证券研究报告|行业专题报告
计算机行业
行业评级 **强于大市（维持评级）**
2024年2月25日

华福证券
HUAFU SECURITIES CO.,LTD.

Sora技术深度解析

证券分析师：
施晓俊 执业证书编号：S0210522050003
研究助理：
李杨玲
王　思

投资要点

- **Sora横空出世引领多模态产业革命。**美国时间2月15日，文生视频大模型Sora横空出世，能够根据文本指令或静态图像生成1分钟的视频。其中，视频生成包含精细复杂的场景、生动的角色表情以及复杂的镜头运动，同时也接受现有视频扩展或填补缺失的帧。总体而言，不管是在视频的保真度、长度、稳定性、一致性、分辨率、文字理解等方面，Sora都做到了业内领先水平，引领多模态产业革命。此外，当 Sora 训练的数据量足够大时，它也展现出了一种类似于涌现的能力，从而使得视频生成模型具备了类似于物理世界通用模拟器的潜力。
- **拆解视频生成过程，技术博采众长或奠定了Sora文生视频领军地位。**从技术报告中，Sora视频生成过程大致由“视频编码+加噪降噪+视频解码”三个步骤组成，视频压缩网络、时空patches、transformer架构、视频数据集等技术与资源在其中发挥了重要作用。
  - ✓ **视频压缩网络**：过往VAE应用于视频领域通常需插入时间层，Sora从头训练了能直接压缩视频的自编码器，可同时实现时间和空间的压缩，既节省算力资源，又最大程度上保留视频原始信息，或为Sora生成长视频的关键因素，并为后续处理奠定基础。
  - ✓ **时空patches**：1）同时考虑视频中时间和空间关系，能够捕捉到视频中细微的动作和变化，在保证视频内容连贯性和长度的同时，创造出丰富多样的视觉效果；2）突破视频分辨率、长宽比等限制的同时显著提升模型性能，节约训练与推理算力成本。
  - ✓ **Transformer架构**：1）相比于U-Net架构，transformer突显Scaling Law下的“暴力美学”，即参数规模越大、训练时长越长、训练数据集越大，生成视频的效果更好；2）此外，在transformer大规模训练下，逐步显现出规模效应，迸发了模型的涌现能力。
  - ✓ **视频数据集**：Sora或采用了更丰富的视频数据集，在原生视频的基础上，将DALL·E3的re-captioning技术应用于视频领域，同时利用GPT保障文字-视频数据集质量，使得模型具有强大的语言理解能力。
- **投资建议：我们认为，在视频压缩网络与时空patches提高计算效率与利用原生视频信息的基础上，transformer或取代U-Net成为扩散模型主流架构。可拓展性更强的transformer需要更为有力的算力支持才能保障视频生成质量，同时相比于大语言模型，视觉数据的训练与推理算力需求更大，因而算力有望成为确定性最高的受益赛道。此外，Sora发布有望形成多模态产业“鲇鱼效应”，激励其他多模态厂商的良性发展。**建议关注：1）AI算力：云赛智联、思特奇、恒为科技、海光信息、寒武纪、景嘉微、中科曙光、浪潮信息、拓维信息、四川长虹、工业富联、神州数码等；2）AI+多模态：万兴科技、虹软科技、当虹科技、中科创达、大华股份、海康威视、漫步者、萤石网络、汉仪股份、美图公司、云从科技。
- **风险提示：**技术发展不及预期、产品落地不及预期、AI伦理风险等。

目录

1.1 Sora横空出世，引领多模态产业革命

- **美国时间2月15日，文生视频大模型Sora横空出世，能够根据文本指令或静态图像生成1分钟的视频。**其中，视频生成包含精细复杂的场景、生动的角色表情以及复杂的镜头运动，同时也接受现有视频扩展或填补缺失的帧。
- **总体而言，不管是在视频的保真度、长度、稳定性、一致性、分辨率、文字理解等方面，Sora都做到了业内领先水平，引领多模态产业革命。**此外，当 Sora 训练的数据量足够大时，它也展现出了一种类似于涌现的能力，从而使得视频生成模型具备了类似于物理世界通用模拟器的潜力。

图表：Sora与业内主流视频生成模型对比

公司名称	生成功能	最长时长	时长可延展	相机控制（平移/变焦）	动作控制
Runway	文生视频 图生视频 视频生视频	4S	√	√	√
Pika	文生视频 图生视频	3S	√	√	√
Genmo	文生视频 图生视频	6S	×	√	√
Kaiber	文生视频 图生视频 视频生视频	16S	×	×	×
Stability	图生视频	4S	×	×	√
Sora	文生视频 图生视频 视频生视频	60S	√	√	√

