微软37页论文逆向工程Sora，得到了哪些结论？

机器之心 · 2024/03/08

原文地址：https://www.linkresearcher.com/theses/976a4e01-a13a-4827-a8c0-582dd8d6ed29

论文

论文标题：Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models

作者：Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, Lichao Sun

期刊：arXiv

发表时间：2024/02/27

数字识别码：arXiv:2402.17177

摘要：Sora is a text-to-video generative AI model, released by OpenAI in February 2024. The model is trained to generate videos of realistic or imaginative scenes from text instructions and show potential in simulating the physical world. Based on public technical reports and reverse engineering, this paper presents a comprehensive review of the model's background, related technologies, applications, remaining challenges, and future directions of text-to-video AI models. We first trace Sora's development and investigate the underlying technologies used to build this "world simulator". Then, we describe in detail the applications and potential impact of Sora in multiple industries ranging from film-making and education to marketing. We discuss the main challenges and limitations that need to be addressed to widely deploy Sora, such as ensuring safe and unbiased video generation. Lastly, we discuss the future development of Sora and video generation models in general, and how advancements in the field could enable new ways of human-AI interaction, boosting productivity and creativity of video generation.

摘要翻译（由计算机程序完成，仅供参考，内容以英文原文为准）：

Sora是一个文本到视频生成的人工智能模型，由OpenAI于2024年2月发布。该模型经过训练，可以根据文本指令生成逼真或富有想象力的场景的视频，并显示出模拟物理世界的潜力。基于公开的技术报告和逆向工程，本文全面回顾了文本到视频人工智能模型的背景、相关技术、应用、剩余挑战和未

> 来方向。我们首先追踪索拉的发展，并研究用于构建这个“世界模拟器”的底层技术。然后，我们详细描述了索拉在从电影制作、教育到营销等多个行业的应用和潜在影响。我们讨论了广泛部署索拉需要解决的主要挑战和局限性，例如确保安全、公正的视频生成。最后，我们讨论了索拉和视频生成模型的未来发展，以及该领域的进步如何能够实现人类人工智能交互的新方式，提高视频生成的生产力和创造力。

> 所属学科：计算机

阅读论文原文

一篇论文回顾 Sora 文生视频技术的背景、技术和应用。

追赶 Sora，成为了很多科技公司当下阶段的新目标。研究者们好奇的是：Sora 是如何被 OpenAI 发掘出来的？未来又有哪些演进和应用方向？

Sora 的技术报告披露了一些技术细节，但远远不足以窥其全貌。

在最近的一篇文章中，微软研究院和理海大学的研究者根据已发表的技术报告和逆向工程，首次全面回顾了 Sora 的背景、相关技术、新兴应用、当前局限和未来机遇。

背景

在分析 Sora 之前，研究者首先盘点了视觉内容生成技术的沿袭。

在深度学习革命之前，传统的图像生成技术依赖于基于手工创建特征的纹理合成和纹理映射等方法。这些方法在生成复杂而生动的图像方面能力有限。

如图 3 所示，在过去十年中，视觉类的生成模型经历了多样化的发展路线。

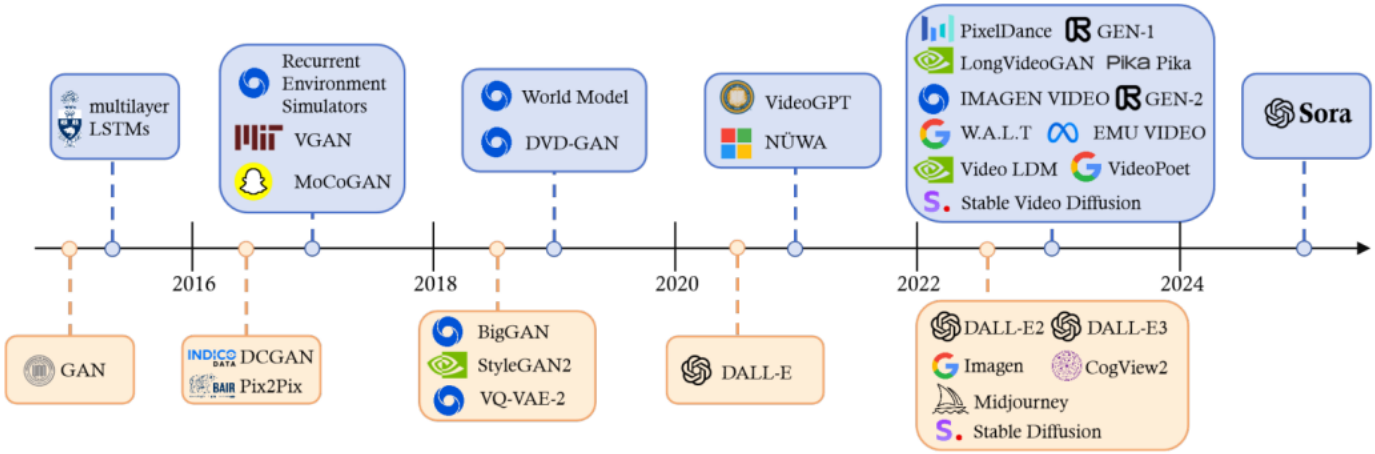


Figure 3: History of Generative AI in Vision Domain.

生成对抗网络（GAN）和变分自动编码器（VAE）的引入标志着一个重要的转折点，因为它在各种应用中都具有非凡的能力。随后的发展，如流模型和扩散模型，进一步增强了图像生成的细节和质量。人工智能生成内容（AIGC）技术的最新进展实现了内容创建的民主化，使用户能够通过简单的文本指令生成所需的内容。

在 BERT 和 GPT 成功将 Transformer 架构应用于 NLP 之后，研究人员尝试将其迁移到 CV 领域，比如 Transformer 架构与视觉组件相结合，使其能够应用于下游 CV 任务，包括 Vision Transformer (ViT) 和 Swin Transformer ，从而进一步发展了这一概念。在 Transformer 取得成功的同时，扩散模型也在图像和视频生成领域取得了长足进步。扩散模型为利用 U-Nets 将噪声转换成图像提供了一个数学上合理的框架，U-Nets 通过学习在每一步预测和减轻噪声来促进这一过程。

自 2021 年以来，能够解释人类指令的生成语言和视觉模型，即所谓的多模态模型，成为了人工智能领域的热门议题。

CLIP 是一种开创性的视觉语言模型，它将 Transformer 架构与视觉元素相结合，便于在大量文本和图像数据集上进行训练。通过从一开始就整合视觉和语言知识，CLIP 可以在多模态生成框架内充当图像编码器。

另一个值得注意的例子是 Stable Diffusion，它是一种多用途文本到图像人工智能模型，以其适应性和易用性而著称。它采用 Transformer 架构和潜在扩散技术来解码文本输入并生成各种风格的图像，进一步说明了多模态人工智能的进步。

ChatGPT 2022 年 11 月发布之后，2023 年出现了大量文本到图像的商业化产品，如 Stable Diffusion、Midjourney、DALL-E 3。这些工具能让用户通过简单的文字提示生成高分辨率和高质量的

新图像，展示了人工智能在创意图像生成方面的潜力。

然而，由于视频的时间复杂性，从文本到图像到文本到视频的过渡具有挑战性。尽管工业界和学术界做出了许多努力，但大多数现有的视频生成工具，如 Pika 和 Gen-2 ，都仅限于生成几秒钟的短视频片段。

在这种情况下，Sora 是一项重大突破，类似于 ChatGPT 在 NLP 领域的影响。Sora 是第一个能够根据人类指令生成长达一分钟视频的模型，同时保持较高的视觉质量和引人注目的视觉连贯性，从第一帧到最后一帧都具有渐进感和视觉连贯性。

这是一个里程碑，对生成式 AI 的研究和发展产生了深远影响。

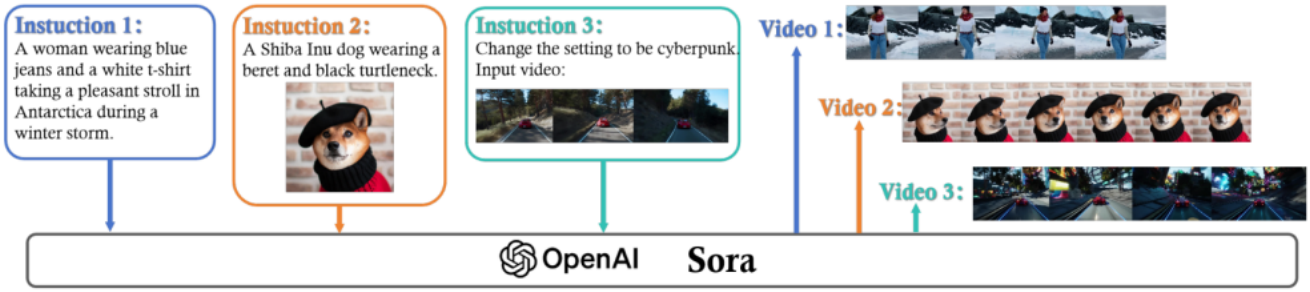


Figure 2: Examples of `Sora` in text-to-video generation. Text instructions are given to the OpenAI `Sora` model, and it generates three videos according to the instructions.

