it actually measures

Public arenas rank blind pairs: which of these two clips looks better. That is one axis. It is not the same as whether the model did what you asked, whether it holds a face, or how often it hands you something unusable.

A model can win on aesthetics and lose your afternoon on prompt adherence. Those are separate properties and they are not correlated in any dependable way.

Demo reels are worse. A demo reel is the top one per cent of thousands of generations, chosen by people who know the model's strengths and steered away from its weaknesses. It tells you what the ceiling looks like. It tells you nothing about your hit rate.

How to actually choose, in an hour

Take one genuinely hard shot from your own list. Same still, same motion instruction. One generation on each of three candidates. Then answer four questions in writing:

- Did it follow the camera instruction, or invent its own move?
- Did the face and the wardrobe survive?
- How bad is the last second?
- What did it cost?

That hour beats every comparison video on the internet, because it is measured on your work rather than on somebody else's best case.

Open weights: what you gain and what you pay

You gain no per-second bill, seed determinism — same seed and settings, same output, which hosted APIs frequently will not promise — no content filter unexpectedly refusing a legitimate shot, and the ability to train a character model of your own. See fine-tuning and running-models-yourself.

You pay in time and VRAM. A comfortable local setup is a 16 to 24GB card. Quantised builds run on 8 to 12GB, slowly — minutes per clip rather than seconds. On a laptop with integrated graphics this is not happening locally at all, and no amount of patience changes that.

The free path that genuinely works: Hugging Face Spaces host browser demos of most open video models with no install, queueing behind everyone else; Google Colab's free tier and Kaggle notebooks give time-limited GPU sessions you can run ComfyUI or a diffusers script in. On a phone, a Space in a browser tab is the entire workflow, and it works.

Free tiers on hosted tools

Most hosted platforms give a small allowance, often refreshing daily. Using several of them is legitimate and it is how a large number of people learn this. Two rules make that allowance go further: spend it on image-to-video rather

than text-to-video, and spend it on real shots from a real list rather than on tests you will throw away regardless.

What survives the next release

Shot list. First frame. One move per clip. Generate many, discard most. Cut in an editor. Know your price per second. Every one of those transfers intact to a model that does not exist yet, which is why they are what this course is mostly about.

BEFORE YOU MOVE ON

A studio switches to the model currently topping the public preference leaderboard and finds it ignores their camera instructions more often than the model they left, even though individual frames look better. What does this most likely reveal?

A. The leaderboard scores which clip looks better in a blind pair, which is a different property from prompt adherence; the two are not reliably correlated.

B. Their prompts were written in the previous model's dialect and simply need retranslating for the new one.

C. The new model defaults to a lower generation tier, and adherence improves on the paid high-quality setting.

D. Leaderboard voting is gamed by the vendors, so the ranking does not reflect real quality at all.

10 · 10 min

What it takes to run one on your own card

THE ONE THING TO KEEP

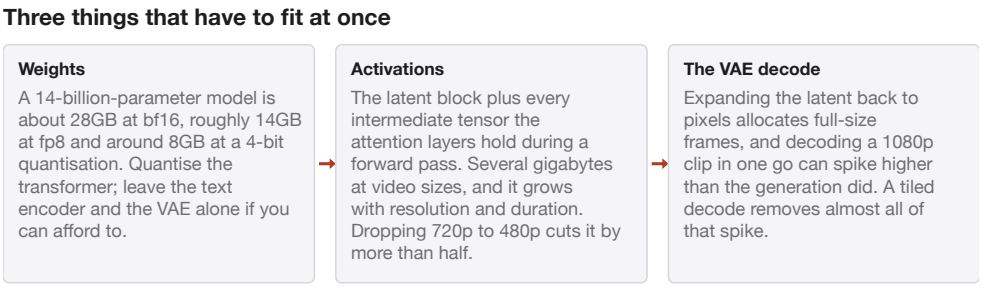
A video model's VRAM requirement is weights plus activations plus the decode, and quantisation and offloading trade speed for fit — which is why a 14-billion-parameter model can run on 8GB slowly or not at all on a phone.

The arithmetic before the download

Three things have to fit in VRAM at once, and people usually count only the first.

1. **Weights.** Parameters times bytes per parameter. A 14-billion-parameter model at bf16 is 2 bytes each: about 28GB. At fp8 it is roughly 14GB. At a 4-bit GGUF quantisation, around 8GB.
2. **Activations.** The latent block plus every intermediate tensor the attention layers hold during a forward pass. This scales with resolution and duration, and at video sizes it is not small — often several gigabytes.
3. **The VAE decode.** Expanding the latent back to pixels allocates the full-size frames. Decoding a 1080p clip in one go can spike higher than the whole generation.

So "the model is 14GB" and "14GB of VRAM is enough" are different claims, and the second one is usually false.



People count the first and are surprised by the other two. Leave three or four gigabytes of headroom over the weights, or the run dies at the decode after you have paid for the whole denoise.

The four things that make it fit

- **Quantisation.** fp8 roughly halves the weights against bf16, with a small quality cost. 4-bit GGUF halves again, with a larger one — motion tends to soften and fine detail degrades before anything looks obviously broken. Quantise the transformer; leave the text encoder and VAE alone if you can afford to.
- **Offloading.** Keep most weights in system RAM and stream the needed blocks to the GPU. ComfyUI's block-swap and the `--lowvram` family of flags do this. It works and it is slow, because you are moving gigabytes across PCIe every step.
- **Tiled VAE decode.** Decode the output in overlapping tiles instead of all at once. Removes the decode spike almost entirely, at the cost of faint seams if the tiles are too small.
- **Smaller outputs.** 480p instead of 720p cuts activation memory by more than half. Generate drafts small. This is the cheapest lever and the one people reach for last.

What runs on what, honestly

Rough shape as of 2026, and it moves:

- **24GB (e.g. a 3090 or 4090).** Comfortable. A 14B model at fp8, 720p, five seconds, in roughly one to three minutes.
- **16GB.** Workable with fp8 and tiled decode. Expect a few minutes per clip.

- **12GB.** Quantised models at 480p, with offloading. Several minutes per clip. Usable for learning and draft passes, painful for production.
- **8GB.** 4-bit quantisations, low resolution, heavy offload. Minutes per clip, frequent out-of-memory failures. Possible, not pleasant.
- **Apple Silicon.** Unified memory helps with fit; MPS support in video pipelines is behind CUDA and considerably slower. A 32GB or 64GB Mac will run things a 12GB PC cannot, and will take longer doing it.
- **Integrated graphics, or a phone.** No. Not slowly — not at all. This is not pessimism, it is the memory arithmetic.

The free GPU paths that actually work

If you do not own a card, three routes are real:

1. **Hugging Face Spaces.** Browser demos of most open video models, no install, no account needed for many. You queue behind everyone else and you cannot change much. On a phone this is the entire workflow, and it genuinely works.
2. **Google Colab free tier.** A T4 with around 15GB, session limits, and disconnections. Enough for quantised models at low resolution. ComfyUI runs in Colab through community notebooks.
3. **Kaggle notebooks.** A weekly GPU quota, typically around 30 hours, on T4s. More generous session length than free Colab and less prone to mid-render disconnection, which matters when a render takes four minutes.

The honest cost of all three is queue time and the risk of losing a session at eighty per cent. Use them to learn and to make your own work. Do not promise a client a deadline that depends on somebody else's free queue.

Why you would bother at all

Running locally is slower and more fiddly than a hosted endpoint. The reasons people do it anyway are specific:

- **No per-second meter.** Experimentation becomes free, which changes how much you experiment.

- **Real determinism.** Same graph, same seed, same output — which hosted APIs will not promise.
- **No unexpected refusal.** Content filters on hosted services occasionally decline legitimate work with no explanation and no appeal.
- **You can train.** A character LoRA needs the weights. That is covered later in the course, and it is the strongest identity tool available.
- **The model does not change under you.** A checkpoint on your disk is the same checkpoint next year.

The setup, in one paragraph

ComfyUI is the practical front end: a node graph, an enormous community library, and workflows shared as JSON files you can drop in. Install it, download a model in the format your VRAM allows, load a known-good workflow from the model's own page rather than building one from scratch, and change one node at a time. Everybody's first week is spent on dependency versions and out-of-memory errors; that is normal and it ends.

More on the general mechanics in running-models-yourself.

Today: look up your GPU's VRAM, then check the model card for the quantisation that fits in it with three or four gigabytes to spare for activations. If nothing fits, open a Hugging Face Space instead and lose nothing.

BEFORE YOU MOVE ON

Someone with a 16GB card downloads a 14-billion-parameter video model quantised to fp8, reasoning that 14GB of weights leaves 2GB spare. The first generation fails with an out-of-memory error during the final stage. What did the estimate miss?

A. fp8 weights are unpacked to bf16 on load, so the model actually occupies 28GB during inference.

B. Video models reserve a fixed fraction of VRAM for the text encoder, which must stay resident throughout generation.

C. The card cannot allocate its full 16GB because the display driver and the operating system reserve part of it, leaving roughly 13GB usable.

D. Activations during attention and the full-size VAE decode need VRAM beyond the weights, and the decode is the largest single spike.

CASE STUDY · MODULE 1

The twenty-second shot the venue insisted on

A two-person wedding and events film unit in Nashik, hired by a banquet venue that wanted one unbroken twenty-second aerial-style reveal of its lawn for the top of its website.

Rehan and Sneha had shot the venue on the ground twice. They had no drone permission for the site and no budget to hire one, so the plan was a generated reveal: start tight on the marigold arch, lift, and pull back to show the lawn, the lit mandap and the hills behind.

The brief said one continuous take. The venue's owner was explicit about it. He had seen a competitor's website with a single flowing shot and he believed cuts would make his lawn look smaller, which is not an unreasonable instinct for a man selling space.

The model gave them five-second clips. They tried the obvious thing first and pressed extend three times.

What came back was a twenty-second file that got worse as it went. The first five seconds were the shot they had paid attention to. By the third extension the marigolds had gone orange-red, the lawn had warmed by what Sneha measured as a noticeable shift on the parade scope, and the fine lattice on the mandap had smoothed into something painted. Nothing in the prompt had changed. Nothing in the settings had changed.

Rehan understood why once he thought about what extend does mechanically. The button takes the last frame of the previous generation — a frame that has already been decoded out of the latent and is therefore already lossy — and re-encodes it as the conditioning image for a fresh run. Three extensions is a photocopy of a photocopy of a photocopy. Contrast climbs, saturation creeps, high-frequency texture dies. There was no prompt that could have prevented it because the damage was not in the prompt.

That left a decision with a cost on each side.

They could deliver the twenty-second take and colour-match the four sections back together in the edit. That is a real technique and it works. But the join sits inside a continuous camera move, which is the worst place for a mismatch to live, because the eye has nothing else to attend to and any residual difference reads as a glitch rather than as a change of shot. Sneha estimated an evening of secondaries, and even then the lattice would still be soft in the last third.

Or they could cut. Four shots, four to five seconds each: the arch, the lift, the lawn, the hills. Each one generated fresh, each one a clean first frame, each one ending before the drift. That would cost them a conversation with a client who had already said no to cuts.

They chose to have the conversation, and they prepared for it rather than arguing. They built both versions. The extended one was ready first. Then they built the cut version and put a single grade and a light grain across all four shots, with a slow push on each so the movement carried through the joins, and laid the venue's own ambience — recorded on a phone at the site, birds and a generator in the distance — continuously underneath so the audio did not break at the cuts.

Then they sent both, unlabelled as to which was which, and asked the owner which looked more expensive.

The honest part of this story is that they were not certain he would pick theirs. The cost of being wrong was a re-edit they would not be paid for, and a client who felt overruled.

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 owner picked the cut version without hesitation and said it looked like it had been shot properly, which was the word he used. He had never objected to

cuts as such; he had objected to a piece that felt small, and four well-graded shots with continuous sound felt larger than one drifting take. Rehan wrote three lines in his notes file afterwards: never more than one extension, generate ten seconds to deliver six, and show two versions rather than argue about one. The venue booked them for a second film in the same month.

Why did the colour and texture degrade across the three extensions when nothing in the prompt or the settings changed?

Extending conditions a new generation on the last decoded frame of the previous one. That frame has already been through the autoencoder once, so it has lost fine texture and picked up whatever bias the decoder has. Re-encoding it feeds that loss back in as the starting point, and the loss compounds with each pass. It is a photocopy of a photocopy: contrast climbs, saturation creeps, and high-frequency detail like the mandap lattice smooths into a painted look. No prompt can address it because the damage happens in the encode-decode round trip, not in the text conditioning.

The team could have colour-matched the four extended sections together in the edit. Why is cutting between them a stronger answer than matching them?

A residual mismatch inside a continuous camera move is glaring, because the eye is tracking one uninterrupted motion and any change in colour or texture arrives as a fault. The same mismatch either side of a cut is largely invisible, because a cut is already a change and the viewer expects the picture to be different. Cutting also lets each shot start from a fresh, sharp first frame rather than a degraded one, so the texture problem is prevented rather than corrected.

What did generating four shots instead of one long take cost them, and was it worth it?

It cost them a difficult conversation with a client who had specified a continuous take, plus the work of building two versions rather than one, plus four first frames instead of one. It was worth it because the alternative was an evening of secondary colour work on a shot that would still have had a soft final third, and because the cut version could be assembled from clips that each ended before the drift. The general lesson is to plan in shots and put the cut where the model is weakest, which is also ordinary film grammar.

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 A progress bar shows coloured fog rather than the first second of your clip. What does that tell you about how the model works?

A. There is no first second yet: the whole block is denoised at once
B. The frames are generated in order, but the decoder holds them back until colour matching finishes
C. Frame interpolation has to complete before any frame can be displayed
D. The provider buffers the file until the whole generation has been paid for

2 In image-to-video, why is the final second of a clip reliably the worst second?

A. Interpolation runs out of reference frames before it reaches the end
B. The model lowers its quality near the end to save compute
C. Error grows with distance from the one fixed anchor, the supplied first frame
D. The autoencoder compresses frames near the end of the clip more aggressively than those at the start

3 A generated clip contains a shop sign whose letters change shape every frame. What does that tell you about the pipeline?

A. The prompt did not name a font, so the model picked a different one each time
B. The sampler needs more steps to resolve the letterforms
C. Guidance was too low for the text instruction to be obeyed
D. The letters were never in the latent; the decoder reinvents them each frame

4 The model ignored your instruction that the woman turns left. You raise guidance from 5 to 9. The most likely result is that the clip

A. shows her turning left, at the cost of slightly stronger contrast in the highlights
B. still does not show her turning left, and is stiffer with cooked colour
C. is unchanged, because guidance only affects the cost of the run
D. runs at a higher step count, which the interface sets automatically

5 You saved the seed, the prompt and every setting for a shot generated on a hosted service. Six weeks later the same call returns a different clip. The most likely explanation is that

A. a seed fixes only the first frame's noise, and the rest is drawn fresh every run
B. seeds expire after a fixed period on hosted services
C. the provider batched, rerouted or quietly updated the model
D. the text encoder normalises prompts differently on each request

6 A 14-billion-parameter model at fp8 is about 14GB. Why is a 14GB card not enough to run it?

A. Activations and the VAE decode have to fit as well, and the decode spikes highest
B. The operating system reserves half of any card's memory for the display
C. The text encoder is larger than the video transformer and has to stay resident for the whole run
D. fp8 weights are expanded back to bf16 when they are loaded

MODULE 2

The first frame, and everything decided in it

In image-to-video the still is the contract: framing, light, wardrobe, style and every flaw persist through the clip and then degrade. This block treats the frame as production work rather than a prompt result — composing for a move that has not happened yet, fixing the subject by hand in a free editor, compositing a real object in before any motion exists, controlling where the shot lands with an end frame, getting a camera move with no video model at all, and reading a frame in advance for the second it will fail.

> BY THE END OF THIS MODULE YOU CAN
>
> Produce a first frame at the model's native ratio and resolution with a stated light direction, a hand-repaired subject and any real object already composited in, predict from that frame alone which part of the clip will fail, and store it with enough provenance to rebuild the shot six weeks later

11 · 9 min

The still frame is where the decisions belong

THE ONE THING TO KEEP

Every decision you can make in a still costs a hundredth of what it costs to make in a clip, so fix the frame before you ever spend on motion.

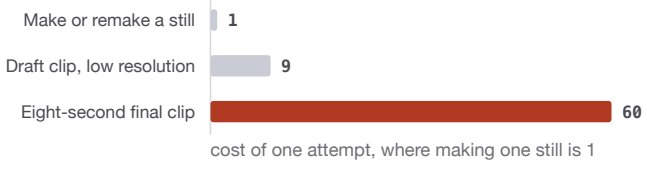
Two ways in, and they are not equal

Text-to-video: you type a description and the model invents the framing, the faces, the wardrobe, the colour, the light and the lens, then animates its own invention. You are gambling on a dozen decisions simultaneously, at video prices.

Image-to-video: you supply the first frame. The model's only job is motion.

Almost everyone doing this professionally works the second way, and the reason is economic before it is aesthetic. A still costs a cent or two and takes seconds. An eight-second clip costs fifty to a hundred times that and takes a minute or two in a queue. Image-to-video moves every hard decision out of the expensive loop and into the cheap one.

What one attempt costs, by where you make the decision



Repairing a still by hand is not on this chart because it costs nothing at all, and it is deterministic besides. Every decision you can move out of the right-hand bar and into the left one is money you keep, which is the whole argument for the still-first workflow.

There is a second reason that matters more. A still is not only cheaper to regenerate — it can be repaired by hand. You can photograph it. You can composite the client's real product into it in Photopea, which runs in a browser tab, opens PSD files and needs no install. You can fix a bad hand in GIMP or

Krita. You cannot do any of that inside a video model. Whatever you fix in the still is fixed for the whole clip.

For making the still, see ai-images. For repairing it, image-editing.

The first frame is a contract

Everything in that frame persists and then degrades. Six fingers in the still means six fingers for eight seconds, getting stranger. A wobbly sign in the still becomes a melting sign. A face you are not quite happy with in the still is a face you will hate by second five.

The frame also fixes the style. The model will not restyle mid-clip — it animates what it is given. So the still is where you decide whether this is a photograph, a gouache painting or a 3D render, and that decision is then locked in a way no prompt word can override.

The practical rule: do not animate a still you would not be happy to publish as a still.

Match the aspect ratio and the resolution

A quiet, common waste of money. You compose a still you like at 3:2. The video model works at 16:9 and either crops it, pads it or stretches it, and the framing you chose is gone.

Generate the still at the ratio the video model outputs — 16:9 for landscape, 9:16 for vertical short-form, 1:1 where a platform wants it — and at or a little above the model's native resolution. Most generate at 720p or 1080p. Feeding in a 4000-pixel still does not buy you detail; it gets downscaled on the way in.

Start and end frames

Several tools let you supply both the first and last frame — Luma's keyframes, Kling's start-and-end frames, and similar controls elsewhere. This is the most underused control in the whole category. It lets you decide where the shot lands, which is what a real camera operator does.

Use it for a reveal, for a match cut into the next shot, or for arriving exactly on a product hero frame.

The failure mode is worth knowing: if the two frames are too different, the model does not invent a camera move between them, it invents a morph. Keep them the same scene, same light, same lens, with a plausible path between. A push in works. A cut from a kitchen to a beach does not.

What text-to-video is still good for

It is not useless. When nothing is decided yet, text-to-video is a fast way to find a look, generate options and see what the model is good at before you commit. Concept phase, not production phase.

There is also a real trade-off on sound. Some models generate synchronised dialogue and ambience along with the picture, and on several of them that native audio path is tied to text prompting rather than to an image you supply. If you need the model's own audio, you may have to give up the control of a fixed first frame. Decide which matters more for that shot.

The free path

- Stills: several image tools give a daily free allowance, and open models run in a browser through a Hugging Face Space with no install. See hugging-face.
- Fixing stills: Photopea in a browser handles layers, masks and PSDs on a laptop that cannot run Photoshop. GIMP and Krita if you can install. On a phone, Snapseed for tone and a browser tab for the rest.
- Video: most hosted tools give a small allowance that refreshes. Spend it on image-to-video, at the lowest resolution, on real shots from your list. You learn far more per free credit that way than by rolling text-to-video prompts.

Today: take one shot you want. Make the still. Fix one thing in it by hand — the label, the hand, the sign. Then animate it, and compare that clip to one

generated from text alone.

BEFORE YOU MOVE ON

A drinks brand's product label comes out as gibberish in every generated clip. The team keeps rewriting the video prompt with more precise wording about the typography. What is the better fix?

A. Add explicit negative-prompt terms for warping and blur so the letterforms stop degrading across frames.
B. Generate at the highest available resolution so the letters have enough pixels to resolve correctly.
C. Composite the real label into the first frame in an image editor, then animate that frame, because the video model propagates what it is handed rather than rendering text it never had.
D. Switch to text-to-video so the model designs the whole bottle and label coherently from scratch.

12 · 9 min

Compose for the frame you will end on

THE ONE THING TO KEEP

A camera move changes the framing, so the still has to be composed for where the shot ends rather than where it starts, and every edge of the frame is a place where the model has to invent what enters.

The still is the first frame, not the shot

People compose a beautiful still, add a slow push-in, and are disappointed. The push-in cropped fifteen per cent off every edge, so the composition they liked is gone by second three and what remains is a tighter, worse version of it.

Compose for the end of the move. If the shot pushes in, start looser than feels right. If it pulls out, start tighter. If it tilts up, leave the headroom you will need at the top. This is ordinary camera-operator thinking and it transfers exactly.

A practical way to do it: decide the move first, then open the still in any image editor and draw a rectangle where the frame will end up. If the composition inside that rectangle does not work, the still is wrong, not the prompt.

Headroom, lead room, and why models get them wrong

Two classical rules the model has no opinion about:

- **Headroom.** The gap above the head. Too much and the subject sinks; too little and they feel cramped. Models trained on centred stock footage default to too much.
- **Lead room.** A person looking or moving to the left needs space on the left to look into. Without it the frame feels like it is pushing them into the edge.

The model will not fix either. It animates the composition it is given. Fix them in the still, where a crop costs nothing.

Every edge is a place things come from

This is the composition rule specific to generative video, and it does not exist in still photography.

When the camera moves, new picture has to enter at the leading edge. The model invents it. It has no idea what was outside your frame, because nothing was — the still ends where it ends.

So a pan to the right invents everything on the right. A pull-out invents a ring of new scene on all four sides. A tilt up invents a ceiling or a sky.

What follows:

- **Put the difficult content in the middle.** Anything with structure — a building edge, a bookshelf, a patterned floor — should not be sitting half in

and half out of frame, because the half that is outside will be invented and will not match.

- **Prefer moves that reveal simple things.** Tilting up into plain sky works. Tilting up into an ornate ceiling produces melted ornament.
- **A push-in invents nothing.** It only crops. That is why it is the safest move in the category and why models default to it.

Orbits and the flat-backdrop problem

An orbit or a parallax move asks the model to show the scene from an angle it was never given. If the still is a subject standing in front of an out-of-focus wall, there is no space behind them to reveal, so the model invents one as it goes and the shot reads as a cardboard cut-out sliding across a painted flat.

The fix is in the still: give the space depth before you ask the camera to move through it. Foreground element, mid-ground subject, background with real recession. A doorway, a table edge, a lamp in front, a corridor behind. A still with three planes orbits convincingly. A still with one does not, regardless of the model.

Aspect ratio is a composition decision, not an export setting

Match the model's native ratio when you make the still — this was covered before and is worth repeating because it is the most common waste. But there is a second point.

If you need both a 16:9 and a 9:16 version of a piece, you cannot crop one from the other and keep the composition. A vertical reframe of a landscape shot either loses the sides or scales everything up until it is soft.

Two honest approaches. Either compose the landscape still with a vertical safe area in mind, keeping everything essential inside a centre column — which costs you the sides as usable composition — or generate the shot twice, once in each ratio, from two stills. The second costs more money and looks considerably better, and for anything with a subject it is usually the right call.

Safe areas, still

If a platform overlays a caption, a profile picture and three buttons on the lower third of a vertical video, then the lower third of your composition is not yours. Check where the interface sits on the platform you are delivering to, and keep faces and text out of it. This is unglamorous and it is the difference between a shot that works in the app and one that works only in your editor.

The free path

Everything here is a crop, a guide and a rectangle. Photopea in a browser does all of it and opens PSDs. GIMP and Krita if you can install. On a phone, any editor with a crop tool and the ability to draw a box. The thinking is the expensive part and it costs nothing.

Today: take a still you intend to push in on. Draw the end frame on it. Decide whether the shot you actually want is the one inside that box.

BEFORE YOU MOVE ON

A generated orbit around a seated figure looks like a flat cut-out sliding across a painted background. What was wrong with the still?

A. The still had only one plane of depth, so the move asked the model to reveal a space the frame never contained.

B. The orbit was too fast for the model's motion range, and a slower rotation would have preserved the geometry of the scene properly.

C. The subject was not centred, so the orbit rotated around the wrong point and the parallax was computed against the wrong axis.

D. The background was out of focus, and blurred regions carry too little signal for the attention layers to keep them consistent across the move.

13 · 9 min

One light direction per scene, written down

THE ONE THING TO KEEP

Light direction is the strongest continuity signal an audience reads and the easiest one to lose between generations, so it should be decided once for a scene, written down, and put into every first frame before any motion is generated.

Why light breaks a cut before a face does

Two shots can contain the same person, the same wardrobe and the same room, and still refuse to cut together because the key light is on the left in one and the right in the other.

Audiences read this instantly and cannot name it. The feeling is "something is off", and they usually blame the performance or the face. Editors call it continuity, and on a real set an entire department exists to protect it.

You have no such department. You have a line in a text file, and it has to do the same job.

The four properties worth deciding

Decide these once per scene, before you make a single still:

1. **Direction.** Where is the key coming from — camera left, camera right, behind, above? Write it as a compass or a clock: "key from camera left, roughly 45 degrees, slightly above eye level".
2. **Quality.** Hard or soft. Hard light means a sharp-edged shadow, a single small source, high contrast. Soft means a wrapped, gradual shadow from a large source. This is the difference between midday sun and an overcast window, and it changes the whole feel.

3. **Colour.** Warm or cool, and whether there is a second colour from a practical light in shot — a lamp, a screen, a sign.
4. **Contrast ratio.** How dark the shadow side is. Low contrast reads as documentary and daylight; high contrast reads as drama and night.

Four short phrases. Put them at the top of your shot list and repeat them in every single still prompt.

Why the prompt words for light actually work

Unlike mood adjectives, light words appeared in captions constantly and precisely: "warm window light from the left", "overhead fluorescent", "backlit at golden hour", "single hard key, deep shadows", "soft overcast daylight". Automatic captioners describe light because light is visible.

So this is one of the places where prompting is genuinely reliable. Name the direction, the quality and the colour, and you will usually get them.

Motivate the light inside the frame

A shot where you can see *why* the light is there holds up better than one where it just exists. A window on the left, a lamp on the table, a screen glowing on the face, a doorway behind.

This matters for a specific technical reason as well as an aesthetic one. If a light source is visible in the still, the model has an anchor for it and tends to keep the lighting stable through the clip. If the light has no visible cause, it drifts — the shadow side brightens, the key wanders, the colour creeps warm — because nothing in the picture is holding it in place.

What the model does badly with light

Two honest limits.

It will not change the light convincingly mid-clip. Ask for a lamp being switched on and you generally get a global brightness ramp rather than a light appearing from a position and casting new shadows. There is no light trans-

port being computed; there are plausible frames being sampled. If a shot needs a lighting change, do it as two shots with a cut, or do it in the grade.

Reflections and speculars drift. The bright point on a metal edge or in an eye moves as the camera moves, and getting that right requires knowing where the light and the surface are in three dimensions. The model approximates. On a wet street or a car body this is where the shot gives itself away, and it is a good reason to prefer matte surfaces in anything you need to hold up.

Checking light without expensive tools

You do not need a trained eye, you need a histogram and a side-by-side.

- Open two stills next to each other at the same size. Mismatches in direction are obvious at a glance and invisible from memory. Always compare, never recall.
- DaVinci Resolve's free tier includes **false colour** and scopes. Drop two stills on a timeline, switch to false colour, and the exposure difference becomes a flat readable map. GIMP's histogram will show you a gross brightness mismatch for nothing.
- On a phone, put two frames in a split-screen or a collage app. Crude, and it works.

The grade can rescue colour, not direction

A warm-versus-cool mismatch between two generations is fixable in the edit: a shot-match pass in Resolve pulls one toward the other in a minute. That is a real safety net and you should use it.

What the grade rescues, and what it cannot

Fixable later, in the edit

- A warm shot against a cool one: a shot-match pulls one toward the other in a minute
- A whole frame that came back too dark or too bright
- Saturation that drifted between two generations
- A colour cast running across a sequence of exteriors

Only fixable by remaking the still

- A key light from camera left in one shot and camera right in the next
- A shadow falling the wrong way across a face
- A hard shadow edge in a scene you decided was soft and overcast
- A highlight sitting on the wrong side of a metal edge

Colour is a correction and direction is geometry. That is the whole distinction, and it is why the four light phrases go into the file before the first still is made rather than into the edit afterwards.

What no grade fixes is a shadow falling the wrong way. That is geometry, not colour, and the only remedy is regenerating the still. Which is why the four phrases go in the file before you start, not after.

Today: write your scene's four light phrases on one line. Put that line into every still prompt for that scene, and put the finished stills side by side before you animate any of them.

BEFORE YOU MOVE ON

Two generated shots of the same character have visibly different colour temperature and the shadow falls on opposite sides of the face. A colourist offers to shot-match them. What will that achieve?

A. Nothing useful, because shot matching operates on the whole frame and cannot separate the subject from the background.

B. Both problems, since matching the colour of the key light also rebalances the apparent direction of the shadows across the two frames.

C. It will fix the colour temperature and leave the shadow direction exactly as it is, so the shots still will not cut together.

D. It will fix the shadows by lifting the dark side of the face, though the colour difference will need a separate secondary correction afterwards.

14 · 10 min

Fix the frame yourself instead of re-rolling it

THE ONE THING TO KEEP

Repairing a still by hand is deterministic and costs one attempt, while re-generating is a fresh random draw that may fix one flaw and introduce two, which is why professionals reach for an editor long before they reach for the regenerate button.

Why the repair pass exists

A generated still comes back almost right. Six fingers, a warped sign, a doubled earring, a chair leg that passes through the floor. The instinct is to regenerate.

Regenerating is a fresh sample. It may fix the hand. It may also change the face, move the light, lose the wardrobe detail you liked and introduce a new flaw somewhere you were not looking. You are rolling dice on the whole frame to fix one square inch of it.

Repairing is deterministic. You fix the hand and nothing else changes. It takes five minutes, it costs nothing, and the result is exactly what you decided rather than what the sampler decided.

This is the single biggest habit difference between people who ship generative video and people who burn credits.

The order to fix things in

Work from most damaging to least, and stop when the frame is good enough to animate.

1. **Hands and contact points.** Fingers, where a hand grips an object, where a foot meets the ground. These become worse under motion, never better.

2. **Faces.** Asymmetric eyes, a wandering pupil, teeth. Small fixes here pay enormously because you are about to watch this face for eight seconds.
3. **Text and logos.** Almost never fixable by nudging. Either paint them out flat and add real text later in the editor, or accept them.
4. **Structural nonsense.** A railing that does not connect, a doorframe with three sides, a reflection that does not match. The eye finds these on the second viewing.
5. **Duplicated background people.** Models produce twins in crowds. One clone stamp each.
6. **Edges and halos.** Leftover fringing where the model composited badly.

The tools, and the honest free path

- **Photopea** runs in a browser tab, opens and saves PSD files, and has layers, masks, clone stamp, healing and selection tools that are close enough to Photoshop that the muscle memory transfers. It is the answer for anyone on a machine that cannot run Adobe, which is most of this readership. Free tier is ad-supported and fully functional.
- **GIMP** if you can install software. Heavier learning curve, no subscription, everything you need.
- **Krita** is better than GIMP for painting — if the repair involves drawing rather than cloning, use it.
- **Photoshop** is named here because job ads name it. It is not required for any technique below.
- On a **phone**: Snapseed's healing tool handles small removals; Photopea works in a mobile browser badly but it works.

The four operations that do ninety per cent of it

Clone stamp. Copy pixels from one part of the image to another. Removes a duplicated person, a stray object, a bad detail. Sample from a region with the same light and the same texture direction, and use a soft brush.

Healing brush. Like clone, but blends the sampled texture into the destination's colour and luminance. Better for skin and skies, worse near hard edges,

because it drags contrast from across the edge.

Selection plus patch. Lasso the bad area, copy a good area from elsewhere in the same image, paste, flip if needed, mask the edges. This is how you replace a bad hand with a good hand from another generation of the same subject.

Paint on a mask. Almost everything else. Paint black to hide, white to reveal, on a layer mask rather than on the pixels, so you can change your mind.

Inpainting as a repair tool

Most image tools offer generative fill on a masked region. It is genuinely useful for the middle case: too big to clone, too small to regenerate.

Two cautions worth carrying. It samples, so it is another dice roll — but a small, cheap, local one, which is a much better bet than re-rolling the whole frame. And the inpainted region comes back through the same autoencoder, so it can be very slightly softer or differently grained than its surroundings. On skin nobody notices. On a hard edge against a flat wall, sometimes they do.

Borrow from your own rejects

Keep the failed generations. When shot four's still has a perfect face and a ruined hand, and shot two's still has the same face and a good hand, the hand is a copy-paste away. Same subject, same light, same model — the pieces fit because they came from the same distribution.

This is why the frame library in a later lesson is not administrative tidiness. It is a parts bin.

What not to repair

If the pose is wrong, the composition is wrong, or the light is wrong, stop repairing. Those are structural and you will spend an hour producing a worse result than a new generation with a better prompt. The repair pass is for local flaws in a frame that is fundamentally right.

The test: can you describe the fix as a shape you could draw around? If yes, repair. If the answer involves the whole frame, regenerate.

Today: take your worst nearly-good still, open it in Photopea, and fix exactly one thing with the clone stamp. Time yourself. Compare that against the cost of four more generations.

BEFORE YOU MOVE ON

A still is perfect except for a malformed left hand. Why do experienced operators mask and repair rather than regenerate with the same prompt?

A. Regeneration is billed per image and the repair is free, so over a thirty-shot project the saving is substantial.

B. The autoencoder degrades the frame slightly on every regeneration, so a frame that has been sampled twice is measurably softer than one sampled once.

C. A repair changes only what you selected, whereas regenerating draws a new sample of the whole frame and may alter the face or the light too.

D. Hand geometry is one of the features the model is least able to improve between attempts, so successive generations will reproduce the same malformation.

15 · 10 min

Put the real object in before you animate it

THE ONE THING TO KEEP

A model has no reference for your client's specific product or person, so the professional split is to generate the environment and the camera move and composite the real thing into the still — because a nearly-right label reads as a counterfeit rather than a stylisation.

The problem, stated precisely

Ask a model for "a bottle of Himalaya face wash on a bathroom shelf" and you get a bottle. It will be the right shape family, the right colour family, and the label will be almost right — which is worse than obviously wrong, because almost right reads as a fake.

The reason is not capability, it is information. The model has no representation of that specific bottle. No amount of prompting, steps or guidance retrieves data that is not there. This is the one limit in the whole subject that does not get better with scale.

The answer is not to try harder. It is to change the division of labour: **the model makes the world, you supply the thing.**

Two places to composite, and how to choose

Into the still, before generating. The object becomes part of the first frame and gets animated with everything else. Cheap, simple, and the object moves naturally with the scene.

Use this when the camera move is small — a slow push, a slight drift, a locked-off frame with ambient motion. The object will soften and may distort slightly as the clip runs, which a short shot and a light move will hide.

Into the finished clip, in the editor. Generate the environment and camera move with a stand-in or an empty surface, then track the real object in afterwards with motion tracking.

Use this when the object must stay pin-sharp and exactly itself, when the camera move is large, or when the client will look at it frame by frame. This is more work and it is the only version that survives scrutiny.

The honest rule: anything a client is paying for a close-up of goes in during the edit, not in the still.

Making a composite hold up in the still

Five things, in order. Get the first three right and most people never notice the rest.

1. **Perspective.** The object's viewing angle must match the scene's. A bottle photographed straight on cannot sit on a table seen from above. Reshoot or regenerate the background — do not try to fix this by distorting, because distortion is exactly what the eye catches.
2. **Light direction.** The key light on the object must come from where the key light in the scene comes from. You decided that direction in the previous lesson; now you shoot the object under it. A window and a piece of white card is enough.
3. **Scale and contact.** The object needs a shadow where it touches the surface, and that shadow needs the same softness as other shadows in the frame. A composite with no contact shadow floats, and floating is the single most common tell.
4. **Colour and contrast.** Match the object's black point and colour cast to the plate. Nine times out of ten this is one curves adjustment.
5. **Grain and edge.** Generated frames are smooth; a real photograph has noise. Add matching grain to the composited object, and soften its cut-out edge by a pixel or two so it does not look laser-cut.

Shooting the object without a studio

This is where thirty seconds of real filming beats an afternoon of generating, and everybody can see that it did.

A phone, a plain wall or a sheet of paper, and daylight from a window. Turn the phone's flash off. Put the object a metre from the wall so its own shadow does not fall on the background. Shoot from the angle the composite needs, not the angle that is comfortable. Take twenty frames and pick one.

Cut it out in Photopea or GIMP with the pen tool for hard edges and a mask for soft ones. Background removal tools handle most objects in one click now and are worth trying first; they fail on hair, glass and thin structures, which is where you take over by hand.

Tracking it into a moving clip

When the object has to go into the finished footage:

- **DaVinci Resolve's Fusion page**, in the free tier, does planar and point tracking, and it is genuinely capable. The workflow is: track a feature in the plate, attach the object to the track, match grade, add grain.
- **Blender** does the same with its motion tracker and is free and open, with a steeper climb.
- **Kdenlive** has basic tracking; enough to stick a logo on a slow move, not enough for a hero product shot.

Generate the plate with a simple, high-contrast trackable feature in it — a table corner, a doorframe, a mark on the wall. A featureless white surface is untrackable, and you will discover this after you have paid for the clip.

The same logic applies to people

A real person's face works the same way. Film them against a plain wall on a phone, key them or mask them, and composite into a generated environment. This is a decades-old technique and it is the honest route to putting a specific real person into a generated scene — with their written permission, which is a separate matter dealt with later in the course.

Today: photograph one real object against a wall in window light, cut it out in Photopea, and drop it into a generated still. Judge only the contact shadow.

BEFORE YOU MOVE ON

A composited product sits in a generated kitchen. The perspective, colour and scale all match, yet it still reads as pasted on. What is the most likely cause?

A. The object was shot at a higher resolution than the generated plate, so its detail level is inconsistent with the surrounding pixels.
B. There is no contact shadow where the object meets the surface, so nothing anchors it to the plane it is supposed to be resting on.
C. The cut-out edge is too soft, and a softer edge than the surrounding focus makes an object read as a separate layer.
D. The generated plate has no grain, while the photograph does, and that texture difference is what the eye registers as a separate element.

16 · 9 min

Deciding where the shot lands

THE ONE THING TO KEEP

Supplying both a first and last frame turns generation from an open-ended sample into an interpolation between two fixed points, which is the most underused control in the category and the one that fails hardest when the two frames are too different.

What the control does

Several tools let you supply two images: the frame the clip starts on and the frame it ends on. Luma calls them keyframes, Kling calls it start and end frame, and similar features exist elsewhere under other names.

Mechanically, both images are encoded and both condition the denoise — one anchoring the beginning of the latent block, one anchoring the end. Everything between is generated to connect them.

This changes the nature of the task. Without an end frame, you are sampling a plausible continuation and hoping it goes somewhere useful. With one, you are asking for a path between two points you chose. It is the difference between letting a camera operator improvise and giving them a start mark and an end mark.

Four things it is genuinely good for

1. **Landing on a hero frame.** Product work, title cards, anything that has to arrive exactly somewhere. Generate the end frame first, perfect it, then decide where the shot starts.
2. **Match cuts.** Make shot A end on a composition and shot B start on the same composition with different content. The cut becomes invisible and the sequence feels designed.
3. **Reveals.** Start on a detail, end on the wide. Or the reverse. Both frames are compositions you controlled rather than a crop the model chose.
4. **Perfect loops.** Set the last frame identical to the first. The clip returns to where it began and can be played endlessly. This is worth real money for backgrounds, social posts and screen content, and it is nearly impossible to achieve any other way.

The failure that catches everyone

If the two frames are too different, the model does not invent a camera move between them. It invents a morph.

This is worth understanding rather than memorising. The model is filling in a smooth path through latent space. If the two endpoints are close — same room, same light, same subject, a plausible camera displacement apart — the smooth path through latent space corresponds to a smooth path through the real world, and you get a camera move. If the endpoints are far apart, the smooth latent path corresponds to nothing physical, and what you see is one

image dissolving into another through a sequence of impossible intermediates.

A kitchen and a beach produce a melt. A wide shot and a close-up of the same subject in the same light produce a push-in.

Rules that keep it working

- Same scene, same light, same time of day.
- Same lens character. A wide-angle start and a telephoto end is a perspective change the model cannot path through.
- A displacement a real camera could achieve in the clip's duration. Two metres in five seconds is fine. Thirty is not.
- Same subject, same wardrobe, same pose family. A person standing and then sitting is a physical action and often works; a person standing and then being a different person does not.
- **Build the end frame from the start frame.** Crop it, or edit it, or generate it with the start frame as a reference. Two independently generated images of "the same" scene are never quite the same scene, and the difference shows up as churn.

Using it without the feature

If your tool has no end-frame control, you can approximate the reveal and the match cut by other means:

- Generate a longer, looser move and find your landing frame in the edit, then trim to it. You pay in generated seconds and you get the frame you wanted.
- For loops, generate a clip with very little motion, then in the editor duplicate it, reverse the copy, and join the two. A ping-pong loop is not the same as a true loop — the motion visibly runs backwards — but on abstract or ambient content nobody notices, and it is free.

What it costs

Two frames means two stills to make and repair rather than one. That is real work. It is still an order of magnitude cheaper than generating a shot six times hoping it ends somewhere usable, and it is the difference between a sequence that feels edited and one that feels assembled from whatever came back.

The free path

End-frame control appears on several hosted tools' free tiers, and open-weights pipelines in ComfyUI expose first-and-last-frame conditioning directly as a node. If you have a free daily allowance, this is the feature to spend it on, because it teaches you more per credit than another round of prompt variations.

Today: make one still, then make its end frame by cropping the same image tighter. Generate with both. You have just directed a push-in to a frame you chose rather than one the model chose.

BEFORE YOU MOVE ON

A designer supplies a start frame of a café interior and an end frame of a rooftop at sunset, hoping for a transition. The result is a churning dissolve. Why?

A. The two frames were generated separately, so their seeds differ and the model cannot reconcile two independent samples into one trajectory.

B. Sunset lighting in the end frame conflicts with the interior lighting, and the model resolves lighting conflicts by cross-dissolving between them.

C. The model fills a smooth path between the two encoded endpoints, and when the endpoints are far apart that smooth path corresponds to no physical camera move.

D. End-frame conditioning only works within the model's temporal attention window, and a scene change exceeds the range over which the two anchors can influence each other.

17 · 9 min

A camera move with no video model at all

THE ONE THING TO KEEP

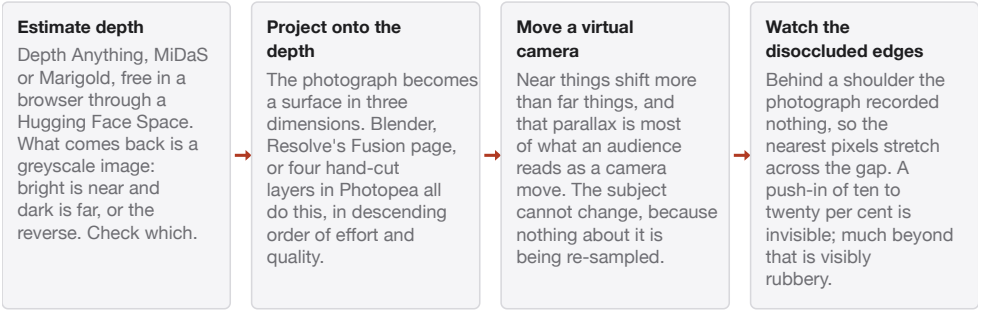
A depth map plus a 2.5D projection gives a camera move on a still image with perfect consistency and zero generation cost, which beats a generated clip outright whenever the subject must stay exactly itself.

The technique nobody mentions

You can move a camera through a photograph without generating any video.

Estimate a depth map for the image, project the pixels onto that depth as a surface, then move a virtual camera. Near things shift more than far things, which is parallax, and parallax is most of what reads as a camera move.

A camera move with no video model in it



Deterministic, free, and strictly better than generation whenever the subject has to stay exactly itself, because it is the original pixels. What it cannot do is move the world — for a person walking or cloth blowing you still need a generation.

This has been standard in documentary and motion graphics for years. It is not a compromise or a poor substitute. For a large class of shots it is strictly better than generation, because:

- **The subject cannot change.** It is the original pixels. A face stays exactly that face, a logo stays exactly that logo, a product stays exactly that product. No drift, no morph, no re-sampling.

- **It costs nothing.** No per-second meter.
- **It is deterministic.** Same input, same output, every time.
- **It works on archival material.** A photograph from 1950 cannot be re-shot and should not be re-generated. It can be moved through.

How to get the depth map

Monocular depth estimation — guessing depth from a single image — got very good very quickly. The models to know:

- **Depth Anything** (v2 and later) is the current practical default: fast, free, strong on ordinary photographs.
- **MiDaS** is older, still fine, and runs on almost anything.
- **Marigold** is slower and more precise on fine structure.

All are available as Hugging Face Spaces — upload an image in a browser, download a depth map, no install, works on a phone. They are also ComfyUI nodes if you run things locally.

What comes back is a greyscale image: bright is near, dark is far, or the reverse depending on convention. Check which.

Building the move

Three routes, all free.

Blender. Import the image as a plane, subdivide it heavily, apply the depth map as a displacement, and animate a camera. Full 3D control, the best-looking result, and the steepest learning curve. Blender is free and open and runs on modest hardware for this kind of work.

A node compositor. DaVinci Resolve's Fusion page, free tier, has 3D nodes: an image plane, a camera, and a displacement driven by the depth map. If you already have Resolve for editing, this costs no new software.

Layered cards, the crude version. Cut the image into three or four depth layers by hand in Photopea or GIMP — foreground, subject, mid, background — paint in what is behind each layer, then animate them at different speeds in

any editor that allows keyframed position and scale. Kdenlive and Shotcut do this. It is how the effect was done before depth models existed and it still works.

Where it breaks, and why

One honest limitation and it is fundamental.

When the camera moves, previously hidden areas come into view — **disocclusion**. Behind a person's shoulder there is nothing, because the photograph never recorded it. The projection stretches the nearest pixels across the gap, which produces a smeared, rubbery edge around foreground objects.

This sets the technique's range: small moves look perfect, large moves fall apart. As a rough guide, a push-in of ten to twenty per cent or a lateral drift of a few per cent of frame width is invisible. Beyond that the stretching shows.

Two mitigations. Paint behind the foreground object in the editor before projecting, so the gap has content to reveal — this is exactly what the layered-cards method does and why it looks better than a naive depth projection. Or use an inpainting pass to fill the hidden regions, which is the same idea automated.

When to choose this over generation

A decision rule you can apply on a shot list:

- **Use parallax** when the subject is specific and must not change, when the move is small, when the source is a photograph you cannot reshoot, or when there is no budget.
- **Use generation** when things need to genuinely move — a person walking, water flowing, cloth blowing, a crowd — because parallax moves the camera, not the world.

Many finished pieces mix both, and nobody watching can tell which shots were which. A title sequence of six archival photographs with slow parallax, cut against two generated establishing shots, is a completely normal and completely honest way to work.

Today: run one photograph through a Depth Anything Space, then build a five per cent push-in from it in whatever tool you already have. Compare the result against a generated clip of the same image.

