中学生能看懂：Sora 原理解读

🤖 原文：https://mp.weixin.qq.com/s/KUnXlDlg-Rs_6D5RFpQbnQ

原创 金色传说大聪明 赛博禅心 *2024-02-17 15:32 广东*

文 / WebPilot Hugo API 图 / DALL·E

写在前面

- Sora 是 OpenAI() 在昨天凌晨发布的超强视频生成AI，旨在探索AI如何在理解真实世界运动和交互方面做得更好
- 本文的文字部分，由 **WebPilot Hugo API** 生成，未经修改（喜
- 本文的例子/演示部分， 来自 Sora 的相关文档，但 Sora 本身并未开放（悲
- 和多位 OpenAI 的朋友进行了确认，目前 Sora 也没灰度体验。
- **Sora 好棒，吹爆！** ╰(｀□′)╯

在深入了解Sora如何处理多样化视觉数据之前，让我们首先想象这样一个生活中的场景：你正在翻看一本世界名胜的相册，这本相册中包含了不同国家、不同风格的景色照片，有的是宽阔的海景，有的是狭窄的巷道，还有的是夜晚灯火辉煌的城市风光。尽管这些照片内容和风格各异，但你能轻松地辨识每一张照片代表的地点和情感，因为你的大脑能够将这些不同的视觉信息统一理解。

现在，让我们将这个过程与Sora处理多样化视觉数据的方式进行对比。Sora面对的挑战就像是需要处理和理解来自世界各地、不同设备拍摄的数以百万计的图片和视频。这些视觉数据在分辨率、宽高比、色彩深度等方面都存在差异。为了让Sora能够像人类大脑那样理解和生成这么丰富的视觉内容，OpenAI开发了一套将这些不同类型视觉数据转换为统一表示形式的方法。

在遗迹上飞的机机

首先，Sora通过一个叫做“视频压缩网络”的技术，将输入的图片或视频压缩成一个更低维度的表示形式，这一过程类似于将不同尺寸和分辨率的照片“标准化”，便于处理和存储。这并不意味着忽略原始数据的独特性，而是将它们转换成一个对Sora来说更容易理解和操作的格式。

接下来，Sora将这些压缩后的数据进一步分解为所谓的“空间时间补丁”（Spacetime Patches），这些补丁可以看作是视觉内容的基本构建块，就像是我们前面相册中的每一张照片都能分解为包含独特景观、颜色和纹理的小片段。这样，不管原始视频的长度、分辨率或风格如何，Sora都可以将它们处理成一致的格式。

通过这种方法，Sora能够在保留原始视觉信息丰富性的同时，将不同来源和风格的视觉数据统一成一种可操作的内部表示形式。这就像你在查看世界名胜相册时，尽管照片多种多样，但你依然能通过相同的方式去理解和欣赏它们。