如图 2 所示，Sora 在准确解读和执行复杂的人类指令方面表现出非凡的能力。该模型可以生成包含多个角色的详细场景，这些角色在错综复杂的背景下执行特定的动作。研究人员认为，Sora 不仅能熟练处理用户生成的文本提示，还能辨别场景中各种元素之间复杂的相互作用。

此外，Sora 的进步还体现在它能够生成具有细微运动和交互描绘的扩展视频序列，克服了早期视频生成模型所特有的短片段和简单视觉渲染的限制。这种能力代表了人工智能驱动的创意工具的飞跃，使用户能够将文字叙述转换成丰富的视觉故事。

总之，这些进步显示了 Sora 作为世界模拟器的潜力，它可以提供对所描绘场景的物理和背景动态的细微洞察。

为了方便读者查阅视觉生成模型的最新进展，研究者在论文附录汇编了近期的代表性工作成果。

Table 1: Summary of Video Generation.

Model name	Year	Backbone	Task	Group
Imagen Video[29]	2022	Diffusion	Generation	Google
Pix2Seq-D[160]	2022	Diffusion	Segmentation	Google Deepmind
FDM[161]	2022	Diffusion	Prediction	UBC
MaskViT[162]	2022	Masked Vision Models	Prediction	Stanford, Salesforce
CogVideo[163]	2022	Auto-regressive	Generation	THU
Make-a-video[164]	2022	Diffusion	Generation	Meta
MagicVideo[165]	2022	Diffusion	Generation	ByteDance
TATS[166]	2022	Auto-regressive	Generation	University of Maryland, Meta
Phenaki[167]	2022	Masked Vision Models	Generation	Google Brain
Gen-1[168]	2023	Diffusion	Generation, Editing	RunwayML
LFDM[140]	2023	Diffusion	Generation	PSU, UCSD
Text2video-Zero[169]	2023	Diffusion	Generation	Picsart
Video Fusion[170]	2023	Diffusion	Generation	USAC, Alibaba
PYoCo[34]	2023	Diffusion	Generation	Nvidia
Video LDM[36]	2023	Diffusion	Generation	University of Maryland, Nvidia
RIN[171]	2023	Diffusion	Generation	Google Brain
LVD[172]	2023	Diffusion	Generation	UCB
Dreamix[173]	2023	Diffusion	Editing	Google
MagicEdit[174]	2023	Diffusion	Editing	ByteDance
Control-A-Video[175]	2023	Diffusion	Editing	Sun Yat-Sen University
StableVideo[176]	2023	Diffusion	Editing	ZJU, MSRA
Tune-A-Video[78]	2023	Diffusion	Editing	NUS
Rerender-A-Video[177]	2023	Diffusion	Editing	NTU
Pix2Video[178]	2023	Diffusion	Editing	Adobe, UCL
InstructVid2Vid[179]	2023	Diffusion	Editing	ZJU
DiffAct[180]	2023	Diffusion	Action Detection	University of Sydney
DiffPose[181]	2023	Diffusion	Pose Estimation	Jilin University
MAGVIT[182]	2023	Masked Vision Models	Generation	Google
AnimateDiff[138]	2023	Diffusion	Generation	CUHK
MAGVIT V2[47]	2023	Masked Vision Models	Generation	Google
Generative Dynamics[183]	2023	Diffusion	Generation	Google
VideoCrafter[81]	2023	Diffusion	Generation	Tencent
Zeroscope[184]	2023	-	Generation	EasyWithAI
ModelScope	2023	-	Generation	Damo
Gen-2[23]	2023	-	Generation	RunwayML
Pika[22]	2023	-	Generation	Pika Labs
Emu Video[185]	2023	Diffusion	Generation	Meta
PixelDance[186]	2023	Diffusion	Generation	ByteDance
Stable Video Diffusion[27]	2023	Diffusion	Generation	Stability AI
W.A.L.T[187]	2023	Diffusion	Generation	Stanford, Google
Fairy[188]	2023	Diffusion	Generation, Editing	Meta
VideoPoet[189]	2023	Auto-regressive	Generation, Editing	Google
LGVI[190]	2024	Diffusion	Editing	PKU, NTU
Lumiere[191]	2024	Diffusion	Generation	Google
Sora[3]	2024	Diffusion	Generation, Editing	OpenAI

技术推演

Sora 的核心是一个预训练的扩散 Transformer。事实证明，Transformer 模型在许多自然语言任务中都具有可扩展性和有效性。与 GPT-4 等强大的大型语言模型（LLM）类似，Sora 可以解析文本并理解

复杂的用户指令。为了提高视频生成的计算效率，Sora 采用了时空潜在 patch 作为其构建模块。

具体来说，Sora 会将原始输入视频压缩为潜在时空表示。然后，从压缩视频中提取一系列潜在时空 patch，以囊括短暂时间间隔内的视觉外观和运动动态。这些片段类似于语言模型中的词 token，为 Sora 提供了详细的视觉短语，可用于构建视频。Sora 的文本到视频生成由扩散 Transformer 模型完成。从充满视觉噪音的帧开始，该模型会对图像进行迭代去噪，并根据提供的文本提示引入特定细节。本质上讲，生成的视频是通过多步完善过程产生的，每一步都会对视频进行完善，使其更加符合所需的内容和质量。

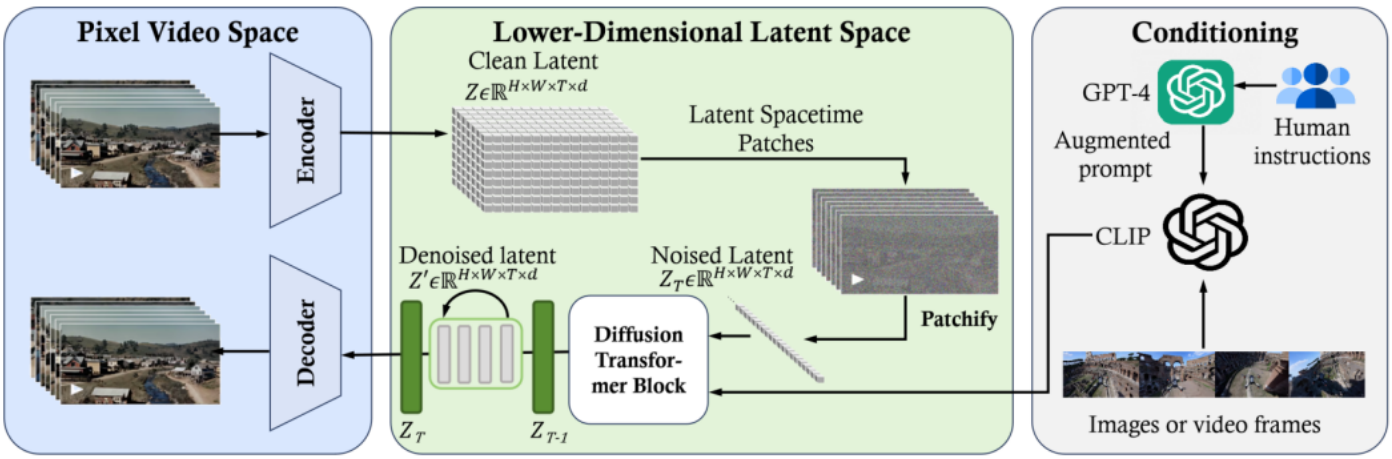


Figure 4: **Reverse Engineering**: Overview of Sora framework

如图 4 所示，Sora 的核心本质是一个具有灵活采样维度的扩散 Transformer。它由三部分组成：（1）时空压缩器首先将原始视频映射到潜在空间。(2) 然后，ViT 处理 token 化的潜在表示，并输出去噪潜在表示。(3) 类似 CLIP 的调节机制接收 LLM 增强的用户指令和潜在的视觉提示，引导扩散模型生成风格化或主题化的视频。经过许多去噪步骤后，生成视频的潜在表示被获取，然后通过相应的解码器映射回像素空间。

