# OpenAI 推出首个文生视频大模型 Sora，引领 AI 文生视频行业跨越式发展

## ——计算机行业跟踪报告

**强于大市**(维持)

2024 年 02 月 18 日

**行业核心观点：**

**文生视频大模型 Sora 重磅发布，可生成长达 1 分钟的视频。**2 月 16 日，OpenAI 推出其首个文生视频大模型 Sora。根据官网介绍，Sora 可以生成长达 1 分钟时长的视频，同时还能保证视频质量，并遵循用户的提示（prompt）。

**投资要点：**

**Sora 是一个扩散 transformer，具有强大的语言理解能力，通过在潜在空间训练 patches 生成视频。**对标 tokens，OpenAI 将视觉数据转换为 patches，有效用于 Sora 大模型训练。Sora 是一种扩散模型，通过给出输入的静态噪声以及相关的文本提示（prompt）等调节信息，训练生成原始的“干净”patches。在推理时，OpenAI 还可以通过在适当大小的网格中排列随机初始化的 patches 来控制生成视频的大小。与 GPT 模型类似，Sora 使用 transformer 架构，释放出卓越的扩展性能。立足 DALL·E 3 和 GPT 模型，Sora 具有强大的语言理解能力，能够生成更加准确遵循用户提示的高质量视频。此外，在固定种子和输入的情况下，可以看到训练计算的增加能显著提升样本视频的质量。

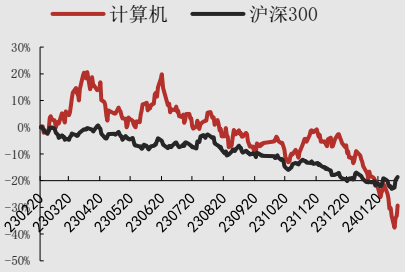
**多维度跨越式突破，视频质量飞跃性提升。**Sora 的采样更具有灵活性，同时改进了框架和构图。Sora 可以采样宽屏 1920x1080p 的视频、垂直 1080x1920 的视频以及介于两者之间的所有视频。这让 Sora 可直接以不同的原始长宽比创建内容。OpenAI 还通过经验发现，在视频的原始长宽比上进行训练可以改善构图和框架。Sora 还支持图生视频、视频生视频，能执行广泛的图像和视频编辑任务，创建完美的循环视频、动画静态图像、向前或向后扩展视频等。在连接视频上，Sora 能将两个输入视频无缝衔接在一起。虽然目前 Sora 仍然有一些缺陷和局限性，但已经开始理解物理意义，并出现许多有趣的涌现能力，如三维一致性。

**重塑 AI 文生视频行业格局，或冲击 AI 文生图赛道。**Sora 在生成视频长度上大幅领先，多角度镜头能力也显著领先行业竞品。同样的 prompt，Sora 生成的视频长度、质量都显著领先。Sora 可以生成可变大小的图像，最高可达 2048×2048 分辨率，图片画质有了大幅提升。我们认为随着文生视频画质能力的提升，图片作为单帧的视频，文生视频领域的产品或将冲击文生图行业。

