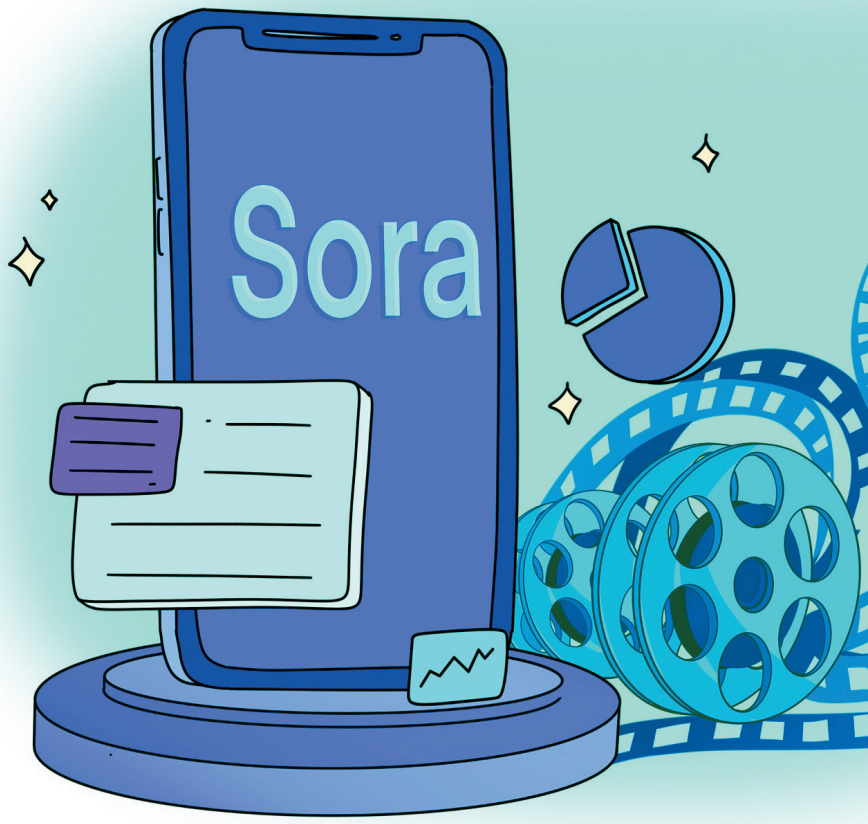
龙年第一热词：

五大问题让你明白，

甲辰龙年迎新，最为吸睛的科技圈大事当属横空出世的文生视频模型Sora。2月16日凌晨，OpenAI官方发布了文生视频模型Sora。与之前市面上的同类工具相比，Sora具有完成更加复杂任务的能力，还能带来更为生动的多视角视频，在生成视频的时长、流畅度和逻辑性等方面优势显着。

有趣的是，Sora发布后，美股知名工具软件公司Adobe股价随即暴跌超7%，图片版权公司Shutterstock跌逾5%，市值一夜蒸发超7000万美元，奈飞、迪士尼等影视公司和数据资产企业Getty Images等股价亦有不同程度的下跌。而国内A股，龙年甫一开市相关AI概念股就掀起涨停潮，不少蹭上Sora概念的小市值公司更是连续斩获涨停。不同市场不同企业股价走势的强烈反差，提醒着大家，对Sora应该有更多的“冷思考”：和过去两年间同样掀起大量讨论的ChatGPT、Midjourney等工具相比，Sora有何过人之处？又是否真能如一些人所鼓吹的那样，“将掀起另一次工业革命”？

■采写：新快报记者 郑志辉
■图片：VCG



1 都是“文转片”，Sora牛在哪？

根据OpenAI发布的示范，只需要给Sora一段二三十字的指令，它就可以生成一段长达一分钟的影片，可以是写实影片，可以是动画，也可以是历史片、黑白片、3D科幻片。

看到这里，一些“AI神教”信奉者已经迫不及待预言，在不久将来，所有人都可以随时随地生成影片，即是说拍摄、绘画、剪片制片的门槛将不再存在。

可是，通过“文字指令”来生成“影片”这件事情，Sora并非全球首家，过去Google、Meta或是创业公司Runway ML都有展示出类似的技术。

不过，在看到公布的几十条示范视频的逼真性和清晰度后，很多人还是被震撼到了。比如那条雪地上撒欢的大狗，毛发丝丝闪亮发光，很具有“大片质感”。这一效果放在电影工业中，像科幻大片《阿凡达》中人物飘逸的长发，那是一大批人一笔笔画了好几个月后，在电脑帮助下制作出来的，Sora却是自动即时生成。再加上“60秒超长长度”“单视频多角度镜头”“充满情感的角色”“高度拟真的细节”和“世界模型”等等优势，对pika、Runway、Stable Video等同行竞品堪称“降维打击”。

有网友在评论时称，“gg Pixar（皮克斯动画制作公司完蛋了）”。连和OpenAI向来不对付的马斯克，在看到Sora作品后，也写下评论谓“gg humans（人类完蛋了）”。