在本节中，研究者对 Sora 所使用的技术进行了逆向工程，并讨论了一系列相关工作。

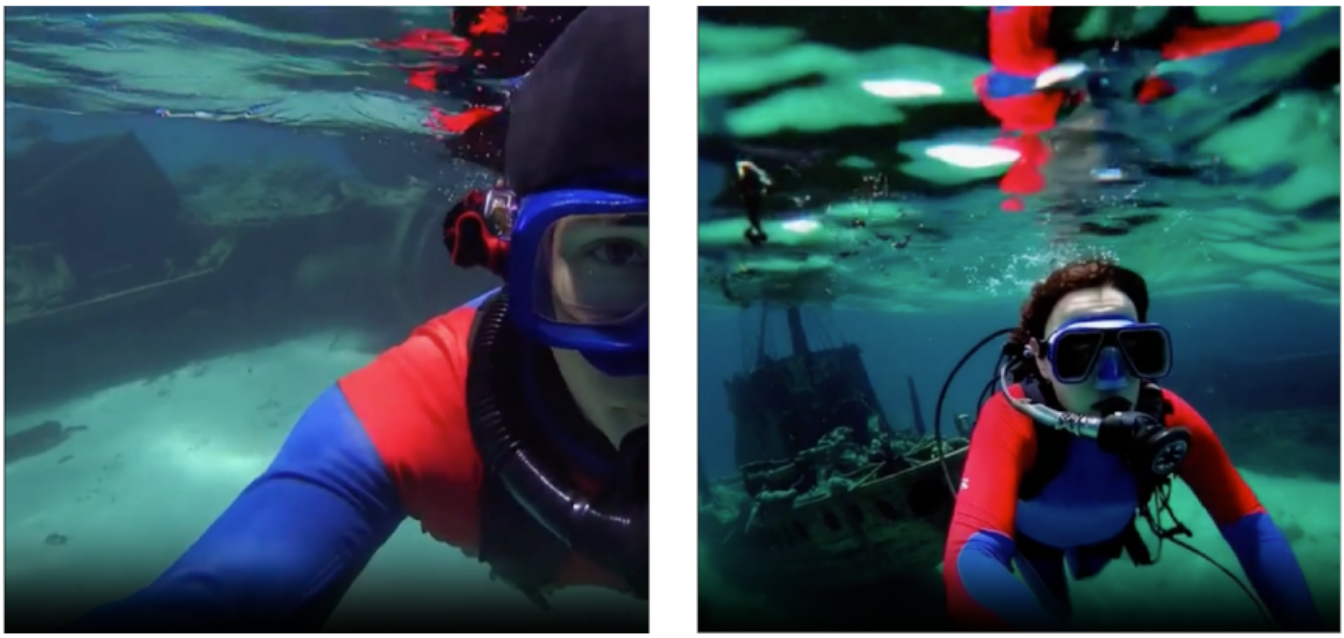
数据预处理

Sora 的一个显著特征是它能够训练、理解和生成原始尺寸的视频和图像，如图 5 所示。而传统方法通常会调整视频大小、裁剪或调整视频的长宽比以适应统一的视频和图像。利用扩散 Transformer 架构，Sora 是第一个拥抱视觉数据多样性的模型，可以以多种视频和图像格式进行采样，范围从宽屏 1920x1080p 视频到垂直 1080x1920p 视频以及介于两者之间的视频，而不影响其原始尺寸。

(a) Vertical (b) Square (c) Horizontal

Figure 5: Sora can generate images in flexible sizes or resolutions ranging from 1920x1080p to 1080x1920p and anything in between.

如图 6 所示，Sora 生成的视频能够更好的展现主题，从而确保在场景中完全捕捉到拍摄对象，而其他视频有时会导致视图被截断或裁剪，导致拍摄对象脱离画面。



(a) Training on videos that are cropped to squares leads to unnatural compositions and framing.

(b) Training in native sizes improves framing.

Figure 6: A comparison between Sora (right) and a modified version of the model (left), which crops videos to square shapes—a common practice in model training—highlights the advantages.

统一视觉表示。为了有效处理不同持续时间、分辨率和高宽比的图像和视频，关键在于将所有形式的视觉数据转换为统一表示。

Sora 处理的过程是这样的：首先将视频压缩到低维潜在空间，然后将表示分解为时空 patch 来对视频进行 patch 化（patchifies）。但是回看 Sora 技术报告，他们仅仅提出了一个高层次的想法，这给研究界的复现带来了挑战。在接下来的章节中，本文尝试对 Sora 的技术路径进行逆向工程，并且借鉴现有文献，讨论可以复现 Sora 的可行替代方案。

首先是视频压缩网络。Sora 的视频压缩网络（或视觉编码器）旨在降低输入数据（尤其是原始视频）的维度，并输出在时间和空间上压缩过的潜在表示，如图 7 所示。根据技术报告中的参考文献， Sora 压缩网络是基于 VAE 或 VQ-VAE 技术的。

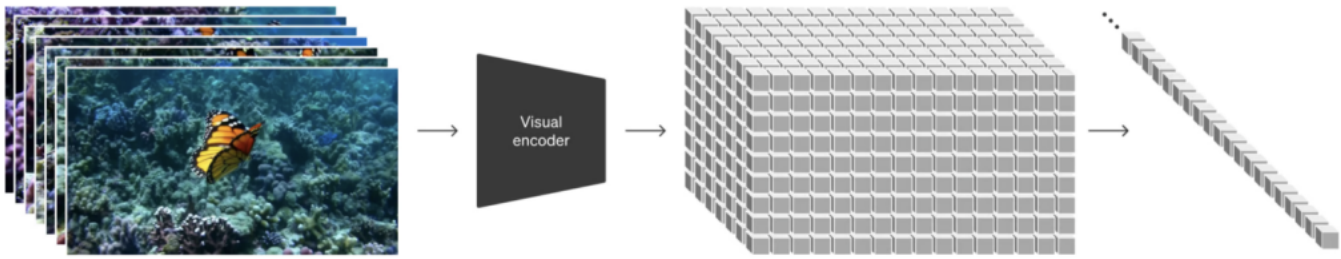


Figure 7: At a high level, Sora turns videos into patches by first compressing videos into a lower-dimensional latent space, and subsequently decomposing the representation into spacetime patches. Source: Sora's technical report [3].

然而，如果不像技术报告中对视频和图像调整大小和裁剪，那么 VAE 将任何大小的视觉数据映射到统一且固定大小的潜在空间挑战巨大。本文总结了两种不同的实现来解决这个问题：

空间 patch 压缩：涉及将视频帧转换为固定大小的 patch，类似于 ViT 和 MAE 中使用的方法（见图 8），然后将其编码到潜在空间中，这种方法对于适应不同分辨率和宽高比的视频特别有效。随后，将这些空间 token 按时间序列组织在一起，以创建时间 - 空间潜在表征。

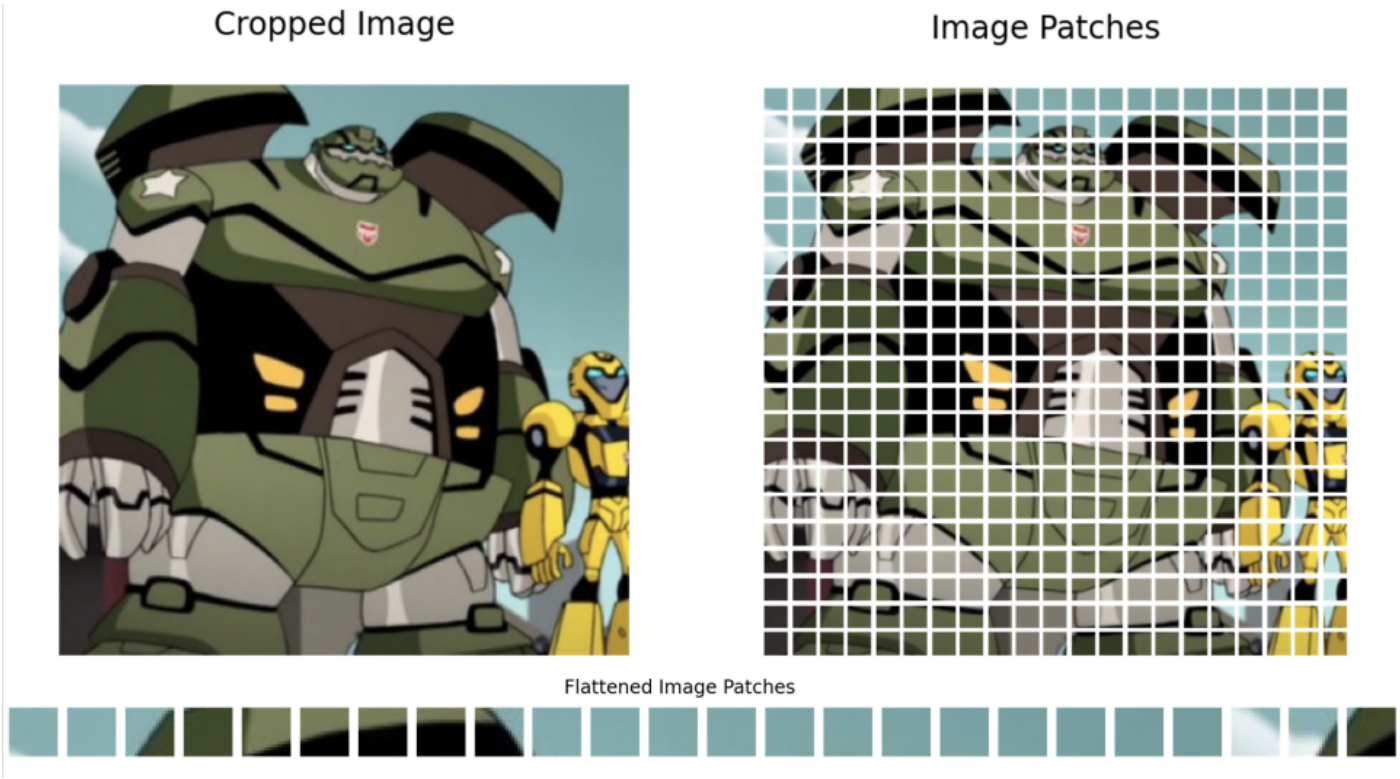


Figure 8: ViT splits an image into fixed-size patches, linearly embeds each of them, adds position embeddings, and feeds the resulting sequence of vectors to a standard Transformer encoder.

