No elephants: Breakthroughs in image generation

When Language Models Learn to See and Create

ETHAN MOLLICK

MAR 30, 2025

Over the past two weeks, first Google and then OpenAI rolled out their multimodal image generation abilities. This is a big deal. Previously, when a Large Language Model AI generated an image, it wasn't really the LLM doing the work. Instead, the would send a text prompt to a separate image generation tool and show you what ca back. The AI creates the text prompt, but another, less intelligent system creates th image. For example, if prompted "*show me a room with no elephants in it, make sure to annotate the image to show me why there are no possible elephants*" the less intelligent image generation system would see the word elephant multiple times and add them the picture. As a result, AI image generations were pretty mediocre with distorted t and random elements; sometimes fun, but rarely useful.

Multimodal image generation, on the other hand, lets the AI directly control the im being made. While there are lots of variations (and the companies keep some of the methods secret), in multimodal image generation, images are created in the same w that LLMs create text, a token at a time. Instead of adding individual words to mak sentence, the AI creates the image in individual pieces, one after another, that are assembled into a whole picture. This lets the AI create much more impressive, exacting images. Not only are you guaranteed no elephants, but the final results of t image creation process reflect the intelligence of the LLM's "thinking", as well as cl writing and precise control.

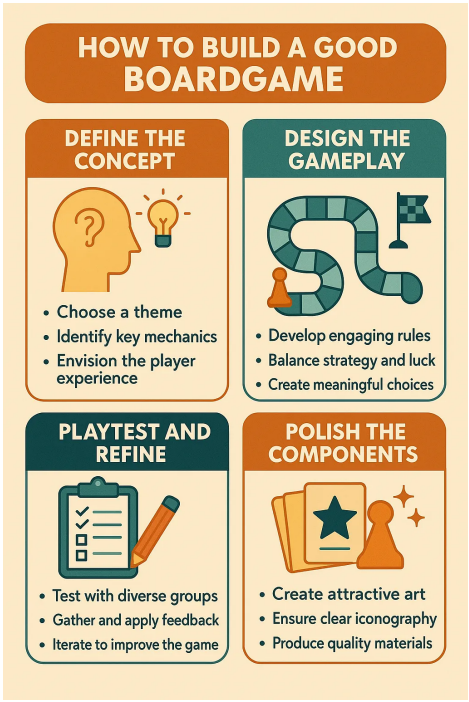
The results of the prompt “show me a room with no elephants in it, make sure to annotate the image to show me why there are no possible elephants” in Microsoft Copilot’s traditional image generator (left), and GPT-4o’s multimodal model (right). Note the traditional model not only shows multiple elephants but also features distorted text.

While the implications of these new image models are vast (and I'll touch on some issues later), let's first explore what these systems can actually do through some examples.

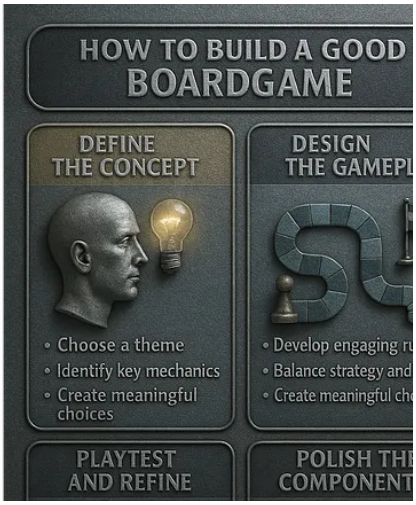
Prompting, but for images

In my book and in many posts, I talk about how a useful way to prompt AI is to trea like a person, even though it isn’t. Giving clear directions, feedback as you iterate, a appropriate context to make a decision all help humans, and they also help AI. Previously, this was something you could only do with text, but now you can do it w images as well.

For example, I prompted GPT-4o “create an infographic about how to build a good boardgame.” With previous image generators, this would result in nonsense, as ther was no intelligence to guide the image generation so words and images would be distorted. Now, I get a good first pass the first time around. However, I did not prov context about what I was looking for, or any additional content, so the AI made all t creative choices. What if I want to change it? Let’s try.

