MAKE THINGS

Making Things with AI

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

First edition, September 2026

Addaly

Series editor: Dr Mustafa Mun

Draft, open for academic review

MAKE THINGS

Making Things with AI

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

Series editor: Dr Mustafa Mun

Addaly

First edition, September 2026 · Draft, open for academic review

Making Things with AI

First edition, September 2026, build 62a26075.

Written by AI models to a written standard, with Dr Mustafa Mun as series editor. Exam keys checked blind by a second model: 3,665 of 3,666 agreed. Academic review invited.

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.

How to cite: *Mun, M. (2026). Making Things with AI. Addaly. addaly.com/learn/making-things-with-ai.*

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

Free to read, free to download, free to print, and free to teach from. There is no paid edition and there never will be.

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. The answers to every check, test and examination question are at the back.

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

Brief contents

1

MODULE 1

How the picture gets made

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

Lesson 1 · 7 min

What the model is actually doing

THE ONE THING TO KEEP

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

Start with a photograph you ruin

Take a photo. Add a little random speckle. Add a little more. Keep going for a thousand rounds and you have television static — the photo is gone. That process is easy and needs no intelligence at all.

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

What training actually does

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

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

Generating is that skill run backwards

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

Two things follow immediately.

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

Where your words come in

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

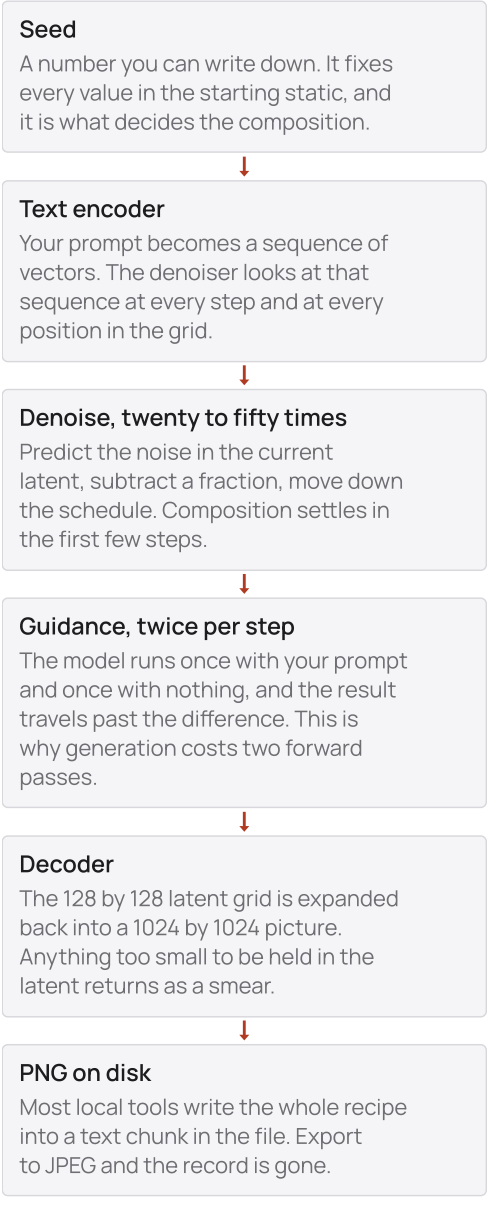
Then a technique called classifier-free guidance sharpens it. The model runs twice per step — once with your prompt, once with nothing — and the difference between the two is amplified. How much you amplify is the guidance scale, usually somewhere around 3 to 8. Set it low and the image drifts off your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

Doing it in a smaller room

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

Figure 1.1

One image, from a seed to a file



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

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

Not every model works this way

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

Why bother knowing this

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

BEFORE YOU MOVE ON

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

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

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

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

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

The answer and why: p. 449

Lesson 2 · 9 min

Noise, steps and the schedule

THE ONE THING TO KEEP

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

Training is destruction, run backwards

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

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

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

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

What a step actually is

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

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

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

Why more steps stop helping

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

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

A practical routine, which costs almost nothing:

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

Run that ladder once for each model you use and you will know its useful range for the rest of the year. Free tools make this easy: ComfyUI and AUTOMATIC1111's WebUI both have a batch mode that varies one number and lays the results out side by side.