资料来源：data learner，新浪新闻图数室，华福证券研究所

图表：Sora和其他模型优势对比总览

OpenAI Sora	能力项	其他模型
60秒	视频时长	最多十几秒
1920x1080与1080x1920之间任意尺寸	视频长宽比	固定尺寸 如16:9,9:16,1:1等
1080P	视频清晰度	upscale之后达到4K
支持	文本生成视频	支持
支持	图片生成视频	支持
支持	视频生成视频	支持
支持	文本编辑视频	支持
向前/向后扩展	扩展视频	仅支持向后扩展
支持	视频连接	不支持
支持	真实世界模拟	支持
强	运动相机模拟	弱
强	依赖关系进行建模	弱
强	影响世界状态(世界交互)	弱

资料来源：data learner，新浪新闻图数室，华福证券研究所

1.2 Sora视频生成过程：视频编码+加噪降噪+视频解码

- **从技术报告中，Sora视频生成过程大致由以下三个步骤组成：**

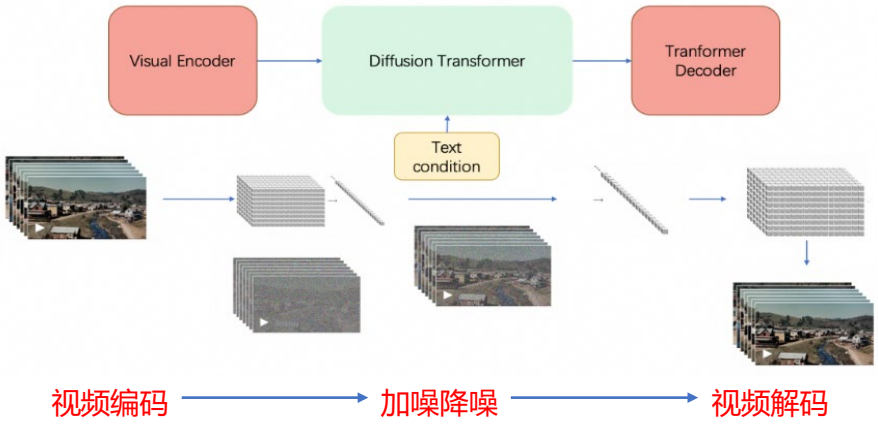
✓ **视频编码：**Visual Encoder将原始视频压缩为低维潜在空间，再将视频分解为时空patches后拉平为系列视频token以供transformer处理。

✓ **加噪降噪：**在transfomer架构下的扩散模型中，时空patches融合文本条件化，先后经过加噪和去噪，以达到可解码状态。

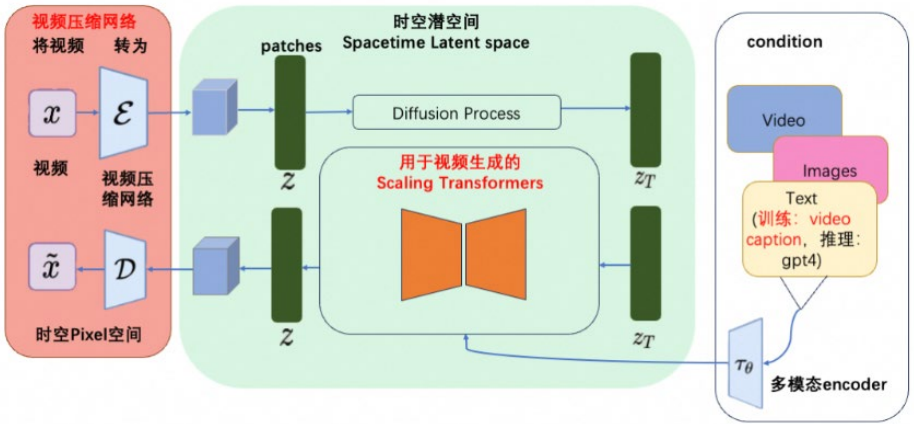
✓ **视频解码：**将去噪后的低维潜在表示映射回像素空间。

- 总体而言，我们认为Sora技术报告虽未能详尽阐述视频生成技术细节，但从参考技术文献中，可初步窥探出时空patches、视频压缩网络、 Transformer技术架构、独特文本标注视频数据集等技术与资源优势，这些或为Sora占据业内领先地位的原因。