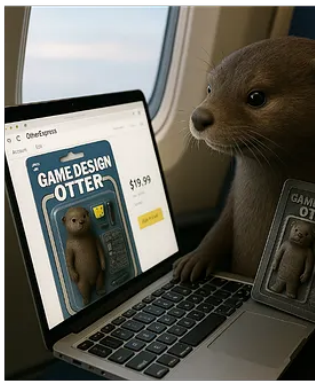


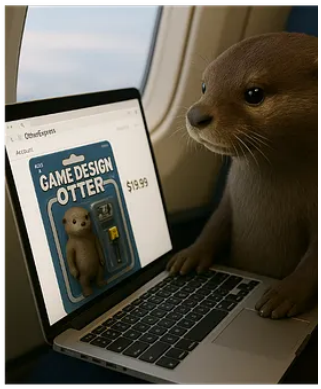
First, I asked it *"make the graphics look hyper realistic instead"* and you can see how it took the concepts from the initial draft and updated their look. I had more changes wanted: *"I want the colors to be less earth toned and more like textured metal, keep everyt else the same, also make sure the small bulleted text is lighter so it is easier to read."* I liked the new look, but I noticed an error had been introduced, the word "Define" had become "Definc" - a sign that these systems, as good as they are, are not yet close to perfect. I prompted "*You spelled Define as Definc, please fix*" and got a reasonable output.



But the fascinating thing about these models is that they are capable of producing almost any image: "*put this infographic in the hands of an otter standing in front of a volcano, it should look like a photo and like the otter is holding this carved onto a metal tablet*"

Why stop there? "*it is night, the tablet is illuminated by a flashlight shining directly at the center of the tablet (no need to show the flashlight)*"— the results of this are more impressive than it might seem because it was redoing the lighting without any sort underlying lighting model. "*Make an action figure of the otter, complete with packaging make the board game one of the accessories on the side. Call it "Game Design Otter" and g it a couple other accessories.*" "*Make an otter on an airplane using a laptop, they are buyin copy of Game Design Otter on a site called OtterExpress.*" Impressive, but not quite righ "*fix the keyboard so it is realistic and remove the otter action figure he is holding.*"



