OPEN AI　SORA 技术报告原文+译文+ 报告总结

文档整理：AI 设计研究室
https://space.bilibili.com/2161614

要点总结

模型路径：

1. 架构为扩散模型（diffusion model）+transformer
2. 训练时先用预训练模型把大量的大小不一的视频源文件编码转化为统一的 patch 表示，把时空要素提取作为 transformer 的 token 进行训练。
3. 模型效果好和超大量的数据集和更多的运算时间息息相关

优势：

1. 人物和背景的连贯性，即时人物运动出了相机范围再回来时还保持同样特征
2. 自然语言的理解程度很高
3. 可以在同一个种子下生成不同尺寸（横向竖向）的视频适配不同设备
4. 可以生成长达 1min 高清视频
5. 可以以文字，图片，视频作为控制要素控制输出结果

不足：

1. 对于物理规则了解较弱，比如吹气后蜡烛不会熄灭，左右不分，玻璃掉落不会碎
2. 对于算力要求较高（猜测）

可以实现：

1. 文生视频，图生视频，图+文生视频，视频修改
2. 视频转绘，视频延伸，视频补全

未来畅想：

1. 重新洗牌 AI 生成视频产业
2. 扩散模型的上限比想象中的高！
3. 全局一致性可以被解决
4. 文字生成 3D 或将迎来突破
5. AR,VR，VIsionPro 新型应用潜力

大神观点：

骆思勉 清华叉院
看完Technical Report的一些想法：

1. Diffusion生成框架的天花板远比我们之前想象的要更高(很可能已经够了)，make diffusion great again! 给Diffusion研究者注入一剂强心剂。从数学理论上来说，Diffusion也是能够几乎拟合任意数据分布的(包括真实世界的连贯性视频)。

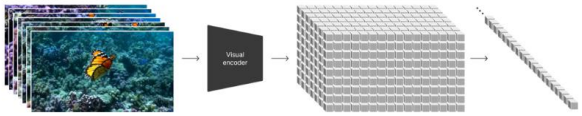
2. Scale is all you need. Scale上去后，在视频生成上能够产生类似在LLM里的涌现现象。包括视频连贯性，3D consistency, Long-range coherence。

3. Physics Prior什么的可能都不需要额外引入。Scale + Data足以。

报告原文 https://openai.com/research/video-generation-models-as-world-simul ators

英文原文	中文翻译
Video generation models as world simulators	**视频生成模型作为世界模拟器**
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.	我们探索了在视频数据上进行大规模生成模型的训练。具体而言，我们联合在可变持续时间、分辨率和宽高比的视频和图像上训练了文本条件扩散模型。我们利用了一个在视频和图像潜在编码的时空块上操作的 transformer 架构。我们最大的模型，Sora，能够生成一分钟高保真度的视频。我们的结果表明，扩展视频生成模型是建立通用物理世界模拟器的一条有前景的道路。
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,1,2,3 generative adversarial networks,4,5,6,7 autoregressive transformers,8,9 and diffusion models.10,11,12 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.	本技术报告关注以下两个方面：(1) 我们将各种类型的视觉数据转换为统一表示的方法，以实现大规模生成模型的训练，以及 (2) 对 Sora 的能力和局限性进行定性评估。模型和实现细节未包含在本报告中。 之前的研究已经探讨了使用各种方法对视频数据进行生成建模，包括循环网络、生成对抗网络、自回归变压器和扩散模型。这些工作通常侧重于某一类视觉数据、较短的视频或固定大小的视频。Sora 是一种视觉数据的通用模型——它可以生成跨越各种持续时间、宽高比和分辨率的视频和图像，高清视频最长可达一分钟。
Turning visual data into patches	**将视觉数据转换成 patch**
We take inspiration from large language models which acquire generalist capabilities by training on internet-scale data.13,14 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	我们受到大型语言模型的启发，这些模型通过在互联网规模的数据上进行训练而获得了通用能力。LLM 范式的成功部分得益于优雅地统一了文本的多种模态——代码、数学和各种自然语言的标记。在这项工作中，我们考虑了生成视觉数据模型如何继承这些好处。而 LLMs 具有文本标记，Sora 具有视觉 patch。patch 已被证明是视觉数据模型的有效表示。

shown to be an effective representation for models of visual data.15,16,17,18 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,19 and subsequently decomposing the representation into spacetime patches.