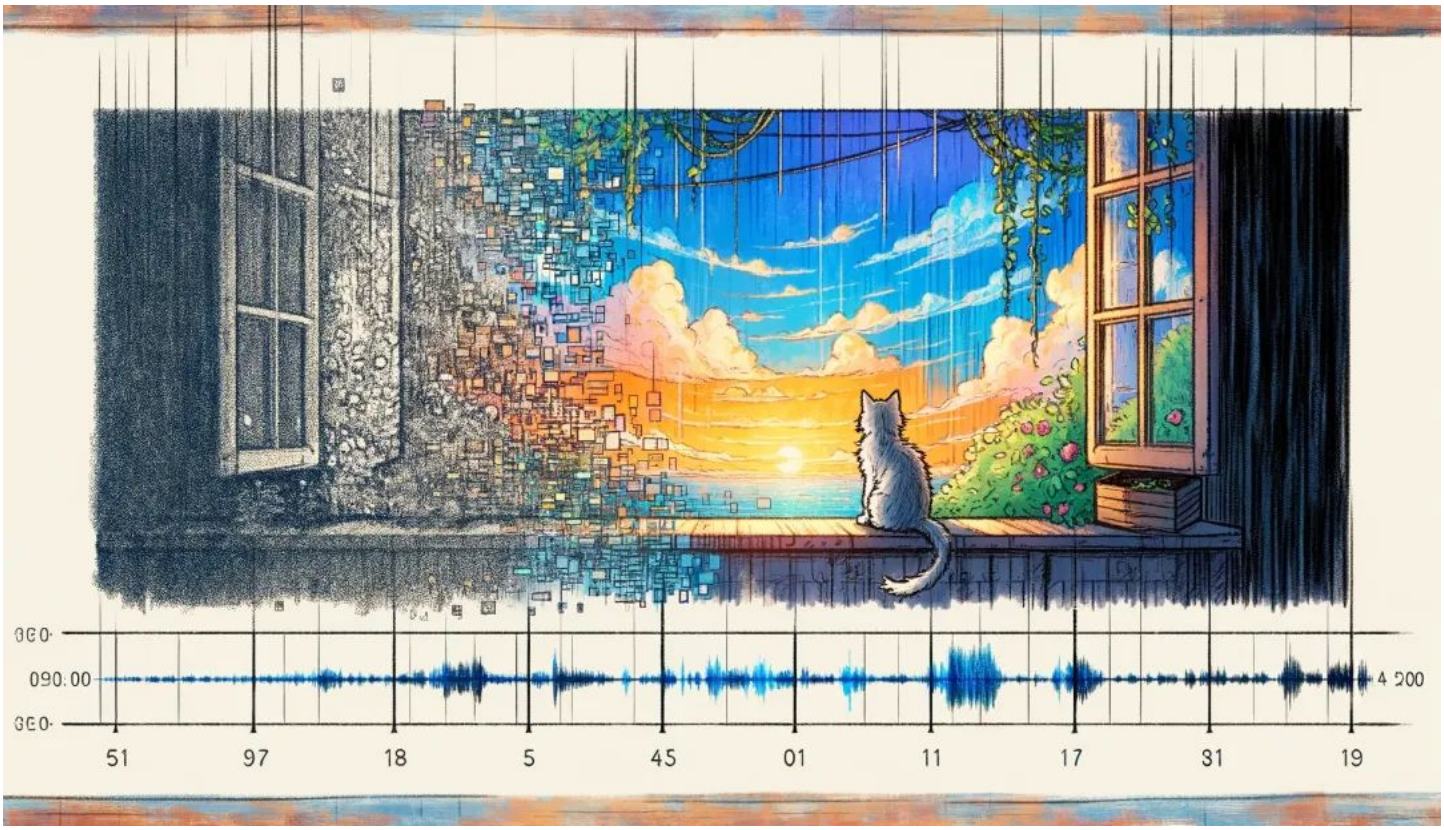
在水下遗迹的蝶蝶（什么鬼！）

这种处理多样化视觉数据的能力，使得Sora在接收到如‘猫坐在窗台上’这样的文本提示时，不仅能理解这个提示背后的意图，还能利用它的内部表示形式，综合利用不同类型的视觉信息，生成与文本提示相匹配的视频或图片。就好比是从全世界的视觉数据中找到那些能够拼凑出你想象中的“猫坐在窗台上”场景的片段，并将它们组合起来，创造出一个全新的视觉作品。

文本条件化()的Diffusion模型

紧接着空间时间补丁的概念，接下来我们探讨Sora如何根据文本提示生成内容的机制。这一过程核心依赖于一种名为“文本条件化的Diffusion模型()”。为了理解这个技术的原理，我们可以用一个日常生活中的比喻来帮助理解：想象你手里有一本涂鸦的草稿本，刚开始时，草稿本上只有随机的斑驳笔迹，看起来毫无意义。但如果你按照某个指定的主题，比如“花园”，逐步地去修改和优化这些斑驳的笔迹，最终，这些无序的线条就会逐渐变成一幅美丽的花园画面。在这个过程中，你的“指定主题”就像是文本提示，而你逐步优化草稿本的过程，就类似于Diffusion模型的工作方式。

具体到Sora的实现，这个过程开始于一段与目标视频同样时长、但是内容完全是随机噪声的视频。可以把这段噪声视频想象成草稿本上那些毫无意义的斑驳笔迹。随后，Sora根据给定的文本提示（比如“一只猫坐在窗台上看日落”）开始“涂改”这段视频。在这个过程中，Sora利用了大量的视频和图片数据学习到的知识，来决定如何逐步去除噪声，将噪声视频转变成接近文本描述的内容。



这个“涂改”过程并不是一蹴而就的，而是通过数百个渐进的步骤完成的，每一步都会让视频离最终目标更进一步。这种方法的一个关键优势在于其灵活性和创造性：同一段文本提示，通过不同的噪声初始状态或通过稍微调整转化步骤，可以生成视觉上截然不同、但都与文本提示相符的视频内容。这就像是多个画家根据同一主题创作出风格各异的画作。

通过这种基于文本条件的Diffusion模型，Sora不仅能生成具有高度创造性的视频和图片，还能确保生成内容与用户的文本提示保持高度一致。无论是模拟真实场景还是创造幻想中的世界，Sora都能依据文本提示“涂改”出惊人的视觉作品。

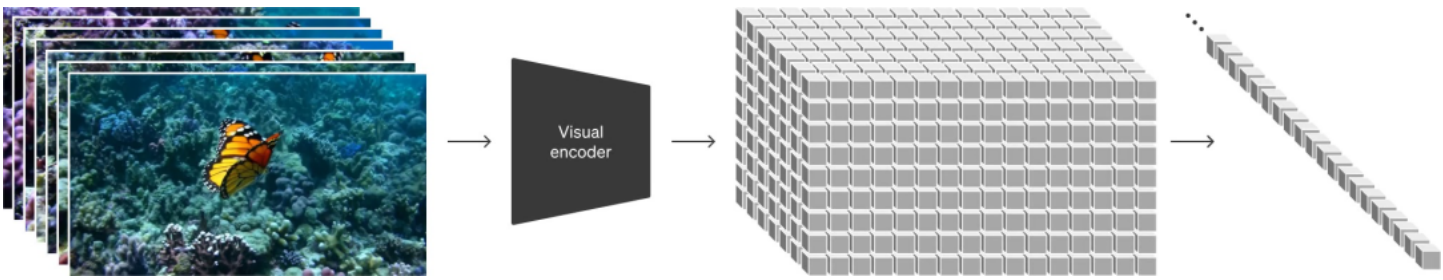
文本条件化的Diffusion模型赋予了Sora强大的理解和创造力，让它能够跨越语言与视觉之间的障碍，将抽象的文字描述转化成具体的视觉内容。这一过程不仅展示了AI在理解自然语言方面的进步，也开辟了视频内容创造和视觉艺术领域的新可能性。