Where the schedule shows up in your results

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

There is a specific, documented flaw worth knowing. Several widely used schedules never quite reach pure noise at their top step — they leave a faint trace of the average brightness of the training set. The model therefore never

learns to produce a genuinely very dark or very bright image, because at generation time it starts from true noise, which it never saw in training. That is the mechanism behind the complaint that certain models "cannot do night". It was a bug in the schedule, not a shortage of night photographs.

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

BEFORE YOU MOVE ON

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

A. The sampler's error is already below what the output file can represent at 20 steps, so extra steps refine a path that has already converged.

B. The model has a fixed internal resolution, so any extra computation is discarded before the image is written.

C. The prompt was too short to give the model enough to do with the additional steps, so it idled.

D. Step counts above roughly 25 are silently capped by most generation tools as a protection against runaway compute costs, so the additional steps were requested but never actually executed on the hardware.

The answer and why: p. 450

CASE STUDY · MODULE 1

Forty thousand posters and a photograph nobody recognised

An illustrative case, built from documented patterns.

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

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

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

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

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

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

The arithmetic is right and the conclusion does not follow. Average compression across a whole training set says nothing about any particular item in it. An image that appeared once contributes a nudge that is averaged away. An image that appeared a thousand times, with much the same caption each time, becomes a reliable low-error target that the model can lower its training loss by learning specifically. This stock photograph had been syndicated to health departments, NGOs and municipal websites across two decades. Its caption was, in effect, the intern′s prompt.

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

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

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

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

YOUR ANSWERS

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

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

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

STOP HERE

Write your answers before you turn the page. How it played out starts on p. 48.

CASE STUDY · MODULE 1

How it played out

Sunita changed the seed.

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

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

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

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

The questions, discussed

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

Inspection asks whether the image looks copied, which is a question about an impression. Reverse image search asks whether this composition already exists somewhere on the indexed web, which is a question about a record outside the file. The second question has an answer that does not depend on

how good the generator was or how carefully anyone looked. It is also the check that catches the specific mechanism at work here: near-copies cluster on images that were duplicated many times in training, and images duplicated many times in training are exactly the ones a search index has seen many times too.

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

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

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

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

MODULE 1 TEST

Check what stuck

6 questions, one right answer each. This one is for you, not for a record. The answers, and the reason behind each, are in the key on p. 448. If two go wrong, re-read the module before the exam.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

B. A prompt naming a very famous single painting.

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

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

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

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

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

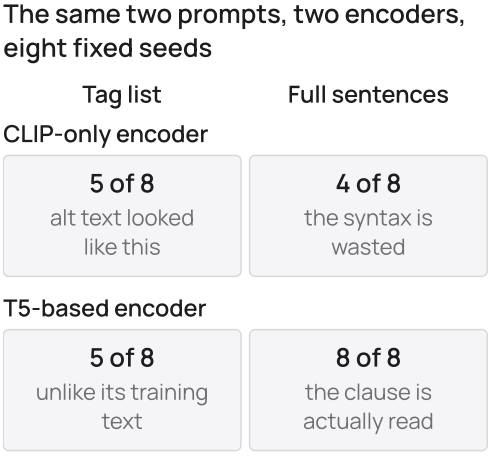
C. The newer model is autoregressive rather than diffusion, and ignores commas.

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

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

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

Figure 2.1



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

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

Style words are still tags, on every model

There is a part of the keyword tradition that survives the transition, and it is worth separating out. Words like `oil painting`, `linocut`, `Kodachrome`, `blueprint`, `medical illustration` are not describing the scene. They are

naming a visual register that co-occurred with those words in captions. These work as tags on every architecture, because they are labels rather than relations.

So the practical shape of a good modern prompt is usually two parts: a sentence describing what is in the frame and where, followed by a short set of register words. Mixing them is fine. Writing forty register words and no sentence is the old habit, and on a current model it produces a picture with a strong look and a vague subject.

The words that were never doing anything

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

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

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

Where free tools help

You can do all of this without paying anyone. Free local interfaces — ComfyUI, AUTOMATIC1111's WebUI, Forge, InvokeAI, Fooocus — all support fixed seeds and batch comparison, and several hosted services offer enough free generations a day to run the tests above. On a phone, Draw Things on iOS

EXAMINATION

Making Things with AI — final examination

This paper covers the whole course:

- how a diffusion model builds a picture
- how prompts and references steer it
- the failures that come from its architecture rather than your wording
- the editing controls that turn a generation into a deliverable
- what video and three-dimensional generation can and cannot do
- the speech and music pipelines and their consent and rights positions
- provenance and disclosure
- the law on training, ownership, likeness and labelling

63 questions, the whole pool. The site draws 25 for a sitting, pass mark 70%. Work any 25 under your own conditions. Answers from page 503.

