【降低噪声】普通人也能一文读懂Sora系列

> 原文档：【降低噪声】普通人也能一文读懂Sora系列

> 💡 **Sora 是能够理解和模拟现实世界的模型的基础，我们相信这一功能将成为实现 AGI 的重要里程碑。——Open AI**

> 注意！注意！Sora很火，但是还未正式对大众开放，但是目前市面上已有Sora的实操课程，请擦亮眼睛，谨慎辨别！

大家好，我是乌鸦哥，一个大厂打工人。
2022年7月开始研究AI绘画，应用到公司课程设计，极大地提升了效率；
2023年，ChatGPT出来后，深入研究，并且将GPT、AI音乐、AI视频等应用到公司的产品和获客;

2023年8月我用AI做视频第1个抖音号，第1个视频涨粉1100+，成功裂变4个账号，2周不到涨粉1500+，多条视频播放25万+。

这些经验让我觉得AI将会带给我们无限的可能，特别是AI视频+自媒体，更是普通人新的机遇。2024年将是AI视频之年，Sora的出现也印证了这个结论，让我们一起深入学习和探索吧！

什么是Sora?

Sora是OpenAI开发的一款革命性的文本到视频生成器，能够根据简单的文字提示创造出最长60秒的视频内容。想象一下，你在沙发上闲逛，忽然脑海中闪过一个电影的点子。你拿起手机，输入一段描述，比如“一个穿着红色羊毛摩托车头盔的30岁太空人，在蓝天和盐沼之上的冒险，电影预告片风格，35mm胶片拍摄，色彩鲜艳。”几分钟后，屏幕上就展现了一段高达一分钟的详细视频。这就是Sora的魔法。

描述：一个穿着红色羊毛摩托车头盔的30岁太空人，在蓝天和盐沼之上的冒险，电影预告片风格，35mm胶片拍摄，色彩鲜艳。
视频如下：

Sora更多视频效果，参考https://openai.com/sora和https://www.tiktok.com/@openai

Sora都有哪些功能?

1. **文字生成视频，时长可到60s**

Prompt: 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.
提示：一位时尚的女士在充满温暖发光的霓虹灯和活泼的城市标志的东京街头漫步。她穿着一件黑色皮夹克，一条长红裙和黑色靴子，手提一只黑色手袋。她戴着太阳镜，涂着红色口红。她走路自信而随意。街道潮湿且具有反射性，创造出色彩斑斓灯光的镜面效果。人来人往。
视频：

2. **图片生成视频**

3. **扩展视频**

Sora 还能够在时间上向前或向后扩展视频。下面是三个视频，它们都是从生成的视频片段开始向后延伸的。因此，这三个视频的开头都不同，但三个视频的结局都是相同的。


San Fidio
San Fidio

我们可以使用此方法向前和向后扩展视频以产生无缝的无限循环。

4. **通过文本编辑视频**

扩散模型使得从文本提示编辑图像和视频的方法层出不穷。下面我们将这些方法之一，SDEdit，应用到Sora上。这种技术使能够零次射击地转换输入视频的风格和环境。

> "零次射击"（Zero-shot）在机器学习和人工智能领域是一个术语，指的是模型在未见过任何特定任务的训练数据的情况下，仍能对该任务进行推断或执行的能力。这种方法依赖于模型对知识的泛化能

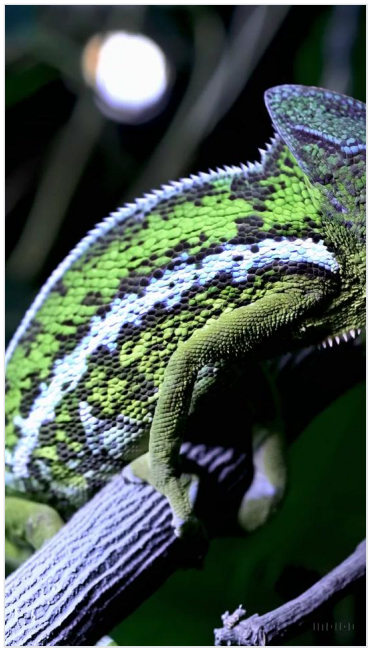
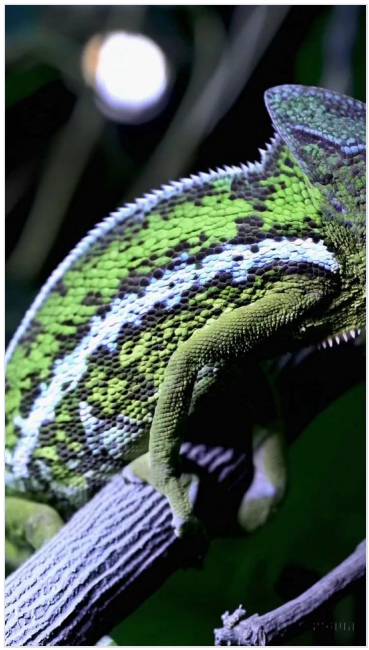
> 力，即使用在其他任务上学到的知识来处理新的、未见过的任务。在上下文中，当提到Sora能够"零次射击"地转换输入视频的风格和环境时，意味着它可以在没有专门为这种风格转换训练过的情况下，仍然能够实现视频风格和环境的转换。这显示了模型在理解和执行与输入相关的新任务方面的强大能力。

