What models can and cannot do

Full course PDF

Learn framing, camera, light, prompts and delivery through short lessons and practice.

7. What models can and cannot do

What does Identity adapter mean in Character and set consistency: the weakest link?

A small trained module that lets a frozen base model read a reference image of a subject as a prompt, IP-Adapter being the common example.

https://shotcraft.video/learn/concepts/character-consistency/

What does Subject fine-tuning mean in Character and set consistency: the weakest link?

Updating model weights, or an adapter on them, from a small image set so a specific person, object or set can be called up by name.

https://shotcraft.video/learn/concepts/character-consistency/

What does Textual Inversion mean in Character and set consistency: the weakest link?

Learning a new embedding, effectively a new word, for a concept while leaving the model itself unchanged.

https://shotcraft.video/learn/concepts/character-consistency/

What does Prior preservation mean in Character and set consistency: the weakest link?

A training term that keeps a fine-tune from overwriting the model's general knowledge of the class your subject belongs to.

https://shotcraft.video/learn/concepts/character-consistency/

What does Character sheet mean in Character and set consistency: the weakest link?

A prepared set of reference frames covering angles, expressions, wardrobe and lighting, used as conditioning across a sequence.

https://shotcraft.video/learn/concepts/character-consistency/

What does Conditioning mean in Conditioning and control: everything you hand the model besides noise?

Any input other than the initial noise that biases what the model generates, including text, images, masks, structure maps and camera trajectories.

https://shotcraft.video/learn/concepts/conditioning-and-control/

What does Cross-attention mean in Conditioning and control: everything you hand the model besides noise?

The layer type that lets a diffusion model read a conditioning signal such as an encoded text prompt while denoising.

https://shotcraft.video/learn/concepts/conditioning-and-control/

What does Structure conditioning mean in Conditioning and control: everything you hand the model besides noise?

Conditioning on a spatial map derived from a reference, such as depth, edges or human pose, to fix geometry while leaving appearance to the model.

https://shotcraft.video/learn/concepts/conditioning-and-control/

What does Conditioning scale mean in Conditioning and control: everything you hand the model besides noise?

A weight that sets how strongly a structure control overrides the model's own tendencies, traded against image quality.

https://shotcraft.video/learn/concepts/conditioning-and-control/

What does Outpainting mean in Conditioning and control: everything you hand the model besides noise?

Generating new content beyond the existing frame edges, used to extend a plate or change aspect ratio.

https://shotcraft.video/learn/concepts/conditioning-and-control/

What does LoRA mean in LoRAs and fine-tuning: teaching a model your subject?

Low-rank adaptation: a training method that freezes the base weights and learns small rank-decomposition matrices inserted into each layer.

https://shotcraft.video/learn/concepts/loras-and-finetuning/

What does Rank mean in LoRAs and fine-tuning: teaching a model your subject?

The size of the low-rank update, which sets how much the adapter can learn and how large the resulting file is.

https://shotcraft.video/learn/concepts/loras-and-finetuning/

What does Base checkpoint mean in LoRAs and fine-tuning: teaching a model your subject?

The specific pretrained model an adapter was trained against and must be loaded with.

https://shotcraft.video/learn/concepts/loras-and-finetuning/

What does Trigger word mean in LoRAs and fine-tuning: teaching a model your subject?

The token a fine-tune binds a subject or style to, used to call that concept up in a prompt.

https://shotcraft.video/learn/concepts/loras-and-finetuning/

What does Overfitting mean in LoRAs and fine-tuning: teaching a model your subject?

Training so far into the dataset that the model reproduces its incidental details and loses responsiveness to prompts.

https://shotcraft.video/learn/concepts/loras-and-finetuning/

What does temporal flicker mean in Temporal consistency: flicker, drift and morphing?

Frame-to-frame instability in texture, brightness or fine detail that was never committed to in the latent.

https://shotcraft.video/learn/concepts/temporal-consistency/

What does identity drift mean in Temporal consistency: flicker, drift and morphing?

A face or character gradually changing across a shot because the model stores no persistent identity.

https://shotcraft.video/learn/concepts/temporal-consistency/

What does texture crawl mean in Temporal consistency: flicker, drift and morphing?

Patterns and fine surfaces reshuffling each frame instead of holding still on the object.

https://shotcraft.video/learn/concepts/temporal-consistency/

What does keyframe cascade mean in Temporal consistency: flicker, drift and morphing?

Generating distant frames first and then filling the gaps with temporal super-resolution models.

https://shotcraft.video/learn/concepts/temporal-consistency/

What does temporal super-resolution mean in Temporal consistency: flicker, drift and morphing?

A separately trained model that increases frame rate by generating intermediate frames.

https://shotcraft.video/learn/concepts/temporal-consistency/

What does Spacetime patch mean in The economics of generated video: why iteration is a budget line?