1 Module 1 · How the picture gets made

Training a diffusion model consists of adding a random amount of noise to a training image and asking the network to predict the noise that was added. Which consequence follows directly from that objective?

A. The model learns to recognise which parts of a mess are mess, rather than storing what objects look like.
B. The model builds an internal library of object shapes that can be queried by name.
C. The model learns a mapping from captions to compositions, which is why prompts control layout so precisely.
D. The model learns to evaluate whether an image is realistic, which is what guidance later amplifies.

2 Module 1 · How the picture gets made

A checkpoint is said to be incapable of producing genuinely dark night scenes. The documented cause is a flaw in its noise schedule. What is the flaw?

A. The schedule spends too few steps at low noise, so dark texture is never resolved.
B. The schedule never quite reaches pure noise, leaving a trace of the training set's average brightness.
C. The schedule applies guidance more strongly at high noise, which lifts the exposure.
D. The schedule was calibrated on a latent with four channels, which cannot represent deep shadow.

3 Module 1 · How the picture gets made

Asking a model trained at 512 pixels for a 1024-wide image in a single pass frequently returns two heads or an extra torso. What is happening?

A. The model is confused about anatomy because it has seen few full-body training images.

B. The decoder tiles the latent and each tile is decoded independently.

C. Each region of the oversized latent independently looks like a plausible start of a subject, and nothing coordinates them.

D. Guidance is applied per region, so distant parts of the frame receive contradictory conditioning.

4 Module 1 · How the picture gets made

'A red cube and a blue sphere' sometimes returns a blue cube and a red sphere. Which part of the machinery makes that possible?

A. The tokeniser splits colour words into pieces that are recombined in the wrong order.

B. The sampler visits latent regions in an order that does not match the prompt's word order.

C. Guidance amplifies colour terms more than noun terms, so colours become mobile.

D. Cross-attention lets every latent cell draw on every token, and nothing ties an adjective to its noun.

5 Module 1 · How the picture gets made

You want to compare two prompts fairly, and you want to be able to reproduce whichever wins. Which sampler family should you use, and why?

A. A deterministic sampler, because it follows a fixed path from the starting noise.

B. An ancestral sampler, because the extra noise averages out over many steps.

C. Either, provided you record the step count, since both converge to the same image.

D. An ancestral sampler, because livelier texture makes differences easier to see.

6 Module 1 · How the picture gets made

You need a poster with correctly spelled text in the image. Which architecture is most likely to manage it, and on what grounds?

A. A GAN, because a single forward pass avoids the drift that accumulates over a denoising loop.

B. A flow-matching model, because a straighter path from noise to data preserves fine structure.

C. Any diffusion model at a high enough guidance scale, because guidance enforces prompt fidelity.

D. An autoregressive token model, because it generates the picture as a sequence with language machinery attached.

Module 1 · How the picture gets made

The module starts on p. 11.

MODULE 1 TEST

Module 1 test, Q1, p. 50

On a modern sampler, thirty steps and eighty steps produce images that are close to identical, and the eighty-step version takes nearly three times as…

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

Why: The denoising loop is a numerical approximation of a smooth path from noise to image, and each step is a guess at where that path goes next. Once the steps are small enough for the guess to be accurate, halving them again reduces an error that was already smaller than an eight-bit image can represent. Modern samplers reach that point at twenty to thirty steps; older ones needed fifty to a hundred, and distilled models get there in one to four.

Module 1 test, Q2, p. 50

A model places the letters of a shop sign in the right order in the latent, and the finished image still shows a smear. Which…

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

Why: Diffusion happens in a compressed latent grid, typically eight times smaller in each direction, so one latent cell stands for an eight-by-eight block of the final picture. A letter two cells wide has no room in the decoder to become a letter shape, and what comes back is the statistical impression of small type. No prompt adds latent resolution. The fixes are structural: generate larger, generate the sign separately, or set the type afterwards in a real font.