Input video（原视频）

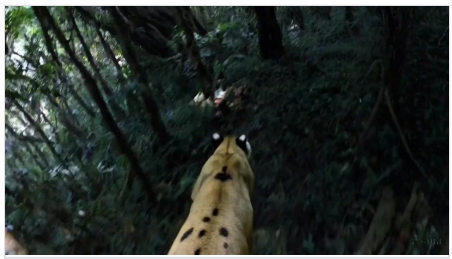
make it go underwater（编辑后视频）

5. **连接视频**

我们还可以使用Sora逐渐在两个输入视频之间进行插值，创建出在完全不同的主题和场景构成之间的无缝过渡。在下面的例子中，中间的视频是在左右两边对应视频之间进行插值最终生成的。






6. 生成图像

Sora也可以生成不同大小的图像——分辨率最高可达2048x2048。

Close-up portrait shot of a woman in autumn, extreme detail, shallow depth of field

https://v88cxopssb.feishu.cn/sync/IF3ZdRWZ0sDJdxb8U7LcYDONnxd

Vibrant coral reef teeming with colorful fish and sea creatures

A snowy mountain village with cozy cabins and a northern lights display, high detail and photorealistic dslr, 50mm f/1.2

7. **新兴的模拟能力**

Sora在大规模训练时，展现出了许多有趣的新兴能力。这些能力使得Sora能够模拟现实世界中人类、动物和环境的某些方面。这些属性的出现，并不需要对三维、对象等进行任何明确的归纳偏见——它们完全是规模效应的现象。

- 3D一致性。Sora能够生成带有动态相机移动的视频。随着相机的移动和旋转，人物和场景元素在三维空间中一致地移动。

- 长期一致性和对象恒定性。对于视频生成系统来说，一个重大挑战一直是在采样长视频时保持时间上的一致性。我们发现，Sora通常（虽然不总是）能够有效地模拟短期和长期依赖性。例如，我们的模型即使在人物、动物和物体被遮挡或离开画面时，也能持续地表示它们。同样，它可以在单个样本中生成同一角色的多个镜头，贯穿整个视频保持他们的外观。

- 与世界互动。

这一功能体现了Sora模型在模拟现实世界动作和其对环境状态影响方面的先进能力。具体来说，这意味着Sora不仅能够生成视觉上吸引人的视频内容，还能在这些内容中模拟出动作和后续效果的逻辑连贯性。

例如，当提到“一个画家在画布上留下新的笔触”这一场景时，Sora能够生成展示画家绘画过程的视频，而且随着时间的推移，这些笔触会持续存在于画布上，就像在现实世界中一样。同样，当模拟“一个人吃汉堡并留下咬痕”的场景时，Sora能够生成不仅展示吃汉堡动作的视频，还能展示汉堡因被咬而状态改变的效果。

这一功能的重要性在于，它展示了Sora在理解和模拟物理世界及其因果关系方面的先进能力。通过模拟这些看似简单的互动，Sora展示了其对细节的注意力以及生成内容的真实性。这不仅对于创造具有沉浸感的视觉内容至关重要，也对于使用视频模拟来探索和解释现实世界事件的能力提出了有意义的展望。

总的来说，"Interacting with the world" 功能是Sora模型在模拟现实世界动作及其影响方面的一个突破，它为未来的内容创造、教育和娱乐等领域开启了新的可能性。

- 模拟数字世界

Sora展示了其在模拟数字世界，特别是视频游戏环境方面的先进能力。具体来说，这意味着Sora能够在高保真度下渲染游戏世界及其动态，并且能够根据基本策略控制游戏中的玩家行为。
以"Minecraft"（我的世界）为例，Sora不仅能够渲染出游戏世界的细节，还能模拟玩家在游戏中的动作和互动，如行走、挖掘或建造。

这一功能的实现，表明Sora不仅能够理解和生成静态图像和动态视频，还能够理解复杂的数字环境和规则，进而在没有任何特定训练样本的情况下（即零样本学习），通过简单的文字提示来生成与特定视频游戏相关的动态场景。这种能力对于AI技术的应用领域意味着巨大的拓展，尤其是在游戏设计、虚拟现实以及数字教育等方面。

通过模拟"Minecraft"等游戏，Sora展现了AI在理解游戏机制和玩家互动方面的潜力，同时也揭示了AI技术在创造复杂、互动性强的数字环境方面的进步。这不仅为游戏开发者提供了新的工具来探索游戏设计的新可能性，也为教育和训练模拟提供了新的视角，使得创建更为丰富和互动的虚拟学习环境成为可能。