BEFORE YOU MOVE ON

A depth-projected push-in on a portrait looks clean at ten per cent but develops a rubbery smear around the subject's shoulder at forty per cent. What is happening?

A. The depth map's precision falls off in the mid-tones, so larger moves expose quantisation steps in the estimated distances.

B. The camera has moved far enough to reveal regions the photograph never recorded, so the projection stretches nearby pixels across the gap.

C. The subdivision of the projection surface is too coarse for displacements of that size, and increasing the mesh density would remove the artefact entirely.

D. Monocular depth estimation is relative rather than metric, so a large camera translation applies an incorrect scale to the near-field displacement.

18 · 9 min

Predicting which second will break

THE ONE THING TO KEEP

Most clip failures are visible in the still before any money is spent, because the same nine features — contact, thin structure, text, reflection, crowd, pattern, transparency, hair and count — fail for reasons you already know.

A checklist you run before you spend

By this point in the course you know the mechanisms. This lesson turns them into a two-minute inspection you run on every still before it becomes a clip.

Open the frame at full size and look for nine things.

1. Contact points. Where a hand grips, a foot presses, a shoulder leans, an object rests. Contact is where a plausible frame becomes an impossible one, and the model has no constraint preventing a finger from passing through a strap. Count the contacts in the still. Each one is a place the clip can fail.

2. Thin structures. Railings, wires, chair legs, bicycle spokes, glasses frames, microphone stands. Thin means near the autoencoder's resolution floor, so these are reinvented each frame and visibly flicker.

3. Text. Any letters at all. If they matter, they are coming out and going back on in the editor.

4. Reflections and speculars. Mirrors, windows, polished floors, car bodies, eyes. These require knowing where the light and surfaces are in three dimensions. The model approximates, and the approximation slides as the camera moves.

5. Background faces. A crowd is a set of faces, each too small to be resolved properly and each subject to the same identity drift as your lead. Background people churn. Blur them in the still, or accept it.

6. Repeating patterns. Tiles, brickwork, fences, striped shirts, keyboards. Regular grids are where the model's failure to count becomes visible. Expect the count to change and the grid to warp.

7. Transparency. Glass, water, smoke, a veil. Transparent things require compositing two depths at once, and the model handles them as a texture rather than as a see-through surface. They tend to solidify or vanish.

8. Hair over a face. A partial occlusion that moves. Combines the two hardest things.

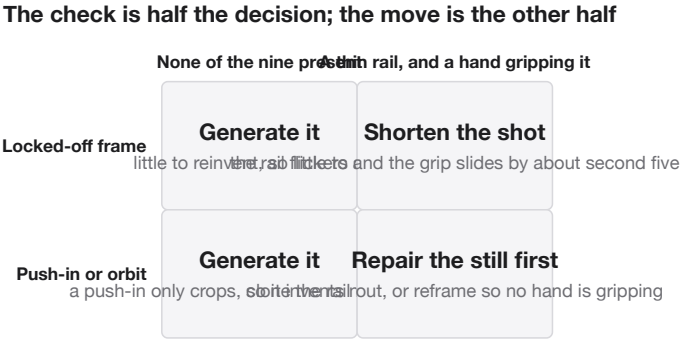
9. Anything you would have to count. Three birds, six candles, two rings, a hand with a specific number of fingers visible. Nothing is counting.

What to do with the list

For each item you find, pick one of four responses:

- **Remove it from the still.** Clone out the wire, blur the crowd, move the reflection out of frame.
- **Reframe.** A tighter shot that excludes the patterned floor. A different angle where the mirror is not visible.
- **Reduce the move.** Every item on the list gets worse with camera motion. A locked-off frame with a slight drift survives things a push-in does not.
- **Shorten the shot.** If the hand will fail at six seconds, the shot is four seconds long.

All four are cheaper than discovering the problem after generating.



Every item on the nine-point list gets worse with camera motion, because a move both lengthens the exposure to drift and asks the model to invent whatever enters at the leading edge. A locked-off frame survives things a push-in does not, which means shortening the move is as legitimate a response as repairing the frame.

Keep a log and find out how good you are

Write your prediction down before generating: which item you think will fail, and when. Then watch the clip and record what actually happened.

Twenty shots of this and you will have something genuinely valuable: a calibrated sense of your own model's weaknesses, in your own subject matter. You will also discover where you are wrong. Most people over-predict text failures and under-predict background churn, because text is the famous problem and background churn is the quiet one.

A plain text file with four columns of your own devising is enough. This is the same habit an editor builds about a lens or a camera body, and it is what sepa-

rates somebody who uses the tool from somebody who knows it.

The two-minute version

If you do nothing else, ask three questions of every still:

1. What is touching what?
2. What is thinner than a fingertip?
3. What would I have to count?

Those three catch most of it, and they take longer to read than to perform.

Honesty about the checklist

This list is a snapshot. Hands and short-range physics have improved measurably between model generations and will keep improving. Text in image models was a standing joke and largely is not any more.

So re-run the list yourself every few months on your own shots. Keep the mechanisms, which do not change, and treat the verdicts as provisional. A course that told you these were permanent would be lying, and a course that told you the next model fixes everything would also be lying — a specific real product is missing information, not missing capability, and that one does not move.

Today: run the nine-point check on the next still you make, write down which item you expect to fail, then generate and see whether you were right.

BEFORE YOU MOVE ON

A still shows a musician holding a violin, with a patterned rug underfoot and a window reflecting the room. Which feature is most likely to fail in a way that no amount of regeneration fixes?

A. The patterned rug, because regular grids warp and the model cannot hold a consistent count of repeating elements.
B. The window reflection, since reflections require a three-dimensional understanding of light and surfaces that the model approximates rather than computes.
C. The fingers on the strings, because contact under motion is where a plausible frame becomes a physically impossible one.
D. The violin's overall shape, as complex curved instruments are under-represented in the aesthetic-filtered training corpus relative to common objects.

19 · 8 min

Every still you keep is a part you do not have to make again

THE ONE THING TO KEEP

A still with its prompt, seed, model version and provenance stored beside it is a reusable part, and a project without that record cannot be revised six weeks later — it can only be redone.

The phone call that makes this matter

A project delivers. Six weeks pass. The client asks for one change to shot nine — the jacket should be navy, everything else stays.

If you kept the still, the prompt, the seed and the model version, that is twenty minutes: open the still, recolour the jacket, regenerate the clip, drop it into the timeline.

If you did not, you are rebuilding shot nine from nothing, in a world where the model has been updated, and you will not match the eight shots around it. That is not a revision, it is a reshoot, and you cannot bill for it because you already delivered.

One change to shot nine, six weeks later

Day 1	The still for shot nine is generated, repaired by hand, and saved as a PNG with a text file beside it: prompt, seed, model version, every setting.
Day 2	Three takes are generated. The chosen one is kept at full quality with its own settings file, and the two rejects go into the rejects folder rather than the bin.
Day 14	The piece is delivered, and the whole project folder is copied somewhere else. One copy in one place is not a backup.
Week 7	The client asks for one change. The jacket in shot nine should be navy, and everything else stays as it is.
Twenty minutes later	The still is reopened, the jacket recoloured, the clip regenerated from the same seed on the same model version, and dropped into the timeline.
The other version	With no record, shot nine is rebuilt from nothing against a model that has been updated since, and it will not match the eight shots around it.

The sidecar costs two seconds at the moment the decision is made and cannot be reconstructed a month afterwards. That is the entire difference between a revision you can bill for and a rebuild you cannot.

Everything below exists to prevent that phone call from being expensive.

What to store next to every frame

A still is not a file. It is a file plus a sidecar. For each frame that gets used:

- The image at full resolution, in a lossless format. PNG or WebP, not a re-compressed JPEG.
- The exact prompt text, copied not retyped.
- The negative prompt, if there was one.
- The model name **and its version string**.
- The seed and every numeric setting.
- The source: generated, photographed, composited, or repaired from another frame.
- If it was repaired or composited, the layered file as well.

A plain `.txt` or `.json` file with the same name as the image is enough. Do not put this in your head, in a chat history, or in a tool's own gallery — tools lose histories, accounts get closed, and chat scrollback is not a filing system.

A folder layout that survives

Nothing clever. Something like:

```
project-name/
  shots.md                 # the shot list, one line per shot
  frames/
    s09_first_v3.png
    s09_first_v3.txt        # prompt, seed, model, settings
    s09_first_v3.psd        # layered, if repaired
    s09_last_v1.png
    s09_last_v1.txt
  clips/
    s09_take4.mp4
    s09_take4.txt           # which frame, which model, which
settings
  audio/
  edit/
```

The naming convention matters more than the structure: **shot number, role, version.** `s09_first_v3` sorts correctly, reads unambiguously, and tells you at a glance that there were two earlier attempts.

Avoid `final`, `final2`, `FINAL_use_this`. Everyone has lived through that and nobody has ever benefited from it.

Keep the rejects

This is the part people skip and later regret.

A rejected still is not waste. It is a parts bin. The generation with the ruined composition but the perfect hand is where tomorrow's hand comes from. The one with the wrong light but the right face is a reference image for the next attempt. Because they came from the same model with the same prompt family,

the pieces actually fit — same colour response, same grain, same rendering of skin.

Keep them in `frames/rejects/`. Delete the whole folder when the project is signed off, if you must, and not before.

Reuse across shots, and across projects

Within a project, the same still often serves several shots: a wide used once as a first frame and once, cropped, as the end frame of a push-in. A background plate reused with a different subject composited in. An establishing frame reversed for a matching shot from the other direction.

Across projects, a small personal library of backgrounds, textures, skies and plates accumulates and is worth real time. Name them by content rather than by the project they came from, or you will never find them again.

The provenance question, which is not administrative

There is a second reason to record the source of every frame, and it is not about convenience.

If a client or a platform later asks what in this video was generated and what was filmed, the honest answer has to come from somewhere. A frame library where every item says `generated` or `photographed` or `composited` answers that question in a minute. A folder of untitled PNGs does not.

Labelling obligations and disclosure rules are dealt with properly later in the course. The record that makes compliance possible gets built here, during production, because it cannot be reconstructed afterwards.

The free path

All of this is folders and text files. No software, no subscription, no database. Version control is optional and generally overkill for binary media, though a plain copy of the project folder at delivery, stored somewhere else, has saved more projects than any tool.

Today: take the last shot you generated and write its sidecar file from memory. Whatever you could not remember is the field that goes in your template.

BEFORE YOU MOVE ON

Why is storing the model's version string alongside a still more than administrative tidiness?

A. Version strings are required by most platforms' synthetic-media disclosure rules, which ask for the specific system used.

B. Different versions price differently, so the record is needed to reconcile a project's costs afterwards.

C. Hosted providers update models silently, so a seed and prompt that reproduced a shot on one version will not reproduce it on the next.

D. The version determines which autoencoder was used, and mixing frames encoded by different autoencoders in one timeline produces inconsistent grain and sharpness.

---

20 · 8 min

Where this is already the right tool

THE ONE THING TO KEEP

Generative video is already the right tool for b-roll, impossible shots, pre-visualisation and short-form, and still the wrong tool wherever a specific person or object has to stay itself.

Six places it is genuinely production-ready

1. **B-roll and cutaways.** Shots nobody studies frame by frame. A city at dawn, a road through fields, a corridor, an abstract texture under a voice-over. Cheaper than stock and specific to your script in a way stock never is, because stock is somebody else's idea of a generic road.

2. **Impossible or unaffordable shots.** An aerial over a place you cannot fly a drone. A street as it looked in 1970. The inside of a machine. Anything where the alternative is a location budget that does not exist.
3. **Concept boards and pre-visualisation.** The clearest honest win, and the one working professionals adopted first. A director can show a client the shot before the shoot. A moving storyboard now costs an hour rather than an animatic team. The output does not need to be broadcast quality — it needs to communicate a shot, and it does that well.
4. **Product motion with the product composited.** Environment, light and camera generated; the real object tracked in. See the previous lesson for why that split is not a compromise but the correct division of labour.
5. **Short-form vertical social.** High volume, low continuity demands, and an audience with genuine tolerance for stylisation. It is also the one format where a phone-only workflow is realistic from generation through to publishing.
6. **Music video and mood pieces.** Where dreamlike inconsistency is a feature rather than a defect, this tool is not a substitute for something else. It is the thing itself.

Where it is still the wrong tool

Say this out loud in week one rather than week four:

- A character-driven narrative with dialogue across many shots.
- Anything where a real person or a real product is the subject.
- Precise physical demonstration — a how-to with hands.
- Anything a client will scrutinise frame by frame.

For all of those, a small live shoot, 2D animation or motion graphics is faster, cheaper and better, and saying so early is a professional service rather than an admission.

What has actually changed in the work

Not "video editors are obsolete", and not "nothing has changed". What changed is specific:

- The cost of a bad first version fell to nearly zero, so far more ideas get tested before anyone commits money.
- Small teams can show a moving version of an idea in a pitch. That used to be available only to productions that were already funded, which is a genuine shift in who gets to compete.
- Stock footage lost part of its market. The generic establishing shot is now generated.
- The scarce skill moved. It was never "can you operate the generator". It is shot design, judgement about which take works, editing, sound and knowing where the tool stops. Those were always the skills. They simply became the entire job faster than anyone expected.

Whether that is good news depends mostly on which half of the job you enjoyed.

A delivery checklist

Before anything leaves your machine:

- Every clip trimmed of its settle and its drift.
- One grade and one grain pass across the whole timeline.
- Real sound under every shot.
- Frame rate conformed to one project rate; upscale once at the end, if at all.
- A visible label on generated content, at the start, in frame — not only in metadata, not only in the caption. Several countries require this and none of them accept a label a viewer will not see.
- Written consent on file for any real face or voice, specific to this use.
- The prompt, seed, first frame and model name saved beside the project for every shot that shipped.

The loop, which is the whole course

Pick one shot from something you are actually making — not a test. Write it as a shot with a duration. Make the still and fix one thing in it by hand. Generate

three takes. Trim them. Put the best one on a timeline with real sound under it. Label it.

That loop is everything in the preceding nine lessons compressed into an afternoon. Everything after this is repetition, better taste, and re-testing the limits list every few months to see which items have quietly moved.

BEFORE YOU MOVE ON

An agency is pitching a sixty-second television commercial featuring a named actor and the client's packaged product. Where does generative video most obviously belong in that job today?

A. Generate the entire spot and composite the real product in at the end, since the product is the only element the model cannot handle.

B. Generate everything except the actor and the product, and shoot those two elements against green screen for compositing.

C. In pre-visualisation — a moving storyboard that wins the pitch and shows the client the shots — with the commercial itself shot conventionally.

D. Nowhere; broadcast work has quality and rights requirements that rule generative video out of the process entirely.

CASE STUDY · MODULE 2

The label that was almost right

A two-person cotton babywear exporter in Tiruppur, making a fifteen-second product film for a trade buyer's portal, with three days before the listing deadline.

Meena ran the business with her brother. The buyer's portal wanted a short motion listing for each product line, and they had eleven lines. Her brother had spent a Saturday generating video from text and had eleven clips of baby-wear that was not theirs.

The garments looked convincing at a glance. They were the right family of colour, the right weight of fabric, the right softness of light. The neck labels were the problem. Their brand mark is a small woven label with three letters and a leaf, and what came back was a small woven label with three letters and a leaf that were not those three letters and not that leaf.

This is the failure that is worse than an obvious one. A garment with no label reads as a stock photograph. A garment with a label that is nearly theirs reads as a counterfeit, which is the single accusation an exporter cannot afford on a buyer's portal.

Meena's brother wanted to try again with a better prompt. He had a theory about describing the leaf more precisely. Meena stopped him, for a reason she could state: the model has no representation of their specific label. It was not in the training data and no amount of describing it retrieves information that is not there. More attempts would produce a different wrong label, not the right one.

So the decision was about the division of labour, and both options had a price.

Option one: generate the whole shot and accept the label. Fast, and they could deliver all eleven that evening. It also meant putting something on an international buyer's portal that misrepresented their own product, which was not a risk about taste.

Option two: change the split. Let the model make the world — the wooden table, the window light, the slow drift of the camera — and put the real garment in themselves. That meant photographing eleven garments properly, cutting them out, and compositing.

They went with the second and did the arithmetic on it first, because three days is not a long time.

The photography took a morning. A sheet of white paper taped to the wall, the garments a metre in front of it so their own shadow did not fall on the background, the window on the left because that was the light direction they had already written at the top of the shot list. Flash off. Twenty frames each, one chosen. A phone.

Cutting out took longer than expected. The one-click background removal handled nine of the eleven and failed on two, both of them knitted pieces with loose fibres at the edge, which is exactly where automatic keying fails. Meena masked those two by hand in Photopea in a browser tab, because the laptop in the office cannot run Photoshop and never could.

Then the composite. Perspective first: the generated table had to be seen from the same height the garment was photographed from, so two of the eleven backgrounds were regenerated rather than distorted, because distortion is precisely what the eye catches. Light direction second, which was already right because they had decided it once. Then the contact shadow under each garment, softened to match the other shadows in the frame, and a curves adjustment to match the black point.

The last step was the one her brother thought was fussy. She added a small amount of grain to each composited garment and softened its cut-out edge by a pixel, because the generated plates are smooth and a real photograph is not, and a laser-sharp edge against a soft plate announces itself.

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

Ten of the eleven listings went up on the second day and the eleventh on the third. The buyer's quality team flagged nothing. Six weeks later a different buyer asked for the same product on a darker surface, and because the cut-out garment files and the sidecar text files for every plate were still in the project folder, that was a new background and an afternoon rather than a reshoot. Meena's written rule afterwards was one line: anything a buyer is paying for a close-up of does not get generated, it gets photographed and put in.

Why would more prompt attempts not have produced the correct label, however precisely the leaf was described?

The model has no stored representation of that specific woven label. Prompting steers the model toward regions of what it already knows; it cannot supply information the model never had. This is the one limit in generative video that does not improve with scale or with more attempts, because it is a missing-information problem rather than a capability problem. The answer is to change the division of labour: generate the environment and the camera move, and supply the real object yourself.

They composited into the still rather than tracking the garment into the finished clip. When would that have been the wrong choice?

Compositing into the still is right when the camera move is small and the shot is short, because the object is animated along with everything else and will soften slightly as the clip runs. It would have been wrong if the buyer were looking at the label frame by frame, if the camera move were large, or if the label had to stay pin-sharp throughout. In those cases the object is tracked into the finished footage in the editor, which is more work and is the only version that survives close scrutiny.

Why did adding grain and softening the cut-out edge by a pixel matter, when neither is visible to a casual viewer?

The generated plate came back through an autoencoder that removed high-frequency detail, so it is smooth. A photograph carries sensor noise. Placing a noisy, laser-edged object onto a smooth plate creates two different surface textures in one frame, which viewers register as a composite without being able to name why. Matching the grain and softening the edge makes the object share the plate's sur-

face properties, which is the same reason a single grain pass across a finished timeline unifies thirty separate generations.

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 Beyond price, what can you do to a still that you cannot do anywhere inside a video model?

A. Make the model restyle the scene partway through the clip
B. Guarantee the same output from the same prompt on a hosted service
C. Raise the resolution beyond the model's native output size
D. Repair it deterministically: fix one hand and change nothing else

2 You compose a still you like at 3:2 and send it to a model whose output is 16:9. What actually happens?

A. The model letterboxes the still and generates inside your original framing
B. It is cropped, padded or stretched to fit, and your framing is gone
C. The model recomposes the scene at the new ratio from the prompt
D. Nothing, because aspect ratio is applied only at export

3 A still is right in every respect except a six-fingered hand. Why do professionals repair it rather than generate again?

A. The model cannot render hands at all, so every hand has to be painted in by a person
B. Regenerating costs credits and repairing does not
C. A repair is deterministic; a new generation is a fresh draw over the whole frame
D. Inpainting returns a higher resolution than the base model produces

4 You supply a kitchen as the first frame and a beach as the last. The model returns a morph rather than a camera move. Why?

A. A smooth path between two distant latent points corresponds to nothing physical
B. End-frame conditioning only works on clips under three seconds
C. The two images were generated at different resolutions
D. The model needs a third keyframe in the middle before it can path between two shots

5 An archival photograph from 1950 must appear to have a slow camera move, and the faces in it must stay exactly themselves. What do you reach for?

A. A trained character model built from the faces in the photograph, then generate
B. An upscaler first, so the model has more detail to work from
C. Image-to-video at a low motion strength, which drifts least
D. A depth map and a 2.5D projection, because it moves the original pixels

6 Which camera move requires the model to invent the least new picture, and why?

A. A pull-out, because the subject stays in frame throughout
B. A push-in, because it only crops what is already there
C. An orbit, because the parallax reveals depth the still already contains
D. A tilt up, because sky is a simple thing to invent

MODULE 3

Directing, not describing

The prompt is a light influence over a great deal of structure, so the words that steer are the ones a shot list uses — size, angle, lens, one move with a speed, a verb for the subject and a note on what stays still. This block builds that vocabulary properly, adds the controls that take motion out of language altogether, and finishes with an experimental protocol that tells you which of your instructions the model is actually following.

> BY THE END OF THIS MODULE YOU CAN
>
> Write a shot instruction naming size, angle, lens character, one camera move with a speed, the subject's verb and what stays still; choose between language, a motion brush and an explicit conditioning signal for a given move; and run a fixed-seed experiment that determines which parts of your instruction the model obeyed

21 · 9 min

Cinematic, 8k, epic gets you the average of everything

THE ONE THING TO KEEP

Name the shot size, one camera move, the subject's verb and what stays still — mood adjectives return the average of everything captioned that way.

Why mood words produce mush

Your prompt is matched against the language that sat beside video in training. Words like cinematic, epic, 8k and hyperrealistic appear underneath an enormous variety of unrelated clips. Ask for them and the model returns something close to the average of all of it: a slow drifting push-in, a teal-and-orange grade, a lens flare, and nobody doing anything in particular.

That is not the model failing. That is the model succeeding at a request that did not locate a shot.

The words that work are the ones that appeared in descriptions of specific shots, because that is what they were attached to: slow dolly in, handheld tracking shot following from behind, static locked-off wide, crane up to reveal, whip pan, rack focus from foreground to background, low angle looking up. These are the terms a shot list uses, and they were in the captions.

Two kinds of word, and what each one returns

Words a captioner would have written	Words scattered across everything
• slow dolly in — a named move, attached to millions of clips that contained it	• cinematic — leaked in from human descriptions of unrelated footage
• medium close-up, eye level — visible facts about the frame	• epic, 8k, hyperrealistic — not observations, so attached to nothing in particular
• warm window light from camera left — light is visible, so captions described it	• she reacts emotionally — an intention, and the model animates actions
• she turns her head to the left and looks down — an action, not an intention	• dynamic camera — the average of every move at once, which is a drift

Almost none of the training captions were written by a person. A vision model described what it could see, tens of millions of times, and that vocabulary is what steers. The rule follows directly: write the caption you would want a machine to produce for the clip you want.

The four things worth naming

1. **Shot size and angle.** Wide, medium, close-up. Eye level, low angle, overhead. This one line does more than any adjective.
2. **One camera move, with a speed.** Slow push in. Not "dynamic camera".
3. **Subject motion, as a verb.** "She turns her head to the left and looks down" beats "she reacts emotionally". The model animates actions, not intentions.
4. **What stays still.** Models like to move everything. If you want a locked-off frame, say the camera does not move and the background is static. A still camera is a request, not a default.

One move per clip

Two camera moves in eight seconds and the model performs neither cleanly — it blends them into a drift. Same for two subject actions. Same for a camera move plus a subject action plus a lighting change.

Pick the one thing the shot is about. This is also just film grammar: a shot has one job. The constraint of the tool and the discipline of good directing point the same way here, which does not happen often.

Preset motion libraries, and what they actually are

Higgsfield is the clearest example of a platform whose main contribution is camera direction rather than raw generation. It offers a large library of named moves and effects — crash zoom, dolly, orbit, bullet time, drone-style flights, car-chase rigs — applied as presets on top of underlying models.

What a preset is, mechanically: a pre-written motion description plus tuned generation parameters, sometimes with an explicit motion-conditioning path. The value is real. It hands you a vocabulary you would otherwise learn by burning credits, and it is repeatable enough to plan a sequence around, which loose prose is not.

The limit is equally real. A preset is a look, not a shot. It applies the same move regardless of whether your frame supports it. Orbit around a subject standing in front of a flat wall and the move immediately reveals that the wall is not a real space — because it is not, it is a painted backdrop the model has to invent as the camera swings.

The same idea appears elsewhere in different clothes: Runway's camera controls and motion brush, where you paint the region of the frame that should move; Kling's motion controls; Pika's effects. All of them exist because language is a poor way to specify a curve, so the interface takes it out of language.

Turn the motion strength down

Nearly every tool has a motion or dynamism setting. High motion means more movement and more artefacts — drift, morphing limbs, background churn — because there are more frames of change to get wrong.

The professional default is lower than feels exciting. A near-still frame with a slow drift, a small parallax and real ambient sound reads as footage. A high-motion clip with a morphing hand reads as AI, immediately, to everybody. You are not being timid; you are choosing the version that survives being watched twice.

Negative prompts

Where the tool offers one, use it for qualities rather than objects: warping, morphing, motion blur, watermark, extra limbs, distorted face. Asking a negative prompt to keep an object out of a scene works about as poorly here as it does with images, for the same reason — the word pulls the thing in and the negation is a weak counterweight.

Today, a ten-minute experiment worth more than a week of reading. Take one still. Generate it three times: locked-off static, slow push in, handheld follow. Same still, same seed if the tool gives you one. Watch what each move costs you in artefacts, and you will know what your motion budget actually is.

BEFORE YOU MOVE ON

Someone prompts a street scene with "cinematic epic drone shot, 8k, hyperrealistic, dynamic camera, dramatic lighting" and gets a slow drifting push-in with a lens flare in which nothing much happens. What best explains the result?

A. The prompt exceeded the model's text encoder length, so the later terms were truncated before generation.

B. Drone-style moves require a high motion-strength setting, and the tool's default suppressed the movement that was asked for.

C. No reference image was supplied, so the model fell back on its own house style instead of the prompt.

D. Those words sit under an enormous variety of unrelated clips, so the model returns roughly the average of everything captioned that way; none of them specify a camera behaviour.

22 · 9 min

The two words that decide most of the frame

THE ONE THING TO KEEP

Shot size and camera angle were named explicitly in the captions the model trained on, so they steer more reliably than any other words, and they carry meaning an audience reads without noticing.

The vocabulary, and why it works here

Auto-captioners describe shot size and angle because both are visible facts about a frame. The words appeared millions of times, attached precisely to what they name. That makes them the most dependable instructions you have.

The standard sizes, from widest:

- **Extreme wide.** The subject is small in a large space. Establishes where.
- **Wide** or **long shot.** Full body, head to feet, with room around.
- **Medium wide.** Knees up. The workhorse of dialogue scenes.
- **Medium.** Waist up.
- **Medium close-up.** Chest up. The interview shot.
- **Close-up.** Shoulders and head.
- **Extreme close-up.** Part of a face, or an object filling frame.

And the angles:

- **Eye level.** Neutral. The default and the one you should use unless you want something.
- **Low angle**, looking up. Gives the subject stature.
- **High angle**, looking down. Diminishes.
- **Overhead** or top-down. Abstracts a scene into a pattern.
- **Dutch** or canted, the horizon tilted. Unease. Use once per project at most.

What each size costs you in a generated clip

This is where film grammar and the machinery interact, and it is worth knowing before you choose.

Wide shots are cheap and safe. The face is forty pixels across, so identity drift is invisible. Hands are too small to go wrong in a way anyone can see. Wides are the single most reliable shot in generative video and you should use more of them than you think.

Close-ups are expensive and risky. Every pixel of the face is being scrutinised by the part of your viewer's brain that is best at faces. Skin texture, teeth and eyes are all near the autoencoder's resolution limit at that scale. Close-ups are where you spend attempts.

Extreme close-ups of objects are surprisingly good. No face, no hands, large simple forms filling frame. A drop of water, a fabric weave, a hand-free detail. These are among the most reliable shots available and they are undervalued because they feel like filler. They are not filler; they are what a real editor cuts to.

Medium shots are the compromise that works. Face large enough to read, small enough to survive. If a shot has to contain a person and has to hold, make it a medium.

Angle and the space behind the subject

A low angle looks up and therefore shows ceiling or sky. A high angle shows floor. An overhead shows the ground plane and nothing else.

Each of those is a region the still may not contain, and if the camera moves the model has to invent it. So angles interact with the composition rule from the previous module: pick the angle in the still, and if you then move the camera, know which new territory the move opens up.

Overhead shots are worth a specific note. They are strong compositions, they hide faces, they hide hands, and the ground plane is usually simple. For a difficult subject, an overhead is often the shot that works when nothing else does.

What the size means to a viewer

Shot size is not decoration. It is how a film controls distance and attention.

A sequence that stays wide keeps the audience outside, observing. A sequence that moves inward as it progresses pulls them in. Cutting from a wide directly to an extreme close-up is a jolt and can be exactly right. Cutting between two shots of nearly the same size on the same subject is a jump cut and reads as a mistake.

That last point has a direct consequence for how you plan generated shots. **Change the size by at least two steps between consecutive shots of the same subject**, or change the angle substantially. A wide to a medium wide is a stumble. A wide to a close-up is an edit.

This also happens to be the rule that hides identity drift, because a big size change gives the eye a new thing to read and less opportunity to compare faces. Grammar and machinery agree again.

Writing it

Put it first in the instruction, because it is the strongest term:

> *Medium close-up, eye level, a woman at a kitchen table turns her head to look out of the window on the left. The camera does not move. Warm window light from camera left.*

Size, angle, subject verb, camera behaviour, light. Five clauses, no adjectives about mood, and every word of it is something an auto-captioner would have written.

The free path

This costs nothing. It is vocabulary and a decision. The most useful free exercise: take one still, generate it four times at four different described sizes, and confirm that your model responds to them. Some models reframe from the supplied still more willingly than others, and knowing which you have changes whether you specify size in the prompt or bake it into the frame.

Today: look at your shot list and mark the size of every shot. If two consecutive shots of the same subject are within one step of each other, change one.

BEFORE YOU MOVE ON

A sequence keeps losing continuity on a character's face between shots. The shot list has: wide, medium wide, medium, medium close-up, close-up — a smooth progression inward. What change would help most?

A. Reverse the order so the sequence moves outward, since identity drift is less noticeable when the face is getting smaller rather than larger.
B. Make the size jumps larger, at least two steps between shots, so the eye gets new information and less chance to compare faces.
C. Hold every shot at the same size and let the angle carry the variation, so the face is rendered at a consistent scale throughout.
D. Generate all five shots from a single still at successively tighter crops, so the face is literally identical in each one.

23 · 9 min

Focal length, depth of field, and the words that carry them

THE ONE THING TO KEEP

Focal length changes the relationship between foreground and background rather than just the field of view, so naming a lens steers the geometry of the frame in a way that no amount of describing the subject can.

What a lens actually changes

Most people think a wide lens shows more and a long lens shows less. That is true and it is the least interesting part.

The real effect is on **relative size between near and far**. A wide lens used close to a subject makes the subject large and the background tiny and distant — space feels stretched. A long lens used far from the same subject makes the background loom up behind them, compressed and close.

This is why a 24mm portrait looks like a confrontation and a 135mm portrait looks intimate, even framed identically. It is not about sharpness or quality. It is about the geometry of the space.

The words that steer

Captions named lenses constantly, in two forms. Both work:

- **Numerically**: "shot on a 35mm lens", "85mm portrait lens", "wide-angle 18mm", "telephoto 200mm".
- **Descriptively**: "wide-angle with visible distortion at the edges", "compressed telephoto background", "shallow depth of field, background falls out of focus".

Use both together for the strongest effect. The number pins the family; the description pins what you actually want from it.

A rough map worth carrying:

- **14-24mm.** Very wide. Edge distortion, dramatic space, everything sharp. Interiors, landscapes, an aggressive close-up.
- **35mm.** The documentary lens. Close to how a room feels to be in.
- **50mm.** Neutral. Roughly human perspective.
- **85mm.** The portrait lens. Flattering, background softly separated.
- **135mm and up.** Compressed. Background looms. Surveillance, isolation, a face picked out of a crowd.

Depth of field and why it helps more than it should

Shallow depth of field — a sharp subject against a soft background — does three useful things at once in generated video.

1. **It directs attention**, which is its normal job.

2. **It hides background churn.** The most common visible failure in generated clips is the background quietly reorganising itself. A background that is a soft wash of colour cannot visibly reorganise. This is an enormous practical benefit and it is free.
3. **It hides background faces.** A crowd rendered as coloured shapes has no identities to drift.

So the honest advice is to use it more than a photographer would, for reasons a photographer would not recognise. "Shallow depth of field, background out of focus" is one of the highest-value phrases in generative video.

The cost is that you lose the location. If the shot has to establish where you are, it needs depth, and you will have to accept the churn or shorten the shot.

Rack focus

Pulling focus from foreground to background mid-shot is a real, nameable move — "rack focus from the foreground bottle to the woman behind it" — and models do attempt it.

What you usually get is a blur that ramps globally rather than a true focal plane travelling through the scene. It reads as a defocus and then a refocus, which is close enough for many shots and obviously wrong for others.

The better route is to do it yourself. Generate the shot with everything acceptably sharp, then apply a keyframed blur in the editor with a mask. Resolve's free tier does this cleanly, and it is deterministic, which the generated version is not.

Lens artefacts, and the honesty of asking for them

Flares, chromatic aberration, vignetting, barrel distortion, anamorphic streaks. Captions named these, so they work as instructions.

They are also a well-known shortcut to making generated material read as photographed, because real lenses do these things and clean renders do not. Used lightly, that is legitimate craft — matching the imperfections of the medium you are pretending to be, exactly as a colourist adds grain.

Used heavily, it becomes the visual equivalent of shouting, and it is one of the reliable markers of amateur generative work. An anamorphic flare in every shot is not cinematic; it is a filter.

If you want the artefact, a better route is often to add it in the edit, where you control the amount and can remove it at the client's request.

Mixing lenses across a sequence

A real production picks a small lens set and stays inside it. Doing the same in generated work pays twice: it looks deliberate, and it reduces one more axis on which two shots can fail to match.

Pick two focal lengths for a project — say 35mm for wides and 85mm for everything closer — and put them in the shot bible. Then the lens is no longer a decision you make thirty times.

The free path

All of this is words, plus a blur node in a free editor. Nothing here requires a subscription. The one thing worth spending a free daily allowance on is a lens test: one still, generated with 24mm, 50mm and 135mm in the instruction, to see whether your model distinguishes them. Some do it well; some ignore focal length entirely and respond only to the depth-of-field description. Knowing which you have saves you writing words that do nothing.

Today: add "shallow depth of field, background softly out of focus" to one shot that has been churning in the background, and compare it against the original.

BEFORE YOU MOVE ON

Adding "shallow depth of field, background out of focus" to a shot reliably improves how finished it looks, beyond the usual aesthetic reasons. What is the mechanical benefit?

A. Defocused regions compress to fewer latent tokens, so more of the model's capacity is available for the sharp subject.
B. A background rendered as a soft wash has no visible structure to reorganise, so the most common generated-video failure becomes invisible.
C. Blur suppresses high-frequency detail that the autoencoder would otherwise discard, so less information is lost at the encode stage.
D. Shallow depth of field appears more often in the aesthetically filtered training corpus, so prompts requesting it land in a denser, better-learned region of the model's distribution.

24 · 9 min

Where people stand, and which way they face

THE ONE THING TO KEEP

Screen direction is a continuity contract across a cut, and because a generated shot has no set and no camera position you have to enforce it in the first frame rather than discovering it on the day.

Blocking is a still-frame decision

Blocking means where people and objects are placed, and how they move through the space. On a set it is worked out with actors. In generative video there are no actors, so blocking is decided entirely in the first frame — and that means you can actually control it, which is unusual in this medium.

The things to decide before generating:

- Where the subject is in frame: left third, centre, right third.
- Which way they face: toward camera, away, left, right.
- What is in front of them and behind them.
- Where they will be at the end of the shot, if they move.

A model asked to animate will otherwise default to what stock footage does: subject centred, facing camera, standing still.

The 180-degree line

The one rule from editing that generative work breaks constantly.

Imagine a line drawn through two people in conversation. Keep the camera on one side of that line and A always appears on the left looking right, B always on the right looking left. Cut across the line and both appear to be looking the same way, and the audience abruptly cannot tell who is talking to whom. The technical name is crossing the line; the experience is a small moment of confusion the viewer cannot explain.

On a set, a script supervisor prevents this. In generative work, nothing prevents it, because each shot is generated independently with no knowledge of the others.

So you enforce it in the stills. Write the direction into every shot on the list:

```
s04  MCU  Priya, screen left, looking RIGHT
s05  MCU  Arjun, screen right, looking LEFT
s06  MCU  Priya, screen left, looking RIGHT
```

Then build each still to match, and check the stills side by side before generating any of them. Two minutes of checking saves a scene that cannot be cut.

Movement direction across a cut

The same rule applies to anything moving. If a car exits frame right in shot one, it should enter frame left in shot two — because it is continuing in the

same direction through an imagined space. Enter frame right and it appears to have turned around.

This matters in generative video more than usual, because you are frequently cutting between shots that were never in the same place at all. Consistent screen direction is a large part of what convinces an audience that a set of unrelated generated clips is one continuous journey. It is, genuinely, the cheapest continuity tool you have: it costs nothing but a note on a shot list.

Staging in depth

Placing people at different distances from camera — one near, one far — rather than side by side does three things:

- It creates the depth planes that make camera moves work, from the composition lesson.
- It lets one shot carry two subjects without either being a tiny background figure.
- It gives the model a foreground element to anchor the frame, which reduces drift.

A figure in soft foreground at the edge of frame, even just a shoulder, stabilises a shot noticeably. This is also how over-the-shoulder shots work, and why they are so useful: the shoulder is a large simple shape that holds, and it hides half the identity problem.

Entrances and exits

Asking a subject to walk into or out of frame is asking the model to invent them mid-clip or dispose of them convincingly. Entrances are harder than exits — something has to appear from a region that does not exist.

Two workable approaches. Put them partly in frame at the start, so they are entering rather than appearing. Or generate the exit only and let the edit handle the entrance by cutting to a shot where they are already there. The cut is free; the invention is not.

Eyelines as a directing instrument

Where a subject looks tells the audience what the next shot contains. A character looking off-frame left, then a cut to a window, and the audience has understood a spatial relationship that was never depicted.

This is enormously powerful in generative work because it lets two completely unrelated clips become a cause and an effect. Generate a person looking down-left. Generate a dog. Cut them together. The audience constructs a room that neither clip contains.

Specify eyeline explicitly in the instruction — "looking down and to the left, out of frame" — because a model left to itself will point the eyes at the lens, which is the one direction that breaks this effect.

Today: take any two-person scene on your list and write screen direction on every line. Then lay the stills side by side and check that nobody has swapped sides.

BEFORE YOU MOVE ON

In a generated conversation, shot A shows Priya on the left looking right and shot B shows Arjun on the left looking right. Cut together, what does the audience experience?

A. A jump cut, because both frames place their subject in the same part of the screen at a similar size.
B. Nothing unusual, since viewers infer the geometry from the dialogue rather than from where the subjects appear on screen.
C. The impression that both are looking at a third person off-screen rather than at each other, because the line has been crossed.
D. A continuity break in the lighting, since reversing the camera position implies the key light now falls from the opposite side of both faces.

25 · 9 min

Taking motion out of language

THE ONE THING TO KEEP

Language is a poor way to specify a curve, so the controls that matter most are the ones that bypass it — a motion strength slider, a painted region, a drawn trajectory — and each of them fails in a way the prompt equivalent does not.

Why these controls exist

Every interface in this category has, at some point, added a control that is not a text box. That is not product decoration. It is an admission that a sentence cannot specify how fast, how far, and in what curve.

The three you will meet:

1. **Motion strength**, a single slider.
2. **Motion brush**, where you paint the part of the frame that should move.
3. **Trajectory or drag control**, where you draw a path an object should follow.

Motion strength, and why the good default is low

A single number controlling how much the clip moves overall. Names vary: motion, dynamism, motion bucket, amount.

The relationship with quality is direct and mechanical. More motion means more change per frame, and every frame of change is another opportunity for drift, morphing limbs, background churn and identity slip. High motion does not just look busier; it is genuinely more likely to break.

The professional default is lower than feels exciting. A near-still frame with a slow drift, a small parallax and real ambient sound reads as footage. A high-

motion clip with a morphing hand reads as generated, immediately, to everybody.

That is not timidity. It is choosing the version that survives being watched twice, which is the only test that matters for anything a client pays for.

A useful calibration: find the motion value at which your model starts producing visible artefacts on your kind of content, then work one or two steps below it. That number is yours and it is worth twenty minutes of a free allowance to find.

Motion brush

Paint a mask over the region that should move, leave the rest alone. Runway popularised it and similar tools exist elsewhere.

What it is good for is specific and valuable: **motion that is local**. Steam rising from a cup while the room stays still. Hair moving while the face does not. Water in the bottom third. A flag. A single figure in a crowd.

What it does not do is make the unpainted region truly static. It biases rather than freezes, so expect some residual drift outside the mask, and check for it rather than assuming.

The technique that gets the most from it: paint less than you think. A small moving region against a still frame is more convincing than a large one, because the stillness is what makes the motion read as real. This is the same principle as a subtle grade — the restraint is where the credibility lives.

Trajectory and drag controls

Draw an arrow, or drag a point from where a thing is to where it should go. The tool converts that into a motion-conditioning signal.

This is the most direct control available and it is also the most brittle. Three honest limits:

- **It controls position, not pose.** Dragging a person across frame moves them; it does not make them walk convincingly while doing so. You often

get a sliding figure.

- **It fights physics.** A drawn path that no real object would take produces exactly the impossible-looking result you asked for.
- **It competes with the prompt.** If the text says one thing and the arrow says another, you get an averaged mess. Make them agree, or leave the motion out of the text entirely when you are drawing it.

Use it for objects rather than people, and for simple translations rather than complex actions.

Camera controls as a separate axis

Some tools separate camera motion from subject motion — sliders or gizmos for pan, tilt, zoom, roll and orbit. Where this exists, use it in preference to describing the move in words, for the obvious reason: a number is unambiguous and a sentence is not.

It also lets you do the thing language cannot, which is ask for a camera move of a specific magnitude. "Slow push in" is a genre. "Dolly in 15%" is an instruction.

Choosing between the three

A decision rule:

- **Whole-frame ambient motion** — slider, set low.
- **One region moves, the rest holds** — brush.
- **A specific object travels a specific path** — trajectory, and expect to fix the result in the edit.
- **The camera moves** — camera controls if they exist, otherwise language.

The free path

Motion strength exists everywhere, including free tiers. Brushes and trajectories are mostly features of paid hosted tools, but the open-weights equivalent is real: ComfyUI workflows expose motion masks and trajectory conditioning as nodes, free, if you can run them. On a phone with only a free hosted al-

lowance, the slider is your whole toolkit — which is fine, because turning it down is the highest-value thing on this page.

Today: generate the same still twice, once at your tool's default motion and once at roughly half of it. Judge which one you would show a client.

BEFORE YOU MOVE ON

An operator paints a motion brush over a cup of steaming tea and leaves the rest of the café frame unpainted. The tea moves well, but the background still drifts slightly. What does this tell you about the control?

A. The brush leaked at its edges, and a tighter mask with a harder edge would have held the background completely still.

B. The mask biases where motion is concentrated rather than freezing everything outside it, so some residual movement remains.

C. Background drift comes from the autoencoder rather than from motion conditioning, so no brush setting could affect it.

D. The unpainted region received a motion strength of zero, so the drift must be interpolation artefacts introduced after generation rather than movement in the clip itself.

26 · 10 min

Depth, pose and flow as instructions

THE ONE THING TO KEEP

A conditioning signal supplies the model with a per-frame structural target — depth, pose skeleton or optical flow — which is why it produces motion a prompt cannot describe, and why it also imports every flaw of the source video.

Beyond language and beyond sliders

The strongest motion control available is not a word or a number. It is another video.

Extract a structural signal from a reference clip — depth per frame, a pose skeleton per frame, or the optical flow between frames — and feed that alongside your prompt and first frame. The model then generates content that follows that structure. This is the video descendant of ControlNet, and it is how most of the striking transformation work you have seen was made.

Three signals, three different jobs.

Depth. A per-frame depth map from the reference. Gives you the reference's camera move and spatial layout, while leaving surfaces free. Good for putting your scene through somebody else's camera move.

Pose. A skeleton — joints and limbs — tracked through the reference. Gives you the reference's performance. This is how a generated character does a specific dance, a specific gesture, a specific walk.

Optical flow or edges. Per-pixel motion, or per-frame outlines. The tightest control and the least freedom; output follows the reference almost exactly, which is what you want for a style transformation and not what you want for anything else.

Why this beats prompting for certain shots

A prompt cannot say "she raises her right hand to shoulder height over 1.4 seconds, pauses, then lowers it". A pose sequence says exactly that, frame by frame, and the model obeys.

So any shot where the *timing* of an action matters is a candidate. Earlier in the course we said action timed to a beat is not currently something you can ask for. With pose conditioning, it is — because you are no longer asking, you are supplying.

The reference can be you, filmed on a phone in a kitchen. That is the thing worth sitting with: the hard part of this technique is free and requires no talent for prompting at all. It requires being willing to stand in front of a phone and do the movement.

What it imports along with the motion

The honest limitation, and it is not small.

A conditioning signal carries everything about the reference's motion, including the parts you did not want:

- **Proportions.** A pose skeleton from a tall person drives a tall character. Applied to a character with different proportions, limbs stretch or the figure warps to reach the joint positions.
- **Camera shake.** Handheld wobble in the reference becomes wobble in the output. Stabilise the reference first, in Resolve or with FFmpeg's `vidstab`, before extracting anything.
- **Framing.** Depth conditioning brings the reference's composition. If the reference has the subject at the left edge, so will your output.
- **Errors.** A pose tracker that loses an arm behind a torso for four frames produces four frames where the generated arm does something impossible. Inspect the extracted signal as a video before you generate from it.

That last point is the one people skip. **Render the conditioning signal and watch it.** It takes thirty seconds and it is where most of these failures are visible.

Where the presets fit

The named camera-move libraries you find on consumer platforms — crash zoom, orbit, bullet time, drone flight — sit on top of this machinery or on top of tuned prompt-and-parameter sets. That is why they are repeatable in a way loose prose is not.

Their value is real: a vocabulary handed to you rather than learned by burning credits. Their limit is equally real: a preset is a look, not a shot. It applies the same move regardless of whether your frame supports it, and an orbit around a subject standing against a flat wall reveals that the wall is not a space — because it is not one.

The free path, which is genuinely complete here

This is one of the areas where open tooling is not a compromise.

- **Depth**: Depth Anything, free, as a Hugging Face Space or a ComfyUI node.
- **Pose**: DWPose and OpenPose, free, same routes. MediaPipe runs pose tracking on a phone in a browser.
- **Flow**: RAFT and similar, free, in ComfyUI.
- **Stabilisation and extraction**: FFmpeg, free, on any machine.
- **The controllable video models themselves** — the open-weights lines with control adapters — run in ComfyUI on the hardware discussed earlier.

The binding constraint is a GPU rather than a budget. Colab and Kaggle sessions close some of that gap; Hugging Face Spaces demonstrating specific control pipelines close more.

When not to use it

If the motion is simple and generic — a slow push-in, a figure standing in a breeze — conditioning is a great deal of machinery for something a sentence already does. Reach for it when the motion is specific, when timing matters, or when you need the same movement repeated across several shots.

Today: film ten seconds of yourself performing one gesture on a phone. Extract a pose sequence in a free Space and watch the skeleton. That file is a piece of direction no prompt could have written.

BEFORE YOU MOVE ON

A character generated from a pose sequence has arms that stretch unnaturally during a reaching movement. The pose extraction looked clean. What is the most likely cause?

A. Optical flow was used instead of a skeleton, so per-pixel motion was applied to a figure with different geometry.

B. The reference performer's limb proportions differ from the generated character's, so the model deforms the figure to reach the joint positions.