图表：Sora视频生成过程图



图表：Sora技术架构猜想



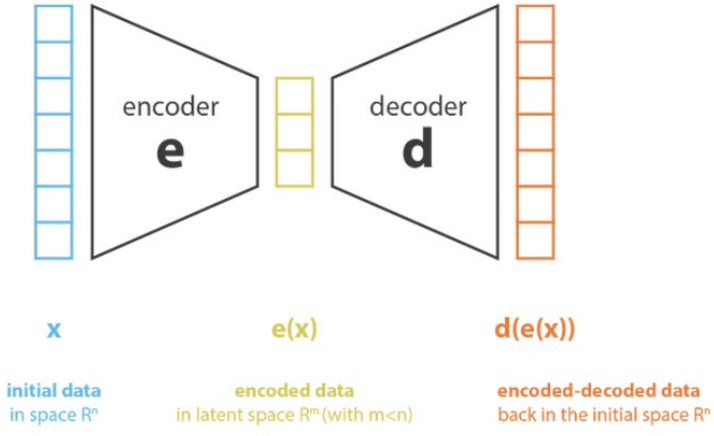
资料来源：魔塔ModelScope社区，华福证券研究所

目录

2.1 视频压缩网络实现降维，或为长视频生成基础

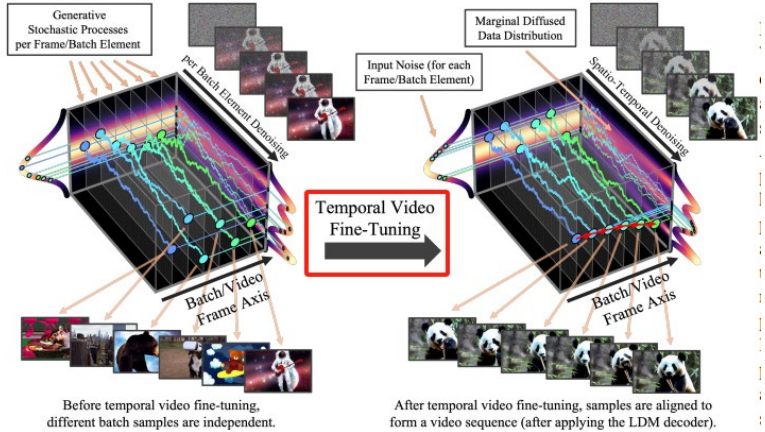
- **OpenAI训练了降低视觉数据维度的网络，该网络接受原始视频作为输入，并输出在时间和空间上都被压缩的潜在表示。**Sora在这个压缩的潜在空间上进行训练，并随后生成视频。与之对应，Sora训练了相应的解码器模型，将生成的潜在表示映射回像素空间。
- **压缩网络本质上是将高维数据映射至低维空间，低维空间中每个点通常对应原始高维数据的潜在表示，在复杂性降低和细节保留之间达到最优平衡点，实现提升视觉保真度的同时降低算力资源消耗的作用。**
- ✓ VAE为图片生成领域的常见图片编码器，应用到视频领域则需要加入时间维度以形成视频框架。例如，2023年发布的VideoLDM通过将视频拆解为每一帧，之后插入时间对齐层，从而实现了视频生成。
- ✓ Sora从头训练了能直接压缩视频的自编码器，既能实现空间压缩图像，又能在时间上压缩视频。**我们认为，在时空维度上压缩视频，既节省了算力资源，又最大程度上保留视频原始信息，或为Sora生成60s长视频的关键因素，并为后续时空patches和transfomer架构处理奠定基础。**

图表：VAE技术原理图，由编码器和解码器组成



资料来源：Medium，Towards Data Science，华福证券研究所

图表：VideoLDM在图片编码器基础上加入时间维度



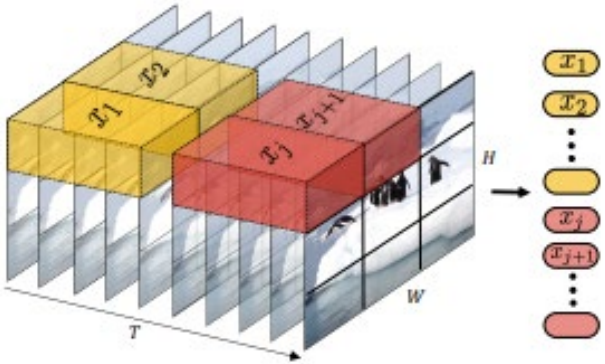
资料来源：Blattmann et al.《Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models》，华福证券研究所