紧接此部分，我们将进入对Sora视频生成过程的进一步探讨，特别是视频压缩网络和空间时间潜在补丁在这一过程中的作用和重要性。

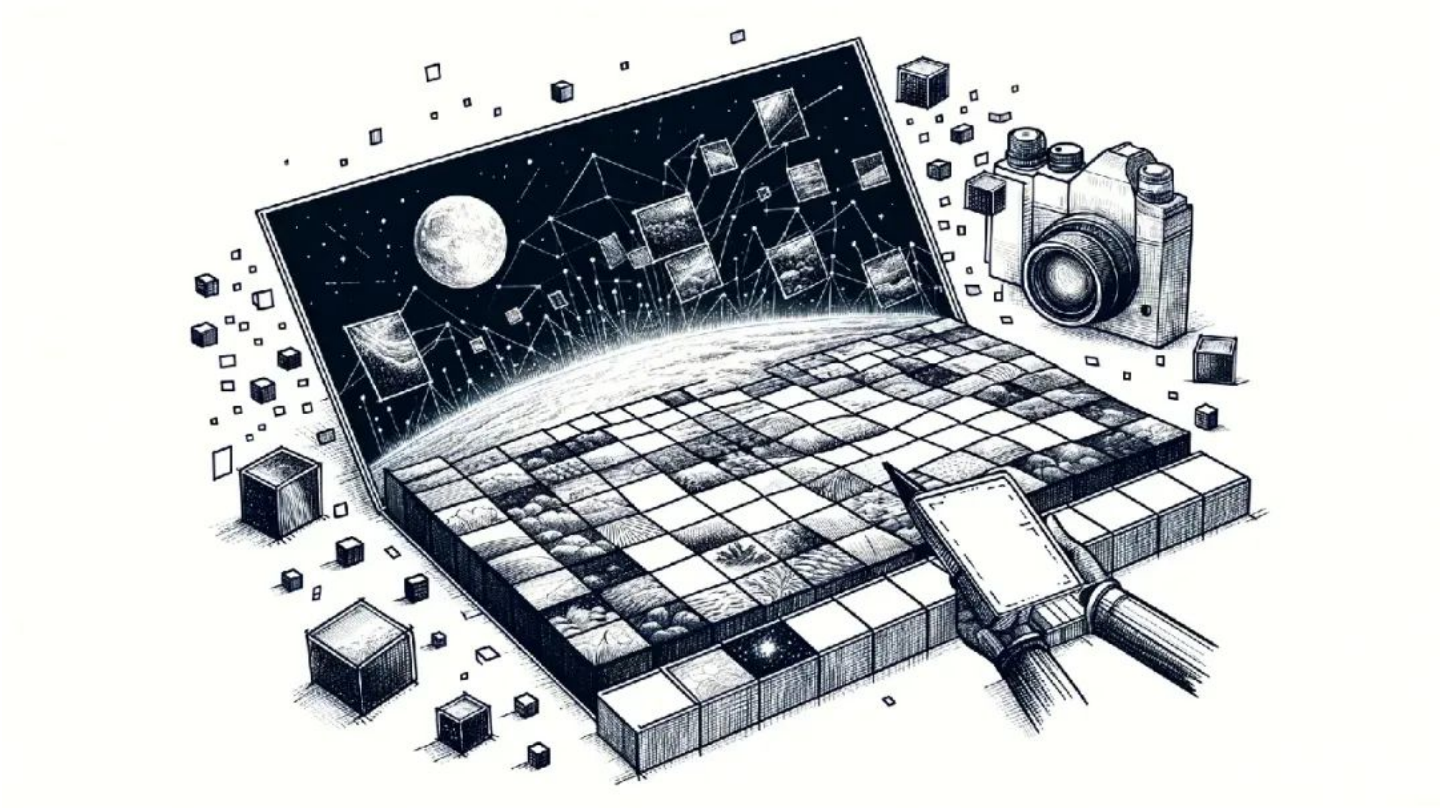
空间时间补丁（Spacetime Patches）

在深入讨论Sora如何通过三个关键步骤生成视频之前，让我们先集中探索一下空间时间补丁（Spacetime Patches）这一概念。这一概念对于理解Sora如何处理复杂视觉内容至关重要。

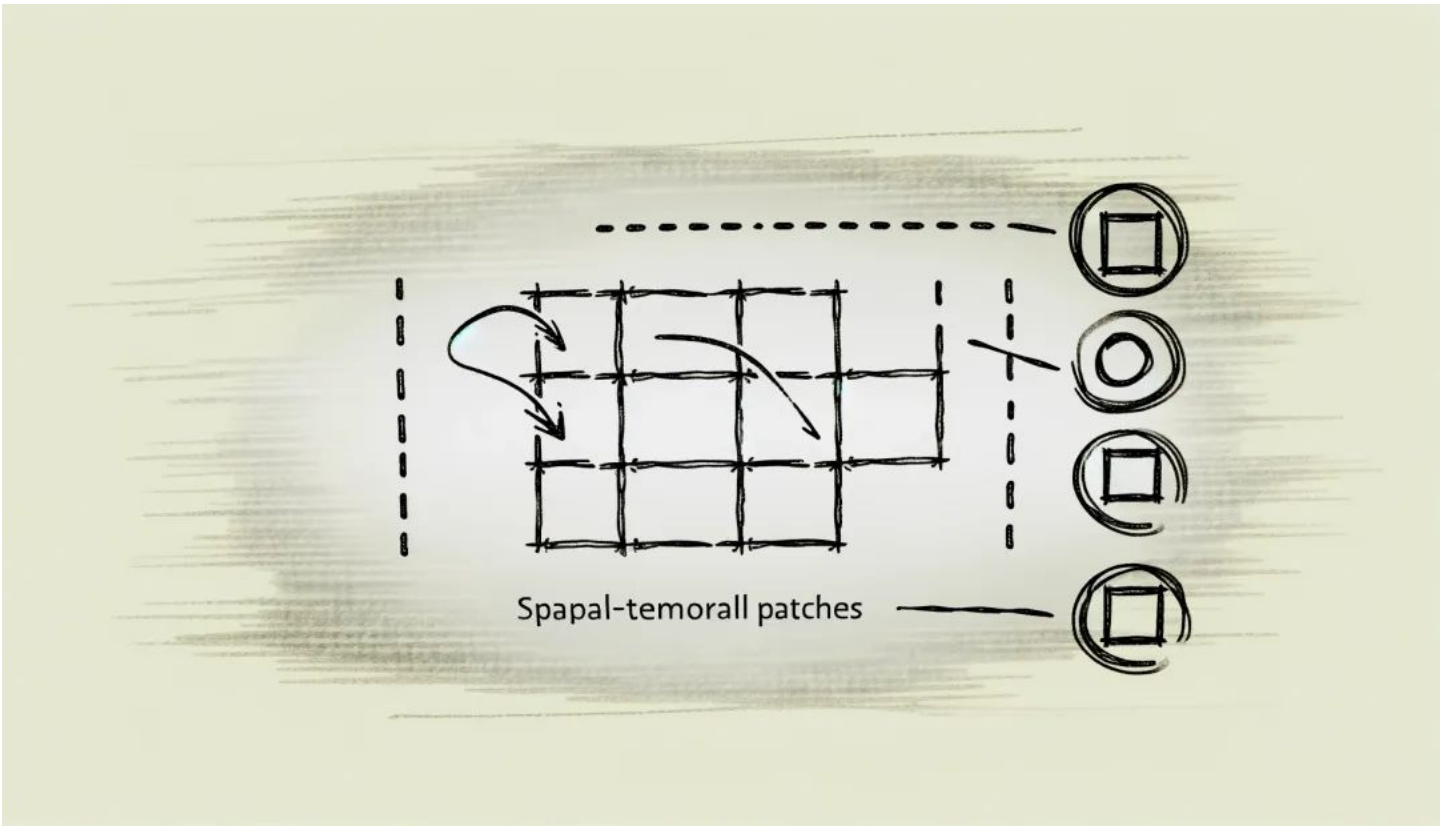
空间时间补丁可以简单理解为将视频或图片内容分解为一系列小块或“补丁”，每个小块都包含了部分时空信息。这种方法的灵感来源于处理静态图像的技术，其中图像被分成小块以便于更有效地处理。在视频处理的背景下，这一概念被拓展到了时间维度，不仅包含空间（即图像的部分区域），还包括时间（即这些区域随时间的变化）。