Video compression network

We train a network that reduces the dimensionality of visual data.20 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.

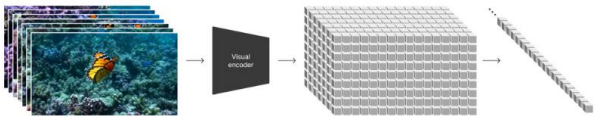
Spacetime Latent Patches

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 model21,22,23,24,25; 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.26 Transformers have demonstrated remarkable scaling properties across a variety of domains, including language modeling,13,14 computer vision,15,16,17,18 and image generation.27,28,29

我们发现，patch 是一种高度可扩展且有效的表示方法，适用于训练不同类型的视频和图像的生成模型。



在高层次上，我们通过首先将视频压缩成低维度潜在空间，然后将表示分解为时空补丁来将视频转换成补丁。

视频压缩网络

我们训练了一个网络来降低视觉数据的维度。这个网络以原始视频作为输入，并输出一个在时间和空间上都被压缩的潜在表示。Sora 在这个压缩的潜在空间上进行训练，并随后生成视频。我们还训练了一个对应的解码器模型，将生成的潜在空间映射回像素空间。

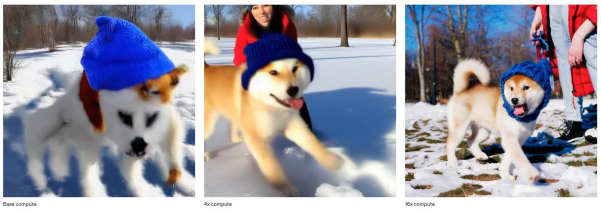
时空潜在补丁

给定一个压缩的输入视频，我们提取一系列的时空补丁，这些补丁充当 transformer 的 tocken。我们基于补丁的表示使得 Sora 能够在不同分辨率、持续时间和宽高比的视频和图像上进行训练。在推理时，我们可以通过将随机初始化的补丁适当地排列在一个大小合适的网格中来控制生成视频的大小。

将 transformer 扩展到视频生成

Sora 是一个扩散模型；给定输入的初始噪声（以及文本提示等条件信息），它被训练为预测原始的"干净"补丁。重要的是，Sora 是一个扩散 transformer。transformer 在多个领域展示了显著的扩展性能，包括语言建模、计算机视觉和图像生成。

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.



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 256x256 resolution. We find that instead training on data at its native size provides several benefits.

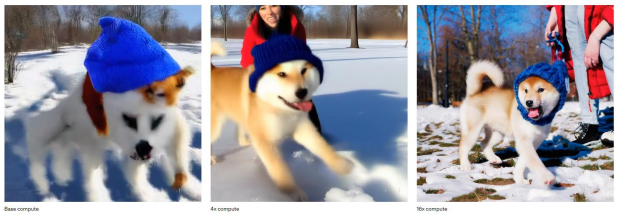
Sampling flexibility

Sora can sample widescreen 1920x1080p videos, vertical 1080x1920 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)s have improved framing.

