

[W&B Fully Connected](#) > [ML News](#)

# Some Details on OpenAI's Sora and Diffusion Transformers

Combining some of the most promising architectures, OpenAI shows off its newest model!

[Brett Young](#)

Last Updated: Feb 16, 2024

OpenAI has taken a notable step forward in the realm of video generation models with the introduction of Sora, a model designed to simulate aspects of the physical world through video. This model diverges from traditional approaches by its ability to generate high-fidelity videos of up to a minute in duration, across various durations, resolutions, and aspect ratios. The foundation of Sora lies in its training on a vast dataset of videos and images, leveraging a "diffusion transformer" architecture that operates on spacetime patches of video.

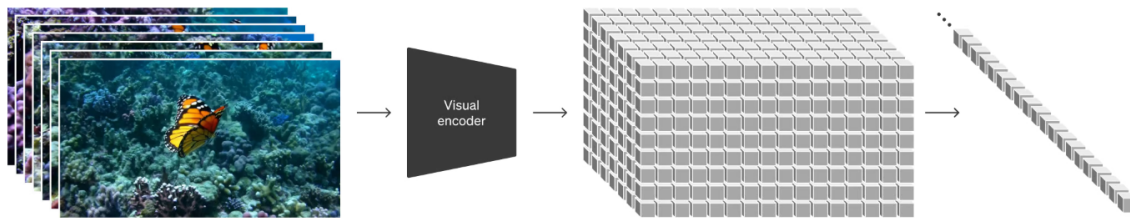
## ▸ Diffusion Transformers

**So you may be wondering, what is a diffusion transformer? - Note, this description has not been officially confirmed by OpenAI but is likely similar at a high level.**

The process of generating high-quality videos using Diffusion Transformers (DiTs) begins with training a Variational Autoencoder (VAE) on the target dataset to create a compact latent representation of the data. This latent representation is then subjected to a forward diffusion process, where noise is incrementally added over several steps, distorting the original data. These noised latents are subsequently split into patches and linearly embedded into vectors, which are

By clicking "Accept All Cookies", you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

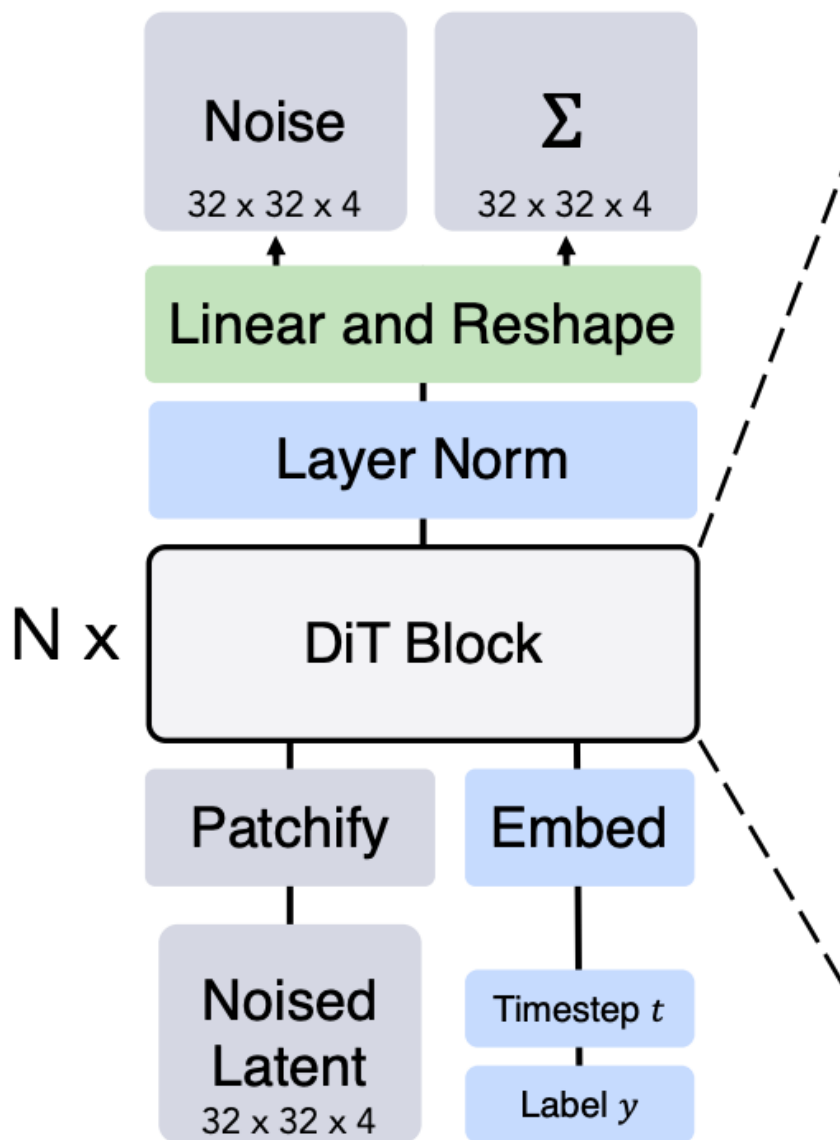
[Cookies Settings](#)[Reject All](#)[Accept All](#)