为了理解空间时间补丁是如何工作的，我们可以借用一个简单的日常生活中的比喻：想象一下，你在观看一部动画电影。如果我们将这部电影切割成一帧帧的静态画面，每帧画面进一步切割成更小的区域（即“补丁”），那么每个小区域都会包含一部分画面的信息。随着时间的推移，这些小区域中的信息会随着物体的移动或场景的变化而变化，从而在时间维度上添加了动态信息。在Sora中，这样的“空间时间补丁”使得模型可以更细致地处理视频内容的每一个小片段，同时考虑它们随时间的变化。

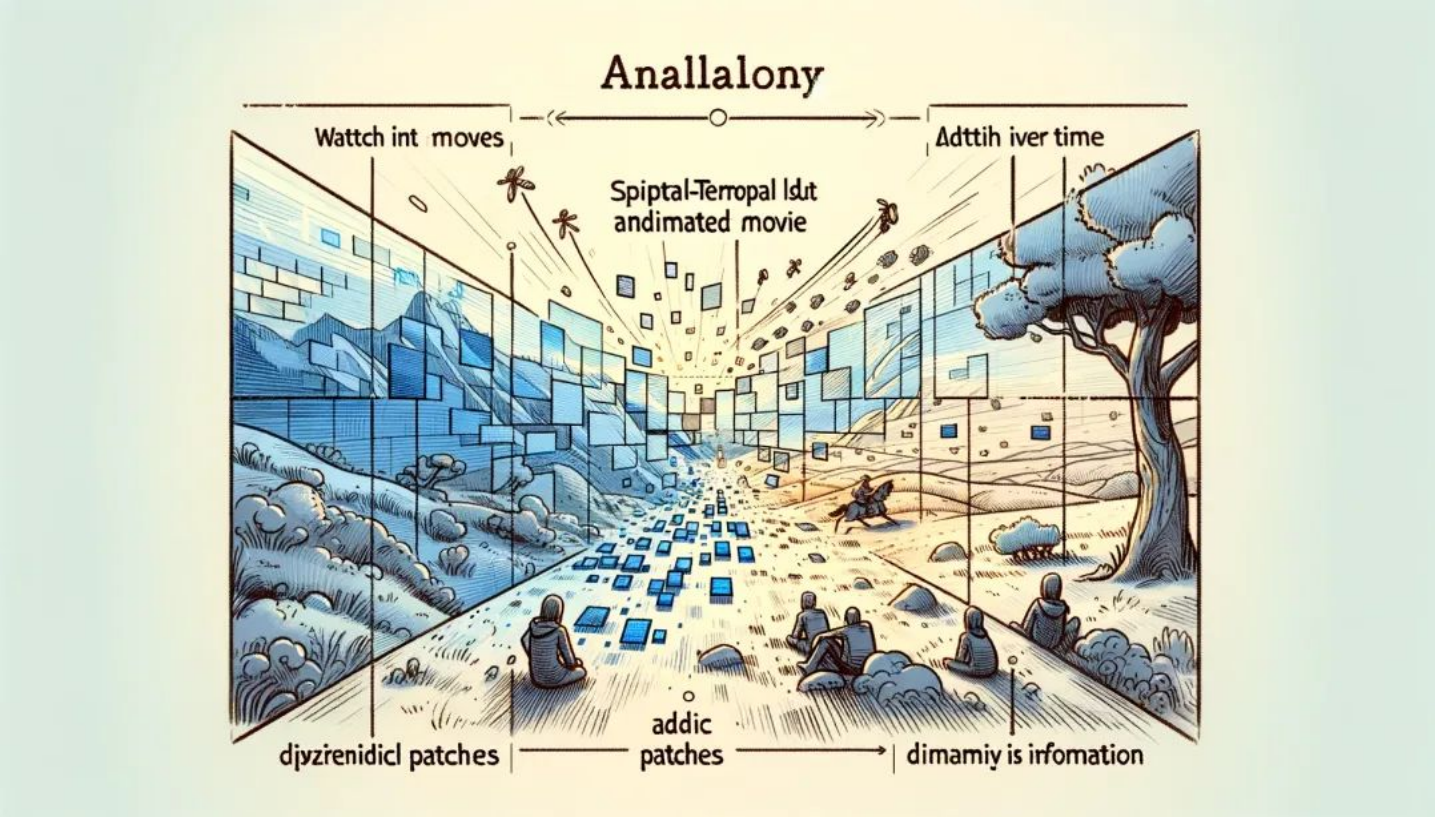


具体到Sora处理视觉内容的过程中，空间时间补丁首先通过视频压缩网络生成。这一网络负责将原始视频数据压缩成更低维度的表示形式，即一个由许多小块组成的密集网络。这些小块即为我们所说的“补丁”，每个补丁都携带了一部分视频的空间和时间信息。



一旦生成了这些空间时间补丁，Sora就可以开始它们的转换过程了。通过预先训练好的转换器（Transformer模型），Sora能够识别每个补丁的内容，并根据给定的文本提示进行相应的修改。例

