

February 15, 2024

# Video generation models as world simulators

[View Sora overview](#)

We explore large-scale training of generative models on video data. Specifically, we train text-conditional diffusion models jointly on videos and images of variable durations, resolutions and aspect ratios. We leverage a transformer architecture that operates on spacetime patches of video and image latent codes. Our largest model, Sora, is capable of generating a minute of high fidelity video. Our results suggest that scaling video generation models is a promising path towards building general purpose simulators of the physical world.

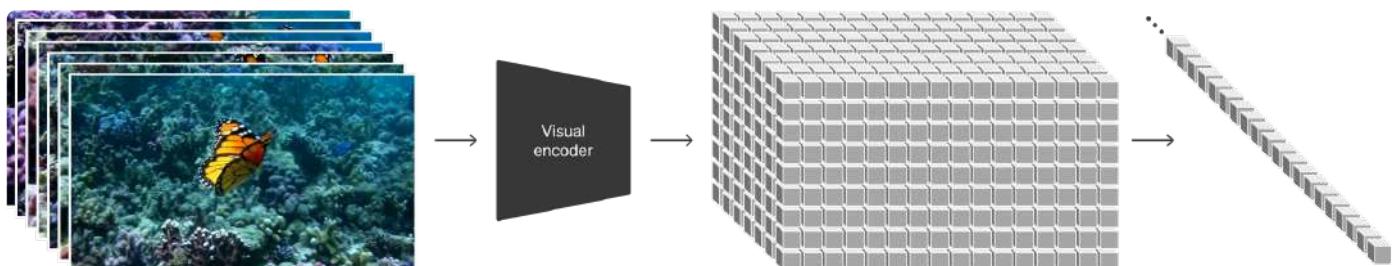
0:00 / 0:59

This technical report focuses on (1) our method for turning visual data of all types into a unified representation that enables large-scale training of generative models, and (2) qualitative evaluation of Sora’s capabilities and limitations. Model and implementation details are not included in this report.

Much prior work has studied generative modeling of video data using a variety of methods, including recurrent networks,<sup>1,2,3</sup> generative adversarial networks,<sup>4,5,6,7</sup> autoregressive transformers,<sup>8,9</sup> and diffusion models.<sup>10,11,12</sup> These works often focus on a narrow category of visual data, on shorter videos, or on videos of a fixed size. Sora is a generalist model of visual data—it can generate videos and images spanning diverse durations, aspect ratios and resolutions, up to a full minute of high definition video.

## Turning visual data into patches

We take inspiration from large language models which acquire generalist capabilities by training on internet-scale data.<sup>13,14</sup> The success of the LLM paradigm is enabled in part by the use of tokens that elegantly unify diverse modalities of text—code, math and various natural languages. In this work, we consider how generative models of visual data can inherit such benefits. Whereas LLMs have text tokens, Sora has visual *patches*. Patches have previously been shown to be an effective representation for models of visual data.<sup>15,16,17,18</sup> We find that patches are a highly-scalable and effective representation for training generative models on diverse types of videos and images.



At a high level, we turn videos into patches by first compressing videos into a lower-dimensional latent space,<sup>19</sup> and subsequently decomposing the representation into spacetime patches.

## Video compression network

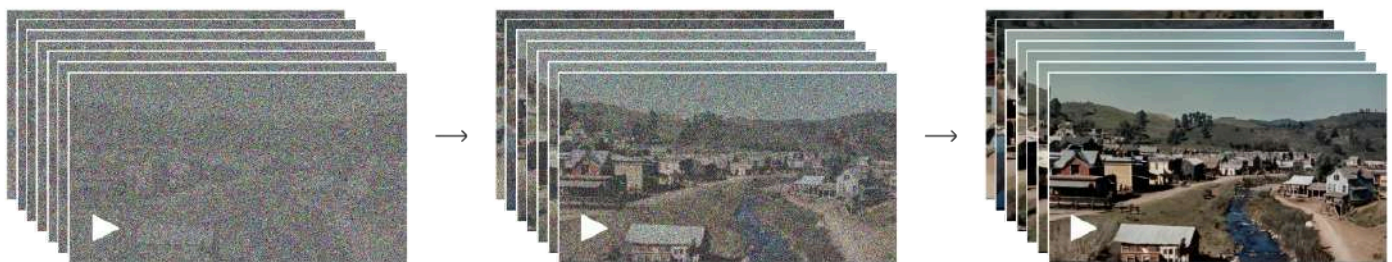
We train a network that reduces the dimensionality of visual data.<sup>20</sup> This network takes raw video as input and outputs a latent representation that is compressed both temporally and spatially. Sora is trained on and subsequently generates videos within this compressed latent space. We also train a corresponding decoder model that maps generated latents back to pixel space.

## Spacetime latent patches

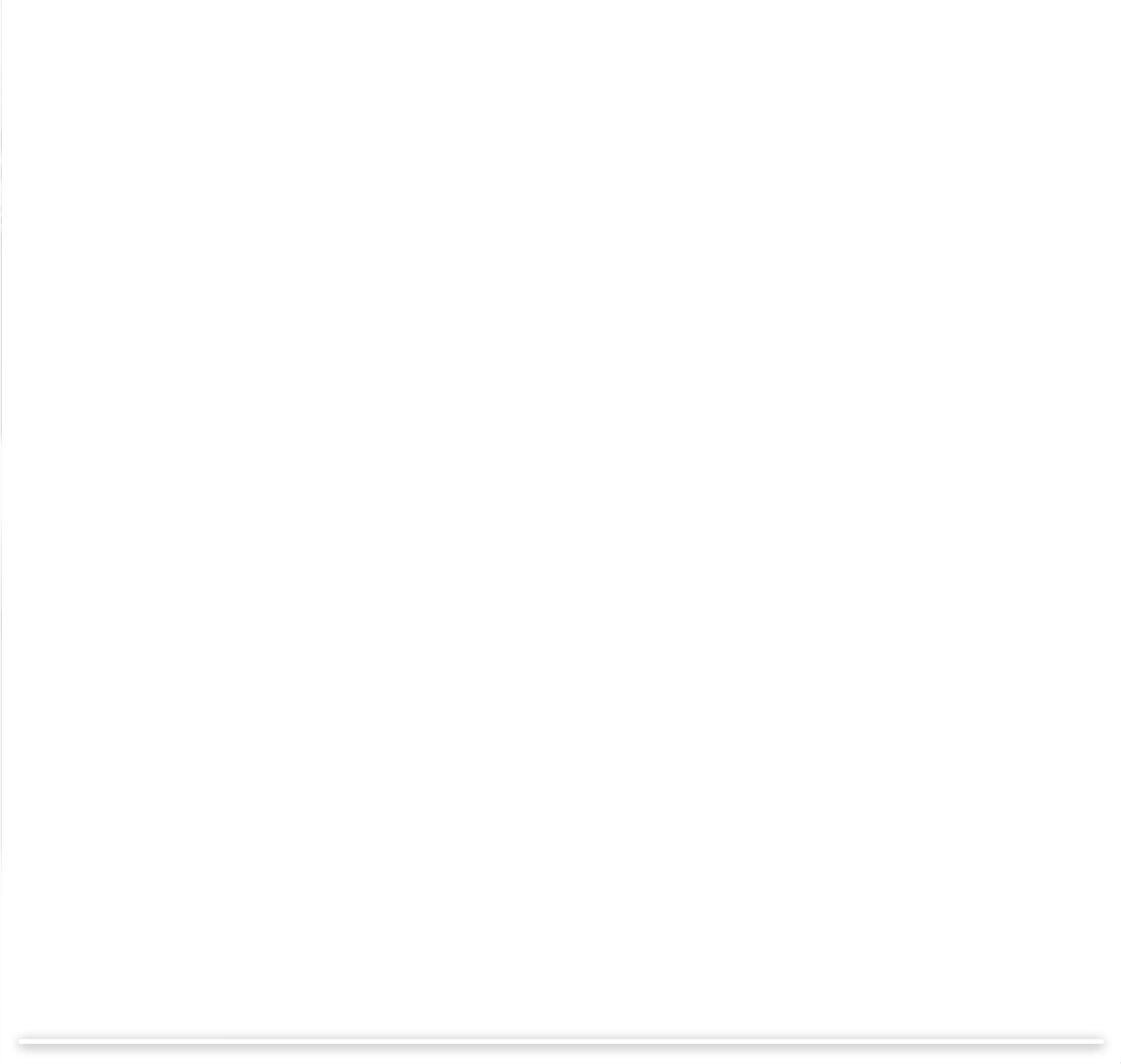
Given a compressed input video, we extract a sequence of spacetime patches which act as transformer tokens. This scheme works for images too since images are just videos with a single frame. Our patch-based representation enables Sora to train on videos and images of variable resolutions, durations and aspect ratios. At inference time, we can control the size of generated videos by arranging randomly-initialized patches in an appropriately-sized grid.

## Scaling transformers for video generation

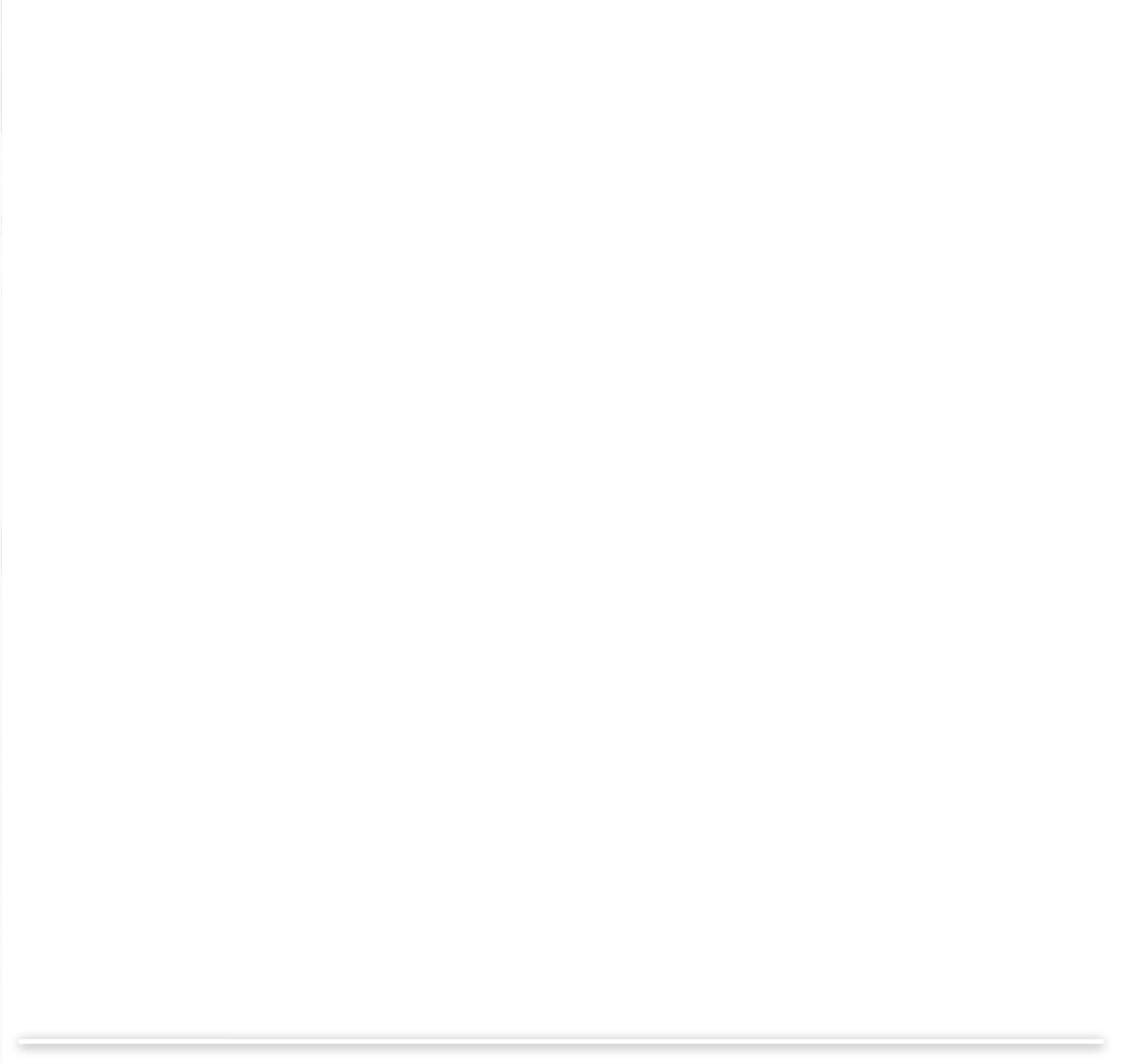
Sora is a diffusion model<sup>21, 22, 23, 24, 25</sup>; given input noisy patches (and conditioning information like text prompts), it's trained to predict the original "clean" patches. Importantly, Sora is a diffusion *transformer*.<sup>26</sup> Transformers have demonstrated remarkable scaling properties across a variety of domains, including language modeling,<sup>13, 14</sup> computer vision,<sup>15, 16, 17, 18</sup> and image generation.<sup>27, 28, 29</sup>