总结：Sora的能力展示了继续扩大视频模型规模是开发高能力物理和数字世界模拟器的有希望的途径，这些模拟器能够复现其中的物体、动物和人类。这表明，通过进一步发展和完善这些模型，我们可以期待在未来实现更加精细和真实的世界模拟，这些模拟不仅能够帮助我们更好地理解和预测现实世界和数字环境中的各种现象，还能够在教育、娱乐、设计和科学研究等多个领域找到应用。

例如，在科学研究中，这种模拟能力可以帮助研究人员创建复杂的模拟环境来测试假设和理论；在教育领域，可以创建互动的虚拟学习环境，让学生以更加直观和实践的方式学习复杂概念；在娱乐领域，可以创造更加丰富和沉浸的虚拟世界和游戏体验。

总之，Sora及类似的视频模型的持续发展，预示着我们在构建能够模拟世界各种方面的高度能力AI系统方面迈出了重要一步。随着技术的进步，我们有理由期待这些模拟器将在提高我们对世界理解、增强人类学习和创造力以及推动科技创新方面发挥重要作用。

Soar的技术原理?

1. **技术版本：**

Sora是一个由OpenAI开发的视频生成模型，它利用先进的人工智能技术，特别是深度学习中的扩散模型和变换器架构，来理解和生成视频内容。这里简单总结一下它的技术原理：

将视觉数据转换成补丁：Sora首先将视频压缩成低维度的潜在空间表示，然后将这些表示分解成时空补丁。这种方法借鉴了大型语言模型处理文本的方式，但Sora应用在视觉数据上。

视频压缩网络：Sora通过一个专门的网络来降低视觉数据的维度，这个网络将原始视频转换为时间和空间上都被压缩的潜在表示。Sora在这个潜在空间上进行训练，并生成新的视频内容。

时空潜在补丁：在得到压缩的视频数据后，Sora提取一系列的时空补丁，这些补丁作为变换器模型的输入令牌。这使得Sora能够处理不同分辨率、持续时间和纵横比的视频和图像。

扩展变换器用于视频生成：Sora作为一个扩散模型，通过预测输入噪声补丁的原始“干净”版本进行训练。它结合了变换器架构，这种架构在处理语言、计算机视觉和图像生成方面已经显示出了显著的扩展性。

通过这些技术原理，Sora能够根据文本提示生成高质量的视频，支持多种分辨率和纵横比，并且能够捕捉视频内容的复杂动态。这种能力使得Sora不仅在内容创造方面有巨大的潜力，也在模拟真实世界和数字世界方面展现出前所未有的能力。

2. **大白话版本：**

想象一下，如果我们能把看到的一切——不管是电影中的精彩场景，还是手机里的随手拍——都变成一个个小方块（就像乐高积木一样），然后通过这些小方块来创造全新的视频。这就是OpenAI的Sora做的事情。它把视频和图片变成了一堆堆的小方块，然后用这些小方块来学习和创造新的视频内容。

接下来，Sora会用一个特别的网络把这些视频压缩，让它们变得更简单，就像是把一本厚重的小说浓缩成精彩的摘要。这样做不仅节省了存储空间，也让生成新视频变得更加快速和高效。

然后，Sora利用这些被压缩的小片段（我们叫它们“补丁”），就像乐高积木一样，来构建和创造全新的视频内容。它可以根据给定的文字提示，比如“夕阳下的海滩”或者“一个人在下雨的街道上跑步”，来生成符合描述的视频场景。

Sora之所以能做到这一点，是因为它使用了一种非常聪明的技术，叫做变换器，这让它能够理解和预测这些小片段应该如何组合在一起，就像是预测下一句歌词或故事情节一样。通过这种方式，Sora不仅可以生成各种大小和形状的视频，还能确保视频的每个部分都能自然衔接，看起来既真实又自然。

总的来说，Sora就像是一个超级聪明的导演，它可以根据你的想法，用视频的“乐高积木”来创造出你想要的任何场景。这种能力让Sora在创造视频内容、模拟真实或虚构世界方面都有着巨大的潜力。

Sora技术报告原文：https://openai.com/research/video-generation-models-as-world-simulators