时间 - 空间 patch 压缩：该技术旨在封装视频数据的空间和时间维度，从而提供全面的表示。该技术不仅仅分析静态帧，还考虑帧间的运动和变化，从而捕获视频的动态信息。3D 卷积的利用成为实现这种集成的一种简单而有效的方法。

图 9 描绘了不同视频压缩方式的比较。与空间 patch 压缩类似，使用具有预定卷积核参数（例如固定内核大小、步幅和输出通道）的时间 - 空间 patch 压缩会导致潜在空间维度也不同。为了缓解这一挑战，空间修补（spatial patchification）所采用的方法在这种情况下同样适用和有效。

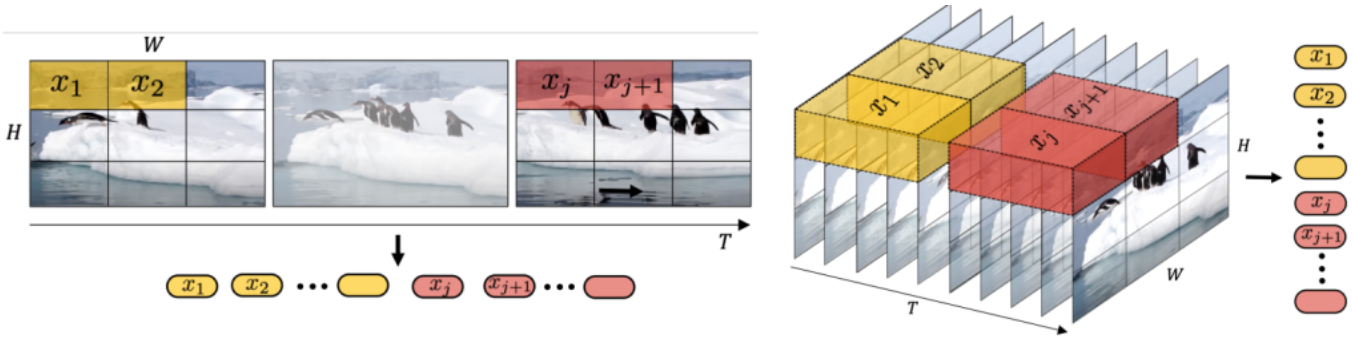


Figure 9: Comparison between different patchification for video compression. Source: ViViT [38]. (**Left**) Spatial patchification simply samples n_t frames and embeds each 2D frame independently following ViT. (**Right**) Spatial-temporal patchification extracts and linearly embeds non-overlapping or overlapping tubelets that span the spatiotemporal input volume.

总的来说，本文基于 VAE 或其变体如 VQ-VQE 逆向工程了两种 patch 级压缩方法，因为 patch 对处理不同类型的视频更加灵活。由于 Sora 旨在生成高保真视频，因此使用了较大尺寸的 patch 或内核尺寸以实现高效压缩。这里，本文期望使用固定大小的 patch，以简化操作、扩展性和训练稳定性。但也可以使用不同大小的 patch，以使整个帧或视频在潜在空间中的尺寸保持一致。然而，这可能导致位置编码无效，并且给解码器生成具有不同大小潜在 patch 的视频带来挑战。

压缩网络部分还有一个关键问题：在将 patch 送入扩散 Transformer 的输入层之前，如何处理潜在空间维度的变化（即不同视频类型的潜在特征块或 patch 的数量）。这里讨论了几种解决方案：

根据 Sora 的技术报告和相应的参考文献，patch n' pack（PNP）很可能是一种解决方案。如图 10 所示，PNP 将来自不同图像的多个 patch 打包在一个序列中。这种方法的灵感来源于自然语言处理中使用的样本打包，它通过丢弃 token 来实现对不同长度输入的高效训练。在这里，patch 化和 token 嵌入步骤需要在压缩网络中完成，但 Sora 可能会像 Diffusion Transformer（扩散 Transformer）那样，为 Transformer token 进一步 patch 化。

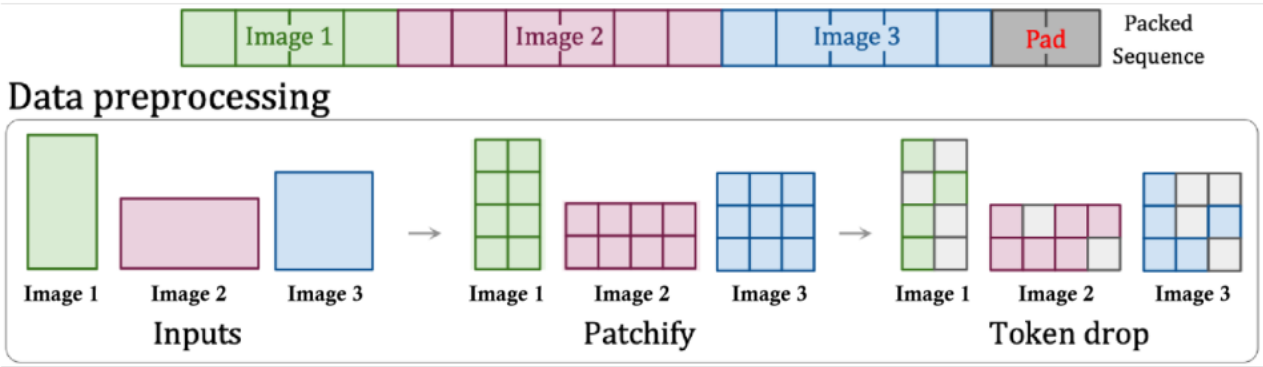


Figure 10: Patch packing enables variable resolution images or videos with preserved aspect ratio.6 Token dropping somehow could be treated as data augmentation. Source: NaViT [40].

无论是否有第二轮修补，都需要解决两个问题：如何以紧凑的方式打包这些 token，以及如何控制哪些 token 应该被丢弃。

对于第一个问题，研究者采用了简单的「贪心」算法，即在第一个序列中添加足够剩余空间的样本。一旦没有样本可以容纳，序列就会被填充 token 填满，从而产生批处理操作所需的固定序列长度。这种简单的打包算法可能会导致大量填充，这取决于输入长度的分布情况。另一方面，可以控制采样的分辨率和帧数，通过调整序列长度和限制填充来确保高效打包。

对于第二个问题，直观的方法是丢弃相似的 token，或者像 PNP 一样，使用丢弃率调度器。不过，值得注意的是，三维一致性是 Sora 的优良特性之一。在训练过程中，丢弃 token 可能会忽略细粒度的细节。因此，研究者认为 OpenAI 很可能会使用超长的上下文窗口并打包视频中的所有 token，尽管这样做的计算成本很高，例如，多头注意力算子在序列长度上表现出二次成本。具体来说，一个长时间视频中的时空潜在 patch 可以打包到一个序列中，而多个短时间视频中的时空潜在 patch 则会串联到另一个序列中。

建模

图像 DiT

传统的扩散模型主要利用包含下采样和上采样块的卷积 U-Net 作为去噪网络骨干。然而，最近的研究表明，U-Net 架构对扩散模型的良好性能并非至关重要。

通过采用更灵活的 Transformer 架构，基于 Transformer 的扩散模型可以使用更多的训练数据和更大的模型参数。沿着这一思路，DiT 和 U-ViT 是第一批将视觉 Transformer 用于潜在扩散模型的作品。

与 ViT 一样，DiT 也采用了多头自注意力层和层范数和缩放层交错的逐点前馈网络。如图 11 所示，DiT 还通过 AdaLN 进行调节，并增加了一个用于零初始化的 MLP 层，将每个残差块初始化为一个恒等函数，从而大大稳定了训练过程。DiT 的可扩展性和灵活性得到了经验验证。