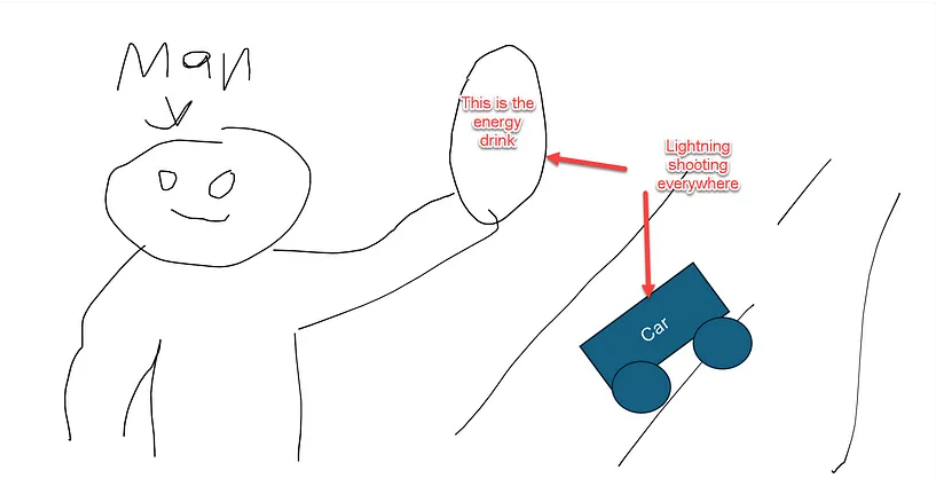


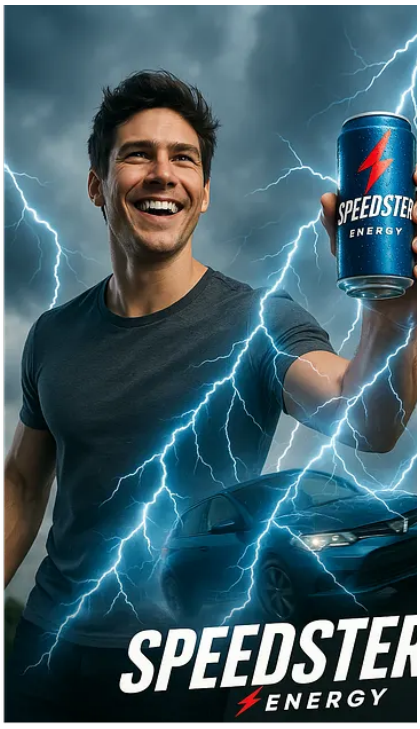
As you can see these systems are not flawless… but also remember that the pictures below are what the results of the prompt “otter on an airplane using wifi” looked lil two and a half years ago. The state-of-the-art is advancing rapidly.

But what is it good for?

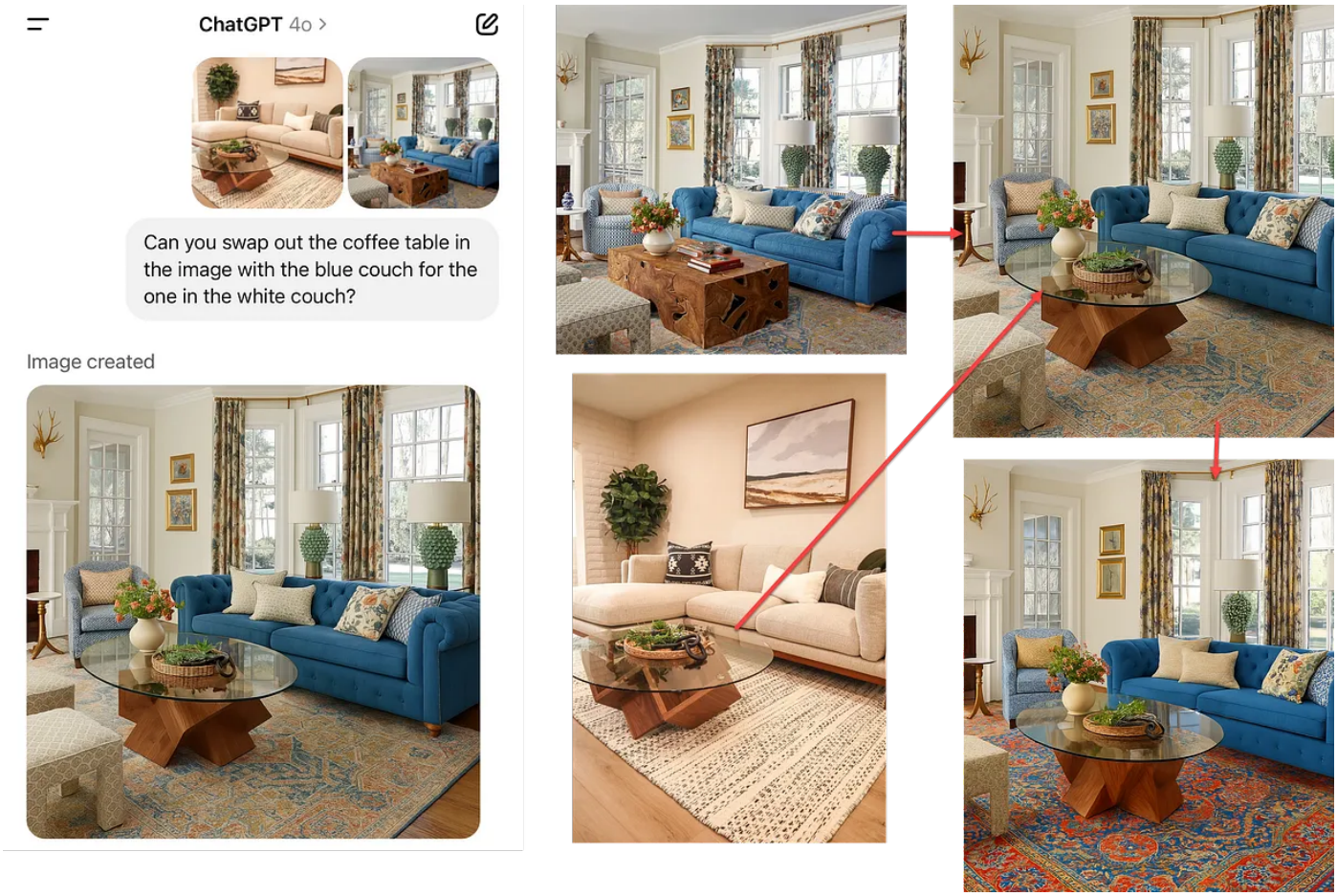
The past couple years have been spent trying to figure out what text AI models are good for, and new use cases are being developed continuously. It will be the same w image-based LLMs. Image generation is likely to be very disruptive in ways we don understand right now. This is especially true because you can upload images that th LLM can now directly see and manipulate. Some examples, all done using GPT-4o (though you can also upload and create images in Google's Gemini Flash):