Sora相对于Pika、runway等文本生成视频AI，优势何在？

跟其他的一些做类似事情的工具比如Runway和Pika相比，Sora就像是那个学霸，无论在哪方面都能做得更好。

1. **清晰度和视频长度：**Sora能够生成高达一分钟的高清视频，这在技术上是一个重大突破。相比之下，其他平台可能在视频长度和清晰度上有限制，时长大多为4s。
2. **灵活的视频参数：**Sora支持生成不同分辨率、纵横比的视频，从竖屏到宽屏都能轻松应对。这种灵活性对于适应不同的展示平台和内容需求至关重要。
3. **人物大幅度运动的模拟：**Sora能够生成显示人物进行复杂、动态运动的视频，如跑步、跳跃或舞蹈，这些运动看起来自然流畅，不会显得生硬或不自然。
4. **场景细节的精细渲染：**Sora在生成视频时能够捕捉和再现丰富的场景细节，无论是自然景观的微妙变化，还是城市环境中的复杂结构，都能以高保真度呈现。
5. **内容一致性的保持：**在生成长视频时，保持场景和角色的一致性是一个挑战。Sora通过高级的算法确保视频内容从开始到结束都保持逻辑上的连贯性和视觉上的一致性。
6. **多角色交互的处理：**Sora能够生成包含多个角色在内的场景，这些角色之间的互动看起来自然和有信服力。这对于创造复杂的社交场景或动作场面尤为重要。
7. **强大的语言理解能力：**Sora采用了类似GPT的技术，自然语言理解极为强大。它甚至可以根据简短的提示，通过生成详细的描述来提高视频内容的相关性和准确性。
8. **模拟现实世界动作的能力：**Sora不仅可以生成静态场景的视频，还能模拟动作对环境的影响，如人物吃东西留下咬痕，这种对细节的捕捉在提高视频真实感方面非常关键。
9. **创造数字世界的能力：**Sora可以模拟视频游戏等数字环境，控制游戏角色并以高保真度渲染游戏世界，这显示了它在理解和生成复杂数字世界方面的强大能力。
10. **大规模训练的优势：**Sora通过大规模训练，能够理解和处理各种视觉数据，这使得它在理解复杂场景和生成高质量视频内容方面具有优势。
11. **内容生成的创新性：**Sora的技术架构允许它在生成视频时展现出创新性，如通过模拟细节增强场景真实感，或是通过创造性地理解文本提示来生成前所未有的内容。

Sora局限性何在？

虽然Sora在视频生成方面展现了许多优势，但它作为一个模拟器仍然存在一些局限性。根据OpenAI提供的信息，这些局限性主要表现在以下几个方面：

1. **物理互动模拟不精确：**Sora在模拟一些基本的物理互动，比如玻璃破碎的过程时，并不总是能够准确捕捉到这些现象的物理本质。这可能是因为模型在理解和生成这类物理效应的复杂性上还有待提高。
2. **对象状态变化不一致：**在模拟吃东西等互动时，Sora不总是能够正确地反映出对象状态的改变，比如食物被吃掉后应有的变化。这表明模型在跟踪和表达对象状态变化方面还存在挑战。

3. **长视频的不连贯性：**在生成较长的视频样本时，Sora有时会产生不连贯的问题，比如突然出现或消失的对象，这影响了视频内容的逻辑性和连贯性。

这些局限性说明，尽管Sora在视频内容生成方面取得了显著进展，但在完全模拟现实世界的物理互动、对象状态变化以及保持长时间视频内容连贯性方面，仍然面临挑战。这些问题的存在可能限制了Sora在某些特定应用场景下的效用，特别是那些对物理真实性和细节连贯性要求较高的场合。

为了克服这些局限性，未来的研究和开发工作可能需要进一步提高模型的物理模拟能力，改善对象状态跟踪机制，并增强长视频生成的逻辑性和连贯性。通过持续的技术迭代和优化，有望逐步解决这些问题，使Sora和类似的AI视频生成模型能够更加精确和自然地模拟现实世界和其互动。

Sora有何影响?

Sora，作为OpenAI开发的先进文本到视频生成AI，对社会、行业和个人创作带来了深远的影响：

1. **媒体和娱乐行业革新：**Sora能够快速生成高质量的视频内容，为电影、电视、广告和游戏行业提供了新的创作工具。这可能降低制作成本，加速创作过程，并为小型创作团队提供与大型制作公司竞争的机会。想象一下，一个小型独立电影工作室想要拍摄一部科幻电影，但预算有限，无法承担昂贵的特效和场景设计费用。Sora可以让他们通过简单的文本描述生成复杂的外星景观或是惊险的太空追逐场景，大大降低制作成本，同时保持影片的视觉冲击力。
2. **教育和培训的变革：**通过生成具体场景的视频，Sora可以为在线教育提供更加丰富和互动的学习材料。这不仅能够提高学习的吸引力和效率，也为远程教育和自学者开辟了新的可能性。比如历史老师想要让学生更直观地了解罗马帝国的生活。通过Sora，老师可以创建详细的、栩栩如生的罗马市场或角斗场的视频，让学生们仿佛穿越回古罗马，增强学习体验和历史的沉浸感。
3. **个人低门槛创作：**Sora降低了高质量视频制作的门槛，使得没有专业视频制作背景的个人也能够创造出吸引人的视觉内容。这促进了内容创作的民主化，激发了更多人的创造潜能。