如，如果文本提示是“雪地中的狗狗奔跑”，Sora将找到与“雪地”和“奔跑的狗狗”相关的补丁，并相应调整它们，以生成与文本提示匹配的视频内容。

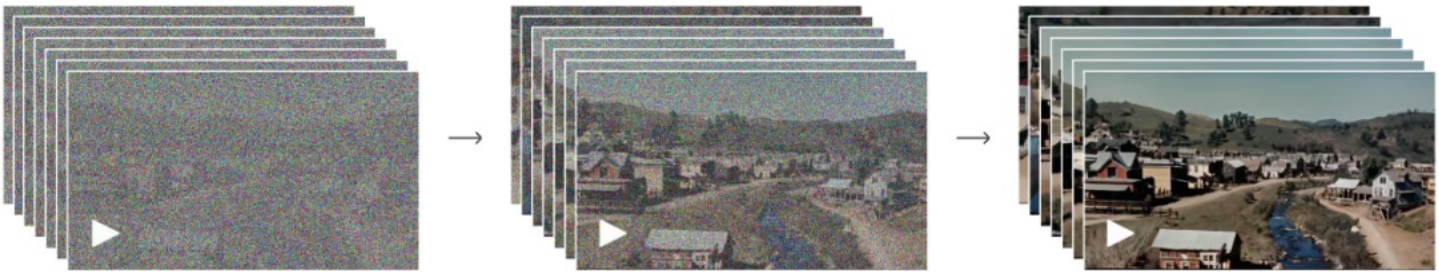


这种基于空间时间补丁的处理方式有几个显著优势。首先，它允许Sora以非常精细的层次操作视频内容，因为它可以独立处理视频中的每一小块信息。其次，这种方法极大地提高了处理视频的灵活性，使得Sora能够生成具有复杂动态的高质量视频，而这对于传统视频生成技术来说是一个巨大的挑战。此外，通过对这些补丁进行有效管理和转换，Sora能够在保证视频内容连贯性的同时，创造出丰富多样的视觉效果，满足用户的各种需求。

随着对Sora视频生成过程的进一步探讨，我们可以看到，空间时间补丁在这一过程中扮演了极其重要的角色。它们不仅是Sora处理和理解复杂视觉内容的基石，也是使得Sora能够高效生成高质量视频的关键因素之一。接下来，我们将更深入地探讨视频压缩网络及其与空间时间潜在补丁之间的关系，以及它们在视频生成过程中的作用和重要性。

视频生成过程

接着上文对于**空间时间补丁**的介绍，我们将详细探讨Sora在视频生成过程中的三个关键步骤：视频压缩网络、空间时间潜在补丁提取以及视频生成的Transformer模型。通过一系列比喻，我们将尝试让这些概念变得更加易于理解。

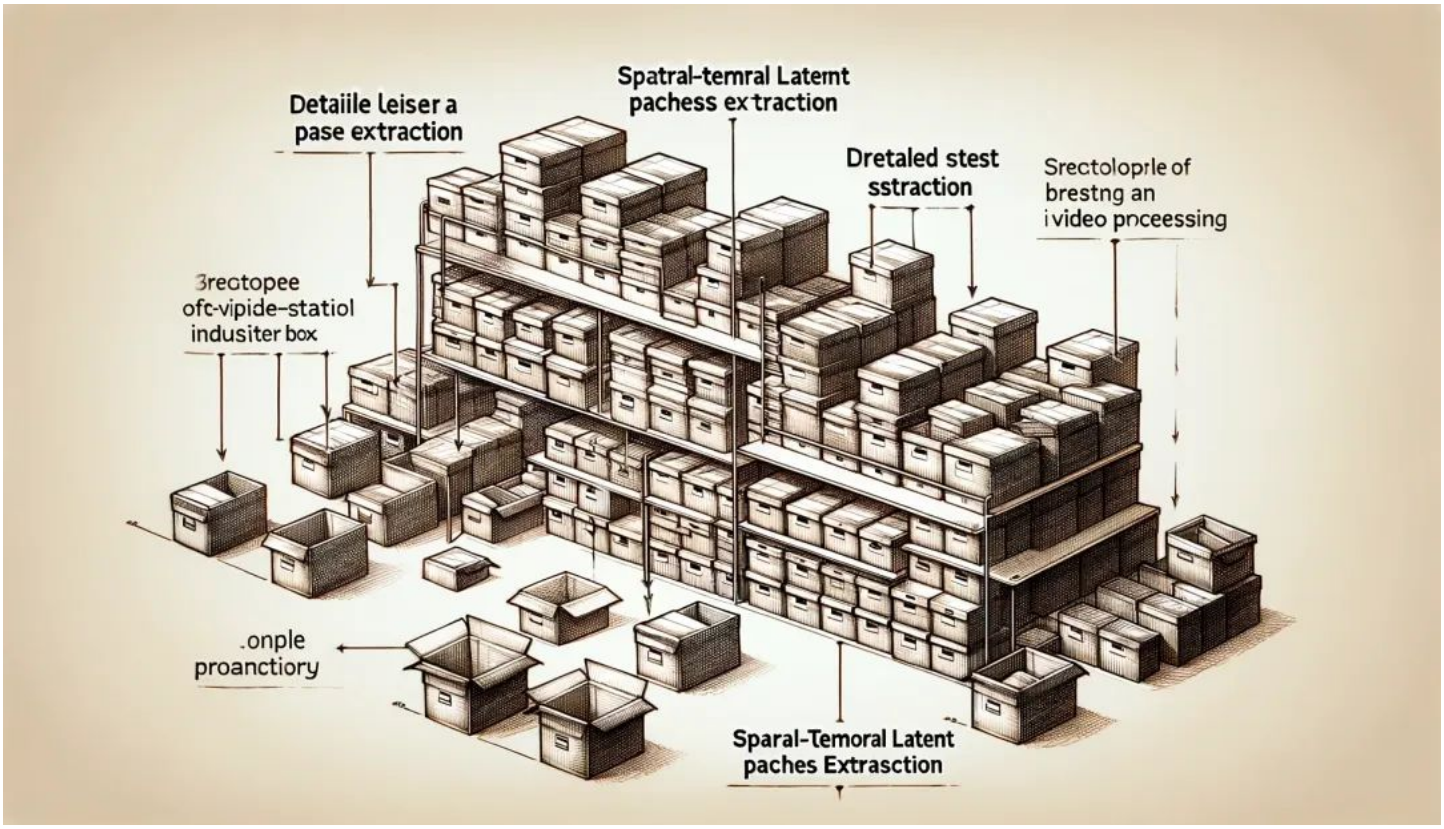


步骤一：视频压缩网络

想象一下，你正在将一间杂乱无章的房间打扫干净并重新组织。你的目标是，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。在这个过程中，你可能会将小物件装入小盒子中，然后将这些小盒子放入更大的箱子里。这样，你就用更少、更有组织的空间存储了同样多的物品。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）。这样，Sora就能在处理时更高效，同时仍保留足够的信息来重建原始视频。