2.2 时空patches统一视频分割，奠定处理和理解复杂视觉内容的基石

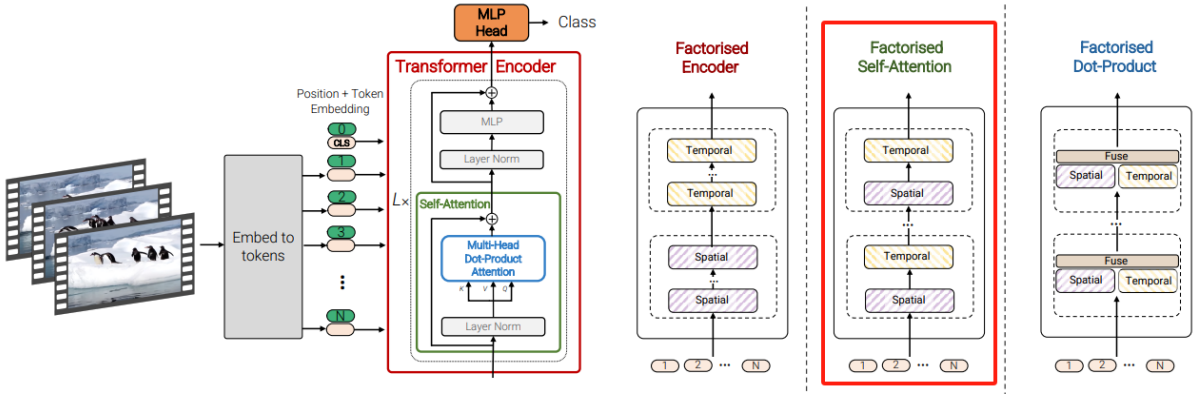
华福证券 HUAFU SECURITIES CO.,LTD.

- Sora借鉴LLM中将文本信息转化为token的思路，针对视频训练视觉patch，实现视觉数据模型的统一表达，实现对多样化视频和图像内容的有效处理和生成，之后**通过视频压缩网络分解为时空patches，允许模型在时间和空间范围内进行信息交换和操作。**
- **从Sora技术报告来看，时空patches或借鉴谷歌ViViT操作。**
- ✓ ViViT借鉴ViT在图片分割上的思路，把输入的视频划分成若干个tuplet，每个tuplet会变成一个token，经过spatial temperal attention进行空间和时间建模获得有效的视频表征token。
- 传统方法可能将视频简单分解为一系列连续的帧，因而忽略了视频中的空间信息，也就是在每一帧中物体的位置和运动。**我们认为，由于连续帧存在时空连续性，Sora的时空patches可同时考虑视频中时间和空间关系，能够更加精准生成视频，捕捉到视频中细微的动作和变化，在保证视频内容连贯性和长度的同时，创造出丰富多样的视觉效果，灵活满足用户的各种需求。**

图表：ViViT将视频划分为若干tuplet



图表：ViViT可利用时空tuplet在时空联合建模

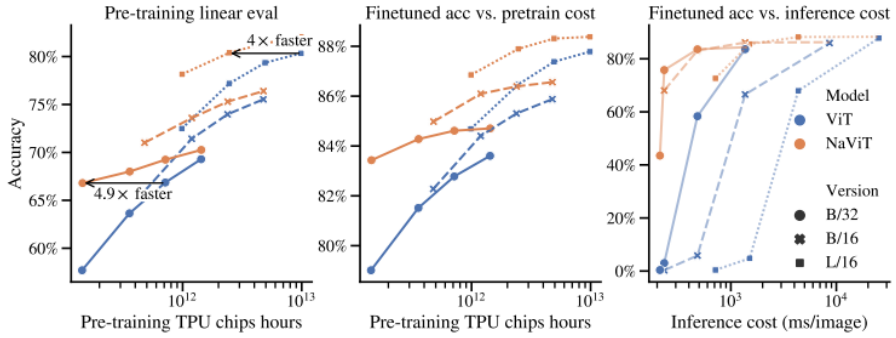


资料来源：Google Research《ViViT: A Video Vision Transformer》，华福证券研究所

2.2 Sora时空patches突破视频长宽比、分辨率等限制

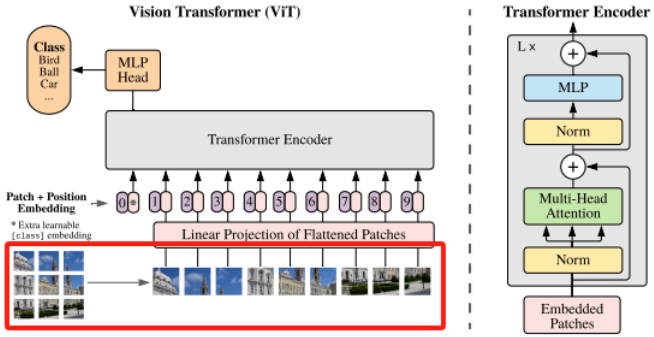
- OpenAI 表示，过去的图像和视频生成方法通常会将视频调整大小、裁剪或修剪为标准尺寸，而这损耗了视频生成的质量。例如，ViT通常需要将图像调整为固定的分辨率与尺寸进行处理，并仅能分解为固定数量的patches，因而限制了灵活处理不同尺寸、分辨率视频的建模。
- **Sora或借鉴谷歌NaViT中“Patch n’ Pack”的方法，在训练效率、模型适应性和推理灵活性等方面具有显著优势。**
- ✓ 1）允许从不同图像中提取多个patch打包在一个序列中，从而实现可变分辨率并保持宽高比。
- ✓ 2）NaViT相比ViT具有较高计算性能。例如，使用四倍少的计算量，NaViT到达顶级ViT的性能。此外，NaViT可以在训练和微调过程中处理多种分辨率的图像，从而在各种分辨率下都能表现出优秀的性能，在推理成本方面给NaViT带来了显著的优势。
- 我们认为，经过patch 化之后，Sora无需对数据进行裁剪，就能够对不同分辨率、持续时间和长宽比的视频和图像的原始数据进行训练，**既极大程度上利用原始信息保障生成高质量图片或视频，又显著提升模型性能，节约训练与推理算力成本。**