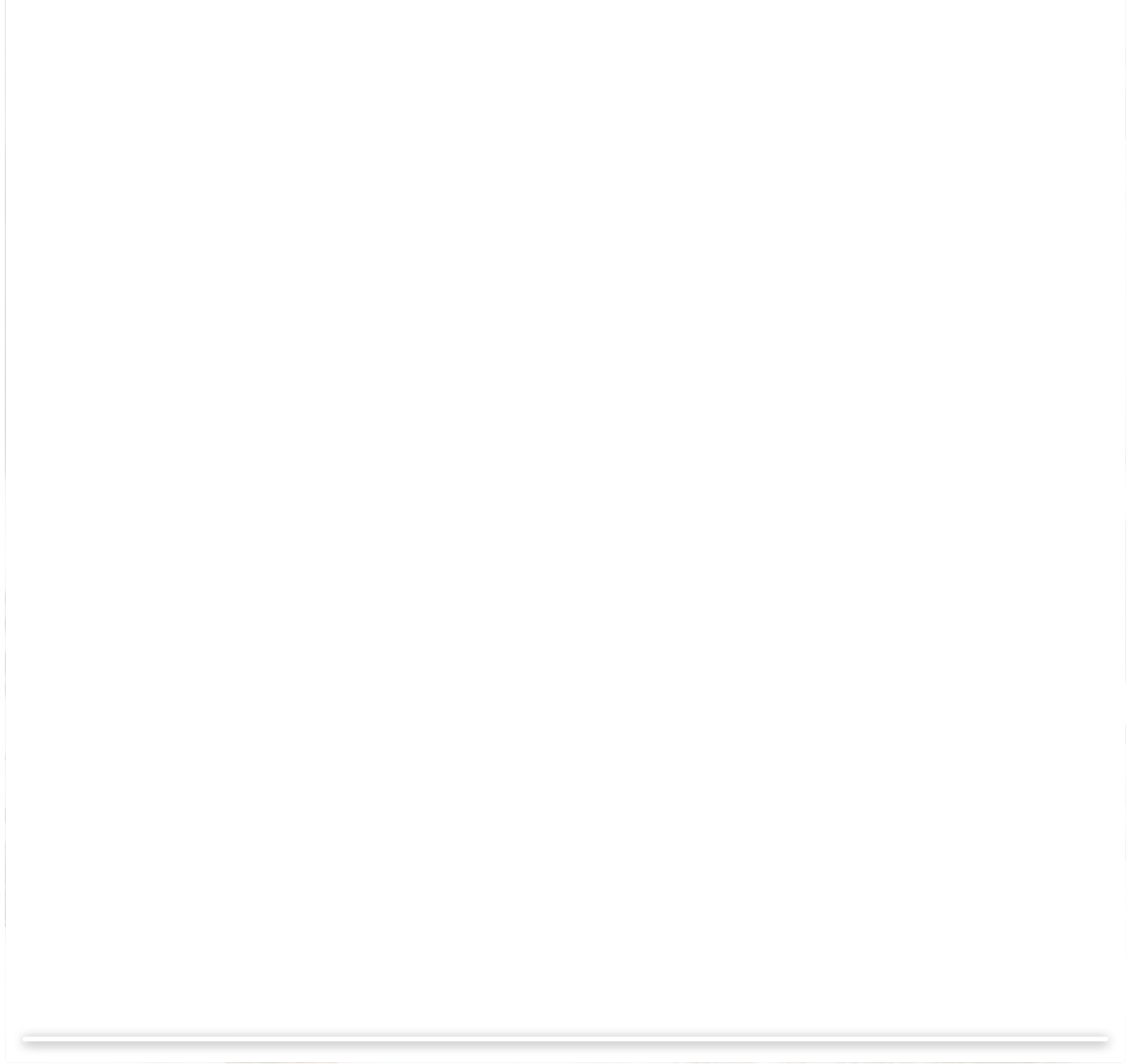
In this work, we find that diffusion transformers scale effectively as video models as well. Below, we show a comparison of video samples with fixed seeds and inputs as training progresses. Sample quality improves markedly as training compute increases.



Base compute



4x compute



32x compute

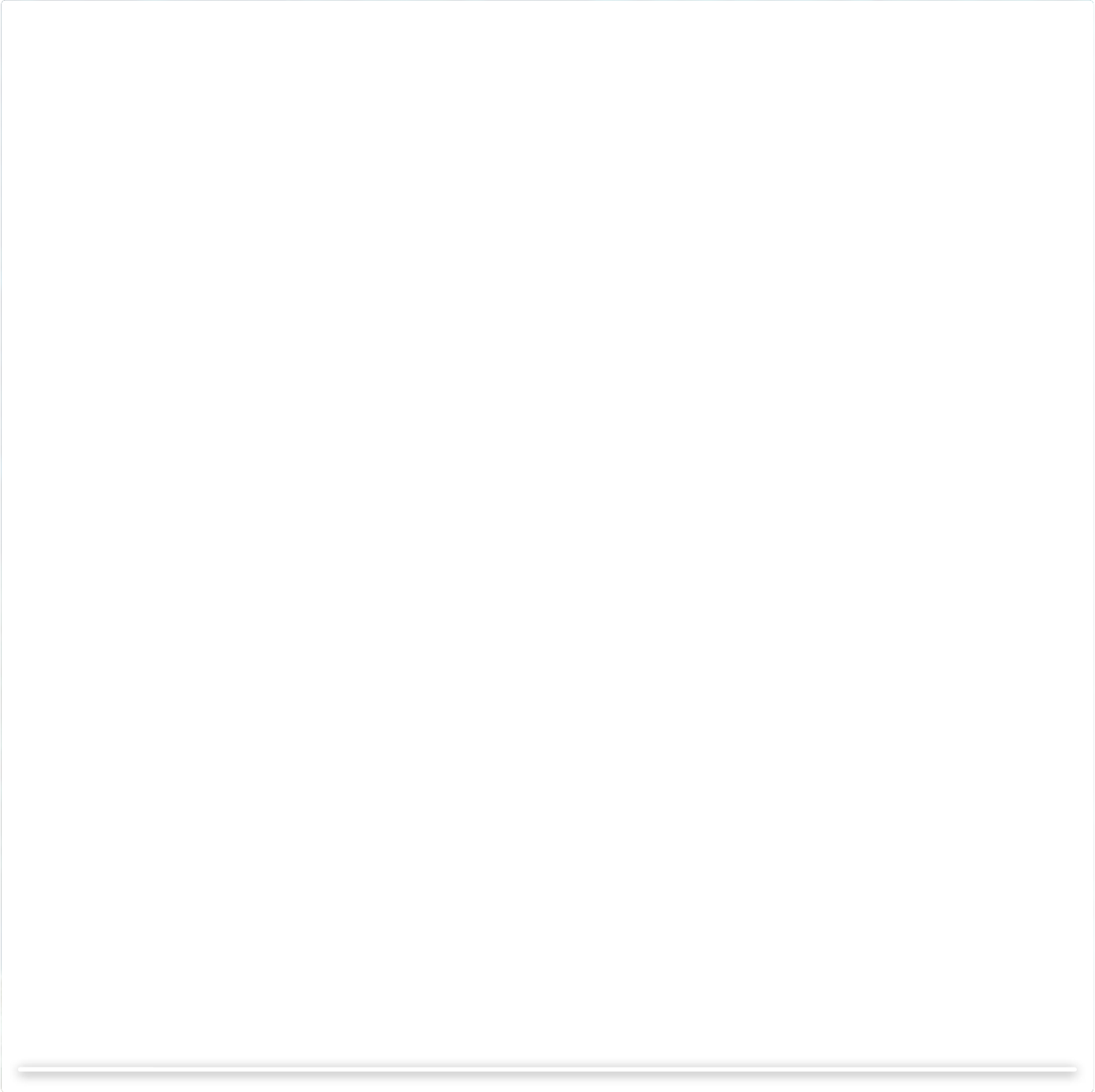
## Variable durations, resolutions, aspect ratios

Past approaches to image and video generation typically resize, crop or trim videos to a standard size—e.g., 4 second videos at 256×256 resolution. We find that instead training on data at its native size provides several benefits.

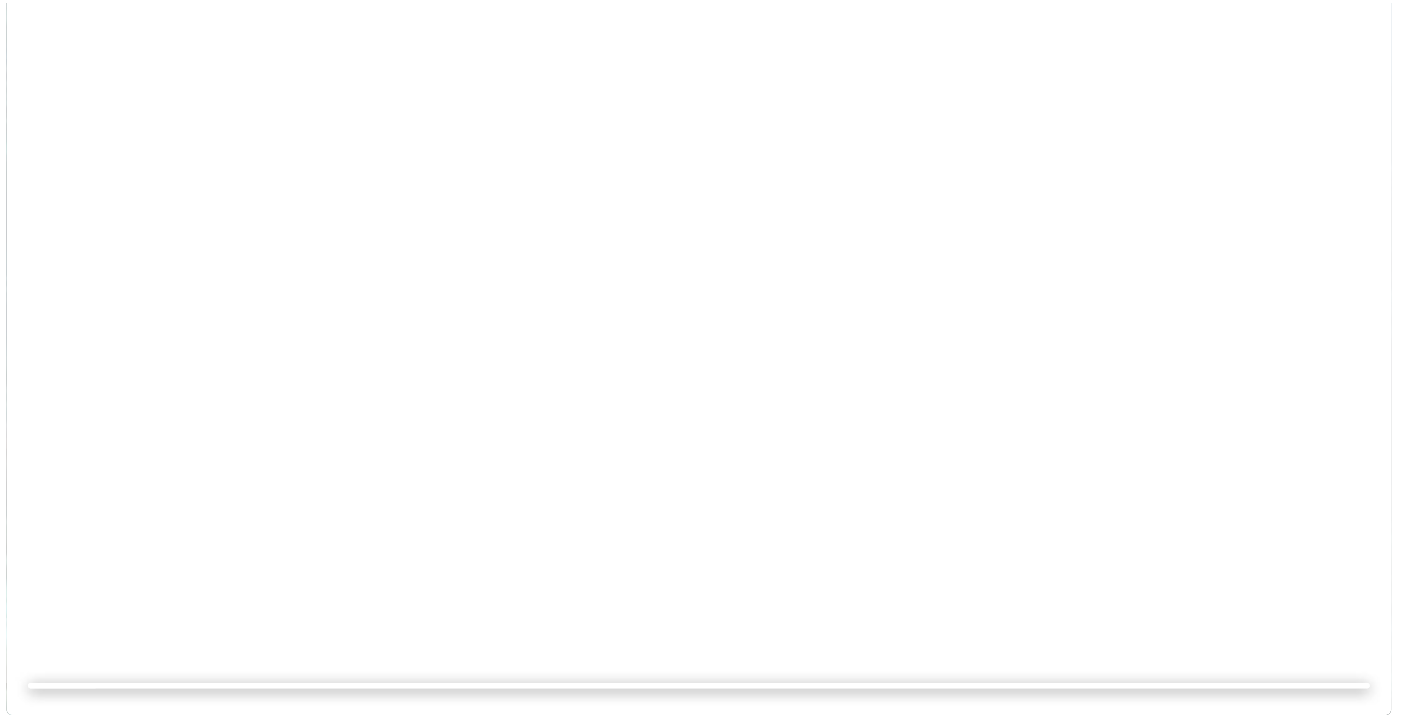
## Sampling flexibility

Sora can sample widescreen 1920×1080p videos, vertical 1080×1920 videos and everything inbetween. This lets Sora create content for different devices directly at their native aspect ratios. It also lets us quickly prototype content at lower sizes before generating at full resolution—all with the same model.





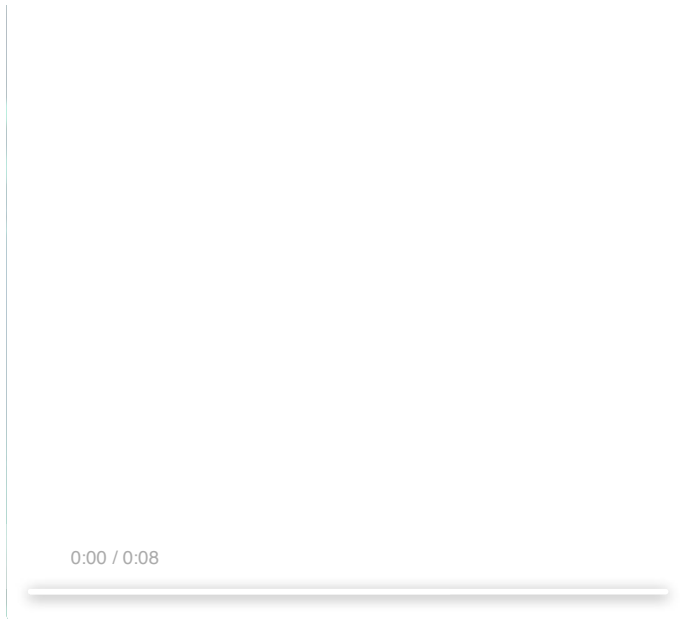
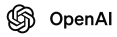




## Improved framing and composition

We empirically find that training on videos at their native aspect ratios improves composition and framing. We compare Sora against a version of our model that crops all training videos to be square, which is common practice when training generative models. The model trained on square crops (left) sometimes generates videos where the subject is only partially in view. In comparison, videos from Sora (right) have improved framing.





## Language understanding

Training text-to-video generation systems requires a large amount of videos with corresponding text captions. We apply the re-captioning technique introduced in DALL-E 3<sup>30</sup> to videos. We first train a highly descriptive captioner model and then use it to produce text captions for all videos in our training set. We find that training on highly descriptive video captions improves text fidelity as well as the overall quality of videos.

Similar to DALL-E 3, we also leverage GPT to turn short user prompts into longer detailed captions that are sent to the video model. This enables Sora to generate high quality videos that accurately follow user prompts.