4. **游戏开发：**对于独立游戏开发者来说，创建一个引人入胜的游戏世界是挑战之一。Sora可以帮助他们通过描述来生成游戏内的环境，比如“神秘的森林在月光下闪耀着银色的光芒”，从而加快游戏开发过程，提升游戏的视觉质量和创意表现。
5. **广告行业：**广告公司需要快速响应市场，为不同的产品和服务制作吸引人的广告。Sora能够根据创意团队的文本提示快速生成展示产品特点的视频，无论是展示一款运动鞋在各种极限环境下的性能，还是创造一个虚拟的幻想世界来宣传最新的视频游戏。
6. **新闻和报道的影响：**Sora提供了一种新的方式来制作新闻报道和纪录片中的视觉内容，特别是在缺乏现场拍摄资源的情况下。然而，这也需要媒体机构采取措施确保内容的准确性和透明度。
7. **社交媒体内容创作：**对于那些热衷于社交媒体的用户，Sora可以帮助他们创作出独特和引人注目的内容。无论是想分享生活点滴，还是传达某种观点或情感，Sora都能提供强有力的视觉支持，让内容更加丰富多彩，增强与观众的互动和连接。
8. **个性化和定制化服务的提升：**随着技术的进步，Sora可能会提供更加个性化和定制化的视频生成服务，满足用户特定的需求和偏好，从而在广告、品牌营销等领域发挥重要作用。
9. **社会和伦理挑战：**随着Sora等技术的发展，也带来了关于内容真实性、版权、以及滥用风险的社会和伦理问题。如何确保技术的负责任使用，防止生成误导性或有害内容，成为了行业、监管机构和社会需要共同面对的挑战。

Sora是世界模拟器?

Sora被称为世界模拟器的原因在于其能力不仅限于简单地根据文本生成视频，而是可以通过视频模型来模拟、理解和重现物理世界和数字世界中的复杂互动和动态变化。这种模拟能力意味着Sora不仅可以创造出视觉内容，还能在一定程度上捕捉和再现现实世界或虚拟世界中的行为、物理规律和社会互动。

模拟物理世界

Sora能够生成显示人物、动物和物体在现实世界中交互的视频，这些交互遵循物理规律，如重力、碰撞和运动。例如，Sora可以生成一个人跳跃或物体掉落的视频，这些视频反映了真实世界的物理特性。

模拟数字世界

Sora同样能够模拟视频游戏或其他数字环境中的场景和互动。这不仅包括视觉呈现，也包括游戏逻辑和角色行为。通过理解和生成这类内容，Sora展现了对复杂数字环境规则的理解能力。

模拟社会和文化互动

通过生成具有复杂社会和文化背景的视频，Sora展示了对人类社会互动和文化表现的模拟能力。这不仅对于娱乐和教育领域有重要意义，也对于社会科学研究提供了新的视角和工具。

未来发展

随着技术的进步，Sora作为“世界模拟器”的能力有望进一步增强，包括更准确地模拟复杂物理现象、更深入地理解和表现人类情感和社会文化互动，以及更广泛地应用于教育、设计、规划和娱乐等

领域。这样的进步将进一步强化Sora在模拟世界各个方面的能力，使其成为研究、教育和创作的强大工具。

总之，Sora之所以被称为视频生成模型中的“世界模拟器”，是因为它的技术和应用展示了对现实世界和数字世界的深入理解和高度模拟能力，这不仅是技术上的创新，也为各行各业提供了新的视角和工具。

Sora将成为实现 AGI 的重要里程碑?

将Sora视为实现通用人工智能（AGI）重要里程碑的理由在于其独特的能力——理解和模拟现实世界。这一功能的重要性体现在以下几个方面：

1. **复杂环境理解：**能够理解和模拟现实世界意味着Sora具备分析和处理复杂环境数据的能力。这种能力是通用人工智能的核心，因为它要求机器不仅仅能执行单一任务，还能理解和适应多变的环境和条件。
2. **跨领域学习与适应：**Sora展示了从文本描述到视频内容生成的跨媒介能力，这种能力体现了模型的多模态理解——即能够整合不同类型信息（如文本、图像和视频）的能力。通用人工智能需要这种跨领域的学习和适应能力，以在不同的环境和任务中应用其知识和技能。
3. **抽象思维与创造力：**通过根据文本提示创造视觉内容，Sora展现了一定程度的抽象思维和创造力。这表明了它不仅能理解文本中的直接描述，还能将其转化为具体的视觉表现。通用人工智能的发展需要这种抽象和创造性思维的能力，以解决新的、未经编程的问题。
4. **动态环境交互：**Sora能够模拟现实世界中物体、人物和环境的相互作用，这种交互能力对于在不断变化的环境中做出适应性反应至关重要。通用人工智能需要能够在现实世界中有效地“行动”，这不仅需要理解环境，还需要能够预测和影响环境变化。