图表：NaViT相比ViT具有显著的计算性能



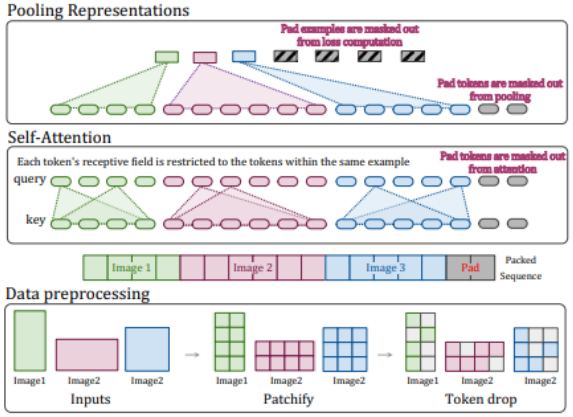
资料来源：Google DeepMind《Patch n’ Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution》，华福证券研究所

图表：ViT需调整图像为标准尺寸并分解为固定数量patches



资料来源：Google Research, Brain Team《AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE》，华福证券研究所

图表：NaViT突破分辨率、宽高比限制



资料来源：Google DeepMind《Patch n’ Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution》，华福证券研究所

2.2 Sora时空patches突破视频长宽比、分辨率等限制

华福证券 HUAFU SECURITIES CO.,LTD.

- **根据技术报告，Sora在原视频训练有以下优势：**

✓ **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格式。这使得Sora能够直接按照不同设备的原生宽高比创建内容。它还允许在使用同一模型生成全分辨率内容之前，快速原型化较小尺寸的内容。

✓ **改进的构图和画面组成：**将Sora与一个版本的模型进行了比较，该模型将所有训练视频裁剪成正方形。在正方形裁剪上训练的模型有时会生成主体只部分出现在视野中的视频。相比之下，来自Sora的视频具有改善的取景。

图表：Sora采样具有较高灵活性

图表：Sora改进的构图和画面组成（右图）

资料来源：OpenAI，华福证券研究所

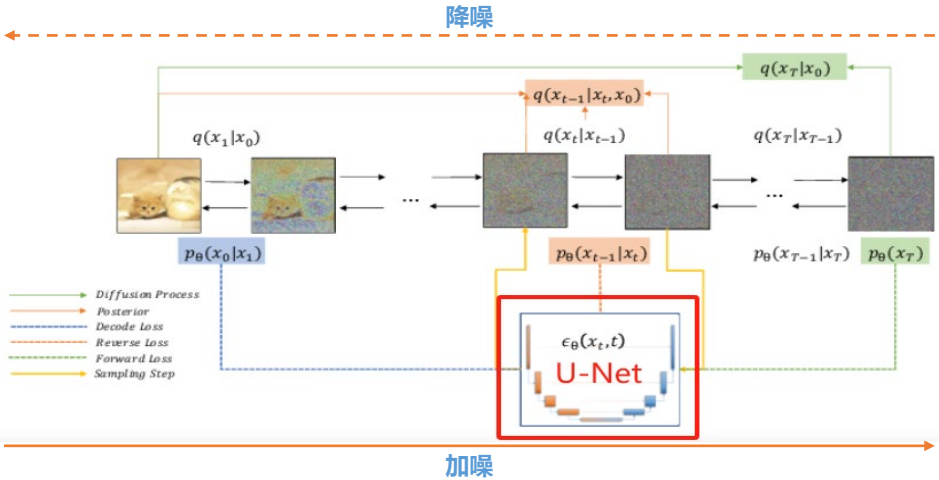
2.3 Transformer架构突显Scalling Law的“暴力美学”

- 扩散模型定义了扩散步骤的马尔科夫链，先通过向真实数据添加随机噪声，后反向学习扩散过程，从噪声中构建所需数据的样本，逐步降噪输出图片或视频。**其中，U-Net为扩散模型的重要架构之一，通过训练U-Net预测噪声，逐步去噪后输入结果。**
- **U-Net为卷积神经网络模型（CNN），在视频生成领域存在需裁剪数据与额外引入时间层等缺陷。**