I can take a hand-drawn image and ask the AI to "*make this an ad for Speedster Energ drink, make sure the packaging and logo are awesome, this should look like a photograph.*" (This took two prompts, the first time it misspelled Speedster on the label). The resu are not as good as a professional designer could create but are an impressive first prototype.

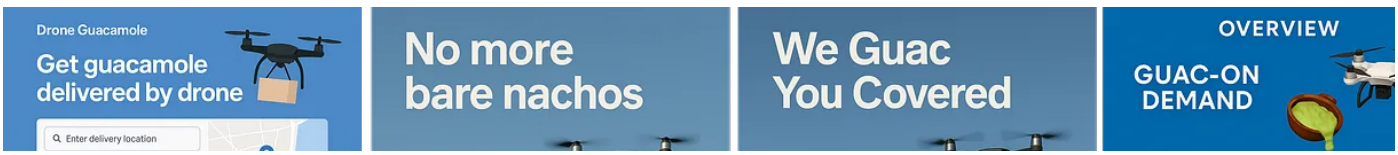




I can give GPT-4o two photographs and the prompt "*Can you swap out the coffee tabl the image with the blue couch for the one in the white couch?*" (Note how the new glass tabletop shows parts of the image that weren't there in the original. On the other hand, the table that was swapped is not exactly the same). I then asked, "*Can you ma the carpet less faded*?" Again, there are several details that are not perfect, but this sc of image editing in plain English was impossible before.



Or I can create an instant website mockup, ad concepts, and pitch deck for my terri startup idea where a drone delivers guacamole to you on demand (pretty sure it is going to be a hit). You can see this is not yet a substitute for the insights of a humar designer, but it is still a very useful first prototype.



Adding to this, there are many other uses that I and others are discovering includin Visual recipes, homepages, textures for video games, illustrated poems, unhinged monologues, photo improvements, and visual adventure games, to name just a few.

Complexities

If you have been following the online discussion over these new image generators, y probably noticed that I haven't demonstrated their most viral use - doing style trans where people ask AI to convert photos into images that look like they were made fo the Simpsons or by Studio Ghibli. These sorts of application highlight all of the complexities of using AI for art: Is it okay to reproduce the hard-won style of other artists using AI? Who owns the resulting art? Who profits from it? Which artists ar in the training data for AI, and what is the legal and ethical status of using copyrighted work for training? These were important questions before multimodal but now developing answers to them is increasingly urgent. Plus, of course, there ar many other potential risks associated with multimodal AI. Deepfakes have been tri to make for at least a year, but multimodal AI makes it easier, including adding the ability to create all sorts of other visual illusions, like fake receipts. And we don't ye understand what biases or other issues multimodal AIs might bring into image generation.

Yet it is clear that what has happened to text will happen to images, and eventually video and 3D environments. These multimodal systems are reshaping the landscape

visual creation, offering powerful new capabilities while raising legitimate questior about creative ownership and authenticity. The line between human and AI creatio will continue to blur, pushing us to reconsider what constitutes originality in a wor where anyone can generate sophisticated visuals with a few prompts. Some creative professions will adapt; others may be unchanged, and still others may transform entirely. As with any significant technological shift, we'll need well-considered frameworks to navigate the complex terrain ahead. The question isn't whether thes tools will change visual media, but whether we'll be thoughtful enough to shape tha change intentionally.