C. The pose tracker sampled at a lower frame rate than the generation, so intermediate joint positions were interpolated linearly through impossible configurations.

D. Pose conditioning overrides the first frame, so the character's build was resampled mid-clip to match the skeleton rather than the supplied still.

27 · 8 min

What a negative prompt can and cannot remove

THE ONE THING TO KEEP

A negative prompt is a second conditioning the model is pushed away from, which works well for qualities of rendering and badly for objects, because naming an object pulls it into the guided direction before the negation subtracts it.

The mechanism, in one paragraph

Classifier-free guidance compares a conditioned prediction against an unconditional one. A negative prompt replaces "unconditional" with "conditioned on this other text", and the model is pushed away from it:

```
prediction = neg + scale * (pos - neg)
```

So a negative prompt does not delete anything. It moves the target away from a region. Whether that helps depends entirely on whether the thing you named is a direction in that space or an object in the picture.

What works: qualities of rendering

Terms describing *how* the image looks rather than *what is in it*:

- `blurry, low quality, compression artefacts, jpeg artifacts`
- `warping, morphing, distorted, deformed`
- `extra limbs, extra fingers, fused fingers`
- `watermark, text overlay, logo, subtitles, timestamp`
- `oversaturated, washed out, flat lighting`
- `flickering, strobing, stuttering`

These work because they name a broad region of bad output that the model has seen a great deal of. Pushing away from that region improves things across the frame.

The video-specific ones — warping, morphing, flickering — are worth having in a default negative, because they describe exactly the failures video models produce.

What fails: objects

Put `no dog` or `dog` in a negative prompt and you will often get a dog.

The reason is worth understanding. Text conditioning does not carry negation well; the token for the concept activates the concept. In the positive prompt, saying "a street with no dogs" reliably produces dogs, for the same reason telling somebody not to think of an elephant does not work. In the negative prompt the subtraction is real but weak, because one noun against a whole conditioned prediction is a small vector.

What happens in practice is that the concept gets activated and only partly subtracted, so you get a dog-shaped smear, or a dog in the background, or a dog-adjacent animal.

The reliable way to remove an object from a frame is to remove it from the frame. Repair the still. Clone it out. This is the point of the repair lesson and it applies with full force here.

Negatives can suppress motion you want

A trap specific to video. A negative list containing `motion blur` and `movement` and `camera shake` will produce a stiffer clip than you intended, because you have pushed away from the region of the space that contains movement generally.

If your clips come back oddly frozen, read your negative prompt before you touch anything else. This is a common and quietly expensive mistake.

Keep the negative list short and about rendering failures. Four to eight terms. Long copied-and-pasted negative lists from image-model forums are a liability

here.

Distilled models often ignore it entirely

A model trained to work at guidance 1 is not doing the extrapolation the formula above describes, so the negative branch may have little or no effect. Some interfaces still show the box.

Test it once: generate with an absurd negative prompt — `blue, sky, person` — and see whether anything changes. If nothing does, stop typing in that box and save yourself the time.

Prompt weighting, briefly

Some tools accept emphasis syntax such as `(wet pavement:1.3)` to push a term harder. Where it exists it is worth knowing, and it obeys the same rule as everything else: it steers emphasis, it does not supply information the model lacks. Weighting `(logo:1.5)` does not make the model know your logo.

Overuse produces the same burnt, frozen output as excessive guidance, because it is the same operation applied to one token.

The honest summary

A negative prompt is a small, free improvement for rendering quality and a poor tool for content control. The hierarchy of what actually controls a generated clip, strongest first:

What actually controls a generated clip

The first frame	A dense, complete specification of the picture. When the image and the text disagree, the image wins, every time.
A conditioning signal	Depth, pose or optical flow from a reference clip: a per-frame structural target, which a sentence cannot be.
Camera and motion controls	A strength slider, a painted region, a drawn path. A number is unambiguous and a sentence is not.
The positive prompt	A few hundred text tokens steering a quarter of a million video tokens. Real, and light.
The negative prompt	A small free improvement to rendering quality, and a poor tool for deciding what is in the frame.

Most people spend their attention in the reverse order, rewriting a prompt to fix something the first frame decided. Spending it in this order is most of what separates a production workflow from an afternoon of tinkering.

1. The first frame.
2. The conditioning signal, if you supplied one.
3. Camera and motion controls.
4. The positive prompt.
5. The negative prompt.

People spend their time in the reverse order. Spending it in this order is most of what separates a production workflow from an afternoon of prompt tinkering.

Today: shorten your negative prompt to six terms about rendering quality, remove anything that names an object or a kind of movement, and see whether your clips get looser and better.

BEFORE YOU MOVE ON

An operator adds "dog" to a negative prompt to keep a dog out of a street scene, and dogs keep appearing in the background. What is the reliable fix?

A. Increase the guidance scale so the subtraction of the negative branch is applied more strongly to the conditioned prediction.
B. Add more related terms — dog, puppy, canine, animal — so the negative vector covers the concept more completely.
C. Remove the dog from the first frame instead, because the still governs content far more strongly than any text branch.
D. Move the term into the positive prompt as "a street with no dogs", since positive conditioning has a stronger effect on the output than negative conditioning does.

28 · 8 min

An experiment protocol that fits on a card

THE ONE THING TO KEEP

Because generation is stochastic, comparing two clips that differ in more than one respect tells you nothing, so every useful conclusion about a model comes from a fixed seed, one changed variable and a written record.

Why most prompt advice is wrong

Somebody changes four words, gets a better clip, and posts the new prompt as a discovery. The clip was better because the random draw was better. The four words did nothing.

This is not carelessness by unusual people. It is the default outcome of how these tools present themselves: one button, one result, no controls. You have to impose the discipline yourself, and almost nobody does, which is why the public prompt literature for video is largely folklore.

The protocol

Five rules. They fit on a card and they are worth more than any prompt guide.

1. **Fix the seed.** If your tool exposes one, set it. If it does not, generate four takes of each condition and compare the sets rather than the individuals.
2. **Fix the still.** Same first frame for every condition. A different still is a different experiment.
3. **Change exactly one thing.** One word, one slider, one setting. Not a rewritten sentence.
4. **Write it down before you look.** What you changed, and what you expect. Recording the expectation is what stops you from reinterpreting the result afterwards.
5. **Compare against the previous file, not your memory.** Put them side by side. Memory of a clip you watched four minutes ago is unreliable in exactly the direction of whatever you hoped for.

A worked example

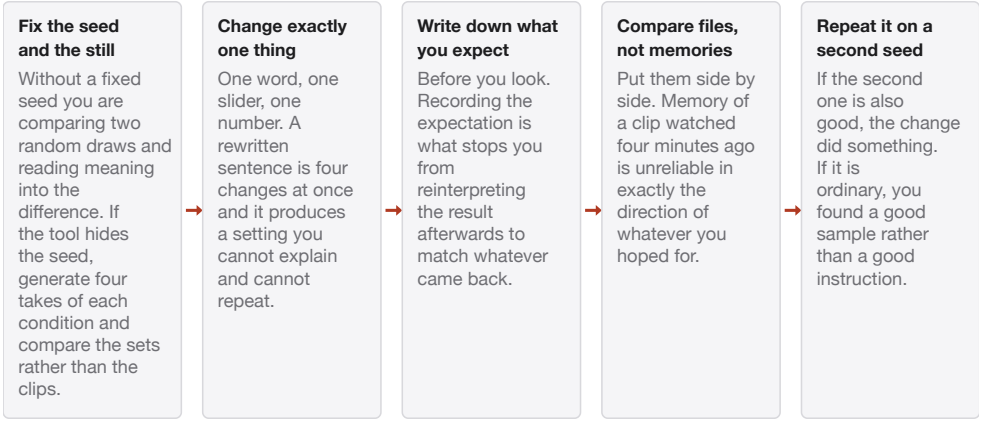
Suppose you want to know whether naming a focal length does anything on your model.

- Still: one frame, fixed. Seed: 42, fixed. Everything else: fixed.
- Condition A: `medium shot, a man at a workbench, static camera`
- Condition B: the same plus `shot on an 85mm lens`
- Condition C: the same plus `shot on a 24mm lens`

Three generations. Then look specifically at the relationship between the subject and the background — does B compress it and C stretch it? If both look identical to A, your model does not respond to focal length and you have just deleted a category of words from your prompts permanently.

That is a real finding, it took fifteen minutes, and it is about your model rather than somebody else's.

The protocol that turns an afternoon into a finding



Generation is stochastic, so two clips that differ in more than one respect tell you nothing at all. This costs one extra generation per conclusion, and it is the difference between building knowledge and collecting superstitions.

What to test first

Your free allowance is finite, so spend it in order of how much the answer changes your work:

1. **Motion strength.** Find the value where artefacts start. This affects every clip you will ever make.
2. **Does the model reframe from the still?** Ask for a close-up from a wide still. If it obeys, size belongs in the prompt. If it does not, size belongs in the frame.
3. **Camera move vocabulary.** Six phrasings, keep the ones that produce a distinct correct move.
4. **Guidance.** Three values. Watch motion die at the top.
5. **Does the negative prompt do anything at all?**
6. **Step count.** Find where improvement stops.

Six experiments, somewhere between twenty and forty generations, and you will know your model better than most people using it professionally.

Keeping the record

A plain text file per model:

```
MODEL: <name and version>, tested 2026-09
- motion: artefacts begin at 7/10; production default 4
- reframing: ignores requested shot size, obeys the still
- moves that work: slow dolly in / handheld follow / static
locked-off
- moves that fail: crane up (becomes a tilt), whip pan (smears)
- guidance: 5 good, 8 freezes motion
- negative prompt: no measurable effect (distilled)
- steps: no visible gain past 22
```

Seven lines. It will save you money every week and it becomes worthless the moment the provider updates the model, which is why it carries a date and why you re-run it after an announced update.

The trap of the impressive result

One last warning. When an experiment produces a beautiful clip, the temptation is to stop and keep the prompt. Resist it for one more generation: run the same condition with a different seed. If the second one is also good, the change did something. If it is ordinary, you found a good sample rather than a good instruction.

This single extra generation is the difference between building knowledge and collecting superstitions.

Today: run experiment one from the list — motion strength — and write the resulting number at the top of your notes file. Use it as your default from now on.

BEFORE YOU MOVE ON

A tester changes a prompt, gets a much better clip, and adopts the new wording. What single extra generation would tell them whether the wording did anything?

A. The same new prompt at a higher step count, to confirm the improvement is not an artefact of under-convergence.

B. The same new prompt with a different seed, to see whether the quality holds across draws or whether they found a good sample.

C. The original prompt with the new seed, to establish whether the seed alone accounts for the difference.

D. The same new prompt with the motion strength lowered, since reduced motion would isolate the contribution of the wording from the contribution of the movement in the clip.

29 · 9 min

The clauses that quietly do nothing

THE ONE THING TO KEEP

Prompt adherence degrades with the number of independent requirements, and the ones dropped first are counts, spatial relations, negations and anything about timing — because a few hundred text tokens are steering a quarter of a million video tokens.

Adherence is a budget, not a guarantee

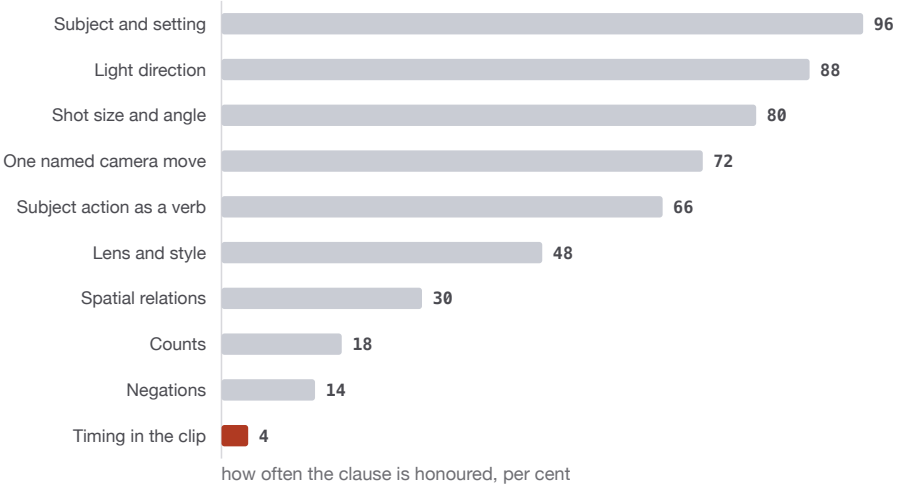
A prompt with one requirement is usually obeyed. A prompt with six requirements typically gets three or four. This is not the model being lazy; it is what happens when a small conditioning signal has to resolve several constraints simultaneously against a large prior.

Which ones survive is fairly predictable, and knowing the order lets you put the important requirement where it will not be dropped.

The reliability order

From most to least reliable, on a typical current model:

Which clauses survive, and which are quietly dropped



The shape matters, not the numbers; measure your own model by listing your requirements as separate lines and ticking them off against the clip. Everything in the bottom half belongs in the still or in the edit, where it is a fact rather than a request.

1. **Subject identity and setting.** What the thing is and roughly where. Almost always obeyed, and largely supplied by your still anyway.
2. **Light direction and quality.** Strong, for the reasons in the lighting lesson.
3. **Shot size and angle.** Strong, though some models reframe less readily from a supplied still.
4. **Camera move.** Reasonably strong if it is one move, named plainly.
5. **Subject action as a simple verb.** Turning, walking, sitting, reaching.
6. **Style and lens character.** Variable by model. Test it.
7. **Spatial relations between objects.** Weak. "The cup to the left of the book" is close to a coin toss.
8. **Counts.** Very weak. Three of anything is a suggestion.

9. **Negations.** Very weak, for the reasons in the previous lesson.
10. **Timing within the clip.** Not available. "Then she looks up" has no mechanism behind it.

The fix is not more words

The instinct on discovering a dropped requirement is to describe it harder — more adjectives, more emphasis, a longer sentence.

This usually makes things worse, because adding tokens dilutes every other token's influence. A 120-word prompt is not six times more specific than a 20-word prompt; it is a longer list from which the model picks a similar number of things to honour.

The things that actually work, in order:

- **Move the requirement into the still.** Counts, spatial relations and object presence all belong in the frame, where they are facts rather than requests. This resolves items 7, 8 and 9 on the list completely.
- **Split the shot.** If two actions must both happen, that is two shots and a cut. This resolves item 10.
- **Supply a conditioning signal.** Pose or trajectory for specific movement.
- **Shorten the prompt.** Cut to the four clauses that matter and let the model handle the rest.

Diagnosing what was dropped

After a disappointing generation, do not rewrite. Do this instead:

1. List your requirements as separate lines.
2. Watch the clip and tick each one.
3. Look at the pattern.

If the dropped ones are counts and spatial relations, that is expected and the fix is the still. If the dropped one is the camera move, try naming it alone in a short prompt to check whether the model knows it at all. If everything was

dropped, suspect guidance set too low or a conditioning image that contradicts the text.

Three minutes of this beats twenty minutes of rewriting, and it produces knowledge rather than a new guess.

The still always wins

Worth stating once more as a standalone rule, because it explains most disappointments.

If the prompt says "a red door" and the still shows a blue door, you get a blue door. The conditioning image is a dense, complete specification of the picture; the text is a few hundred tokens of suggestion. When they disagree, the image wins and the text produces a slight, unhelpful pull toward the thing it named.

So do not use the prompt to correct the still. Fix the still.

What this means for how you write

A good video prompt is short, ordered and free of wishes. Something like:

> *Wide shot, eye level. A man walks slowly away from camera down a wet alley. Slow dolly in. Hard key from camera right, deep shadows. Everything else static.*

Five clauses, all of them in the reliable half of the list. Nothing counted, nothing negated, nothing timed, no spatial relation between objects, no adjective about mood.

Everything that could not be requested reliably has been moved into the still or into the edit, which is where it belonged.

Today: take your longest prompt, list its requirements as separate lines, and move every count, negation and spatial relation out of the text and into the frame.

BEFORE YOU MOVE ON

A prompt asks for "three lanterns hanging to the left of the doorway, no people in shot". The result has five lanterns, mostly on the right, and a figure in the background. What is the correct response?

A. Add emphasis weighting to the lantern count and move the negation to a negative prompt, so both constraints are conditioned more strongly.

B. Rewrite the prompt more precisely and at greater length, describing the position of each lantern relative to the doorway and the emptiness of the street.

C. Build the lanterns, their positions and the empty street into the first frame, and delete all three requirements from the text.

D. Raise the guidance scale, since all three failures are adherence failures and guidance is the control that governs how closely the prompt is followed.

CASE STUDY · MODULE 3

Eighteen students and one free allowance

A polytechnic media lab in Nagpur with one shared account on a hosted video tool, a free daily allowance of about twenty generations, and eighteen second-year students who all wanted to use it.

The lab had no budget line for credits. What it had was a free tier that refreshed each morning, a spreadsheet for booking slots, and a lecturer, Farid, who had watched the same thing happen for two terms.

The pattern was consistent. A student would get a slot, generate a clip, dislike it, rewrite the entire prompt, generate again, and repeat until the allowance was gone. Then they would post the final prompt on the class group as a discovery. The next student would copy it, get something quite different, and conclude that the tool was unreliable.

By the end of term the group chat held about forty prompts, no two of which agreed, and not one of which told anybody anything about the model. Farid described it accurately in a staff meeting as forty superstitions.

The cause was not laziness. Generation is stochastic, and two clips that differ in more than one respect cannot tell you which difference mattered. Every student had been comparing two random draws and reading meaning into the gap. Rewriting four words at once and keeping the better result produces a prompt you cannot explain and cannot repeat.

So at the start of the third term Farid proposed something the students did not like. For the first two weeks, nobody would make anything. The whole allowance would go on a fixed protocol: one still, one seed, one changed variable per generation, a written expectation before each result was viewed, and a shared notes file.

The cost of that was real and the students named it immediately. Two weeks of a fourteen-week term with nothing to show. No clips for their portfolios. Several of them had been waiting since first year to make something and were being told to run experiments instead.

The cost of the alternative was less visible, which is why Farid had to argue for it. Another term of forty superstitions, and every student rediscovering the same facts privately and wrongly.

They compromised on ten days. Six experiments were assigned, one per pair, each on the same supplied still with the seed fixed at a value written on the whiteboard.

The first pair tested motion strength, generating at four values and finding where visible artefacts began. The second tested whether the model would reframe from the supplied still at all when asked for a close-up. The third tested six phrasings of camera moves, one word changed each time, keeping the ones that produced a distinct and correct move. The fourth ran guidance at three values. The fifth tested whether the negative prompt did anything measurable by supplying an absurd one. The sixth found the step count at which improvement stopped being visible.

Each pair wrote one line of findings on the shared file and demonstrated it to the class with two clips side by side rather than describing it.

The experiment that changed the most behaviour was the second one. The model ignored requested shot sizes almost entirely and animated whatever framing the still already had. That single finding deleted a whole category of words from everybody's prompts and moved shot size out of the text and into the first frame, where it is a fact rather than a request.

The fifth pair found that the negative prompt box did nothing at all on their model, which is what a distilled model at guidance one does. Nine students stopped typing into it.

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 notes file ended at seven lines with a date on it. Across the remaining ten weeks the lab's recorded generations per finished clip fell from about nine to about three and a half, which meant the same free allowance produced roughly two and a half times as much finished work. Two students objected that the fortnight had cost them a project, and one of them was right: she had a deadline and lost it. Farid now runs the protocol in the first week rather than the third, and re-runs the six experiments whenever the provider announces an update, because a notes file without a date is worth nothing.

Why did forty prompts posted to a group chat produce no usable knowledge about the model?

Each prompt was arrived at by changing several things at once against an unfixed seed, so nobody could attribute the difference in output to any particular change. Generation is stochastic: a better clip is frequently just a better random draw. Without a fixed seed, a fixed still and exactly one changed variable, a comparison carries no information, and what accumulates is folklore rather than findings.

State the cost on both sides of Farid's proposal, and why the experiment about shot size repaid the fortnight on its own.

The cost of running the protocol was ten days of a fourteen-week term with no finished work, portfolios delayed, and at least one student missing a deadline. The cost of not running it was another term of contradictory prompts and every student paying to rediscover the same facts wrongly. The shot-size experiment repaid it because finding that the model ignores requested sizes and obeys the supplied frame removes an entire category of words from every future prompt and relocates the decision to the still, where it costs a hundredth as much to change.

Why does the notes file carry a date, and what should happen when the provider announces a model update?

Every finding is about one model version, with one set of weights, defaults and safety behaviour. A hosted provider may update the model, change a default scheduler or alter a filter without notice, and any of those can invalidate the measured numbers. The date is what tells a later reader whether the file still describes the model they are using. After an announced update the six experiments are re-run, because an out-of-date notes file is more dangerous than none: it is trusted.

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 Why does a prompt of "cinematic, epic, 8k" tend to return a slow drifting push-in with a teal and orange grade?

A. Aesthetic filtering removed every clip that was captioned with them
B. Those words are on a blocked list and are silently dropped
C. They sat under unrelated clips, so you get their average
D. The model reserves its best quality for prompts that request it plainly

2 An orbit preset is applied to a still of a subject standing against a flat, out-of-focus wall. The result reads as a cardboard cut-out sliding across a painted flat. Why?

A. The still has one depth plane, so there is no space behind the subject to reveal
B. The preset conflicts with a negative prompt suppressing warping
C. The model cannot perform lateral motion, only push-ins
D. Orbit presets need a higher motion strength than most hosted tiers permit you to set

3 Putting "dog" in the negative prompt frequently produces a dog-shaped smear rather than no dog. What is happening?

A. Negative prompts are applied only after the final denoising step
B. The model reads a negative prompt as a list of things to include for contrast
C. The tool concatenates the negative prompt onto the positive one
D. The token activates the concept, and the subtraction is a weak counterweight

4 Your clips keep coming back oddly frozen. Before you touch guidance or steps, what should you read?

A. The seed, in case it is being reused across takes
B. The negative prompt, for terms like movement or motion blur
C. The step count, because too few steps flatten motion
D. The model's native frame rate, in case interpolation has been disabled

5 A prompt asks for exactly three lamps and the clip contains four. The reliable fix is to

A. put three lamps in the first frame
B. write "exactly three lamps, not four" in the prompt
C. add "four lamps" to the negative prompt
D. raise guidance until the count is obeyed

6 You change four words in a prompt and the resulting clip is better. What have you actually learned?

A. That at least one of the four words was the effective one
B. That the model responds to longer prompts
C. That those four words together are an improvement and the prompt should be kept
D. Nothing yet: the draw may have been better rather than the prompt

MODULE 4

Making separate shots belong to each other

Nothing in a video model remembers your character, your room or your look between generations, so continuity is something you impose from outside. This block works through every mechanism available — a character sheet, reference conditioning, a trained identity model, reusable plates, a written look file, and repair after the fact — and finishes with the arithmetic nobody quotes: what a per-shot success rate does to the odds of getting eight matching shots in a row.

> BY THE END OF THIS MODULE YOU CAN
>
> Hold a character, a location and a look across a multi-shot sequence using a character sheet, reference conditioning or a trained model as appropriate, decide which of those a given job justifies, and state the compounding probability of a matching sequence from a measured per-shot hit rate

30 · 10 min

A cut breaks on the light before it breaks on the face

THE ONE THING TO KEEP

Identity holds because the first frame holds it, and a cut breaks on mismatched light and wardrobe long before it breaks on the face.

Why identity drifts

The model has no stored representation of your character. Every generation is a fresh sample conditioned on whatever you supplied. Two generations from the same description land on two nearby but different points, and nearby is not the same.

Faces are where this becomes intolerable because human vision is specialised for faces. A three per cent change in the distance between the eyes reads as a different person. A thirty per cent change in the shape of a chair reads as the same chair. The model is no worse at faces than at furniture; you are far better at noticing.

The four mechanisms, in order of strength

1. **A fixed first frame.** The strongest thing available, and free. If shot two begins from a still containing the face you want, the face survives that shot. This is the single biggest argument for the still-first workflow from the earlier lesson.
2. **Reference or element features.** Kling's elements, Runway's references, Higgsfield's character tools, Sora's consent-gated cameos. You supply images of a subject and the model conditions on them. These give you a consistent type of person and a consistent wardrobe. They do not reliably give you the same actor across ten shots.

3. **A trained character model.** A LoRA on an open-weights base is the strongest identity lock there is. It costs you a dataset of fifteen to thirty varied images, a training run, and the local setup from the last lesson. See fine-tuning.
4. **Description alone.** The weakest. "A woman in her forties with short grey hair" produces a genre of woman, differently each time.

The workflow that holds up

Build a character sheet before you generate a single clip:

- The face at frontal, three-quarter and profile.
- Full body in the exact wardrobe, front and back.
- The same lighting direction in every one of them.

Then every shot's first frame is built from that sheet — generated with the reference where the tool supports it, and composited by hand in Photopea or GIMP where it does not. Head from one image, body from another, background from a third. Compositing feels like defeat the first time you do it. It is faster than the twelfth generation, and it is deterministic, which no amount of re-rolling is.

Lighting and wardrobe break the cut before the face does

This is the part almost nobody warns you about. Two shots can have a genuinely matching face and still refuse to cut together, because the key light is on the left in one and the right in the other, or the shirt is a slightly different blue, or the sun is at a different height.

Editors call this continuity and audiences read it instantly without being able to name it. The feeling is "something is off", and people usually blame the face when the face was fine.

So decide the light direction for the scene before you start, put it in every first frame, and check the wardrobe colour by putting the stills side by side rather than by remembering.

Colour drift between generations can be pulled together afterwards. In DaVinci Resolve's free tier, put two clips on the timeline, use the shot-match to move one toward the other, then correct the skin tone by eye. That fixes a warm-versus-cool mismatch. It does not fix a different jacket.

Shoot around the problem

The professional answer is not to solve identity. It is to need less of it. Design the sequence so the face carries less of the load:

- Over-the-shoulder and back-of-head shots.
- Hands, feet, objects, details, the thing being looked at.
- Wide shots where the face is forty pixels across.
- Reaction cut to something other than the character.
- Fewer shots of the character than your instinct says.

A three-shot scene where the face is clearly visible once is far more achievable than a six-shot scene where it is visible every time. It is also, independently, better editing — which is the pattern this whole subject keeps producing.

The honest number

Even with references, a character sheet and matched lighting, expect to reject a substantial share of generations for identity alone, and expect the odds to compound across a sequence: a per-shot success rate that feels acceptable becomes a low probability of getting eight consecutive shots that all match.

Anyone showing you a character-consistent, ten-shot narrative without mentioning an enormous discard pile is showing you the survivors and not the attempt.

Today: build a character sheet for one person — three angles, one wardrobe, one light direction — before you generate any motion at all.

BEFORE YOU MOVE ON

A team locks a character with reference images. The faces match convincingly across six shots, but cut together the sequence still feels wrong and the client cannot say why. What is the most likely cause?

A. The face match is worse than it appears, and audiences are detecting differences too small to consciously identify.
B. The key light direction and the exact wardrobe colour differ between shots, which audiences read as a continuity break even when the face is right.
C. The clips were generated at different frame rates and the timeline is conforming some of them, producing judder.
D. The shots have not been graded to a single LUT, so the colour science is inconsistent across the sequence.

31 · 9 min

The document you build before you generate anything

THE ONE THING TO KEEP

A character sheet is a fixed set of images of one person under one light in one wardrobe, and it works because every later first frame is built from it rather than described afresh.

What it is, and what it is not

A character sheet is a small set of images of your character, made once, that every shot in the project is built from. Animation studios have used them for a century. They exist because a description is not a specification, and a drawing is.

It is not a mood board and not a set of nice pictures. It is a reference document with a job: any first frame you need later should be constructible from it,

by generation with reference, by compositing, or by hand.

The minimum sheet

Six images. Fewer and you will be describing rather than referencing.

1. **Frontal head and shoulders**, neutral expression, eyes to camera.
2. **Three-quarter**, the most useful angle for almost every shot.
3. **Profile.**
4. **Full body, front**, in the exact wardrobe, showing footwear.
5. **Full body, back.** Needed for over-the-shoulder and walking-away shots, which are the ones you will lean on hardest.
6. **Medium shot in context** — the character in the actual location, at the actual light.

All six at the same light direction, the same quality, the same colour temperature. That is the part people get wrong: six lovely images under six different lights is not a character sheet, it is six characters.

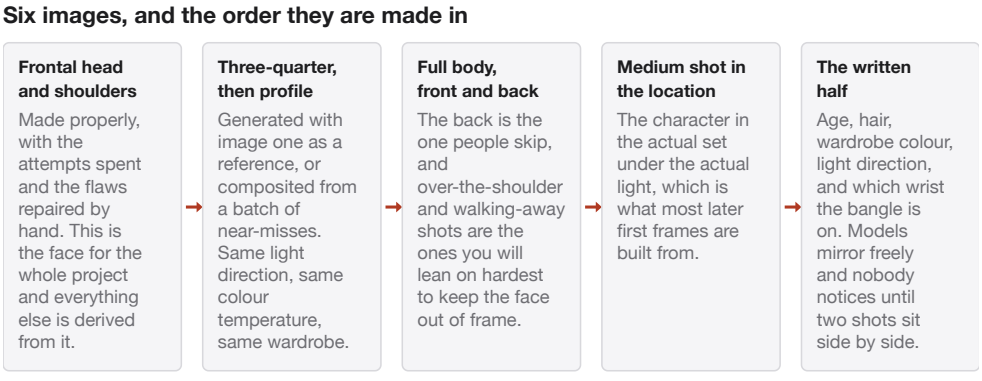
Building it

Make image one properly. Spend the attempts. Repair it by hand. This is the face for the whole project and everything else derives from it.

Then produce the others from it:

- With a reference feature, if your image tool has one — supply image one and ask for the three-quarter.
- By generating a batch and keeping the ones that match, if it does not.
- By hand, in Photopea or GIMP, for the composites: the correct head from one generation on the correct body from another.

The hand route feels like defeat the first time. It is faster than the twelfth generation, and it is deterministic, which no amount of re-rolling is.



Six lovely images under six different lights is not a character sheet, it is six characters. Put all of them on one canvas at the same size and fix every mismatch before you generate a second of video, because a mismatch here becomes a mismatch in every shot derived from it.

What goes in the written half

The sheet is images plus a short text file, because some things are easier to check as words than to read off a picture:

```
PRIYA
age ~34, medium-brown skin, short black hair tucked behind left
ear
small mole above right eyebrow
WARDROBE: olive green cotton kurta, sleeves to elbow,
  dark grey trousers, brown leather sandals, thin silver bangle
LEFT wrist
LIGHT: soft key from camera left, ~45 deg, slightly above eye
level,
  warm (window), low contrast
PROPS: canvas shoulder bag, mustard, worn on RIGHT shoulder
```

Note the sides. Left and right are where continuity dies, because a model will mirror things freely and you will not notice until two shots are next to each other.

This file goes into every still prompt for the character, in some compressed form. It is the thing you paste.

Design the character to be reproducible

A point that saves more time than any technique. Some characters are far easier to hold than others, and you usually get to choose.

Easier: distinctive silhouette, strong single hair shape, one saturated wardrobe colour, one memorable accessory, no fine pattern, no small jewellery, no glasses.

Harder: subtle features, elaborate patterned clothing, long loose hair, fine jewellery, glasses, anything with text on it.

Glasses deserve a specific mention: thin frames sit near the autoencoder's resolution floor and flicker, and the reflections on the lenses drift. If a character can be written without glasses, write them without glasses.

This is not a compromise imposed by the tool. It is costume design, and costume designers have always chosen silhouettes that read at distance and colours that survive the grade.

Checking the sheet before you commit

Put all six images on one canvas, at the same size, and look at them together. Specifically check:

- Is the light coming from the same side in every one?
- Is the wardrobe the same colour, not merely a similar colour?
- Is the accessory on the same wrist, the same shoulder?
- Does the hair part on the same side?

Any mismatch here becomes a mismatch in every shot that derives from that image. Fixing it now is one regeneration; fixing it later is a scene.

The free path

A browser image tool with a daily allowance for the generations, Photopea for the composites and repairs, and a text file. Total cost zero, total time a couple of hours for a character you will use for a whole project.

Today: build the six images for one character, put them on a single canvas side by side, and fix every mismatch before you generate one second of video.

BEFORE YOU MOVE ON

An artist produces six beautiful reference images of a character, each with excellent lighting suited to its own composition. Why is this not yet a usable character sheet?

A. Six images is too few to condition a reference feature reliably; fifteen to thirty are needed for the model to generalise the identity.

B. Reference features need consistent framing rather than consistent lighting, so the images should all be shot at the same distance instead.

C. The differing light means each image specifies a different key direction, so shots derived from them will not cut together.

D. Images generated separately carry different seeds, and without a shared seed the underlying facial geometry will differ between them in ways that accumulate across a sequence.

32 · 9 min

What a reference image is actually doing

THE ONE THING TO KEEP

Reference features condition on an embedding of the supplied images rather than storing the person, so they reliably reproduce a type — build, colouring, wardrobe, general impression — while the exact facial geometry keeps re-sampling within that type.

The feature, under several names

Kling calls them elements. Runway calls them references. Higgsfield has character tools. Sora has consent-gated cameos. Image models have their own versions, which matter here because your first frames come from them.

The common mechanism: you supply one or more images of a subject, an encoder turns them into a conditioning vector, and that vector is injected alongside the text and the first frame. The model is pulled toward the region of its output space that matches the reference.

That phrasing — *pulled toward a region* — is the whole lesson. It is not retrieval and it is not copying. There is no stored image of your person being pasted in.

What this gets you, reliably

- Build, height, apparent age.
- Skin tone, hair colour, hair length and general shape.
- Wardrobe colour and rough garment type.
- The overall impression somebody would describe in a sentence.

This is genuinely valuable. A reference gets you a consistent *type* of person across shots, which is far better than a description, and for background characters, crowds and anyone who is not the lead it is entirely sufficient.

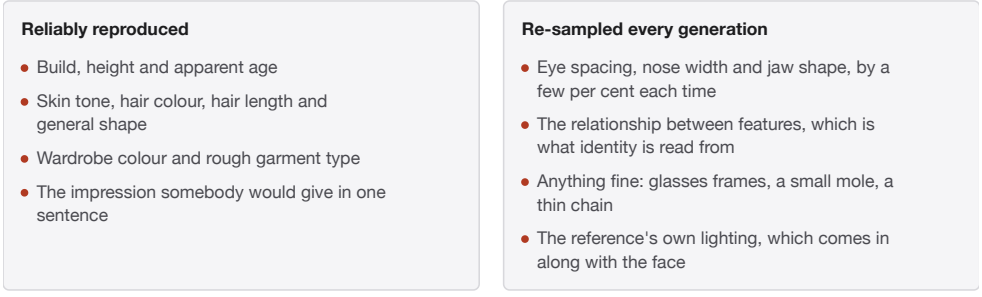
What it does not get you

The exact face. Over several shots, the facial geometry re-samples within the region the reference defines. Eye spacing, nose width, jaw shape and the relationship between features shift by a few per cent each time.

A few per cent is under the threshold for a chair and well over it for a face. This is the gap between "a woman who looks like her" and "her", and it is where character-consistency demos stop being impressive if you watch them carefully.

The honest formulation: references give you a consistent character *design*. They do not give you a consistent *actor*.

A reference gives you a design, not an actor



A few per cent of change is under the threshold for a chair and well over it for a face, which is exactly where the character-consistency demos stop being impressive if you watch them carefully. Use the reference to build a good first frame, then generate from the frame.

Getting the most out of them

Five things that measurably help:

1. **Multiple reference images, varied angles, consistent light.** Your character sheet is exactly this. One frontal image produces a model that only knows the front.
2. **Plain backgrounds in the references.** A busy background leaks into the conditioning. The model has no clean separation between "the person" and "the pixels around the person".
3. **Reference plus first frame, not reference alone.** The first frame is still the stronger signal. Use the reference to build a good first frame, then

generate from that frame, and the reference has done its most valuable work before the video model starts.

4. **Do not mix references of different people.** Two subjects in one conditioning produces a blend, and the blend is stable and wrong.
5. **Crop the references consistently.** Head-and-shoulders references produce head-and-shoulders influence. If you need the wardrobe held, include a full-body reference.

Where the leakage happens

A specific, common frustration worth naming. Supply a reference photographed outdoors at midday, then ask for an indoor evening shot. The output comes back oddly bright and cool, because the reference carried its lighting in with it.

The conditioning encodes the whole image, not a disentangled notion of identity. Same with background colour, same with the camera's apparent lens.

So make your references under the light you intend to use, which your character sheet already specifies. If you need the character under two different lights, make two sheets.

Consent-gated references

Some products require a person to record a verification video before their likeness can be used — a short clip, live, proving they are the person and that they agree. Sora's cameo mechanism is the best-known example.

This is a good design and it is worth understanding as a norm rather than as a feature. It moves consent from a checkbox you tick about somebody else to an act that person performs themselves. Whether a given product implements it well is a separate question; the shape is right.

It does not replace your own paperwork on a commercial job, which is dealt with later in the course.

The free path

Open-weights image pipelines expose reference conditioning — IP-Adapter and its descendants — in ComfyUI, free, and several Hugging Face Spaces demonstrate it in a browser. On the video side, reference features are currently more common on hosted tools, so the free route is to do the reference work at the still stage, where the free options are strong, and let the video model simply animate the frame you built.

That sequencing is not a workaround. It is the better workflow for cost reasons anyway.

Today: generate the same shot twice with the same reference images and compare the two faces at full size, side by side. The gap you see is the gap you will be managing all project.

BEFORE YOU MOVE ON

A reference image photographed outdoors at midday is used to generate an indoor evening shot, and the result comes back too bright and too cool. Why?

A. The conditioning encodes the whole reference image, including its lighting, rather than a disentangled representation of the person.

B. Reference conditioning is applied after the text conditioning in the pipeline, so it overrides any lighting described in the prompt.

C. Daylight-balanced images have a wider colour gamut, so the encoder assigns them higher weight than the darker regions the prompt requested.

D. The model interprets a high-key reference as a request for high-key output, because aesthetic filtering during training favoured brighter clips and biased the association.

33 · 10 min

When a LoRA is worth the afternoon

THE ONE THING TO KEEP

Training a small adapter on twenty images teaches the base model one specific identity as a real concept, which is the strongest consistency tool available and only worth it when the character will appear across enough shots to repay a fixed setup cost.

What a LoRA is, in one paragraph

A LoRA — low-rank adaptation — is a small set of extra weights trained on top of a frozen base model. Instead of retraining billions of parameters, you train a few million that nudge the model's behaviour toward one concept. The file is typically tens to a few hundred megabytes and loads alongside the base model at generation time.

For identity, this is the difference in kind. A reference *points toward* a region. A LoRA *teaches* the model a concept, which then behaves like every other concept it knows: it can be posed, lit, dressed, put at a different angle, and it stays itself.

The dataset, which is where the quality comes from

Fifteen to thirty images is the usual range. More is not automatically better; varied and clean beats numerous.

What the set needs:

- **Varied angles.** Frontal, three-quarter both sides, profile, slightly above, slightly below.
- **Varied expressions.** Neutral, speaking, a small smile. A set of thirty neutral frontals produces a model that can only do one face.

- **Varied framing.** Close-ups and at least a few full-body shots, or the model will not know what the body looks like.
- **Varied backgrounds and light.** This is the opposite of the character sheet rule and it is deliberate. If every training image has the same background, the model learns the background as part of the identity. Variety is what teaches it which parts are the person.
- **Consistent identity.** Every image must be the same person. One wrong face poisons the set.

Captions matter. Caption what varies and not what is constant. If every image is your character, do not describe her face in every caption — describe the angle, the setting, the clothing. What you leave undescribed is what the trigger word absorbs.

The training run

On video models, the practical route today is usually to train the identity on the *image* side and use those images as first frames, or to train a LoRA for an open-weights video model where adapter support exists.

Rough shape for an image LoRA, which is the accessible case:

- 20 images, 1500 to 3000 training steps, a learning rate around 1e-4, rank 16 to 32.
- Twenty to sixty minutes on a mid-range GPU.
- Free on Colab or Kaggle for smaller models; several Hugging Face Spaces offer one-click training.
- Tools: `kohya_ss`, `diffusers` training scripts, or a ComfyUI trainer node.

Overtraining is the common failure and it has a recognisable signature: the character appears in the right pose and the right clothes in every image, because the model memorised the training photographs rather than the person. If your outputs all look like your dataset, train fewer steps or gather more varied images.

When it is worth it

Honest arithmetic. A LoRA costs you: gathering a dataset, a training run, and the local setup. Call it half a day the first time and an hour once you have done it before.

It repays when:

- The character appears in more than roughly ten shots.
- The face is visible and close in several of them.
- The project will have revisions, so shots must be rebuildable months later.
- You will reuse the character across multiple projects — a brand mascot, a recurring presenter, a series.

It does not repay for a three-shot social clip, a background character, or anything where a reference feature already gets you close enough.

The part that is not technical

If the character is based on a real person, you need their written permission, specific to this use, before you build a dataset of their face. A trained model of somebody's likeness is a more consequential object than a single generated image: it is a reusable capability to produce new pictures of them indefinitely.

Treat it accordingly. Say what it will be used for, for how long, and what happens to the file afterwards. "We will delete the model at the end of the campaign" is a commitment you can make and should keep.

This is dealt with properly in the module on law and disclosure. It is raised here because the technical lesson is where the temptation arrives.

The free path

Genuinely complete. Open base models, free training scripts, free GPU sessions on Colab and Kaggle, and Hugging Face Spaces that will train a small LoRA from an upload. The binding constraint is a session that does not disconnect, which is the usual cost of free compute.

More general ground in fine-tuning and running-models-yourself.

Today: count how many shots in your project show one character's face at medium or closer. If the number is above ten, you have your answer about whether to train.

BEFORE YOU MOVE ON

A LoRA trained on thirty photographs of a character produces images that always show her in the same jacket, in front of the same bookshelf, in one of three poses. What went wrong?

A. The rank was set too low, so the adapter lacked the capacity to separate identity from context and collapsed onto the most frequent configuration.

B. The captions described the character's face in every image, so the trigger word absorbed the pose and setting instead of the identity.

C. The dataset lacked variation in setting and pose, and the model learned those as part of the identity rather than as incidental context.

D. The learning rate was too high for the number of steps, which drove the adapter into a sharp minimum that reproduces the training distribution exactly rather than generalising from it.

34 · 9 min

Building a room that stays the same room

THE ONE THING TO KEEP

A location is easier to hold than a face because a plate can be literally reused — generate the space once, then derive every shot in that scene from the same image by cropping, reframing and compositing.

The advantage nobody exploits

Everybody worries about faces. Locations get less attention and they are where the cheap win is, because a location does not have to be regenerated at all.

Generate the room once, at high resolution, as a wide establishing image. That image is now your set. Every other shot in the scene is derived from it:

- The medium shot is a crop of it, upscaled, with the subject composited in.
- The reverse angle is a new generation that uses it as a reference for colour, wardrobe of the walls, and light.
- The insert of the table is an extreme crop.
- The background behind the character in shot six is a piece of it, blurred.

This is how a real production works. They build one set and shoot it from several angles. You can do the same thing, and unlike a real production, your set costs one generation.

Generate the plate wide and large

Two rules for the establishing image:

1. **Wider than any shot you need.** You cannot crop out to reveal more. Generate more room than the script requires.

2. **Higher resolution than the video model's output.** If the video model runs at 720p, make the plate at 2048 pixels or more, so a crop to a medium shot still has enough pixels. Image models go much higher than video models, which is exactly why this workflow exists.

Then keep it in the frame library as `s0x_plate`, and derive from it rather than regenerating.

The reverse angle, which is the hard one

Cropping gives you every shot pointing the same way. The reverse — looking back from the other side of the room — is new information and the model has to invent it.

Three approaches, in order of reliability:

- **Avoid needing it.** Design the scene so the camera stays on one side. A great many scenes work this way and it is the cheapest answer.
- **Generate the reverse with the plate as reference**, describing the continuity explicitly: same room, same light from the same side, same wall colour, same floor. Then put the two side by side and check the light direction has not flipped, which is the failure you will actually get.
- **Build it.** If the room matters enough, block it out roughly in Blender — boxes for furniture, a light in the right place — render two angles, and use those renders as the structural base for both generations. This sounds like a lot until you have wasted an afternoon on the alternative. Blender is free and this level of use needs no artistry, only patience.

What changes and what does not

When a location does drift between generations, it drifts in a predictable order:

1. **Small objects move or vanish.** A cup on a table is not the same cup.
2. **Patterns change.** The rug, the tiles, the bookshelf contents.
3. **Wall colour shifts.** Usually warmer.

4. **Architecture stays roughly right.** Doors and windows tend to survive, because they are large.

So dress the set sparsely. A room with three objects in it holds. A room with forty holds nothing, and every one of those forty is a chance for a viewer to notice something moved.

This is another place where the constraint of the tool and good production design agree: a simple, strongly designed space reads better and reproduces better.

Exteriors, weather and time of day

Outdoor scenes carry an extra axis. The sun's position, cloud cover and the colour of the sky must match across every shot or the scene falls apart.

Put it in the shot bible as one line — "late afternoon, sun low from the west, thin high cloud, long soft shadows" — and repeat it verbatim in every prompt for the scene. Exteriors are also where the grade rescues the most, because a global colour shift across the sequence pulls varying skies toward each other convincingly.

What the grade does not fix is shadow length and direction. Decide once.

The free path

An image model with a daily allowance for the plate, Photopea for crops and composites, Blender for a rough blockout if you need reverses. Nothing on this page requires a subscription, and the plate-reuse workflow is a pure saving: one generation instead of six.

Today: generate one location as a wide plate at the highest resolution you can, then produce three different shot sizes from it by cropping alone. Compare the cost against generating three shots.

BEFORE YOU MOVE ON

Why should a location plate be generated wider and at higher resolution than any shot that will use it?

A. Higher resolution reduces the autoencoder's compression ratio, so derived frames lose less detail when the video model encodes them.

B. Video models accept larger inputs than they output, so the extra pixels give the model more information to condition on during generation.

C. Because tighter shots can be cropped out of a wide plate, but a wider shot cannot be recovered from a tight one.

D. Because an upscaler applied to the plate before cropping will distribute detail more evenly across the derived shots than upscaling each crop separately afterwards.

35 · 8 min

The blue that is not quite the same blue

THE ONE THING TO KEEP

Audiences detect wardrobe and prop discontinuity faster than facial drift because clothing is a large area of uniform colour, and colour differences of a few per cent across a large area are exactly what human vision is built to notice.

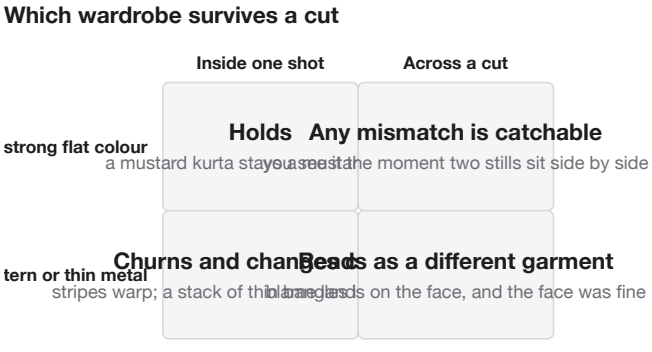
Why clothes break cuts

A face is small, complex and read by specialised machinery. A shirt is large, flat and read by the general colour system, which is extremely good at comparing two large uniform areas.

So a shirt that shifts from a slightly greenish blue to a slightly purplish blue between shots is noticed immediately, even by somebody who could not tell

you what changed. "Something feels off" is the usual report, and the face gets blamed when the face was fine.

This is the most under-rated continuity risk in generative video, and it is also one of the most fixable.



A shirt is a large area of uniform colour read by the general colour system, which is extremely good at comparing two large flat areas. A face is small and complex. That is why wardrobe drift is noticed first and why the face gets blamed for it.

Name the colour, do not describe it

"A blue shirt" is a family. The model picks a member of it each time.

Write the colour as narrowly as language allows: "a deep indigo cotton shirt, almost navy", "a mustard-yellow kurta", "a faded brick-red jacket". Narrow descriptions land in a smaller region and vary less between samples.