2 技术上有何过人之处？

在技术层面，据记者了解，Sora的核心技术主要包括Diffusion Transformer架构和时空patches。

OpenAI的技术报告显示，基于Diffusion Transformer，从一开始看似静态噪声影片出发，经过多步骤的噪声去除过程逐渐生成影片。而时空patches将不同类型的视觉数据转化为统一的表现形式。同时，该模型对语言有着深刻的理解，能够准确地演绎提示内容，并生成情感表达充分且引人注目的角色。

这可能不太好理解，源码资本在一份报告中通过三个步骤的一系列比喻进行了解释，让大众读者更容易明白一些：

第一步，想象一下，你正在对一间杂乱无章的房间打扫整理，方法是用尽可能少的盒子装下所有东西，同时确保日后能快速找到所需之物。视频压缩网络正是遵循这一原理。它将一段视频的内容“打扫和组织”成一个更加紧凑、高效的形式（即降维）。

接下来，你会为每个盒子编写一张清单。这样，当你需要找回某个物品时，只需查看对应的清单，就能快速定位它在哪个盒子里。在Sora中，类似的“清单”就是空间时间潜在补丁。通过视频压缩网络处理后，Sora会将视频分解成一个个小块，这些小块含有视频中一小部分的空间和时间信息，就好像是对视频内容的详细“清单”。这让Sora在之后的步骤中能针对性地处理视频的每一部分。

最后，想象你和朋友一起玩拼图游戏。这个游戏的目标是根据一段故事来拼出一幅图，每人负责一部分。最终，大家将各自的拼图部分合并，形成一幅完整的图画。在Sora的视频生成过程中，Transformer模型正扮演着类似的角色。它接收空间时间潜在补丁（即视频内容的“拼图片”）和文本提示（即“故事”），然后决定如何将这些片段转换或组合以生成最终的视频，从而讲述文本提示中的故事。

通过这三个关键步骤的协同工作，Sora能够将文本提示转化为具有丰富细节和动态效果的视频内容。不仅如此，这一过程还极大地提升了视频内容生成的灵活性和创造力，使Sora成为一个强大的视频创作工具。

浙商证券电子首席分析师蒋高振概括说：“Sora主要有四方面的突破：其一，视频生成视频。Sora可以依据原视频，用自然语言对其进行修改，以达到更换环境、天气等元素的目的。其二，向过去拓展视频。过去类似产品主要是向未来拓展，而Sora可以向过去拓展，同时保持结尾的一致性。其三，视频拼接。Sora可将不同种类的视频拼接至同一视频的统一场景下。其四，具有交互反馈。”他认为，此次Sora在时长和效果上，均有了更加接近人类拍摄视频的效果。

3 Sora是否被“神化”了？

如果只是“文生片”的超能力，Sora不太可能成为今天这么火爆的话题，关键点是OpenAI的技术报告最后提到，当模型在大规模数据上训练后，模型表现出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。

也就是说，模型训练前并没有给它输入一些物理规则，然而模型在接受大规模数据的训练后，自然而然学习到了这些物理规律。

例如，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。视频主体在暂时地被遮挡或者离开画面后，后续也能继续存在，并且也能保证主体在多个镜头画面中，保持外形的一致性。

Sora有时还能够模拟以简单方式影响世界状态的行为。例如，画家可以在画布上留下随时间持续的新笔触，或者一个人吃汉堡时能留下咬痕。

技术报告最后的结论是，Sora的这些示例，无论是模拟真实场景还是虚拟场景，大部分都体现了物理规律。这表明基于Transformer的Diffusion模型，是发展世界模型的一条可行道路。

这一结论直接被英伟达AI研究院首席研究科学家Jim Fan解读成，“这是一个数据驱动的物理引擎。它是对许多世界的模拟，无论是真实的还是幻想的。”他认为，Sora是一个可学习的模拟器，或“世界模型”。

正是为此，国内外学术界、产业界展开了激烈争论。

激进者如360集团董事长周鸿祎认为，Sora展现的不仅仅是一个视频制作的能力，它展现的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。“一旦AI接上摄像头，把所有的电影和视频都看一遍，对世界的理解将远远超过文字学习，这就离AGI（通用人工智能）真的不远了，不是10年20年的问题，可能一两年就可以实现。”

Meta首席科学家杨立昆则质疑说：“仅凭能够根据提示生成逼真的视频，并不能说明系统真正理解了物理世界。生成过程与基于世界模型的因果预测不同，生成式模型只需要从可能性空间中找到一个合理的样本即可，而无需理解和模拟真实世界的因果关系。”

猎豹移动董事长傅盛表示，Sora这次的重大突破并不代表技术上的重大升级，更可以理解成是一个暴力美学。文生图和文生视频引擎，目前对世界的理解还停留在初级水平，可能只是通过生成符合人类感官的图像来表现，而不是真正的理解。