What generation actually is

Full course PDF

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

1. What generation actually is

What does forward process mean in Diffusion models: generation by denoising?

The fixed schedule of adding noise to a real image, used to manufacture training examples.

https://shotcraft.video/learn/concepts/diffusion-basics/

What does reverse process mean in Diffusion models: generation by denoising?

The learned sequence of denoising steps that turns pure noise into a sample.

https://shotcraft.video/learn/concepts/diffusion-basics/

What does noise prediction mean in Diffusion models: generation by denoising?

The model's actual output at each step: an estimate of the noise in the current input, not the finished image.

https://shotcraft.video/learn/concepts/diffusion-basics/

What does score mean in Diffusion models: generation by denoising?

The direction in which a noisy image should move to become more like real data; the quantity score-based formulations estimate.

https://shotcraft.video/learn/concepts/diffusion-basics/

What does sample mean in Diffusion models: generation by denoising?

One generation drawn from the model, which differs run to run because the starting noise differs.

https://shotcraft.video/learn/concepts/diffusion-basics/

What does latent mean in Latent space: working at proxy scale?

A compressed, learned representation of an image or clip that the diffusion model operates on instead of pixels.

https://shotcraft.video/learn/concepts/latent-space/

What does latent diffusion mean in Latent space: working at proxy scale?

Running the denoising process inside a pretrained autoencoder's compressed space rather than in pixel space.

https://shotcraft.video/learn/concepts/latent-space/

What does compression ratio mean in Latent space: working at proxy scale?

How many pixels of input each unit of latent stands for, trading detail for speed and length.

https://shotcraft.video/learn/concepts/latent-space/

What does token mean in Latent space: working at proxy scale?

One position the model attends over; in image and video models, one patch of latent.

https://shotcraft.video/learn/concepts/latent-space/

What does channels mean in Latent space: working at proxy scale?

The many values stored per latent cell, describing content rather than colour.

https://shotcraft.video/learn/concepts/latent-space/

What does Seed mean in Seeds and reproducibility: re-running a generation?

The integer that initialises the pseudo-random number generator producing the starting noise for a generation.

https://shotcraft.video/learn/concepts/seeds-and-reproducibility/

What does Generator mean in Seeds and reproducibility: re-running a generation?

The stateful random-number object a pipeline draws from, which must be re-seeded per call to repeat a result.

https://shotcraft.video/learn/concepts/seeds-and-reproducibility/

What does Determinism mean in Seeds and reproducibility: re-running a generation?

The property that identical inputs produce identical outputs, which in practice holds only within a fixed software, hardware and model configuration.

https://shotcraft.video/learn/concepts/seeds-and-reproducibility/

What does Version pinning mean in Seeds and reproducibility: re-running a generation?

Fixing the exact model checkpoint and library versions used for a job so it can be rerun later.

https://shotcraft.video/learn/concepts/seeds-and-reproducibility/

What does Generation log mean in Seeds and reproducibility: re-running a generation?

The record of prompt, seed, model version, parameters and conditioning asset hashes needed to reconstruct a delivered frame.

https://shotcraft.video/learn/concepts/seeds-and-reproducibility/

What does attention mean in Transformers and DiT: why the backbone changed?

An operation where each position pulls a weighted mixture of information from all other positions.

https://shotcraft.video/learn/concepts/transformers-and-dit/

What does patch mean in Transformers and DiT: why the backbone changed?

A fixed-size square of latent treated as one token by the transformer.

https://shotcraft.video/learn/concepts/transformers-and-dit/

What does DiT mean in Transformers and DiT: why the backbone changed?

Diffusion Transformer: a latent diffusion model whose denoising backbone is a transformer rather than a U-Net.

https://shotcraft.video/learn/concepts/transformers-and-dit/

What does U-Net mean in Transformers and DiT: why the backbone changed?

The convolutional encoder-decoder backbone used by earlier diffusion models, which processes the frame through a hierarchy of resolutions.

https://shotcraft.video/learn/concepts/transformers-and-dit/

What does scaling behaviour mean in Transformers and DiT: why the backbone changed?

The observed relationship between compute or model size and output quality, which lets teams predict the return on more spending.

https://shotcraft.video/learn/concepts/transformers-and-dit/

What does encoder mean in The VAE: a learned codec between pixels and latents?

The half of the autoencoder that turns pixels into a latent, used during training and for image-to-video conditioning.

https://shotcraft.video/learn/concepts/vae/

What does decoder mean in The VAE: a learned codec between pixels and latents?

The half that turns a finished latent back into pixels, run once at the end of generation.

https://shotcraft.video/learn/concepts/vae/

What does 3D VAE mean in The VAE: a learned codec between pixels and latents?