Better still, do not rely on language. Put the garment in the still by compositing or by deriving each first frame from the character sheet, where the colour is already fixed as pixels rather than as a request.

Design the wardrobe to be holdable

The same logic as character design, and it is legitimate costume work:

- **One strong colour per character.** It reads at distance, it survives the grade, and it makes mismatches obvious enough that you catch them.
- **Avoid fine patterns.** Stripes, checks and small florals are below the encoder's resolution and will churn, warp and change count. A large bold pattern survives; a small busy one does not.

- **Avoid text on clothing.** Same reason as text anywhere.
- **Keep accessories few and large.** A single wide bangle holds. A stack of thin ones becomes a different stack every shot.
- **Decide the side.** Bag on which shoulder, watch on which wrist, hair parted which way. Write it down, because models mirror freely.

Props that have to persist

Any object a character carries through several shots is a continuity liability, for the reason established earlier: there is no object memory, and an item that leaves frame and returns has been re-sampled.

Three responses:

- **Reduce the number of persistent props to as close to one as the story allows.**
- **Keep the prop in frame** rather than letting it exit and return.
- **Composite it.** If the prop is specific — a particular book, a particular phone, the client's product — it goes in by hand, which was the conclusion of the compositing lesson and applies here unchanged.

Catching mismatches before you generate the clip

The check takes two minutes and it is the whole defence.

Put every first frame for a scene on one canvas, at the same size, side by side. Then look at:

- The garment colour as a block. Squint, or apply a heavy blur to all of them at once — blurring removes the detail and leaves the colour, which is exactly what you are comparing.
- Which side the accessory is on.
- The hair shape.
- The prop's presence and position.

Doing this in Photopea takes as long as opening the files. Doing it after generating costs you the generations.

What the grade can and cannot rescue

A warm-cool mismatch across the whole frame is fixable: a shot-match in Resolve's free tier pulls one shot toward another in a minute, and a single grade across the timeline unifies the rest.

A garment that is a different colour while the rest of the frame matches is not fixable that way, because a global correction would break everything else. It needs a secondary — isolating that colour range and shifting it alone — which Resolve's free tier does support and which is fiddly, imprecise on moving subjects, and sometimes the only thing standing between you and a reshoot.

The better answer is not to need it.

The one that always catches people

Mirroring. A model will happily flip a composition, and the character's bag moves to the other shoulder while everything else looks fine. It is invisible in a single shot and glaring across a cut.

Check sides explicitly, every time. It is the single highest-yield item on the list.

Today: put every first frame for one scene side by side, apply a heavy blur to all of them, and compare the wardrobe colours as blocks.

BEFORE YOU MOVE ON

Why does a slightly mismatched shirt colour between two shots get noticed faster than a slightly mismatched face?

A. Clothing occupies a large uniform area, and comparing two large flat colours is something ordinary colour vision does extremely well.

B. Faces are processed by specialised machinery that tolerates small variation as natural change in expression and lighting.

C. Colour drift between generations is larger in magnitude than geometric drift, so the shirt simply changes more than the face does.

D. The autoencoder preserves large flat regions more faithfully than complex ones, so any colour change in the garment is rendered with full fidelity while facial change is smoothed away by the decoder.

36 · 9 min

Fixing a face after the clip exists

THE ONE THING TO KEEP

Face restoration and face replacement operate per frame on a finished clip, which makes them useful for small drift and dangerous for large changes — and the same tooling that rescues your own shot is the tooling behind non-consensual synthetic media.

Two different tools with similar names

Face restoration improves a face that is already the right face but degraded — soft, low resolution, slightly malformed. Models such as GFPGAN and CodeFormer take a face region and reconstruct plausible detail: eyes, teeth, skin texture, edges.

Face replacement puts a different face onto the body in the footage. Different job, different ethics, dealt with below.

Both run per frame on a finished clip. Both are widely available, free, and easy.

When restoration genuinely helps

A face at a distance in a 720p generation is perhaps eighty pixels across. The encoder discarded most of its detail and the decoder invented a soft approximation. Restoration is the honest fix for exactly this, and it is the difference between a usable wide shot and an unusable one.

Good cases:

- Faces in the middle distance, soft but correct.
- A close-up where eyes and teeth are mushy but the geometry is right.
- Upscaling a lower-resolution generation where the face is the limiting factor.

Where it fails, and the failure is specific

Temporal flicker. Restoration runs per frame. Each frame's face is reconstructed independently, so the invented detail differs slightly from frame to frame. The result is a face that boils — skin texture crawling, eye highlights popping, teeth shifting. On a still it looks great; in motion it announces itself.

Mitigations, in order:

- Use a lower restoration strength. CodeFormer's fidelity weight trades detail against faithfulness to the input; keeping it closer to the input reduces flicker sharply.
- Blend the restored face back at partial opacity over the original. Fifty per cent often looks better than a hundred.
- Use video-aware restoration where available, which considers neighbouring frames.
- Restore fewer frames: on a short shot, restore and check rather than restoring the whole timeline by default.

Identity shift. Aggressive restoration invents a more conventionally attractive, more symmetrical, younger face. It is pulling toward the average of its training data. On a character with distinctive features this quietly replaces them, and on non-white faces the pull toward a training-set average has been a documented and repeatedly criticised problem. Check the output against your character sheet rather than against your impression.

Face replacement, and the line

The technical description is short: detect the face, align it, swap in a face generated or transferred from a source, blend the edges, run per frame.

Legitimate uses exist and are real. Holding your own trained character across shots where the generation drifted. Replacing a background actor's face with a synthetic one to avoid needing their release. Fixing one shot in a sequence where everything else matched.

The illegitimate use is the one this technology is famous for, and the mechanics are identical. There is no technical difference between the operation that rescues your shot and the operation that puts a real person into a video they were never in.

So the line is not technical, it is about permission and disclosure:

- A real person's face requires that person's written, specific permission. Consent to be filmed is not consent to be synthesised.
- The output requires a visible label, for reasons dealt with properly in the final module.
- Many jurisdictions now legislate here specifically, and the direction of travel is one way. Non-consensual intimate synthetic imagery is criminal in a growing number of places, and personality-rights actions over synthetic likeness have succeeded in several courts since 2023.

If you find yourself reasoning about why a particular case is an exception, that is the signal to stop and ask the person.

The free path

GFPGAN, CodeFormer and the current successors are free and open, available as Hugging Face Spaces for browser use and as ComfyUI nodes locally. FFmpeg extracts frames and reassembles them. Resolve's free tier blends a restored sequence back over the original with an opacity slider.

None of this needs a paid tool, which is precisely why the ethical question cannot be left to whoever can afford the software.

Today: restore one face in a soft wide shot at low strength, blend it back at fifty per cent, and watch the result in motion rather than as a frame.

BEFORE YOU MOVE ON

A restored face looks excellent in a single frame but appears to crawl and boil when the clip plays. What causes this?

A. The restoration model was trained on still photographs, so it produces detail at a spatial frequency the video codec cannot encode without shimmering.

B. Restoration runs independently on each frame, so the detail it invents differs slightly every time and that difference reads as movement.

C. Restoration increases contrast in the face region, and the interpolation step then amplifies those differences into visible flicker between frames.

D. The face detector's bounding box shifts by a pixel or two between frames, so the alignment differs and the reconstructed face is resampled at a slightly different position each time.

37 · 8 min

Deciding the look once, in writing

THE ONE THING TO KEEP

A look is a set of decisions about colour, contrast, texture and lens that should be made once and written down, because a piece where every shot was styled individually reads as a collection rather than as a film.

Why this needs a document

Over thirty shots generated across several sessions, taste drifts. Monday's shots are cooler than Thursday's. A shot generated after looking at a reference is greener than the one before it. None of the drift is visible while you work and all of it is visible on the timeline.

The fix is the same one productions have always used: decide the look once, write it down, and refer to the document rather than to your memory.

What goes in it

Seven lines. It fits on a card.

```
LOOK — project name
1. PALETTE: desaturated, cool shadows, warm skin. No pure
blacks.
2. CONTRAST: medium-low. Lifted blacks, no clipped highlights.
3. LENS: 35mm wides, 85mm anything closer. Shallow focus
throughout.
4. TEXTURE: light 35mm grain over the whole timeline, added in
the edit.
5. MOTION: low. Slow moves only. No handheld except shots 12
and 19.
6. TIME: late afternoon throughout. Sun low, from the west.
7. FORMAT: 16:9, 24fps, delivered 1080p.
```

Every one of those is a decision you would otherwise make thirty times, differently.

Look references, and how to use them without copying

Collect three to five reference frames from existing work that has the look you want — a film still, a photograph, a painting. Put them on one page.

Use them as a target to check against, not as a thing to feed into a model. The distinction matters practically and ethically. Practically: a reference image fed as style conditioning brings its composition and its subject matter along with its palette, which is rarely what you want. Ethically: deriving a palette and a contrast approach from existing work is normal craft, and generating output designed to reproduce a specific living artist's recognisable style is a different act with contested legal status, dealt with in the final module.

The safe and better workflow: look at the reference, describe what you see in words — "cool shadows, warm skin, lifted blacks, soft key" — and put those words in your look file. You have extracted the decision rather than the image.

Applying the look at three stages

You can put the look into the piece in three places, and doing it in the wrong one costs you.

In the still. Colour, contrast and lens character baked into each first frame. Strong, consistent, and expensive to change — if the client wants it warmer, you regenerate everything.

In the prompt. Weak and variable. Useful as reinforcement, useless as the only mechanism.

In the grade. One adjustment layer across the whole timeline in the editor. Weakest control over what the shot contains, total control over how it looks, and changeable in thirty seconds at any point.

The working answer is **neutral stills, unified grade**. Generate frames that are correctly exposed and reasonably neutral, then put the look on in the edit.

It keeps the expensive stage flexible and the cheap stage decisive, which is the same principle as the whole course.

The exception is anything stylised enough that the model has to know about it — a clip meant to look like gouache, or like 1970s film stock with bloom and halation. That has to be in the still, because no grade invents it.

Grain, once, at the end

A single grain pass over the finished timeline does more for cohesion than any other single operation, and the mechanism is worth restating: the autoencoder removed real texture, so every clip has the same suspiciously smooth surface. One grain layer gives them all the same *new* surface, which the eye reads as one medium rather than thirty renders.

Do it once, on the whole sequence. Doing it per clip gives you thirty slightly different grains, which is worse than none.

The free path

A text file, a page of reference images, and DaVinci Resolve's free tier for the grade and grain. Kdenlive and Shotcut will do a simple grade and a grain overlay. On a phone, CapCut has filters and a grain overlay and will apply one to every clip, which is crude and serves the purpose.

Today: write the seven lines for whatever you are working on, then check your last three generations against them and note where you have already drifted.

BEFORE YOU MOVE ON

Why is it usually better to generate neutral stills and apply the look in the grade, rather than baking the colour treatment into each first frame?

A. Graded footage compresses better on delivery, so applying colour late produces a smaller file at the same perceived quality.
B. A grade is one adjustment across the timeline that can be changed in seconds, while baked-in colour means regenerating every shot to alter it.
C. Colour described in a prompt is unreliable, so stills generated with a look specified will vary more than neutral ones and make matching harder.
D. The video model re-encodes the first frame through its autoencoder, which shifts saturated colours unpredictably and makes any baked-in grade inconsistent between shots.

38 · 9 min

The arithmetic nobody puts in the demo reel

THE ONE THING TO KEEP

Per-shot success rates multiply, so a hit rate that feels acceptable on one shot becomes a low probability across a sequence — which is why the professional response is to reduce the number of shots that must match rather than to improve the rate.

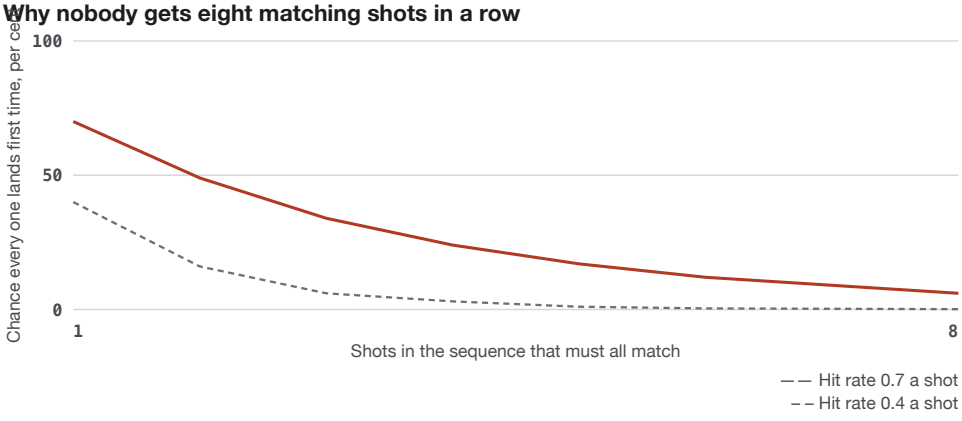
The number that decides your project

Suppose a shot has to hold a character's face convincingly, and your generations do that acceptably seven times out of ten. That feels fine.

Now you need eight consecutive shots that all match. The probability that all eight land, if attempts are roughly independent, is 0.7 to the power of 8, which is about 0.06. Six per cent.

You will not get eight in a row. You will generate each shot until it works, which is the real workflow — but the arithmetic tells you the *cost*: at a 0.7 hit rate you generate on average 1.4 takes per shot, so eleven or twelve generations for the sequence, plus the ones you reject for other reasons.

Drop the hit rate to 0.4 — which is realistic once a shot has a specific requirement — and it is two and a half takes per shot, twenty generations, and a meaningfully worse chance that any of them are actually good rather than merely acceptable.



You will not try for eight in a row; you will generate each shot until it works. What this arithmetic actually prices is the takes — about 1.4 generations a shot at a rate of 0.7, and about 2.5 at 0.4. The professional response is to reduce the number of shots that have to match, not to chase the rate.

Measure your own rate

The number above is invented. Yours is not, and it takes a week of normal work to find.

Keep a tally. For every shot: how many generations before one you would ship, and what the rejections were for.

```
s04  3 takes  (1 face drift, 1 hand)
s05  1 take
s06  6 takes  (3 face drift, 2 background churn, 1 too much mo-
tion)
s07  2 takes  (1 wardrobe colour)
```

After twenty shots you have two things: a hit rate you can quote in a budget, and a ranked list of what actually causes your rejections. Most people are surprised by the second. The thing you worry about is rarely the thing that costs you.

Reduce the exposure, not the rate

You have limited control over the per-shot rate. You have complete control over how many shots have to match.

This is the professional response and it is why it keeps appearing in this course:

- **Fewer shots of the character.** A three-shot scene where the face is clearly visible once is far more achievable than a six-shot scene where it is visible every time.
- **Over-the-shoulder and back-of-head shots.** The face is not in them, so it cannot drift.
- **Wides where the face is forty pixels across.** Drift is present and invisible.
- **Cutaways.** Hands, objects, the thing being looked at, a detail of the room. Each one is a shot that carries the scene without exposing the hard thing.
- **Reaction cut to something other than the character.**

Every one of those is also, independently, better editing. A scene told entirely in shots of a face is weak filmmaking regardless of how it was produced.

Where the rate genuinely improves

In order of effect:

1. **A trained identity model.** The largest single improvement available.
2. **First frames derived from a character sheet** rather than described.
3. **Shorter shots.** Drift grows with distance from the anchor, so a four-second shot holds where an eight-second one does not.
4. **Lower motion.** Fewer frames of change, fewer chances to break.

5. **Reference conditioning.** Real, and weaker than the above.

Notice that three of the five are decisions about the shot rather than about the model. That ratio is the lesson.

What this means for a quote

If you are billing, the hit rate is the number your price is built on, and it belongs in the conversation rather than hidden inside an hourly rate.

Quote in shots with a stated number of included attempts. "Thirty shots, up to four attempts each, further revisions billed per shot" is a sentence that protects both sides: it tells the client that attempts cost money, and it tells you what you agreed to when they ask for the twelfth version of shot nine.

The money module deals with this properly. It is raised here because the hit rate is the input.

The honest note about demos

Anyone showing a character-consistent, ten-shot narrative without mentioning a discard pile is showing you the survivors. That is not dishonesty exactly — it is a showreel, and showreels have always been the top one per cent — but it sets an expectation that costs beginners real money and real confidence.

The rate is not a measure of your skill. It is a property of the tool, and the skill is in designing work that does not depend on it being high.

Today: start the tally. Twenty shots from now you will have a number you can put in a quote, and a ranked list of what actually goes wrong in your work.

BEFORE YOU MOVE ON

A team measures a 0.7 per-shot hit rate for holding a character's face and needs an eight-shot scene. What does the compounding arithmetic most usefully tell them?

A. That the scene is not achievable, since a six per cent chance of eight consecutive successes is too low to plan around.
B. That the hit rate must be raised above 0.9 before an eight-shot scene can be attempted at a reasonable cost.
C. That they should budget roughly twelve generations for the scene, and reduce how many shots must show the face.
D. That the shots should be generated in a single batch with a shared seed, since independent attempts are what cause the probabilities to multiply and a shared draw would correlate them.

39 · 10 min

Six things it cannot do, and the reason for each

THE ONE THING TO KEEP

Each failure has a mechanism — no simulation, no symbols, no persistent objects, no reference for your specific thing — and the mechanism tells you whether to wait for a better model or change the shot.

Hands doing precise tasks

Threading a needle, tying a lace, typing, playing a guitar, counting notes. The mechanism is contact and occlusion. The model learned the statistics of what hands look like; it did not learn the constraint that a finger cannot pass through a string. Contact events are exactly where a plausible-looking frame

becomes an impossible one, and hands make contact constantly, at speed, while occluding each other and the object.

This is not a resolution problem and more sampling steps will not fix it.

Work around it: cut away at the contact. Hand approaches, cut, result. A century of film grammar exists for precisely this reason, because real crews could not always film the contact either.

Readable text

Signs, packaging, screens, anything with letters in the frame. Text is a symbolic system with a small alphabet and strict rules; the model has pixel statistics. Image models improved on this enormously. Video models are behind, and there is an extra reason: a letterform must stay identical across a hundred and fifty frames, and one that shifts by two pixels a frame reads as melting even when each individual frame is fine.

Work around it: composite the text in afterwards, in the editor, as an actual title with an actual font. Or put it in the first frame and keep the camera nearly still — which sometimes survives and usually does not.

Physical cause and effect

A glass that falls and shatters into the right number of pieces. A ball that bounces lower each time. Cloth that settles and stays settled. Fire that reacts when the wind changes. The model interpolates plausible-looking frames; it does not simulate. There is no state, no mass, no conservation of anything.

It gets short, common, heavily represented physics roughly right — a person walking, water running downhill, hair moving in a breeze — because those patterns are everywhere in the training data. Anything unusual it gets wrong, confidently.

Work around it: single events, short, and never in slow motion. Slow motion is where fake physics is most visible, because you have given the viewer time to check.

A specific real product

Your client's bottle, with their label, their cap, their exact proportions. The model has no reference for that object. It produces a member of the category with a label that is almost right — and almost right is worse than obviously wrong, because it reads as a counterfeit rather than as a stylisation.

Work around it: this is a compositing job, not a generation job. Generate the environment and the camera move; mask and track the real product in. Or shoot the product on a phone against a plain wall, which is one of the places where thirty seconds of real filming beats an afternoon of generating and everyone can see it did.

Continuity over minutes

Covered earlier, but the mechanism is worth stating once more because it explains four of the other five items. There is no persistent object memory and no scene model. Every frame is inferred, the only anchor is the conditioning image, and error grows with distance from that anchor. Anything that requires the same object to be the same object several shots later is fighting the architecture, not the settings.

Action timed to a beat

"She looks up on the fourth beat" is not currently something you can ask for. Duration maps loosely onto the request and the timing of an action within a clip is not controllable.

Work around it: generate longer than you need and find the frame in the edit, then slip the clip until the action lands. That is what an editor does with real footage anyway.

Two things people over-claim in both directions

Not all of this gets fixed by the next model. Hands and short-range physics have improved measurably from one generation of models to the next, and will keep improving. A specific real product is different in kind — it is not

a capability gap, the information simply is not available to the model, and no scale of training changes that.

Not all of it is permanent either. Text in image models was a standing joke in 2022 and mostly works now. So write down the mechanism rather than the verdict, and re-test this list yourself every few months on your own shot rather than believing anyone, including this lesson.

The most valuable thing you can tell a client or a team is where the tool stops. Being the person who says on day one that a shot is not achievable is worth considerably more than being the person who can drive the tool and discovers it on day fourteen with the deadline in sight.

BEFORE YOU MOVE ON

A brand asks for a hero shot of their bottle rotating slowly on a plinth. The generated bottle looks convincing at a glance, but the label is subtly wrong and the cap is the wrong shape. What is the right diagnosis and fix?

A. The model has no reference for that specific object and can only produce a member of the category, so the real bottle should be shot or photographed and composited into the generated environment.

B. The prompt did not describe the label and cap precisely enough; a detailed written specification of both will resolve it.

C. The generation resolution is too low for the label to resolve, so a higher-resolution tier will render it correctly.

D. Feeding the model a folder of brand photographs will teach it the bottle, after which it will render it accurately.

CASE STUDY · MODULE 4

Six shots, one face, and the arithmetic

A three-person agency in Kochi pitching a sixty-second film for a regional bank, built around one recurring character — a shopkeeper — appearing in six shots across two locations.

The pitch was won on a moving storyboard. The production was where the trouble started.

Divya, who was directing, had generated the character once and liked her: a woman of about fifty in a dark green saree behind a counter. She built the rest of the shots by describing her again in each prompt.

By the third shot it was obvious that this was not the same person. Not wildly different — the same build, the same colouring, the same green — but the eye spacing had moved, the jaw was rounder, and a colleague looking at the three stills together said, without being prompted, that the middle one was the shopkeeper's sister.

Divya already understood why. A reference or a description conditions the model toward a region of its output space. It reproduces a type reliably: build, colouring, wardrobe, the impression somebody would give in one sentence. It does not store the person, so the facial geometry re-samples within that region every time, by a few per cent. A few per cent is nothing on a chair and is the difference between two people on a face.

Then she did the arithmetic that decided the job. Her measured hit rate for a shot that had to hold this face at medium size was about two in three. Six shots that all had to match, at two in three, is a probability of about nine per cent that a first attempt at each lands. She was not going to try for six in a row — nobody does — but the number told her what the sequence would cost in takes, and it was roughly two and a half generations per face shot before anything else went wrong.

There were two routes and the agency spent a morning on the choice.

The first was to train a small character model. Twenty varied images of the shopkeeper, an hour or two of training on a free GPU session, and the base model would then know this specific identity as a concept that can be posed, lit and dressed. It is the strongest identity tool available. It cost them roughly half a day for somebody who had not done it before, a local setup, and a session that might disconnect at eighty per cent. It also raised a question they had not thought about: the character had been generated rather than based on a real person, so there was no consent problem — but Divya wrote that check into the process afterwards anyway, because the next brief might not be so simple.

The second was to need the face less. Build a proper character sheet — six images, one light direction, one wardrobe, front and back — derive every first frame from it, and then redesign the sequence so that the face carried less of the load. An over-the-shoulder for the opening. A wide where the face is forty pixels across. Two cutaways: hands counting notes, and the thing she is looking at. One clean medium shot where the face is genuinely visible and where they would spend the attempts.

The cost of the second route was a conversation with the client about a storyboard that had already been approved with six shots of a face in it.

What made the decision was the number of shots, not the technology. A trained model repays itself when a character appears in more than roughly ten shots, or across several projects, or where revisions are likely. This was six shots in one film.

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 built the character sheet and redesigned the sequence. Four of the six shots no longer showed the face clearly, and the client approved the change in

a fifteen-minute call because the new version was described to them as a better-edited scene, which it also genuinely was. The one clean medium shot took five generations. The whole sequence took eleven, against the twenty-odd Divya had projected. The rejection tally, which she kept for the first time on this job, was not what she expected: face drift caused four of the rejections and mismatched saree colour caused five, because a large flat area of dark green is exactly what human vision compares best. The next brief, a series with the same character across three films, got the trained model.

A reference image reproduced the shopkeeper's build, colouring and saree accurately but not her face. Why is that the expected outcome rather than a failure?

Reference conditioning encodes the supplied images into a vector and pulls generation toward the matching region of the output space. It is not retrieval and nothing is stored or pasted in. Everything that is broad — build, height, colouring, hair shape, wardrobe colour, general impression — sits well inside that region and comes back consistently. The exact relationship between facial features is fine detail that re-samples within the region each time, by a few per cent. Human vision is specialised for exactly that few per cent, so the same drift that is invisible on furniture reads as a different person.

Why did six shots rather than sixteen decide against training a character model?

A trained adapter has a fixed setup cost — gathering and captioning a varied dataset, a training run, and a local or free-GPU session that may fail — of roughly half a day the first time. That cost is repaid when the character appears often enough: across more than about ten shots, in several close shots, across multiple projects, or where revisions months later must match. For six shots in one film the fixed cost is not recovered, and a character sheet plus a redesigned sequence achieves enough at a fraction of the effort.

The rejection tally showed more takes lost to saree colour than to facial drift. What should the team change because of that?

They should fix the wardrobe as pixels rather than as words: derive every first frame from the character sheet where the green is already decided, or composite the garment, instead of describing it in each prompt. They should also check the stills side by side with a heavy blur applied, which strips detail and leaves colour

blocks to compare. More generally, the tally is the point of keeping one — effort should go where the rejections actually are, and that is rarely where the team assumed.

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 Two shots of the same character have genuinely matching faces and still refuse to cut together. What has usually broken?

A. The focal length named in each prompt
B. The number of denoising steps used for each generation
C. The frame rate, if the two clips were generated at different native rates
D. The light direction or the wardrobe colour

2 What does a trained character adapter do that reference conditioning cannot?

A. It stores the reference photographs and pastes the face into each frame
B. It teaches the base model the identity as a concept it can pose and light
C. It removes the need for a first frame
D. It guarantees that the same seed returns the same face on a hosted service

3 Why is a location easier to hold across a sequence than a face?

A. The plate can be literally reused: crop it, reframe it, composite into it
B. Models were trained on more interiors than on people
C. Architecture is simple geometry that the model can simulate
D. Rooms contain fewer high-frequency details than a face does, so less can drift

4 You run face restoration over a soft wide shot. Each frame looks better and the face boils in motion. Why?

A. The blend opacity was set too high for the codec to carry
B. Restoration has to be applied before the clip is interpolated rather than after
C. The restorer runs at a lower frame rate than the clip
D. Each frame is restored on its own, so invented detail changes every frame

5 Your measured hit rate on a face shot is about 0.7, and the scene needs eight shots that all match. What is the professional response?

A. Accept a six per cent chance and rebuild the sequence if it fails
B. Generate all eight repeatedly until every one of them matches
C. Reduce the number of shots in which the face has to match
D. Raise the motion strength so the shots feel more alike

6 Across a cut, your character's shoulder bag has moved to the other shoulder. What happened, and what is the check?

A. The model lost track of the prop and re-sampled it, so add the bag to the negative prompt
B. The composition was mirrored; check sides explicitly, on stills, side by side
C. The wardrobe colour drifted; run a secondary correction in the grade
D. The seed changed between the two shots; reuse one seed for the sequence

MODULE 5

The half of video that is not picture

Sound is the cheapest quality upgrade in the pipeline and the one most people skip. This block covers the talking head honestly — what lip-sync is actually reshaping and where it fails, what performance transfer captures that phonemes do not — and then the audio nobody sees: cloned and recorded voice, room tone, spot effects, music licensing you can defend, and why a well-chosen sound makes a viewer forgive physics the model got wrong.

> BY THE END OF THIS MODULE YOU CAN
>
> Build the audio half of a generated video — a speaking character whose consent you can document, a voice recorded or synthesised with a stated quality trade-off, room tone and spot effects under every shot, and music whose licence you could show a platform — and explain why sound repairs perceived physical errors that no regeneration would fix

40 · 10 min

Lip-sync is easy; the permission is not

THE ONE THING TO KEEP

Permission to film is not permission to synthesise, and the only label that survives a screenshot is the one burned into the frame.

Two families of talking-head tool

Audio-driven lip-sync. You supply a video or a still plus an audio track, and the model reshapes the mouth region to match the phonemes. The older open-source line — Wav2Lip and its descendants, LatentSync and similar — does the mouth only, and does it well enough at small sizes. Hosted products such as Hedra, HeyGen and Synthesia, and the sync features inside the larger platforms, animate the mouth plus some head movement, blinks and eye direction.

Performance transfer. You record yourself acting on a webcam and the model maps that performance onto a generated character. Runway's Act-One is the best-known example of the class. This is substantially better than lip-sync alone, because the timing, the pauses and the eyebrows come from a person who understood the line. Delivery is most of acting, and no phoneme model invents delivery.

Where they break, and why

- **Profile angles.** The mouth has to be visible and roughly frontal. Past about thirty degrees off-axis the model is guessing at geometry it cannot see, and the jaw starts sliding. Keep the head near frontal for any shot that has to speak.
- **Plosives.** B, P and M require the lips to close completely. Many models under-close them, and the result reads as badly dubbed foreign footage —

most viewers cannot name what is wrong but all of them notice. Slowing the delivery slightly improves it.

- **Teeth and tongue.** The interior of the mouth is still where close-ups fall apart. Stay at a medium shot and the problem largely disappears.
- **Long takes.** The head loses its anchor over time, exactly as everything else does. Cut every few seconds.

Voice

Cloning a voice from a short sample is a solved problem, and it is cheap. The free path is real: open models such as Coqui XTTS, Piper and Kokoro run on modest hardware and produce serviceable synthetic speech. Hosted services like ElevenLabs are better and cost money. Audacity — free, runs on almost anything, including very old machines — is enough to trim, level, remove a hum and normalise. Recording a real voice on a phone in a quiet room, cleaned in Audacity, still beats most synthetic speech for anything longer than a few sentences.

The line, stated plainly

You need permission to use a real person's face or voice. This is not a preference, a style question or a thing that becomes acceptable at sufficient production value.

- Get it in writing, and get it specific: what the likeness will be used for, on what platforms, for how long, and whether it may be altered or made to say new things.
- Consent to appear is not consent to be synthesised. A client who filmed an interview with an employee last year did not thereby acquire the right to generate new sentences in that employee's voice. Those are different acts and the first does not license the second.
- Public figures and dead people are not a loophole. India's courts have issued a run of personality-rights orders since 2023 protecting named actors' likeness and voice. Denmark moved to give people a right in their own likeness. Many US states have publicity-rights statutes, and the 2025

TAKE IT DOWN Act obliges platforms to remove non-consensual intimate imagery, synthetic included. The direction of travel is one way.

- The consent mechanism inside some products — Sora requires a person to record a verification video before their likeness can be used as a cameo — is a decent model for how to think about this. It is not a substitute for your own paperwork on a commercial job.

That is the shape of it, not advice for your case. A lawyer in your own country is the person who answers it for a specific job, and the answer differs by country more than people expect.

The labelling duty, which is separate

Consent is about permission. Labelling is about the audience, and it is a distinct obligation that applies even when everyone involved agreed.

Several countries now require synthetic media to be labelled visibly. China's rules require both a label a person can see or hear and an implicit label inside the file's metadata. India amended its IT Rules for synthetically generated information in a way that prescribes visibility — a label covering a defined portion of the frame and the opening portion of the audio, plus a duty on platforms to verify declarations. The EU AI Act carries a transparency obligation covering deepfake disclosure. Timetables have shifted more than once, so confirm the current position rather than trusting any date, including this one.

Content Credentials, built on the C2PA standard, and invisible watermarks such as SynthID are real and worth keeping in the file. Both are fragile. A screenshot destroys them. Many social re-encodes strip them. The label that survives contact with the internet is the one burned into the frame.

So put it in the video, at the start, where a viewer sees it before they believe it. It costs you a title card and about two seconds.

Today: if you have a project with a real person in it, write down four things — whose face, whose voice, what you hold in writing, and what the video says on

screen about being generated. Any blank is your next task, before the next generation.

BEFORE YOU MOVE ON

A client supplies interview footage of an employee filmed last year, with a signed release for that shoot, and asks you to generate three new sentences in her voice for an updated cut. What is the correct read?

A. She is already in the footage under a signed release, so her voice is licensed to the client for this project and the new lines are covered.

B. It is acceptable as long as the new sentences are things she genuinely said elsewhere in the raw interview.

C. The earlier release does not cover generating new speech, so you need fresh, specific permission naming that use before anything is made.

D. Adding an on-screen AI-generated label resolves it, since the audience is then not being deceived.

41 · 9 min

What the model is actually reshaping

THE ONE THING TO KEEP

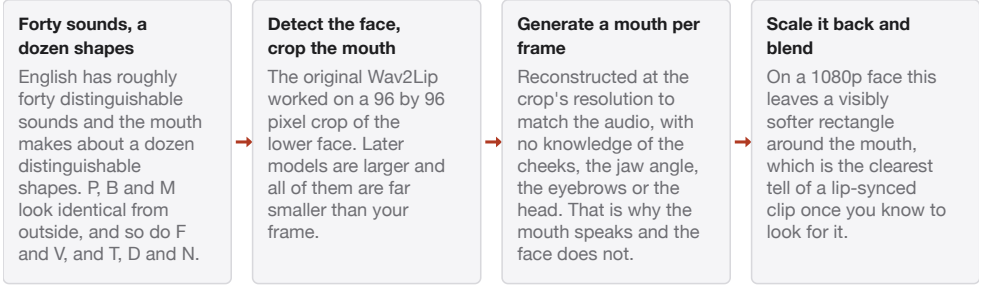
Audio-driven lip-sync maps sound to a small set of visually distinct mouth shapes and inpaints the mouth region frame by frame, which is why it works at medium shots and falls apart on plosives, profiles and close-ups.

Phonemes and visemes

Speech contains roughly forty distinguishable sounds in English. The mouth makes far fewer distinguishable shapes — around a dozen. Those shapes are called visemes, and several phonemes share one: **p**, **b** and **m** look identical from outside, as do **f** and **v**, and as do **t**, **d** and **n**.

A lip-sync model learns the mapping from audio features to those shapes, then reconstructs the mouth region of each frame to match. The Wav2Lip family and its successors work essentially this way: detect the face, crop the lower half, generate a mouth that matches the audio, blend it back.

What a lip-sync model actually reshapes



Everything these tools do well and badly follows from this description. A medium shot with clean, slightly slowed speech works. A close-up, a profile past about thirty degrees, or a run of plosives does not, and no setting changes that.

Everything about how these tools behave follows from that description.

Why it is only the mouth

The model changes a small rectangle and leaves the rest of the frame alone. That is why the technique is cheap, why it works on any footage including real footage, and why it produces a specific uncanny quality: the mouth is speaking and the face is not.

Real speech involves the cheeks, the jaw angle, the eyes narrowing on emphasis, the eyebrows, the head moving slightly on stressed syllables. A mouth-only sync gives you none of that, and viewers register the absence as "dubbed" without being able to say why.

Newer hosted products — Hedra, HeyGen, Synthesia and the sync features inside larger platforms — animate more: head motion, blinks, some eye direction. Better, and still a generated performance rather than a captured one.

The four places it breaks

Plosives. For **b**, **p** and **m** the lips must close completely. Many models under-close them, and the result reads exactly like badly dubbed foreign footage. Most viewers cannot name what is wrong; all of them notice. Slowing the delivery slightly improves it, because the closure gets more frames.

Profiles. The mouth has to be visible and roughly frontal. Past about thirty degrees off-axis the model is reconstructing geometry it cannot see, and the jaw slides. Keep any speaking shot near frontal.

Close-ups. The interior of the mouth — teeth, tongue, the dark space behind — is where these models are weakest, and a close-up shows all of it. Stay at a medium shot and most of the problem disappears. This is also why so much AI-presenter content is framed at chest height: it is not a style choice.

Long takes. Drift, as everywhere. The reconstructed mouth region slowly diverges from the surrounding skin in colour and texture. Cut every few seconds.

Resolution, and the reason for the soft square

Many open lip-sync models operate on a small crop — 96 by 96 pixels in the original Wav2Lip, larger in later ones. That crop is generated at low resolution and scaled back up into the frame.

On a 1080p face, that leaves a visibly softer rectangle around the mouth. It is the clearest tell of a lip-synced clip and you can see it on a great deal of published content once you know to look.

Mitigations: work at a shot size where the face is small enough that the crop covers it adequately; run a face restoration pass afterwards at low strength, accepting the flicker trade-off from the previous module; or use a higher-resolution model and pay for it.

Getting a better result from the audio side

Half of lip-sync quality is the audio you feed it, and this part is free:

- **Clean speech, no music, no background.** The model is extracting phonetic features; anything else is noise to it. Sync against the dry voice track and add music afterwards in the edit.
- **Consistent level.** Compress and normalise the voice first. Audacity does both in about a minute and runs on very old machines.
- **Slightly slower delivery.** More frames per phoneme, better closures, fewer smears.
- **Trim leading and trailing silence.** Some models produce odd mouth movement over silence.

The free path

Wav2Lip, LatentSync and their current successors are open and free, available as Hugging Face Spaces for browser use and as local installs. Audacity handles every audio preparation step above. FFmpeg splits and rejoins audio and video.

The paid products are better, particularly on head motion and resolution. They are not a prerequisite for learning any of this, and a medium shot with clean audio from a free model beats a close-up from a paid one.

Today: sync thirty seconds of clean speech to a medium shot and to a close-up of the same character, and compare where the eye goes in each.

BEFORE YOU MOVE ON

A lip-synced clip reads as badly dubbed even though the timing is correct. The character is saying a line with several b and p sounds. What is the most likely cause?

A. The model under-closes the lips on plosives, which require full closure, and incomplete closure is the classic signature of poor dubbing.
B. The audio was normalised too aggressively, so the transients that mark plosives were flattened and the model had no cue to close the lips on.
C. The viseme mapping is ambiguous for b, p and m, so the model selected the wrong mouth shape for those phonemes.
D. The frame rate is too low to represent the brief closure, and interpolating the clip to a higher frame rate before syncing would restore the missing frames of contact.

42 · 9 min

Acting is timing, and timing has to come from somewhere

THE ONE THING TO KEEP

Performance transfer drives a generated face from a recorded one, capturing the pauses, emphasis and eyebrow movement that carry meaning — which is why it beats lip-sync on anything with emotional content and why the person in front of the webcam matters more than the model.

What it does differently

Instead of deriving mouth shapes from audio, performance transfer derives the whole facial performance from a video of a person acting. You record yourself on a webcam delivering the line; the model maps your expression, head movement, blinks and eye direction onto a generated character.

Runway's Act-One is the best-known example of the class, and similar features exist elsewhere under other names.

The difference in output quality is not subtle, and the reason is not technical sophistication. It is that **delivery is most of acting**, and no phoneme model invents delivery.

What delivery actually consists of

Watch somebody say a sentence they mean. The mouth is doing the least interesting part.

- **Pauses.** Where somebody stops, and for how long. A quarter-second before a word changes what the word means.
- **Emphasis.** Which syllable gets weight, and the tiny head movement that accompanies it.
- **Eyebrows.** Raised on a question, drawn together on difficulty, lifted on surprise. These carry as much meaning as the words.
- **Blinks.** People blink at clause boundaries and at moments of thought. Regular mechanical blinking reads as wrong.
- **Eye direction.** Looking away while thinking, back on the point being made.
- **Micro-stillness.** Genuine listening is very still. Generated faces tend to fidget.

A lip-sync model produces none of these. A performance transfer produces all of them, because a person did them.

Which means the webcam is the important part

This is the unintuitive conclusion and the practically useful one. The quality of your talking-head shot depends mostly on how well the person in front of the webcam performs, and only secondarily on the model.

So:

- **Perform the line properly.** Read it aloud several times first. Know what it means. Decide where the pauses are.
- **Sit still and frontal**, well lit, camera at eye level. The model tracks your face, so give it a face to track: even light, no strong shadow across the features, no backlight.
- **Do several takes** and pick the one where the delivery is right. This is free — it is a webcam recording, not a generation.
- **Overplay very slightly.** Expression loses a little amplitude in transfer, as it does in animation generally.

If you cannot act, find somebody who can. A friend who does theatre will produce a better result than any amount of prompt engineering, and they will do it in ten minutes.

The limits, honestly

- **Proportions transfer, and they should not.** A driver with a long face drives the target toward a long face. Pick a driver whose head shape is roughly compatible, or expect the character to shift toward you.
- **Head rotation range is limited.** Large turns lose tracking. Keep within perhaps thirty degrees of frontal.
- **Occlusion breaks it.** Hands near the face, hair falling across it, glasses catching the light.
- **It is still a generated face.** Skin behaves slightly wrongly in motion, and long takes drift like everything else. Cut.
- **Duration limits apply** exactly as they do to any generation.

Where it fits in a workflow

A talking-head sequence that works usually looks like this:

1. Record the audio properly, as a separate clean track — a phone in a quiet room beats a webcam microphone every time.
2. Record the performance on the webcam, listening to or delivering the same line.

3. Transfer the performance onto the character.
4. Replace the webcam audio with the clean recording in the edit, aligned to the performance.
5. Cut every few seconds, using cutaways and reaction shots so no single generation has to run long.

Step four is the one people skip, and it is the difference between something that sounds like a video call and something that sounds produced.

The consent question is unchanged

Performance transfer onto a *generated* character is your own face driving a fictional one, which is straightforward. Performance transfer onto a *real person's* likeness is synthesising that person saying something they did not say, which requires their written, specific permission and a visible label on the output.

The technique does not change the rule and the quality of the result does not change it either. If anything, a better result raises the stakes, because a convincing synthetic performance is exactly what the disclosure obligations exist for.

The free path

Performance transfer is currently mostly a hosted, paid feature, and that is an honest limitation rather than something to talk around. The free approximations: open lip-sync plus a separately generated head-motion pass; pose and face-landmark conditioning in ComfyUI, which is fiddlier and does work; or the oldest answer, which is to film a real person and composite them, covered earlier in the course.

Recording the audio well, which is half the result, is free on any phone.

Today: record yourself saying one line three ways — flat, with a pause before the last word, and with an eyebrow. Watch the three back. That difference is what transfer captures and lip-sync cannot.

BEFORE YOU MOVE ON

Two teams produce a talking-head shot of the same generated character saying the same line. One uses audio-driven lip-sync with excellent audio; the other uses performance transfer from an amateur webcam recording. The second reads as more convincing. Why?

A. Performance transfer operates at higher resolution than lip-sync crops, so the second result has visibly better detail around the mouth.

B. Audio-driven models introduce a small latency between sound and mouth movement, and viewers are extremely sensitive to audio-visual desynchronisation.

C. Lip-sync reconstructs only the lower face, so the upper face remains at whatever expression the source frame had, which viewers interpret as emotional flatness rather than as a technical limitation of the crop.

D. The pauses, emphasis and eyebrow movement that carry a line's meaning came from a person, and audio alone contains no instruction for them.

43 · 9 min

Synthetic voice, and where it stops being good enough

THE ONE THING TO KEEP

Text-to-speech has solved pronunciation and timbre but not interpretation, so synthetic voice holds up for short informational lines and degrades over long-form narration where a listener starts noticing that nothing is being meant.

What is solved and what is not

Modern text-to-speech produces speech that is clear, correctly pronounced, and in a convincing timbre. Cloning a recognisable voice from a short sample is a solved problem and it is cheap.

What is not solved is interpretation: deciding what a sentence means and delivering it accordingly. A synthetic voice reading a paragraph applies plausible-sounding prosody derived from the text's surface. A human reading the same paragraph applies prosody derived from understanding it.

Over one sentence the difference is invisible. Over three minutes it accumulates into a specific fatigue — the listener stops believing anyone is there. That is the honest boundary of the technology and it is not about audio quality.

The quality ladder, with the free rungs named

From most accessible upward:

- **Piper.** Small, fast, runs on a Raspberry Pi or an old laptop, entirely offline. Robotic by current standards and perfectly usable for short utterances, accessibility and prototyping. Free and open.
- **Kokoro.** A recent small open model with a notably good quality-to-size ratio. Runs locally and in a browser Space. Free.
- **Coqui XTTS.** Multilingual, supports voice cloning from a short sample, runs locally on modest hardware. Free and open; the company behind it wound down, which is a genuine caution about long-term maintenance rather than about the model's usefulness today.
- **Hosted services**, of which ElevenLabs is the best known. Clearly better on expressiveness and on long-form stability. Paid, with small free tiers.

For most of this readership, a free model plus careful text preparation gets within touching distance of a paid one on short lines. On a five-minute narration, the paid ones pull ahead noticeably.

Getting more out of any of them

The text is an instrument. Most people type a paragraph and press generate.

- **Punctuate for breath, not for grammar.** A comma inserts a short pause. A full stop inserts a longer one. Adding commas where you want the speaker to breathe works better than any setting.

- **Break long sentences.** Synthetic prosody degrades across long clauses. Two short sentences outperform one long one.
- **Spell out anything ambiguous.** Numbers, dates, acronyms, units. "2026" may be read as a number rather than a year; "twenty twenty-six" is not ambiguous. "Dr" is doctor or drive.
- **Generate line by line, not paragraph by paragraph.** You can then regenerate the one bad line rather than the whole take, and assemble in the editor with pauses of your own choosing.
- **Use the emotion or style controls** where the model has them, and test them rather than trusting their names.

Cloning, and the permission it requires

Cloning a voice needs seconds of audio. That is the technical fact and it is why the rule has to be firm rather than proportionate to effort.

A person's voice is theirs. Using it requires their written, specific permission: what it will say, where it will appear, for how long, and whether new sentences may be generated. Recording somebody for one purpose does not license synthesising them for another. This is the same principle as the likeness rule and it is stated again because the ease of cloning makes it easy to forget that anything was taken.

Voice is also increasingly legislated for directly — several US states' publicity-rights statutes name voice explicitly, and courts in a number of countries have treated a distinctive voice as protectable. Where you are and where your audience is both matter, and neither this lesson nor any lesson is legal advice.

For your own voice, cloning is straightforward and genuinely useful: record a clean sample once, then produce corrections and pickups without setting up a microphone again.

Languages and accents

A practical point for this audience. Multilingual models vary enormously in quality by language, and quality in a language's *majority* accent does not pre-

dict quality in others. Hindi and Hinglish output from a model trained mostly on English can be intelligible and audibly wrong in rhythm.

Test the specific language and accent you need before building a project on it. If the result is not good, a real speaker recorded on a phone is better, cheaper and faster than fighting the model.

Cleaning up whatever you generate

Synthetic speech benefits from the same treatment as recorded speech: trim the silences, normalise the level, apply gentle compression, and cut any low-frequency rumble. Audacity does all of it for free on any machine, and it takes two minutes.

The next lesson covers that chain properly, because it applies identically to a real recording.

Today: generate one line three times — as a paragraph, as a single sentence, and with commas inserted where you want breaths. Listen to all three.

BEFORE YOU MOVE ON

A three-minute synthetic narration is clear and correctly pronounced, but listeners describe it as tiring. What is the underlying cause?

A. The model applies prosody inferred from the surface of the text rather than from understanding it, so nothing in the delivery is being meant.

B. Synthetic speech lacks the low-frequency content of a real voice, and the absence of that body over a long listen produces fatigue.

C. Generating a long passage in one pass causes the voice's timbre to drift, and listeners register the drift as strain.

D. Pauses are inserted at punctuation rather than at clause boundaries determined by meaning, which gives the narration a rhythm that never matches the argument being made.

44 · 9 min

A phone, a quiet room, and four minutes in Audacity

THE ONE THING TO KEEP

Most of recorded voice quality comes from the room rather than the microphone, so a phone in soft furnishings beats an expensive microphone in a hard empty room, and a fixed four-step cleanup chain handles the rest.

The room is the equipment

Amateur recordings sound amateur mostly because of reflections. Sound bounces off hard parallel surfaces, arrives at the microphone a few milliseconds after the direct sound, and the result is the hollow, distant quality everybody recognises as "recorded in a room".

No processing removes it convincingly. Reverb is far easier to add than to take away, and every de-reverb tool trades artefacts for improvement.

