AI Explained · Images, Voice and Video, Part 1

Why Did AI Move the Window?

By Dr Peter Fields

THE 2-MINUTE SUMMARY

An image generator has to make the whole scene hang together. Some models do that by working from noise towards a picture. Others build the image through successive predictions.

A model can draw a convincing sheep and learn facts about sheep from its training. It acquires those associations without the experience of watching an animal graze or raising one.

For an edit, give the tool the original, mark the area if you can and say what to preserve. If a particular detail must stay exact, keep that part of the original in a conventional editor.

The chair and the missing window pane

Imagine asking for a picture of a sunlit Georgian study. The tool gives you a wooden desk, a tall bookcase and a leather armchair blocking the window. You ask it to remove the chair. The chair disappears. So does one of the window panes, and the desk now has a drawer where there was none.

You meant “change this one thing”. The job also requires the tool to work out what should appear where the chair stood. The floor behind it was hidden. The tool has to supply plausible floorboards, continue the skirting and decide what to do with the shadow.

How a picture comes together

A text generator can build a sentence one token at a time. A token is a small piece of text. A picture has to make sense across the whole scene. The desk legs have to meet the floor, the window must sit in the wall and the shadows should agree with the direction of the light.

One way to make a picture is called diffusion. During training, the model learns to reverse the effect of adding noise to examples. When generating a picture from scratch, it begins with noise and makes a series of adjustments, guided by your description. That gives it a way to build up a scene in which the features fit together.[1,2]

Can a model know what a sheep is?

Ask for a photograph of a sheep in a field and you may get a remarkably convincing animal. A system trained on images and descriptions can learn that words like “sheep” tend to go with certain shapes and textures. Training that also includes text can teach it that sheep are living creatures, that they eat grass, and that “lamb” can refer to a young sheep or to meat. Those are useful connections. A model learns from the material it is given. It has no memory of parting wet wool with its fingers or watching a lamb find its feet.[3]

Geek Box · What is latent space?

Some image generators do their work on a compact numerical version of a picture. Researchers call this a latent representation. It keeps enough of the shapes, colours and textures for a separate part of the system to turn the result into visible pixels. Working on the smaller representation can make high-resolution generation less costly.[4]

Do all image generators work this way?

Some image models work more like a text generator building a sentence. OpenAI's 2025 GPT-4o image system card, for example, described a system that builds pictures through successive predictions. The same kind of picture can be produced in different ways.[5]

How to remove the chair and keep the room

Start with the picture you want to keep. Gemini and other conversational image tools can accept follow-up instructions, but the app may also have a separate option to generate a fresh picture. Some systems rewrite your request before passing it on. The result alone gives little away about which part of the process caused an unwanted change.[6]

If the app has a brush or selection tool, mark the chair, including its shadow if you want that gone as well. A selection, sometimes called a mask, guides the edit. Depending on the tool, the changes can go beyond the area you marked.[4,6]

"Remove the leather armchair and its shadow. Continue the floorboards and skirting where the chair stood. Keep the window, desk, bookcase and camera angle as they are."

- **Give it one job at a time.** If you also want a different wall colour, save that for the next edit. Keep a copy of the version you liked so you can go back to it.[6]

- **Compare the result to the original.** Look at the floorboard joins, the edges of the window and the position of the desk. If the window has changed, try the edit again from the version you wanted to keep.

- **Keep exact details in a separate layer.** For a logo, a product label or wording that has to be right, add it separately in a conventional editor. The generated scene can supply the background while you control the details that need to match.

The sheep only has to look right for one moment in a still image. In Part 2, it has to lower its head and take a bite. That brings us to two new questions. How does a moving scene hold together? What rights do you have in what you make?

Sources and notes

1. Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan and Surya Ganguli (2015), Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML 32, pp. 2256–2265, especially §2 and Fig. 1. Conference paper

2. Jonathan Ho, Ajay N. Jain and Pieter Abbeel (2020), Denoising Diffusion Probabilistic Models, NeurIPS 33, pp. 6840–6851, especially §§2–3 and Algorithms 1–2. Conference paper

3. Alec Radford, Jong Wook Kim, Chris Hallacy and colleagues (2021), Learning Transferable Visual Models From Natural Language Supervision, ICML 38, pp. 8748–8763, especially §2 and Fig. 1. The study tests image–text associations, rather than the sheep example. Conference paper

4. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer (2022), High-Resolution Image Synthesis with Latent Diffusion Models, CVPR, pp. 10684–10695, especially §§3, 4.1 and 4.5. Conference paper

5. OpenAI (2025), Addendum to GPT-4o System Card: Native Image Generation, introduction, pp. 1–2. System card

6. OpenAI, Image generation and Image prompting. Google, Nano Banana image generation, section on image editing. Documentation checked 18 November 2026. Image generation · Image prompting · Gemini image editing