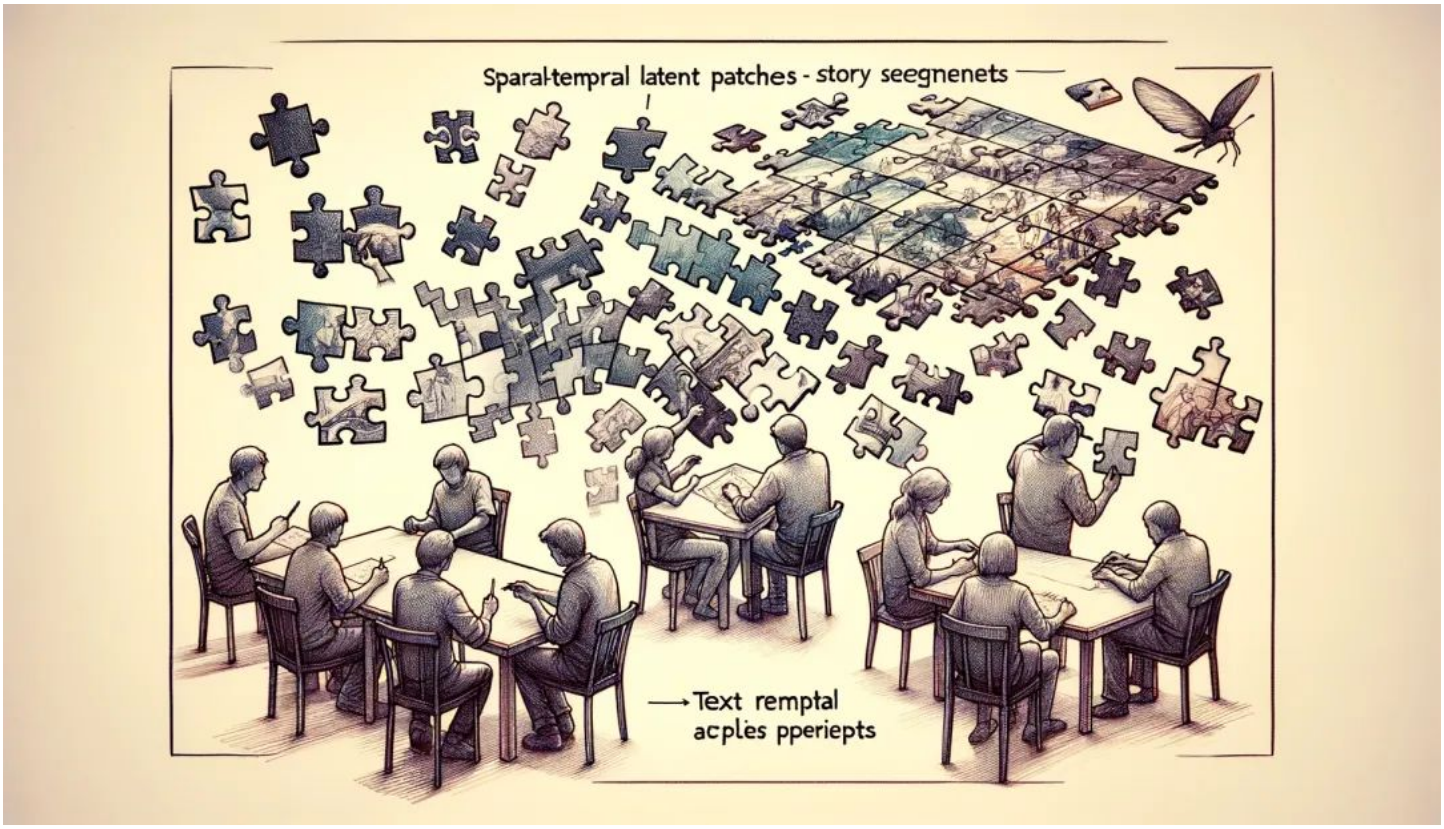
步骤二：空间时间潜在补丁提取

接下来，如果你想要细致地记下每个盒子里装了什么，可能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。



步骤三：视频生成的Transformer模型

最后，想象你和朋友一起玩拼图游戏，但游戏的目标是根据一段故事来拼出一幅图。你们先将故事拆分成若干段落，每人负责一段。然后，你们根据各自负责的故事段落选择或绘制出拼图的一部分。最终，大家将各自的拼图部分合并，形成一幅完整的图画，讲述了整个故事。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。

技术特点与创新点

接下来，我们将深入了解Sora的技术特点与创新点，以便更好地理解它在视频生成领域的领先地位。

支持多样化视频格式

首先，Sora展现出了对多样化视频格式的支持力度。举例来说，无论是宽屏的1920x1080p视频、垂直的1080x1920视频，还是其他任意比例的视频，Sora都能够应对自如。这种能力使得Sora能直接为不同设备生成其原生比例的内容，从而适应多变的观看需求。此外，Sora还能在较低分辨率下快速原型内容，然后再全分辨率下生成，所有这些都在同一个模型下完成。这个特点不仅提高了内容创作的灵活性，也极大地简化了视频内容的生成流程。

扁的龟龟

方的龟龟

长的龟龟

改进的视频构图和框架

进一步地，Sora在视频构图和框架上也展示了明显的改进。通过在原生比例上进行训练，Sora可以更好地掌握视频的构图和框架设计，与那些将所有训练视频裁剪成正方形的模型相比，Sora能够更加准确地保持视频主题的全貌。例如，对于宽屏格式的视频，Sora可以确保主要内容始终处于观众视线中，而不会像某些模型那样，只显示主题的一部分。这不仅提高了生成视频的视觉质量，也提升了观看体验。