Module 1 test, Q3, p. 51

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

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

Why: Guidance takes the difference between the prompted and unprompted predictions and travels several times that distance past either of them. At high scale the resulting latent sits outside the distribution the autoencoder ever learned to decode, so the decoder is asked to render values it has never seen. Clipping and posterisation are what an autoencoder does with impossible input. The step count does not change, and the burnt look is a dial turned too far rather than a model defect.

Module 1 test, Q4, p. 51

You send a colleague your prompt, seed, model file, sampler, step count, guidance and resolution. She runs it on her own machine and gets a…

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

Why: Everything in the loop is deterministic arithmetic, but not bit-identical arithmetic across different hardware. Small floating-point differences at step one are amplified by twenty-nine further steps into a visibly different image. Bit-identical reproduction across machines is not something these pipelines promise; what is reproducible is the same machine with the same versions. Seeds do behave differently across models, but she is using the same model file, and a stochastic sampler would also break repeatability on her own machine, which is not what is described.

Module 1 test, Q5, p. 51

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

Answer B: A prompt naming a very famous single painting.

Why: Memorisation is driven by duplication in the training data, not by prompt length or guidance. An image that appeared once is averaged away; one that appeared hundreds of times with a consistent caption becomes a target the optimiser can profitably learn exactly. Famous paintings, film posters, book covers and stock photographs are the duplicated cases. A living illustrator with a thin online presence usually yields a loose style imitation rather than a copy of any one work, which raises a different question, since style is generally outside copyright while a reproduction is not.

Module 1 test, Q6, p. 52

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

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

Why: Prompt style should match the captions a model was trained on. Alt text reads like a keyword list, so a model trained on it is a bag-of-concepts machine and keyword lists suit it. Recaptioning — passing every training image through a vision-language model that writes a long literal description — means the newer model learned from prose, and a tag list is unlike anything in its training set. It will cope, and the relational and spatial vocabulary it was specifically trained on goes unused. Token limits on recaptioned models are generally longer, not shorter.

THE LESSON CHECKS

Lesson 1, p. 16

You generate an image with a fixed seed and like everything except one background element. You change a single word in the prompt and generate…

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

Why: An image model does not draw.

A Only the part of the picture…: Treats the seed as fixing the content of the image rather than only the random starting noise.

B Nothing changes, because with the seed…: Assumes the seed alone determines the result, forgetting the prompt steers every step.

D Only fine detail changes, because the…: Imagines the noise already contains a layout that the prompt merely decorates at the end.

Lesson 2, p. 19

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

Answer A: The sampler's error is already below what the output file can represent at 20 steps, so extra steps refine a path that has already converged.

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

B The model has a fixed internal…: Invents a hard internal limit; the loop genuinely runs sixty times, and the results are near-identical because the numerical path has converged, not because work was thrown away.

C The prompt was too short to…: Treats steps as time the model fills with prompt content; step count controls how finely the denoising path is approximated and is independent of prompt length.

D Step counts above roughly 25 are…: A plausible-sounding safety story with no basis; the steps really run, which is exactly why the slower version cost three times as much.

Lesson 3, p. 23

A model generates a street scene at its native 1024 pixels. The shop sign in the middle distance is a convincing smear of letter-like marks…

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

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

A The text encoder truncated the prompt…: Confuses two different limits; truncation would drop the request entirely, whereas here the model clearly attempted a sign and produced letter-shaped marks.

C The model was not trained on…: Contradicts the evidence in the image; letter-like marks appear precisely because the model has seen a great many signs.

D Small text within a generated scene…: Attributes an ordinary resolution artefact to moderation; safety filters block or blur whole outputs, they do not selectively degrade small type.

MAKE THINGS

Making Things with AI

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

WHAT IT TEACHES	A plain, honest tour of generative media.
WHAT IS INSIDE	Nine modules and 84 lessons, 29 figures drawn from data, nine case studies, nine module tests, a 63-question examination, and every answer with the reason behind it.
LICENCE	CC BY-NC-SA 4.0. Copy, print, share and adapt it for any non-commercial purpose, with credit, under the same licence. There is no paid edition.
THE COURSE	Free to read, with the lessons, the films and the examination, at **addaly.com/learn/making-things-with-ai**

Scan to open the course

Addaly

First edition, September 2026 · build 62a26075