在这项工作中，我们发现扩散变压器在视频模型中也能有效地扩展。在下面，我们展示了随着训练进行，具有固定种子和输入的视频样本的比较。随着训练计算量的增加，样本质量显著提高。



可变持续时间、分辨率、宽高比

过去的图像和视频生成方法通常将视频调整为标准大小，例如，4 秒的视频以 256x256 分辨率。我们发现，与其这样处理，训练原始大小的数据提供了几个好处。

采样灵活性

Sora 可以采样宽屏 1920x1080p 视频、竖屏 1080x1920 视频以及介于两者之间的所有内容。这使得 Sora 可以直接以原生宽高比为不同设备创建内容。它还使我们能够在生成全分辨率之前，快速原型化低分辨率的内容——而且只需使用同一个模型。

改进的构图和组合

我们凭经验发现，以视频的原生宽高比进行训练可以改善构图和组合。我们将 Sora 与我们的模型的一个版本进行比较，该版本将所有训练视频裁剪为正方形，这是训练生成模型时的常见做法。在使用正方形裁剪训练的模型（左侧）有时会生成主体仅部分可见的视频。相比之下，Sora 生成的视频（右侧）具有改进的构图。

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 330 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 231 and DALL·E 330 images.

A Shiba Inu dog wearing a beret and black turtleneck.

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.

语言理解

训练文本到视频生成系统需要大量具有对应文本标题的视频。我们将 DALL·E 3 引入的重新标题技术应用到视频中。我们首先训练一个高度描述性的标题模型，然后使用它为我们训练集中的所有视频生成文本标题。我们发现，训练在高度描述性视频标题上可以提高文本的准确性以及视频的整体质量。

类似于 DALL·E 3，我们还利用 GPT 将用户简短提示转换为更详细的长标题，然后将其发送给视频模型。这使得 Sora 能够生成高质量的视频，准确地遵循用户的提示。

使用图像和视频作为输入 prompt

我们在上述所有结果和我们的登陆页面上展示的都是文本到视频的样本。但是 Sora 也可以使用其他输入来提示，例如预先存在的图像或视频。这种能力使得 Sora 能够执行各种图像和视频编辑任务——创建完美循环的视频，给静态图像添加动画，将视频向前或向后延伸等等。

把 DALL·E 图像变成动画

Sora 能够生成基于 DALL·E 231 和 DALL·E 330 图像的视频，只需提供图像和提示作为输入。下面我们展示了基于这些图像生成的示例视频。

A Shiba Inu dog wearing a beret and black turtleneck.

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 four videos that were all extended backward in time starting from a segment of a generated video. As a result, each of the four videos starts different from the others, yet all four videos lead to the same ending. We can use this method to extend a video both forward and backward to produce a seamless infinite loop.

Video-to-video editing

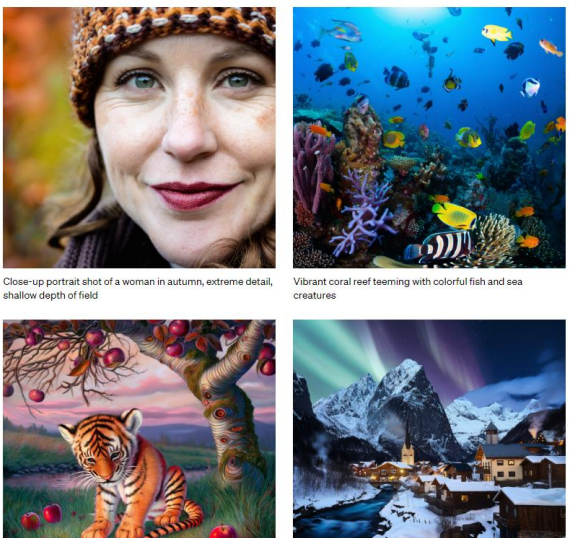
Diffusion models have enabled a plethora of methods for editing images and videos from text prompts. Below we apply one of these methods, SDEdit,32 to Sora. This technique enables Sora to transform the styles and environments of input videos zero-shot.

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 2048x2048 resolution.