Sora作为一个能够理解和模拟现实世界的模型，其发展和完善将推动人工智能从专用AI向通用人工智能（AGI）的转变。这种转变意味着AI将不再仅限于执行特定任务，而是能够更广泛地理解和参与我们的世界，从而解决更复杂的问题，并在多种情境下提供帮助。这一切都指向了实现AGI的关键路径之一，即发展能够全面模拟人类理解和创造力的AI系统。

其他观点

💡 360创始人 周鸿祎

1.AGI实现时间缩短:由于科技的发展,AGI的实现时间可能会从10年缩短到1年。

2.科技竞争与人才密度:科技竞争最终取决于人才密度和深度积累。 OpenAI等拥有核心技术的公司在A领域的实力强劲,这表明创业公司不能仅依赖AI技术,还需要有创新和创意。

3.AI对行业的颠覆:AI可能不会立即颠覆所有行业,但它能激发人们的创造力。例如,Sora可能对广告业、电影预告片和短视频行业产生重大影响,但不太可能迅速击败 Tiktok,更可能成为其创作工具。

4.国内A发展水平:尽管国内大模型的发展水平接近GPT-3.5,但与GPT4.0相比仍有一年半的差距。OpenAI可能还持有一些未公开的技术优势。

5.大语言模型的能力:大语言模型不仅仅是填空机,它们能够完整地理解世界知识。Sora通过结合大语言模型(LLM)和 Diffusion技术,实现了对现实世界的理解和模拟,这代表了未来AI的发展方向。

6.AI对物理世界的理解:Sora展示了AI对物理世界的模拟能力,这将对机器人具身智能和自动驾驶等领域产生重大影响,因为这些领域需要对世界有深入的理解。

7.视频学习的重要性: OpenAI训练Sora时可能会阅读大量视频,这将帮助AI更好地理解世界。通过视频学习,AI对世界的理解将远超文字学习,这可能会加速AGI的实现。

💡 心识宇宙ceo 陶芳波

1.Dall-E会被弃用,所有的视觉生成统一到Sora,Sora未来成为世界模型的基础,负责图片、视频、3D。

2.下一步,Sora被统-到LLM,构造下一代 Foundation model(真正的GPT-5?),LLM负责抽象世界,Sora负责现实世界,但架构统一,参数融合(配上各种特异 decoders,如 diffusion),智力的借用也更融合。

3.再往后,不再有各种AIG-X模型,世界范围内统一一套 Foundation Mode和 Foundation Agent的标准和参数,每几年迭代一次,成为 World AI Kernel/世界智能内核。所有人基于该标准,来开发AI-Native应用。

结论
1. OpenAI可能是世界的福音,但是是当下AI创业的噩梦。
2.比特的大一统,比能量(力)的大一统来的更快。

💡 润米咨询创始人 刘润

1、用Sora生成的视频，并不总是能“咬就会有痕”。它“有时”也会出错。但这已经很厉害，很可怕了。因为“先记忆，再预测”，这种理解世界的方式，是人类理解世界的方式。这种方式有个名字，叫：世界模型。2、什么是，世界模型？我举个例子。

3、你的“记忆”中，知道一杯咖啡的重量。所以当你想拿起一杯咖啡时，大脑准确“预测”了应该用多大的力。于是，杯子被顺利拿起来。你都没意识到。但如果，杯子里碰巧没有咖啡呢？你就会用很大的力，去拿很轻的杯子。你的手，立刻能感觉到不对。然后，你的“记忆”里会加上一条：杯子也有可能是空的。于是，下次再“预测”，就不会错了。你做的事情越多，大脑里就会形成越复杂的世界模型，用于更准确地预测这个世界的反应。这就是人类与世界交互的方式：世界模型。

3、关于世界模型，如果感兴趣，我建议你读一本书，叫《千脑智能》。

4、回到Sora。Sora的技术文档里有一句话：

Our results suggest that scaling video generation models is a promising path towards building general purpose simulators of the physical world.

翻译成中文就是：

我们的结果表明，扩展视频生成模型是向着构建通用物理世界模拟器迈进的有希望的路径。

5、什么意思？意思就是说，OpenAI最终想做的，其实不是一个“文生视频”的工具，而是一个通用的“物理世界模拟器”。也就是世界模型，为真实世界建模。

6、而Sora，只是验证了，这条道路可行的一个里程碑。

7、如果从“视频”中，可以开始学习物理的规律了，那么，未来可以不可以从“摄像头”里学习呢？如果也可以的话，那么，给AI装一双“眼睛”，让他满世界跑，会发生什么？如果也可以的话，那么，把全世界的公共摄像头，都开放给OpenAI，会发生什么？

7、Sora的出现，可能意味着，通用人工智能（AGI），正在加速到来。

8、这才是OpenAI，真正想做的事情。

9、所以，这时你就能理解，为什么Sam Altman要筹集7万亿美金，重塑全球AI芯片的基础设施了。7万亿，相当于全球GDP的10%，能买2.5个微软，4个英伟达，或者11.5个特斯拉。为什么？因为，通往通用人工智能的道路上，需要大量、大量、大量的算力。

