Menu

Creating video from text

Sora is an AI model that can create realistic and imaginative scenes from text instructions.

Read technical report

All videos on this page were generated directly by Sora without modification.

We're teaching AI to understand and simulate the physical world in motion, with the goal of training models that help people solve problems that require real-world interaction.

Capabilities Safety Research

Introducing Sora, our text-to-video model. Sora can generate videos up to a minute long while maintaining visual quality and adherence to the user's prompt.

1 of 9

ompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and
mated city signage. She wears a black leather jacket, a long red dress, and black boots,...
re

Today, Sora is becoming available to red teamers to assess critical areas for harms or risks. We are also granting access to a number of visual artists, designers, and filmmakers to gain feedback on how to advance the model to be most helpful for creative professionals.

We're sharing our research progress early to start working with and getting feedback from people outside of OpenAI and to give the public a sense of what AI capabilities are on the horizon.

ompt: Historical footage of California during the gold rush.

Sora is able to generate complex scenes with multiple characters, specific types of motion, and accurate details of the subject and background. The model understands not only what the user has asked for in the prompt, but also how those things exist in the physical world.

ɔmpt: The camera follows behind a white vintage SUV with a black roof rack as it speeds a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicks up fro... ɹre

The model has a deep understanding of language, enabling it to accurately interpret prompts and generate compelling characters that express vibrant emotions. Sora can also create multiple shots within a single generated video that accurately persist characters and visual style.

1 of 8

ompt: Tour of an art gallery with many beautiful works of art in different styles.

The current model has weaknesses. It may struggle with accurately simulating the physics of a complex scene, and may not understand specific instances of cause and effect. For example, a person might take a bite out of a cookie, but afterward, the cookie may not have a bite mark.

The model may also confuse spatial details of a prompt, for example, mixing up left and right, and may struggle with precise descriptions of events that take place over time, like following a specific camera trajectory.

1 of 5

ɔmpt: Step-printing scene of a person running, cinematic film shot in 35mm.

ɜakness: Sora sometimes creates physically implausible motion.

Safety

We'll be taking several important safety steps ahead of making Sora available in OpenAI's products. We are working with red teamers—domain experts in areas like misinformation, hateful content, and bias—who will be adversarially testing the model.

We're also building tools to help detect misleading content such as a detection classifier that can tell when a video was generated by Sora. We plan to include C2PA metadata in the future if we deploy the model in an OpenAI product.

In addition to us developing new techniques to prepare for deployment, we're leveraging the existing safety methods that we built for our products that use DALL·E 3, which are applicable to Sora as well.

For example, once in an OpenAI product, our text classifier will check and reject text input prompts that are in violation of our usage policies, like those that request extreme violence, sexual content, hateful imagery, celebrity likeness, or the IP of others. We've also developed robust image classifiers that are used to review the frames of every video generated to help ensure that it adheres to our usage policies, before it's shown to the user.

We'll be engaging policymakers, educators and artists around the world to understand their concerns and to identify positive use cases for this new technology. Despite extensive research and testing, we cannot predict all of the beneficial ways people will use our technology, nor all the ways people will abuse it. That's why we believe that learning from real-world use is a critical component of creating and releasing increasingly safe AI systems over time.

1 of 10

ompt: The camera directly faces colorful buildings in Burano Italy. An adorable dalmation
ks through a window on a building on the ground floor. Many people are walking and...
re

Research techniques

Sora is a diffusion model, which generates a video by starting off with one that looks like static noise and gradually transforms it by removing the noise over many steps.

Sora is capable of generating entire videos all at once or extending generated videos to make them longer. By giving the model foresight of many frames at a time, we've solved a challenging problem of making sure a subject stays the same even when it goes out of view temporarily.

Similar to GPT models, Sora uses a transformer architecture, unlocking superior scaling performance.

We represent videos and images as collections of smaller units of data called patches, each of which is akin to a token in GPT. By unifying how we represent data, we can train diffusion transformers on a wider range of visual data than was possible before, spanning different durations, resolutions and aspect ratios.

Sora builds on past research in DALL·E and GPT models. It uses the recaptioning technique from DALL·E 3, which involves generating highly descriptive captions for the visual training data. As a result, the model is able to follow the user's text instructions in the generated video more faithfully.

In addition to being able to generate a video solely from text instructions, the model is able to take an existing still image and generate a video from it, animating the image's contents with accuracy and attention to small detail. The model can also take an existing video and extend it or fill in missing frames. Learn more in our technical report.

Sora serves as a foundation for models that can understand and simulate the real world, a capability we believe will be an important milestone for achieving AGI.

Research Leads

Bill Peebles & Tim Brooks

Systems Lead

Connor Holmes

Contributors
Clarence Ng
David Schnurr
Eric Luhman
Joe Taylor
Li Jing
Natalie Summers
Ricky Wang
Rohan Sahai
Ryan O'Rourke
Troy Luhman
Will DePue
Yufei Guo

Special Thanks

Bob McGrew, Brad Lightcap, Chad Nelson, David Medina, Gabriel Goh, Greg Brockman, Ian Sohl, Jamie Kiros, James Betker, Jason Kwon, Hannah Wong, Mark Chen, Michelle Fradin, Mira Murati, Nick Turley, Prafulla Dhariwal, Rowan Zellers, Sam Altman, Sandhini Agarwal, Sarah Yoo, Srinivas Narayanan & Wesam Manassra

Communications
Elie Georges
Justin Wang
Kendra Rimbach
Niko Felix
Thomas Degry
Veit Moeller

Legal
Che Chang
Fred von Lohmann
Gideon Myles
Tom Stasi

External Engagement

Alex Baker-Whitcomb, Allie Teague, Anna Makanju, Anna McKean, Becky Waite, Brittany Smith, Chan Park, Chris Lehane, David Duxin, David Robinson, James Hairston,

Jonathan Lachman, Justin Oswald, Krithika Muthukumar, Lane Dilg, Leher Pathak,
Ola Nowicka, Ryan Biddy, Sandro Gianella, Stephen Petersilge, Tom Rubin & Varun Shetty

Executive Producer

Aditya Ramesh

Built by OpenAI in San Francisco, California
Published February 15, MMXXIV

Social
Twitter
YouTube
GitHub
SoundCloud
LinkedIn

Back to top