The unit a video transformer processes, a small block of latent covering an area of frame across a span of frames.

https://shotcraft.video/learn/concepts/video-economics/

What does Context length mean in The economics of generated video: why iteration is a budget line?

The maximum number of tokens a model can attend over at once, which sets the longest clip it can generate in one pass.

https://shotcraft.video/learn/concepts/video-economics/

What does Compression ratio mean in The economics of generated video: why iteration is a budget line?

How much smaller the latent representation is than the pixels, traded directly against retained fine detail.

https://shotcraft.video/learn/concepts/video-economics/

What does Denoising steps mean in The economics of generated video: why iteration is a budget line?

The number of iterative passes over the whole latent clip, each one a full-cost forward pass.

https://shotcraft.video/learn/concepts/video-economics/

What does Accept rate mean in The economics of generated video: why iteration is a budget line?

The proportion of generations a human keeps, which is the number that actually determines cost per finished shot.

https://shotcraft.video/learn/concepts/video-economics/

What does Native audio mean in Video model capability matrix?

The model returns a soundtrack with the picture, generated in the same pass, rather than leaving you to add sound afterwards.

https://shotcraft.video/reference/model-capabilities/ · 2026-08-17

What does Negative prompt field mean in Video model capability matrix?

A separate field for what the model must leave out. Where there is none, exclusions have to sit inside the prompt itself, where they compete for length and are read as description.

A field is not the same as an effect. Some models accept the field and act on it weakly. See the prompt length guide for how much room the field gets, which differs from the prompt's own budget.

https://shotcraft.video/reference/model-capabilities/ · 2026-08-17

What does Image as start frame mean in Video model capability matrix?

The generation can be anchored to a still you supply, which is how you control composition, cast and colour instead of describing them.

https://shotcraft.video/reference/model-capabilities/ · 2026-08-17

What does Last frame or keyframes mean in Video model capability matrix?

The generation can be given an end frame, or frames at set points, so a shot lands somewhere specific. This is what makes a cut between two shots match.

https://shotcraft.video/reference/model-capabilities/ · 2026-08-17

What does Longest single generation mean in Video model capability matrix?

Seconds in one call, before extension. Longer clips are usually assembled from several calls, and the joins are visible.

Models you run yourself are usually specified as a frame count and a frame rate rather than a duration, and the ceiling belongs to the inference pipeline rather than to the weights. Those rows are left not established unless the repository states a figure in seconds.

https://shotcraft.video/reference/model-capabilities/ · 2026-08-17

What does Open weights mean in Video model capability matrix?

The weights are published, so the model can run on your own hardware. It does not mean you may use the output commercially. Read the licence.

https://shotcraft.video/reference/model-capabilities/ · 2026-08-17

What does API mean in Video model capability matrix?

The vendor offers a first-party programmatic endpoint. A model marked no may still be reachable through a reseller.

https://shotcraft.video/reference/model-capabilities/ · 2026-08-17

How do Conditioning and Structure conditioning differ?

Conditioning: Any input other than the initial noise that biases what the model generates, including text, images, masks, structure maps and camera trajectories. Structure conditioning: Conditioning on a spatial map derived from a reference, such as depth, edges or human pose, to fix geometry while leaving appearance to the model.

https://shotcraft.video/learn/concepts/conditioning-and-control/

Explain why this claim is wrong: the model remembers the character from the last generation.

It has no persistent memory of your cast or set between generations; continuity comes from what you re-supply as conditioning each time.

https://shotcraft.video/learn/concepts/character-consistency/

Explain why this claim is wrong: one good reference image is enough.

Single references pin a look from one angle and one light; identity fails first on profile, extreme angles, and lighting far from the reference, which is why fine-tuning methods use small sets rather than single frames.

https://shotcraft.video/learn/concepts/character-consistency/

Explain why this claim is wrong: a DreamBooth-style fine-tune is strictly better than an adapter.

A fine-tune changes the weights and can degrade everything else the model knew, which is why DreamBooth carries a prior-preservation loss; an adapter such as IP-Adapter leaves the base model frozen and stays compatible with other controls.

https://shotcraft.video/learn/concepts/character-consistency/

Explain why this claim is wrong: consistency means the face.

Wardrobe detail, set dressing, prop position, hair, light direction and eyeline all drift independently, and audiences read a change of light direction across a cut as fast as they read a change of face.

https://shotcraft.video/learn/concepts/character-consistency/

Explain why this claim is wrong: this is a solved problem in current models.

Cross-shot continuity remains the weakest part of the pipeline and is currently bought with references, fine-tunes and human selection rather than with model memory.

https://shotcraft.video/learn/concepts/character-consistency/

Explain why this claim is wrong: a good enough prompt can specify anything.

In practice text is one conditioning channel competing with image, structure and motion inputs, and spatial precision comes from the image-side overlays, not from more adjectives.

https://shotcraft.video/learn/concepts/conditioning-and-control/

Explain why this claim is wrong: image-to-video means the model keeps your image intact.