So fix it at the source, for nothing:

- **Record in the softest room you have.** A bedroom with a bed, curtains and a wardrobe. Not a kitchen, not a bathroom, not an empty office.
- **Record close.** A hand's width from the mouth, slightly off-axis so breath does not hit the microphone directly. Proximity means the direct sound dominates the reflections, which is most of the battle.
- **Hang something soft behind you** if the wall is bare. A duvet over a door, a coat on a hook. This is not a joke; it is what a blanket fort does and it works.
- **Turn things off.** Fan, air conditioner, fridge, phone notifications. Listen for thirty seconds before you record and you will hear three things you had stopped noticing.

The phone is fine

A modern phone's microphone is genuinely good, and its voice recorder app records at a quality that survives a video edit. Use the native voice memo app rather than recording through the camera, because the camera microphone is optimised for a scene rather than for a voice at close range.

If you have a wired earphone set with an inline microphone, that is often better again, because it sits close to the mouth and away from handling noise.

A USB microphone improves things. It improves them less than moving to a softer room does.

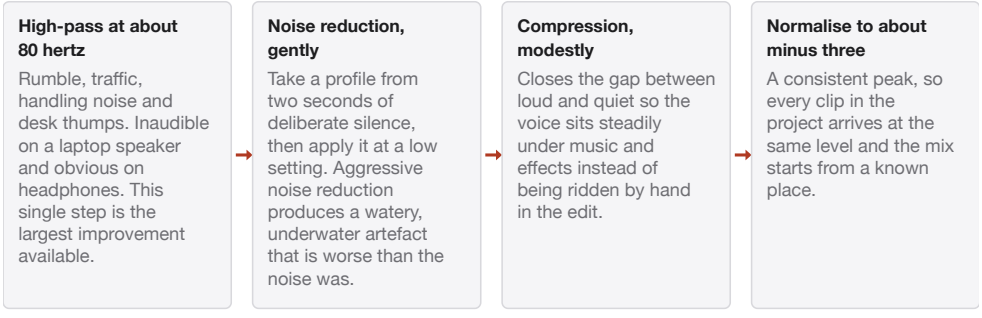
The four-step chain

In Audacity, which is free, open, and runs on machines that will not run anything else. Do these in order.

1. **High-pass filter at around 80Hz.** Removes rumble, traffic, handling noise and desk thumps that you cannot hear on a laptop speaker and that will be audible on headphones. This single step is the largest improvement available.
2. **Noise reduction.** Select two seconds of silence, take a noise profile, then apply it to the whole file at a gentle setting. Aggressive noise reduction produces a watery, underwater artefact that is worse than the noise was.
3. **Compression.** Reduces the gap between loud and quiet, so the voice sits steadily under music and sound effects. Audacity's compressor with modest settings is enough.
4. **Normalise** to a consistent peak, around -3dB, so every clip in the project is at the same level.

Four steps, about two minutes once you know where the menus are. Save it as a macro in Audacity and it becomes one click.

Four steps in Audacity, and what each removes



About two minutes once you know where the menus are, and one click after you save it as a macro. The room still matters more than the whole chain: a phone in a bedroom with curtains and a bed beats a good microphone in a hard empty office.

Record more than you need

Three things to capture every time, because going back is expensive:

- **Thirty seconds of room tone.** Silence in the actual room, recorded deliberately. You need this to fill gaps and to smooth edits, and it is the subject of the next lesson.
- **Three takes of every line.** Storage is free; a second session is not.
- **A handclap at the start** if you are recording audio separately from a video. It gives you a visible spike and an audible transient to align against, which is what a clapperboard has always been for.

Editing voice

Cut on the breath, not on the word. A cut placed in the middle of a breath sounds natural; a cut immediately after a word ends sounds clipped. Leave a little air.

Remove long pauses and keep short ones. Perfectly even pacing sounds synthetic, which is an interesting reversal — the thing that makes recorded speech sound human is precisely the irregularity that a text-to-speech model struggles to produce.

Do not remove every breath. A narration with no breathing in it is unsettling in a way listeners cannot identify.

Why this beats synthesis for anything long

Stated plainly, because it runs against the direction of the rest of the course: for narration longer than a few sentences, a real voice recorded on a phone in a soft room and cleaned in Audacity is better than almost any synthetic voice you can access for free, and better than several you would pay for.

It is also faster. Ten minutes of recording plus two minutes of processing against an hour of regenerating lines that are nearly right.

Use synthesis for what it is good at — short utterances, many languages, pickups, placeholder tracks while you edit — and record the narration.

Today: record the same paragraph in your kitchen and in your softest room, run both through the four-step chain, and listen on headphones.

BEFORE YOU MOVE ON

Two recordings of the same line: one on a phone in a curtained bedroom, one on a good USB microphone in a bare-walled office. The phone recording sounds better. Why?

A. Phone microphones apply voice-optimised processing that flatters speech, while a plain USB microphone captures the signal without enhancement.

B. Reflections from hard parallel surfaces arrive milliseconds after the direct sound, and no microphone quality compensates for a room that does that.

C. USB microphones are typically more sensitive, so they pick up more background noise from the office environment than a phone would.

D. The phone was held closer to the mouth, and proximity raises the level of the direct signal, which is the single factor that determines how present a voice sounds in a recording.

45 · 8 min

Why silence is the worst sound

THE ONE THING TO KEEP

Generated clips arrive with no sound at all, and true digital silence reads as broken, so a continuous ambience bed under every shot is the single cheapest thing that makes generated footage read as filmed.

What digital silence does to a viewer

Real life has no silence. There is always something — air movement, distant traffic, the building, the refrigerator, your own ears. A recording with no sound at all is an experience nobody has ever had outside a laboratory.

So when a video's audio drops to absolute zero, the viewer's reaction is not "how peaceful". It is that something has broken. Even in a cut between two shots, a moment of true silence reads as a fault.

Generated video arrives with no audio whatsoever. Every clip is that fault, until you fix it.

The ambience bed

The fix is one continuous audio track running underneath the whole piece, unbroken across every cut. Not per-shot sound — one bed.

This does two things at once. It removes the silence, and because it does not change at the cut points, it glues shots together. A sequence of unrelated generated clips over one continuous room tone is perceived as one place. That is an enormous amount of cohesion for one track.

Where to get it:

- **Record it yourself.** Thirty seconds of the actual room, or a street, or a field, on a phone. Free, specific, and better than anything generic.

- **Freesound**, which is a large community library under various Creative Commons licences. Check the licence on each file; some require attribution and some do not permit commercial use.
- **The BBC sound effects archive**, which is extensive and available under its own terms — read them, because they are not a blanket free-for-commercial-use licence.
- **Your own phone held out of a window.** Genuinely the most used technique in low-budget production.

Loop it, crossfade the loop point, and lay it under everything at a low level — quiet enough that you notice it only when it stops.

Spot effects

On top of the bed go the specific sounds: footsteps, a door, a cup set down, a car passing, cloth movement, a keyboard.

The rule that matters is **sync**. A sound that lands on the frame of the action is invisible and convincing. A sound two frames late is noticed. Zoom in on the waveform and place it against the picture, rather than dropping it approximately.

You need fewer than you think. Three or four well-placed effects in a ten-second shot is a lot. A wall of sound effects reads as a game rather than as footage.

Recording your own foley

The traditional answer and still the best one for anything specific. Foley is performed sound: somebody making the noise in sync with the picture.

A phone and household objects cover most of it. Footsteps recorded on the actual surface. Cloth movement by rubbing a sleeve near the microphone. A cup on a table is a cup on a table. Rain on a window is rice on a sheet of paper, which is a hundred-year-old trick that still works.

Recording it yourself takes twenty minutes, sounds specific rather than stock, and has no licence attached to it at all — which, on a commercial job, is worth

more than the twenty minutes.

Perspective

Sound has distance. A wide shot should sound further away than a close-up: quieter, with more reflection and less high frequency. Matching audio perspective to shot size is a small adjustment — a level change and a slight high-frequency roll-off — and it is one of the things that separates work that sounds designed from work that sounds assembled.

The same principle applies across a cut. If you cut from inside a room to outside it, the ambience should change at the cut. If it does not, the cut reads as a glitch rather than as a move.

The level nobody sets

A practical figure worth having. Dialogue sits around -12dB to -6dB peak. Ambience sits perhaps 20 to 30dB below it — low enough that it is felt rather than heard. Music under dialogue sits lower than beginners set it, usually around -25dB.

If you cannot hear the ambience at all on a laptop speaker, that is probably correct. Check on headphones.

The free path

Audacity for everything, Freesound and your own phone for material, and whatever editor you are using for placement. Resolve's free tier includes a full audio page. Kdenlive and Shotcut handle multiple audio tracks. CapCut on a phone will layer audio and is enough for short-form.

Nothing on this page costs money, and it is the largest perceived-quality improvement available anywhere in the pipeline.

Today: record thirty seconds of the room you are in, loop it under three generated clips, and watch them with and without it.

BEFORE YOU MOVE ON

A sequence of six unrelated generated clips is cut together. Adding one continuous ambience track underneath makes the sequence feel like a single place. Why does one track achieve that?

A. The ambience does not change at the cut points, so the audience infers that the shots share an acoustic space even though the pictures do not match.
B. Continuous audio masks the visual discontinuities at each cut by occupying attention that would otherwise notice them.
C. The track provides a consistent noise floor, which normalises the perceived audio quality of clips generated at different times.
D. Removing digital silence eliminates the impression that the file is faulty, and once that impression is gone viewers judge the sequence on its picture continuity alone.

46 · 9 min

Music you could defend if somebody asked

THE ONE THING TO KEEP

Music is the easiest part of a soundtrack to obtain and the easiest to get legally wrong, because a licence has to cover the specific use, and AI-generated music sits on unresolved ground about ownership and training data.

Four sources, with what each actually permits

Commercially licensed libraries. Artlist, Epidemic Sound, Musicbed and similar. You pay a subscription and get a licence covering specified uses. Read what it covers: many subscriptions license content published while you are subscribed and continue to cover it afterwards, some do not, and most treat broadcast and paid advertising differently from social posting. This is the safe route for client work and it costs money.

Creative Commons. Free Music Archive, ccMixter, parts of Freesound, and a great deal of independent work. Free, and the licence terms vary per track and matter:

- **CCo** — no conditions at all.
- **CC BY** — attribution required, which means actually crediting the artist in the description or on screen.
- **CC BY-NC** — non-commercial only. A client video is commercial. A monetised channel is commercial. This catches people constantly.
- **CC BY-SA** — share-alike, which can carry obligations onto your finished work.

Always check the specific track's licence on the page you downloaded it from, not the site's general description.

AI-generated music. Suno, Udio and their competitors, plus open models. Fast, specific to your piece, and legally unsettled — see below.

Music you make. Free open tools genuinely cover this: **LMMS** for composition, **Audacity** for arrangement of loops, **Ardour** for a full multitrack session. If you play anything, a phone recording of it has no licence questions at all.

The unsettled ground, described rather than resolved

AI music generation sits in the middle of a live dispute and it would be dishonest to tell you otherwise.

The disagreement has two parts. **Training**: whether using copyrighted recordings to train a generative model requires a licence. Major labels have brought actions against AI music companies; some have been settled or restructured into licensing arrangements, others continue. The answer differs by jurisdiction and no jurisdiction has produced a settled body of law.

Output ownership: whether what the model produces is protectable, and who by. In some countries a work requires human authorship to attract copyright, which puts a purely prompted output on uncertain footing. Platform terms of service often purport to assign output rights to the user, which is a

contract between you and that company, and is not the same thing as copyright existing.

What you can do about it practically:

- Read the specific service's terms for what it says about commercial use, and keep a copy of what it said on the day you generated.
- Prefer models and services that state their training sources, where any do.
- On client work, say in writing which music is AI-generated, so the client is making the decision with you rather than discovering it later.
- Where the stakes are high — broadcast, a paid campaign, anything with real exposure — licensed library music is the boring choice and it is the right one.

This is a description of a disagreement, not advice about your situation, and the position in your country may differ from all of the above.

Platform detection is a separate problem

Independently of whether you have a valid licence, automated content-identification systems on the large platforms may flag your upload. A legitimately licensed track can be claimed because a rights database contains it; an AI-generated track can be flagged for resembling something.

Keep the licence documentation — the receipt, the licence page, the track page — filed with the project. A dispute is resolved by producing it, and the process is much easier when the file is in the project folder rather than in an inbox from eight months ago.

Using music well, which is mostly using less of it

- **Start it late.** Music from frame one is a common amateur signature. Let the piece begin with ambience and bring music in at the first real beat of the story.
- **Duck it under speech.** Lower the music where anyone talks. A keyframed level change is enough; automatic ducking exists in Resolve's free tier.

- **Cut picture to the music or music to the picture, and decide which.** Doing neither produces the characteristic feeling that the two are unrelated.
- **Silence is a tool once the bed exists.** Pulling the music out at a moment is a strong effect, and it is available to you precisely because the ambience is still there underneath.

Today: find the licence text for whatever music you last used and read the commercial-use clause. Then file it in the project folder.

BEFORE YOU MOVE ON

A freelancer uses a CC BY-NC track in a video for a paying client, crediting the artist in the description. What is the problem?

A. Attribution in a description is insufficient for CC BY-NC, which requires an on-screen credit within the work itself.

B. Creative Commons licences cannot be applied to sound recordings, only to compositions, so the track's licence does not cover the recording used.

C. The share-alike obligation in the licence would require the finished client video to be released under a compatible licence, which the client is unlikely to accept.

D. The NC clause excludes commercial use, and a video made for a paying client is commercial regardless of whether attribution was given.

47 · 8 min

The viewer believes the sound

THE ONE THING TO KEEP

When sound and picture disagree, perception leans on the sound, which is why a well-synced impact makes a physically wrong collision read as correct and why sound design is a repair tool as well as a finishing one.

Audio wins arguments with vision

This is not a metaphor. It is a measurable perceptual effect.

The McGurk effect is the famous demonstration: a video of a mouth saying one syllable dubbed with a different syllable is heard as a third thing that neither source contains. Vision and hearing are being combined, and the combination is involuntary — knowing about the effect does not switch it off.

Related findings run through the whole perception literature. Sound changes the perceived timing of visual events. Sound changes the perceived force of an impact. A single flash accompanied by two beeps is often perceived as two flashes.

For generative video this has a direct, exploitable consequence: **a sound can repair a picture**.

Where this pays

Recall the list of things the model cannot do. Several of them are substantially fixable with audio.

Impacts and collisions. A glass hitting a floor and breaking into the wrong number of pieces is one of the model's weaknesses. Put a real, sharp, correctly-timed breaking-glass sound on the exact frame of contact and the shot reads. The eye had a moment of ambiguity and the ear resolved it.

Footsteps. Generated walking often has slightly wrong weight — the foot lands and the body does not respond correctly. Footsteps recorded on the right surface, placed on the contact frames, supply the weight the picture lacks.

Contact events generally. A hand touching an object, a door closing, a cup set down. These are exactly the moments the model handles worst, and they are also the moments that carry the clearest sounds.

Off-screen causes. The strongest version. A sound implies an event the picture never had to show. A car crash heard over a reaction shot is a car crash, and you generated nothing.

The technique, precisely

Three things and they are all about precision.

1. **Find the exact frame of contact.** Step through the clip frame by frame. The contact frame is usually one or two earlier than you think, because you are reacting to the frame where the change is obvious rather than where it began.
2. **Place the transient on that frame.** The transient is the sharp attack at the start of the sound, which is not necessarily the start of the file. Zoom into the waveform and align the spike, not the edge.
3. **Match the sound's character to the picture.** A light object needs a light sound. Nothing exposes a fake more than a heavy impact on a small event — and this is worth watching for, because stock libraries are full of exaggerated sounds designed for trailers.

Two frames late reads as wrong. On the frame, it disappears into the picture and does its work.

What sound cannot fix

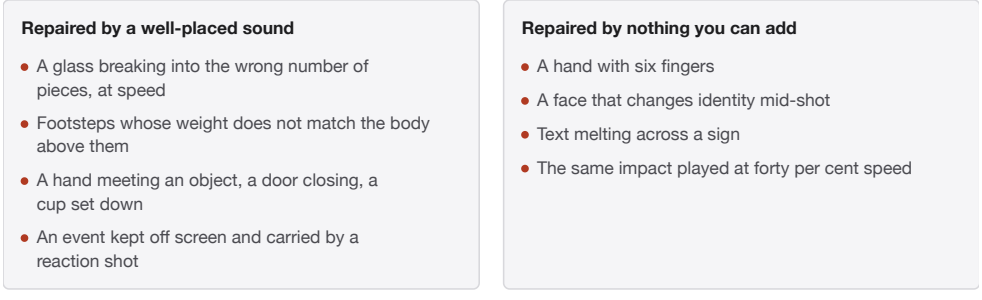
Honesty about the limit. Sound resolves ambiguity; it does not overwrite something clearly seen.

- A hand with six fingers is not fixed by any sound.

- A face that changes identity mid-shot is not fixed by any sound.
- Text that melts is not fixed by any sound.
- Slow motion defeats the whole technique. The audience has been given time to look, the ambiguity is resolved visually, and no audio cue competes with that. This is the mechanism behind the earlier advice never to put fake physics in slow motion.

The rule: sound repairs fast, ambiguous, brief events. It does nothing for sustained, clearly visible errors.

What a sound on the right frame repairs



Sound resolves ambiguity; it does not overwrite something clearly seen. Slow motion removes the ambiguity by giving the viewer time to check, which is the mechanism behind the rule never to put generated physics in slow motion.

Designing the cut around this

Once you know sound carries impacts, you can build shots that rely on it:

- Cut away *at* the contact and let the sound continue over the next shot. The audience assembles an event they never saw. This is the oldest trick in the medium and it works perfectly here.
- Put difficult physics off-screen entirely and cover it with sound plus a reaction.
- Let a sound start before its picture — an L-cut — so the next shot arrives already explained.

These are standard editing moves. They are worth restating here because in generative work they are not merely stylish, they are how you get shots you could not otherwise generate.

The free path

Everything: Audacity to trim and shape, Freesound or your own phone for material, and any editor with frame-accurate placement. Resolve's free tier shows waveforms at frame level. Kdenlive and Shotcut do too.

The skill is precision, and precision costs attention rather than money.

Today: take a generated clip with a contact event that looks slightly wrong. Find the contact frame, place one real sound on it, and watch it again.

BEFORE YOU MOVE ON

A generated shot of a mug being set down on a table looks slightly wrong — the contact reads as weightless. Adding a recorded mug-on-wood sound on the contact frame fixes the impression. Why does this work?

A. The sound draws attention away from the contact moment, so the viewer's eye is elsewhere when the error occurs.

B. The added transient masks the frame-to-frame inconsistency in the generated contact, because an abrupt sound reduces temporal acuity in vision for a short window afterwards.

C. The sound establishes the material of both objects, and once a viewer knows what the objects are made of they interpret the motion as consistent with that weight.

D. Audio and visual information are combined involuntarily, and where the picture is briefly ambiguous the sound determines what is perceived.

48 · 8 min

When the model makes the sound too

THE ONE THING TO KEEP

Native audio generation produces synchronised ambience and dialogue in one pass, which is remarkable and usually unusable as a final track, because you cannot separate, replace or adjust anything in it afterwards.

What is now possible

Several flagship models generate a soundtrack along with the picture: ambience matched to the scene, footsteps roughly in sync, and in some cases spoken dialogue with lip movement that matches. Google's Veo line is the most prominent example of the capability.

This is a genuine technical achievement and it is genuinely useful. It is also, for most production work, a draft rather than a deliverable, and the reason is not quality.

The reason is that it is baked

The audio arrives as a single mixed track welded to the picture. That means:

- **You cannot adjust the balance.** If the ambience is too loud under the dialogue, it is too loud permanently.
- **You cannot replace one element.** A wrong footstep sound cannot be swapped without replacing everything.
- **You cannot extend it across a cut.** Each clip has its own audio, which stops at the clip boundary — which is exactly the digital silence problem from two lessons ago, arriving at every edit point.
- **You cannot match it to the next shot.** Two generated clips have two unrelated ambiences. Cutting between them produces an audible jump at every cut, which is more noticeable than a visual mismatch.

- **You cannot localise it.** A dialogue line in one language is not translatable without regenerating the shot.

A mixed track is the end of an audio process, not the middle of one. Receiving it at the start of your edit means you have inherited somebody else's final mix for a film that does not exist yet.

Where it is the right answer

Be specific rather than dismissive, because there are real cases:

- **Single-clip social content.** One shot, no cuts, posted as is. No continuity problem exists, so the objections above do not apply.
- **Previsualisation and pitching.** A moving storyboard with sound communicates far better than one without, and nobody is grading it.
- **A reference track.** Generate with audio, then use that audio as a guide for what to build properly — it tells you what the model thought the scene sounded like, which is sometimes a good idea you would not have had.
- **Speed over control.** A deadline where a finished-enough clip now beats a better clip tomorrow.

The trade-off with the first frame

The awkward part, and it is worth knowing before you plan a shoot.

On several models the native audio path is tied to text prompting rather than to a supplied first frame. If you want the model's audio, you may have to give up image-to-video — which means giving up the still, the composite, the repaired hand, the character sheet and most of what this course has built.

That is a large price for a soundtrack you will probably replace. Decide per shot, and be clear that it is a trade rather than a free addition.

Extracting and rebuilding

If you do generate with audio and want to work properly afterwards:

```
# separate the audio from a generated clip
ffmpeg -i shot.mp4 -vn -acodec pcm_s16le shot_audio.wav
# strip audio from the picture for a clean edit
ffmpeg -i shot.mp4 -an -c:v copy shot_silent.mp4
```

Then treat the extracted audio as source material: pull out the one ambience you liked, loop it as a bed, and build the rest in the normal way. Source separation tools — Demucs and similar, free and open — will split a mixed track into rough stems, which sometimes rescues a good element from a bad mix.

Where this is going, without predicting it

The obvious improvement would be generated audio delivered as separate stems: dialogue, ambience, effects, each on its own track. That is what a production actually needs, and it is a reasonable thing to hope for rather than a thing to plan around.

Until it exists, the working assumption stands: generate the picture, build the sound. It is more work, it is entirely free, and it is the half of the job where a beginner can match a professional result with a phone and an afternoon.

Today: generate one clip with native audio, extract the audio with FFmpeg, and see whether any single element in it is worth keeping as a bed.

BEFORE YOU MOVE ON

Why is a model's natively generated soundtrack usually a draft rather than a deliverable for a multi-shot piece?

A. Generated audio is produced at a lower sample rate than the picture's frame rate requires, so it drifts out of sync over a long timeline.

B. Native audio is generated from the text prompt rather than from the picture, so it describes the scene the prompt asked for rather than the scene that was produced.

C. Licensing terms for generated audio differ from those for generated video on most platforms, so the soundtrack cannot be used commercially even where the picture can.

D. It arrives as one mixed track per clip, so nothing can be rebalanced or replaced and each cut produces an audible jump between unrelated ambiences.

CASE STUDY · MODULE 5

The retired doctor's voice

A small public-health communication unit attached to a district hospital in Solapur, producing a six-part Marathi video series on childhood immunisation for waiting-room screens.

The unit had forty minutes of archival footage of Dr Kulkarni, a paediatrician who had worked at the hospital for thirty-one years and retired in 2024. The footage was from an old awareness campaign. In the district he is close to a household name, and the unit's lead, Anagha, knew that a message in his voice would be listened to in a way that a message in hers would not.

The new series needed thirty-one new sentences he had never said. The archival footage contained none of them.

The technical question was trivial. Cloning a recognisable voice from a short clean sample is a solved problem and it is cheap; forty minutes is a great deal more than anyone needs. A colleague had a working clone in an afternoon, and it was good enough that Anagha could not reliably tell it from the original on a laptop speaker.

The question that mattered was whether they were allowed to, and here the unit disagreed with itself for a week.

One argument was that Dr Kulkarni had signed a release for the original campaign. He had agreed to appear, on camera, speaking about immunisation, for the hospital. This was the same hospital, the same subject, and a message he had spent his career delivering.

Anagha's position was that consent to appear is not consent to be synthesised. He agreed to say particular sentences. He did not agree to have new sentences generated in his voice, and the two are different acts, one of which did not exist as a practical possibility when he signed. Recording somebody for one purpose does not license synthesising them for another.

The cost of her position was not abstract. Dr Kulkarni is elderly and lives with his daughter in Pune. Asking meant a delay of at least two weeks on a series

with a district health office deadline. It also meant he might say no, in which case the unit would lose the single strongest asset it had, and the alternative — Anagha reading the lines herself — would be a measurably less persuasive series.

The cost of the other position was that if he objected later, the hospital would be the institution that generated a retired doctor's voice without asking, on medical advice, in a district where people trust him. No amount of production value survives that.

They asked. The letter said what would be generated, that the sentences would be new ones he had not spoken, which platforms it would appear on, for how long, and that the voice model would be deleted at the end of the campaign.

His reply distinguished two things the unit had been treating as one. He agreed to his voice. He refused any use of his face — he did not want to be shown saying things he had not said, and he was clear that a voice-over is a voice-over while a moving face is a claim about a person being somewhere. He also asked that the videos say plainly that the narration was synthesised.

That last request turned out to be the easiest to honour and the one the unit had been least prepared for.

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 series was built with generated illustrative footage, no synthetic face anywhere, Dr Kulkarni's cloned voice on the narration, and a Marathi title card at the head of every episode stating in plain words that the voice was synthesised with his permission. The card is burned into the frame rather than left in metadata, because a screenshot destroys metadata and several platforms strip it on re-encode. The delay was sixteen days. The signed letter, the deletion commitment and its date live in a consent folder beside the project. One

episode was later clipped and reposted by a local news account; the label survived, because it was in the picture.

The doctor had signed a release for the original footage. Why did that not cover generating new sentences in his voice?

A release to appear covers the particular performance that was recorded: those words, that occasion, that purpose. Synthesising new speech is a different act — it produces statements the person never made and never reviewed, with their authority attached. Consent has to be specific about what will be generated, where it will appear, for how long, and whether new sentences may be produced. The ease of cloning does not change this; if anything it makes the rule more important, because nothing about the process signals to anyone that something was taken.

Dr Kulkarni agreed to his voice and refused his face. Is that distinction coherent, and what does it imply for the production?

It is coherent and it is the distinction he explained: a narration is understood as a voice laid over pictures, while a moving face is a claim that a person was present and said something. A synthetic face carries a stronger implicit assertion and is harder for a viewer to discount. For the production it meant building the series around generated illustrative footage and graphics with no talking head at all, which is also the easier half of the pipeline to do well.

Why was the disclosure burned into the frame rather than carried in the file's metadata or in the post caption?

Metadata labels and invisible watermarks are worth keeping, but both are fragile: a screenshot destroys them and many platform re-encodes strip them. A caption belongs to the post rather than to the video, so a clip that is downloaded and reposted arrives with nothing. A title card in the picture travels with every copy of the file, which is exactly what happened when a news account reposted an episode. Several jurisdictions also now require a label a viewer can actually perceive, not only one a tool can read.

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 client filmed an employee's interview last year and now wants new sentences generated in that employee's voice. What does last year's release cover?

A. The recorded performance only; synthesis is a separate permission
B. Synthesis, provided the new sentences are on the same subject
C. Nothing, because a release older than twelve months lapses automatically
D. Any use of the recording, including speech derived from it

2 Audio-driven lip-sync holds up at a medium shot and falls apart on a close-up. What is the mechanism?

A. The crop is upscaled less at close range, so artefacts are hidden
B. Close-ups are always generated at a lower resolution than medium shots
C. Close-ups consume more of the model's attention budget
D. The mouth interior is weakest, and a close-up shows all of it

3 What does performance transfer capture that an audio-driven lip-sync cannot?

A. The character's identity across a longer take
B. A higher resolution around the mouth region than lip-sync models produce
C. The pauses, emphasis, eyebrows and blinks that carry the meaning
D. Correct pronunciation of unusual words

4 Why is one continuous ambience bed worth more than a set of well-chosen per-shot sounds?

A. It is louder than spot effects and masks the codec's compression artefacts
B. It does not change at the cuts, so unrelated clips read as one place
C. Platforms reject uploads whose audio track falls to zero
D. It removes the need to sync effects to contact frames

5 A generated glass shatters into the wrong number of pieces. When does a well-placed sound rescue the shot?

A. At speed, where the event is brief and ambiguous
B. Only if the sound is louder than the ambience bed
C. In slow motion, where the viewer has time to hear it properly
D. Always, because audio overrides vision in every case

6 A flagship model returns a clip with its own synchronised soundtrack. Why is that usually a draft rather than a deliverable?

A. The model's audio path only works on clips shorter than about four seconds
B. The audio is at a lower sample rate than platforms accept
C. It is one mixed track: nothing can be rebalanced or carried across a cut
D. Generated audio carries no licence, so it cannot be published

MODULE 6

From generations to a finished piece

Most of the difference between footage and output happens after the last generation. This block is the post-production half in order: conforming mixed frame rates onto one timeline, selecting and assembling, hiding cuts where the eye is already moving, pulling thirty independent colour responses into one world, laying grain that replaces the texture the encoder removed, adding real text, retiming without exposing fake physics, and exporting to a spec that survives a platform's re-encode — including on a machine that cannot run a professional editor.

> BY THE END OF THIS MODULE YOU CAN
>
> Take a folder of mixed-rate generations to a delivered file — conformed, assembled, cut on motion, colour-matched under one grade and one grain pass, titled, sound-finished and exported to a named platform specification — and name the free tool that performs each step on the hardware you actually have

49 · 10 min

The generator makes footage; the editor makes the video

THE ONE THING TO KEEP

Trim the settle and the drift, unify everything with one grade and one grain pass, and put real sound underneath — that is most of the difference between footage and output.

Nobody ships a raw generation

What separates work that looks made from work that looks generated is almost entirely the edit: pace, cut points, sound, colour continuity, and a ruthless discard rate.

Expect shooting ratios. A documentary crew shoots thirty minutes for one that airs. You will generate somewhere between four and ten clips for every one that ships. That is not failure, it is the normal shape of the job, and the sooner you budget for it the sooner you stop taking each bad generation personally.

The same thirty generations, before and after an afternoon

Straight out of the generator	After trim, grade, grain and sound
Every clip carries its quarter-second settle and its drifting last second	Each shot ends one moment before it stopped working
Thirty independent decisions about white balance and contrast	One shot-match pass, then a single grade over the whole timeline
A smooth, denoised surface that reads as synthetic before anything else does	One grain layer, so every clip shares a surface it did not have alone
No audio at all, so every cut lands in true digital silence	A continuous ambience bed, unbroken across every cut

None of the right-hand column costs money, and most of it is subtraction. Expect to generate four to ten clips for every one that ships: that is the normal shape of the job rather than a sign that you are doing it badly.

Conform the frame rate before you import anything

Models output at different frame rates — commonly 24, 25 or 30 — often at 720p or 1080p, and sometimes at odd durations like 5.1 seconds. Drop mixed rates onto one timeline and the mismatched clips judder, because frames are being duplicated or dropped to fit.

Pick the project frame rate first: 24 for a film feel, 30 for social platforms. Set it before importing. In DaVinci Resolve, changing the project frame rate once media is on the timeline is awkward or blocked depending on version, and re-conforming later is a genuine nuisance.

Upscale at the end, once, on the finished cut — not per clip. Upscaling a 720p generation to 4K adds no information. It adds a sharpening signature, and doing it thirty times gives you thirty slightly different signatures.

The three moves that make generated clips read as footage

1. **Trim the head and the tail by default.** Drop roughly the first quarter-second and the last second of every clip. You are cutting off the settle and the drift, which is where a clip announces itself.
2. **Unify colour and texture.** Different generations drift warm or cool independently. One shot-match pass, then a single grade on an adjustment layer across the whole timeline, pulls everything into one world. Then put a light film grain over the entire sequence. Grain is astonishingly effective here, because it gives every clip the same surface texture and it hides the smooth, plastic, denoised quality that reads as synthetic more than any other single property.
3. **Sound carries the picture.** Real room tone, real footsteps, real traffic under a generated street and the brain accepts what it is seeing. Free sources: Freesound, the BBC sound effects archive under its own terms, and your own phone held out of a window. Clean it in Audacity. This is the cheapest quality upgrade in the entire pipeline and it is the one most people skip.

Cut on motion, and cut before the failure

A cut placed during movement is far less visible than a cut on a static frame — the eye is already tracking change. And you decide where the failure lands. If the hand goes wrong at six seconds, the shot is five seconds long and nobody will ever know there was a sixth.

The free toolchain, end to end

- **DaVinci Resolve**, free tier: a professional editor, colourist and audio suite in one application, free permanently, no watermark on output. Honest limits — the free version withholds some effects and noise reduction, has restrictions around certain high-resolution and codec outputs, and it is heavy software that wants a reasonable GPU and 16GB of RAM.
- **Kdenlive** and **Shotcut**: open source, much lighter, and they run happily on older Windows and Linux machines that Resolve will refuse. For cuts, titles and audio they are not worse. For colour work they are clearly behind.
- **CapCut** on a phone: the whole assembly is genuinely achievable on a phone, and for vertical short-form it is often faster than a desktop. Its free tier and watermarking on certain templates have moved around over time, so check before you deliver something commercial.
- **Audacity** for audio cleanup on any machine at all.

More depth on all of this in video-editing, and on sound design and motion graphics in motion-and-audio.

Keep four things per shot

For every clip that ships, save the prompt, the seed, the first frame and the model name in a plain text file next to the project. Six weeks later a client will ask for one small change to shot nine, and without those four things you are not editing, you are starting over.

Today: take three clips you already generated. Trim the head and tail off each, put a single grade and a light grain across all three, and lay any real ambient sound underneath. Compare that against the raw exports.

BEFORE YOU MOVE ON

An editor assembles twelve individually good generated clips and the result still reads as a slideshow of AI clips rather than a piece of film. Which single change does the most?

A. Upscale every clip to 4K before assembly so the sequence carries more perceived production value.
B. Put one grade and a light grain across the whole timeline and lay real sound under every shot, so the clips share a texture and a world.
C. Regenerate the four weakest clips until each one matches the quality of the best.
D. Generate longer clips so the sequence needs fewer cuts and the joins become less noticeable.

---

50 · 9 min

Decide the frame rate before you import anything

THE ONE THING TO KEEP

A timeline has one frame rate and one resolution, and every clip that does not match is resampled to fit — so the settings you choose before importing determine whether your footage judders for the rest of the project.

The ten minutes that save an afternoon

Before a single clip is imported, four decisions:

1. **Frame rate.** 24 for anything meant to read as film. 25 for PAL-region broadcast. 30 for most social platforms. 60 only if the content is fast and the platform supports it.
2. **Resolution.** Match your delivery, not your source. 1080p is almost always right; 4K only if somebody has asked for it and paid for it.
3. **Aspect ratio.** 16:9, 9:16, or 1:1. This was decided when you made the stills.
4. **Colour management.** Leave it off unless you know you need it. A simple Rec.709 project is correct for almost all of this work, and colour-managed workflows introduce problems that are only worth solving on a graded film.

In DaVinci Resolve these live in Project Settings, and changing the frame rate after media is on the timeline ranges from awkward to blocked depending on version. Set it first.

What happens when rates do not match

A 30fps clip on a 24fps timeline has to lose one frame in five. A 24fps clip on a 30fps timeline has to repeat one frame in four.

On a static shot nobody notices. On a slow pan it is a visible stutter — the motion advances, pauses, advances — and once you have seen it you cannot stop seeing it. The name for this is judder and it is the most common technical fault in amateur edits.

Generative video makes it likely, because models output at different native rates and you will be mixing several models in one project. Check every clip's rate on import.

```
# what am I actually dealing with
ffprobe -v error -select_streams v:0 \
  -show_entries stream=r_frame_rate,width,height,duration \
  -of default=noprint_wrappers=1 shot.mp4
```

Conforming properly

If a clip is at the wrong rate, do not let the timeline drop frames. Convert it deliberately, with one of three methods, in ascending order of quality:

- **Duplicate or drop frames.** What the timeline does by default. Fast, judders.
- **Frame blending.** Blends adjacent frames into the intermediate ones. Soft on fast motion, smooth on slow motion, and generally the safe default.
- **Optical flow.** Estimates motion and warps pixels. Best result, and it produces the occlusion-boundary halos from the interpolation lesson. Resolve's free tier includes it; set it per clip in the inspector.

```
# conform 30 to 24 with blending
ffmpeg -i in.mp4 -filter:v "minterpolate=fps=24:mi_mode=blend"
-c:a copy out.mp4
```

Choose per clip rather than globally. A static shot can be dropped frames with no cost; a slow pan needs flow.

Resolution and scaling

Mixing 720p and 1080p generations on one timeline is normal. What matters is where the scaling happens.

Scale up once, at the end, on the finished cut — not per clip. Upscaling thirty clips individually gives you thirty slightly different sharpening signatures, and the difference between them is visible at the cuts.

And upscaling adds no information. It invents plausible detail. A 720p generation upscaled to 4K is a 720p generation with a 4K file size and an opinion about what the missing pixels looked like. Deliver at the resolution you generated unless somebody has a specific reason.

Odd durations and proxies

Generated clips arrive at 5.1 or 4.8 seconds because latent frames divided by frame rate does not land round. Nothing is wrong; trim to what you need.

If your machine struggles with playback, generate proxies — low-resolution copies the editor uses for editing and swaps for the originals at export. Resolve does this automatically; FFmpeg does it in a loop:

```
for f in *.mp4; do ffmpeg -i "$f" -vf scale=640:-2 -c:v libx264
-crf 28 "proxy_$f"; done
```

This is the single most effective thing you can do to make an old laptop usable for editing, and it costs disk space rather than money.

The folder structure that survives

Set it up before you import, not when it hurts:

```
edit/
  project.drp            # or .kdenlive, .mlt
  footage/               # the generations, unmodified
  audio/
  graphics/              # titles, logos, the label card
  exports/
```

Never edit from your Downloads folder. Every editor stores paths to media, and a moved or renamed file becomes a missing clip, usually on the day of delivery.

Today: run ffprobe over every clip in your project and write down how many different frame rates you have. Then set the project rate and conform deliberately.

BEFORE YOU MOVE ON

A slow pan generated at 30fps is placed on a 24fps timeline and plays with a visible stutter. What is the correct fix?

A. Conform the clip deliberately with frame blending or optical flow, rather than letting the timeline drop one frame in five.
B. Change the project frame rate to 30 so the clip plays natively, and conform the 24fps clips instead since they are fewer.
C. Retime the clip to 80 per cent speed so its frames align with the 24fps timeline grid without any being discarded.
D. Re-export the source with a constant frame rate, because generated clips often carry a variable frame rate and it is the variability rather than the mismatch that produces the stutter.

51 · 8 min

Choosing takes before you start cutting

THE ONE THING TO KEEP

Selecting takes and assembling a rough cut are separate passes with different judgements, and collapsing them is why edits stall — you cannot decide whether a shot works in the sequence while you are still deciding whether the shot works.

Four passes, in order

Editors work in passes because each one asks a different question and mixing them produces paralysis.

1. **Selects.** For each shot, which take? Judge the take alone.
2. **Assembly.** Put the chosen takes in order, full length, no trimming. Does the sequence work?
3. **Rough cut.** Trim to length. Find the cut points.

4. **Fine cut.** Frame-level adjustment, then colour, then sound, then titles.

Beginners try to do all four at once and get stuck on shot three for an hour. The passes exist because each one's decisions are only makeable once the previous one is settled.

The selects pass

You generated four takes of shot nine. Watch all four, in full, before choosing.

What you are judging, in this order:

- **Did it do the thing the shot is for?** Not whether it is pretty. If shot nine exists to show her noticing the door, the take where she notices the door wins even if another is prettier.
- **Where does it break, and when?** Note the timecode. A take that fails at six seconds is fine if the shot is four seconds long.
- **Does it match its neighbours?** Light direction, wardrobe, screen direction.
- **Then aesthetics.**

Mark the chosen take and keep the rejects. A reject often becomes a cutaway, an insert, or a source for a repaired frame later.

Write the choice down next to the shot on your list. Do not rely on file names alone; `s09_take4` does not tell you why.

The stringout

Put every selected take end to end, untrimmed, in script order. This is a stringout and it is the most useful ugly thing in editing.

It will be far too long and it will feel wrong. That is correct. What it tells you, which nothing else does:

- Whether the sequence makes sense at all.
- Which shots are missing. You will find two or three, every time, and finding them now costs a generation rather than a re-edit.

- Which shots are unnecessary. You will find more of these than missing ones.
- Where the piece sags.

Watch it once, all the way through, without stopping to fix anything. Take notes on paper. Stopping to fix is the beginning of the hour on shot three.

Cutting the length down

From stringout to rough cut, the operations are almost all subtraction:

- **Trim the head and tail of every clip.** Roughly a quarter-second off the front for the settle, and a second off the back for the drift. This is not a stylistic choice; it is removing the parts where the generation announces itself.
- **Cut the shots that were doing nothing.** If removing a shot does not hurt, it was not working.
- **Shorten everything.** Almost every rough cut is improved by being ten per cent shorter. Commercial editing runs two to four seconds a shot; your instinct for shot length is probably long.

Pace is a pattern, not a speed

A sequence of shots all the same length reads as mechanical. Vary them: a long establishing shot, three short ones, a longer one to let something land.

The simplest useful pattern is to let the shot after a fast run breathe. Three quick cuts then a held shot gives the audience somewhere to arrive, and it is the difference between energy and noise.

In generative work this pattern has a practical bonus: short shots hide drift, and the one long shot can be the one you generated most carefully.

Keeping the decision record

At the end of the assembly, your shot list should carry, per shot: the chosen take, the reason, and the intended length. That document is what lets you an-

swer a client's note about shot nine three weeks later without rewatching everything.

It is also what tells you, honestly, how many generations the project actually took — the number your next quote depends on.

Today: build a stringout of everything you have generated for one project and watch it once, straight through, taking notes on paper without touching the timeline.

BEFORE YOU MOVE ON

What does a stringout — every selected take, untrimmed, in order — reveal that a partially trimmed rough cut does not?

A. The true final duration of the piece, which cannot be estimated once clips have been trimmed to different lengths.

B. Where each generation begins to drift, since the untrimmed tails are still present and can be compared across shots.

C. How well the shots colour-match, because viewing them at full length under identical conditions makes differences in grade and exposure easier to detect than brief trimmed fragments would.

D. Whether the sequence holds together and which shots are missing or unnecessary, before any trimming decisions have foreclosed the structure.

52 · 8 min

Cut on motion, and one frame before the failure

THE ONE THING TO KEEP

A cut placed while the eye is already tracking change is far less visible than one on a static frame, which gives you both a craft rule and a way to place every cut exactly where a generation stops working.

Why motion hides a cut

When the eye is tracking movement it is already processing change, and a discontinuity arrives as one more change among several. On a static frame, the cut is the only change and it is perceived as an event.

Editors have used this for a century — cut on the action, meaning cut during a movement rather than before or after it. A character begins to stand in shot A, and shot B picks them up mid-rise. The movement carries across the cut and the join becomes almost invisible.

The same rule protects you from mismatches. Two shots whose lighting differs slightly will cut together more happily on movement than on stillness, because the eye has something else to do.

Cutting on the action, practically

The technique: find the frame where the movement begins in shot A, then find the corresponding point in shot B, and overlap slightly. In practice, cut a few frames *after* the action starts in A, and pick up in B a few frames *later* than feels right. Movement reads as continuous even when the geometry does not quite match, and audiences forgive a great deal inside a moving cut.

In generative work you rarely have two shots of the same action, so the usable version is broader: **cut while something in the frame is moving**,

whether or not the movement continues. A cut at the moment a car passes, a door swings, a figure turns — all of them absorb the join.

Cutting before the failure

The rule specific to this medium, and it is the reason the edit rescues so much.

You know where each take breaks, because you noted it in the selects pass. The shot ends one moment before that. If the hand goes wrong at six seconds, the shot is five seconds long, and nobody will ever know there was a sixth second.

This is not damage control. It is what editors have always done with real footage — end the shot before it stops working — and it is why the shot list should carry an intended length that you are willing to shorten.

Two consequences worth planning around: generate longer than you need, and never plan a sequence where one shot has to carry an exact duration. Music-video work is the exception and it is harder for exactly this reason.

J-cuts and L-cuts

The two most useful tools for making a sequence feel connected rather than assembled.

- **J-cut**: the audio of shot B starts before its picture. You hear the next scene before you see it.
- **L-cut**: the audio of shot A continues over the picture of shot B. You see the next scene while still hearing the last.

Both work because the audio and picture cuts no longer land in the same frame, so there is no single moment where everything changes. They are the difference between a sequence that feels edited and one that feels stapled.

In generative video they do an extra job: because each generated clip has its own audio situation and its own ambience, overlapping audio across the cut disguises the fact that the shots were never in the same place.

Cuts that read as mistakes

Three to avoid:

- **The jump cut.** Two shots of the same subject at nearly the same size and angle. Unless it is a deliberate style, it reads as a fault. Change size by at least two steps or change the angle substantially.
- **Crossing the line.** Covered in the directing module. Enforce screen direction in the stills.
- **The cut on a static held frame with no motivation.** If nothing is moving and nothing has prompted a change, the cut arrives as arbitrary. Give the audience a reason: a look, a sound, a movement.

The transition question

Use cuts. A dissolve means a passage of time or a change of place; a fade means an ending. Everything else in the transitions menu is a 1998 wedding video.

There is one generative-specific exception worth knowing. If two shots genuinely will not cut together — mismatched light, a colour you could not fix — a very short dissolve of four to eight frames can blur the join enough to pass. It is a patch, it is visible if you look, and it is better than the cut it replaces. Use it sparingly and never as a default.

Today: take two shots that refused to cut together and move the cut point to a frame where something in the outgoing shot is moving. Compare.

BEFORE YOU MOVE ON

Two generated shots have slightly mismatched colour that you could not fully correct. Moving the cut point to a frame where a figure is turning makes the join much less noticeable. Why?

A. The eye is already processing change, so the discontinuity arrives as one change among several rather than as the only event in the frame.
B. Motion blur during the turn reduces the effective colour saturation in those frames, narrowing the gap between the two shots.
C. Cuts on movement are shorter in perceived duration, so the viewer has less time to register the two shots as separate images.
D. The turning figure occupies the centre of attention, and foveal attention on a moving subject suppresses peripheral colour processing where the mismatch is most visible.

53 · 10 min

Thirty renders, one world

THE ONE THING TO KEEP

Each generation makes its own independent decisions about colour and contrast, so a piece is unified by a two-stage process — shot matching to bring every clip to a neutral common ground, then one grade over all of it — and never by grading each clip to taste.

Why every clip is a different colour

Nothing carries colour decisions between generations. Two clips from the same still, the same prompt and the same model will differ slightly in white balance, contrast and saturation, because those properties emerge from the sample rather than being set by a parameter.

Across thirty clips generated over a week, that variation is large. On a timeline it reads as amateurism more reliably than almost any other fault, because audiences associate consistent colour with produced work.

Two stages, in this order

The mistake is to grade each shot until it looks nice. That produces thirty attractive shots that do not belong together.

Stage one: match. Bring every clip to a neutral common ground. Same black level, same white point, same rough contrast. Nobody sees this stage; its output looks flat and boring and correct.

Stage two: grade. One look, applied across everything at once, on an adjustment layer. This is where the piece gets its colour.

The reason for the order: a look applied to inconsistent clips amplifies the inconsistency, because a contrast curve pushes two different starting points further apart.

What sits above the clips on a finished timeline

Grain, one pass	A single layer over the whole sequence, set so you can only see it on a paused frame. Thirty separate grains is worse than none, because now the cuts have one more thing changing across them.
The grade, one layer	The look from your look file, on one adjustment layer across everything. Modest: a strong grade on generated footage draws attention to the footage.
Shot matching, per clip	Black level first, then white point, then mid-tones, then saturation, read off the waveform rather than off your eyes. Nobody sees this stage; its output is flat and correct.
The generations	Trimmed of the settle and the drift, conformed to one project frame rate, and otherwise untouched.