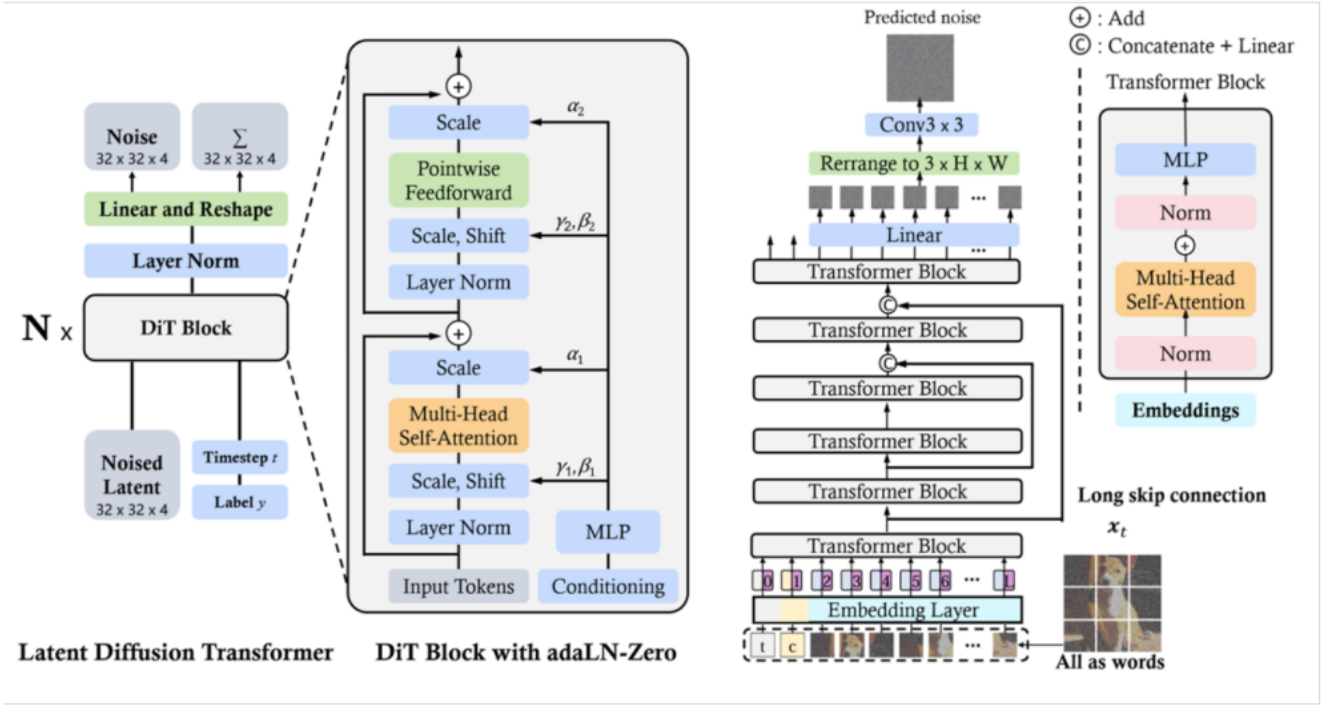


Figure 11: The overall framework of DiT (left) and U-ViT (right)

在 U-ViT 中，如图 11 所示，将包括时间、条件和噪声图像片段在内的所有输入都视为 token，并在浅层和深层 Transformer 层之间提出了长跳跃连接。结果表明，基于 CNN 的 U-Net 中的下采样和升采样算子并非总是必要的，U-ViT 在图像和文本到图像生成方面取得了破纪录的 FID 分数。

与掩蔽自编码器（MAE）一样，掩蔽扩散 Transformer（MDT）也在扩散过程中加入了掩码潜在模型，以明确增强图像合成中对象语义部分之间的上下文关系学习。

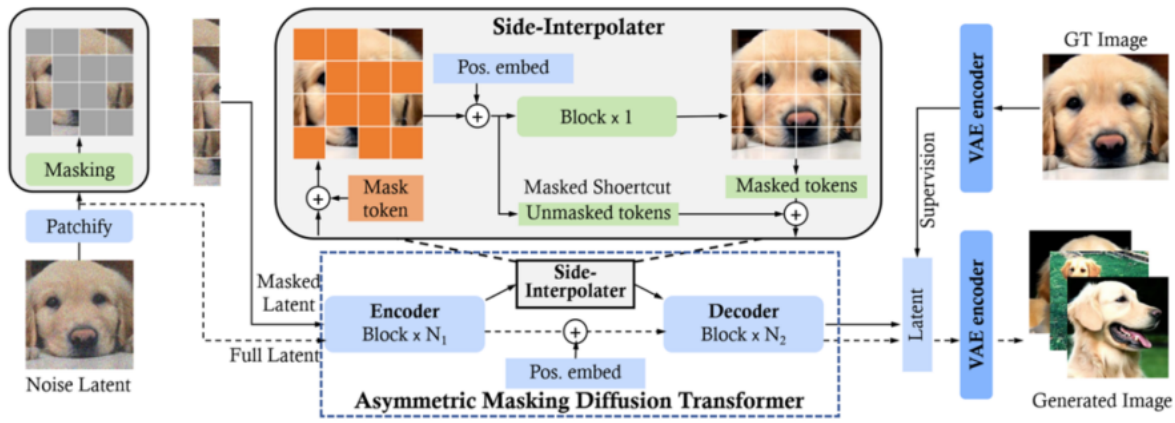


Figure 12: The overall framework of Masked Diffusion Transformer (MDT). A solid/dotted line indicates the training/inference process for each time step. Masking and side-interpolater are only used during training and are removed during inference.

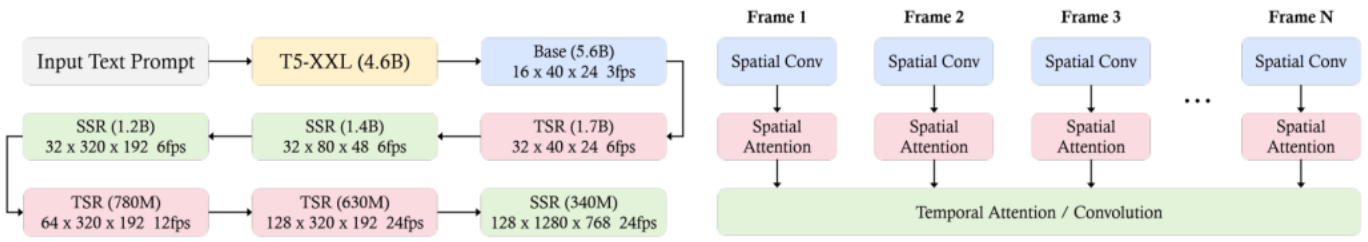
具体来说，如图 12 所示，MDT 在训练过程中使用边缘插值（side-interpolated）进行额外的掩蔽 token 重建任务，以提高训练效率，并学习强大的上下文感知位置嵌入进行推理。与 DiT 相比，MDT 实现了更好的性能和更快的学习速度。Hatamizadeh et al. 没有使用 AdaLN（即移位和缩放）进行时间条件建模，而是引入了 Diffusion Vision Transformers (DiffiT)，它使用与时间相关的自注意力（TMSA）模块对采样时间步长内的动态去噪行为进行建模。此外，DiffiT 采用两种混合分层架构，分别在像素空间和潜在空间进行高效去噪，并在各种生成任务中取得了新的先进成果。总之，这些研究表明，利用视觉 Transformer 进行图像潜在扩散取得了可喜的成果，为面向其他模态的研究铺平了道路。

视频 DiT

在文本到图像（T2I）扩散模型的基础上，一些近期研究专注于发挥扩散 Transformer 在文本到视频（T2V）生成任务中的潜力。由于视频的时空特性，在视频领域应用 DiT 所面临的主要挑战是：i) 如何将视频从空间和时间上压缩到潜在空间，以实现高效去噪；ii) 如何将压缩潜在空间转换为 patch，并将其输入 Transformer ；iii) 如何处理长序列时空依赖性，并确保内容一致性。

这里将讨论基于 Transformer 的去噪网络架构（该架构旨在时空压缩的潜在空间中运行）下文详细回顾了 OpenAI Sora 技术报告参考文献列表中介绍的两项重要工作（Imagen Video 和 Video LDM）。

Imagen Video 是谷歌研究院开发的文本到视频生成系统，它利用级联扩散模型（由 7 个子模型组成，分别执行文本条件视频生成、空间超分辨率和时间超分辨率）将文本提示转化为高清视频。



(a) **Cascaded diffusion models.** The cascaded sampling pipeline with a base diffusion model and six up-sampling models that operate spatially and temporally. The text embeddings are injected into all the diffusion models.

(b) **Video U-Net space-time separable block.** Spatial operations are performed independently over frames with shared parameters, whereas the temporal operation mixes activations over frames. Temporal attention is only used in the base model for memory efficiency.

Figure 13: The overall framework of Imagen Video. Source: Imagen Video [29].