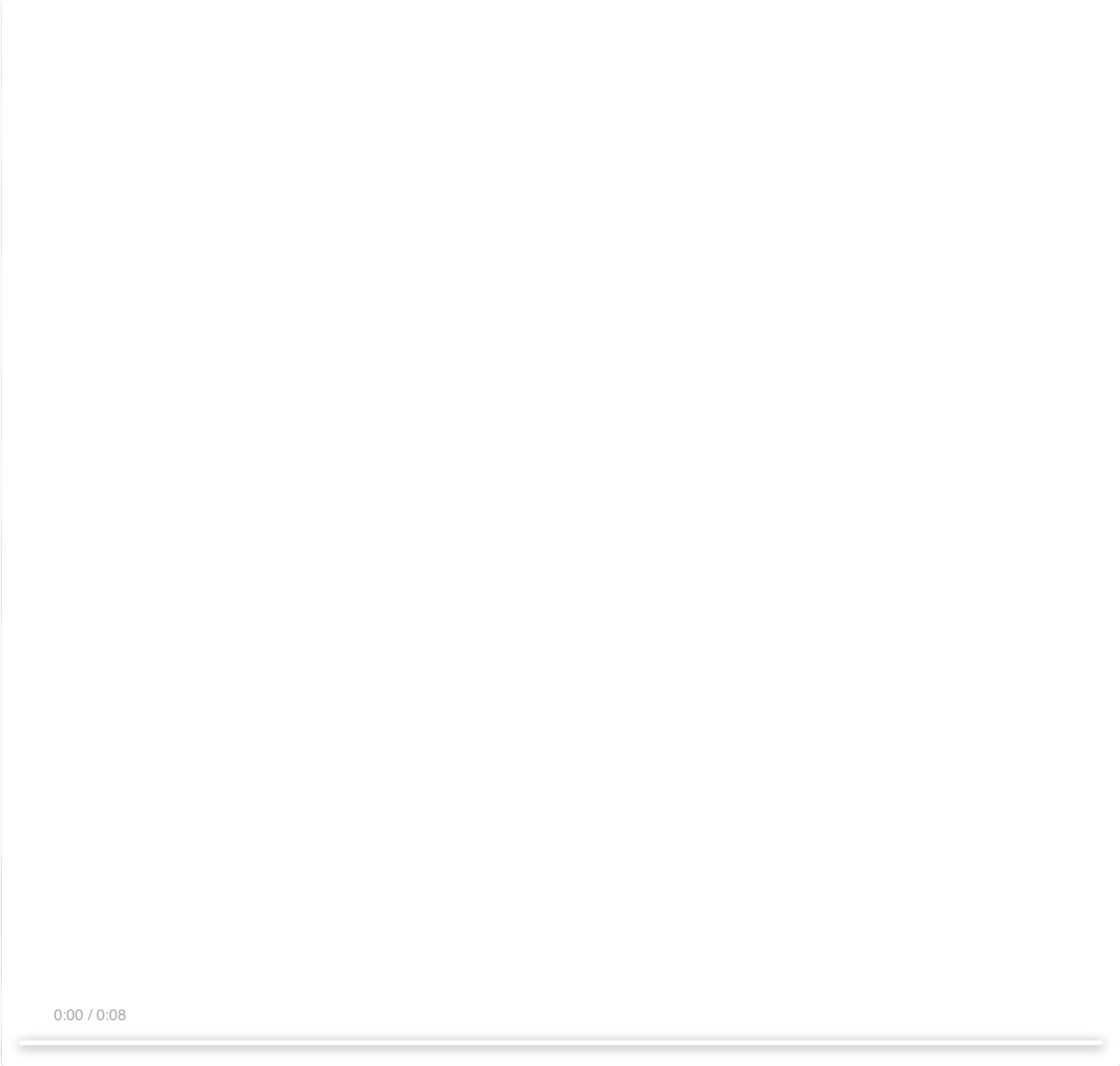
## Prompting with images and videos

All of the results above and in our [landing page](#) show text-to-video samples. But Sora can also be prompted with other inputs, such as pre-existing images or video. This capability enables Sora to perform a wide range of image and video editing tasks—creating perfectly looping video, animating static images, extending videos forwards or backwards in time, etc.

## Animating DALL·E images

Sora is capable of generating videos provided an image and prompt as input. Below we show example videos generated based on DALL·E 2<sup>31</sup> and DALL·E 3<sup>30</sup> images.



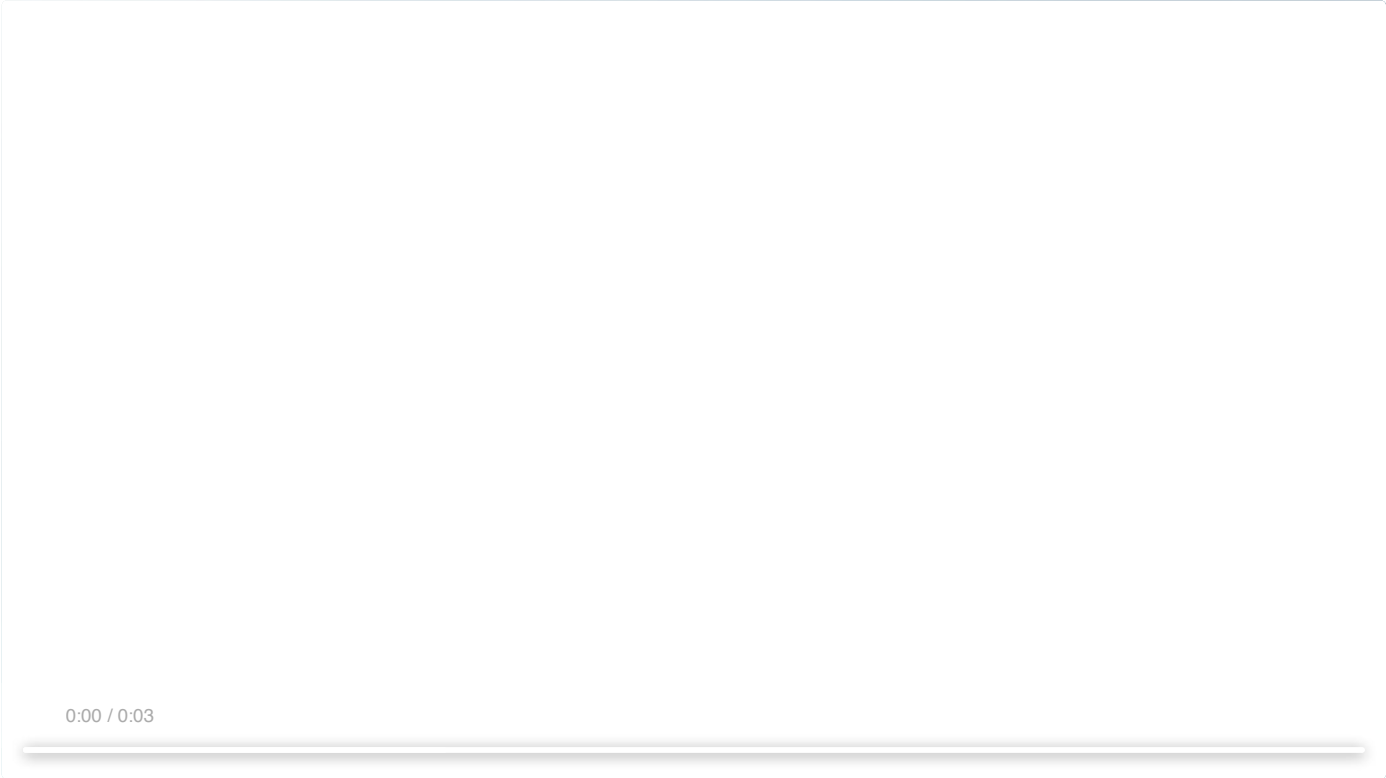


A Shiba Inu dog wearing a beret and black turtleneck.



0:00 / 0:08

Monster Illustration in flat design style of a diverse family of monsters. The group includes a furry brown monster, a sleek black monster with antennas, a spotted green monster, and a tiny polka-dotted monster, all interacting in a playful environment.



An image of a realistic cloud that spells "SORA".





0:00 / 0:09

In an ornate, historical hall, a massive tidal wave peaks and begins to crash. Two surfers, seizing the moment, skillfully navigate the face of the wave.

## Extending generated videos

Sora is also capable of extending videos, either forward or backward in time. Below are three videos that were all extended backward in time starting from a segment of a generated video. As a result, each of the three videos starts different from the others, yet all three videos lead to the same ending.

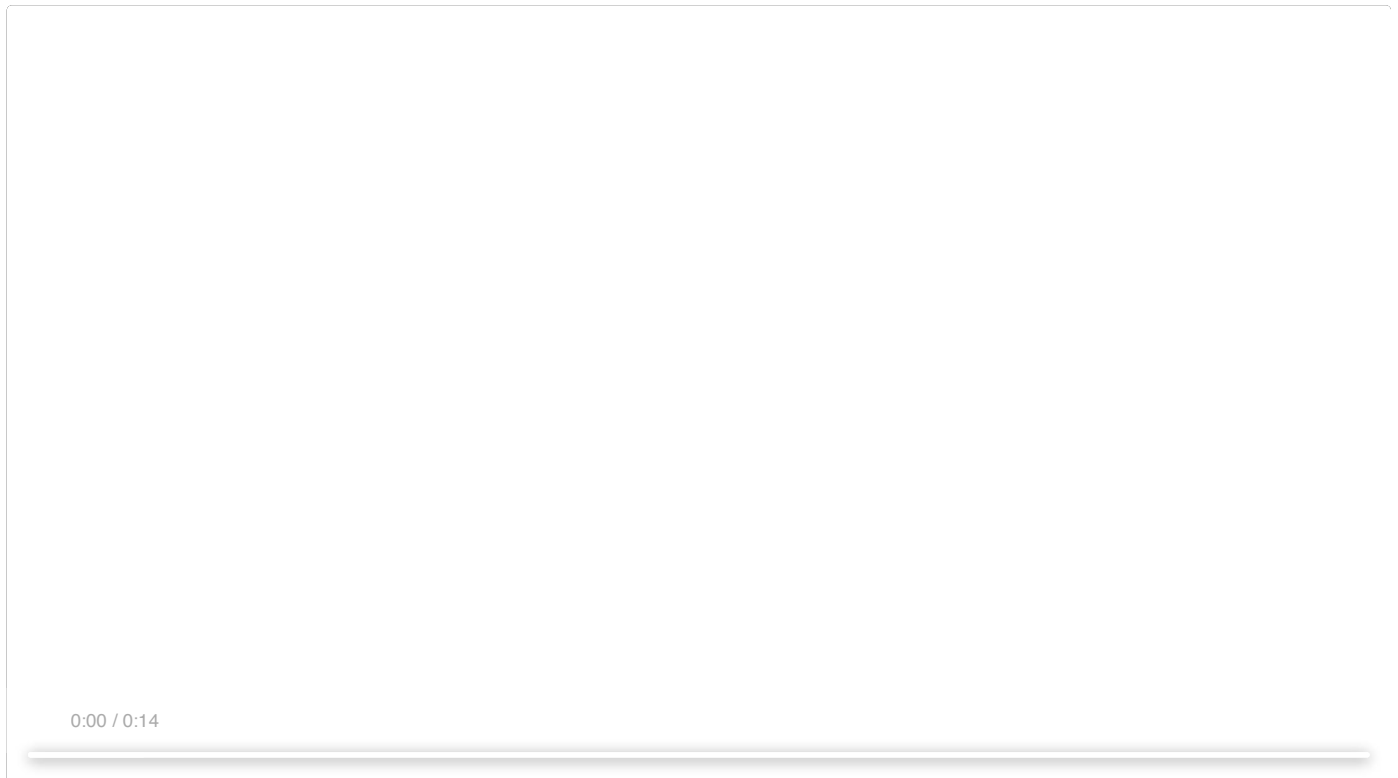




00:00  00:20



We can use this method to extend a video both forward and backward to produce a seamless infinite loop.



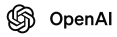
0:00 / 0:14

## Video-to-video editing

Diffusion models have enabled a plethora of methods for editing images and videos from text prompts. Below we apply one of these methods, SDEdit,<sup>32</sup> to Sora. This technique enables Sora to transform the styles and environments of input videos zero-shot.