The order is the lesson. A look applied to clips that have not been matched amplifies the mismatch, because a contrast curve pushes two different starting points further apart. Grading each shot until it looks nice gives you thirty attractive shots that do not belong together.

Doing stage one

Pick a hero shot — the one closest to what you want, usually a well-exposed mid-shot — and match everything to it.

In DaVinci Resolve's free tier:

1. Put the hero shot and the shot you are matching side by side, using the split-screen view on the colour page.
2. Right-click the target clip and use **Shot Match** to that hero. It does most of the work in one click.
3. Fix what it got wrong by hand, in this order: **black level first**, then white point, then mid-tone brightness, then saturation.

Black level first matters. If the blacks do not sit at the same level, every subsequent adjustment is fighting a mismatch you have not fixed. Use the waveform scope: find the bottom of the trace in each clip and line them up.

Kdenlive and Shotcut have colour-balance and curve filters but no shot-match. The manual method works: set the scopes side by side, match blacks, match whites, match saturation. Slower, same result.

Reading scopes instead of your eyes

Your eyes adapt within seconds, which makes them useless for comparison. After thirty seconds of looking at a warm shot, a neutral one looks blue.

Scopes do not adapt:

- **Waveform** shows brightness across the frame. Use it for black level and white point.
- **Parade** shows red, green and blue separately. Use it for colour cast — if the blue trace sits higher at the bottom than the red, the shadows are blue.
- **Vectorscope** shows hue and saturation. Skin tones fall along a specific line on it, which is the most useful single fact in colour correction: if your skin tones sit on the skin line, the shot is close to right.

Resolve's free tier has all three. This is professional colour instrumentation at no cost, and most people using it never open them.

Stage two, the grade

One adjustment layer over the whole timeline, carrying the look from your look file. Typically:

- A contrast curve, usually lifting the blacks slightly so nothing is pure black.
- A colour shift — warm highlights and cool shadows is the most common, and the most over-used.
- A saturation adjustment, usually downward.

Keep it modest. A strong grade on generated footage draws attention to the footage, which is the opposite of what you want.

The secondary you will eventually need

When one element is wrong while the rest of the frame is right — a garment that came out a different blue in one shot — a global correction cannot help.

A secondary isolates a colour range or a shape and corrects only that. Resolve's free tier supports qualifiers and windows, including tracking a window across a moving subject.

It is fiddly, it is imprecise on anything with motion blur or soft edges, and it is sometimes the only thing between you and regenerating the shot. Budget an hour when you need one, and treat needing one as a signal that your still-checking pass failed.

Check on more than one screen

A laptop screen, a phone and, if possible, a television. Generated footage graded on a bright laptop routinely looks crushed and muddy on a phone in daylight, which is where most of your audience is.

If you can only check on one screen, check on the phone.

Today: open the waveform scope, put three of your clips on a timeline, and line up their black levels. Do nothing else and watch how much more like one

piece they look.

BEFORE YOU MOVE ON

An editor grades each generated clip individually until each looks attractive, then plays the sequence. It reads as inconsistent. What was the process error?

A. The clips should have been graded in sequence order rather than individually, so each could be matched to the one before it as the work progressed.

B. Individual grades were applied to the clips rather than to an adjustment layer, so they cannot be revised together as a group later.

C. Visual adaptation means the editor's eyes adjusted between clips, so each grade was made relative to the previous shot rather than to an absolute reference, compounding the differences across the timeline.

D. Grading to taste optimises each shot separately, so nothing brings the clips to a common ground before a shared look is applied on top.

54 · 8 min

Putting back the texture the encoder removed

THE ONE THING TO KEEP

Generated frames are smooth because the autoencoder discarded high-frequency detail, so a single grain pass across the finished timeline replaces a missing surface property and unifies clips more effectively than any colour operation.

Why grain works so well here

Every clip went through the same lossy encoder, which removed fine texture and returned a smooth, denoised surface. That smoothness is one of the strongest cues that something is synthetic — stronger than colour, stronger than motion, and largely unconscious.

Real captured images have noise. Film has grain, digital sensors have shot noise, and both are irregular, high-frequency, and present everywhere.

Adding grain does two things at once:

1. **It replaces a missing property**, so the image reads as captured rather than rendered.
2. **It gives every clip the same surface.** Thirty generations with one grain layer over them share a texture they did not have individually, which is a form of matching that colour work cannot achieve.

One pass, over everything

The rule that matters: apply grain once, to the finished timeline, on an adjustment layer above everything.

Applying it per clip gives you thirty grains with slightly different characteristics, and the difference is visible at every cut — which is worse than no grain at all, because now the cuts have an extra thing changing across them.

In Resolve: an adjustment clip on the top track spanning the whole timeline, with Film Grain on it. In Kdenlive or Shotcut: a grain filter on a track above, or a grain video file on a track above set to overlay or soft light blend at low opacity. In CapCut: a grain overlay applied to the whole project.

Getting the amount right

The two things people get wrong are amount and size.

Amount. You should not be able to see the grain when watching normally. You should see it if you pause and look. If the grain is noticeable in motion, halve it. A useful test: toggle the layer on and off — the footage should look slightly wrong with it off and unremarkable with it on.

Size. Grain has a spatial scale, and that scale must relate to the output resolution. Grain sized for 4K applied to a 1080p export looks like fine noise; grain sized for 1080p and then upscaled looks like blotches. Set it for the resolution you are delivering at, and if you change delivery resolution, revisit it.

There is also a delivery consideration. Grain is high-frequency detail, which is exactly what video compression removes first, so heavy grain plus a low-bitrate platform re-encode produces a smeary mess. Keep the grain light and the bitrate adequate, which the export lesson deals with.

Halation and bloom

Real lenses and real film do something else generated footage does not: bright areas spill slightly into their surroundings. Film halation gives a warm red glow around highlights; lens bloom gives a general soft spread.

A small amount of this sells a shot as photographed more than most people expect. In Resolve, the Glow effect at low intensity restricted to highlights. In any editor, duplicate the clip, blur the copy heavily, restrict it to bright areas with a luminance key, and blend it back at low opacity.

The same warning as grain: subtle. Visible bloom is a filter.

Defocus and vignette

Two more properties of real optics worth adding deliberately rather than asking the model for.

Defocus. A keyframed blur, masked, gives you the rack focus that the model performs badly. It is deterministic and adjustable, which the generated version is not.

Vignette. Real lenses darken slightly toward the corners. A very light vignette focuses attention and reads as lens character. A heavy one reads as an Instagram filter from 2012.

Denoise before regrain, sometimes

A counter-intuitive professional move worth knowing. If a generated clip has visible compression artefacts or inconsistent noise of its own, cleaning it first and then adding uniform grain produces a better result than adding grain on top of a mess.

Resolve's free tier does not include the full noise reduction tools, which is one of its real limitations. The free route is FFmpeg's `hqdn3d` or `nlmeans` filters:

```
ffmpeg -i in.mp4 -vf hqdn3d=1.5:1.5:6:6 -c:v libx264 -crf 18
clean.mp4
```

Then grain the clean version in the edit. This is worth doing only when a clip genuinely has a noise problem, not as a default.

Today: put a single grain layer over your whole timeline, set it until you can only see it on a paused frame, and toggle it on and off while watching.

BEFORE YOU MOVE ON

An editor applies a film grain effect individually to each of thirty generated clips, matching the settings carefully each time. What goes wrong?

A. Each clip's grain is generated independently, so its pattern differs from its neighbours and the change becomes one more thing moving across every cut.

B. Per-clip grain accumulates where clips overlap in transitions, producing visibly heavier texture at every dissolve.

C. Grain applied before the grade is altered by the subsequent contrast curve, so identical settings produce different visible amounts once the look is applied.

D. Rendering grain thirty times rather than once multiplies the processing load, so the export takes substantially longer without any improvement in the result.

55 · 8 min

Real type, added afterwards

THE ONE THING TO KEEP

Text belongs in the editor rather than in the generation, because a title is vector type that stays identical for every frame while generated letterforms are reinvented each frame from a latent that never held them.

The rule and its reason

Any text that has to be readable is added in post. Not because models are bad at text in some vague way, but because a letterform is a symbolic shape below the encoder's preserved resolution, reinvented each frame, and a letter that shifts by two pixels per frame reads as melting even when each individual frame is acceptable.

A title in your editor is rendered from a font. It is identical in every frame. There is no mechanism by which it can drift.

So: shop signs, packaging, screen content, subtitles, lower thirds, end cards. All of it goes on afterwards.

Titles on a generated background

Three practical points that make the difference between a title that sits in the shot and one that sits on top of it:

Contrast. Text must be readable against the busiest frame it covers, not against the frame you happened to pause on. Check the whole shot. If contrast fails anywhere, add a subtle shape behind the text, or a soft gradient, or darken that region of the picture slightly.

Motion. A locked-off title over a moving background reads as an overlay. Matching the title's movement to the shot's — a slight drift in the same direc-

tion, a scale that follows a push-in — integrates it. This is one line of keyframes and it is what separates broadcast titling from a caption.

Safe margins. Keep text within roughly ninety per cent of the frame, and further in on vertical video where interface elements cover the lower third. Check the actual platform.

Replacing text inside a shot

When the generated sign has to say something specific:

1. Generate the shot with the sign blank, or with any text, and note that you will replace it.
2. Track the sign's surface in the editor. Resolve's Fusion page does planar tracking in the free tier; Blender's tracker is free and capable.
3. Attach your real text to the track, corner-pinned to the surface.
4. Match the grade, then add grain to the text so it shares the surface texture of the rest of the frame.

Step four is the one that makes it work. Clean vector type over grained footage announces itself immediately.

Generate the plate with a trackable feature — an edge, a corner, a mark. A featureless surface cannot be tracked and you will discover this after paying for the clip.

Fonts, and the licence you did not read

Font licences are real and they catch freelancers regularly. A font free for personal use is not free for a client's advertisement.

Safe free sources:

- **Google Fonts.** Open licence, commercial use permitted, enormous range, and strong coverage of Indic scripts — Noto covers most writing systems, which matters if you are working in Hindi, Bengali, Tamil or Telugu.
- **The Open Font Library** and **Fontshare** for more distinctive faces.

Check the licence file that ships with any font you download. Keep a copy in the project folder, for the same reason you keep the music licence.

Subtitles, which are not optional any more

Most social video is watched without sound. Subtitles are not an accessibility afterthought; they are how the majority of your audience receives the words.

The free route: generate a transcript with **Whisper**, which is open, free, and runs locally or as a Hugging Face Space, then import the resulting SRT into your editor and style it. Resolve, Kdenlive and CapCut all import SRT.

Always read the transcript before you use it. Automatic transcription gets names, technical terms and accented speech wrong, and a confidently wrong subtitle is worse than none.

The label card

The piece of typography this course keeps returning to.

Several countries now require synthetic media to carry a label a viewer can actually see. Metadata labels and invisible watermarks are real and worth keeping in the file, and both are fragile — a screenshot destroys them and many platform re-encodes strip them. The label that survives contact with the internet is the one burned into the frame.

So put it in the video: a title card at the start, in frame, before a viewer has had time to believe anything. Legible, on screen long enough to read, not hidden in the corner at four pixels tall.

It costs you two seconds and a text layer. The module on law and disclosure deals with what the rules currently ask for and how much they differ by country; the craft point here is simply that it belongs in the titling pass, with everything else you want a viewer to read.

Today: build your label card as a reusable title in your editor and save it as a template. You will use it on everything.

BEFORE YOU MOVE ON

Why is a title added in the editor immune to the melting that affects generated on-screen text?

A. Editors render text at the delivery resolution rather than the model's native resolution, so the letterforms have enough pixels to stay legible.

B. Text layers are composited after the grade, so they are unaffected by the colour and contrast operations that destabilise generated detail.

C. The title is rendered from a font and is therefore bit-identical in every frame, while generated letterforms are re-sampled from a latent that never contained them.

D. Vector outlines are stored as mathematical curves rather than as pixels, so they survive the video codec's compression without the quantisation errors that degrade generated text between frames.

56 · 8 min

Slow motion is where fake physics goes to be caught

THE ONE THING TO KEEP

Slowing generated footage gives the viewer time to verify physics the model only approximated, so retiming should be used to speed things up and to find a moment, and almost never to dwell on a physical event.

Why slow motion exposes generated footage

A generated collision, splash or impact is plausible at speed and wrong on inspection. There is no simulation underneath — no mass, no conservation of anything — only frames that look reasonable in sequence.

At normal speed the viewer gets a fraction of a second and the sound carries the event, for the reasons in the previous module. Slow it down and you have

handed them time to check, removed the sound's power to resolve ambiguity, and directed their attention at exactly the frames the model was least sure about.

So the rule is firm: **do not put generated physics in slow motion.** If a shot has an impact, a splash, cloth settling, or anything breaking, play it at speed or faster.

What retiming is good for

Three legitimate uses.

Speeding up. Generated footage is often slightly too slow, because models default to gentle drifting motion. A clip at 110 or 120 per cent frequently reads better and no one notices. This is the most common retime in this kind of work.

Finding a moment. You generated ten seconds and the expression you want happens at 6.2 seconds. Slipping the clip — moving the content within a fixed timeline position — puts that moment where the cut needs it. This is not a speed change at all and it is the most useful retiming operation there is.

Hiding drift. A clip that degrades toward the end can be sped up in its final second so the bad frames pass quickly. Crude, and it works.

How the editor makes the frames

Three methods, and choosing wrongly is where retimed shots go wrong:

- **Nearest frame.** Repeats or drops frames. Stutters on slow-downs. Fine for small speed increases.
- **Frame blend.** Averages neighbouring frames. Soft ghosting, which is forgiving and reads as motion blur at modest ratios.
- **Optical flow.** Estimates motion and generates genuinely new intermediate frames. Best for smooth slow-downs, and it produces the occlusion halos discussed earlier — most visible exactly where a limb crosses a body, which is where a slowed generated clip is already under scrutiny.

A practical guide: up to 150 per cent speed, any method. Slowing to 50 per cent, use flow and inspect the result. Below 50 per cent on generated footage, do not.

Speed ramps

A ramp changes speed within the shot — fast, then slowing into a moment, then fast again. Done well it is invisible and gives weight to a beat. Done badly it is the single clearest marker of an amateur edit.

Two rules keep it honest:

- **Ramp on motion, not on stillness.** A speed change during movement is absorbed; on a static frame it looks like buffering.
- **Ease the change.** An instant jump from 200 per cent to 50 per cent reads as a glitch. Every editor lets you curve the speed change; use it.

And remember what slowing does to generated physics. A ramp that slows into an impact is exactly the shot this lesson says not to make.

Audio and retiming

Speed changes pitch-shift audio unless you tell the editor otherwise. For a small speed change on ambience nobody notices. For anything with speech, detach the audio before retiming and keep it at normal speed, or the voice goes up.

This is another argument for building sound separately: a soundtrack assembled from your own material is unaffected by whatever you do to the picture's speed.

Freeze frames

A freeze is a retime to zero and it is worth a note. Freezing a generated frame is safe — it is just a still — but the transition into and out of it exposes any drift immediately, because the eye suddenly has a fixed reference to compare the moving frames against.

If you freeze, freeze on a frame you have inspected, and consider adding a small push-in on the frozen image so it does not sit completely dead.

Today: take a generated clip with any physical event in it, slow it to 40 per cent, and watch what you can now see. Then put it back.

BEFORE YOU MOVE ON

A shot of a glass falling and breaking looks convincing at normal speed and obviously wrong at 40 per cent. What accounts for the difference?

A. Optical-flow interpolation introduces artefacts when generating the intermediate frames, and those artefacts are what reads as wrong rather than the original motion.

B. Slowing removes the sound's ability to land on the contact frame, and without the audio cue the visual event loses the weight that made it read as real.

C. Slowing gives the viewer time to check physics the model only approximated, and directs attention at the frames it was least certain about.

D. The model generates fewer distinct frames during a fast event than during a slow one, so a slowed clip is displaying repeated or near-identical frames that betray the lack of underlying simulation.

57 · 9 min

The last render, and what the platform does to it

THE ONE THING TO KEEP

Every platform re-encodes what you upload, so your export should be a high-bitrate master that gives their encoder good material to work from rather than a small file that has already thrown away the detail.

What happens after you upload

You upload a file. The platform transcodes it into several versions at different bitrates for different connections. Your viewers never see your file; they see the platform's re-encode of it.

That single fact determines the whole export strategy. You are not producing a file for viewers, you are producing source material for somebody else's encoder. Give it good material.

The corollary people get wrong: exporting small to "save upload time" throws away detail before the platform's encoder has seen it, and the platform then compresses your already-compressed file. Two lossy passes instead of one, and it shows in exactly the places generated footage is already weak — grain, fine texture, gradients in skies.

The settings that matter

- **Codec.** H.264 is universally compatible and correct for nearly everything. H.265 is more efficient and less universally accepted. Use H.264 unless you have a reason.
- **Bitrate.** The number that actually decides quality. For 1080p delivery, 15 to 25 Mbps. For 4K, 40 to 60. These are higher than most default presets and they are the single most effective change you can make.

- **Profile and level.** High profile. Leave the level automatic.
- **Audio.** AAC at 320kbps stereo. Audio is a small share of the file and there is no reason to economise.
- **Frame rate.** Whatever the project is. Do not change it at export.
- **Colour.** Rec.709. Do not export HDR unless somebody specifically asked and can play it back.

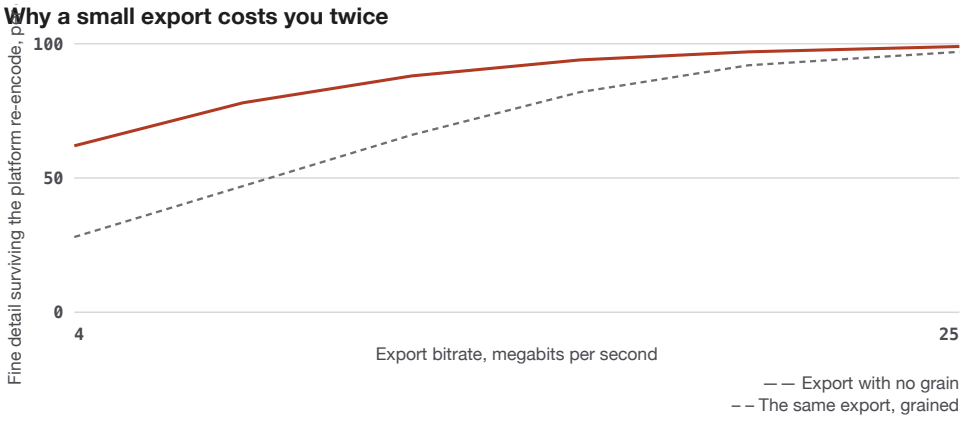
```
# a reliable 1080p master
ffmpeg -i timeline.mov -c:v libx264 -preset slow -crf 16 \
  -pix_fmt yuv420p -c:a aac -b:a 320k master.mp4
```

`yuv420p` is not optional for wide compatibility. A file exported in 4:4:4 or 4:2:2 will refuse to play on a surprising number of devices and platforms, and it is a common and confusing delivery failure.

Grain and bitrate interact

Worth its own note because it bites people who just learned about grain. Grain is high-frequency detail and it is the first thing a codec discards. Heavy grain at a low bitrate produces a smeared, blocky mess where the grain used to be.

So: keep the grain light, keep the bitrate up, and check your export at full size before delivering. If the grain has turned to mush, either raise the bitrate or reduce the grain.



You are not making a file for viewers, you are making source material for the platform's encoder, and grain is the first thing any codec discards. Export a 1080p master at 15 to 25 megabits, keep the grain light, and check the busiest shot at full size before you deliver.

Platform specifications

Vertical platforms want 1080×1920, 30fps, and they overlay interface on the lower portion. YouTube accepts almost anything and prefers higher bitrates than its own guidance suggests. Broadcast wants exactly what the broadcaster's spec document says and will reject anything else.

Two habits:

- **Read the current spec before delivering**, because they change and a two-year-old blog post is not a spec.
- **Export a master first**, at high bitrate in your project's native format, and derive platform versions from it. Never export a platform version as your only copy — it is already compressed for somebody else's pipeline and it cannot be un-compressed.

Two exports, always

Keep a master and deliver a copy. The master is high bitrate, full resolution, no platform-specific cropping, no burned-in subtitles. It is what you go back to when the client wants a square version or a different edit.

File it with the project, with the date in the name.

The delivery checklist

Before anything leaves your machine:

- Every clip trimmed of its settle and its drift.
- One grade and one grain pass across the whole timeline, not per clip.
- Frame rate conformed; upscaled once at the end, if at all.
- Real sound under every shot, with no moment of true digital silence.
- Levels checked: dialogue around -12dB to -6dB peak, music well below it.
- A visible label on generated content, at the start, in frame.
- Written consent on file for any real face or voice, specific to this use.
- Licences filed for every piece of music and every font.
- The prompt, seed, first frame and model version saved beside the project for every shot that shipped.
- Watched once, end to end, on a phone.

Today: export your current project at your editor's default preset and again at 20 Mbps, and compare the two at full size on the busiest shot.

BEFORE YOU MOVE ON

Why is exporting a small, heavily compressed file for upload a mistake even though the platform will compress it anyway?

A. Platforms allocate a bitrate proportional to the uploaded file's bitrate, so a small upload receives a lower-quality transcode.

B. Compressed uploads are transcoded more aggressively because the platform cannot detect the original quality and defaults to its lowest tier.

C. A small file lacks the high-frequency information the encoder uses to decide where to allocate bits, so its rate-control decisions are made on degraded input and distribute quality poorly across the frame.

D. The platform's encoder then compresses already-compressed material, so detail is discarded twice and artefacts compound rather than being applied once.

58 · 9 min

The whole pipeline on a phone or an old laptop

THE ONE THING TO KEEP

Every step of post-production has a version that runs on modest hardware, and the constraint that actually bites is playback performance rather than capability — which proxies solve for free.

The honest situation

DaVinci Resolve is free and it is heavy. It wants a reasonable GPU and 16GB of RAM, and on a machine with integrated graphics and 8GB it will install, open, and then stutter through playback in a way that makes editing genuinely unpleasant.

That is a real barrier and talking around it does not help anybody. So here is what actually works on what.

The tool ladder

Resolve free tier, if the machine can run it. A professional editor, colourist and audio suite in one, free permanently, no watermark. Real limits in the free version: some effects and the full noise reduction are withheld, and there are restrictions around certain high-resolution and codec outputs.

Kdenlive and **Shotcut**, open source, much lighter, happy on older Windows and Linux machines that Resolve refuses. For cuts, titles, audio and basic colour they are not worse. For colour work they are clearly behind — no shot match, weaker scopes, no tracked secondaries.

Olive and **Avidemux** for very constrained machines, where Avidemux is really a trimmer rather than an editor.

CapCut on a phone. The whole assembly is genuinely achievable: multiple tracks, audio, titles, subtitles, filters, speed. For vertical short-form it is often

faster than a desktop. Its free tier and watermarking on certain templates have moved around over time, so check before delivering something commercial.

FFmpeg on anything at all, including a phone through Termux. Not an editor, and it does conform, trim, concatenate, retime, denoise, scale, mix audio and export. Every operation in the preceding lessons except the creative choices has an FFmpeg equivalent.

Proxies, which change everything

The single most effective intervention on a slow machine. Edit with low-resolution copies, swap in the originals at export.

```
mkdir -p proxy
for f in footage/*.mp4; do
  ffmpeg -i "$f" -vf scale=640:-2 -c:v libx264 -preset veryfast
-crf 28 \
    -c:a aac "proxy/$(basename "$f")"
done
```

Resolve and Kdenlive both support proxy workflows natively, where the editor manages the swap. Doing it manually works too: edit the proxies, then relink to the originals before export.

A 640-pixel proxy plays smoothly on almost anything, and every editing decision except colour and fine detail can be made on it.

Working in stages

If the machine cannot hold a whole project, do not make it. Split the work:

1. Conform and trim every clip with FFmpeg first, so the editor receives clean, matched material.
2. Edit picture only, with proxies, no effects.
3. Export that cut, then do the grade as a second pass on the flattened export.

4. Do sound as a third pass in Audacity against the exported picture, then mux the finished audio back in.

```
ffmpeg -i picture.mp4 -i final_mix.wav -c:v copy -c:a aac -b:a
320k -shortest out.mp4
```

This is how editing worked before machines could play everything at once, and it still works. It is slower and it removes the memory pressure entirely.

What actually needs a good machine

Honestly, three things:

- **Real-time playback of many layers.** Solved by proxies and by flattening between stages.
- **Noise reduction and heavy effects.** FFmpeg's denoisers cover most of the need at the cost of a render wait.
- **Generating the video in the first place.** This is the real constraint, and it is why hosted tools and free GPU sessions exist. Editing is not the expensive part.

Notice that none of them is the editing itself. A sequence of cuts, a grade, titles and a sound mix are not computationally demanding tasks; they became demanding because editors render everything live.

The phone-only workflow, stated as a real option

Generate through a browser on a hosted free tier or a Hugging Face Space. Repair stills in Photopea in a mobile browser or Snapseed. Assemble in CapCut. Record voice and ambience on the phone's own microphone. Clean audio in a mobile audio editor, or accept the phone recording as it is. Add subtitles with the editor's own transcription. Export and publish.

That is a complete pipeline. It is not a lesser version of the work with a caveat attached — short-form vertical video made this way is the dominant format on the internet, and the constraints of the phone match the constraints of the format.

Today: make proxies for your current project with the loop above and edit the proxies. Time how much less you wait.

BEFORE YOU MOVE ON

An editor on an 8GB laptop finds playback unusable in a free professional editor. What single change most improves the situation without changing tools?

A. Edit with low-resolution proxy copies and relink to the originals before export, so playback decodes small files instead of full-resolution ones.

B. Reduce the timeline resolution to half in the project settings, which lowers the decoding load for playback while preserving full quality at export.

C. Convert all footage to an intra-frame codec so each frame decodes independently, removing the cost of reconstructing frames from a group of pictures.

D. Render effects to disk as they are added rather than previewing them live.

CASE STUDY · MODULE 6

The four-K admissions film

A coaching centre in Kota making a three-minute admissions film for its website and for parents' phones, edited on a five-year-old laptop with integrated graphics by a first-year student paid by the hour.

The centre's owner had two instructions. The film had to be in 4K, because a rival's film said 4K on the video player, and it had to be finished by Sunday.

The student editing it, Prakash, had thirty-one generated clips in a folder. They came from two different tools. Nine were 720p at 24 frames per second, nineteen were 1080p at 30, and three were 1080p at 25, which nobody had asked for and which arrived because a free tier had defaulted that way. Several were 5.1 seconds long.

His first attempt was to do exactly what he had been told. He upscaled each clip to 4K individually, dropped them on a timeline, and discovered two things within an hour.

The laptop could not play it back. Not slowly — it stalled, and every trim took three attempts to judge.

And the cuts looked wrong in a way he could not immediately name. Each clip had been upscaled separately, so each carried a slightly different sharpening signature, and at every cut the texture of the picture changed. Some shots also juddered on slow moves, because the 30 and 25 frames-per-second clips were having frames dropped to fit a 24 frames-per-second timeline he had set without thinking about it.

He started again and made two decisions that both had a cost.

The first was to edit on proxies. A loop through FFmpeg turned all thirty-one clips into 640-pixel copies in about four minutes, and the laptop played those without complaint. The cost was that he could not judge colour or fine detail while cutting, and that he would have to relink to the originals before export and hope he had not broken the paths. He kept the originals in one folder and never moved them, which is the whole discipline.

The second was to tell the owner that the film would be 1080p.

That conversation was the difficult one, and Prakash prepared for it the way the venue argument in another case was prepared for: he built evidence rather than an argument. He exported thirty seconds at 4K from upscaled 720p sources, and thirty seconds at 1080p from the originals with one grade and one light grain pass across the whole sequence. He put both on the owner's phone, which is where parents would watch it.

The 4K version was larger, softer and, on a phone, visibly worse: the upscaler had invented detail that did not survive the platform's re-encode, and the per-clip sharpening differences were still visible at the cuts.

The rest was conforming. He set the project to 30 frames per second, because the film was for phones and a website rather than for anything that needed to read as film. He converted the 24 and 25 frames-per-second clips deliberately with frame blending rather than letting the timeline drop frames, checking each one — a static shot can lose frames with no cost, a slow pan cannot. He trimmed roughly a quarter-second from the head of every clip and a second from the tail, which removed the settle and the drift and took forty seconds off the running time before he had made a single editorial decision.

Then one shot-match pass to a hero clip, one grade on an adjustment layer, one grain layer over everything set so it was only visible on a paused frame, and an ambience bed recorded on his own phone in the centre's corridor running unbroken underneath every cut.

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 film delivered on Sunday at 1080p, exported as a high-bitrate master at 20 megabits with a separate vertical version derived from it. The owner's objection to 1080p lasted until he watched both on his own phone. What con-

vinced him was not the resolution argument but the grain: the 1080p version looked, in his words, like it had been filmed. Prakash's note to himself afterwards was that the two operations that made the most difference — one grade and one grain pass across the whole timeline — took nine minutes together, and that the afternoon he had spent upscaling thirty-one clips individually had made the film worse.

Why did upscaling each clip individually to 4K make the cuts look wrong?

An upscaler does not restore detail, it invents plausible detail, and it leaves a sharpening signature. Run thirty-one times on thirty-one different clips, it produces thirty-one slightly different signatures, so the texture of the picture changes at every cut — an extra difference across a join that the eye reads as inconsistency. If upscaling is done at all it belongs once, at the end, on the finished cut, so the whole piece shares one signature.

Prakash set the project to 30 frames per second and then converted the mismatched clips deliberately. What would have happened if he had simply left them on the timeline?

A timeline has one frame rate, and any clip that does not match is resampled by duplicating or dropping frames. A 24 frames-per-second clip on a 30 frames-per-second timeline repeats one frame in four; the result is judder, which is invisible on a static shot and an obvious stutter on a slow pan. Converting deliberately — frame blending as a safe default, optical flow where the motion needs it, per clip rather than globally — removes it. Deciding the project rate before importing anything is what makes this a ten-minute job rather than a re-conform.

The grain pass changed the owner's mind about resolution. Why does one grain layer do more for perceived quality than extra pixels?

Every clip went through an autoencoder that discarded high-frequency detail, leaving a smooth, denoised surface that reads as synthetic more strongly than colour or motion does. Real captured images carry noise. One grain layer over the finished timeline replaces that missing surface property and gives every clip the same texture, which is a form of matching colour work cannot achieve. Extra pixels from an upscaler add file size and invented detail that the platform's re-encode discards anyway.

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 30 frames-per-second clip sits on a 24 frames-per-second timeline and a slow pan stutters. What is happening?

A. The clip's colour space does not match the project's
B. The codec is dropping keyframes to hold the target bitrate
C. One frame in five is being discarded to fit the project rate
D. Optical flow is interpolating frames that were never generated

2 Why is trimming roughly a quarter-second off the head and a second off the tail of every clip a default rather than a matter of taste?

A. It brings every clip to the same length, which is what the assembly pass needs
B. It removes the settle and the drift, which is where a clip announces itself
C. Platforms trim the first and last second on upload anyway
D. Shorter clips compress better at any given bitrate

3 Why does shot matching come before the grade rather than after it?

A. A contrast curve pushes two different starting points further apart
B. Scopes only read correctly on an ungraded timeline
C. Matching is done per clip and most editors apply clip effects alphabetically
D. The grade is destructive and cannot be undone once applied

4 Why is grain applied once over the finished timeline rather than to each clip?

A. Grain has to sit below the grade in the layer order
B. Grain filters are expensive to render and a single pass is much faster
C. Thirty grains differ slightly, so every cut gains one more change
D. Per-clip grain is discarded by the export codec

5 A shot contains a generated splash. Which retime is the mistake?

A. Slipping the clip so the splash falls on the cut point
B. Speeding the last second so the drift passes quickly
C. Speeding it to 115 per cent to tighten the pace
D. Slowing it to 40 per cent so the splash lands

6 Why does exporting a small file to save upload time make the final picture worse rather than the same?

A. Platforms penalise small uploads by serving them at a lower playback quality tier
B. The platform re-encodes your already-compressed file: two lossy passes, not one
C. Small files lose their colour space metadata
D. The platform's encoder ignores any file below a minimum bitrate

MODULE 7

Money, clients and running the job

Generative video is the first creative tool in years with a meter running, which changes how a project has to be planned, quoted and executed. This block builds the production apparatus: a shot list that can be batched, an animatic that gets approval before anything expensive happens, a draft-then-final discipline, queue management through an API, the arithmetic of subscription against top-up, a quote priced in shots with a stated attempt allowance, and the record that makes a revision six weeks later a small job rather than a reshoot.

> BY THE END OF THIS MODULE YOU CAN
>
> Plan, price and run a generative video job end to end — a shot list with durations and dependencies, an approved animatic before any final generation, a draft pass at the cheapest tier, a kill number per shot, a quote expressed in shots with included attempts, and a project record that makes a later revision cheap

59 · 9 min

An afternoon of curiosity can cost more than a day of work

THE ONE THING TO KEEP

Price your work in shots with a fixed number of included attempts, because the cost driver is generations thrown away, not hours spent.

The first creative tool in years with a meter running

Open Photoshop and you can experiment for eight hours at no marginal cost. Generative video is not like that. Every button press spends money, and the interface deliberately does not show you money, it shows you credits.

As of 2026, list prices sit roughly between about ten cents and seventy-five cents per generated second, depending on model, resolution and tier. That range moves constantly; check the current page rather than trusting the number here.

The figure that actually matters is not price per second. It is price per **usable** second. At a one-in-six hit rate — normal once a shot has a specific requirement — an eight-second clip listed at two dollars really cost twelve, plus an hour of your attention. A thirty-shot piece at that rate is a real invoice, and people discover this at shot nineteen.

Credits are not money, and that is the point

Every platform prices in credits. Different models and resolutions cost different credits per second. Credits usually expire at the end of the billing month. Aggregators add a margin on top of the underlying model's price.

None of this is dishonest, but all of it puts distance between the click and the cost. So close the distance manually, once: work out what a single eight-second 1080p generation on your chosen model costs in your own currency.

Write it on paper and put it next to the screen. That one number changes behaviour more than any advice in this lesson.

A discipline that keeps the bill survivable

1. **Shot list first.** You cannot batch what you have not planned, and unplanned generation is where budgets die.
2. **Draft pass at the cheapest setting.** Lowest resolution, shortest duration, fastest model or explicit turbo tier. At this stage you are testing whether a motion idea works, not making the shot. A draft that costs a tenth of a final tells you nine times as much per rupee.
3. **Fix the still, not the prompt.** Regenerating a still costs on the order of one per cent of regenerating a clip. Almost every failure has a cheaper diagnosis one step upstream.
4. **Set a kill number before you start.** Three attempts per shot. On the fourth you are not allowed to generate again — you change the still, change the shot, or cut it from the list. Without a kill number, one stubborn shot quietly eats the budget for eight others, and it is always a shot that did not matter.
5. **Batch.** Queue every shot at once instead of watching each one render. You make better decisions reviewing eight results together than one at a time, and it breaks the drip of just-one-more that is how afternoons disappear.
6. **Final pass at full quality on the chosen take only.** Same still, same prompt, same seed where the tool allows it.

Subscription against pay-as-you-go

A subscription is cheaper per second if you finish a project inside the billing month, and worse than useless if you do not, because unused credits generally expire rather than roll over. Pay-as-you-go top-ups — through the platforms themselves or through fal.ai and Replicate — cost more per second and nothing at all in a month where you are not working.

For a student, or anyone with irregular work, top-up is usually the honest choice. Read the rollover terms before subscribing; most do not roll over, and the ones that do usually cap it.

The free path, and what it costs instead

Free daily allowances on hosted tools, Hugging Face Spaces, and Colab or Kaggle GPU sessions cost nothing in money. They cost queue time and session limits, and a free session that dies at eighty per cent through a render is a real loss of an hour.

Use them for learning and for your own work without hesitation. Do not promise a client a deadline that depends on somebody else's free queue.

Quoting client work

Quote generation as a line item, and quote it in shots rather than in hours. Something like: thirty shots, up to four attempts included per shot, further revisions billed per shot.

That sentence exists to protect you from the client who asks for the twelfth version of shot nine without knowing that each version costs money. Hourly rates hide the cost driver from both of you. Per-shot pricing puts it where the client can see it and make a choice.

Today: open your tool's pricing page and work out, in your own currency, what one eight-second 1080p generation costs. Multiply by six. That is your real cost per delivered shot, and it is the number your quote has to be built on.

BEFORE YOU MOVE ON

A freelancer quotes a thirty-shot AI video by the hour and finishes sixty per cent over budget, having spent most of the overage on two shots the client kept revising. What structural change actually fixes this?

A. Switch to a cheaper model so each generation costs less and the overage shrinks proportionally.
B. Add a contingency percentage on top of the hourly rate to absorb difficult shots.
C. Move to a monthly subscription so generations are effectively free at the margin once the fee is paid.
D. Quote per shot with a stated number of attempts included, and bill further revisions per shot, because the cost driver is generations rather than hours.

60 · 9 min

The document everything else is built from

THE ONE THING TO KEEP

A shot list with a duration on every line converts an open-ended creative task into a countable one, which is what makes batching, quoting and stopping possible.

Why this is the highest-value document

Without a shot list you generate until you run out of money or patience. With one you know how many shots exist, how long the piece is, what each shot needs, and therefore what it costs.

It is not an AI workflow. It is the same document a crew has always worked from, and it transfers unchanged.

A three-minute explainer is roughly thirty shots. Write those thirty lines before you generate anything. On a phone, in a notes app, is fine.

What a line contains

Seven fields, and you can fit them on one line each:

```
#   SIZE  DESCRIPTION                          SEC  SOURCE
NEEDS        STATUS
s01 EWS   city rooftops at dawn, slow push in  4    gen      –
done t2
s02 MS    Priya at desk, turns to window       3    gen
char sheet   takes 1–3
s03 ECU   hands on keyboard                    2    gen      –
–
s04 MS    reverse, window, light on desk       3    plate
s01 grade    –
s05 CU    Priya reacts                         2    gen
char sheet   blocked
```

- **Number.** So you can refer to it in a conversation.
- **Size.** From the shot-size lesson. Check no two consecutive shots of the same subject are within a step of each other.
- **Description.** What happens, in the caption style the model responds to.
- **Seconds.** The intended screen duration, not the generated duration.
- **Source.** Generated, plate reuse, parallax, filmed, stock, graphic. Not everything is generated, and marking which is which is half the budget.
- **Needs.** Dependencies: a character sheet, a plate, a composited product, a consent form.
- **Status.** Where it stands.

Durations, and the number they add up to

Write a duration on every line and total the column. That number is your piece's length and almost nobody's first shot list produces the length they expected.

If you are cutting to music or to a voiceover, do it the other way: record or find the audio first, mark where each shot has to land, and derive the durations from that. Editing to a fixed audio length with shots of arbitrary duration is where projects fall apart, because generated clips do not arrive at the length you asked for.

Mark the dependencies

The *needs* column is what lets you order the work. Shots needing a character sheet cannot start until the sheet exists. Shots derived from a plate cannot start until the plate is generated. Shots with a real product cannot start until the product is photographed.

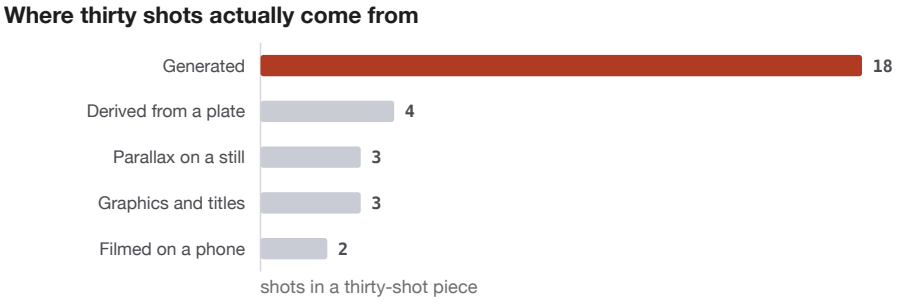
Sort by dependency and you get a work order:

1. Build the assets — character sheet, plates, photographed objects.
2. Generate the stills for every shot.
3. Check the stills side by side for continuity.
4. Draft-generate everything at the cheap tier.
5. Edit the drafts into an assembly.
6. Final-generate only the shots that survived.

Step six is where most of the saving lives, and it only exists because step five told you which shots survived.

Mark what is not generated

The most valuable column for a budget. On a typical thirty-shot piece a realistic breakdown might be: eighteen generated, four derived from plates, three parallax moves on stills, two filmed on a phone, three graphics.



Twelve of the thirty cost nothing to generate, and nobody watching can tell which were which. A shot list with no source column quietly quotes a generation price for a title card, and it hides the twelve lines that were never the expensive part.

Twelve of thirty cost nothing to produce. Nobody watching can tell which were which, and a shot list that does not distinguish them quietly quotes generation prices for graphics.

Keep it in a plain format

A text file, a markdown file, or a spreadsheet. Not in a chat window, not in your head, not in a tool's own project view.

The reason is that this document is also your quote, your progress tracker, your revision reference and your post-mortem. It gets read by you six weeks later and possibly by a client. It needs to outlive whatever software you used this month.

Revise it out loud

Read the list to somebody, or to yourself, shot by shot, as if describing the finished video. Two things emerge every time: shots that do nothing, and a missing shot that the sequence obviously needs.

Finding those now costs an edit to a text file. Finding them after generating costs the generations.

Today: write the shot list for whatever you are making, total the duration column, and count how many lines say generated.

BEFORE YOU MOVE ON

Why does marking a *source* column on a shot list — generated, plate, parallax, filmed, graphic — materially change a project's budget?

A. Because shots not produced by generation cost nothing to make, and a list without the distinction prices the whole piece at generation rates.

B. Because different sources require different frame rates, and conforming mixed sources adds processing time that must be budgeted for.

C. Because generated shots need a higher attempt allowance, so the column lets you apply different revision terms to different shots in the quote.

D. Because platforms require disclosure of which portions of a video were synthetically generated, so the column forms the basis of the labelling statement at delivery.

61 · 9 min

Get the yes before you spend the money

THE ONE THING TO KEEP

An animatic made from stills costs almost nothing and carries the same approval as a finished cut, which moves the expensive client conversation to before generation rather than after it.

The conversation that costs the most

A client watches a finished thirty-shot piece and says the pacing is wrong, or that they imagined it warmer, or that shot nine should be a close-up. Each of those is now a re-generation.

The same conversation, held over an animatic, costs a rearranged text file.

So the discipline is simple: **nothing expensive happens until somebody has approved something cheap.**

The animatic

An animatic is the shot list made watchable. Stills on a timeline, each held for its intended duration, with a scratch voiceover and temporary music.

Build it from what you already have:

1. The first frames you generated for every shot — they cost a cent or two each.
2. Each one held on screen for the duration on the shot list.
3. A scratch voiceover, recorded on your phone in one take.
4. Temporary music at roughly the right feel.
5. Titles as plain text.

That is an afternoon at most and it costs essentially nothing. What it buys:

- **The real duration.** Not the total of your estimates, but what it feels like to sit through.
- **Pacing problems**, which are invisible on a list and obvious in motion.
- **Missing shots**, which appear as moments where the voiceover has nothing to sit under.
- **A client decision**, made on something concrete rather than on a description.

Add a small push-in on each still using the parallax technique and the animatic starts to read as a film rather than as a slideshow. That is free and it changes how seriously a client watches it.

Getting approval that holds

Send the animatic with a specific, short list of what you are asking them to approve:

> *Please confirm: the order of shots, the length, the script, and the look of the four key frames attached. Once approved, changes to structure or shot count are billable per shot.*

That sentence is not aggressive, it is clarity. Clients change their minds because nobody told them when changing their mind becomes expensive. Telling them is a service, and it is how you avoid the conversation where you absorb twelve regenerations out of embarrassment.

Get the approval in writing, by email. A verbal yes on a call is not a record.

Key frames as the look approval

Separately from the animatic, send three or four finished-quality stills — the shots that carry the look. Those are the approval for style, colour and character design.

This is standard practice in animation and advertising, where it is often called a key-frame or style-frame approval, and it exists for exactly the reason it matters here: agreeing the look on four images is cheap, and discovering a disagreement on thirty clips is not.

Pre-visualisation as the deliverable

Worth naming as a service in its own right, because it is currently one of the strongest commercial uses of this technology.

A director pitching a commercial, a studio testing a sequence, an agency selling an idea — all of them need a moving version of a shot before anyone commits to a shoot. That used to require an animatic team or a funded production. Now it takes an afternoon.

The output does not have to be broadcast quality. It has to communicate a shot, and it does that well. Selling pre-visualisation is honest work at an honest price, and it does not depend on the generated footage being good enough to ship.

When the client is you

The same discipline applies to personal work and it is harder to enforce, because there is nobody to send it to.

Build the animatic anyway and watch it once, the next day. The gap of a night is what gives you the outside view a client would have had. Almost everybody discovers the piece is too long.

Today: put your existing stills on a timeline at their intended durations, record a scratch voiceover on your phone, and watch the result before generating anything else.

BEFORE YOU MOVE ON

A studio builds an animatic from first frames and a phone voiceover before generating any clips. Beyond saving money, what does the animatic reveal that a written shot list cannot?

A. Which shots will drift, since holding each still for its full duration exposes the frames that are structurally weak.

B. Whether the colour and lighting match between shots, because the frames are played in sequence rather than examined individually.

C. How the piece feels to sit through — its real pacing and where it sags — which is a property of duration in sequence rather than of any list.

D. The final cost of the project, since each still on the timeline corresponds to one generated clip and the count becomes concrete at that point.

62 · 8 min

Test the idea at a tenth of the price

THE ONE THING TO KEEP

A draft generation at low resolution on a distilled model answers whether a motion idea works, which is a different question from whether the shot is good, and answering it cheaply first is where most of the saving in a project lives.

Two different questions

When you generate a shot you are really asking one of two things:

1. **Does this idea work at all?** Does the camera move suit the frame, does the subject motion read, is the shot even a shot?
2. **Is this specific take good enough to ship?**

These cost wildly different amounts to answer, and most people pay the second price to get the first answer.

A draft — lowest resolution, shortest duration, fastest or distilled model — answers question one for perhaps a tenth of the price. It tells you nine times as much per unit of money, because you can ask it nine times.

What a draft is for, and what it is not

A draft tells you: whether the move suits the composition, whether the subject motion reads, roughly where the clip starts drifting, whether the shot has a purpose in the sequence.

A draft does not tell you: the final quality, whether the face holds at full resolution, or what the finished grade will look like. Turbo models have reduced motion range and weaker prompt adherence, so a draft that looks calm may come back livelier at full quality, and one that ignored an instruction may obey it.

Be clear which question you are answering. Treating a draft as a preview of the final is how people get surprised twice.

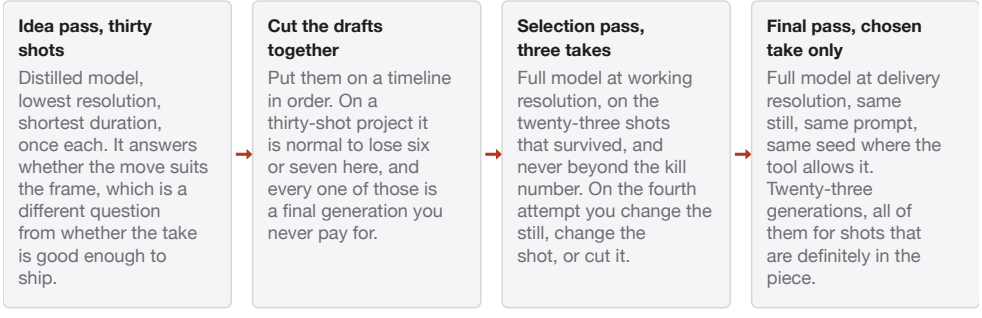
The ladder

A typical project moves through three tiers:

1. **Idea pass.** Distilled model, lowest resolution, shortest duration. Every shot on the list, once. Cheap enough to do the whole list.
2. **Selection pass.** Full model, working resolution, on the shots that survived the assembly. Three or four takes each.
3. **Final pass.** Full model, delivery resolution, same still, same prompt, same seed where the tool allows it, on the chosen take only.

