文生视频模型吹响进击号角

本报记者 宋婧

2024年春节档，科技厂商“AI大片”的压轴戏是近期OpenAI亮出的视频模型Sora，它掀起了“文生视频热”，同时也吹响了新一轮大模型进击的号角。

据悉，受益于Sora的大火，OpenAI的估值狂飙275%，在最新一轮融资中有望超过800亿美元，跃升为仅次于字节跳动（2250亿美元）和SpaceX（1500亿美元）的全球第三大独角兽企业。

为什么偏偏是Sora？

明明还没正式开放，Sora已经传遍国内外科技圈。走在东京街道上的时尚女郎、缓缓前进的舞龙队伍、踏雪而来的猛犸、海上自行车比赛……在各大社交媒体上，Sora的Demo（演示）视频被反复播放，登顶热搜榜。与之形成鲜明对比的是，几乎同时发布的谷歌Gemini 1.5 PRO却似乎被人们遗忘在了角落。

360总裁周鸿祎高度评价了Sora。他指出，Sora的面世意味着实现AGI（通用人工智能）的时间将从10年缩短到一两年。在他看来，Sora只是小试牛刀，它展现的不仅仅是视频制作能力，而是大模型对真实世界有了理解和模拟之后，会带来新的成果和突破。SpaceX创始人马斯克则直接在社交媒体上发布了“gg humans（人类输了）”的感叹。“gg”是电子竞技中常用的术语，意为“Good Games”，通常用来表示对对手的认可和自己认输的态度。

实际上，AI视频生成模型并不是一个新鲜事物。谷歌早已发布了零镜头视频生成模型VideoPoet，百度也推出了视频生成模型UniVG，腾讯发布了视频生成模型Video-Crafter2，阿里有自研的视频生成模型Animate Anyone，甚至AI初创企业Pika的视频生成平台Pika 1.0也率先面向所有用户开放了网页端访问权限。为什么偏偏是OpenAI发布的Sora“一炮而红”？

从业内反应来看，Sora最令人震撼的技术突破莫过于视频时长的巨大提升。Sora能生成长达1分钟的视频，远超市面上其他的AI视频模型。此前，Runway能够生成4秒的视频，用户可以将其最多延长至16秒，这已经是AI生成视频在2023年所能达到的最长时长纪录；Stable Video能提供4秒的视频，Pika则可提供3秒的视频。

Sora实现视频时长的突破，背后的大功臣是其采用的Diffusion transformer模型。该模型融合了扩散模型与自回归模型的双重特性，在训练GPT这类大语言模型的时候，OpenAI把句子拆分成tokens（词符），放到transformer进行训练。在Sora中，则是将不同尺寸、分辨率的视频拆分成patch（视觉补丁），把patch当作tokens放到transformer进行训练。训练完成后再通过解码，把tokens“渲染”成人们可以看得懂的像素。

另一个震撼性突破在于Sora展示出了对物理世界部分规律的理解，这是过去文生视频模型的一大痛点。专家分析指出，Sora带有“世界模型”的特质，这让其在逼真度上更胜一筹。

所谓“世界模型”便是对真实的物理世界进行建模，让机器能够像人类一样，对世界产生一个全面而准确的认知。这一特质会使AI视频生成更流畅、更符合逻辑。比如，咬一口饼干，饼干上一定会留下齿痕，这样的逻辑对于人类来说非常简单，而要让AI模型领悟前后两帧画面之间的逻辑关联则非常困难。它需要从大量数据中去学习和掌握生成语言、图像或视频的某种方法，从而产生难以解释的“涌现”能力。

“Sora的成功并非偶然。”Forrester副总裁、研究总监戴鲲在接受《中国电子报》记者采访时表示，这背后有四大推动因素。首先，近期不同领域的最新技术研究进展是促使Sora实现技术突破的关键。其次，OpenAI从2016年起就将生成式模型作为战略方向，长期的技术创新投入积累是其成功的核心要素。最后，高质量的海量数据和大规模高性能硬件投入是必要保证。

Stability AI的CEO埃马德·莫斯塔克（Emad Mostaque）在社交平台上感慨称“奥特曼（OpenAI的创始人兼CEO）真是一个魔术师”，并表示Sora可以被视为AI视频的GPT3，将在未来几年内得到扩展、细化、调整和优化。