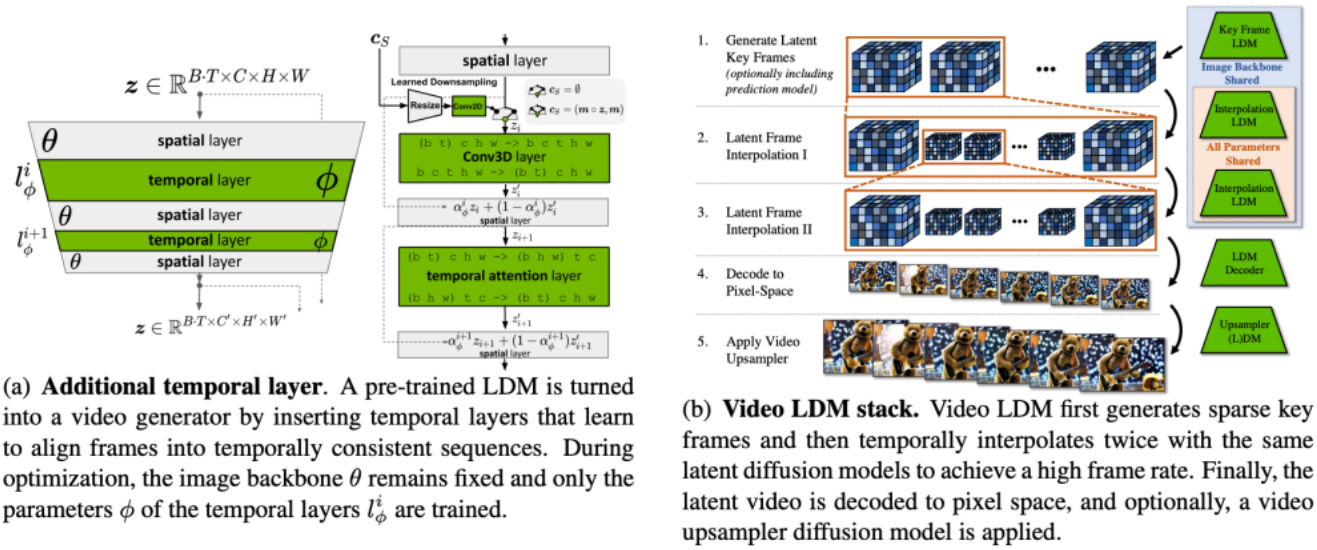
如图 13 所示，首先，冻结的 T5 文本编码器会根据输入的文本提示生成上下文嵌入。这些嵌入对于将生成的视频与文本提示对齐至关重要，除了基础模型外，它们还被注入级联中的所有模型。随后，嵌入信息被注入基础模型，用于生成低分辨率视频，然后由级联扩散模型对其进行细化以提高分辨率。

基础视频和超分辨率模型采用时空可分离的 3D U-Net 架构。该架构将时间注意力层和卷积层与空间对应层结合在一起，以有效捕捉帧间依赖关系。它采用 v 预测参数化来实现数值稳定性和条件增强，以促进跨模型的并行训练。

这一过程包括对图像和视频进行联合训练，将每幅图像视为一帧，以利用更大的数据集，并使用无分类器引导来提高提示保真度。渐进式蒸馏法用于简化采样过程，在保持感知质量的同时大大减少了计算负荷。将这些方法和技术相结合，Imagen Video 不仅能生成高保真视频，而且还具有出色的可控性，这体现在它能生成多样化的视频、文本动画和各种艺术风格的内容。

Blattmann et al. 建议将二维潜在扩散模型转化为视频潜在扩散模型（Video LDM）。为此，他们在 U-Net 主干网和 VAE 解码器的现有空间层中添加了一些临时时间层，以学习如何对齐单个帧。这些时间层在编码视频数据上进行训练，而空间层则保持固定，从而使模型能够利用大型图像数据集进行预训练。LDM 的解码器可进行微调，以实现像素空间的时间一致性和时间对齐扩散模型上采样器，从而提高空间分辨率。

为了生成超长视频，作者对模型进行了训练，以预测未来帧的上下文帧数，从而在采样过程中实现无分类器引导。为实现高时间分辨率，作者将视频合成过程分为关键帧生成和这些关键帧之间的插值。在级联 LDM 之后，使用 DM 将视频 LDM 输出进一步放大 4 倍，确保高空间分辨率的同时保持时间一致性。这种方法能以高效的计算方式生成全局一致的长视频。此外，作者还展示了将预先训练好的图像 LDM（如稳定扩散）转化为文本到视频模型的能力，只需训练时间对齐层，即可实现分辨率高达 1280 × 2048 的视频合成。



(a) **Additional temporal layer**. A pre-trained LDM is turned into a video generator by inserting temporal layers that learn to align frames into temporally consistent sequences. During optimization, the image backbone θ remains fixed and only the parameters ϕ of the temporal layers l_ϕ^i are trained.

(b) **Video LDM stack.** Video LDM first generates sparse key frames and then temporally interpolates twice with the same latent diffusion models to achieve a high frame rate. Finally, the latent video is decoded to pixel space, and optionally, a video upsampler diffusion model is applied.

Figure 14: The overall framework of Video LDM. Source: Video LDM [36].

语言指令跟随

为了提高文本到视频模型遵循文本指令的能力，Sora 采用了与 DALL・E 3 类似的方法。

DALL・E 3 中的指令跟随是通过一种描述改进方法来解决的，其假设是模型所训练的文本 - 图像对的质量决定了最终文本 - 图像模型的性能。数据质量差，尤其是普遍存在的噪声数据和省略了大量视觉信息的简短标题，会导致许多问题，如忽略关键词和词序，以及误解用户意图等。描述改进方法通过为现有图像重新添加详细的描述性描述来解决这些问题。该方法首先训练图像描述器（视觉语言模型），以生成精确的描述性图像描述。然后，描述器生成的描述性图像描述将用于微调文本到图像模型。

具体来说，DALL・E 3 采用对比式描述器（CoCa），联合训练具有 CLIP 架构和语言模型目标的图像描述器。该图像描述器包含一个图像编码器、一个用于提取语言信息的单模态文本编码器和一个多模态文本解码器。它首先在单模态图像和文本嵌入之间采用对比损失，然后对多模态解码器的输出采用描述损失。由此产生的图像描述器将根据对图像的高度详细描述进行进一步微调，其中包括主要对象、周围环境、背景、文本、风格和色彩。通过这一步骤，图像描述器就能为图像生成详细的描述性描述。文本到图像模型的训练数据集由图像描述生成器生成的重新描述数据集和真实人工编写数据混合而成，以确保模型捕捉到用户输入。

这种图像描述改进方法带来了一个潜在问题：实际用户提示与训练数据中的描述性图像描述不匹配。DALL・E 3 通过上采样解决了这一问题，即使用 LLM 将简短的用户提示改写成详细而冗长的说明。这确保了模型在推理时接收到的文本输入与模型训练时的文本输入保持一致。

为了提高指令跟踪能力，Sora 采用了类似的描述改进方法。这种方法是通过首先训练一个能够为视频制作详细说明的视频描述器来实现的。然后，将该视频描述器应用于训练数据中的所有视频，生成高质量的（视频、描述性描述）对，用于微调 Sora，以提高其指令跟随能力。

Sora 的技术报告没有透露视频描述器是如何训练的细节。鉴于视频描述器是一个视频到文本的模型，因此有很多方法来构建它：

一种直接的方法是利用 CoCa 架构来制作视频描述，方法是获取视频的多个帧，并将每个帧输入图像编码器，即 VideoCoCa。VideoCoCa 以 CoCa 为基础，重新使用图像编码器预训练的权重，并将其独立应用于采样视频帧。由此产生的帧 token 嵌入会被扁平化，并连接成一长串视频表示。然后，生成式池化层和对比池化层会对这些扁平化的帧 token 进行处理，二者是用对比损失和描述损失联合训练的。

其他可用于构建视频描述的方法包括 mPLUG-2、GIT、FrozenBiLM 等。

最后，为确保用户提示与训练数据中的描述性描述格式一致，Sora 还执行了额外的提示扩展步骤，即使用 GPT-4V 将用户输入扩展为详细的描述性提示。

然而，Sora 训练描述器的数据收集过程尚不清楚，而且很可能需要大量人力，因为这可能需要对视频进行详细描述。此外，描述性视频描述可能会对视频的重要细节产生幻觉。本文作者认为，如何改进视频描述器值得进一步研究，这对提高文本到图像模型的指令跟踪能力至关重要。

提示工程

文本提示

文本提示工程对于指导文本视频模型制作出既具有视觉冲击力又能精确满足用户规格的视频至关重要。这就需要制作详细的描述来指导模型，以有效弥合人类创造力与人工智能执行能力之间的差距。

Sora 的提示涵盖了广泛的场景。近期的作品（如 VoP、Make-A-Video 和 Tune-A-Video）展示了提示工程如何利用模型的自然语言理解能力来解码复杂指令，并将其呈现为连贯、生动和高质量的视频叙事。

如图 15 所示，「一个时髦的女人走在霓虹灯闪烁的东京街头……」就是这样一个精心制作的文本提示，它确保 Sora 生成的视频与预期的视觉效果非常吻合。提示工程的质量取决于对词语的精心选择、所提供细节的具体性以及对其对模型输出影响的理解。例如，图 15 中的提示详细说明了动作、设置、角色出场，甚至是所期望的场景情绪和氛围。

Figure 15: A case study on prompt engineering for text-to-video generation, employing color coding to delineate the creative process. The text highlighted in blue describes the elements generated by Sora, such as the depiction of a stylish woman. In contrast, the text in yellow accentuates the model's interpretation of actions, settings, and character appearances, demonstrating how a meticulously crafted prompt is transformed into a vivid and dynamic video narrative.