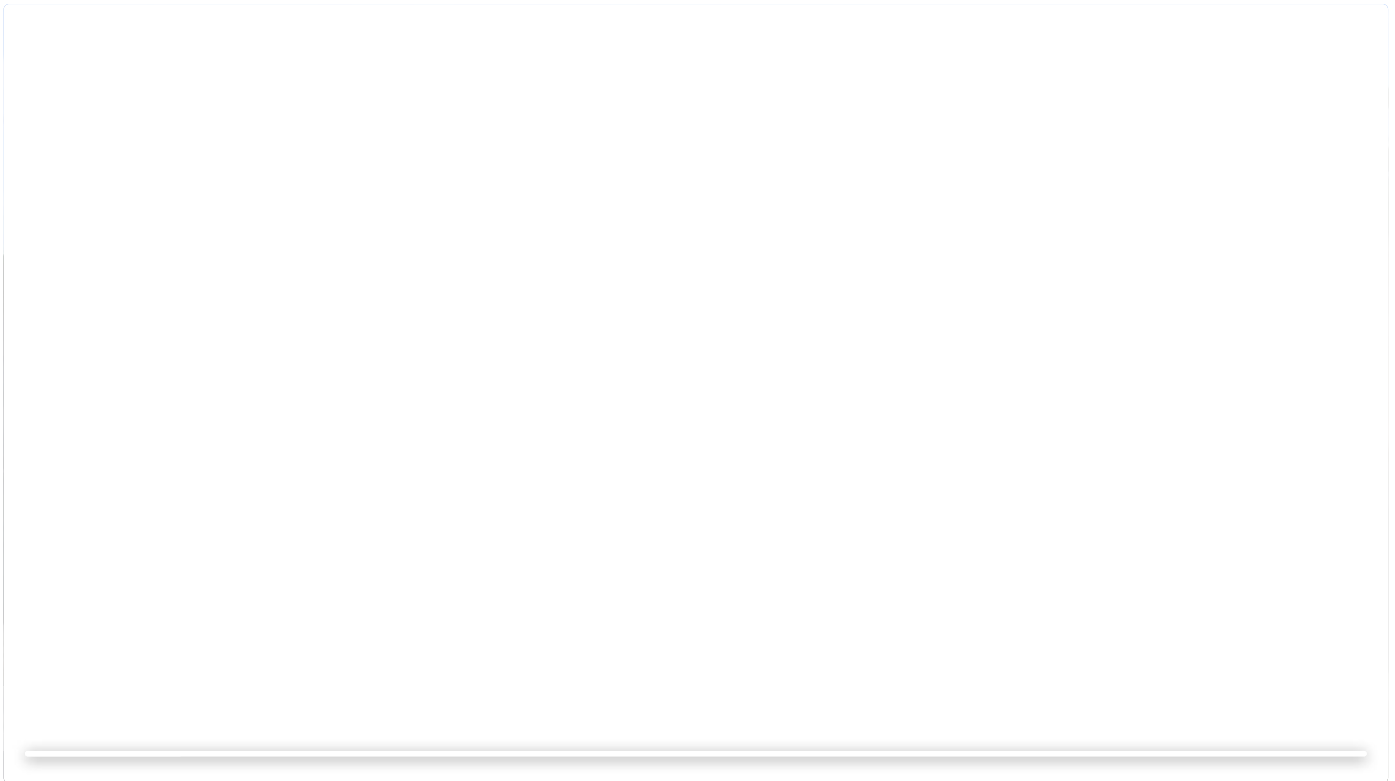
Input video

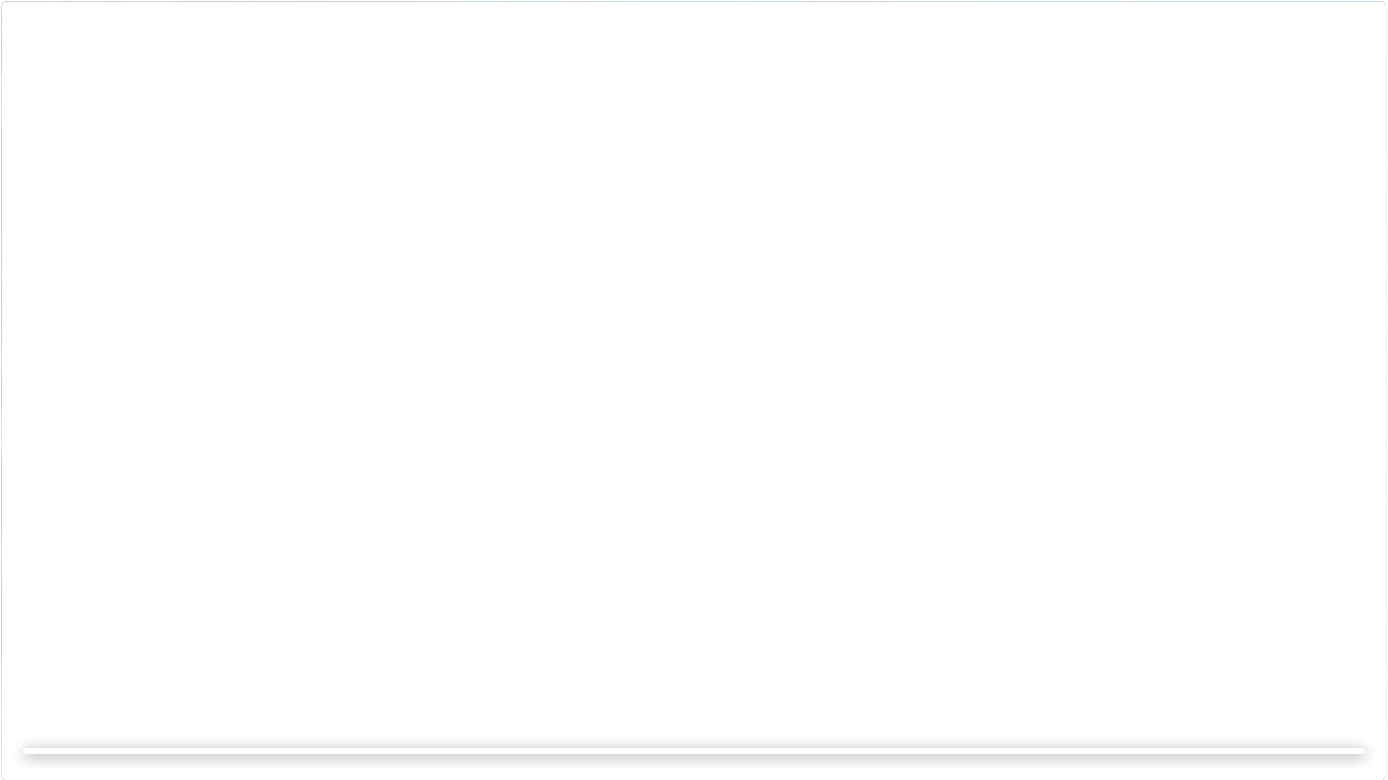
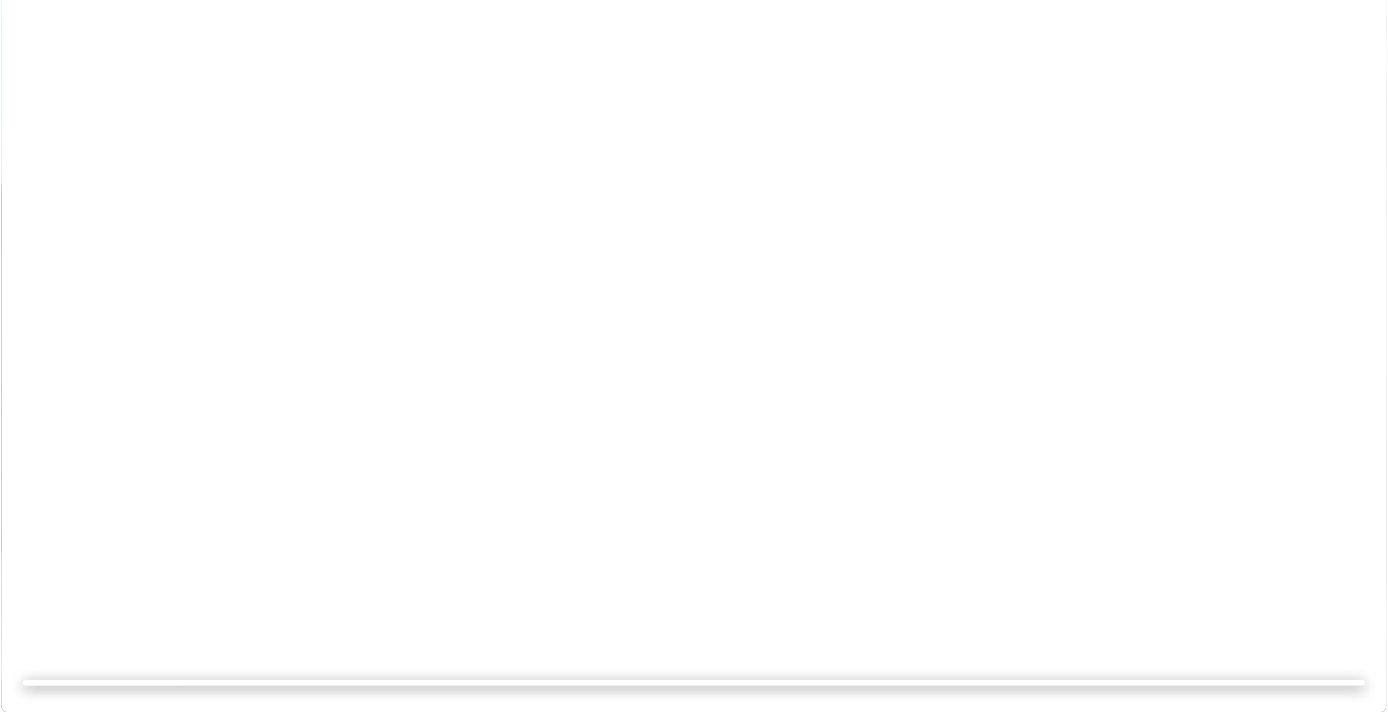
[change the setting to be in a lush jungle](#)

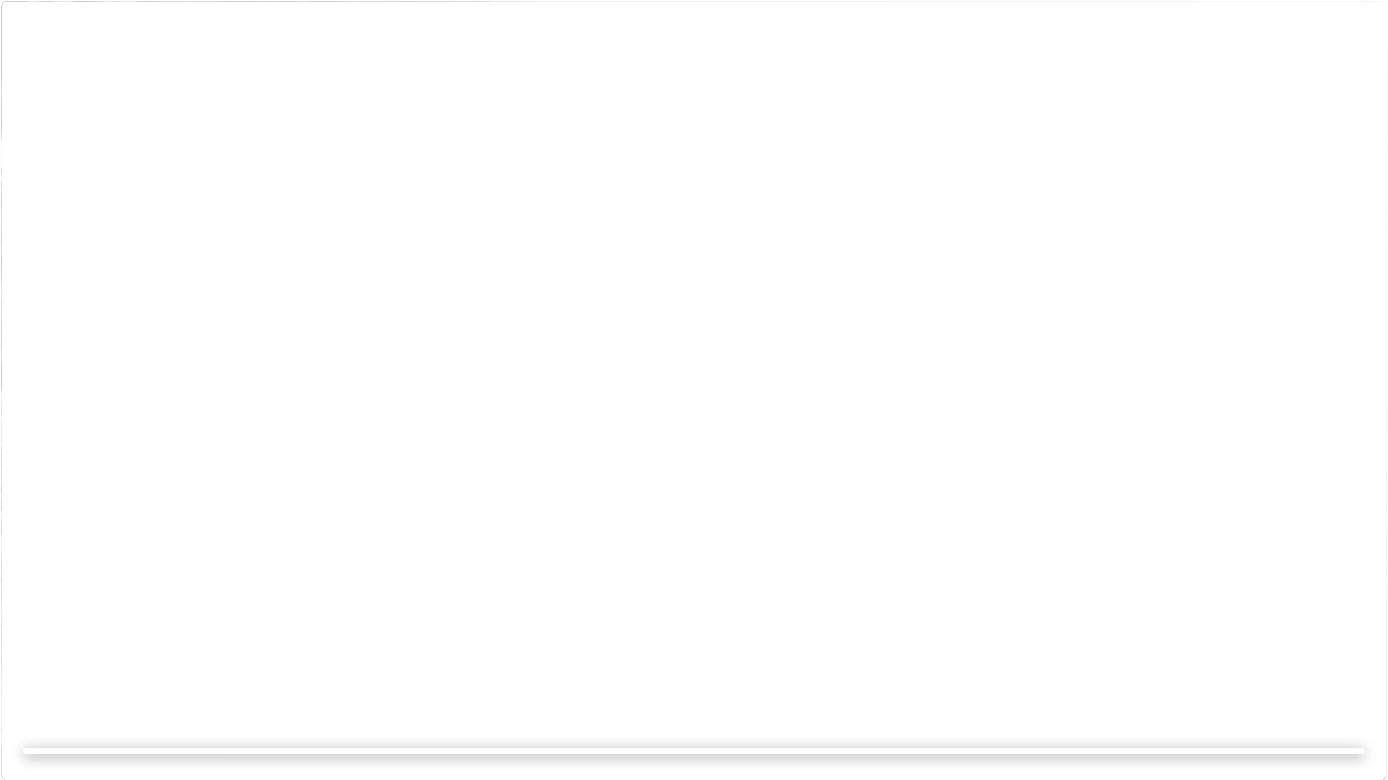
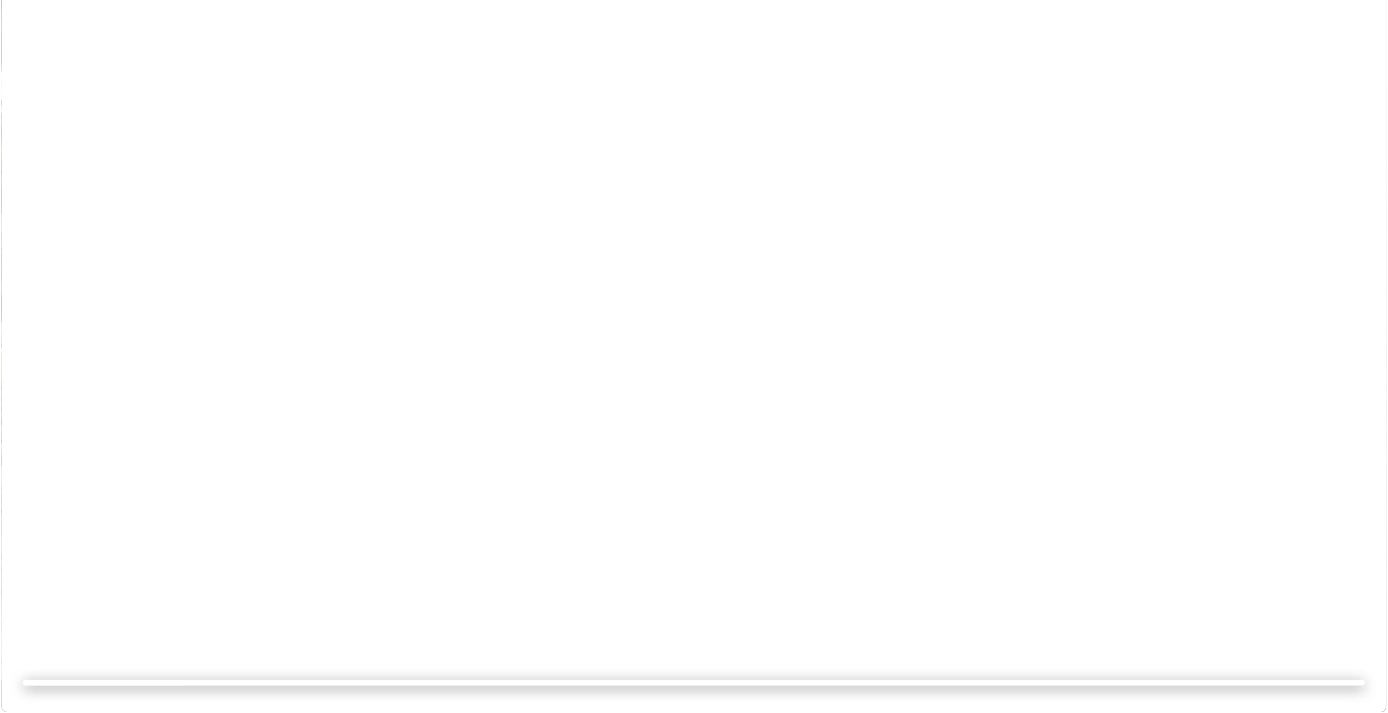


## Connecting videos

We can also use Sora to gradually interpolate between two input videos, creating seamless transitions between videos with entirely different subjects and scene compositions. In the examples below, the videos in the center interpolate between the corresponding videos on the left and right.