✓ 1）卷积神经网络由于架构限制，存在分辨率与长宽比约束，输入与输出的结果均需调整至标准化大小，可能产生性能损失与效率低下等问题。

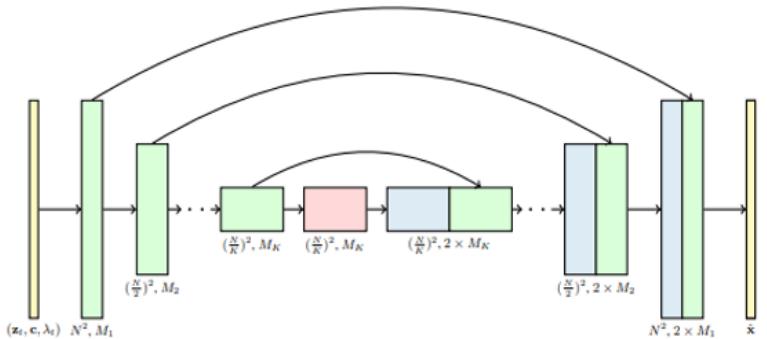
✓ 2）U-Net 的去噪模型在处理视频数据时，需额外加入一些和时间维度有关的操作，比如时间维度上的卷积、自注意力。在该过程涉及到时间注意力块嵌入位置问题，因而或较难处理长视频较多帧数的时间嵌入。

图表：基于U-Net架构的DDPM模型



资料来源：Semantic Scholar， Cao et al.《A Survey on Generative Diffusion Model》，华福证券研究所

图表：加入时间注意力块的3D U-Net



资料来源： Ho et al.《Video Diffusion Models》，华福证券研究所

2.3 Transformer架构突显Scalling Law的“暴力美学”

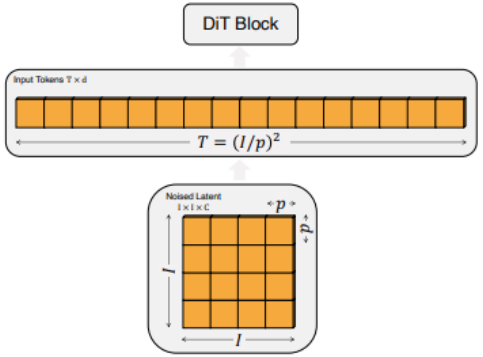
- **OpenAI 在 2020 年首次提出了模型训练的秘诀——Scaling Law。根据 Scaling Law，模型性能会在大算力、大参数、大数据的基础上像摩尔定律一样持续提升，不仅适用于语言模型，也适用于多模态模型。**
- **Sora替换U-Net为DiT的transformer作为模型架构，具有两大优势：**

✓ 1）transformer可将输入视频分解为3D patch，类似DiT将图片分解为图块，不仅突破了分辨率、尺寸等限制，而且能够同时处理时间和空间多维信息；

✓ 2）transformer延续了OpenAI的Scaling Law，具有较强的可拓展性，即参数规模越大、训练时长越长、训练数据集越大，生成视频的效果更好。例如，Sora随着训练次数的增加，小狗在雪地里的视频质量显著提升。

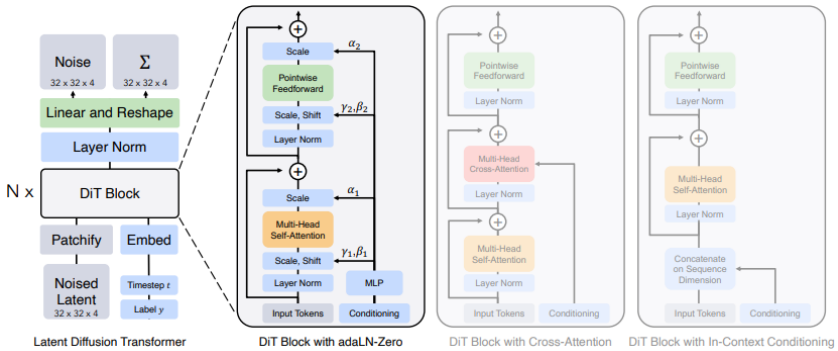
- U-Net为扩散模型主导架构，主要系Transformer 中全注意力机制的内存需求会随输入序列长度而二次方增长，高分辨率图像处理能力不足。在处理视频这样的高维信号时，这样的增长模式会让计算成本变得非常高。然而，**我们认为，OpenAI背靠微软云计算资源，具有较强的算力禀赋支持其再次打造“ChatGPT”时刻的Sora，此外通过视频网络空间降维技术可起到节约算力资源的作用，进一步促成Sora的成功与巩固OpenAI的龙头地位。**