10、Sora来了，通用人工智能还会远吗？

11、这个世界正在发生着难以想象的变化。看似很远，但又瞬间近在眼前。

网易有道CEO周枫

十天时间，Sora的创新与潜在应用大家已经讨论得挺充分了，我读了不少。周末把Sora和SD 3相关文献读了一下，所以我就补充一些技术沿革和产品思路和理解吧。首先是相关技术的一个简单脉络，然后是产品端的几点想法。

一、DiT及相关技术

Sora和SD 3都是基于Diffusion Transformer（DiT）这个新的图像生成技术，这是Sora作者之一William Peebles的成果，文章是Scalable Diffusion Models with Transformers，2022年12月上了arXiv，正式发表于ICCV（2023年10月）。

DiT这个技术被OpenAI和Stable.ai两大当红AI公司选中作为重要项目的基础，首先当然是性能足够好，下图是作者给出的ImageNet数据集生成效果指标，最重要的FID指标（越小越好）比之前的SOTA LDM一下从3.6降低到了2.27，可以说是质的飞跃。ImageNet数据集生成效果指标.png

另外我理解DiT也比较符合技术趋势，**这几年AI界一大思路就是技术的统一化**，语言生成、翻译、语音识别、图像识别都在往Transformer上面靠，而且纷纷成功，唯有图像生成这个领域还在用UNet这样的卷积模型，而DiT成功地把图像生成用Transformer做出来，并大幅改进SOTA，所以被两大公司同时选中，就不奇怪了。

我们把图像生成和相关技术的发展过程简单捋一下的话，大概是这个样子：

GAN（2014）：这个是现代图像生成技术的开端，2014年的时候深度学习已经火了起来，但是当时主要的能力是识别图像，GAN的作者Ian Goodfellow（Bengio的学生，在Google, Apple都工作过）打破了这个限制，通过GAN就可以让算法具有想像力，完成文生图的任务。

扩散模型（Diffusion Model，2020）：GAN带来了对图像生成的很多兴趣，但有两大缺点，计算量大和难以控制，所以随后出现了很多其它图像生成算法，当前胜出的就是扩散模型。奠基的文章是Berkeley的Ho, Jain，Abbeel的Denoising Diffusion Probabilistic Models（2020）。扩散模型的双向统计图模型.png

Diffusion是一个经典的把统计理论应用到AI中的算法，正向的扩散过程将图片逐步变成随机噪声，而逆向的扩散过程就从随机噪声生成图片，我借机回顾了一下在Berkeley Evans Hall上的随机过程课，所谓“扩散过程”，就是“连续马尔可夫过程”，和一滴墨水放到水里这样的布朗运动有类似性。而通过加入足够强的限制（每次扩散必须是小幅高斯噪声），就使得这个随机过程具有良好的数学性质和刻画能力，通过深度学习SGD训练，就能得到生成图像的逆向模型了。

潜扩散模型（Latent Diffusion Model，2021）：扩散模型可以生成高质量的图像，这是相比GAN的优点，但是因为图像数据维度很高（每一个像素点就是3个颜色维度），所以训练和推理起来都还比较困难。因此扩散模型提出后第二年（2021年），就有降低复杂度的成果出现，Rombach等人的High-Resolution Image Synthesis with Latent Diffusion Models是最重要的文章。

这里有一个思路是一直贯穿到当前包括DiT在内的技术的，就是**降维（Dimensionality Reduction）**以及**潜空间（Latent Space）**，名字不一样，意思差不多，这是AI中关键的一个通用思路，就是当一个问题太复杂，数据量太大，或者有很多对人不重要的数据细节的时候，可以用各种纯数学、自适应、模型训练等等的方法，来将数据转化为更低维度的数据（可以理解为“压缩”）。而一种降维的常用方法，就是使用所谓“潜空间（Latent Space）”或者“潜模型（Latent Model）”，意思就是在原始数据的背后，想办法找到“潜在”数据，用潜在的数据来代表原始数据。举个例子，一段说话的语音数据，波形数据是原始数据，而说话的音素（比如拼音）就是潜空间数据（这就是语音识别经典算法隐马尔可夫模型HMM的思路）。这就是在DiT和Sora中使用的图像块（Patches）方法的出发点，通过将图像分成小块，就可以用潜空间向量来表示图像，大幅降低数据的维度和复杂度，让视频生成问题变成可解决的问题。

Vision Transformer（2020）：2020年Google Brain的Vision Transformer（ViT）技术出现时，也是非常轰动的，业界热火朝天地讨论了很多个月，论文的题目叫An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale，就是玩了一下“一图胜千言”的梗，说“一图胜16*16个词”。ViT这个名字和DiT就差一个字母，的确也是DiT的直接前身之一。我记得当时ViT让大家印象深刻的点，主要是首次让Transformer这个NLP技术，可以被应用到领域之外图像识别的场景。而在ViT里，就已经引入了图像块（Patches）这个重要的技巧，将图像转化成Tokens来处理。所以大家看到Sora的技术报告中介绍Patches，