奔走的车车

语言理解与视频生成

Sora对文本的深度理解能力是其另一个重要特点。利用先进的文本解析技术，Sora可以准确理解用户的文本指令，并根据这些指令生成具有丰富细节和情感的角色以及生动的场景。这种能力使得从简短的文本提示到复杂视频内容的转换变得更加自然和流畅，无论是复杂的动作场景还是细腻的情感表达，Sora都能够精确捕捉并展现。

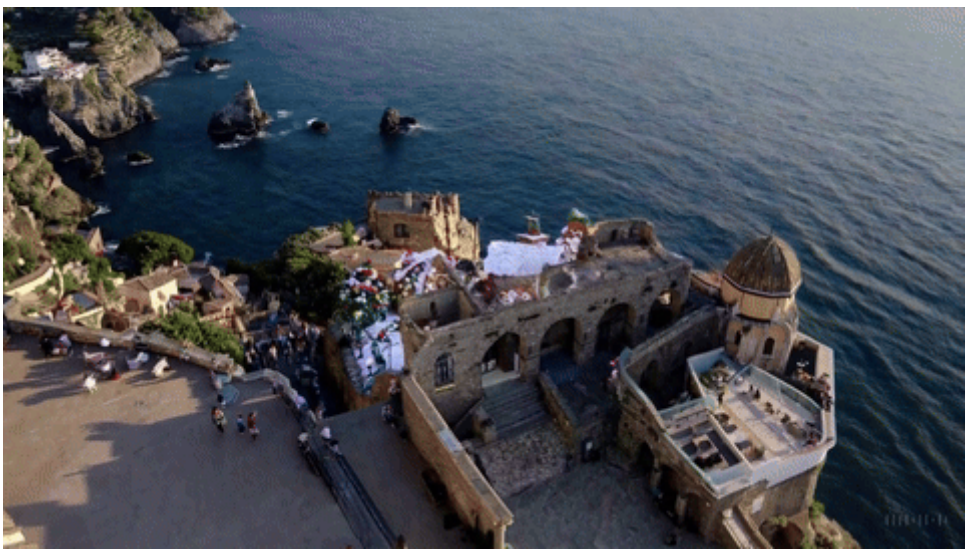
好吃的堡堡

多模态输入处理

最后，Sora的多模态输入处理能力也不容忽视。除了文本提示外，Sora还能够接受静态图像或已有视频作为输入，进行内容的延伸、填充缺失帧或进行风格转换等操作。这种能力极大地扩展了Sora的应用范围，不仅可以用于从零开始创建视频内容，也可以用于已有内容的二次创作，为用户提供更多的创意空间。

第一个输入

第二个输入



视频合成，启动！

通过上述四个方面的技术特点与创新点，Sora在视频生成领域确立了其领导地位。无论是在视频格式的支持、视频构图的改进，还是在语言理解与多模态输入处理上，Sora都展现出了其强大的能力和灵活性，使其成为不同领域创意专业人士的有力工具。

Sora不仅可以生成具有动态摄像机运动的视频，还能模拟简单的世界互动。例如，它可以生成一个人走路的视频，展现出3D一致性和长期一致性。

模拟能力

Sora的模拟能力在AI视频生成领域中展现出了独特的优势。以下是其在模拟真实世界动态和互动方面的关键能力：

3D一致性

Sora能够生成展现动态摄像机运动的视频，这意味着它不仅能捕捉到平面图像中的动作，还能以3D的视角呈现物体和人物的运动。想象一下，当摄像机围绕一个正在跳舞的人物旋转时，你可以从不同的角度看到这个人的动作，而人物的每一个动作和背景都能保持在正确的空间位置上。这种能力展现了Sora对三维空间理解的深度，使得生成的视频在视觉上更加真实和生动。

一直转转的山山

长期一致性

在生成长视频时，保持视频中的人物、物体和场景的一致性是一项挑战。Sora展示了在这方面的卓越能力，能够在视频的多个镜头中准确保持角色的外观和属性。这不仅包括人物的外表，还包括他们的行为和与环境的互动。例如，如果一个视频中的角色开始时穿着红衣服，那么即使在视频的不同部分中，这个人物的衣着也会保持一致。同样，如果视频描绘了一个人物从一张桌子走向另一张桌子，即使视角发生了变化，人物与桌子的相对位置和互动也会保持准确，体现了Sora在维持长期一致性上的强大能力。

总在张望的狗狗

世界交互模拟

更进一步，Sora还能模拟人物与环境之间简单的互动，比如一个人走路时脚下的尘土飞扬，或是在绘画时画布上颜色的变化。这些细节虽小，却极大地增强了视频内容的真实感。例如，当一个角色在视频中画画，Sora不仅能生成动作本身，还能确保每一笔都在画布上留下痕迹，这些痕迹随着时间的推移而累积，展现了Sora在模拟真实世界互动方面的细腻处理。

我画不出的花花

通过这些技术特点，Sora能够在生成视频内容时，不仅模拟动态的视觉效果，还能捕捉到更深层次的，与我们日常生活经验一致的互动模式。尽管在处理复杂的物理交互和长时间一致性上仍存在挑战，但Sora在模拟简单世界互动方面已经展现出了显著的能力，为未来AI技术的发展开辟了新的路径，特别是在理解和模拟真实世界动态这一领域。

讨论与局限性

尽管Sora作为OpenAI最新发布的视频生成AI模型，在模拟真实世界动态和互动方面取得了显著进步，但它仍然面临一些局限性和挑战。以下是Sora目前的主要局限性及探讨如何克服这些挑战的途径。

物理世界模拟的局限性

Sora虽然能够生成具有一定复杂度的动态场景，但在模拟物理世界的准确性方面仍然存在局限。例如，对于复杂的物理互动，如玻璃破碎的精细过程，或是涉及精确力学运动的场景，Sora有时无法准确再现。这主要是因为Sora目前的训练数据中缺乏足够的实例来让模型学习这些复杂的物理现象。