图像提示

图像提示为即将生成的视频内容和其他元素（如人物、场景和情绪）提供了视觉锚点。此外，文字提示还可以指示模型将这些元素动画化，例如，添加动作、互动和叙事进展等层次，使静态图像栩栩如生。通过使用图像提示，Sora 可以利用视觉和文本信息将静态图像转换成动态的、由叙事驱动的视频。

图 16 展示了人工智能生成的视频：「一只头戴贝雷帽、身穿高领毛衣的柴犬」、「一个独特的怪物家族」、「一朵云组成了 SORA 一词」以及「冲浪者在一座历史悠久的大厅内驾驭潮汐」。这些例子展示了通过 DALL・E 生成的图像提示 Sora 可以实现哪些功能。

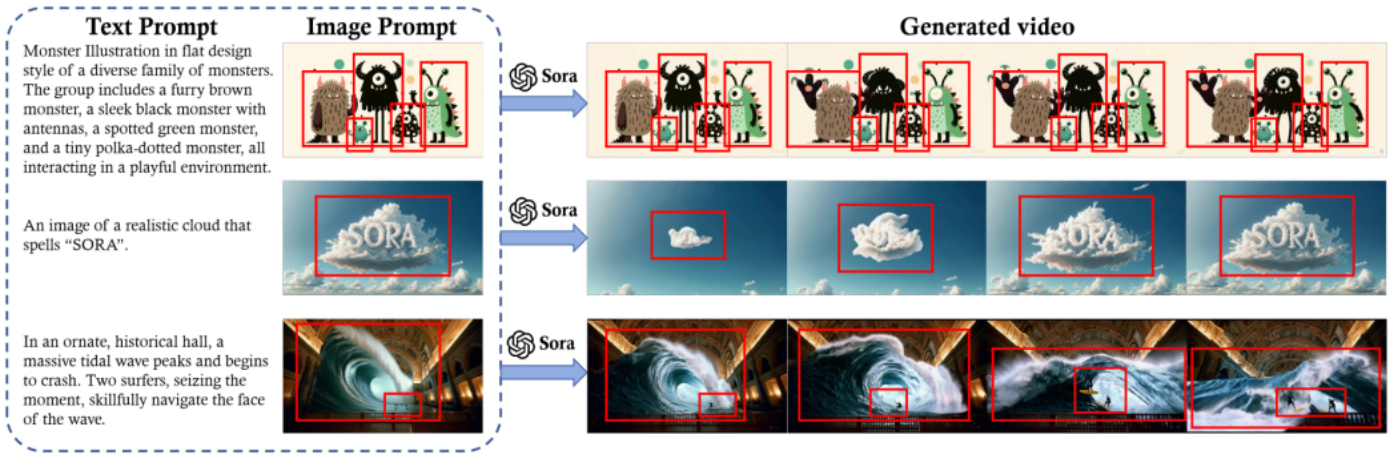


Figure 16: This example illustrates the image prompts to guide Sora's text-to-video model to generation. The red boxes visually anchor the key elements of each scene—monsters of varied designs, a cloud formation spelling "SORA", and surfers in an ornate hall facing a massive tidal wave.

视频提示

视频提示也可用于视频生成。最近的研究（如 Moonshot 和 Fast-Vid2Vid）表明，好的视频提示需要「具体」而「灵活」。这样既能确保模型在特定目标（如特定物体和视觉主题的描绘）上获得明确的指导，又能在最终输出中允许富有想象力的变化。

例如，在视频扩展任务中，提示可以指定扩展的方向（时间向前或向后）和背景或主题。在图 17 (a) 中，视频提示指示 Sora 向后延伸一段视频，以探索导致原始起点的事件。如图 17（b）所示，在通过视频提示执行视频到视频的编辑时，模型需要清楚地了解所需的转换，例如改变视频的风格、场景或氛围，或改变灯光或情绪等微妙的方面。在图 17 (c) 中，提示指示 Sora 连接视频，同时确保视频中不同场景中的物体之间平滑过渡。

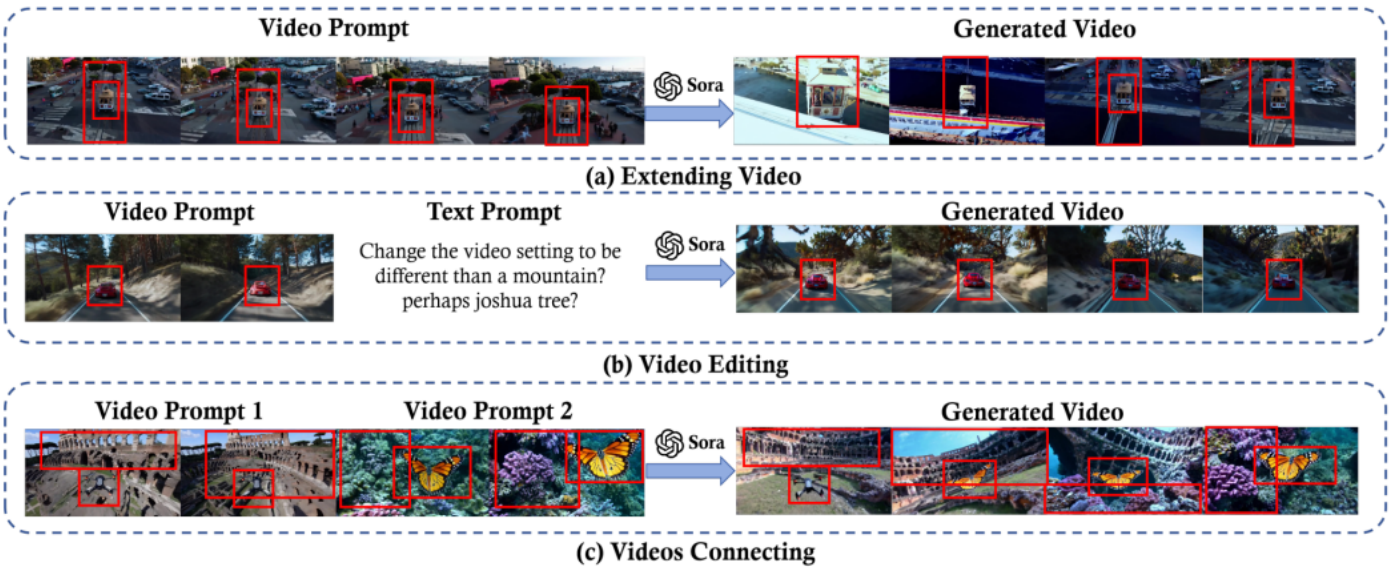


Figure 17: These examples illustrate the video prompt techniques for `Sora` models: (a) Video Extension, where the model extrapolates the sequence backward the original footage, (b) Video Editing, where specific elements like the setting are transformed as per the text prompt, and (c) Video Connection, where two distinct video prompts are seamlessly blended to create a coherent narrative. Each process is guided by a visual anchor, marked by a red box, ensuring continuity and precision in the generated video content.