Sora并非完美无瑕

“与大语言模型相比，文生视频模型实现难度显然更大。在技术实现上面临的挑战主要体现在数据复杂性、计算资源需求和多模态融合三方面。”戴鲲表示。

以数据复杂性为例，一方面，大语言模型处理的是文本序列，而视频由连续帧组成，每一帧都是一个高维图像，文生视频模型需要同时处理空间和时间两个维度的数据，不仅需要理解单个帧内的像素级关系，还要捕捉帧与帧之间的动态变化和时序依赖，确保生成的视频能够平滑过渡和动作自然，这要求模型具备极高的时空推理能力及对目标对象行为模式的理解。另一方面，大规模高质量的标注视频数据集比大规模文本数据集更难获取，视频数据涉及对颜色、亮度、运动轨迹等多种视觉特征的编码和解码，它的存储和预处理也更为复杂。

与此同时，算力资源的供给也是一个大问题。视频生成涉及大量的视觉信息处理，所需计算量远超文本生成。模型可能需要在数以亿计的参数上进行训练，消耗巨大的GPU算力资源。此外，文生视频模型需要结合音频、文本等多个模态信息，这就需要模型能够有效融合不同类型的输入信号，并输出相应的跨模态内容，这无疑将大大增加模型设计和训练的难度。

现阶段的Sora并非完美无瑕。细心的网民们也在公开的Demo视频中扒出了不少生成式AI的漏洞，比如随着时间推移，有的人物、动物或物品会消失、变形或者生出分身；或是出现一些违背物理常识的画面，比如穿过篮筐的篮球、悬浮移动的椅子。

OpenAI在技术报告中坦诚地公布了Sora的不成熟之处，表示Sora可能难以准确模拟复杂场景的物理原理，可能无法理解因果关系，可能混淆提示的空间细节，可能难以精确描述随着时间推移发生的事件，如遵循特定的相机轨迹等。

英伟达高级科学家Jim Fan指出，目前Sora对涌现物理的理解是脆弱的，远非完美，仍会产生严重且不符合常识的幻觉，还不能很好地掌握物体间的相互作用。

“Sora对真实世界的模拟能力还有很大的提升空间，就目前的展示内容来看，并不意味着它已经‘读懂了’物理规律。”多年从事计算机视觉研究的上海交通大学人工智能研究院副教授王韫博认为。

图灵奖得主、Meta首席AI科学家杨立昆（Yann LeCun）在社交平台上表示：“一个AI模型可以生成逼真的视频，但并不代表这个AI可以理解世界。”他曾提出过生成式模型不适合处理视频的观点，并指出目前最有希望“落地”的是图像识别模型，并不是生成式模型。

此外，Sora的出现也进一步加剧了人们对于AI伦理和安全治理方面的担忧。中国人民大学哲学院教授、国家发展与战略研究院研究员刘永谋指出，在AI短视频构建的世界中，显然不能将眼睛看到的东西作为判断依据。Sora的应用，无疑将进一步加剧当代社会的“后真相”状况，真实与虚拟的边界进一步模糊，甚至完全被消解。“这需要我们高度警惕。”刘永谋说道。

DCCI互联网研究院院长刘兴亮表示，随着AI生成内容与现实之间的界限变得越来越模糊，如何确保内容的真实性和透明性成为一个重要问题。此外，版权、隐私和数据安全等问题也需要得到妥善解决。社会必须面对这些挑战，通过制定相关政策、法律和伦理准则来确保技术的健康发展，同时保护个人和社会的利益不受侵害。

OpenAI仍在进击中

当前，OpenAI的估值在Sora的驱动下，正在大幅飙升。市场预计，在最新一轮由风投公司Thrive Capital牵头的融资中，OpenAI的估值有望超过800亿美元。而作为对比，在OpenAI去年年初发布Chat-GPT的时候，该公司的估值为290亿美元。

然而，OpenAI在生成式AI领域的野心显然不止于此。除了先后甩出“ChatGPT”和“Sora”两张“王炸”，攻破自然语言模型和视频生成模型两座“堡垒”，OpenAI还公布了筹资7万亿美元建立“芯片帝国”的计划。这笔巨额投资相当于美国GDP（国内生产总值）的25%，中国GDP的40%，全球GDP的10%。

OpenAI CEO奥特曼透露，目前OpenAI每天生成约1000亿个单词，需要大量的GPU（图形处理器）芯片进行训练计算——这或许是奥特曼“造芯”的重要原因之一。此前，他曾多次“抱怨”AI芯片短缺问题，称目前英伟达的芯片产能已不足以满足未来的需求。

据业内人士估算，ChatGPT训练一次大约需要2.5万块英伟达A100芯片。如果训练GPT-5，则还需要5万块英伟达H100芯片。市场分析认为，随着GPT模型的不断迭代升级，未来GPT-5或将出现无“芯”可用的情况。所以对于OpenAI而言，下场造芯是顺理成章，也是必然选择。