其实是从这个4年前的成果里来的。当然，Vision Transformer和Diffusion Transformer完成的任务方向相反，前者是图像的识别，后者是图像的合成。ViT在图像块序列中工作.png

DiT（Diffusion Transformer，2022）：Sora的前两个作者William Peebles和Tim Brooks，都是Berkeley教授Alyosha Efros的博士，DiT是Peebles的博士课题，他去年才刚刚毕业，这里插一句，**Sora体现了OpenAI强大的资源和执行力，基于DiT这么新的技术，快速研究视频合成这样的新问题，在短时间内一直推进到产出相对成熟的产品**。很了不起。从DiT技术本身来说，将本来就在Berkeley发端的Diffusion Model推上了新的高度：就像这篇文章开头说的，大家都在探索AI模型的统一化，而DiT成功地将之前一直不能统一的图像生成问题统一到用Transformer解决。

二、Sora和SD 3产品

关于这两个具体的产品，我列一些观察和想法：

1. **Sora的背后没有物理引擎，**有的是DiT架构的Diffusion Transformer，大家能看到的对场景模拟的真实性，是Transformer在大数据量的情况下的强大刻画能力的体现，和大语言模型涌现出逻辑推理等能力是类似的现象。对比来看，可以理解成Sora就在Stable Diffusion这样的文生图模型之外，加了一个时间维度。而Patches这样的降维方法，已经是业内比较成熟的方法。视频因为是运动的，讲故事能力更强，所以观感上让我们很震撼，但从机器来看，既然每一个像素就有几维向量，再加一个时间复杂度维度，并不是那么本质的变化。当然，虽然方法类似，数据还是高维了很多，很多问题要解决，而且实现起来工程难度是非常大的。
2. **“世界模拟”和“通用人工智能AGI”是愿景**。怎么理解OpenAI把Sora定义为“世界模拟”？网上有很多讨论，认为Sora做世界模拟不现实，这样的讨论我觉得就偏颇了，我倾向于认为“世界模拟”是一个非常好的项目愿景，和“通用人工智能”作为整体的愿景一样起到非常正面的作用。愿景既是也不是产品目标，就像“人人平等”一样，是努力的方向，重要的是引发的思考，带来的激励作用，以及能聚集的资源，这两个都是具有号召力的愿景，而且不是完全达不到，所以是非常好的。
3. **Stable Diffusion 3后续版本有希望成为Sora的开源平替**。从目前公开的信息来看，这两个产品从技术架构上有相当的类似性，都是基于DiT架构，而且SD 3承诺了会继续开源（目前还没有），Stable也说会具有视频和3D的能力，和之前的SD版本相比，这是一个新的技术的基础，后续有更多的升级的空间。值得关注。

在哪获取Sora最新消息？

目前Sora还没有正式开放，一切以OpenAI 官网为准。

OpenAI（https://twitter.com/OpenAI）

Sam Altman (OpenAI创始人, https://twitter.com/sama)

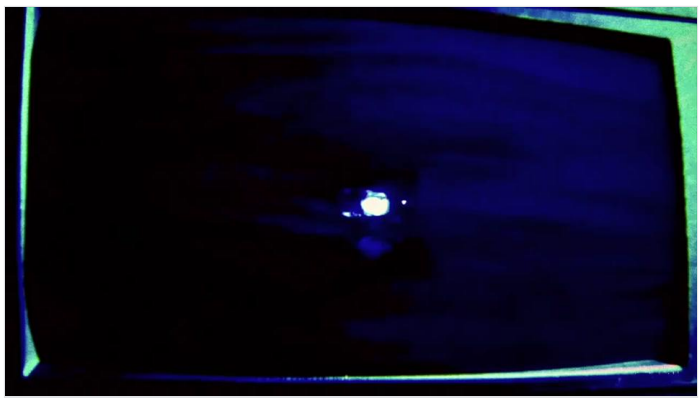
Tim Brooks (研发 Leader, https://twitter.com/_tim_brooks)

Bill Pebbles (研发 Leader, https://twitter.com/billpeeb)

最新资讯

1、OpenAI最新发布了与视觉艺术家、设计师、创意总监和电影制作人等专业人士合作，用 Sora 创作的视频。原文地址：https://openai.com/blog/sora-first-impressions

视频合集：



> 💡 **如果说GPT让AI掌握了文字与语言，拥有了初步的思维能力；那么Sora则是在教会AI理解真实物理世界的运行法则，让它拥有视觉等感官，感受、理解、模拟这个世界。**
>
> **——乌鸦哥**

Sora学习交流

本文由乌鸦哥撰写，引用请注明作者和出处。有什么问题可在评论区留言。