SORA 的震撼

就像滚热的油锅中，撒进了一大把海盐——Sora 来了。

当地时间 2 月 15 日，人工智能研究公司 OpenAI，发布了首个视频生成模型 Sora ——“世界模拟器”。当 Sora 视频亮相时，“世界”被“模拟”后的震撼扑面而来：仅仅依靠几句文字提示语，Sora 自动生成了雪地里撒欢的大狗、咖啡里破浪的帆船、街景里雪舞的樱花，惟妙惟肖、纤毫毕露、几可乱真。

Sora 带来的震撼，让许多人惊愕不已。

人工智能（AI）发展至今，本质上是机器通过模仿、学习人类的智能，接近、达到（甚至超过）人类的智能，以帮助减轻人类的劳动、提高人的能力。它是人类主导、模拟人类智能的科学，人通过设计学习路径——建模，让机器反复学习以具备特定能力。

这种运行方式，近乎于手工作坊。想要生产什么样的产品，就提供什么样的学习图纸，然后机器对应反复学习，由此具备设定的能力。Sora 的意义在于，只需要提供一些提示性的文字语言，它会自动生成人类想要的场景。这就意味着，Sora 会理解人类的思想！

Sora 亮相的那天，大家最初都惊讶于文（字）生视频的逼真性和清晰度。这当然不容易。比如那条雪地上撒欢的大狗，毛发丝丝闪亮发光，十分清晰逼真。如果跟现在的类似竞品 pika、Runway、Stable Video 等对比，几乎是信息时代与石器时代的差别。

类似效果，以前需要人工一笔笔画出毛发，然后建模渲染，以形成活生生的形象，成品也不如 Sora。科幻大片《阿凡达》中人物飘逸的长发、大海里汹涌的海浪，就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 能自动即时生成如此生动的视频图像，当然不容易。

Sora 更震撼的，是视频产生、生成的过程。它依据给出的文字提示，能理解其中蕴含的信息，准确地转换（想象）成匹配的图像场景，最终形成符合人类意图可长达一分钟的视频。Sora 具备的理解能力、从文字到图像的想象力，已接近人类特有的复杂想象判断能力，这是最关键、最有意义的。

比如人类对海浪的认知，并不需要通过一帧帧图像建模来实现，只要见过大海的人，马上就会想象生成图景。Sora，就具备了某些这样的从文字想象到图景的能力。它反映了对物理世界的理解，已经从文字进到图像，从图像进到对这个世界 3D 环境的理解。相比一年多前同样由 OpenAI 发布的 chatGPT，已经从文字理解、文字解读的一维层面，跃升到文字直接生成视频的二维和三维层面。这是质的巨大飞跃。

Sora 是怎样得到这种能力的呢？

根据技术团队透露的信息，Sora 的诞生，有着诸多与众不同、与以往不同的方式。

第一是解构视频。将极大数量的各类视频（可视数据）碎片化，转化为可统一标识的特定编码，便于输入信息时认知。第二，视频与文字的巨大不同在于有复杂的格式差异，Sora 着重细化了不同分辨率、持续时间和纵横比的视频和图像的解析和标识，方便应用时可组合成不同需求的视频。第三，将可视数据转换成数据包。这是非常重要的一步，碎片化、精细化的数据，只有通过一定组合的数据包，才能被有效、可扩展的运用。第四，建立高度描述性的转译员模型。它具有两方面作用，一方面大量训练学习带有文本解读的视频，理解每一帧画面包含的文本意义；另一方面，接受文本传递的信息，学习训练得到相应的画面和图像。

Sora 还充分利用了一年多前诞生的 GPT 技术，将用户提供的简短文字提示，先由 GPT 转换为更长的详细描述，再发送给视频模型，这大大提高了按照用户提示生成高质量视频的精准度。从这个意义上来说，GPT 实际上是 Sora 得以诞生的关键一步。

Sora 的出现，是人工智能领域一次重要的进步。它自动解析文字描述，用真实物理定律孪生虚拟数字世界，重构真实世界与虚拟空间互动。它能够将人们的想象力转化为生动的动态画面，将文字的魔力转化为视觉的盛宴。它也预示着一个全新视觉叙事时代的到来，将给传媒、影视、教育等诸多行业，带来印象深刻的变化。

“以前不相信是真的，现在不相信是假的。”