图表：DiT中将图片分解为图片块



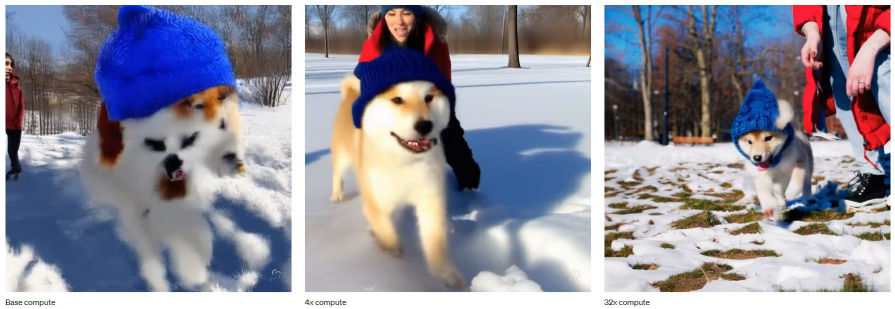
资料来源：Peebles & Xie《Scalable Diffusion Models with Transformers》，华福证券研究所

图表：采用transformer架构的DiT



资料来源：Peebles & Xie《Scalable Diffusion Models with Transformers》，华福证券研究所

图表：随着计算次数增加Sora生成的视频质量明显提升

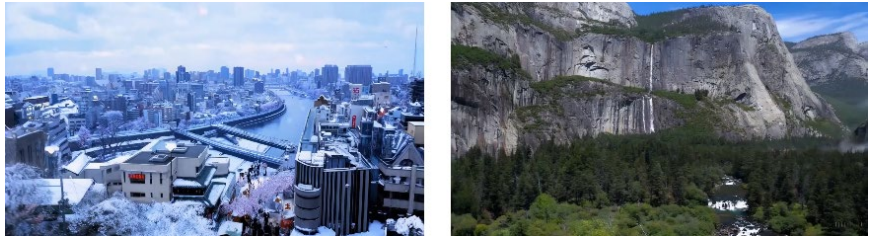


资料来源：OpenAI，华福证券研究所

2.3 Sora在Transformer大规模训练下涌现模拟能力

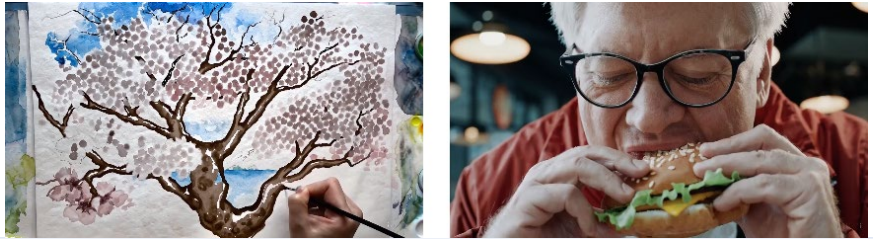
- **Sora在大规模训练的“暴力美学”下，未经过明确的3D、物体等归纳信息的训练，逐步显现出规模效应，迸发了模型的涌现能力：**
- ✓ **3D一致性：**Sora能够生成具有动态相机运动的视频。随着相机的移动和旋转，人物和场景元素在三维空间中保持一致地移动。
- ✓ **长距离连贯性和物体持久性：**Sora通常能够有效地建模短距离和长距离依赖关系。例如，即使在人、动物和物体被遮挡或离开画面时，也能持续保持它们的存在；在单个样本中生成同一角色的多个镜头，并在整个视频中保持其外观。
- ✓ **与世界互动：**Sora有时可以模拟一些简单的动作来影响世界的状态。例如，画家可以在画布上留下随时间持续存在的新笔触，或者一个人可以吃一个汉堡并留下咬痕。
- ✓ **模拟数字世界：**Sora可以在同时控制《我的世界》中的玩家采用基本策略的同时，还能以高保真度渲染世界及其动态。

图表：3D一致性

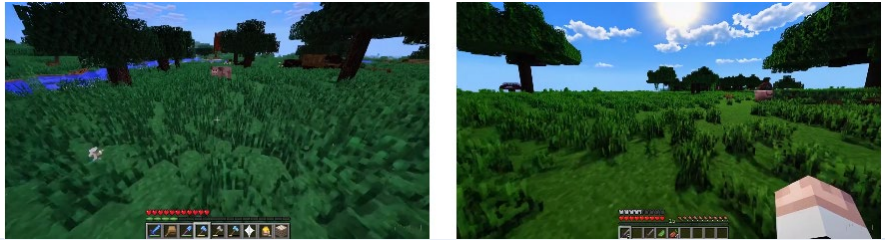


图表：长距离连贯性和物体持久性

图表：与世界互动



图表：模拟数字世界



资料来源：OpenAI，华福证券研究所

2.4 数据来源或更为丰富，视频重标注技术展示强大语言理解能力

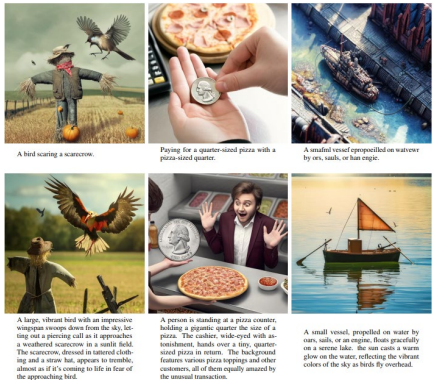
- **缺乏丰富的视频数据集以及如何对视频标注文本为文生视频的主要难点之一。**从流行的Gen-2、Emu Video等应用来看，这些模型通常先利用CLIP技术训练生成文本-图像对，之后加入时间层对视频进行标注，因而或许面临视频数据质量保证问题。
- **Sora训练数据集具有如下特点：**