**投资建议：**1）AI 文生视频行业发展带动 AI 行业应用落地的机遇；2）AI 行业发展对算力、光模块等基础设施的持续需求；3）AIGC 在媒体、游戏等行业的加速落地带来的投资机遇。

**风险提示：**AI 产业发展不及预期；AI 带来的版权、隐私及技术风险；国内 AI 应用落地不及预期；中美科技摩擦风险。

**行业相对沪深 300 指数表现**




**数据来源：**聚源，万联证券研究所

**相关研究**

Q4 基金重仓略微超配，前十大重仓股组成不变

人工智能行业应用多点开花

工信部就《国家人工智能产业综合标准化体系建设指南》公开征求意见

**分析师：** 夏清莹
执业证书编号：S0270520050001
电话：075583223620
邮箱：xiaqy1@wlzq.com.cn

# 正文目录

# 1 OpenAI 发布 Sora，AI 文生视频大模型跨越性突破

## 1.1 OpenAI 首个文生视频大模型 SORA 重磅推出

**文生视频大模型Sora重磅发布，可生成长达1分钟的视频。**2月16日，OpenAI推出其首个文生视频大模型Sora。根据官网介绍，Sora可以生成长达1分钟时长的视频，同时还能保证视频质量，并遵循用户的提示（prompt）。

图表1：Sora 一分钟展示视频的 prompt 及部分截图

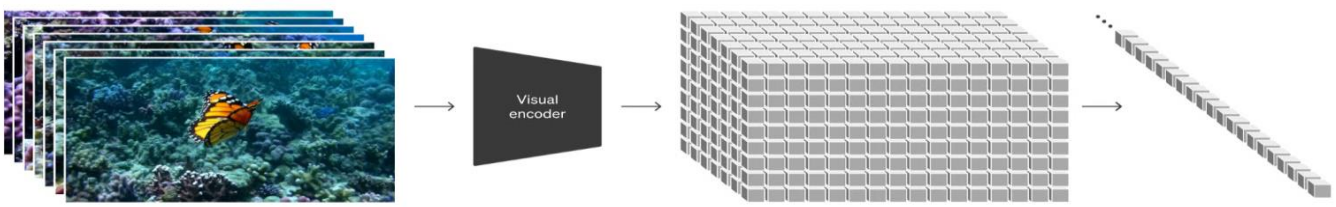
| Prompt（提示） | 视频部分截图 |
| --- | --- |
| A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.<br><br>翻译：一位时尚的女人走在东京的街道上，街道上到处都是温暖的发光霓虹灯和动画城市标志。她身穿黑色皮夹克，红色长裙，黑色靴子，背着一个黑色钱包。她戴着墨镜，涂着红色口红。她自信而随意地走路。街道潮湿而反光，营造出五颜六色的灯光的镜面效果。许多行人四处走动。 |  <br>  <br>  <br>  |

*资料来源：OpenAI，万联证券研究所*

*注：翻译内容来自Microsoft Edge网页自带翻译。*

**将视觉数据转换为patches，有效用于Sora大模型训练。**LLM范式的成功部分受益于使用tokens，tokens能够将文本的多种模态（代码、数学、各种自然语言）统一起来。OpenAI基于LLMs使用文本tokens的灵感，将所有视觉数据转化为patches，在Sora中实现类似的效果。根据OpenAI的介绍，patches此前就已经被证明是视觉数据模型的有效表示，同时OpenAI还发现，patches在训练生成不同类型视频和图像模型中是一种高度可扩展且有效的表示。

图表2：Sora 将视觉数据转换为 patches 的示意图




*资料来源：OpenAI，万联证券研究所*

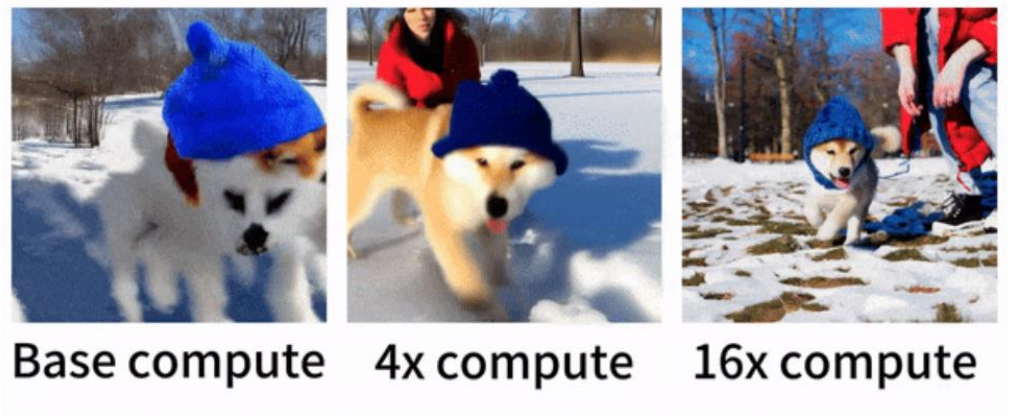
**Sora是一个扩散transformer（diffusion transformer），通过在潜在空间训练patches生成视频。**具体来看视频生成的过程，**1）首先将视频压缩到低维的潜在空间：**OpenAI训练了一个降低视觉数据维度的网络，通过这个网络原始视频会在时间和空间上都被压缩，并输出为潜在表示；**2）用时空潜在patches训练Sora：**Sora在这个压缩后的潜在空间中接受训练，基于从原始视频中提取的时空潜在patches，OpenAI能够使得Sora对不同分辨率、持续时间和长宽比的视频和图像进行训练（图像相当于单帧视频）；**3）解码生成新视频：**OpenAI训练了对应的解码器模型，将Sora在潜在空间训练生成的视频（潜在表示）映射回像素空间；在推理时，OpenAI还可以通过在适当大小的网格中排列随机初始化的patches来控制生成视频的大小。**Sora是一种扩散模型，**通过给出输入的静态噪声以及相关的文本提示（prompt）等调节信息，训练生成原始的“干净”patches。与GPT模型类似，Sora使用transformer架构，释放出卓越的扩展性能。