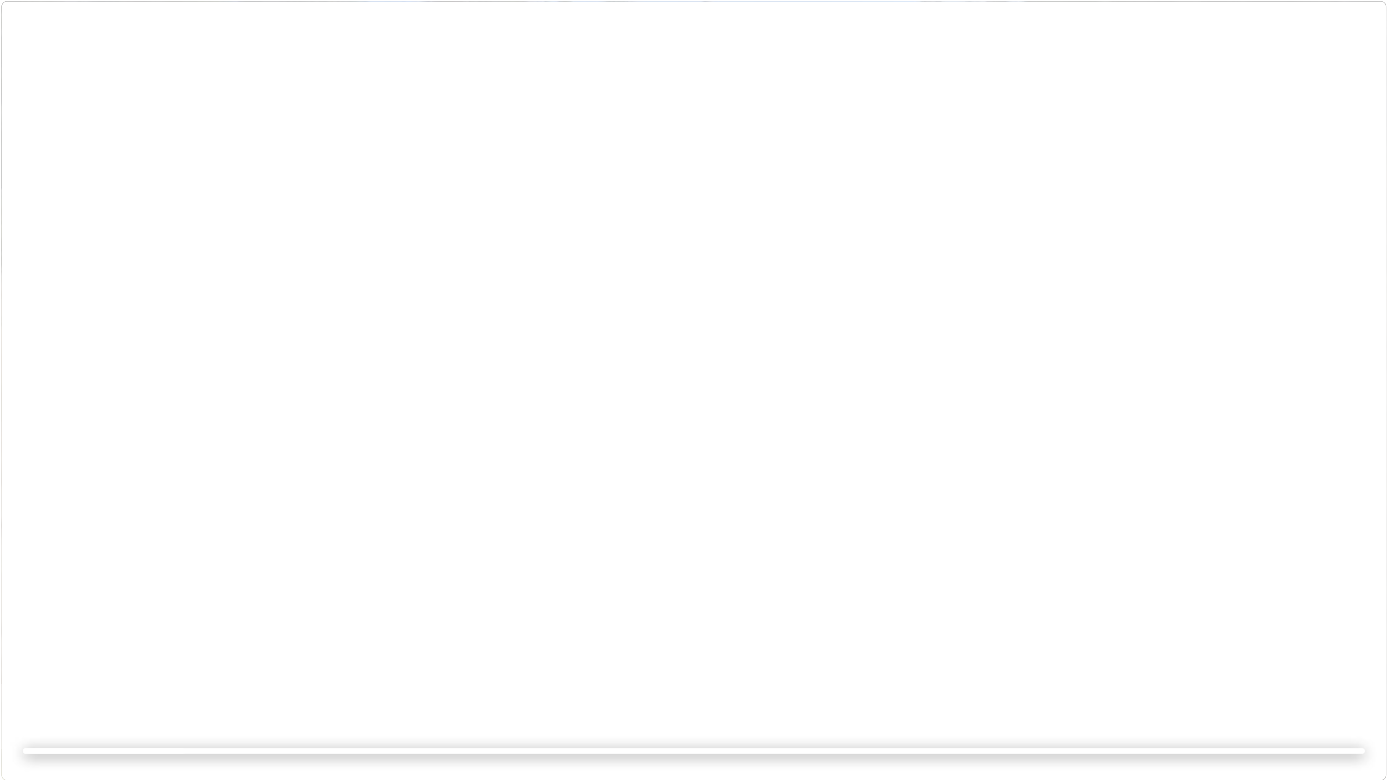
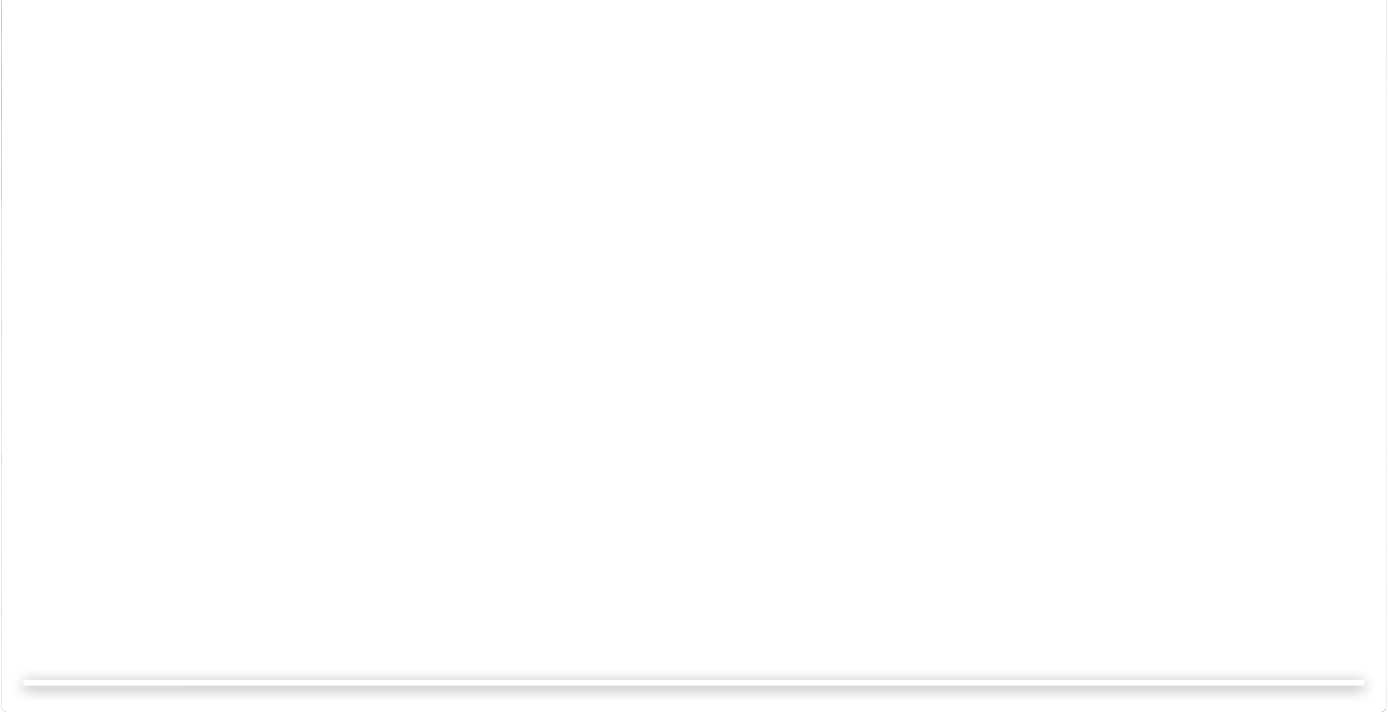


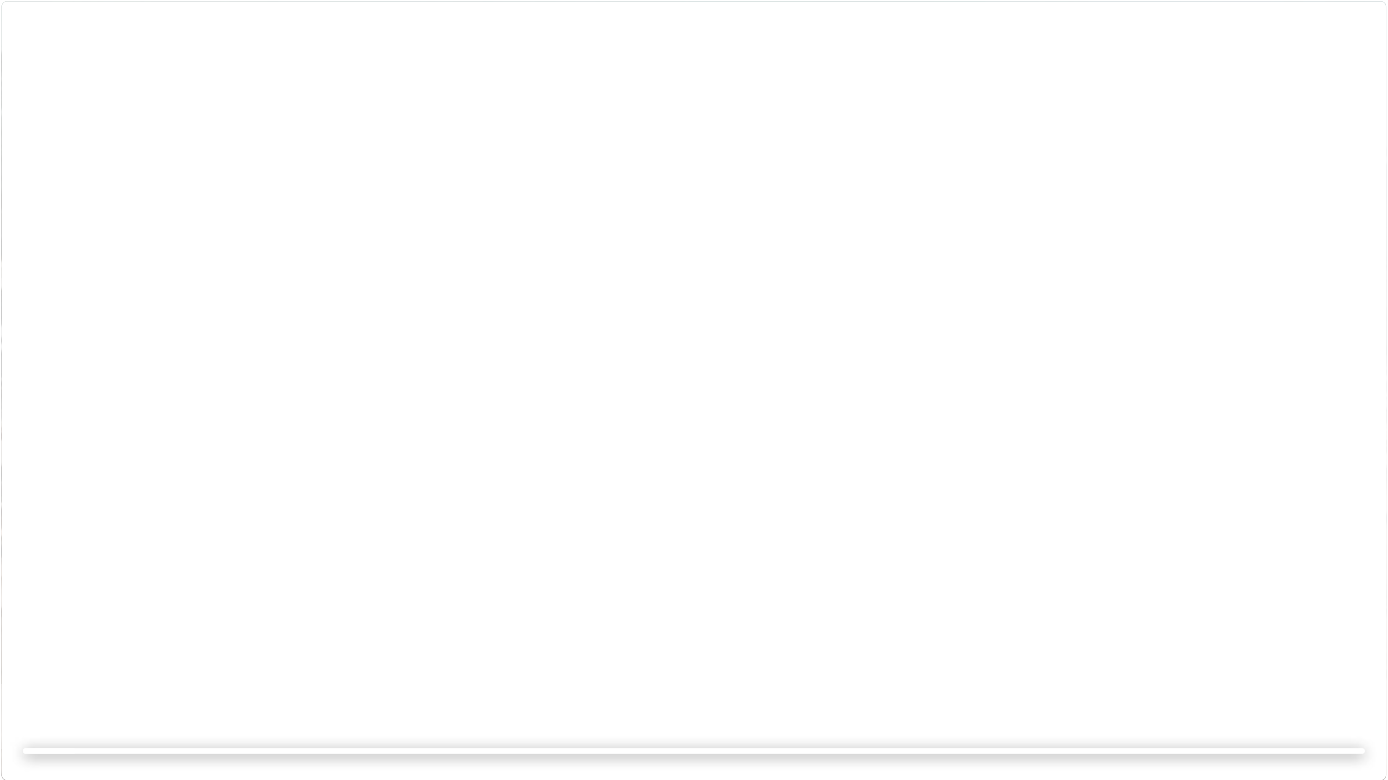
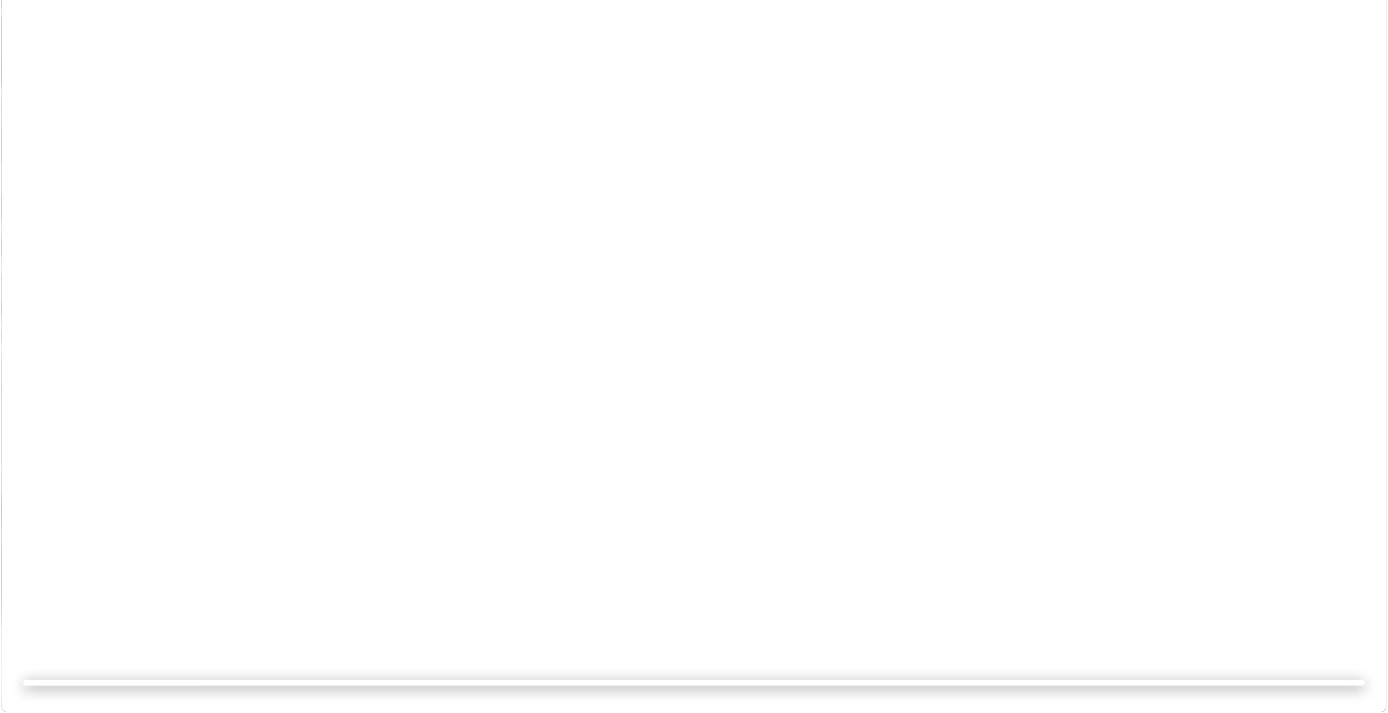