1月20日消息，阿尔特曼正在与中东阿布扎比G42基金、日本软银集团等全球投资者筹集超过80亿美元资金，成立一家全新的AI芯片公司，目标是利用资金建立一个工厂网络来制造芯片，直接对标英伟达，目前谈判仍处于早期阶段。1月25日，奥特曼在韩国与存储芯片龙头SK海力士、三星电子集团的高管会面，重点提及构建“AI芯片联盟”，双方或将在AI芯片设计、制造等方面与三星和SK集团合作。

除了建厂和供应链合作之外，OpenAI还至少投资了3家芯片公司，包括美国知名算力芯片公司Cerebras（致力于简化芯片制造流程）、芯片初创企业Rain Neuromorphics（擅长算法训练）、Atomic Semi（致力于简化芯片制造流程，实现快速生产，降低芯片成本）。

作为科技圈的“网红”公司，OpenAI的一举一动都会引发业界的高度关注。从自然语言模型ChatGPT到视频生成模型Sora，再到AI芯片产业链，OpenAI在生成式AI领域的布局将有助于确立其在算法和算力上的优势，进而向AGI高地发起总攻。不过，正如杨立昆所言，人工智能技术仍需在抽象表征空间中不断探索和发展。OpenAI能否继续保持领先地位，抢先奔赴AGI的未来，仍有待时间揭晓。

Sora爆火背后：与其忧心忡忡不如埋头实干

> OpenAI视频模型Sora的爆火在春节后迅速引燃整个行业，一时间“史诗级”“地表最强”等溢美之词不绝于耳，甚至引发了中美大模型差距拉大的话题。诚然，Sora所展现的文生60秒一镜到底视频的能力，是生成式AI的巨大飞跃，但放大到中美大模型差距拉大也大可不必。

低头细看，国内在大模型应用上完全不输，技术上差距的一时拉大代表不了永远，也代表不了所有，与其忧心忡忡不如埋头实干。

正视差距 也要正视机遇

如果单论差距，产品只是一个维度上的展现，大模型的综合实力体现在技术能力、数据、算力、人才、应用等多个方面，其中“卡脖子”的是算力和数据质量。在算力方面，我国正在努力构建自主可控的产业链。在数据质量方面，国家数据局等17个部门联合印发了《“数据要素×”三年行动计划（2024—2026年）》，指出要以推动数据要素高水平应用为主线，带动数据要素高质量供给、合规高效流通。在算力和数据质量上，国内各方都在合力追赶。

而从全球技术进步的角度来看，每一次大的技术浪潮都来自一个现象级应用。如果Sora是这个现象级应用，那么它会推动全球而不是单一国家的技术应用创新。中国在大模型技术应用和产业化方面也有着独特的优势。

实际上，如果你问ChatGPT“Sora发布是否意味着中美大模型差距拉大”这样的问题，它的回答都会相当辩证。ChatGPT表示，虽然Sora代表着美国新一代大模型取得的显著进步，但这并不能和中美大模型差距拉大画等号。ChatGPT指出，中国在数据资源方面有一定优势，并持续加大对人才的培养和引进力度。另外，在中文语境下，中国的大模型技术可能会更适用或更具竞争力。

此外，ChatGPT表示，OpenAI发布Sora在某种程度上会加剧中美大模型竞争，但也会激发中国企业加快技术创新和应用落地的步伐，推动中国的大模型技术不断向前发展。由此可见，至少在这个问题上，人工智能工具看问题要更冷静、更客观。

强化大模型 应用优势

应用是中国大模型的优势环节并不是自吹自擂，国内庞大的互联网用户基数、互联网信息化普及程度、产业链配套能力、应用场景、政策支持等都给大模型落地提供了丰富的土壤。《2023AI大模型应用中美比较研究》报告指出，美国在基础模型上保持领先，而中国则在应用上更胜一筹。基于国内庞大的下游需求，国内涌现出大量大模型应用相关创业项目，以及各垂直领域成熟企业的大模型应用。

成熟商用领域大模型应用落地比较顺畅，因为场景明确、需求明确。大模型应用是基于用户真实需求推动，用户对大模型能力的感受会更直观，未来商业化空间也更大。

相比较通用大模型，有能力的垂直企业倾向于研发大模型自用。原因是通用大模型综合能力强，但未必适用于企业所在的垂直领域。比如夸克是一个比较值得观察的案例。与百度等做通用大模型的思路不同，夸克在2023年11月发布的大模型主要是为了服务自身业务，根据自身业务衍生出通识、医疗、教育等垂类模型。正是因为业务明确，在大模型发布后的不到100天里，夸克便已经发布了四款大模型产品：元知、健康助手、AI讲题助手、AI PPT。这四款产品分别在搜索、医疗健康、教育学习、职场办公业务上为用户提供服务。