A video autoencoder that compresses along time as well as height and width.

https://shotcraft.video/learn/concepts/vae/

What does reconstruction artifact mean in The VAE: a learned codec between pixels and latents?

A defect introduced by decoding rather than by the diffusion process, so it repeats across seeds.

https://shotcraft.video/learn/concepts/vae/

What does checkerboard artifact mean in The VAE: a learned codec between pixels and latents?

Regular grid patterning produced by certain upsampling layers in a decoder.

https://shotcraft.video/learn/concepts/vae/

What does spacetime patch mean in The time axis: spacetime patches and the cost of duration?

A token covering a small area of frame across a short span of frames, the basic unit a video model operates on.

https://shotcraft.video/learn/concepts/video-time-axis/

What does sequence length mean in The time axis: spacetime patches and the cost of duration?

How many tokens one generation contains, growing with resolution and duration.

https://shotcraft.video/learn/concepts/video-time-axis/

What does context length mean in The time axis: spacetime patches and the cost of duration?

The maximum number of tokens a model was trained to process at once, which caps clip length.

https://shotcraft.video/learn/concepts/video-time-axis/

What does joint spatial-temporal attention mean in The time axis: spacetime patches and the cost of duration?

Attention computed across all positions in space and time at once, rather than separately per frame.

https://shotcraft.video/learn/concepts/video-time-axis/

What does temporal layers mean in The time axis: spacetime patches and the cost of duration?

Blocks added to a pretrained image model so it can relate frames to each other, an alternative to dicing along time.

https://shotcraft.video/learn/concepts/video-time-axis/

Explain why this claim is wrong: the model draws the image progressively, like a painter adding strokes to a blank canvas.

In fact every step operates on the whole frame at once. What changes between steps is how much noise is left in the frame, not how much of the picture has been filled in.

https://shotcraft.video/learn/concepts/diffusion-basics/

Explain why this claim is wrong: the noise is a random ingredient the model sprinkles in for variety.

The noise is the starting material and the training signal. The model's only output is an estimate of the noise present, and the picture is the residue after those estimates are subtracted.

https://shotcraft.video/learn/concepts/diffusion-basics/

Explain why this claim is wrong: the forward noising process is something the model learns.

Only the reverse direction is learned. Adding noise is fixed arithmetic, which is what gives training a gradeable right answer at every noise level.

https://shotcraft.video/learn/concepts/diffusion-basics/

Explain why this claim is wrong: diffusion models retrieve and blend training images.

No image library is consulted at generation time. What is stored is a function mapping a noisy frame to a noise estimate, and the resemblance to training data comes from that function.

https://shotcraft.video/learn/concepts/diffusion-basics/

Explain why this claim is wrong: more steps means a more detailed picture.

Sampling converges. Past the point where each step moves the frame a negligible distance, extra steps buy render time and nothing else.

https://shotcraft.video/learn/concepts/diffusion-basics/

Explain why this claim is wrong: the latent is a small version of the picture, so a good upscaler could recover the missing detail.

It is not an image at all, but a grid of learned descriptions. Detail that was never encoded can only be invented, not recovered.

https://shotcraft.video/learn/concepts/latent-space/

Explain why this claim is wrong: a larger output resolution means the model worked at that resolution.

Denoising happens on the latent grid; pixel dimensions come from the decoder. Asking for more pixels does not add latent capacity.

https://shotcraft.video/learn/concepts/latent-space/

Explain why this claim is wrong: latent space is a map you can navigate by moving in a direction toward a concept.

Some directions are meaningful, but the tidy pictures people show are low-dimensional projections, and projections of high-dimensional data mislead about distance and grouping.

https://shotcraft.video/learn/concepts/latent-space/

Explain why this claim is wrong: small text fails because the model cannot spell.

Spelling is part of it, but compression is the harder limit. When a single token covers a 32 by 32 pixel patch, there is no room to carry the shape of a small glyph.

https://shotcraft.video/learn/concepts/latent-space/

Explain why this claim is wrong: all models share one latent space, so latents are portable.

Each autoencoder defines its own; matching an existing latent space is a deliberate engineering goal, not a default.

https://shotcraft.video/learn/concepts/latent-space/

Explain why this claim is wrong: the same seed gives the same picture.

The same seed gives the same picture only with the same model version, parameters and, in practice, the same hardware and library stack.

https://shotcraft.video/learn/concepts/seeds-and-reproducibility/

Explain why this claim is wrong: a seed is a style or a look.