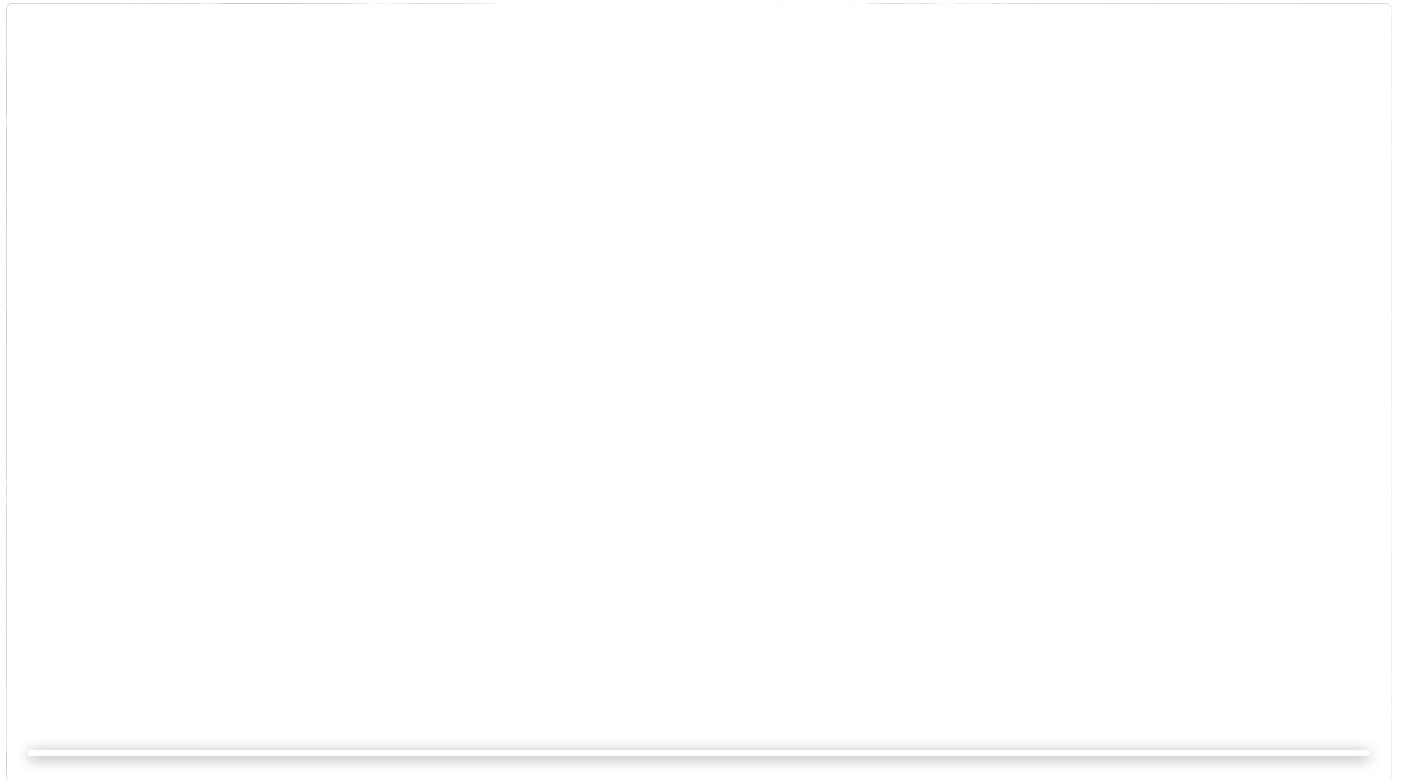
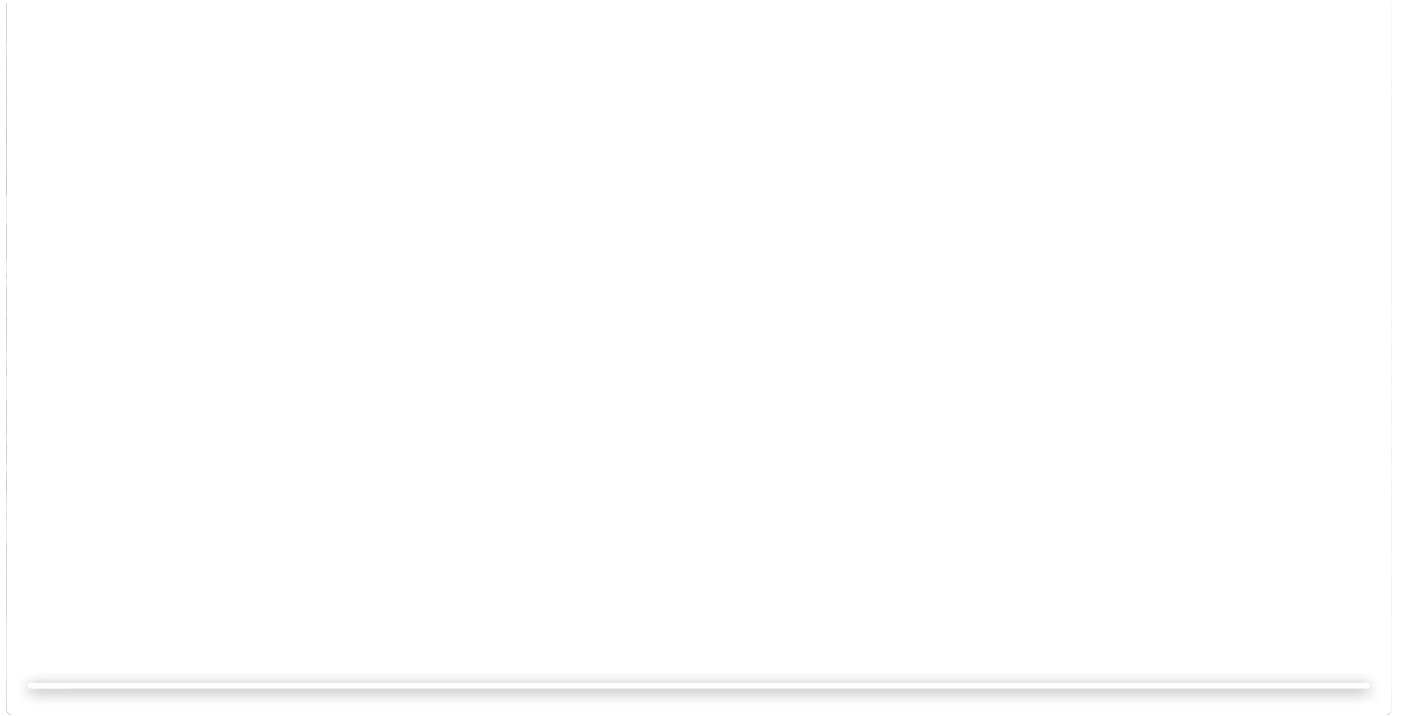












## Image generation capabilities

Sora is also capable of generating images. We do this by arranging patches of Gaussian noise in a spatial grid with a temporal extent of one frame. The model can generate images of variable sizes—up to 2048×2048 resolution.



Close-up portrait shot of a woman in autumn, extreme detail, shallow depth of field





Vibrant coral reef teeming with colorful fish and sea creatures





Digital art of a young tiger under an apple tree in a matte painting style with gorgeous details



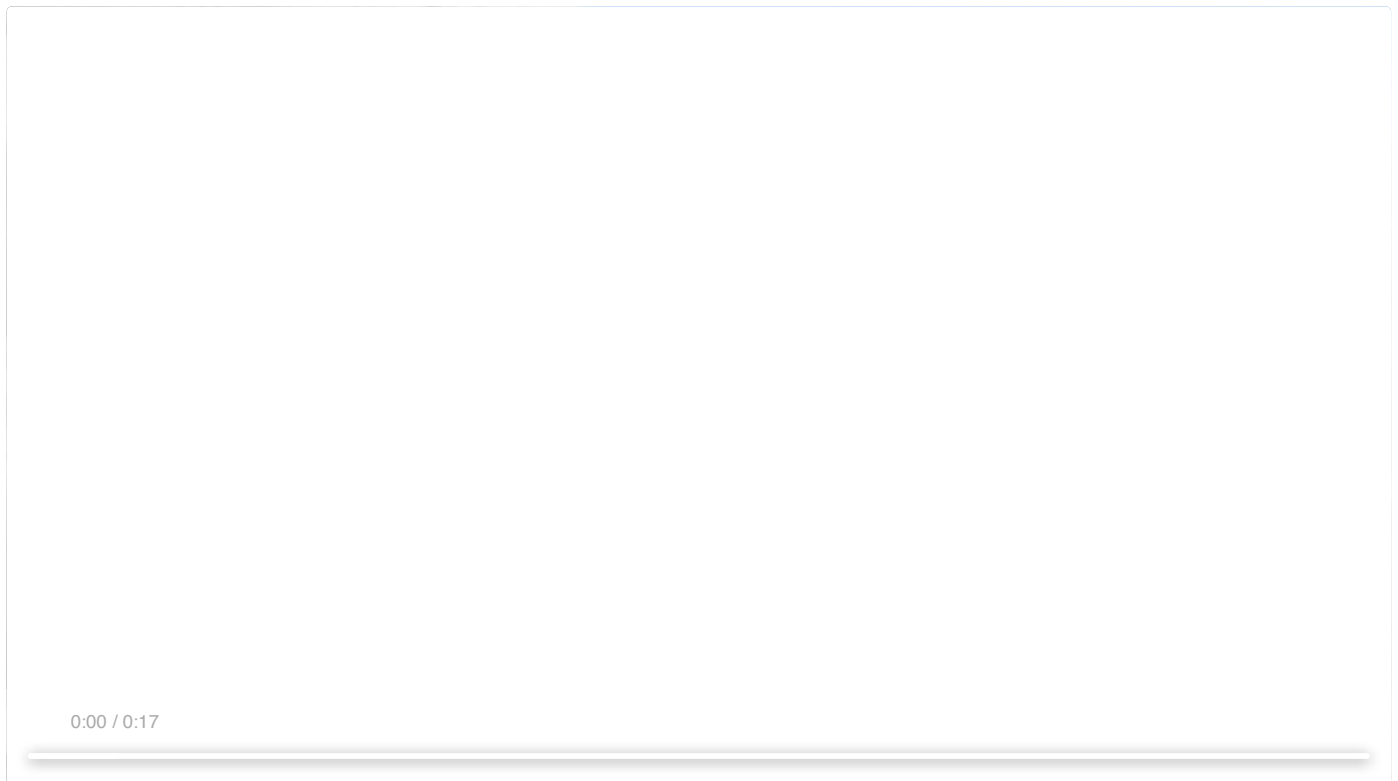
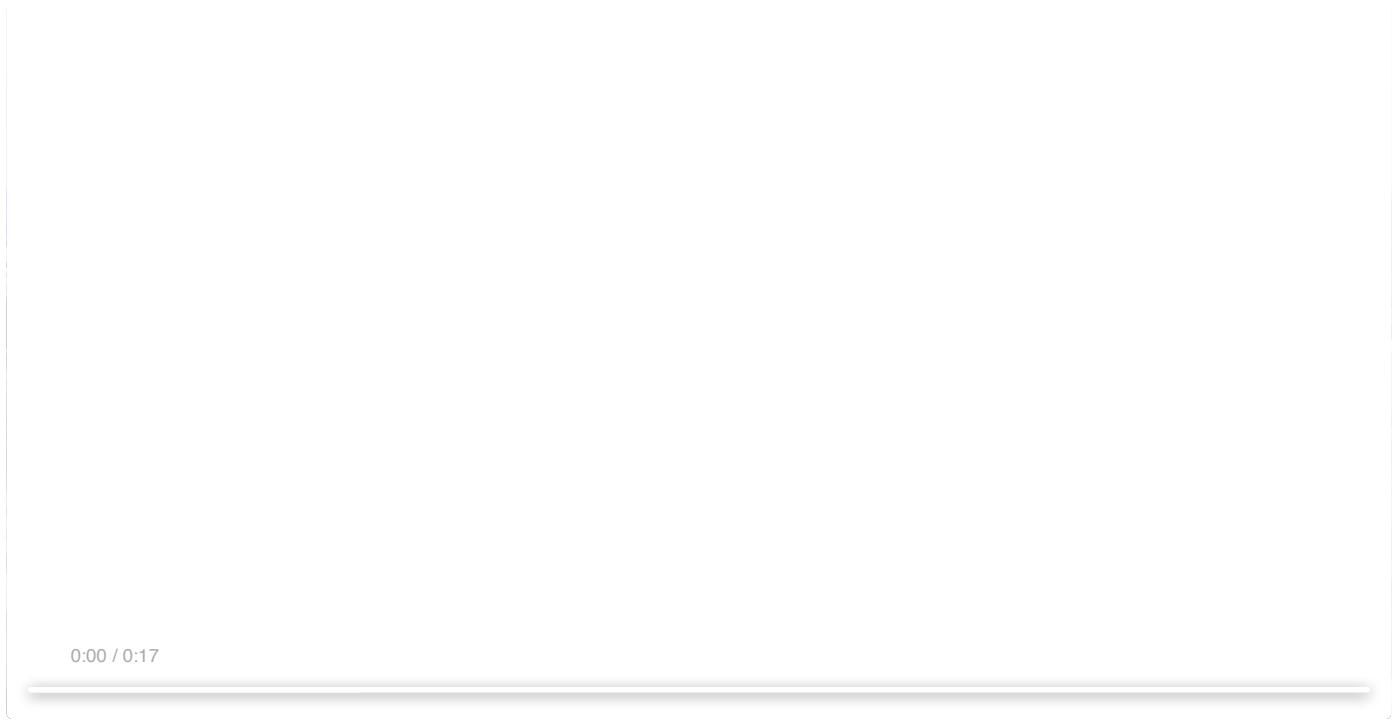
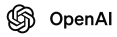


A snowy mountain village with cozy cabins and a northern lights display, high detail and photorealistic dslr, 50mm f/1.2