用户对夸克大模型产品的反馈也比较直观。据了解，夸克健康助手发布后，用户对健康助手的回答满意度相当高，主要原因是夸克健康助手答案的准确性、逻辑性都比较高。在用户群体中有相当比例的用户是医学生，能够获得专业群体青睐，至少说明夸克在解决幻觉率上做得比较好。

准确性高的原因是因为夸克意识到，要解决大模型的应用问题，关键要保证知识的正确性。根据夸克技术负责人的表述，通过模型预训练、人类对齐、模型改进、知识增强等组合，夸克大模型将医疗健康类问答内容幻觉率降低至5%，处在业内领先水平。另外，这也涉及高质量语料库的问题。夸克自身的搜索业务、夸克文档的积累等，都给夸克训练大模型提供了天然土壤，夸克的训练语料在绝对数量上虽然不是最大的，但胜在质量比较高。

在硬件产品和工业领域，也有很多大模型落地场景在推进中。比如在传统PC基础上，联想已经发布了多款AI PC，这是将大模型能力加持到PC上的进化产物；前不久在CES 2024上吸引了诸多目光的口袋AI设备Rabbit R1，也是由中国开发者开发。在工业领域，阿里巴巴的千问大模型就被应用于工业机器人，提高了机器人的任务推理和执行能力。

所以我们缩小差距的机会有吗？答案是有，需要在优势领域扩大优势，落后领域奋起追赶。因为即使在大模型应用领域，目前所处的阶段也比较初级，离真正大规模商用还有较远距离。但整个大模型的竞争远远未到终局，这时候仰望星空是为了明确方向，更重要的是要脚踏实地。（赵 晨）

春晚首次将手机AR技术应用于实体场景

本报记者 谷月

观众拿着手机对着西安的地标性建筑一扫，瞬间“穿越”了时空。跟随着诗人李白的步伐，进入长安城内，舞狮、舞龙创意水袖的舞动灵韵，鼓乐交响的激昂绵长，大唐不夜城再现盛世长安的恢宏景象。“将进酒，杯莫停，与君歌一曲，请君为我倾耳听。”千人齐诵《将进酒》，豪放洒脱，荡气回肠。今年中央广播电视总台春晚，陕西西安分会场的《山河诗长安》登上热搜榜第一，让无数网友直呼：“太燃了！”

“穿越”时空来到大唐不夜城，与诗仙千年再聚，这些全靠科技加持。

据了解，《山河诗长安》节目是中央广播电视总台春晚首次将手机增强现实（AR）技术应用于实体场景。

记者从抖音旗下的视频制作公司火山引擎方面获悉，在《山河诗长安》节目中，火山引擎通过场景结构的精准重建，利用自研视觉定位技术进行贴合度打磨，实现了让李白稳稳地站在大唐不夜城的屋顶弹琵琶，而非“飘”在空中的真实效果。为了保证观众在手机端可以流畅、稳定且实时地欣赏到直播画面，火山引擎通过对庞大的影视级文件进行了模型减面、重展UV、改贴图、绑定动画、引擎适配等优化，保障了手机端的互动流畅和稳定。

据悉，此次应用在手机端的AR技术属于地标AR，是近年来在实体空间，尤其是文化旅游场景中备受瞩目的新型体验方式。通过识别特定地标，激发互动，将地标、文化元素与创新思维进行超越现实的虚拟融合，为大众构建一场奇妙的体验之旅。

而在李白动作逼真、行动流畅的背后，虚拟动作捕捉技术同样立下了汗马功劳。

据中央广播电视总台春晚视觉导演江宇昊介绍，由于李白是个虚拟形象，为了实现最完美的匹配度和真实的效果，他的动作、画面角度都需要多种光影技术进行渲染。

元客视界相关负责人在接受《中国电子报》记者采访时表示，元客视界利用“FZMotion光学运动捕捉系统”，为手机端AR场景中李白的动画表演提供了稳定且精准的动作数据支撑，并且动画制作采用了动作捕捉技术和实时动画预演相结合的方式，更便于导演查看动画角色的动作、走位和表情。

“这段爆燃的影像还只能在屏幕端一饱眼福，希望有朝一日，观众在实景现场也能真正看到虚拟李白，那效果就更好了。”有网友憧憬道。

社长：张立 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元 广告部：010-88558848/8808 发行部：010-88558777 广告许可证：京石工商广登字20170003号 发行单位：中国电子报社 印刷单位及地址：经济日报印刷厂 北京市西城区白纸坊东街2号