It uses the frame as a starting condition and re-synthesises everything after it, so wardrobe, set dressing and background text drift as the clip runs.

https://shotcraft.video/learn/concepts/conditioning-and-control/

Explain why this claim is wrong: ControlNet-style depth and pose conditioning give you camera and lighting control.

They constrain where surfaces and limbs are; the lens character, exposure and light direction still come out of the model's learned appearance.

https://shotcraft.video/learn/concepts/conditioning-and-control/

Explain why this claim is wrong: stacking more controls always helps.

Multiple conditionings can contradict each other, and the standard advice with multi-ControlNet setups is to mask the inputs so they do not overlap and to tune each conditioning weight separately.

https://shotcraft.video/learn/concepts/conditioning-and-control/

Explain why this claim is wrong: inpainting a face fixes it.

Inpainting resamples inside the mask with the same priors that produced the error, so the usual fix is a better reference or a different seed rather than a repeated local repair.

https://shotcraft.video/learn/concepts/conditioning-and-control/

Explain why this claim is wrong: a LoRA is a smaller model.

It is a set of additional low-rank weights that only works on top of the specific base model it was trained against, so the base checkpoint is part of the deliverable.

https://shotcraft.video/learn/concepts/loras-and-finetuning/

Explain why this claim is wrong: more training images always give a better LoRA.

A small, clean, varied set beats a large set with repeated backgrounds or inconsistent lighting, because the model will happily learn the wallpaper along with the face.

https://shotcraft.video/learn/concepts/loras-and-finetuning/

Explain why this claim is wrong: full fine-tuning is the quality ceiling and LoRA is the compromise.

The original LoRA work reports performance on par with or better than full fine-tuning at a fraction of the trainable parameters, and adapters have the practical advantage of composing with other controls.

https://shotcraft.video/learn/concepts/loras-and-finetuning/

Explain why this claim is wrong: if the LoRA works it will work at any weight.

Adapter strength is a dial with a usable window; too low and identity slips, too high and the adapter overrides prompt, composition and camera behaviour.

https://shotcraft.video/learn/concepts/loras-and-finetuning/

Explain why this claim is wrong: overfitting shows up as bad images.

It usually shows up as loss of control: the trigger word stops being optional, the poses repeat, and the model can no longer place the subject anywhere it did not see during training.

https://shotcraft.video/learn/concepts/loras-and-finetuning/

Explain why this claim is wrong: flicker is a bug awaiting a patch.

It follows from the model re-deriving detail it never committed to. Scale and architecture have reduced it considerably and have not removed it, and benchmarks still measure it as a standing dimension.

https://shotcraft.video/learn/concepts/temporal-consistency/

Explain why this claim is wrong: the model tracks characters and objects from frame to frame.

It keeps frames mutually plausible. Identity is not stored, which is why a face can drift inside one shot and will not match across shots by default.

https://shotcraft.video/learn/concepts/temporal-consistency/

Explain why this claim is wrong: higher resolution reduces morphing.

Resolution and temporal stability are separate axes, and a crisper frame often makes crawl and drift easier to see.

https://shotcraft.video/learn/concepts/temporal-consistency/

Explain why this claim is wrong: interpolating between keyframes gives consistent motion.

It gives smooth motion. Lumiere's authors argue that generating distant keyframes and filling between them makes global consistency harder, which is why they generate the whole duration at once.

https://shotcraft.video/learn/concepts/temporal-consistency/

Explain why this claim is wrong: continuity within a shot and continuity across shots are the same problem.

Within a shot, attention between latents does real work. Across shots nothing is shared, so matching has to be bought with references, fine-tunes and selection.

https://shotcraft.video/learn/concepts/temporal-consistency/

Explain why this claim is wrong: video costs about as much as an image times the frame count.

Tokens span space and time and attention cost grows faster than linearly with sequence length, so doubling duration costs more than double.

https://shotcraft.video/learn/concepts/video-economics/

Explain why this claim is wrong: the model generates frame by frame, so a longer clip is just more of the same work.

Sampling is iterative over the whole clip at once, and every denoising step processes the full spacetime latent.

https://shotcraft.video/learn/concepts/video-economics/

Explain why this claim is wrong: compute is the dominant cost of a generative sequence.

Human selection is usually the larger line, because someone must watch and reject the majority of generations and that time does not scale down.

https://shotcraft.video/learn/concepts/video-economics/

Explain why this claim is wrong: generating at final resolution from the start saves a step.

The standard economy is to prototype small and finish large, which the Sora report describes as an explicit benefit of training on native sizes.

https://shotcraft.video/learn/concepts/video-economics/

Explain why this claim is wrong: more compression is free efficiency.

LTX-Video is explicit that its high compression ratio inherently limits the representation of fine detail, which is why the model puts effort back into the decoder.

https://shotcraft.video/learn/concepts/video-economics/

Progress can be exported to a file and imported again. With no server, that file is the only real guarantee — browsers do clear storage.