## Emerging simulation capabilities

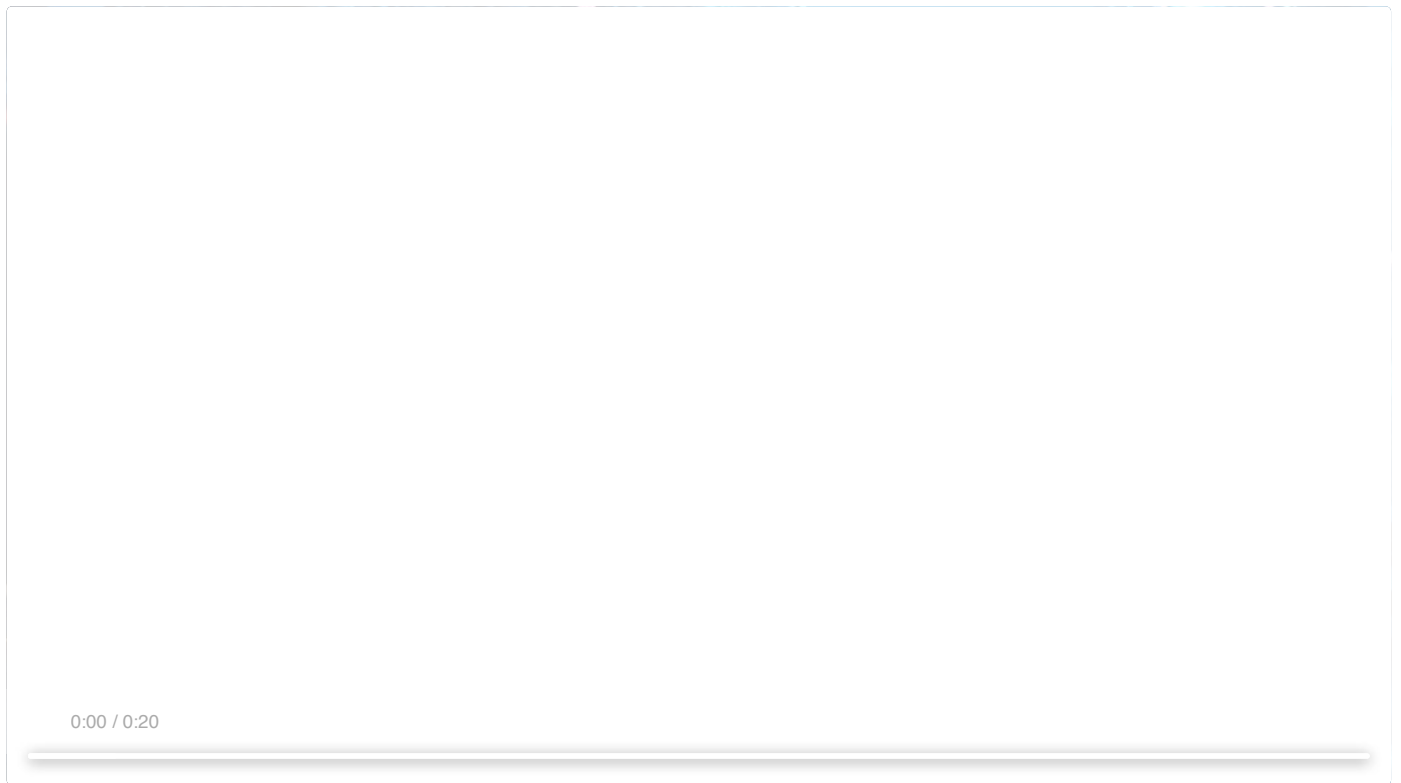
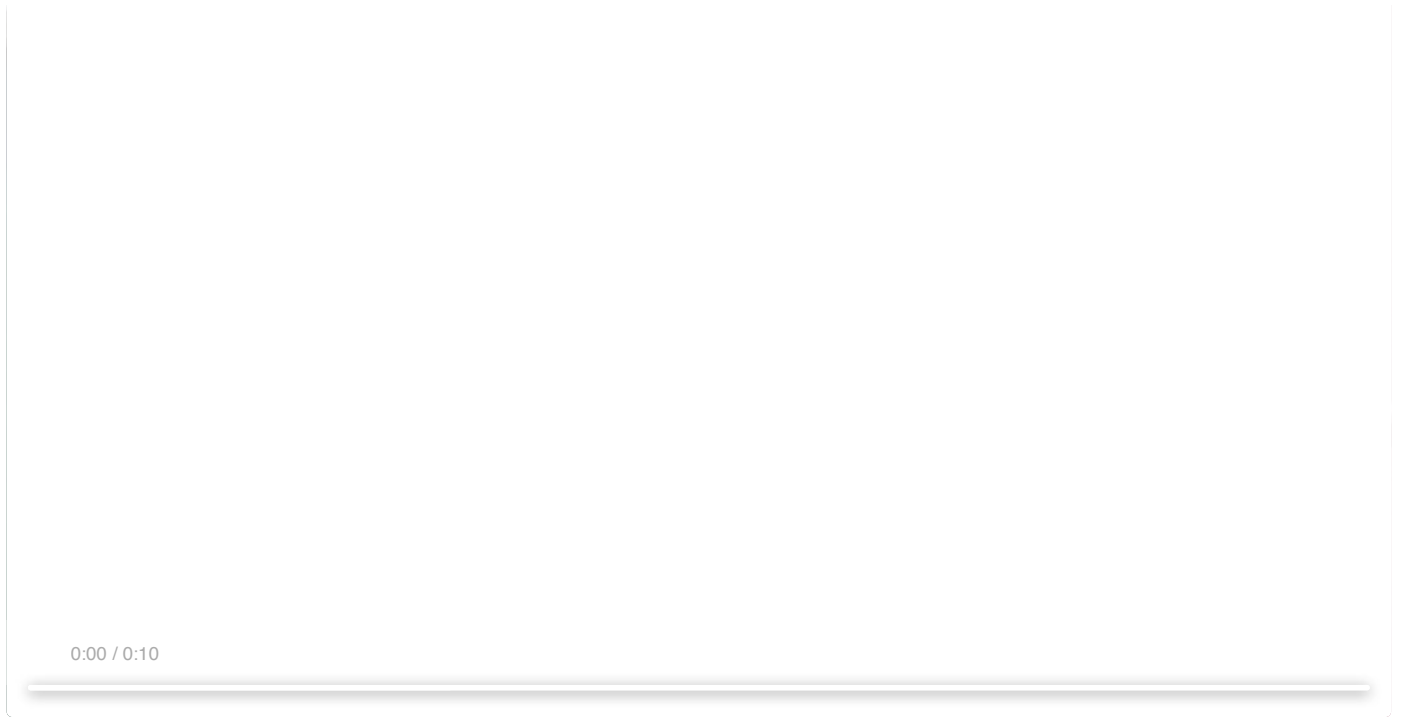
We find that video models exhibit a number of interesting emergent capabilities when trained at scale. These capabilities enable Sora to simulate some aspects of people, animals and environments from the physical world. These properties emerge without any explicit inductive biases for 3D, objects, etc.—they are purely phenomena of scale.

**3D consistency.** Sora can generate videos with dynamic camera motion. As the camera shifts and rotates, people and scene elements move consistently through three-dimensional space.

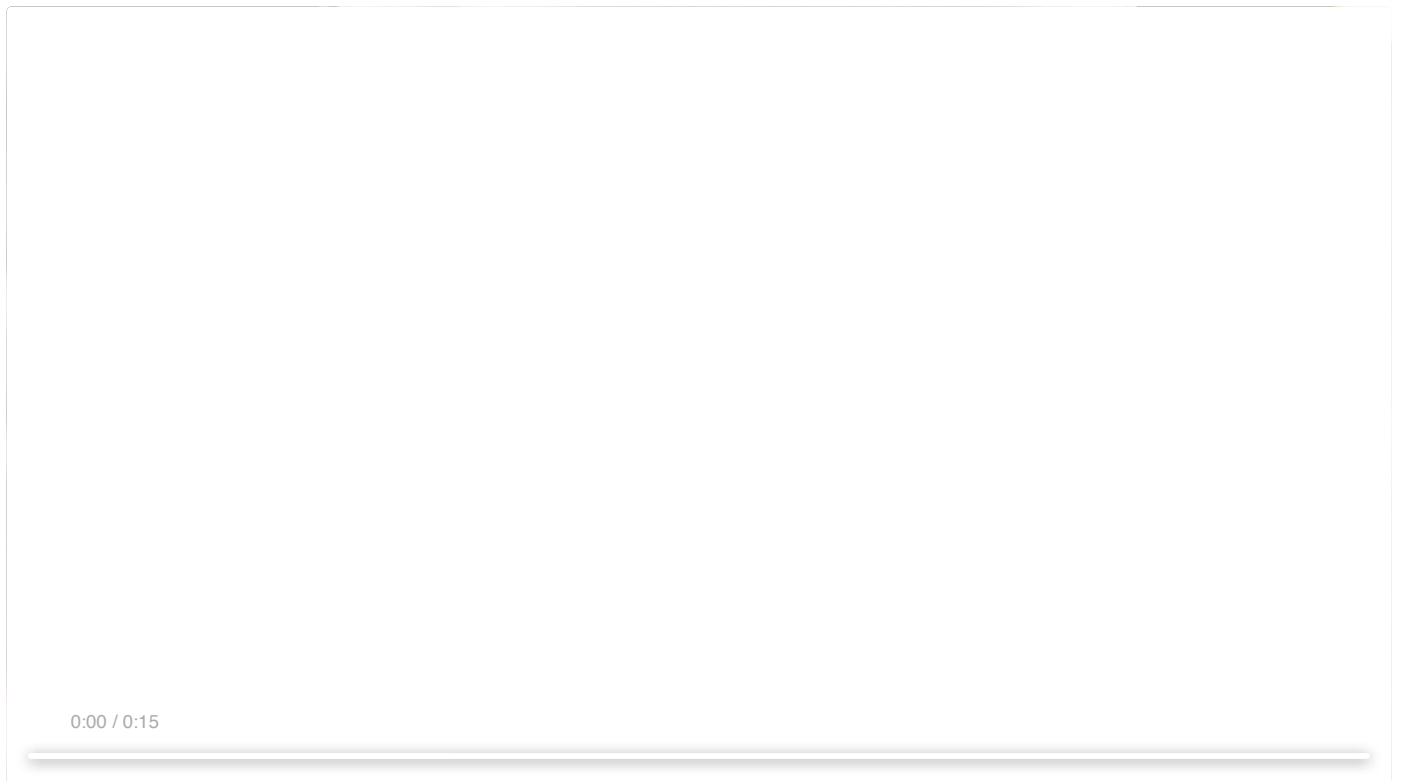
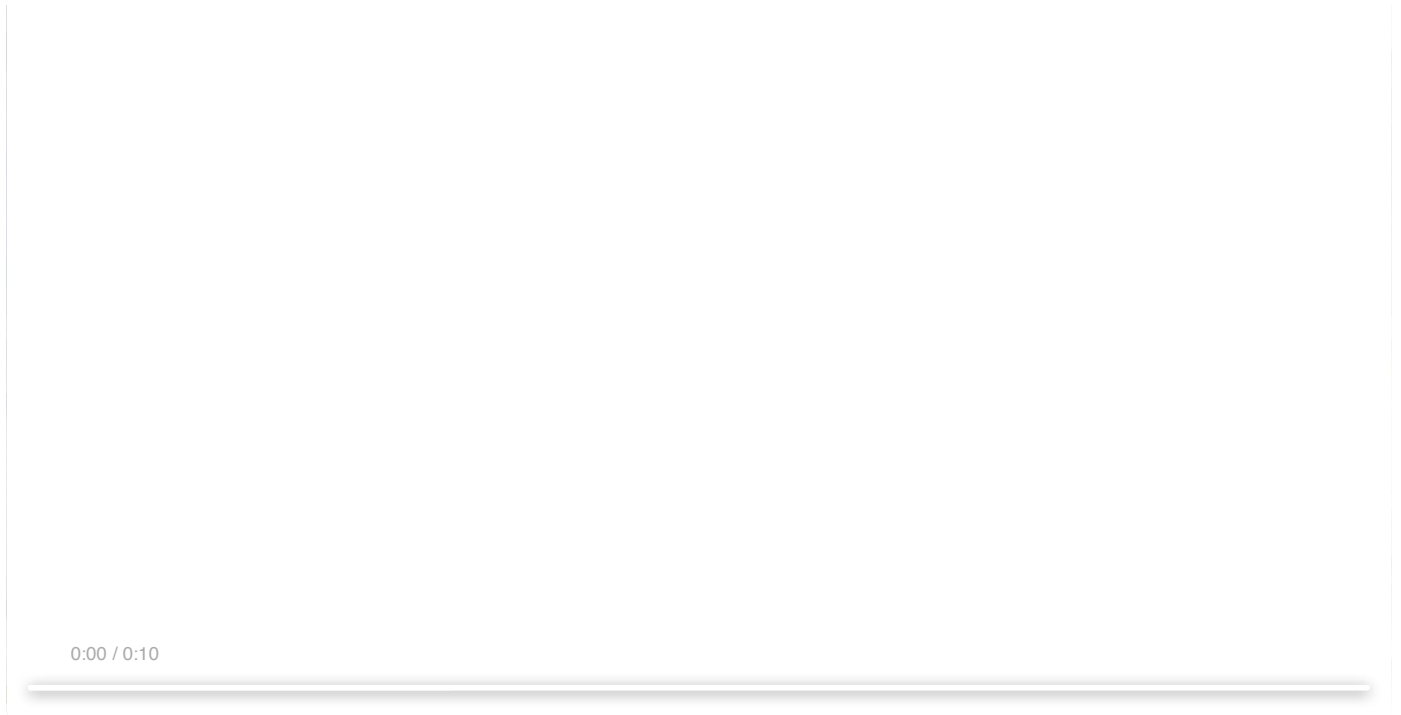


**Long-range coherence and object permanence.** A significant challenge for video generation systems has been maintaining temporal consistency when sampling long videos. We find that Sora is often, though not always, able to effectively model both short- and long-range dependencies. For example, our model can persist people, animals and objects even when they are occluded or leave the frame. Likewise, it can generate multiple shots of the same character in a single sample, maintaining their appearance throughout the video.