It is an index into the noise the process starts from; it correlates with composition and framing for a fixed prompt, but it carries no meaning that transfers to a different prompt or model.

https://shotcraft.video/learn/concepts/seeds-and-reproducibility/

Explain why this claim is wrong: reusing the same generator object in a batch loop pins the result.

Generator state is consumed as it is used, so each call in the loop draws different noise unless you re-seed before every call.

https://shotcraft.video/learn/concepts/seeds-and-reproducibility/

Explain why this claim is wrong: enabling deterministic algorithms makes runs bit-identical everywhere.

Deterministic modes remove specific sources of variation and cost throughput, while the underlying framework still refuses to guarantee identity across releases, commits or platforms.

https://shotcraft.video/learn/concepts/seeds-and-reproducibility/

Explain why this claim is wrong: keeping the prompt is keeping the revision trail.

Without the model version and every parameter, a saved prompt reproduces a family of images rather than the shot that was approved.

https://shotcraft.video/learn/concepts/seeds-and-reproducibility/

Explain why this claim is wrong: attention means the model looks where a human would look.

Attention is weighted retrieval between positions inside the network. It sometimes correlates with salience, but it is not a gaze and not an explanation.

https://shotcraft.video/learn/concepts/transformers-and-dit/

Explain why this claim is wrong: transformers replaced diffusion.

DiT keeps the diffusion training objective and replaces only the backbone that performs the denoising.

https://shotcraft.video/learn/concepts/transformers-and-dit/

Explain why this claim is wrong: a token in an image or video model is an object, a word or a meaningful unit.

It is a fixed-size patch of latent, drawn on a grid with no regard for what happens to be inside it.

https://shotcraft.video/learn/concepts/transformers-and-dit/

Explain why this claim is wrong: the DiT result proves bigger is always better.

It reports a scaling trend between compute per forward pass and FID on image benchmarks. That is a prediction about spending, not a guarantee about any particular shot.

https://shotcraft.video/learn/concepts/transformers-and-dit/

Explain why this claim is wrong: U-Nets are obsolete.

They remain in wide use for images and in some video pipelines, and Lumiere's space-time U-Net is a recent design. The shift to transformers is about scaling behaviour, not about one architecture being wrong.

https://shotcraft.video/learn/concepts/transformers-and-dit/

Explain why this claim is wrong: the VAE is part of the diffusion model.

It is a separate network, trained separately and sometimes swapped independently, so two pipelines can share a diffusion model and differ in their decoder.

https://shotcraft.video/learn/concepts/vae/

Explain why this claim is wrong: every artifact comes from the diffusion model, so another seed will clear it.

Decoder artifacts are deterministic given the latent, which is why grid patterning, texture buzz and waxy skin reappear across every seed.

https://shotcraft.video/learn/concepts/vae/

Explain why this claim is wrong: a video VAE is just an image VAE applied frame by frame.

Early pipelines did that and it leaves motion between consecutive frames unsmooth. Current video models compress along time too, which is what a 3D VAE means.

https://shotcraft.video/learn/concepts/vae/

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

It buys speed and duration and costs fine detail, which is why LTX-Video states the limitation directly and gives extra work to the decoder to compensate.

https://shotcraft.video/learn/concepts/vae/

Explain why this claim is wrong: the decoder simply enlarges the latent.

It reconstructs pixels from a learned description, and in some designs it also performs the last denoising step, so it is making decisions about texture rather than scaling anything.

https://shotcraft.video/learn/concepts/vae/

Explain why this claim is wrong: a video model generates frames one after another the way a camera exposes them.

Most current models denoise the entire clip at once, with all frames present in the same latent at every step.

https://shotcraft.video/learn/concepts/video-time-axis/

Explain why this claim is wrong: doubling the clip length doubles the cost.

Token count grows with duration and attention cost grows faster than linearly in token count, so the bill grows worse than double.

https://shotcraft.video/learn/concepts/video-time-axis/

Explain why this claim is wrong: frame rate is effectively free because interpolation is cheap.

In a spacetime-patch model, temporal resolution consumes the same token budget as spatial resolution. Some pipelines do add temporal super-resolution afterwards, but that is a separate model with its own failure modes.

https://shotcraft.video/learn/concepts/video-time-axis/

Explain why this claim is wrong: the model maintains an internal 3D scene and renders frames from it.

It holds spacetime patches and attention between them. Any apparent geometry is a byproduct of appearance statistics, not a maintained representation.

https://shotcraft.video/learn/concepts/video-time-axis/

Explain why this claim is wrong: a longer clip is simply a longer version of the same generation.

Length changes what fits in context, which changes how the model can hold the start and end of a move together.

https://shotcraft.video/learn/concepts/video-time-axis/

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