虽然以前关于提示工程的研究主要集中在 LLM 和 LVM 的文本和图像提示上，但预计研究者们对视频生成模型的视频提示的兴趣会越来越大。

应用

随着以 Sora 为代表的视频扩散模型技术取得突破，其在不同研究领域和行业的应用正在迅速加速。

本文作者指出，这项技术的影响远远超出了单纯的视频创作，为从自动内容生成到复杂决策过程的各种任务提供了变革潜力。

在论文的第四章中，全面探讨了视频扩散模型的当前应用，希望为实际部署方案提供一个广阔的视角（图 18）：

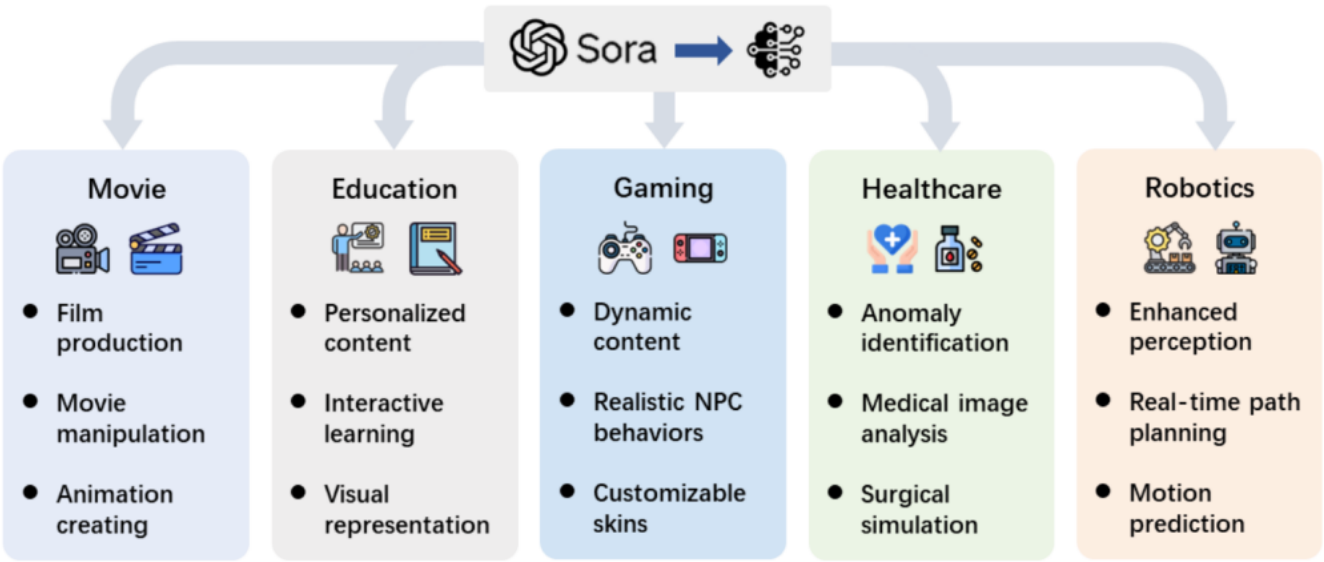


Figure 18: Applications of Sora.

提高模拟能力：对 Sora 进行大规模训练，是因为它能够出色地模拟物理世界的各个方面。尽管没有明确的三维建模，但 Sora 通过动态摄像机运动和远距离连贯性表现出三维一致性，包括物体持久性和模拟与世界的简单交互。此外，Sora 还能模拟类似 Minecraft 的数字环境，在保持视觉保真度的同时由基本策略控制，这一点非常有趣。这些新出现的能力表明，可扩展视频模型可以有效地创建人工智能模型，以模拟物理和数字世界的复杂性。

- 提高创造力：想象一下，通过文字勾勒出一个概念，无论是一个简单的物体还是一个完整的场景，都能在几秒钟内呈现出逼真或高度风格化的视频。Sora 可以加速设计过程，更快地探索和完善创意，从而大大提高艺术家、电影制作人和设计师的创造力。
- 推动教育创新：长期以来，视觉辅助工具一直是教育领域理解重要概念不可或缺的工具。有了 Sora，教育工作者可以轻松地将课堂计划从文字变成视频，吸引学生的注意力，提高学习效率。从科学模拟到历史剧，可能性是无限的。

- 增强可访问性：提高视觉领域的可访问性至关重要。Sora 通过将文字描述转换为可视内容，提供了一种创新的解决方案。这种功能使包括视觉障碍者在内的所有人都能积极参与内容创建，并以更有效的方式与他人互动。因此，它可以创造一个更具包容性的环境，让每个人都有机会通过视频表达自己的想法。
- 促进新兴应用：Sora 的应用领域非常广泛。例如，营销人员可以用它来制作针对特定受众描述的动态广告。游戏开发商可以利用它根据玩家的叙述生成定制的视觉效果甚至角色动作。

具体而言，以下几个行业将面临变革：

影视

传统上，创作电影是一个艰巨而昂贵的过程，往往需要数十年的努力、尖端的设备和大量的资金投入。先进视频生成技术的出现预示着电影制作进入了一个新时代，从简单的文本输入中自主生成电影的梦想正在成为现实。事实上，研究人员已经涉足电影生成领域，将视频生成模型扩展到电影创作中。

MovieFactory 应用扩散模型从 ChatGPT 制作的精心脚本中生成电影风格的视频，这是一个重大飞跃。在后续研究中，MobileVidFactory 只需用户提供简单的文本，就能自动生成垂直移动视频。Vlogger 则让用户可以制作长达一分钟的 Vlog。

Sora 能够毫不费力地生成引人入胜的电影内容，这是这些发展的缩影，标志着电影制作民主化的关键时刻。它们让人们看到了一个人人都能成为电影制作人的未来，大大降低了电影行业的准入门槛，并为电影制作引入了一个新的维度，将传统的故事讲述方式与人工智能驱动的创造力融为一体。这些技术的影响不仅仅是简单化。它们有望重塑电影制作的格局，使其在面对不断变化的观众喜好和发行渠道时，变得更加容易获得，用途更加广泛。

游戏

游戏产业一直在寻求突破逼真度和沉浸感界限的方法，但传统游戏开发往往受到预先渲染的环境和脚本事件的限制。通过扩散模型效果实时生成动态、高保真视频内容和逼真音效，有望克服现有的限制，为开发人员提供工具来创建不断变化的游戏环境，对玩家的行为和游戏事件做出有机的反应。这可能包括生成不断变化的天气条件、改变地貌，甚至即时创建全新的设置，从而使游戏世界更加身临其境、反应更加灵敏。一些方法还能从视频输入中合成逼真的冲击声，增强游戏音频体验。

将 Sora 集成到游戏领域后，就能创造出无与伦比的身临其境的体验，吸引并吸引玩家。游戏的开发、玩耍和体验方式都将得到创新，并为讲故事、互动和沉浸式体验带来新的可能性。

医疗

尽管具有生成能力，但视频扩散模型在理解和生成复杂视频序列方面表现出色，因此特别适用于识别人体内的动态异常，如早期细胞凋亡、皮肤病变进展和不规则人体运动，这对早期疾病检测和干预策略至关重要。此外，MedSegDiffV2 等模型利用 Transformer 的强大功能，以前所未有的精度分割医学影像，使临床医生能够在各种成像模式中精确定位感兴趣的区域，提高准确性。

将 Sora 集成到临床实践中，不仅有望完善诊断流程，还能根据精确的医学影像分析提供量身定制的治疗方案，实现患者护理的个性化。然而，这种技术整合也带来了一系列挑战，包括需要采取强有力的数据隐私措施和解决医疗保健中的伦理问题。

机器人

视频扩散模型目前在机器人技术中发挥着重要作用，它展示了一个新时代：机器人可以生成和解释复杂的视频序列，以增强感知和决策。这些模型释放了机器人的新能力，使它们能够与环境互动，以前所未有的复杂度和精确度执行任务。将网络规模扩散模型引入机器人学，展示了利用大规模模型增强机器人视觉和理解能力的潜力。潜在扩散模型被用于语言指导的视频预测，使机器人能够通过预测视频格式的行动结果来理解和执行任务。此外，视频扩散模型能够创建高度逼真的视频序列，创新性地解决了机器人研究依赖模拟环境的问题。这样就能为机器人生成多样化的训练场景，缓解真实世界数据匮乏所带来的限制。

将 Sora 等技术整合到机器人领域有望取得突破性发展。通过利用 Sora 的强大功能，未来的机器人技术将取得前所未有的进步，机器人可以无缝导航并与周围环境互动。

局限性

最后，研究者指出了 Sora 这项新技术存在的风险问题和局限性。

随着 ChatGPT 、GPT4-V 和 Sora 等复杂模型的快速发展，这些模型的能力得到了显著提高。这些发展为提高工作效率和推动技术进步做出了重大贡献。然而，这些进步也引发了人们对这些技术可能被滥

用的担忧，包括假新闻的产生、隐私泄露和道德困境。因此，大模型的可信度问题引起了学术界和工业界的广泛关注，成为当下研究讨论的焦点。

虽然 Sora 的成就凸显了人工智能的重大进步，但挑战依然存在。在描绘复杂动作或捕捉微妙面部表情方面，该模型还有待改进。此外，减少生成内容中的偏见和防止有害的视觉输出等道德方面的考虑也强调了开发人员、研究人员和更广泛的社区负责任使用的重要性。确保 Sora 的输出始终安全、无偏见是一项主要挑战。

但伴随着视频生成领域的发展，学术界和工业界的研究团队都取得了长足的进步。文本到视频竞争模式的出现表明，Sora 可能很快就会成为动态生态系统的一部分。这种合作与竞争的环境促进了创新，从而提高了视频质量并开发了新的应用，有助于提高工人的工作效率，使人们的生活更具娱乐性。