✓ 数据来源或更为丰富。Sora技术报告未披露训练数据的详细情况，而我们认为从其涌现能力表现来看，Sora在训练数据中或许容纳了众多电影、纪录片、甚至游戏引擎等合成数据。

✓ 原生视频处理。不对视频/图片进行裁剪等预处理，从而保证Sora生成的灵活性。

✓ Sora建立在过去DALL · E3和GPT模型的研究基础之上，构建视频re-captioning，使得模型具有强大的语言理解能力。原始的文本可能并不能很好的描述视频，可以通过re-captioning的方式为视觉训练数据生成高度描述性的字幕。因此，该模型能够在生成的视频中更忠实地遵循用户的文字提示。

图表：DALL·E 3利用文本重新标注技术渲染更好结果



资料来源：Betker et al.《Improving Image Generation with Better Captions》，华福证券研究所

图表：主流文生视频模型数据集情况

公司	产品	推出时间	模型	架构	数据集	文本条件生成方法
Runway	Gen-2	2023.06	扩散模型	U-Net	2.4亿张图片 640万视频片段	CLIP
Meta	Emu Video	2023.11	扩散模型	U-Net	3400万视频-文本对	CLIP
Stability AI	Stable Video Diffusion	2023.11	扩散模型		6亿个样本数据集	CoCa、V-BLIP、LLM、CLIP

资料来源：澎湃、AI新智界，量子位，Runway《Structure and Content-Guided Video Synthesis with Diffusion Models》，GenAI, Meta《EMU VIDEO: Factorizing Text-to-Video Generation by Explicit Image Conditioning》Stability AI《Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets》，华福证券研究所

图表：Sora根据文本说明生成高质量视频

a toy robot wearing purple overalls and cowboy boots taking a pleasant stroll in Johannesburg, South Africa during a colorful festival

资料来源：OpenAI，华福证券研究所

目录

3 投资建议

- **我们认为，在视频压缩网络与时空patches提高计算效率与利用原生视频信息的基础上，transformer或取代U-Net成为扩散模型主流架构。可拓展性更强的transformer需要更为有力的算力支持才能保障视频生成质量，同时相比于大语言模型，视觉数据的训练与推理算力需求更大，因而算力有望成为确定性最高的受益赛道。此外，Sora发布有望形成多模态产业“鲇鱼效应”，激励其他多模态厂商的良性发展。**
- **建议关注：**
  - ✓ 1）AI算力：云赛智联、思特奇、恒为科技、海光信息、寒武纪、景嘉微、中科曙光、浪潮信息、拓维信息、四川长虹、工业富联、神州数码等。
  - ✓ 2）AI+多模态：万兴科技、虹软科技、当虹科技、中科创达、大华股份、海康威视、漫步者、萤石网络、汉仪股份、美图公司、云从科技

目录

4 风险提示

- **产品落地不及预期。**
  - ✓ 垂直领域产品推出速度缓慢，商业化进行较慢。
- **技术迭代不及预期。**
  - ✓ AI多模态技术发展未能取得新的突破。
- **AI伦理风险。**
  - ✓ AI技术滥用导致的数据安全、隐私安全等问题。

分析师声明及一般声明

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

一般声明

华福证券有限责任公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

特别声明及投资声明评级

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	评级	评级说明
公司评级	买入	未来6个月内，个股相对市场基准指数涨幅在20%以上
	持有	未来6个月内，个股相对市场基准指数涨幅介于10%与20%之间
	中性	未来6个月内，个股相对市场基准指数涨幅介于-10%与10%之间
	回避	未来6个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来6个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来6个月内，行业整体回报高于市场基准指数5%以上
	跟随大市	未来6个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来6个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的6~12个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中，A股市场以沪深300指数为基准；香港市场以恒生指数为基准；美股市场以标普500指数或纳斯达克综合指数为基准（另有说明的除外）。

华福证券
HUAFU SECURITIES CO.,LTD.

诚信专业 发现价值

联系方式

华福证券研究所 上海
公司地址：上海市浦东新区浦明路1436号陆家嘴滨江中心MT座20楼
邮编：200120
邮箱：hfyjs@hfzq.com.cn