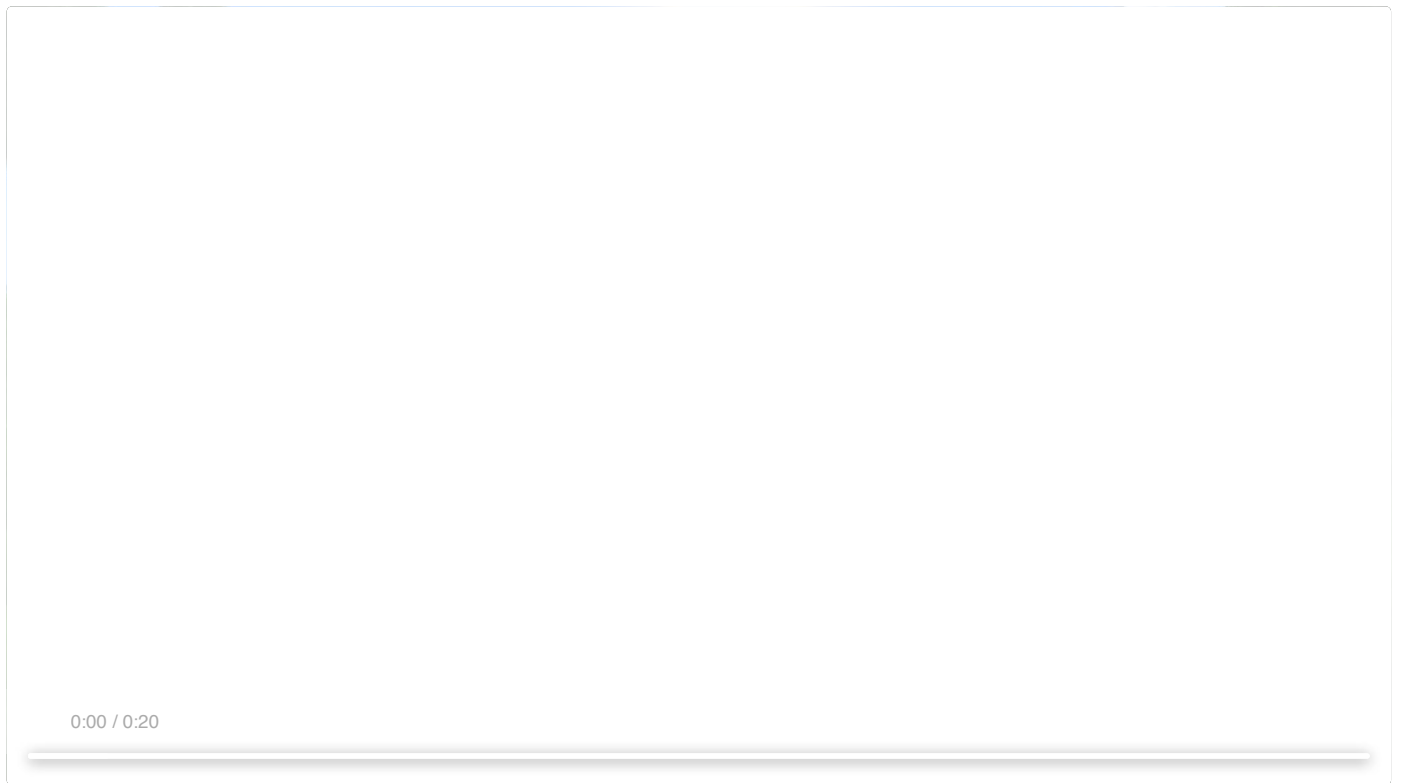
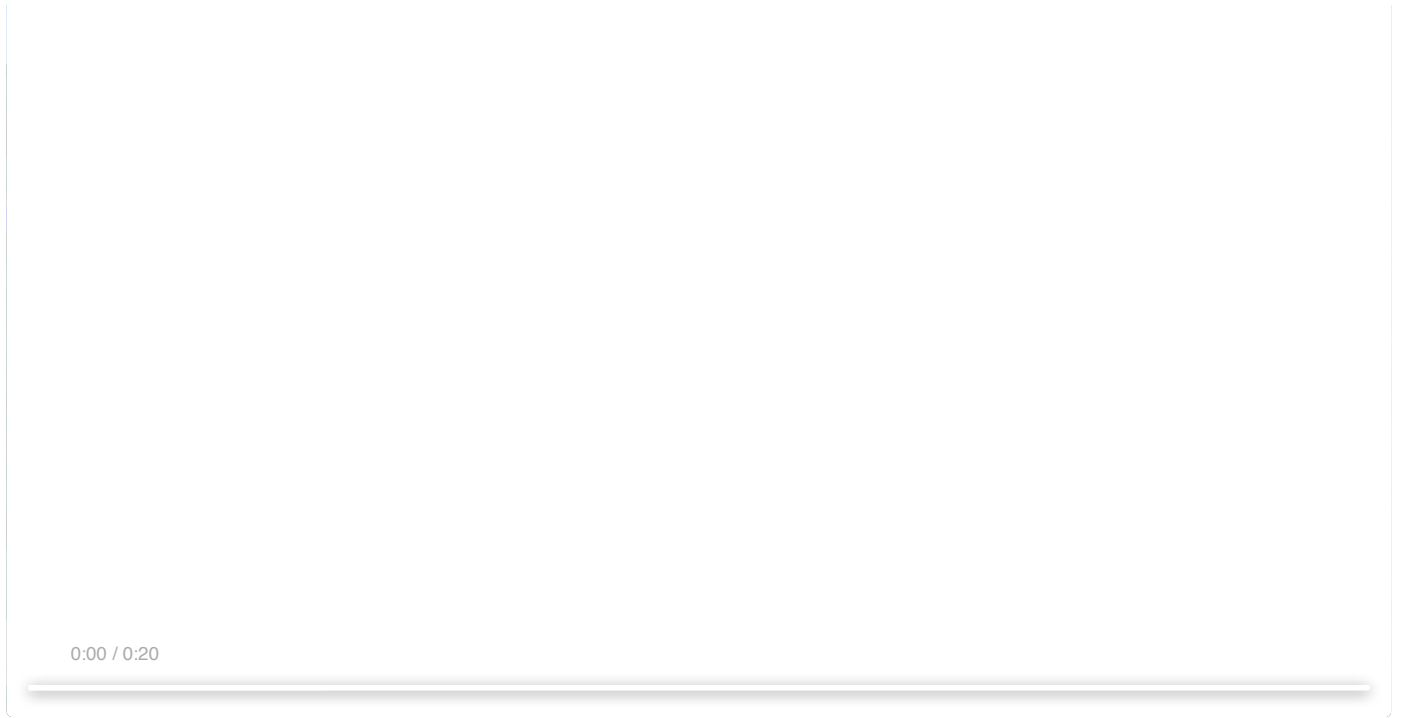




**Interacting with the world.** Sora can sometimes simulate actions that affect the state of the world in simple ways. For example, a painter can leave new strokes along a canvas that persist over time, or a man can eat a burger and leave bite marks.

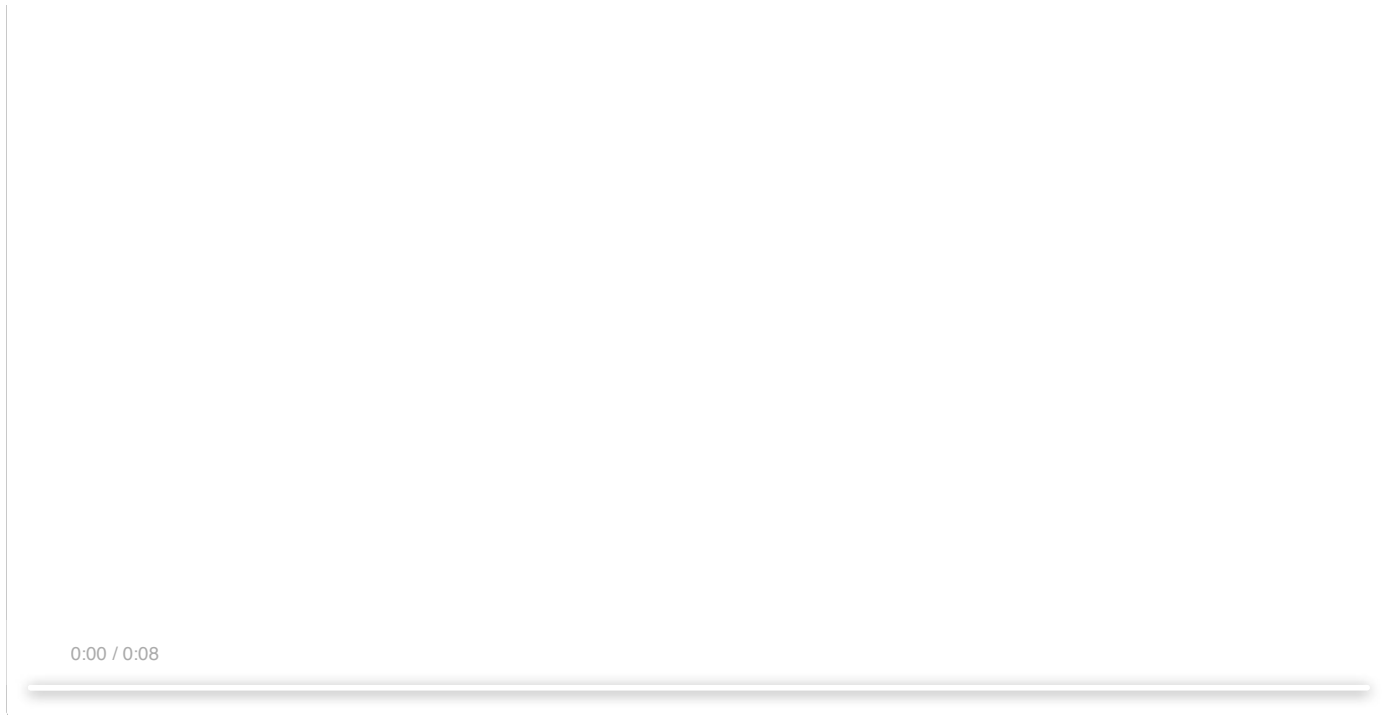


**Simulating digital worlds.** Sora is also able to simulate artificial processes—one example is video games. Sora can simultaneously control the player in Minecraft with a basic policy while also rendering the world and its dynamics in high fidelity. These capabilities can be elicited zero-shot by prompting Sora with captions mentioning “Minecraft.”

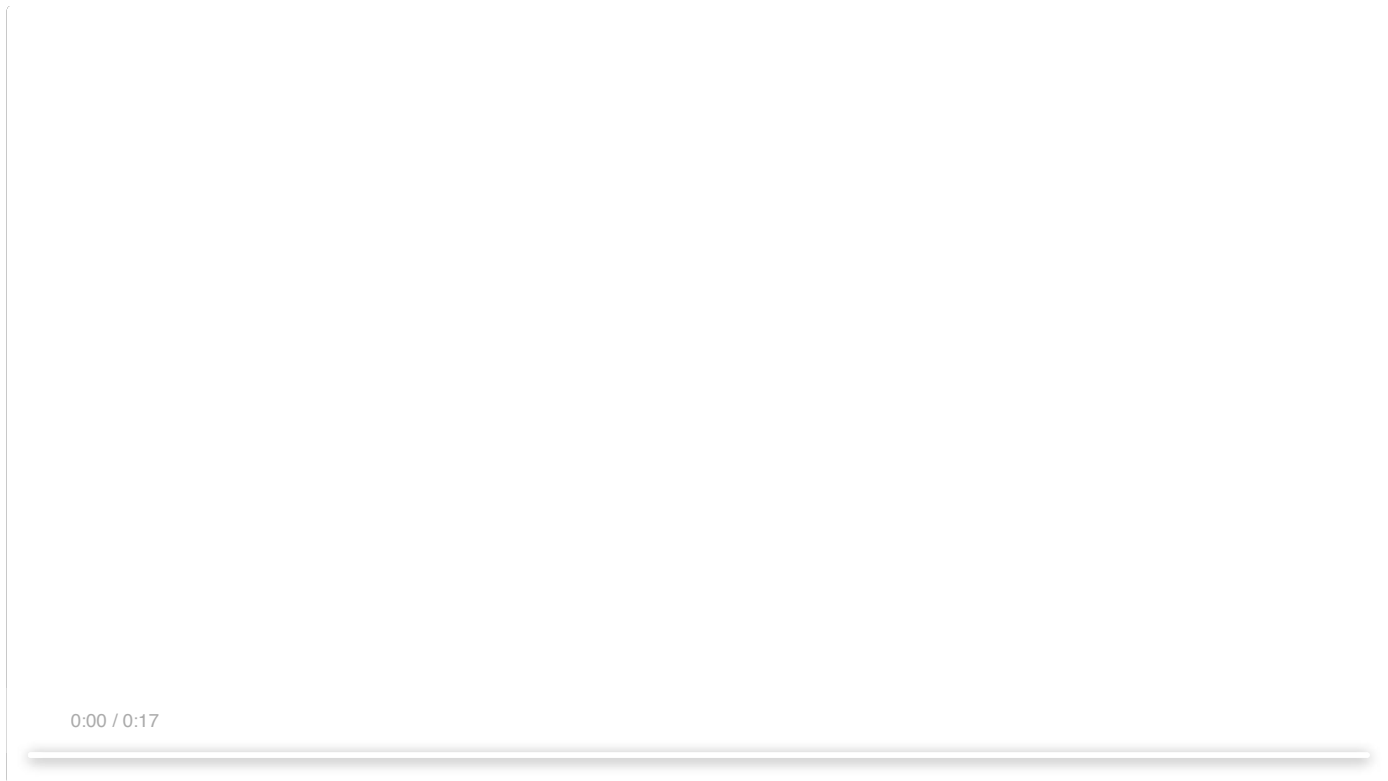


These capabilities suggest that continued scaling of video models is a promising path towards the development of highly-capable simulators of the physical and digital world, and the objects, animals and people that live within them.

## Discussion



Sora currently exhibits numerous limitations as a simulator. For example, it does not accurately model the physics of many basic interactions, like glass shattering. Other interactions, like eating food, do not always yield correct changes in object state. We enumerate other common failure modes of the model—such as incoherencies that develop in long duration samples or spontaneous appearances of objects—in our [landing page](#).



We believe the capabilities Sora has today demonstrate that continued scaling of video models is a promising path towards the development of capable simulators of the physical and digital world, and the objects, animals and people that live within them.