延长生成的视频

Sora 还能够延长视频，无论是向前还是向后延长。下面是四个视频，它们都是从一个生成的视频片段开始向时间的后方延长。因此，这四个视频的开头各不相同，但最终都会导向相同的结尾。我们你也可以用这个方法扩展一个视屏的头和尾让他首尾相连成一个无限循环的视频。

视频到视频编辑

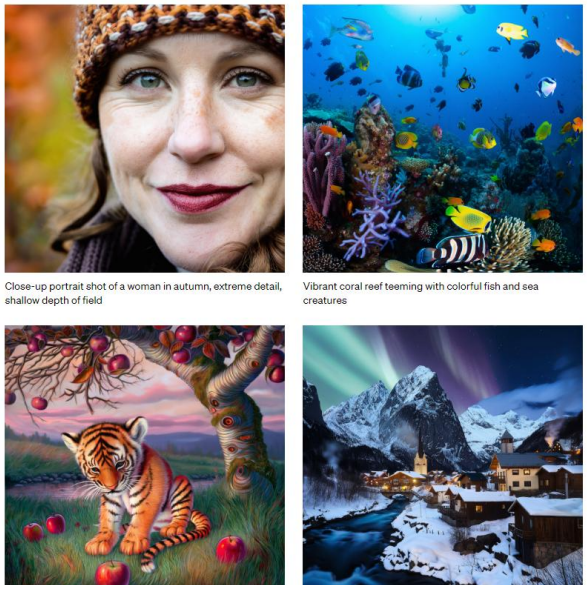
扩散模型已经为从文本提示编辑图像和视频提供了大量方法。下面我们将其中一种方法，SDEdit，应用到 Sora 上。这种技术使得 Sora 能够在零样本情况下转换输入视频的风格和环境。

连接视频

我们还可以使用 Sora 逐渐插值两个输入视频之间，从而在完全不同的主题和场景构图的视频之间创建无缝的过渡。在下面的示例中，中间的视频在左侧和右侧对应视频之间进行插值。

图像生成能力

Sora 也能够生成图像。我们通过将高斯噪声的补丁以一个帧的时间范围排列成空间网格来实现这一点。该模型可以生成不同尺寸的图像，分辨率高达 2048x2048。



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.”

涌现出模拟的能力

我们发现，在大规模训练时，视频模型表现出许多有趣的新兴能力。这些能力使得 Sora 能够模拟来自物理世界的一些人、动物和环境的方面。这些属性是在没有任何明确的归纳偏见的情况下出现的，比如对 3D、物体等——它们纯粹是规模现象。

3D 一致性。Sora 可以生成具有动态摄像机运动的视频。随着摄像机的移动和旋转，人物和场景元素在三维空间中保持一致的移动。

长程连贯性和对象永恒性。

对于视频生成系统来说，一个重要挑战是在采样长视频时保持时间一致性。我们发现，Sora 通常能够有效地模拟短期和长期依赖关系，尽管并非总是如此。例如，我们的模型可以在人、动物和物体被遮挡或离开画面时仍然保持其持久性。同样地，它可以在一个样本中生成同一个角色的多个镜头，并在整个视频中保持其外观。

与世界进行交互。Sora 有时可以模拟一些简单方式影响世界状态的动作。例如，一个画家可以在画布上留下持续一段时间的新笔触，或者一个人可以吃掉一个汉堡并留下咬痕。


模拟数字世界。

Sora 还能够模拟人工过程，一个例子是视频游戏。Sora 可以同时使用基本策略控制《我的世界》中的玩家，并以高保真度呈现世界及其动态。这些能力可以通过提示 Sora 提到“Minecraft”的标题来零样本激发。

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.

这些能力表明，持续扩展视频模型是发展高度能力的物理世界和数字世界模拟器，以及其中的物体、动物和人的有前景的途径。

讨论

目前，Sora 作为模拟器表现出了许多限制。例如，它并不能准确地模拟许多基本交互的物理特性，比如玻璃破碎。其他交互，比如吃食物，并不总是产生正确的物体状态变化。我们在我们的登陆页面上列举了模型的其他常见失败模式——例如，在长时间样本中发展的不一致性或对象的突然出现。

我们相信，Sora 目前的能力证明了持续扩展视频模型是发展能力强大的物理世界和数字世界模拟器，以及其中的物体、动物和人的有前景的途径。