The saving comes from the fact that tier one covers thirty shots and tier three covers the twenty that made the cut, at three or four takes rather than ten.

Three tiers, and where the saving actually is



One hundred and twenty-two generations in total, and the thirty cheapest are what decided the other ninety-two. Running the whole list at full quality from the start means paying final prices for the seven shots that were always going to be cut.

Edit the drafts

The step that makes the whole thing work. Put the draft clips on a timeline and cut them together.

This is a moving version of the piece at a small fraction of the cost, and it does the job the animatic did one level up: it tells you which shots are unnecessary, which are missing, and which are the wrong length. Every shot you delete at this stage is a final generation you never pay for.

On a thirty-shot project it is normal to cut six or seven shots at this point. That is a fifth of your generation budget recovered for the price of an afternoon.

The kill number

Set it before you start: **three attempts per shot.**

On the fourth attempt you are not allowed to generate again. You must do one of three things:

- Change the still.
- Change the shot — a different size, a different move, a simpler idea.
- Cut it from the list.

Without a kill number, one stubborn shot quietly eats the budget for eight others, and it is always a shot that did not matter. With one, the failure is bounded and the decision is made while you still have money.

The kill number is the single most effective budget control in this whole subject, and it is free.

Fix the still, not the prompt

Restating it because this is where it costs money. Regenerating a still is on the order of one per cent of the cost of regenerating a clip. Almost every clip failure has a cheaper diagnosis one step upstream.

When a generation disappoints, the first question is always "is there something wrong with the frame?" and the answer is usually yes.

The free path

Free daily allowances are exactly right for the idea pass. Spend them on real shots from the real list at the cheapest settings rather than on tests you will throw away, and you learn far more per free credit.

A phone plus two or three free tiers plus an afternoon is a complete idea pass for a thirty-shot project, and it costs nothing at all.

Today: pick your kill number, write it at the top of your shot list, and hold to it on the next shot that fights you.

BEFORE YOU MOVE ON

A draft generated on a distilled model at low resolution comes back calm and slightly stiff. The operator concludes the motion idea does not work and cuts the shot. What is the error?

A. Low resolution reduces the latent's capacity for motion, so any draft will understate movement regardless of which model produced it.
B. Distilled models have reduced motion range, so a calm draft does not establish that the full model would be calm.
C. Drafts are generated at a shorter duration, so the clip was cut before the motion had time to develop into what the shot needed.
D. The draft used a different seed from the one the final pass would use, so the two are separate random draws and no conclusion transfers between them.

63 · 9 min

Queue everything, then judge it together

THE ONE THING TO KEEP

Generating one clip at a time invites the just-one-more loop that destroys budgets, while queueing a batch and reviewing the results together produces better decisions and a countable spend.

Why one at a time is expensive

Watching a single generation render creates a specific behaviour. It finishes, it is nearly right, you adjust one word, you press generate, you watch again. An

afternoon disappears and the spend is invisible because it arrived in small pieces.

Batching breaks the loop mechanically. Queue eight shots, walk away, come back to eight results, and judge them together.

The decisions are also better. Comparing eight results side by side tells you which are genuinely good; watching one at a time tells you only whether each is better than the one before it, which is a much weaker judgement.

Batching without an API

Most hosted tools let you queue several generations. Use it even when the interface makes it awkward:

- Write all your prompts and stills first, as a batch of work, before opening the tool.
- Submit them all.
- Close the tab.
- Come back later and download everything.

That last step matters more than it sounds. Staying on the page is what re-creates the loop.

When an API is worth it

If you are producing regularly, driving generation from a script is a different way of working and it is not difficult.

fal.ai and **Replicate** both expose most current models behind a single API with pay-as-you-go billing, which means one account rather than six. The platforms' own APIs exist too, with their own SDKs.

A minimal loop over a shot list:

```
import json, pathlib, fal_client   # or replicate

shots = json.load(open("shots.json"))   # [{id, image, prompt,
seconds}, ...]
out = pathlib.Path("takes"); out.mkdir(exist_ok=True)

for s in shots:
    for take in range(3):
        r = fal_client.run("MODEL-SLUG", arguments={
            "image_url": s["image"],
            "prompt": s["prompt"],
            "duration": s["seconds"],
            "seed": 1000 + take,
        })
        (out /
f'{s["id"]}_t{take}.json').write_text(json.dumps(r))
        print(s["id"], take, r.get("video", {}).get("url"))
```

What this buys, beyond speed:

- **Every parameter is recorded**, because it is in the script. Your provenance problem is solved as a side effect.
- **Seeds are explicit and varied deliberately**, so three takes are three known draws rather than three unknown ones.
- **Repeatable runs.** Re-running the script with one changed field is an experiment, not a session.
- **Spend is countable.** Shots times takes times price, known before you run it.

Rate limits, failures and refunds

Things that will happen and should not surprise you:

- **Rate limits.** Providers cap concurrent jobs. Handle it with a small delay and a retry rather than by hammering.
- **Failed generations.** A job returns an error or a black clip. Most providers do not charge for a hard failure, and some charge for a completed-but-bad one. Check your usage page against your expected spend after the first batch.

- **Content filter refusals.** Hosted models occasionally decline legitimate work with no explanation. Budget for it and keep an open-weights fallback for anything the filters are awkward about.
- **Silent model updates.** Pin a version string in the call if the provider supports it.

Always write the response to disk, including the failures. A log of what failed and why is what tells you whether a shot is unlucky or impossible.

Local batching

If you run models yourself, batching is the default rather than a feature. ComfyUI queues as many jobs as you give it and works through them, and the marginal cost of a fourth take is electricity.

That changes the economics of experimentation completely, and it is the strongest practical argument for a local setup even if your final renders happen on a hosted model.

The habit that survives any tool

Prepare work in batches, submit, leave, review together, decide. It applies to a scripted API run, a hosted queue, or a free tier where you submit three jobs a day and collect them tomorrow.

The tool changes; the loop that wastes money does not.

Today: prepare eight shots fully before opening your tool, submit all eight, close the tab, and review them together an hour later.

BEFORE YOU MOVE ON

Beyond throughput, what is the main decision-quality benefit of queueing eight generations and reviewing them together?

A. Batch submission uses a shared seed sequence, so the eight results are comparable in a way independently submitted jobs are not.
B. Reviewing results in a group lets you judge which are genuinely good, rather than only whether each is better than the one before it.
C. Queued jobs are processed at the same model version, eliminating the risk that a silent update changes the output partway through a session.
D. Batch pricing is typically lower per second than interactive generation, so the same review can be done across more takes for the same spend.

64 · 8 min

The arithmetic nobody does before subscribing

THE ONE THING TO KEEP

Credits usually expire at the end of a billing period, so a subscription is cheaper per second only if you actually finish a project inside that window and is worse than useless in a month where you do not work.

The two models

Subscription. A monthly fee buys a credit allowance at a lower effective price per second. Unused credits generally expire at the end of the period; where they roll over, it is usually capped and usually only while you remain subscribed.

Pay-as-you-go top-up. You buy credits when you need them, at a higher price per second. Most top-up credits do not expire, or expire over a much

longer horizon. Available directly and through aggregators such as fal.ai and Replicate.

Doing the sum

The question is not which is cheaper per second. It is which is cheaper for how you actually work.

Work out three numbers:

1. **Your monthly usage in seconds of generation**, from your last two real projects. Not your optimistic estimate.
2. **The subscription's effective price per second**, which is the monthly fee divided by the credits it includes, converted into seconds at your usual model and resolution.
3. **The top-up price per second** for the same model and resolution.

Then: subscription cost per month against top-up price times your actual usage.

The result surprises people, because the comparison people run in their head is the subscription's price per second against the top-up's, which ignores that the subscription charges you whether you generate or not.

Irregular work changes the answer completely

For a student, a freelancer with uneven months, or anyone learning, the pattern is usually: three weeks of nothing, one intense week, then nothing again.

A subscription over that pattern pays full price for the quiet months and expires unused credits in them. Top-up costs more per second during the intense week and nothing at all otherwise, and over a year it is frequently cheaper in total.

The honest general advice: **if your work is irregular, top up.** If you generate every week, price the subscription against your real usage and subscribe only if the sum works.

What to read before subscribing

Four clauses, and they are not always on the pricing page:

- **Rollover.** Do unused credits carry over, and is there a cap?
- **What happens on cancellation.** Do remaining credits survive, and for how long?
- **Commercial-use rights.** Are they tied to the tier? Some services grant commercial rights only above a certain plan, which matters enormously if you are billing a client.
- **Queue priority and resolution limits.** Free and low tiers are often capped at lower resolutions or slower queues, which affects whether the plan can actually do your work.

Screenshot what the terms said on the day you subscribed. Terms change, and a dispute is much easier with a record.

Aggregators: the margin and what it buys

fal.ai, Replicate and consumer aggregators resell access to several underlying models. You pay a margin over the model's own price.

What the margin buys is real: one balance instead of six, one interface, the ability to switch models without opening new accounts, and no monthly commitment. For someone testing several models, or working irregularly, that is often worth more than the margin costs.

What it costs, besides money, is one more step of distance from the model's own controls, and a dependency on a company that is not the one building the model.

The number to write on paper

Whatever you choose, do this once: work out what a single eight-second generation at your working resolution costs in your own currency. Write it on paper and put it next to the screen.

Credits exist to put distance between the click and the cost. Closing that distance manually changes behaviour more than any budgeting advice, and it takes five minutes.

Then multiply by your measured takes-per-shot from the continuity module. That product is your real cost per delivered shot, and it is the number a quote has to be built on.

The free path, and its real cost

Free daily allowances, Hugging Face Spaces and Colab or Kaggle sessions cost nothing in money and cost queue time and session limits. A free session that dies at eighty per cent through a render is a real loss of an hour.

Use them without hesitation for learning and for your own work. Do not promise a client a deadline that depends on somebody else's free queue.

Today: work out your cost per eight-second generation in your own currency, multiply by your takes-per-shot, and write both numbers where you will see them.

BEFORE YOU MOVE ON

A freelancer works intensively for one week a month and does nothing the other three. The subscription's per-second price is half the top-up price. What should they consider before subscribing?

A. Whether the subscription tier includes commercial-use rights, since freelance work requires them and lower tiers sometimes exclude them.
B. Whether the subscription offers priority queueing, since an intensive week concentrates demand and queue time becomes the binding constraint.
C. That the fee is charged in the quiet months too and the credits expire unused, so the real comparison is total annual cost.
D. That intensive weeks may exceed the included allowance, at which point overage is billed at top-up rates and the effective saving on the excess disappears entirely.

65 · 9 min

Price the thing that actually costs money

THE ONE THING TO KEEP

The cost driver in generative video is generations discarded rather than hours worked, so a quote expressed in shots with a stated attempt allowance puts the real variable where the client can see it and decide about it.

Why hourly rates fail here

An hourly rate hides the cost driver from both sides. The client asks for the twelfth version of shot nine without knowing that each version costs money and attention; you absorb it, or you have an awkward conversation about hours that does not describe what actually happened.

The honest unit is the shot, and the honest variable is attempts.

The shape of a quote

> ***Thirty shots**, up to **four attempts** included per shot.*
>
> *Further revisions billed at **X per shot**.*
>
> *Includes: shot list, animatic, generation, edit, grade, sound, one delivery format.*
>
> *Excludes: music licensing, voice talent, additional delivery formats.*
>
> *Timeline: animatic in five days; approval required before final generation.*

Six lines. Everything a dispute would later be about is in them.

The attempt allowance is the important one, and it is not a restriction — it is information. It tells the client that attempts are the cost, which is a fact about

the medium they have no way of knowing.

Pricing the shot

Build the number from your own measurements rather than from a rate card:

- **Generation cost.** Cost per generation times your measured takes-per-shot. You have both numbers from the previous lessons.
- **Still preparation.** Making and repairing the first frame, which is often longer than the generation.
- **Edit share.** Total edit time divided by shot count.
- **Overhead and margin.**

Then sanity-check against what the work is worth to the client rather than what it cost you. A shot that saves a location shoot is worth what the location shoot would have cost, not what your credits cost.

The clauses that prevent arguments

Five things worth writing down, each of which corresponds to a real dispute:

Approval gates. "Final generation begins after written approval of the animatic. Structural changes after that point are billed per shot." Without this you will re-generate a finished piece for free.

What a revision means. Distinguish a *note* (make shot nine warmer — a grade, free) from a *re-shoot* (make shot nine a close-up — a new generation, billed). Clients do not know these are different and will not guess.

Who supplies what. Logos as vectors, products as photographs or physical items, brand fonts with their licence, any real person's written consent. Late supply is the most common cause of a missed deadline and it is rarely your fault.

Consent and likeness. "The client warrants that written consent has been obtained for any real person's likeness or voice supplied to us." This is not you giving legal advice; it is you declining to be the party who did not ask.

Disclosure. State that generated content will carry a visible label, and that platform and legal requirements may compel it. Put it in the quote so nobody is surprised at delivery.

Saying no early

The most valuable professional act in this subject is telling a client on day one that a shot is not achievable.

From the limits lesson: a specific real product as a generated hero, a character-driven dialogue scene across many shots, precise physical demonstration with hands, anything that will be scrutinised frame by frame. For all of those, a small live shoot, 2D animation or motion graphics is faster, cheaper and better.

Being the person who says this on day one is worth considerably more than being the person who can drive the tool and discovers it on day fourteen with the deadline in sight. It also wins work, because it reads as expertise rather than as reluctance.

What to do when they ask for a rush

Generation time is queue time, and queue time is not something you control. A rush means paying for a faster tier, running more takes in parallel, or accepting fewer attempts per shot.

Offer the trade explicitly: "We can deliver in three days at two attempts per shot rather than four, which raises the chance that some shots are merely acceptable." That is an honest statement about the medium and it lets the client choose.

Keep the record with the quote

When the project ends, put the actual numbers next to the quoted ones: shots delivered, takes used, hours spent. After three projects that file is the most accurate pricing tool you will ever own, and it is more useful than any published rate card because it is about your work at your hit rate.

Today: rewrite your last quote in shots with an attempt allowance, and see how differently it reads.

BEFORE YOU MOVE ON

Why does stating an attempt allowance per shot protect both the freelancer and the client, rather than only the freelancer?

A. It caps the freelancer's exposure to unlimited revisions while giving the client a contractual minimum number of takes to expect.

B. It converts a variable cost into a fixed one, so both parties can plan against a known figure rather than an open-ended one.

C. It tells the client that attempts are the cost driver, which is a fact about the medium they otherwise have no way of knowing before asking for a twelfth version.

D. It establishes a per-shot price that can be applied consistently to later work, so a client commissioning a second project can budget from the first one's terms without renegotiating the rate.

---

66 · 8 min

What to keep, and the day it saves you

THE ONE THING TO KEEP

A project record is what turns a later revision into a twenty-minute job instead of a reshoot, and it has to be assembled during production because most of it cannot be reconstructed afterwards.

The four things per shipped shot

For every clip that made the final cut, keep:

1. **The first frame**, as a lossless image file.
2. **The exact prompt**, copied rather than retyped.

3. **The model name and version string**, plus every numeric setting and the seed.
4. **The output file itself**, at the highest quality you received.

That fourth item is the primary record, not a fallback. The others let you get close; the file is the only thing that is certain, because providers update models and a seed that worked last month may not work today.

Everything else that belongs in the folder

By the end of a job, the project folder should contain:

```
project/
  shots.md               # the list, with final takes and dura-
tions marked
  quote.pdf              # what was agreed
  approvals/             # the emails approving animatic and
key frames
  consent/               # signed releases for any real face or
voice
  licences/              # music, fonts, stock — the actual li-
cence text
  frames/                # stills + sidecars, including rejects
until sign-off
  clips/                 # every generation, with its settings
file
  audio/
  edit/
  exports/
    master_2026-09-21.mp4
    delivered_1080_2026-09-21.mp4
```

None of this is bureaucracy. Each folder corresponds to a question somebody asks later: what did we agree, who said yes, did we have permission, can we prove the music was licensed, can you change shot nine.

Provenance, which you will be asked about

At some point a client, a platform or a broadcaster will ask what in this video was generated and what was filmed.

If every frame's sidecar says `generated`, `photographed`, `composited` or `stock`, the answer takes a minute. If it does not, the answer is a guess, and a guess is not something you want to put in writing about a disclosure question.

The labelling obligations themselves are dealt with in the next module. The record that makes compliance possible is built here, during production, because it cannot be assembled afterwards.

The post-mortem, which takes fifteen minutes

When the project delivers, before you forget:

```
PROJECT: name, delivered 2026-09-21
Quoted: 30 shots, 4 attempts each
Actual: 27 shots delivered, 71 generations, 2.6 takes/shot
Rejections: face drift 21, background churn 14, motion too
strong 9,
            wardrobe colour 5, other 3
Time: stills 6h, generation 4h, edit 11h, sound 3h, revisions
2h
What cost more than expected: reverse-angle shots in scene 3
What to do differently: build the plate first; three attempts,
not four
```

Six lines. After three projects this file prices your next quote more accurately than any rate card, and its rejection tally tells you where to spend your improvement effort — which is almost never where you assumed.

Archiving

At sign-off:

- Delete the reject frames if space matters. Not before — a client note two days after delivery is common and the parts bin is what answers it quickly.
- Keep the master export, the project file, the frames, the sidecars, the approvals, the consent forms and the licences.
- Copy the whole folder somewhere else. One copy in one place is not a backup, and this is the point in the process where people discover that.

- Note an expiry on anything time-limited: a consent that covers twelve months, a music licence tied to a subscription, a model whose terms changed.

Why this is a professional differentiator

Anyone can generate a clip. Very few people can, eight months later, produce the frame, the prompt, the model version, the consent form and the music licence for shot nine within ten minutes.

That capability is what a client is actually buying when they hire somebody rather than doing it themselves, and it is entirely free to provide. It is also what makes a revision a small billable job rather than an unpaid rebuild.

The version that fits on a phone

If the full structure feels like too much, keep three things and you will recover most of the benefit: a folder per project rather than a downloads pile, a sidecar text file beside every shipped frame, and the master export saved twice in two places. Everything else in this lesson is an elaboration of those three, and a project with only those three is already ahead of most freelance practice.

The habit that makes it stick is writing the record at the moment the decision is made rather than at the end. A prompt copied when you use it takes two seconds; the same prompt reconstructed a month later takes twenty minutes and is probably wrong.

Today: write the post-mortem for the last thing you finished, even roughly. The rejection tally is the interesting line.

BEFORE YOU MOVE ON

Why is the delivered output file, rather than the prompt and seed, the primary record of a shipped shot?

A. The file preserves the grade and sound work applied afterwards, which a re-generation from the prompt would not reproduce.

B. Prompts and seeds are frequently mistranscribed, so the file is the only copy that cannot contain an error.

C. Providers update models without notice, so the same prompt and seed may no longer reproduce the clip, while the file is unconditionally what shipped.

D. Seeds are only meaningful in combination with a specific model version, and version strings are frequently omitted from a provider's response, making the settings record incomplete by default rather than merely stale.

CASE STUDY · MODULE 7

The quote that said no to two shots

A freelance video maker in Guwahati, asked by a state tourism board to produce a two-minute film with a six-week deadline and a fixed budget already approved by a committee.

The brief arrived as a list of twenty-eight shots and one sentence about the budget. Bhaskar had been freelancing for four years and had done three generative jobs. He had never written a quote that refused anything, and he was about to.

Two of the twenty-eight shots were not achievable and he could say why.

Shot nine was a hero close-up of a named tea brand's packet, held and turned. The model has no reference for that packet. It would produce a member of the category with a label that is nearly right, which on a government tourism film promoting a specific producer is not a stylistic issue.

Shot twenty-two was a forty-second conversation between two characters across seven cuts, matching eyelines and wardrobe throughout. Seven shots that all had to match, at his measured hit rate of about one in two for a face at medium size, is arithmetic nobody would like: roughly two generations a shot before anything else went wrong, and a real chance that the sequence still would not cut together on light and wardrobe rather than on the faces.

He could have quoted for all twenty-eight and discovered both problems in week four with the deadline visible. Freelancers do this constantly, and the reason is not stupidity: saying no early feels like admitting you cannot do the work, and there is a live fear that the board will simply hire somebody who says yes.

He wrote the quote the other way round and put the two refusals at the top, each with an alternative and a price.

For shot nine: photograph the actual packet against a wall in window light, generate the environment and the camera move, and track the real packet in

during the edit. Half a day more than a generated shot, and it survives being looked at.

For shot twenty-two: film it. Two people, one afternoon, a phone on a tripod and a borrowed lapel microphone. Cheaper than generating it, faster, and better.

Then he changed the shape of the rest of the quote. His previous jobs had been priced hourly, which hides the cost driver from both sides: the client asks for a twelfth version of shot nine without knowing that each version costs money, and the freelancer either absorbs it or has an awkward conversation about hours that does not describe what happened.

The new quote priced shots. Twenty-six shots, up to four attempts included per shot, further revisions billed per shot. Included: shot list, animatic, generation, edit, grade, sound, one delivery format. Excluded: music licensing, and any additional aspect ratio. Final generation begins after written approval of the animatic, and structural changes after that point are billed per shot.

He built the number from his own records rather than from a rate card. His cost per eight-second generation, in rupees, multiplied by his measured takes per shot, plus still preparation, plus the edit divided across the shot count, plus overhead.

The risk he was taking was specific. A committee comparing three quotes would see one that delivered twenty-six shots where two others delivered twenty-eight, and committees do not always read the reasons.

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 awarded him the job and quoted his refusal paragraph back to him in the meeting as the reason. The person chairing it had been burned the pre-

vious year by a supplier who agreed to everything and delivered a product close-up that the tea producer rejected. The animatic — stills on a timeline at their intended durations with a scratch voiceover recorded on his phone — was approved in eight days, and cut four shots before any final generation was paid for. The filmed conversation took one afternoon. His post-mortem recorded 22 shots delivered, 58 generations, 2.6 takes a shot, and one line under what to do differently: three attempts included, not four.

Why is a quote expressed in shots with an attempt allowance more honest than an hourly rate for this kind of work?

The cost driver in generative video is generations discarded, not hours worked. An hourly rate hides that from the client, who has no way of knowing that asking for another version of a shot spends money, and it hides it from the freelancer, who ends up absorbing attempts or arguing about hours that do not describe what happened. Pricing in shots with a stated allowance puts the real variable in front of the client so they can make a decision about it, and it defines in advance what both sides agreed to.

Bhaskar refused two shots at the top of a competitive quote. What was the risk, and why did it pay?

The risk was that a committee comparing quotes would see twenty-six shots against a rival's twenty-eight and choose on the count without reading the reasons. It paid because naming where the tool stops is the most valuable thing a professional can tell a client, and because both refusals came with a cheaper, better alternative rather than a shrug. Being the person who says on day one that a shot is not achievable is worth more than being the person who discovers it on day fourteen, and to a client who has been burned before it reads as expertise.

The animatic cut four shots before any final generation was paid for. Why is that stage worth an afternoon of unpaid-looking work?

An animatic is the shot list made watchable: the first frames held for their intended durations, with a scratch voiceover and temporary music. It costs almost nothing because the stills already exist, and it reveals the real running time, the pacing problems, the missing shots and the unnecessary ones — none of which are visible on a list. Every shot deleted at that stage is a final generation never paid for, and it moves the expensive client conversation to a point where changing your mind costs a rearranged text file.

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 An eight-second clip is listed at two dollars and your hit rate on that kind of shot is about one in six. What did the delivered shot cost?

A. About twelve dollars, plus the attention
B. About four dollars, because most failures are caught in the draft pass
C. It cannot be estimated, because hit rates are not measurable
D. Two dollars, because failed generations are refunded

2 You set a kill number of three attempts. What does that require you to do on the fourth?

A. Buy more credits and continue with the same still
B. Switch to a different model and try the shot again
C. Change the still, change the shot, or cut it
D. Raise the resolution so the take is worth keeping

3 Your work comes in bursts: three quiet weeks and one intense week. Which billing model is usually cheaper across a year?

A. A subscription, because credits roll over between periods
B. Either one, because the totals come out the same across twelve months
C. A subscription, because its effective price per second is lower
D. Top-up, because you pay nothing in the months you do not work

4 Why does a shot list need a column recording how each shot will be produced?

A. Because platforms require a per-line disclosure of which shots were generated
B. Because a third of the shots often cost nothing to make
C. Because the editor needs it to conform frame rates
D. Because generated shots have to be listed last in the work order

5 What does an animatic tell you that a written shot list cannot?

A. Approval of the final grade and colour treatment
B. An accurate estimate of the generation cost of each shot
C. The real running time, and where the piece sags
D. The model's hit rate on each type of shot

6 Of the four things kept for every shipped shot, which is the primary record and why?

A. The output file, because providers update models
B. The prompt, because it can be re-run anywhere
C. The model version string, because it can be pinned in a later API call
D. The seed, because it reproduces the clip exactly

EXAMINATION

Video With AI — final examination

This paper covers the whole course: how a latent video model turns a prompt and a first frame into a clip, what belongs in the still rather than in the prompt, how to write a shot instruction a model can act on, what it costs to hold a character and a location across a sequence, the audio half of the job and the consent and labelling that go with it, the edit that turns generations into a piece, and how to plan, price and record a paid job. Twenty-five questions are drawn at random from a larger pool, so no two attempts are the same paper. The pass mark is 70 per cent. There is no timer: read the question twice if you need to. If you do not pass, you can retake it after thirty minutes with a fresh set of questions, as many times as you like — a failed attempt is not recorded against you anywhere. A teacher can open the next course for you regardless of this paper, and that override is recorded as an override rather than disguised as a pass. Every question asks you to explain a mechanism rather than recall 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. What the machine is actually doing

Why is a ten-second clip considerably more than twice the price of a five-second one?

A. Longer generations are billed at a premium rate by every provider

B. Longer clips are placed in a lower-priority queue, and skipping that queue costs extra credits

C. Attention work rises with the square of the token count, and the block must fit at once

D. The autoencoder has to be run twice for clips over eight seconds

2 1. What the machine is actually doing

After three uses of the extend button, the room is a different colour and the fine texture has gone flat. What is the mechanism?

A. Each extension re-encodes an already-decoded frame, so the loss compounds
B. Extensions are generated at a lower resolution to keep the price down
C. The model reuses the same seed and drifts toward its unconditional prediction
D. The prompt's influence decays across successive generations

3 1. What the machine is actually doing

Most current video autoencoders are causal. What practical consequence does that have for image-to-video?

A. The last frame is encoded without temporal compression instead of the first
B. Motion can only run forwards, so reverse camera moves have to be flipped in the edit
C. The clip can be streamed out as it is generated, one frame at a time
D. Your supplied still enters the latent with less loss than any frame after it

4 1. What the machine is actually doing

A person walks behind a pillar and emerges wearing a different shirt. No setting prevents this. Why not?

A. The occlusion mask is recalculated at each denoising step
B. Nothing in the model holds an object's identity; the tokens carrying it stopped existing
C. The attention window is too small to span the width of the pillar and the frames either side
D. The prompt did not specify the shirt in enough detail for it to persist

5 1. What the machine is actually doing

Why does adding a fourth clause to a video prompt so often change nothing at all?

A. A few hundred text tokens are steering a quarter of a million video tokens

B. Cross-attention layers are switched off after the first half of the denoising run

C. The extra clause is added to the negative branch by default in most tools

D. Most interfaces truncate prompts after three clauses

6 1. What the machine is actually doing

Asking a model for "a montage, then it cuts to the street" returns a morph or a camera move. What in the training explains that?

A. Cuts are filtered out at inference by the safety layer in most hosted tools

B. Aesthetic scoring removed clips containing hard cuts

C. Scene detection split the training data into single unbroken shots

D. Captioners were instructed not to describe edits

7 1. What the machine is actually doing

A clip interpolated from 16 to 24 frames per second shows a soft halo hugging a hand as it crosses in front of a body. What produced it?

A. The autoencoder's tiled decode leaving a seam at the tile boundary

B. At an occlusion boundary some pixels exist in one frame and not the next

C. The model generated the hand at a different resolution from the body

D. Motion blur was added by the interpolator and then sharpened again at export

8 1. What the machine is actually doing

You move a working prompt from a full model to its distilled turbo variant and keep guidance at 6. The output is burnt and almost still. Why?

A. Guidance is applied per frame on distilled models and per clip on full ones

B. The turbo variant uses a different text encoder that rejects long prompts

C. Turbo variants require twice the step count to reach the same quality

D. The distilled model was trained to need no guidance, so 6 overshoots badly

9 2. The first frame, and everything decided in it

You compose a still you are happy with, then add a slow push-in. By second three the composition is worse. What should you have done?

A. Generated at a higher resolution so the crop had more pixels

B. Named the push-in as a percentage instead of as a phrase in the prompt

C. Lowered the motion strength so the crop was less aggressive

D. Composed for where the frame ends rather than where it starts

10 2. The first frame, and everything decided in it

An orbit around a subject standing in front of an out-of-focus wall reads as a cut-out sliding across a painted flat. What fixes it?

A. A negative prompt containing warping and morphing

B. A still with a foreground, a mid-ground subject and a background that recedes

C. A slower orbit, so the model has more frames in which to invent the space behind

D. Generating at the model's native aspect ratio instead of a cropped one

11 2. The first frame, and everything decided in it

Why does a visible light source in the frame — a window, a lamp, a screen — make the lighting more stable through the clip?

A. The picture contains an anchor for the light, so it drifts less
B. Bright regions are encoded at higher precision by the autoencoder
C. Prompt terms describing light are weighted higher when a source is present
D. The model computes light transport from any source it can see

12 2. The first frame, and everything decided in it

Which test tells you whether a flaw in a still should be repaired by hand or fixed by generating again?

A. Whether the still was generated from text or from another image as a reference
B. Whether the flaw is inside the safe area of the frame
C. Whether you can describe the fix as a shape you could draw around
D. Whether the tool offers inpainting for that region

13 2. The first frame, and everything decided in it

You have composited a real product into a generated still. Which failure makes it read as fake most reliably?

A. A slight difference in saturation between object and plate
B. No contact shadow where the object meets the surface
C. A cut-out edge that is one pixel too soft
D. Grain on the object that does not match the grain on the background plate

14 2. The first frame, and everything decided in it

Why should an end frame be built from the start frame rather than generated independently?

A. The tool will reject two images that do not share a seed
B. A derived end frame costs nothing, while a generated one is billed twice
C. Independently generated images are stored at different bit depths and colour primaries
D. Two independent generations are never the same scene, and the gap shows as churn

15 2. The first frame, and everything decided in it

In a frame library, why is the output file kept as the primary record rather than the seed and settings?

A. Storage is cheaper than the compute needed to regenerate a clip months later
B. Seeds are not portable between different tools
C. The others let you get close; the file is the only thing that is certain
D. The file contains the prompt and the seed in its metadata anyway

16 2. The first frame, and everything decided in it

A client wants a hero close-up of their own branded packet, and a generic dawn aerial over a city. Which is generative video already right for, and why?

A. The aerial: nobody studies it frame by frame and nothing has to be itself
B. The packet, because a close-up is a small, simple composition
C. Neither, because any work commissioned by a paying client belongs in a live shoot
D. Both, if you spend enough attempts on the packet

17 3. Directing, not describing

Which instruction is a model most likely to act on, and why?

A. "Her expression conveys a complicated mixture of relief and regret"
B. "She reacts emotionally to the news she has just heard"
C. "She turns her head to the left and looks down"
D. "A deeply moving, emotionally resonant performance"

18 3. Directing, not describing

Why should consecutive shots of the same subject differ in size by at least two steps?

A. A near-identical size reads as a jump cut, and a large change also hides drift
B. Models reframe more accurately at larger size differences
C. The autoencoder handles large subjects and small subjects with different fidelity
D. Small size changes cost more generations to match

19 3. Directing, not describing

Beyond directing attention, what does shallow depth of field do for a generated clip?

A. It makes the subject easier to track for later compositing
B. It lowers the effective motion strength, which reduces artefacts right across the frame
C. It reduces the number of tokens the model has to attend over
D. A soft wash of colour cannot visibly reorganise itself, so background churn disappears

20 3. Directing, not describing

Why does crossing the line happen so easily in generative work, and where do you prevent it?

A. Models mirror shots at random, so it cannot be prevented, only corrected later in the grade
B. Each shot is generated with no knowledge of the others, so direction goes in the stills
C. The 180-degree rule only applies to footage shot with a real camera
D. It is caused by the negative prompt suppressing the words left and right

21 3. Directing, not describing

You paint a motion brush over the steam above a cup and leave the rest of the frame unpainted. What should you expect?

A. The unpainted region is biased toward stillness, with some residual drift
B. The brush is ignored unless the prompt also describes the motion
C. Only the painted region is generated; the rest is copied from the first frame
D. The unpainted region is frozen exactly, frame for frame

22 3. Directing, not describing

You drag a trajectory arrow to move a person across the frame and get a figure that slides rather than walks. Why?

A. Trajectory conditioning is applied only to the final quarter of the denoise
B. The arrow was drawn faster than the clip's duration allows
C. Trajectory control specifies position, not pose
D. The prompt and the arrow were in agreement, which over-constrains the motion

23 3. Directing, not describing

Before generating from a pose sequence extracted from a reference clip, what should you always do?

A. Convert the reference to the model's native frame rate
B. Render the extracted signal and watch it as a video
C. Match the reference's aspect ratio to the output
D. Strip the reference's audio so it does not condition the generation

24 3. Directing, not describing

A 120-word prompt drops half its requirements. What is the most effective response?

A. Repeat the dropped requirements at the end of the prompt
B. Raise guidance until adherence improves across all of the clauses at once
C. Add emphasis weights to the clauses that were dropped
D. Cut to four clauses and move counts and spatial relations into the still

25 4. Making separate shots belong to each other

Why must every image on a character sheet share one light direction?

A. The autoencoder encodes evenly lit faces with less loss

B. Image models refuse to accept a set of references shot under different conditions

C. Mixed lighting confuses the face detector used by reference features

D. Every later first frame derives from the sheet, so a mismatch reaches every shot

26 4. Making separate shots belong to each other

You supply a reference photographed outdoors at midday and ask for an indoor evening shot. The result is oddly bright and cool. Why?

A. The model averages the reference's exposure against the prompt's description of the light

B. The conditioning encodes the whole image, not a disentangled notion of identity

C. Reference features override the first frame's exposure by design

D. Daylight-balanced images are encoded with a different colour primary

27 4. Making separate shots belong to each other

A character sheet wants one consistent light and background. A training dataset for an adapter wants varied ones. Why the opposite rule?

A. Variety is what teaches the model which parts of the image are the person

B. Consistent datasets overfit to the trigger word rather than to the face

C. The adapter learns lighting separately from identity in a second training pass

D. Training runs need more data variety to reach a stable loss

28 4. Making separate shots belong to each other

Every output from your new character adapter shows the same pose, the same clothes and the same three-quarter angle. What has happened?

A. The learning rate was set too low, so only the strongest features were ever learned

B. The rank was set too low to capture pose variation

C. It is overtrained: the model memorised the photographs rather than the person

D. The trigger word collided with an existing concept in the base model

29 4. Making separate shots belong to each other

When a location drifts between generations, what changes first, and what does that imply for the set?

A. Wall colour goes first, so choose neutral paint

B. Small objects move or vanish first, so dress the set sparsely

C. Doors and windows go first, so avoid openings in frame

D. Floor patterns go first, so shoot at a high angle to exclude them

30 4. Making separate shots belong to each other

Why is a finely striped shirt a worse costume choice than a plain saturated one?

A. Patterns are penalised by the aesthetic scoring the model was trained against

B. Striped fabric produces moiré whenever the clip is interpolated up to a higher frame rate

C. Stripes confuse the optical flow used by motion controls

D. The stripe is below the encoder's preserved resolution, so it warps and changes count

31 4. Making separate shots belong to each other

Why is the working answer "neutral stills, unified grade" rather than baking the look into every first frame?

A. Neutral frames generate faster because they contain less contrast information
B. Graded stills condition the model toward higher contrast output
C. It keeps the expensive stage flexible and the cheap stage decisive
D. A grade applied in the edit survives platform re-encoding better

32 4. Making separate shots belong to each other

Face replacement can rescue a drifted shot of your own character. What separates that from the misuse the technique is famous for?

A. Permission and disclosure, because the operation itself is identical
B. Whether the source face is generated or photographed
C. Whether the tool used is an open-source model or a hosted commercial product
D. The resolution at which the replacement is performed

33 5. The half of video that is not picture

A lip-synced clip reads exactly like badly dubbed foreign footage. Which failure is most likely?

A. The head is moving too much for the mouth region to be tracked reliably
B. The audio was compressed before it was fed to the model
C. Plosives are under-closed, so the lips never fully meet on b, p and m
D. The clip was generated at a lower frame rate than the audio

34 5. The half of video that is not picture

Your generated character shifts toward your own head shape after a performance transfer. What caused it?

A. The driver's proportions transfer along with the performance
B. The character was generated at a different aspect ratio from the webcam
C. The model re-sampled the identity because the take ran longer than four seconds
D. The transfer strength was set above its default

35 5. The half of video that is not picture

Synthetic speech holds up over a single line and tires a listener over three minutes. What is the boundary?

A. Voice models drift in timbre in much the same way video models drift in identity
B. Long passages exceed the context the text encoder can attend over at once
C. Audio quality degrades as the generated file gets longer
D. Prosody is derived from the text's surface rather than from understanding it

36 5. The half of video that is not picture

Which change to your script most improves a text-to-speech read, and why?

A. Writing in the present tense so the model chooses a livelier delivery
B. Punctuating for breath and breaking long sentences into short ones
C. Using the style or emotion preset that matches the content of the line
D. Generating a whole paragraph at once so the prosody stays consistent

37 5. The half of video that is not picture

Someone has a good USB microphone and a hard, empty office. What improves the recording most?

A. Moving to a bedroom with a bed and curtains, and recording close
B. Applying a stronger noise reduction afterwards in Audacity
C. Recording through the camera app so the picture and sound stay locked
D. Raising the input gain and moving further from the microphone

38 5. The half of video that is not picture

Why is the high-pass filter the first step of the voice cleanup chain rather than the last?

A. Noise reduction cannot build a profile from a file containing low frequencies
B. It reduces the file size before the slower steps run
C. It removes rumble that the later steps would otherwise amplify and spread
D. Audacity applies filters in menu order regardless of what you choose

39 5. The half of video that is not picture

You find a perfect track licensed CC BY-NC for a client's promotional video. What is the position?

A. Usable, if you credit the artist in the description
B. Not usable: a client's promotional video is a commercial use
C. Usable, because the client rather than you is the commercial party
D. Usable if the video is not monetised on the platform it is posted to

40 5. The half of video that is not picture

You are placing a breaking-glass effect. Where exactly does it go?

A. Two frames early, so the ear leads the eye
B. At the loudest point of the file, wherever that falls in the clip
C. At the start of the sound file, on the frame the glass leaves the hand
D. With the sound's transient aligned to the contact frame

41 6. From generations to a finished piece

Why are selects and assembly separate passes rather than one?

A. Assembly requires the final grade, which selects does not
B. Timelines can only hold one take per shot, so the selects pass must finish first
C. Editors are billed separately for each pass on professional jobs
D. You cannot judge a shot in the sequence while still judging the shot itself

42 6. From generations to a finished piece

A stringout is far too long and feels wrong. What is it for?

A. Estimating the export time for the finished piece

B. Showing which shots are missing, which are unnecessary, and where it sags

C. Checking that every clip shares one frame rate before the conform pass begins

D. Giving the client an early version of the cut to approve

43 6. From generations to a finished piece

Two shots with slightly mismatched light cut together better on movement than on stillness. Why?

A. The eye is already processing change, so the cut is one change among several

B. Editors' software applies a short automatic dissolve during motion

C. The codec allocates more bits to moving frames, which hides the difference well

D. Motion blur averages the two shots' colour across the join

44 6. From generations to a finished piece

What extra job do J-cuts and L-cuts do in a piece assembled from generated clips?

A. They let the grade change gradually rather than at the cut point

B. They reduce the number of clips that have to be colour matched

C. Overlapping audio disguises the fact that the shots were never in one place

D. They allow two clips generated at different frame rates to sit next to each other

45 6. From generations to a finished piece

Why match colour on scopes rather than by eye?

A. Scopes measure the values the codec will actually encode
B. Your eyes adapt in seconds, so a neutral shot looks blue after a warm one
C. Scopes read the clip before any grade is applied
D. Laptop screens cannot display the full range that the scopes are able to measure

46 6. From generations to a finished piece

A clip has visible compression artefacts of its own. Why denoise it before adding the timeline's grain?

A. Denoising after grain would remove the grain as well
B. The export codec allocates bits to the noise before it allocates them to detail
C. Grain and compression noise cancel out if applied in the wrong order
D. Uniform grain over an uneven mess looks worse than grain over a clean image

47 6. From generations to a finished piece

You track a real title onto a generated sign. Which step makes it stop announcing itself?

A. Corner-pinning it to the plane rather than placing it flat on the frame
B. Increasing the type size so it reads at a glance
C. Adding grain to the text so it shares the frame's surface
D. Animating the title with a slight drift in the same direction as the shot

48 6. From generations to a finished piece

Proxies make an old laptop usable for editing. What do they not solve?

A. Judging colour and fine detail while you cut
B. Scrubbing a long timeline without stalling
C. Keeping a multi-track audio mix in sync during playback
D. Real-time playback of several layers at once

49 7. Money, clients and running the job

Why does writing your per-generation cost in your own currency on a piece of paper change behaviour?

A. It lets you compare aggregators against the underlying model's own pricing page
B. It satisfies the record-keeping most clients require in a quote
C. Credits are designed to put distance between the click and the cost
D. Providers publish prices only in credits, so the conversion is otherwise unavailable

50 7. Money, clients and running the job

Beyond speed, why does queueing eight shots and reviewing them together produce better work than generating one at a time?

A. Comparing eight results shows which are good, not merely which is better
B. Batches are billed at a lower rate per second
C. A queue prevents the content filter from refusing a legitimate shot mid-session
D. Batched jobs are given a higher priority by most providers

51 7. Money, clients and running the job

What does driving generation from a script give you that a web interface does not, besides speed?

A. Guaranteed determinism across runs on a hosted provider
B. A lower price per second than the same model through its own product
C. Access to model versions that the interface does not expose
D. Every parameter recorded, because it is in the script

52 7. Money, clients and running the job

A client asks for shot nine to be warmer, and for shot twelve to be a close-up instead of a wide. How should a quote treat those two requests?

A. Both are revisions and both are billable at the per-shot rate

B. The first is a note and free; the second is a re-shoot and billed

C. Both are included if they fall within the attempt allowance

D. Neither is billable once the animatic has been approved in writing

53 7. Money, clients and running the job

Which brief should a freelancer decline on day one, with an alternative rather than a shrug?

A. A forty-second dialogue scene across seven matched shots

B. A corridor cutaway under a voiceover

C. A previsualisation sequence to show a client before a shoot

D. A dawn aerial over a city the client cannot fly a drone over

54 7. Money, clients and running the job

A client asks for a three-day turnaround on a job quoted at six. What is the honest offer?

A. Accept the deadline and raise the per-shot price to cover the additional risk involved

B. Accept and absorb the extra attempts to keep the relationship

C. Name the trade: fewer attempts per shot, so more shots land merely acceptable

D. Refuse, because generation time cannot be compressed at all

55 7. Money, clients and running the job

A post-mortem records the reason for every rejected take. What is that tally actually for?

A. Proving to a client that the attempt allowance was used fairly
B. Showing where to spend improvement effort, which is rarely where you assumed
C. Establishing which model to use on the next project
D. Supporting a refund claim with the provider for every failed generation in the batch

56 7. Money, clients and running the job

Why does a quote include a clause about the client warranting consent for any real likeness or voice they supply?

A. Platforms require the clause before a video can be uploaded
B. It substitutes for the written consent forms in the project folder
C. It transfers every part of the legal responsibility away from you permanently
D. It makes you not the party who failed to ask, and raises it in time

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. Nobody is hiding the long version from you — B

A — Attributes a compounding image-quality loss to instruction-following, when the prompt is re-applied identically at each step.

C — Invents a fallback mechanism and confuses a training window with a different technique entirely.

D — Reaches for a billing behaviour where a signal-processing mechanism explains the result.

2. Your clip is compressed fifty times before anything happens — B

A — Treats a representational limit as a convergence problem, and invents a step count at which a discarded detail reappears.

C — Assumes the failure is in conditioning rather than in the pixel budget; a perfectly specified sign fails identically.

D — Names a real mechanism that causes other drift, but text shimmers even inside a single attention window because the detail was never in the latent.

3. The whole clip is made at once, not frame by frame — D

A — Blames the interface for an architectural property; a local run has the same block and the same problem.

B — Invents a progressive per-frame refinement order that flow-matching video diffusion does not use.

C — Half right about the decoder but wrong about the latent: no group of latent frames finishes ahead of the others.

4. Why a chair stays a chair and a face does not — A

B — Assumes a window boundary landed exactly at the occlusion; the same failure occurs well inside a window.

C — Imagines the text conditioning weakening over time; cross-attention applies equally to every frame in the block.

D — Misapplies the VAE limit, which affects fine detail rather than the colour of a large object.

5. The prompt vocabulary was written by whoever captioned the training clips — D

A — Assumes montages were merely rare rather than structurally absent; scene segmentation removed them entirely.

B — Blames captioning for something segmentation had already decided before any caption was written.

C — Describes a plausible-sounding constraint, but attention would happily represent a discontinuity if the data had contained any.

6. Same settings, different clip, and why that is not a bug — C

A — Invents a mechanism; the seed is usually a genuine RNG seed, which is exactly why the failure is surprising.

B — Treats sampling as irreducibly random; a fully pinned local run repeats bit for bit.

D — Names a real but secondary effect that would usually produce a near-identical clip rather than a visibly different one.

7. The four numbers, and which of them you should touch — A

B — Frames the failure as adherence succeeding; the figure is not walking, so adherence is worse, not better.

C — Confuses an out-of-distribution extrapolation with an under-converged one; more steps sharpen the same frozen clip.

D — Invents a threshold behaviour; the unconditional prediction is used at every scale, and it is the extrapolation away from it that causes the damage.

8. Generated at sixteen, delivered at twenty-four — B

A — Attributes a post-generation artefact to the VAE, which compresses uniformly and had finished long before interpolation ran.

C — Invents a multiple-frame blend; the problem appears at integer multiples too, and only at boundaries.

D — Blames generation for an artefact that reappears when you interpolate real camera footage of the same move.

9. The ranking will be wrong by the time you read this — A

B — Assumes a leaderboard measures everything a model is good at, so the fault must lie with the user.

C — Reaches for a settings explanation where a difference in what was measured accounts for it.

D — Substitutes bad faith for a straightforward difference in measurement scope.

10. What it takes to run one on your own card — D

A — Describes a dequantisation that some runtimes do per-layer but not for the whole model at once, which would make fp8 pointless.

B — Overstates the text encoder, which is comparatively small and can be unloaded after the prompt is encoded.

C — Names a real overhead of a few hundred megabytes and inflates it into the whole explanation.

11. The still frame is where the decisions belong — C

A — Treats a representational limit as an artefact-suppression problem that a parameter can clean up.

B — Assumes text failure is a sampling and detail problem rather than a symbolic one.

D — Moves a decision the team can fully control into one they cannot control at all.

12. Compose for the frame you will end on — A

B — Treats a missing-information problem as a speed problem; a slow orbit reveals the same absent space more gradually.

C — Assumes an explicit rotation axis exists; the model has no geometric pivot to get right or wrong.

D — Blames defocus, but a sharp single-plane backdrop orbits just as badly — the problem is the absence of recession, not of detail.

13. One light direction per scene, written down — C

A — Overstates the limitation; a shot match does help, and secondaries can isolate a subject when needed.

B — Assumes shadow direction is a colour property; it is geometry baked into the pixels.

D — Inverts which problem is tractable; lifting shadows flattens the face without moving where the light comes from.

14. Fix the frame yourself instead of re-rolling it — C

A — Names a real but minor cost; stills are cheap, and the argument would collapse if they were free.

