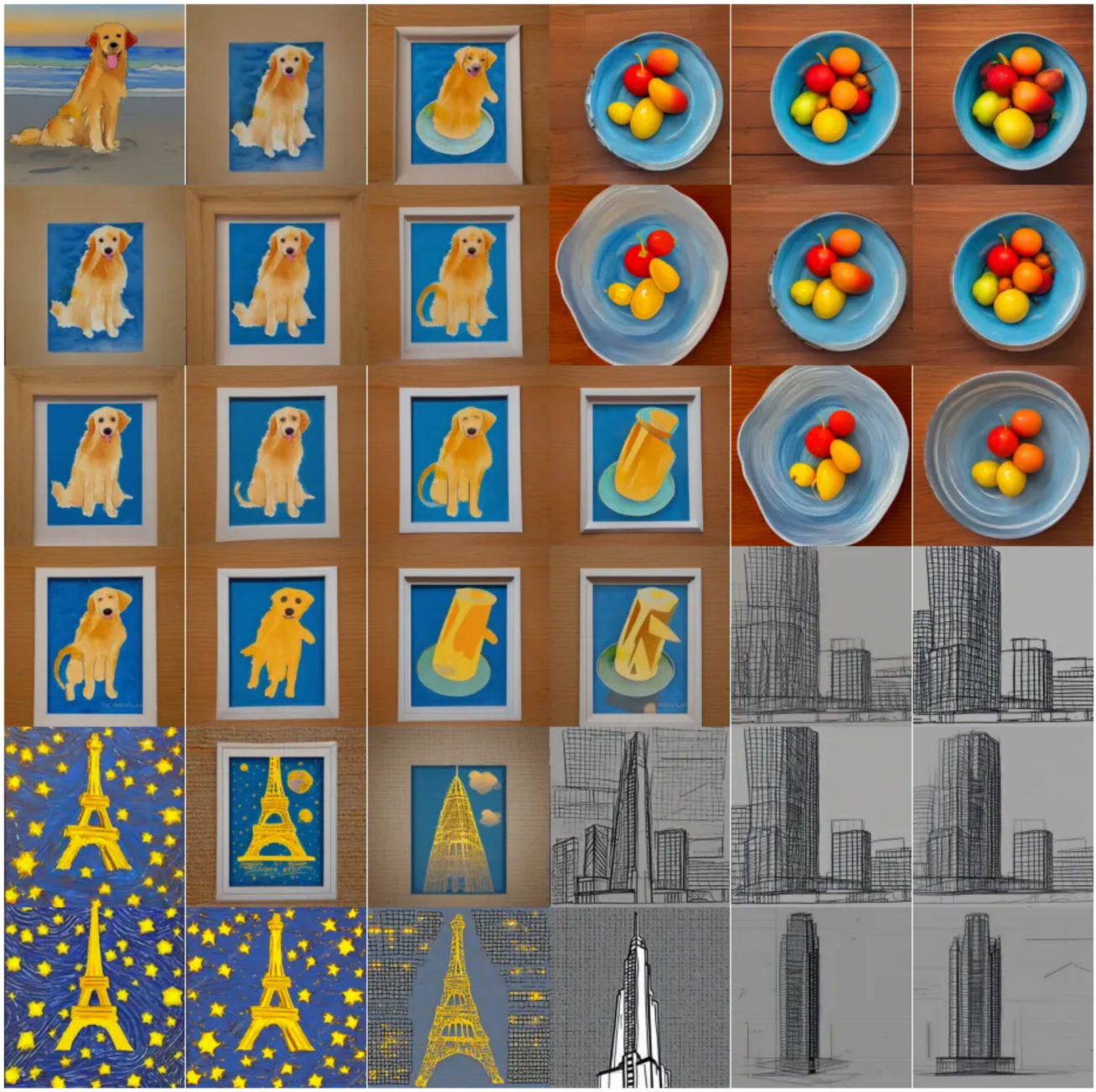
A “complex collage tool?”

The plaintiffs in the class-action lawsuit describe Stable Diffusion as a “complex collage tool” that contains “compressed copies” of its training images. If this were true, the case would be a slam dunk for the plaintiffs.

But experts say it’s not true. Erik Wallace, a computer scientist at the University of California, Berkeley, told me in a phone interview that the lawsuit had “technical inaccuracies” and was “stretching the truth a lot.” Wallace pointed out that Stable Diffusion is only a few gigabytes in size—far too small to contain compressed copies of all or even very many of its training images.

In reality, Stable Diffusion works by first converting a user’s prompt into a latent representation: a list of numbers summarizing the contents of the image. Just as you can identify a point on the Earth’s surface based on its latitude and longitude, Stable Diffusion characterizes images based on their “coordinates” in the “picture space.” It then converts this latent representation into an image.

Let’s make this concrete by looking at an example from this excellent article about Stable Diffusion’s latent space.



Enlarge

If you ask Stable Diffusion to draw "a watercolor painting of a Golden Retriever at the beach," it will produce a picture like the one in the upper-left corner of this image. To do this, it first converts the prompt to a corresponding latent representation—that is, a list of numbers summarizing the elements that are supposed to be in the picture. Maybe a positive value in the 17th position indicates a dog, a negative number in the 54th position represents a beach, a positive value in the 73rd position means a watercolor painting, and so forth.

I just made those numbers up for illustrative purposes; the real latent representation is more complicated and not easy for humans to interpret. In any event, though, there will be a list of numbers that correspond to the prompt, and Stable Diffusion uses this latent representation to generate an image.

The pictures in the other three corners were also generated by Stable Diffusion using the following prompts:

- Upper right: "a still life DSLR photo of a bowl of fruit"
- Lower left: "the eiffel tower in the style of starry night”
- Lower right: "an architectural sketch of a skyscraper”

The point of the six-by-six grid is to illustrate that Stable Diffusion’s latent space is continuous; the software can not only draw an image of a dog or a bowl of fruit, it can also draw images that are “in between” a dog and a bowl of fruit. The third picture on the top row, for example, depicts a slightly fruit-looking dog sitting on a blue dish.

Or look along the bottom row. As you move from left to right, the shape of the building gradually changes from the Eiffel Tower to a skyscraper, while the style changes from a Van Gogh painting to an architectural sketch.

The continuous nature of Stable Diffusion’s latent space enables the software to generate latent representations—and hence images—for concepts that were not in its training data. There probably wasn’t an “Eiffel Tower drawn in the style of ‘Starry Night’” image in Stable Diffusion’s training set. But there were many images of the Eiffel Tower and, separately, many images of “Starry Night.” Stable Diffusion learned from these images and was then able to produce an image that reflected both concepts.