The core of the DiTs model involves processing these tokens through a series of transformer blocks that can incorporate additional conditioning information, such as class labels or text descriptions, to guide the generation process toward specific outcomes. DiTs specifically aim to predict the noise that was added during the forward diffusion process. This prediction of noise enables the model to effectively reverse the diffusion process by iteratively denoising the latent representations. If you are interested in more of the details, I recommend checking the references mentioned in OpenAI's announcement.

By clicking "Accept All Cookies", you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.





## Latent Diffusion Transformer

In the final stages, the reverse diffusion process uses the predicted noise to gradually reconstruct the clean latent representation of the data. These denoised latents are then passed through the decoder part of the VAE to generate the final image output. This intricate process, from initial data encoding to the reverse diffusion and final decoding, exemplifies the model's ability to generate coherent and detailed images by learning to reverse the noise addition process, showcasing the

By clicking "Accept All Cookies", you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.





## ▸ Re-captioning

Sora also incorporates advanced techniques for language understanding, applying re-captioning methods to improve the fidelity of text-to-video generation. This capability is further enhanced by leveraging GPT models to expand short user prompts into detailed captions, guiding the video generation process more effectively. From what it seems (based on limited information), this technique seems similar to the visual instruction tuning method used by LLaVA, but expanded to videos.

## ▸ Video and Image Editing

Beyond generating standalone video samples, Sora demonstrates proficiency in a range of video and image editing tasks. It can animate static images, create perfectly looping videos, and extend videos in time. Additionally, Sora shows emerging simulation capabilities, such as 3D consistency, long-range coherence, and object permanence, indicating its potential as a tool for simulating complex scenarios from the physical world.

## ▸ Limitations

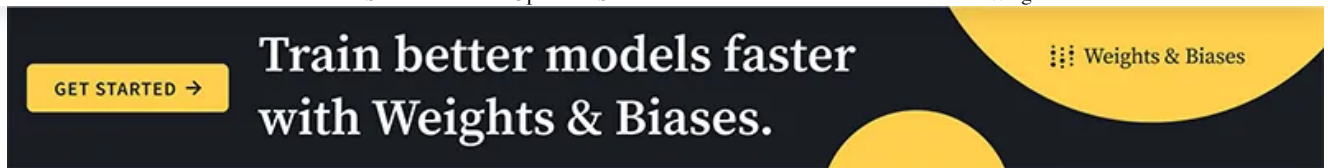
Despite its advances, Sora is not without limitations. The model sometimes struggles with accurately modeling physics for certain interactions and maintaining consistency in long-duration samples. However, the research team remains optimistic about the potential of scaling video models like Sora for developing capable simulators of the physical and digital worlds.

The Announcement:

<https://openai.com/research/video-generation-models-as-world-simulators#fn-15>

By clicking "Accept All Cookies", you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.





Created with ❤️ on Weights & Biases.

<https://wandb.ai/byyoung3/ml-news/reports/Some-Details-on-OpenAI-s-Sora-and-Diffusion-Transformers--Vmldzo2ODQwNTM3>


---

Made with Weights & Biases. [Sign up](#) or [log in](#) to create reports like this one.

Never lose track of another ML  
project. **Try W&B today.**

[SIGN UP](#)

[TRY W&B NOW](#)

 **Weights & Biases**  
Get weekly updates with the latest ML news.

[Subscribe](#)

---

#### PRODUCTS

[Dashboard](#) [Sweeps](#) [Artifacts](#) [Reports](#) [Tables](#)

#### QUICKSTART

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.



Copyright ©2024 Weights & Biases. All rights reserved.

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