B — Confuses regeneration with re-encoding a decoded output; a fresh generation from the same prompt is not a second-generation copy.

D — Assumes the flaw is deterministic; hands vary considerably between samples, which is exactly why re-rolling tempts people.

15. Put the real object in before you animate it — B

A — Names a real mismatch that is usually fixed by downscaling, and which rarely produces the floating quality described.

C — Inverts the usual error: composites more often fail from edges that are too hard and clean, not too soft.

D — Identifies a genuine tell, but the stated scenario has everything matched except the anchoring, and grain alone rarely makes an object float.

16. Deciding where the shot lands — C

A — Blames the seed; two frames from the same seed but different scenes fail identically.

B — Picks out one real difference and invents a specific dissolve behaviour for it; the failure occurs with matched lighting too.

D — Invokes attention range, but both anchors condition the whole block regardless of how different the scenes are.

17. A camera move with no video model at all — B

A — Names a real depth-map weakness that produces banding and terracing rather than the directional stretching described.

C — Mesh density affects smoothness of the surface but cannot supply pixels for an area that was never photographed.

D — True about relative depth and irrelevant here; rescaling the displacement changes the amount of parallax, not the missing information behind the shoulder.

18. Predicting which second will break — C

A — A genuine weakness, but rugs often survive a short static shot and a reframe removes the problem cheaply.

B — Also genuine, and usually solvable by reframing the window out of shot rather than by regenerating.

D — Invents a data-scarcity claim; the body of a violin is a large smooth form and is among the things models render most reliably.

19. Every still you keep is a part you do not have to make again — C

A — Overstates current disclosure requirements, which generally ask whether content is synthetic rather than which version produced it.

B — A real bookkeeping use, but billing records already carry it and it would not affect whether a revision matches.

D — Describes a plausible mismatch, but the decode has already happened — the delivered frames carry whatever texture they carry regardless of what you record.

20. Where this is already the right tool — C

A — Treats the product as the sole obstacle while ignoring that the actor's likeness and shot-to-shot continuity are the harder constraints.

B — Assumes generated and filmed elements will match in light, lens and grain without the continuity control that is precisely what is missing.

D — Over-corrects from a real limitation on the finished deliverable to a blanket exclusion from the production process.

21. Cinematic, 8k, epic gets you the average of everything — D

A — Locates the fault in a token limit rather than in what the words were attached to in training.

B — Treats a vocabulary problem as a slider problem, so the fix becomes a setting rather than a rewrite.

C — Confuses a model's stylistic default with the effect of averaging over thousands of vaguely captioned clips.

22. The two words that decide most of the frame — B

A — Invents a directional asymmetry; drift is equally visible whichever way the progression runs.

C — Produces a sequence of jump cuts, and a constant scale makes the face easiest of all to compare between shots.

D — Solves identity at the cost of every shot sharing one pose and one moment, which reads as a zoom rather than a scene.

23. Focal length, depth of field, and the words that carry them — B

A — Invents an adaptive token budget; the latent grid is uniform regardless of what is in focus.

C — Reverses the causality: the encoder discards that detail anyway, and nothing is preserved by blurring it first.

D — Probably true and not the point; the benefit is visible on any model regardless of how its corpus was filtered.

24. Where people stand, and which way they face — C

A — Describes a different error; the subjects are different people, so this is not a jump cut.

B — Underrates a spatial cue that audiences read pre-consciously; dialogue does not repair a crossed line.

D — Names a real consequence of reversing a camera on a set, but here the stills were generated independently and lighting was never tied to camera position.

25. Taking motion out of language — B

A — Treats residual drift as a masking error; tightening the mask does not convert a bias into a freeze.

C — Assigns drift to the wrong stage; the encoder is lossy but it is the generation that moves the background.

D — Assumes the mask sets a hard zero and then invents a post-processing explanation for what is plainly generated motion.

26. Depth, pose and flow as instructions — B

A — Contradicts the premise, which states a pose sequence was used; flow would distort the whole frame, not the limbs alone.

C — Names a real sampling issue that produces jerky timing rather than sustained limb stretching.

D — Overstates the conditioning's authority; the still still governs appearance, which is precisely why the conflict resolves as deformation.

27. What a negative prompt can and cannot remove — C

A — Amplifies the whole extrapolation, which burns contrast and freezes motion long before it removes the animal.

B — Broadens a weak subtraction and risks removing wanted animals and fur textures without reliably removing the dog.

D — Correct that the positive branch is stronger, and wrong about negation: naming the concept positively activates it more reliably still.

28. An experiment protocol that fits on a card — B

A — Tests a different variable and leaves the original confound — the change of wording — entirely unexamined.

C — Close to right and backwards: it tests the old wording on a draw that has not yet been shown to be favourable.

D — Changes a second variable in an experiment whose problem is that too many variables moved at once.

29. The clauses that quietly do nothing — C

A — Applies more pressure to two requirement types that are weak for structural reasons rather than for lack of emphasis.

B — Adds tokens, which dilutes every other clause; length is the opposite of the fix here.

D — Names the adherence control correctly and ignores that counts, spatial relations and negations resist it while motion and colour degrade.

30. A cut breaks on the light before it breaks on the face — B

A — Reaches for an unfalsifiable explanation instead of the visible continuity error sitting in front of them.

C — Reaches for a technical conform issue where a physical continuity error is what the eye is reporting.

D — Treats a physical continuity error as a grading step, which would unify tone without fixing a light coming from the wrong side.

31. The document you build before you generate anything — C

A — Quotes the dataset size for training a character model, which is a different technique with different requirements.

B — Picks a genuine but secondary property; varied framing across the sheet is useful, varied lighting is not.

D — Attributes identity to the seed; a shared seed with a different prompt produces a different face, and a different seed with a strong reference can produce the same one.

32. What a reference image is actually doing — A

B — Invents an ordering rule; the two are combined, and the prompt's lighting is weakened rather than overridden outright.

C — Invokes gamut and weighting in a way the encoder does not work, and would predict the same problem in reverse, which does not occur.

D — Names a plausible dataset bias, but the effect persists on models with different filtering and tracks the specific reference rather than a general preference.

33. When a LoRA is worth the afternoon — C

A — Low rank limits expressiveness but produces a weak, vague identity rather than memorised training scenes.

B — Half right about captioning: describing the constant is the error, but it leaves identity vague rather than producing memorised scenes.

D — Describes overtraining in plausible terms, and overtraining on a varied dataset still yields varied output — the constraint here is what the data contained.

34. Building a room that stays the same room — C

A — Misunderstands the encoder, which compresses to a fixed latent size regardless of the source resolution fed in.

B — Reverses what happens: oversized stills are downscaled on the way in, so the surplus pixels never reach the model.

D — Describes a real workflow preference about where to upscale, which is a separate question from why the plate is generated oversized in the first place.

35. The blue that is not quite the same blue — A

B — Inverts the known asymmetry: face-processing machinery is unusually intolerant of small geometric change, not tolerant of it.

C — Assumes the drift magnitudes differ; both drift by similar proportions, and it is detection rather than magnitude that separates them.

D — Correct that flat regions encode well, but the decoder does not smooth facial drift away — it reproduces it, and viewers still notice it more slowly.

36. Fixing a face after the clip exists — B

A — Blames the codec for an artefact that is present in the uncompressed frames before any encoding happens.

C — Names two real steps and invents a causal link; flicker appears whether or not interpolation is applied.

D — Describes a genuine secondary contributor that is usually smoothed away by tracking, while the independent reconstruction of detail remains regardless.

37. Deciding the look once, in writing — B

A — Names an unrelated encoding consideration that does not depend on when the grade was applied.

C — True about prompt reliability and beside the point: the look can be baked into a still deterministically by grading the still itself.

D — Overstates a small encoding shift into the main reason, when the decisive factor is how cheaply the decision can be revised later.

38. The arithmetic nobody puts in the demo reel — C

A — Misreads the probability as a one-shot gamble; in practice you regenerate failures rather than accepting the first run.

B — Sets an arbitrary threshold and points at the one variable that is hardest to move, rather than at the number of exposed shots.

D — Treats a shared seed as a way to correlate outcomes across different prompts and stills, which it is not.

39. Six things it cannot do, and the reason for each — A

B — Believes a description can substitute for a reference when the target is one specific physical object.

C — Treats a missing-reference problem as a detail problem that more pixels can recover.

D — Confuses conditioning on reference images with training, and assumes either would hold an exact geometry through a full rotation.

40. Lip-sync is easy; the permission is not — C

A — Confuses consent to record with consent to synthesise, treating one signature as covering any later use of the same person.

B — Makes the truthfulness of the words the test, when the act of manufacturing her speech is what needs permission.

D — Collapses two separate duties, treating a transparency obligation to viewers as if it supplied the subject's permission.

41. What the model is actually reshaping — A

B — Names a real audio processing risk; the closure failure occurs on well-levelled audio too, because it is a reconstruction limit.

C — Those three phonemes share one viseme, so ambiguity between them cannot produce a wrong shape — the shape is the same.

D — Interpolation invents frames from existing ones and cannot add a closure the source never contained.

42. Acting is timing, and timing has to come from somewhere — D

A — Names a real resolution difference in many implementations, which would not explain why the performance reads as more convincing rather than merely sharper.

B — Sync offset is a genuine failure mode, and the premise specifies excellent audio with correct timing.

C — Describes the mechanism accurately and stops one step short: the question is what the second method supplies, not only what the first omits.

43. Synthetic voice, and where it stops being good enough — A

B — Invents a spectral deficiency; modern synthesis reproduces the full range, and the fatigue persists after any equalisation.

C — Long-form instability is real in some models, and it produces audible wobble rather than the flatness described here.

D — Identifies one symptom of the real problem and presents it as the whole cause, when correcting the punctuation still leaves the emphasis unmeant.

44. A phone, a quiet room, and four minutes in Audacity — B

A — Some processing exists, and it would not overcome a room problem — the same phone in the bare office sounds worse too.

C — Sensitivity affects the noise floor rather than the hollow quality, and a quiet bare room still sounds hollow.

D — Proximity genuinely helps and is part of the advice, but the stated comparison is between rooms, and a close USB microphone in the bare office still sounds worse.

45. Why silence is the worst sound — A

B — Attributes the effect to distraction rather than to inference; viewers still see the cuts and read them as edits within one space.

C — Describes a real levelling benefit that would apply equally to six separate ambience clips, which do not produce the same cohesion.

D — Correct about silence reading as a fault, and it explains why silence is bad rather than why one unbroken bed unifies the shots.

46. Music you could defend if somebody asked — D

A — Invents a placement requirement; the licence specifies attribution, not where it must appear.

B — Fabricates a restriction; CC licences are routinely and validly applied to recordings.

C — Describes the SA condition accurately and attaches it to a licence that does not carry it — BY-NC has no share-alike term.

47. The viewer believes the sound — D

A — Explains the effect as distraction; viewers watch the contact and now perceive it as correct rather than missing it.

B — Borrows a real masking phenomenon from audition and applies it to vision, where no such suppression accounts for the effect.

C — Material identification is part of what the sound conveys, and the perceptual integration happens below the level of any inference about materials.

48. When the model makes the sound too — D

A — Invents a sync-drift mechanism; the audio is generated against the same clip and stays locked to it.

B — Overstates the disconnection; the audio is generated jointly with the picture and is usually well matched to it within a clip.

C — Asserts a terms-of-service distinction that is not generally how these products are licensed.

49. The generator makes footage; the editor makes the video — B

A — Assumes resolution is what reads as finish, when upscaling adds a sharpening signature and no information.

C — Treats the problem as per-clip quality when the failure is the mismatch between clips.

D — Blames the cuts for a discontinuity that lives in colour and texture, and buys more drift in exchange.

50. Decide the frame rate before you import anything — A

B — Solves one clip by breaking the others, and assumes the delivery spec is negotiable when it is usually not.

C — Aligns the frames arithmetically and changes the speed of the action, which is a different shot rather than a conform.

D — Variable frame rate is a real nuisance from some sources, and a constant 30 on a 24 timeline still has to lose a frame in five.

51. Choosing takes before you start cutting — D

A — Reverses the relationship: a stringout is always far longer than the finished piece and tells you little about final duration.

B — A real incidental benefit that belongs to the selects pass, where each take was already watched in full.

C — Colour comparison is genuinely easier on longer clips, and it is a later pass whose decisions do not depend on structure being settled.

52. Cut on motion, and one frame before the failure — A

B — Motion blur does soften detail, and the colour difference between the shots is unchanged by it.

C — Treats a cut as having duration; the cut is instantaneous, and what varies is how much competing change surrounds it.

D — Invents a specific suppression mechanism; the effect holds when the mismatch is in the centre of frame too.

53. Thirty renders, one world — D

A — Chained matching accumulates drift across thirty clips and still lacks a common reference to return to.

B — Names a real workflow inconvenience that affects revising the work rather than why the result looks inconsistent.

C — Describes a genuine reason eyes are unreliable, and the inconsistency would remain even with perfect vision because no common ground was established.

54. Putting back the texture the encoder removed — A

B — Would be true at dissolves, which are a small fraction of the timeline and not where the problem is reported.

C — A real ordering consideration that is solved by grading first, and which does not explain why one global pass succeeds where thirty matched ones fail.

D — Names a genuine inefficiency that has no bearing on how the finished footage looks.

55. Real type, added afterwards — C

A — Resolution helps legibility and would not stop identical letterforms from being reinvented each frame if they were generated.

B — Compositing order is real and irrelevant; grading does not cause generated text to drift.

D — Correct about how fonts are stored and wrong about the cause: codec quantisation is not what makes generated text melt.

56. Slow motion is where fake physics goes to be caught — C

A — Flow artefacts are real and secondary; the shot also fails when slowed by frame blending or by generating at a higher frame rate.

B — Names one real contributing factor, and the shot still fails when the audio is re-timed to match perfectly.

D — Assumes the model varies its frame output by event speed; it generates a fixed number of frames regardless of what happens in them.

57. The last render, and what the platform does to it — D

A — Invents a proportional allocation; platforms encode to their own targets regardless of what arrives.

B — Attributes a detection-and-penalty behaviour to platforms that they do not have.

C — Describes rate control in plausible detail and overcomplicates the point, which is simply that lossy compression applied twice loses more than once.

58. The whole pipeline on a phone or an old laptop — A

B — Helps somewhat and still decodes the full-resolution source files, which is where most of the load actually is.

C — Genuinely eases decoding and produces very large files that then strain disk throughput on a modest machine.

D — Background rendering helps with effects and leaves raw playback of untouched clips just as heavy.

59. An afternoon of curiosity can cost more than a day of work — D

A — Reduces unit cost while leaving the unbounded number of revisions that produced the overage untouched.

B — Prices the symptom into every job instead of capping the driver on the job that has it.

C — Believes a subscription makes generations free rather than pre-purchased, capped and expiring.

60. The document everything else is built from — A

B — Conforming is real and is a minor labour cost rather than the thing that changes a budget.

C — A sensible use of the column that follows from the main point rather than being it.

D — The column genuinely helps with provenance, and disclosure obligations do not determine what the project costs to produce.

61. Get the yes before you spend the money — C

A — A still held on screen shows nothing about how it will behave under motion, which is what drift describes.

B — Genuinely useful and available from laying the stills side by side, without needing them on a timeline at all.

D — The count was already on the shot list, and the animatic adds timing rather than a number that was previously unavailable.

62. Test the idea at a tenth of the price — B

A — Resolution affects spatial detail; motion range is a property of the model and its training, not of the latent's width.

C — Shorter drafts do truncate motion, and the stiffness described is a character of the output rather than an absence of time.

D — Seed differences make individual takes incomparable, while systematic properties of a model still transfer across draws.

63. Queue everything, then judge it together — B

A — Batching does not impose a shared seed, and eight shots with different stills would not be comparable even if it did.

C — A real provenance concern that batching does not actually protect against over a long queue.

D — Asserts a batch discount that these services generally do not offer.

64. The arithmetic nobody does before subscribing — C

A — A genuine and important clause to check, and it does not address the pricing question that was asked.

B — Relevant to whether the week is workable and unrelated to whether the plan is cheaper.

D — Describes a real overage mechanic that reduces the saving on part of the usage rather than reframing the comparison.

65. Price the thing that actually costs money — C

A — Describes the commercial effect on both parties without touching what the client learns from it.

B — The cost is only fixed within the allowance; beyond it the variability returns, which is the point of stating it.

D — A real convenience of per-shot pricing that concerns future projects rather than what the allowance communicates in this one.

66. What to keep, and the day it saves you — C

A — True of a graded export and not of the raw generation, which is what is being archived at this stage.

B — Transcription errors are avoidable by copying, and a correct prompt still fails to reproduce the shot on an updated model.

D — Correct that a seed needs its version, and the version can be recorded — the deeper problem is that even a complete record can stop reproducing.

Module test — 1. What the machine is actually doing

1. A

Generation starts with the entire latent block full of noise and refines every latent frame together, step by step. At step seven, the first frame and the last frame are both seven steps along. A preview is a partially denoised block being decoded, which is why it looks like fog. The same fact explains why you cannot add a bit more onto the end and why the cost is fixed before any picture exists.

2. C

The first frame is given to the model. Every later frame is inferred, anchored only by the frames before it rather than by anything fixed, so error accumulates with distance from that anchor. The practical response is to generate long and deliver short: treat the first quarter-second and the last second as offcuts, the way a printer treats bleed.

3. D

The autoencoder compresses by roughly eight times in each spatial direction, so anything finer than about an eight-pixel block at output resolution is not carried. A letterform is exactly that fine. The decoder produces a plausible replacement independently for each frame, which is why text melts. More steps sharpen what the model decided to depict; they cannot restore information the encoder never kept. Text goes on afterwards, in the editor.

4. B

Guidance extrapolates past the conditioned prediction, away from the unconditional one. Pushing harder drives every frame toward the strongest single interpretation of the prompt, and the strongest interpretation does not change over time, so motion freezes while contrast and saturation climb. Guidance cannot supply an instruction the model did not act on. The fix is the first frame or the phrasing.

5. C

A seed fixes the starting noise and nothing else. Reproducibility also requires identical weights down to the checkpoint, identical precision, step count, scheduler, tokenisation, batch shape and GPU kernels. A hosted endpoint batches with other requests, may route to a different GPU generation and may update the model, and providers generally reserve the right to all three. This is why the output file itself is the primary record, not the seed.

6. A

Three things occupy VRAM at once: the weights, the activations the attention layers hold during each forward pass, and the full-size frames allocated when the latent is decoded back to pixels. The decode of a 1080p clip in one go can peak higher than the generation did. Leave three or four gigabytes over the weights, or use a tiled decode, which removes almost all of that spike.

Module test — 2. The first frame, and everything decided in it

1. D

A still can be opened in an editor. You can clone out a stray wire, patch a bad hand from another generation, composite a real product in, and paint on a mask. None of that exists inside a video model, where your only lever is another sample. Whatever you fix in the still is fixed for the whole clip, and the repair costs one attempt rather than a fresh draw over the entire frame.

2. B

The conditioning image has to match the shape of the latent the model generates into, so a mismatched still is fitted by crop, pad or stretch before the model sees it. The composition you chose is not what gets animated. Generate the still at the ratio the video model outputs, and at or a little above its native resolution: a 4000-pixel still is downscaled on the way in, so those pixels were paid for and deleted.

3. C

Regenerating is a new sample. It may fix the hand, and it may also move the light, change the face, lose the wardrobe detail you liked and add a flaw somewhere you were not looking. You are rolling dice on the whole frame to fix one square inch. A clone stamp or a patch changes that square inch and nothing else, costs nothing, and gives you what you decided rather than what the sampler decided.

4. A

Both frames are encoded and both condition the denoise, one anchoring each end of the block, and everything between is generated to connect them. When the end-points are close — same room, same light, a plausible camera displacement apart — the smooth latent path corresponds to a smooth path through the real world, and you get a camera move. When they are far apart it corresponds to nothing, and you watch one image dissolve into another.

5. D

A depth projection moves a virtual camera over the original pixels, so the subject cannot change: no drift, no re-sampling, no morph, and no per-second bill. The limit is disocclusion — behind a shoulder the photograph recorded nothing, so a large move stretches the nearest pixels across the gap. A push-in of ten to twenty per cent is invisible. Generation is for when the world has to move, not the camera.

6. B

When a camera moves, new picture enters at the leading edge and the model has to invent it, because nothing existed outside your frame. A pull-out invents a ring of scene on all four sides; a tilt up invents a ceiling or a sky; an orbit asks for the scene from an angle it was never given. A push-in only crops, which is why it is the safest move in the category and why an under-specified prompt tends to return one.

Module test — 3. Directing, not describing

1. C

Training captions were written by a vision model describing what it could see, so they name shot sizes, moves, light and lens. Mood adjectives are not observations, and they appear only where human-written text leaked in, scattered across wildly unrelated footage. Ask for them and you get the average of that scatter — which, after aesthetic and motion filtering, is gentle stock footage with a commercial grade.

2. A

A preset is a look, not a shot: it applies the same move whether or not the frame supports it. An orbit asks the model to show the scene from an angle it was never given, and if the still is a subject in front of a blurred wall there is no space behind them to reveal, so the model invents one as it goes. The fix is in the still — foreground, midground and a background with real recession.

3. D

A negative prompt replaces the unconditional prediction with one conditioned on that text, and the model is pushed away from it. Text conditioning does not carry negation, so the token for the concept activates the concept, and one noun against a whole conditioned prediction is a small vector. The reliable way to remove an object from a frame is to remove it from the frame: repair the still.

4. B

A negative list containing motion blur, movement or camera shake pushes the generation away from the whole region of the space that contains movement. It is a common and quietly expensive mistake, usually inherited from a long list copied from an image-model forum. Keep the negative list to four to eight terms about rendering failures — warping, morphing, flickering — and nothing about objects or kinds of motion.

5. A

Counts sit near the bottom of the adherence order, along with spatial relations and negations, because a few hundred text tokens are steering a quarter of a million video tokens and nothing in the model is counting. More words dilute every other token. Moving the requirement into the still turns it from a request into a fact, and the image wins every disagreement with the text.

6. D

Generation is stochastic, so two clips that differ in more than one respect carry no information about which difference mattered. Without a fixed seed you are comparing two random draws. Fix the seed and the still, change exactly one thing, write down what you expect before you look, and when a result is good, run it once more on a second seed. That extra generation is the difference between a finding and a superstition.

Module test — 4. Making separate shots belong to each other

1. D

A key light on the left in one shot and the right in the next reads instantly as wrong, and so does a shirt that has shifted from a greenish blue to a purplish one. Audiences cannot name either and usually blame the face. Decide the light direction once per scene, write it down, put it in every first frame, and compare wardrobe as colour blocks by blurring the stills and looking at them side by side.

2. B

A reference encodes your images into a vector and pulls generation toward a region, which reproduces a type — build, colouring, wardrobe, general impression — while the exact facial geometry re-samples within that region each time. A small trained adapter teaches the model the identity as a concept, so it behaves like anything else the model knows: it can be posed, lit, dressed and angled, and it stays itself.

3. A

A face has to be regenerated for every shot. A room does not: generate it once, wide and at a resolution above the video model's output, and derive every other shot in the scene from that one image by cropping, reframing and compositing. That is how a real production uses a set. The hard case is the reverse angle, which is new information — avoid needing it, or block the room out roughly in Blender.

4. D

Restoration reconstructs plausible detail — skin texture, eye highlights, teeth — independently for each frame, and the invented detail differs slightly every time. The result is a face that crawls. Lower the restoration strength, blend the restored sequence back over the original at partial opacity, or use a video-aware restorer. Watch the result in motion, never as a still, because a still is where it always looks good.

5. C

Per-shot rates multiply: 0.7 to the power of eight is about six per cent. You will not attempt eight in a row, you will generate each until it works, so what the arithmetic really prices is the takes. You have little control over the rate and complete control over the exposure. Over-the-shoulder shots, wides where the face is forty pixels across, and cutaways all carry the scene without exposing the hard thing, and they are better editing anyway.

6. B

Models mirror compositions freely, and a flip is invisible in a single shot and glaring across a cut. Write the sides down in the character file — bag on which shoulder, watch on which wrist, hair parted which way — and check every first frame for a scene side by side before generating any of them. It is the single highest-yield item on the continuity list and it costs two minutes.

Module test — 5. The half of video that is not picture

1. A

Consent to appear covers the words that were actually said, on that occasion, for that purpose. Synthesising new sentences produces statements the person never made and never reviewed, with their authority attached. Permission has to be in writing and specific: what will be generated, where it will appear, for how long, and whether new sentences may be produced. The ease of cloning does not make anything about it automatic.

2. D

These models crop the lower face, generate a mouth matching the audio at the crop's resolution, and blend it back. Teeth, tongue and the dark space behind them are the hardest part of that reconstruction, and a close-up shows all of it, at a size where the upscaled crop also leaves a visibly softer rectangle. Framing at chest height is not a style choice in AI-presenter content; it is the thing that makes it work.

3. C

Lip-sync derives mouth shapes from audio and changes a small rectangle, so the mouth speaks and the face does not. Performance transfer drives the whole facial performance from a recorded one: where somebody paused, which syllable got weight, the eyebrow on a question, the blink at a clause boundary. Delivery is most of acting and no phoneme model invents it, which is why the person in front of the webcam matters more than the model.

4. B

True digital silence is an experience nobody has outside a laboratory, so it reads as a fault — and generated clips arrive with no audio at all. A single unbroken bed removes the silence and, because it does not change at the edit points, glues shots together: a sequence of unrelated generations over one room tone is perceived as one place. It is the cheapest perceived-quality gain in the whole pipeline.

5. A

Perception combines sound and vision involuntarily, and where vision is ambiguous the ear resolves it. A sharp, correctly-timed break placed on the exact contact frame makes a physically wrong shatter read as correct. Slow motion destroys the effect: the ambiguity is resolved visually because the viewer has been given time to check, which is the mechanism behind never putting generated physics in slow motion.

6. C

A mixed track is the end of an audio process, not the middle of one. You cannot lower the ambience under the dialogue, swap a wrong footstep, extend the bed across a cut, match it to the next shot or translate the line. Worse, each clip's audio stops at its own boundary, so every edit point becomes the digital-silence problem. Generate the picture and build the sound, or accept it for single-clip social content and previsualisation.

Module test — 6. From generations to a finished piece

1. C

A timeline has one frame rate and anything that does not match is resampled by dropping or repeating frames. That is judder, and it is invisible on a static shot and obvious on a slow pan. Decide the project rate before importing anything, check each clip with ffprobe, and conform deliberately — frame blending as a safe default, optical flow where the motion needs it, per clip rather than globally.

2. B

Motion starts slightly wrong and corrects, which is the settle, and the face loosens and the background churns near the end, which is the drift. Both are properties of how the clip was made rather than of the shot you wanted. Cutting them is the same thing an editor does with real footage: end the shot before it stops working. Generate longer than you need so the trim is free.

3. A

Stage one brings every clip to a neutral common ground — same black level, same white point, same rough contrast — and its output is meant to look flat and boring. Stage two puts one look over everything at once. Applied in the other order, the look amplifies whatever inconsistency remains, because a curve maps two different inputs to two more different outputs. Match black level first, read off the waveform rather than off your eyes.

4. C

Grain works because every clip came through the same lossy encoder and lost its fine texture, so one layer gives them all the same new surface, which is a kind of matching that colour work cannot do. Applied per clip you get thirty slightly different textures and the difference shows at every join, which is worse than no grain. Set the amount so you only see it on a paused frame, and size it for the delivery resolution.

5. D

There is no simulation under a generated splash — no mass and no conservation of anything — only frames that look reasonable in sequence. At speed the viewer gets a fraction of a second and the sound carries the event. Slowing it hands them time to check, removes the sound's power to resolve the ambiguity, and points their attention at exactly the frames the model was least sure about.

6. B

Viewers never see your file; they see the platform's re-encode of it. Your export is source material for somebody else's encoder, so it should be generous. Exporting small throws detail away before that encoder has seen it, and the loss lands exactly where generated footage is weakest: grain, fine texture and gradients in skies. For 1080p, 15 to 25 megabits, H.264, yuv420p, and audio at 320 kilobits.

Module test — 7. Money, clients and running the job

1. A

The figure that matters is not price per second but price per usable second. At one in six, six generations were bought for one shot that shipped. That is the number a quote has to be built on, and it is why credits are the wrong unit to think in: work out what one generation costs in your own currency, multiply by your measured takes per shot, and write both numbers on paper beside the screen.

2. C

Without a kill number one stubborn shot quietly eats the budget for eight others, and it is always a shot that did not matter. With one, the failure is bounded and the decision is forced while you still have money. All three permitted moves are upstream of the generator: regenerating a still costs on the order of one per cent of regenerating a clip, and almost every clip failure has a cheaper diagnosis one step earlier.

3. D

The comparison people run in their head is the subscription's price per second against the top-up's, which ignores that a subscription charges you whether you generate or not, and that unused credits generally expire at the end of the period rather than rolling over. Compare the monthly fee against the top-up price times your actual measured usage from your last two real projects, not your optimistic estimate.

4. B

On a typical thirty-shot piece, a realistic breakdown is eighteen generated, four derived from a reused plate, three parallax moves on stills, two filmed on a phone and three graphics. Twelve of the thirty cost nothing to generate and nobody watching can tell which were which. A list without that column quietly quotes a generation price for a title card and hides where the budget actually goes.

5. C

An animatic is the stills you already made, held for their intended durations, with a scratch voiceover from your phone and temporary music. It costs almost nothing and it shows what a list cannot: what it feels like to sit through, where the pacing fails, which shots are missing and which are doing nothing. Every shot deleted at that stage is a final generation never paid for.

6. A

A seed fixes the starting noise and nothing else, and a hosted provider may batch, reroute or update the model between your two runs. The prompt, the seed and the version string let you get close. The file you were given is the only thing that is certain, which is why it is kept at the highest quality you received rather than treated as a fallback.

Video With AI — final examination

1. C ***1. What the machine is actually doing***

The model generates the clip as one object so that frame ninety can agree with frame twelve about where the wall is. Doubling the duration roughly doubles the tokens and roughly quadruples the attention work, and the whole latent block has to sit in GPU memory while it is denoised. Past a point it does not fit at all. Five to ten seconds is where quality, price and memory currently meet.

2. A ***1. What the machine is actually doing***

Extend takes the last frame of the previous generation — already decoded out of the latent and therefore already lossy — and re-encodes it as the conditioning image for a fresh run. That is a photocopy of a photocopy. Contrast climbs, saturation creeps and high-frequency texture smooths into a painted look. Nothing in the prompt can prevent it. Colour-match the sections in the edit, or better, treat each as a separate shot and cut.

3. D ***1. What the machine is actually doing***

A causal encoder computes each latent frame from pixel frames at or before it, never from later ones, so the very first frame is encoded on its own while every later group is squeezed together in time. Your still is therefore the sharpest frame in the clip before generation even starts, which is a large part of why image-to-video works as well as it does — and part of why the first quarter-second looks subtly different.

4. B ***1. What the machine is actually doing***

There is no variable anywhere holding the red mug or the blue shirt. There are tokens whose values correlate across frames because the model was trained on footage where objects persist, which makes persistence a statistical habit rather than a stored fact. While something is hidden, the tokens that carried it are gone; when it reappears the model re-samples something consistent with context. Design around it: cut, rather than letting an object leave and return.

5. A ***1. What the machine is actually doing***

The prompt enters through cross-attention as a few hundred text tokens against an enormous number of video tokens whose structure is mostly determined by the first frame and the model's priors. It is a light steering influence, not a specification. That ratio is also why the still wins every disagreement with the text, and why moving a requirement into the frame works when describing it harder does not.

6. C ***1. What the machine is actually doing***

Long videos were cut at shot boundaries before training, so the training unit is one continuous shot and the model has essentially never seen an edit. It has no representation of a cut to produce. That is not a limitation to work around; it is the correct division of labour. Cuts are made in an editor, which is also where you place them one moment before each generation stops working.

7. B ***1. What the machine is actually doing***

Interpolators estimate motion between two real frames and warp pixels along it. Where one object passes in front of another, some pixels exist in frame A and not in frame B, so there is nothing to warp and the interpolator smears or invents. The same failure appears on very fast motion, where the match falls outside the search range, and on anything appearing or disappearing. Interpolate calm footage freely and be suspicious of interpolated fast action.

8. D ***1. What the machine is actually doing***

A distilled model imitates a many-step model in four to eight steps and is generally trained to work near guidance 1. Feeding it the value tuned on the full model pushes far past the region where its predictions are valid, which produces clipped highlights, posterised edges and frozen motion. Distillation also narrows the motion range and weakens adherence on unusual requests, which is why a draft answers a different question from a final.

9. D ***2. The first frame, and everything decided in it***

A camera move changes the framing, so the still has to be composed for the end of the move. If the shot pushes in, start looser than feels right; if it pulls out, start tighter; if it tilts up, leave the headroom you will need. Draw the end frame as a rectangle on the still before you generate, and if the composition inside that rectangle does not work, the still is wrong rather than the prompt.

10. B ***2. The first frame, and everything decided in it***

An orbit asks the model to show the scene from an angle it was never given. If there is no space behind the subject, it invents one as the camera swings, and the invention does not hold together. Three real depth planes — a doorway or table edge in front, the subject, a corridor or room behind — give the move something to be a move through. A still with one plane will not orbit convincingly on any model.

11. A ***2. The first frame, and everything decided in it***

Nothing is being simulated; plausible frames are being sampled. But if the cause of the light is in the picture, the model has something holding the lighting in place, and the shadow side, the key direction and the colour wander less. With no visible cause there is nothing anchoring it and the light creeps. This is an aesthetic rule that happens to be a technical one, which is why motivated light is worth the trouble.

12. C ***2. The first frame, and everything decided in it***

The repair pass is for local flaws in a frame that is fundamentally right: a hand, an earring, a duplicated background figure, a bad edge. If the answer involves the whole frame — the pose is wrong, the composition is wrong, the light is wrong — you will spend an hour producing something worse than a new generation with a better prompt. Repair what you can draw a boundary around; regenerate the rest.

13. B ***2. The first frame, and everything decided in it***

A composite with no contact shadow floats, and floating is the single most common tell. Get perspective, light direction and contact right and most people never examine the rest. Colour, grain and edge softness matter and they are corrections; a missing shadow is a claim about physics that the eye rejects immediately. The shadow also has to match the softness of the other shadows in the frame.

14. D ***2. The first frame, and everything decided in it***

The model is asked to find a smooth path between two encoded endpoints. If the endpoints differ in ways you did not intend — a slightly different wall colour, a shifted lamp, a different lens character — the path has to travel through those differences too, and you see it as churn through the middle of the clip. Crop the end frame from the start frame, edit it, or generate it using the start frame as a reference.

15. C ***2. The first frame, and everything decided in it***

A seed fixes the starting noise, and everything else in the path has to match too: weights, precision, scheduler, tokenisation, batch shape and kernels. A hosted provider may batch, reroute or silently update the model between your two runs, and usually reserves the right to. The prompt, seed and version string let you rebuild something close. The file you were given is the only thing that cannot change under you.

16. A ***2. The first frame, and everything decided in it***

B-roll, cutaways, impossible or unaffordable shots, previsualisation and short-form social are where the tool is already the right one. A specific real product is the case that does not improve with scale, because the model has no reference for that packet and produces a member of the category with a label that is almost right — which reads as a counterfeit. That shot is a compositing or filming job, not a generation job.

17. C ***3. Directing, not describing***

The model animates actions, not intentions. An auto-captioner described what it could see, so verbs of movement appeared attached precisely to footage containing them, while descriptions of inner state did not appear at all. Naming the shot size, one camera move with a speed, the subject's verb and what stays still is four clauses that all sit in the reliable half of the adherence order.

18. A ***3. Directing, not describing***

A wide to a medium wide is a stumble; a wide to a close-up is an edit. That is ordinary film grammar, and in generative work it pays a second time: a big size change gives the eye something new to read and much less opportunity to compare two faces. Grammar and machinery agree, which happens often enough in this subject to be worth noticing.

19. D ***3. Directing, not describing***

The most common visible failure in a generated clip is the background quietly rearranging itself, and a background rendered as soft colour has no structure to rearrange. It also dissolves background faces into shapes with no identities to drift. The cost is that you lose the location, so a shot that has to establish where you are needs depth — and then you accept the churn or shorten the shot.

20. B ***3. Directing, not describing***

On a set a script supervisor protects screen direction. In generative work nothing does, because every shot is an independent sample. Write the direction on every line of the shot list — who is screen left, who is looking which way — build each first frame to match, and lay the stills side by side before generating any of them. Two minutes of checking saves a scene that cannot be cut.

21. A ***3. Directing, not describing***

A motion brush biases the generation rather than freezing anything, so check the unpainted area rather than assuming it held. It is genuinely valuable for local motion — steam, hair, water in the bottom third, a flag, one figure in a crowd — and the technique that gets the most from it is to paint less than you think, because the stillness around the motion is what makes the motion read as real.

22. C ***3. Directing, not describing***

A drawn path tells the model where something should be, not what its body should be doing on the way. It also fights physics if the path is one no real object would take, and it competes with the prompt, so a text description of a different motion averages into a mess. Use it for objects and simple translations; for a specific human movement, supply a pose sequence instead.

23. B ***3. Directing, not describing***

A conditioning signal carries everything about the reference's motion, wanted or not: proportions, handheld wobble, framing, and the tracker's own errors. A pose tracker that loses an arm behind a torso for four frames gives you four frames where the generated arm does something impossible. Watching the skeleton takes thirty seconds and is where most of those failures are visible. Stabilise the reference before extracting, too.

24. D ***3. Directing, not describing***

Adding tokens dilutes every other token's influence, so a long prompt is a longer list from which a similar number of things get honoured. Counts, spatial relations and object presence belong in the frame, where they are facts rather than requests. Two actions that must both happen are two shots and a cut. Specific movement belongs in a conditioning signal. What is left is a short, ordered instruction free of wishes.

25. D ***4. Making separate shots belong to each other***

The sheet is not a mood board; it is the parts bin every first frame is built from, by reference, by compositing, or by hand. Six lovely images under six different lights is not a character sheet, it is six characters, and any mismatch in it becomes a mismatch in every shot derived from it. Fixing it at the sheet stage is one regeneration; fixing it later is a scene.

26. B ***4. Making separate shots belong to each other***

A reference is turned into a vector that describes everything about those pixels — the person, the light on them, the background colour, the apparent lens — and the model is pulled toward that whole region. Identity is not separated out. Make the references under the light you intend to use, which your character sheet already specifies, and build a second sheet if the character needs a second lighting setup.

27. A ***4. Making separate shots belong to each other***

If every training image shares a background and a light, the model learns those as part of the identity and reproduces them whenever the concept is invoked. Varying background, framing, angle and expression is what isolates the person from everything around them. Caption what varies and not what is constant, because whatever you leave undescribed is what the trigger word absorbs.

28. C ***4. Making separate shots belong to each other***

Overtraining has a recognisable signature: the outputs look like the dataset. The model has learned those particular images instead of the identity behind them, so it can no longer pose, light or dress the character. Train fewer steps, or gather a more varied set — different angles, expressions, framings, backgrounds and light. Fifteen to thirty varied, clean images beats a larger set of near-duplicates.

29. B ***4. Making separate shots belong to each other***

The order is small objects, then patterns, then wall colour, with architecture surviving best because doors and windows are large. A room with three objects in it holds; a room with forty holds nothing, and every one of those forty is a chance for a viewer to notice something moved. A simple, strongly designed space both reads better and reproduces better, which is ordinary production design.

30. D ***4. Making separate shots belong to each other***

Fine patterns sit near the autoencoder's resolution floor and are reinvented each frame, so they churn, warp and change count. A large bold pattern survives; a small busy one does not. The same reasoning covers thin jewellery, glasses frames and text on clothing. One strong colour per character reads at distance, survives the grade, and makes any mismatch obvious enough that you catch it.

31. C ***4. Making separate shots belong to each other***

A look baked into the stills is strong and consistent, and changing it means regenerating everything. A look on one adjustment layer is changeable in thirty seconds at any point in the job. The exception is anything stylised enough that the model has to know about it — gouache, or 1970s stock with bloom and halation — because no grade invents that. Otherwise generate neutral and decide the look in the edit.

32. A ***4. Making separate shots belong to each other***

There is no technical difference between the operation that fixes your shot and the one that puts a real person into a video they were never in. The line is a real person's written, specific permission, and a visible label on the output. Many jurisdictions now legislate here directly and the direction of travel is one way. If you find yourself reasoning about why your case is an exception, stop and ask the person.

33. C ***5. The half of video that is not picture***

B, P and M require the lips to close completely, and many models under-close them. Viewers cannot name what is wrong and all of them notice. Slowing the delivery slightly gives the closure more frames and improves it measurably. Feeding a clean, dry, compressed and normalised voice track rather than a mix also helps, because the model is extracting phonetic features and anything else is noise to it.

34. A ***5. The half of video that is not picture***

Performance transfer maps a recorded face's movement onto a generated one, and proportions come along with the movement. A driver with a long face pulls the target toward a long face. Pick a driver whose head shape is roughly compatible, keep within about thirty degrees of frontal, keep hands and hair clear of the face, and cut every few seconds like everything else.

35. D ***5. The half of video that is not picture***

Pronunciation and timbre are solved. Interpretation is not: a synthetic voice applies plausible prosody derived from the words, while a person applies prosody derived from knowing what the sentence means. Over one sentence the difference is invisible; over three minutes it accumulates until the listener stops believing anyone is there. For anything long, record a real voice in a soft room and clean it in Audacity.

36. B ***5. The half of video that is not picture***

The text is the instrument. A comma inserts a short pause and a full stop a longer one, so punctuating where you want the speaker to breathe works better than any setting, and synthetic prosody degrades across long clauses. Spell out anything ambiguous — twenty twenty-six rather than 2026 — and generate line by line so you can regenerate one bad line instead of a whole take.

37. A ***5. The half of video that is not picture***

Amateur recordings sound amateur mostly because of reflections off hard parallel surfaces, and no processing removes reverb convincingly. Soft furnishings absorb it, and recording close — a hand's width away, slightly off-axis — makes the direct sound dominate what reflections remain. A phone in a soft room beats a good microphone in a hard one, which is why the room is the equipment.

38. C ***5. The half of video that is not picture***

Rumble, traffic, handling noise and desk thumps live below about 80 hertz, are inaudible on a laptop speaker and obvious on headphones. Leaving them in means the compressor pumps on them and the normalise step sets the peak from a thump rather than from the voice. Removing them first is the single largest improvement in the chain, and the whole four-step sequence takes about two minutes.

39. B ***5. The half of video that is not picture***

The NC clause excludes commercial use, and a client video is commercial however the platform treats it. Attribution satisfies BY, not NC. Check the licence on the specific track's own page rather than the site's general description, keep a copy of it in the project folder, and for anything with real exposure use a licensed library track, which is the boring and correct choice.

40. D ***5. The half of video that is not picture***

The transient is the sharp attack at the start of the sound, which is not necessarily the start of the file, so zoom into the waveform and align the spike rather than the edge. The contact frame is usually one or two earlier than it feels, because you react to the frame where the change is obvious rather than where it began. On the frame it disappears into the picture; two frames late reads as wrong.

41. D ***6. From generations to a finished piece***

Each pass asks a different question. Selects asks which take did the thing the shot is for, where it breaks, and whether it matches its neighbours. Assembly asks whether the sequence works. Collapsing them is why edits stall on shot three for an hour: you are trying to answer both at once and neither can be settled while the other is open.

42. B ***6. From generations to a finished piece***

Every selected take, untrimmed, in script order. It is the most useful ugly thing in editing because it reveals what no list can: whether the sequence makes sense, the two or three shots you are missing, the larger number you do not need, and where it drags. Watch it once straight through, taking notes on paper, without stopping to fix anything.

43. A ***6. From generations to a finished piece***

On a static frame the cut is the only change and is perceived as an event. During movement the eye is already tracking change and the join arrives as one more. In generative work you rarely have two shots of the same action, so the usable version is broader: cut while something in the frame is moving — a car passes, a door swings, a figure turns.

44. C ***6. From generations to a finished piece***

A J-cut brings the next scene's sound in early; an L-cut carries the last scene's sound over. Either way the audio and picture cuts no longer land on the same frame, so there is no single moment where everything changes. In generative work that also hides the join between clips with unrelated ambience and no shared location, which is most of what makes a sequence feel edited rather than stapled.

45. B ***6. From generations to a finished piece***

Visual adaptation makes memory of a shot you watched a minute ago useless for comparison. Scopes do not adapt: waveform for black level and white point, parade for a colour cast, vectorscope for hue and saturation with its skin line. Resolve's free tier has all three, which is professional colour instrumentation at no cost, and most people using it never open them.

46. D ***6. From generations to a finished piece***

The point of one grain pass is that every clip ends up with the same surface. A clip that arrives with its own uneven artefacts does not share that surface, and adding grain on top gives you two textures at once. Clean it first — FFmpeg's hqdn3d or nlmeans will do it free — then grain the clean version in the edit. Do this only when a clip genuinely has a noise problem, not as a default.

47. C ***6. From generations to a finished piece***

Clean vector type over grained footage is instantly readable as an overlay, because everything around it has a texture that it does not. Match the grade, then grain the text. Corner-pinning and matched motion matter too, and neither survives the moment a viewer notices that the letters are the only perfectly clean thing in the frame. Generate the plate with a trackable feature in it in the first place.

48. A ***6. From generations to a finished piece***

A 640-pixel proxy plays smoothly on almost anything and every editing decision except colour and fine detail can be made on it. Those two you make later, on the originals, after relinking — which is also why the original footage folder is never moved or renamed. The constraint that actually bites on a slow machine is playback, not capability, and proxies are free.

49. C ***7. Money, clients and running the job***

Every platform prices in credits, different models cost different credits per second, credits usually expire at the end of a billing period, and aggregators add a margin. None of that is dishonest and all of it separates the button from the money. Closing the distance manually, once, does more than any budgeting advice, and it takes five minutes.

50. A ***7. Money, clients and running the job***

Watching a single render finish invites the just-one-more loop: nearly right, adjust a word, generate, watch. An afternoon disappears and the spend arrives in invisible pieces. Batching breaks the loop mechanically, and the judgement improves: side-by-side comparison is a much stronger test than sequential comparison against your memory of the previous clip.

51. D ***7. Money, clients and running the job***

The provenance problem is solved as a side effect: prompts, seeds, durations and model slugs are all in the file. Seeds become explicit and deliberately varied rather than unknown draws, a re-run with one changed field is an experiment rather than a session, and the spend is countable before you start — shots times takes times price. Always write the responses to disk, including the failures.

52. B ***7. Money, clients and running the job***

Warmer is a grade: an adjustment layer, thirty seconds, no generation. A different shot size is a new first frame and a new set of generations. Clients do not know these are different and will not guess, so the quote has to say so. The same document should name the approval gate after which structural changes are billed, and who supplies logos, products, fonts and consent forms.

53. A ***7. Money, clients and running the job***

Character-driven dialogue across many shots is where the compounding hit rate and the absence of object memory both bite, and a small live shoot is faster, cheaper and better. The other three are exactly what the tool is already right for. Saying so on day one wins work, because it reads as expertise; discovering it on day fourteen with a deadline in sight does not.

54. C ***7. Money, clients and running the job***

Generation time is queue time, and queue time is not yours to control. A rush means paying for a faster tier, running more in parallel, or accepting fewer attempts per shot. Stating the trade explicitly is an honest description of the medium and lets the client choose, which is better for both of you than a quiet gamble on the hit rate.

55. B ***7. Money, clients and running the job***

Most people are surprised by the tally. The thing you worry about is rarely the thing costing you takes: teams that expect face drift find wardrobe colour, and teams that expect text failures find background churn. Written next to the quoted numbers — shots delivered, generations used, takes per shot, hours by stage — it prices your next quote better than any rate card.

56. D ***7. Money, clients and running the job***

The clause is not legal advice and it is not a shield against everything. It does two useful things: it puts the question in front of the client at the point where it can still be answered, and it records that you asked. The signed releases still go in a consent folder beside the project, with any time limit noted, because that is what answers the question eight months later.