图表3：Sora通过扩散还原视频的示意图

*资料来源：OpenAI，万联证券研究所*

**训练计算的增加可以显著提升视频质量。**在固定种子和输入的情况下，可以看到训练计算的增加能显著提升样本视频的质量。

图表4：不同训练计算生成的样本视频对比




*资料来源：OpenAI，万联证券研究所*

**立足DALL·E 3和GPT模型，Sora具有强大的语言理解能力，能够生成更准确、更高质量的视频。**OpenAI将DALL·E 3的re-captioning技术应用在Sora中，先训练一个高度描述性的字幕生成器模型，然后使用它为训练集中所有视频生成文本字幕，通过对高度描述性视频字幕进行训练，能够提供文本的保真度以及视频的整体质量。此外，OpenAI还利用GPT将简短的用户prompt转化为较长的详细字幕，然后发送到视频模型中，使得Sora能够生成更加准确遵循用户提示的高质量视频。

### 1.2 多维度跨越式突破，视频质量飞跃性提升

**Sora的采样更具有灵活性，同时改进了框架和构图。**过去的图像和视频生成方法通常需要调整大小、进行裁剪或者是将视频剪切到标准尺寸，例如4秒的视频分辨率为256x256。**而OpenAI的研究发现在原始大小的数据上进行训练，采样更具灵活性、同时可以提高视频质量。**Sora可以采样宽屏1920x1080p的视频、垂直1080x1920的视频以及介于两者之间的所有视频。这让Sora可直接以不同的原始长宽比创建内容。Sora还支持在生成全分辨率的内容之前，以较小的尺寸快速创建内容原型——所有内容都使用相同的模型。OpenAI还通过经验发现，在视频的原始长宽比上进行训练可以改善构图和框架。研究团队将Sora与其他模型的一个版本进行比较，该版本将所有训练视频裁剪为方形。在正方形裁剪上训练的模型有时会生成仅部分显示主题的视频（左图），相比之下，来自Sora的视频有改进的帧内容（右图）。

图表5：使用正方形裁剪（左）与使用原始大小（右）的训练视频效果对比

*资料来源：OpenAI，万联证券研究所*

**Sora还支持图生视频、视频生视频，能够向前/向后扩展视频。**Sora能基于DALL·E 3的图像生成视频，还能执行广泛的图像和视频编辑任务，创建完美的循环视频、动画静态图像、向前或向后扩展视频等。以下是Sora从一段生成的视频向后拓展出的三个新视频，可以看到新视频的开头各不相同，但拥有相同的结尾。

图表6：向后扩展视频示意


*资料来源：OpenAI，万联证券研究所*

**在连接视频上，Sora能将两个输入视频无缝衔接在一起。**如下图所示，左上图是村庄，右下图是海洋，看似毫不相关的场景，Sora通过逐渐放大河流，合理、无缝地从左上→左下→右上→右下实现两个视频的转化连接。

图表7：从左上图逐渐转化至右下图的场景示意

*资料来源：OpenAI，万联证券研究所*

**虽然目前Sora仍然有一些缺陷和局限性，但已经开始理解物理世界，并出现许多有趣的涌现能力。**Sora目前还存在许多局限性，例如不能准确模拟许多基本交互的物理现象，如玻璃碎裂；如吃食物，并不总能产生正确的物体状态变化。但我们认为Sora已经接触到了世界模型的范畴。Sora能够生成具有多个角色、特定类型的运动以及主题和背景的准确细节的复杂场景。该模型不仅了解用户在提示中要求的内容，还了解这些东西在物理世界中的存在方式。OpenAI发现，视频模型在经过大规模训练后，会表现出许多有趣的新能力。这些能力使Sora能够模拟物理世界中的人、动物和环境的某些方面。这些特性的出现没有任何明确的三维、物体等归纳偏差，它们纯粹是规模现象。例如三维一致性：Sora可以生成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的移动是一致的。

图表8：Sora 三维一致性示意图

*资料来源：OpenAI，万联证券研究所*