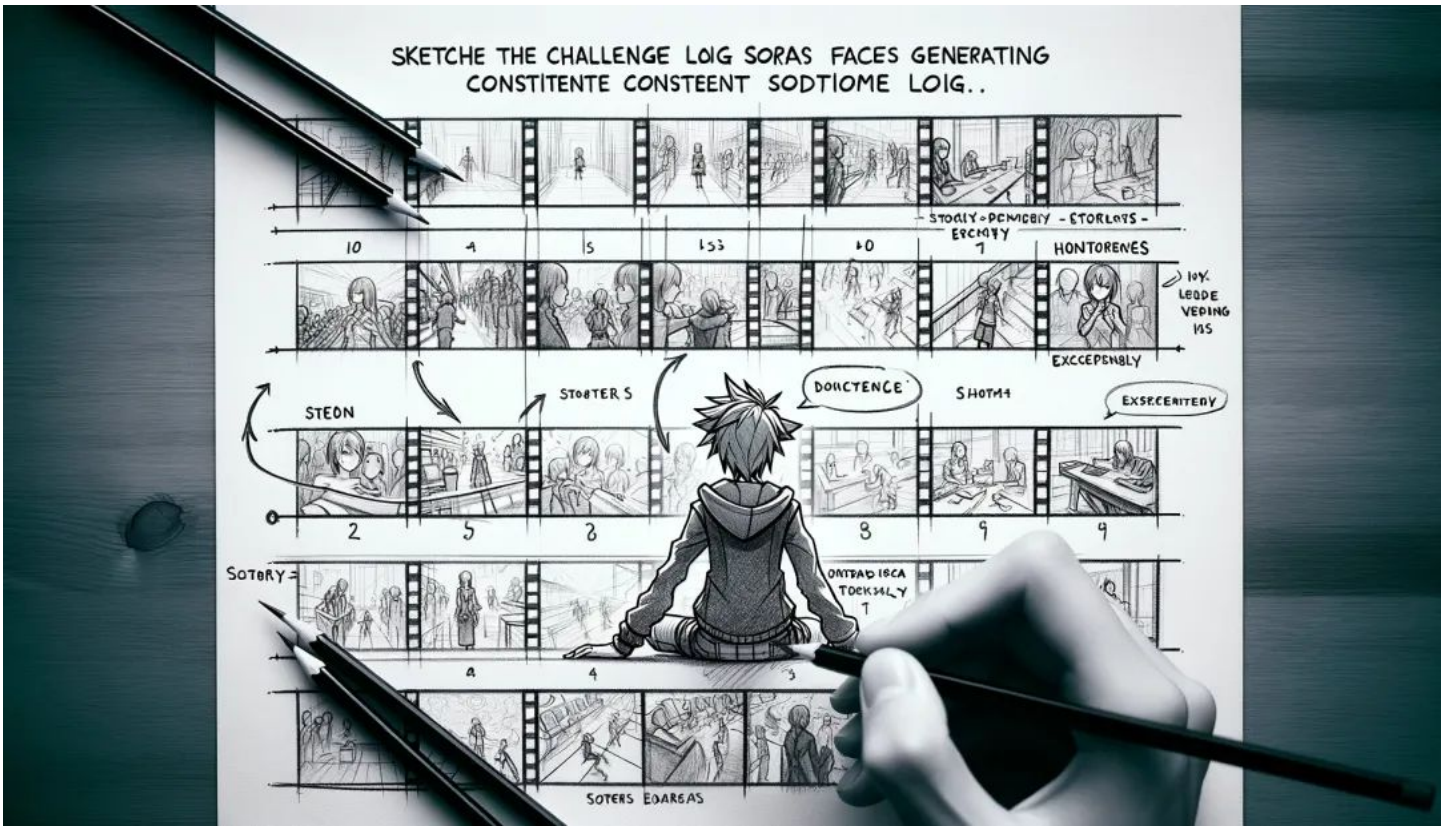
碎碎的杯杯

克服挑战的策略：

- **扩大训练数据集**：集成更多包含复杂物理互动的高质量视频数据，以丰富Sora学习的样本。
- **物理引擎集成**：在Sora的框架中集成物理引擎，让模型在生成视频时能参考物理规则，提高物理互动的真实性。

长视频生成的困难

Sora在生成长时间视频时面临的另一个挑战是如何保持视频内容的长期一致性。对于较长的视频，维持人物、物体和场景的连续性和逻辑一致性变得更加困难。Sora有时可能会在视频的不同部分产生矛盾，例如，人物的衣着突然变化，或是场景中物体的位置不一致。



克服挑战的策略：

- **增强时间连续性学习**：通过改进训练算法，增强模型对时间连续性和逻辑一致性的学习能力。
- **序列化处理**：在视频生成过程中，采取序列化处理的方法，按照时间顺序逐帧生成视频，确保每一帧都与前后帧保持一致性。

准确理解复杂文本指令

虽然Sora在理解简单的文本指令并生成相应视频方面表现出色，但对于复杂的、含有多重含义或要求精确描绘特定事件的文本指令，模型有时会遇到困难。这限制了Sora在更加复杂创意内容生成上的应用。

克服挑战的策略:

- **改善语言模型**：提升Sora内嵌的语言理解模型的复杂度和准确性，使其能够更好地理解和分析复杂文本指令。
- **文本预处理**：引入先进的文本预处理步骤，将复杂的文本指令分解为简单的、易于模型理解的多个子任务，逐一生成，最后综合为完整视频。

训练与生成效率

Sora作为一个高度复杂的模型，其训练和视频生成的时间效率是一个不容忽视的挑战。高质量视频的生成通常需要较长的时间，这限制了Sora在实时或快速反馈场景中的应用。

克服挑战的策略：

- **优化模型结构**：对Sora的架构进行优化，减少不必要的计算，提高运行效率。
- **硬件加速**：利用更强大的计算资源和专门的硬件加速技术，缩短视频生成的时间。

总的来说，Sora在视频生成和模拟真实世界互动方面的表现虽然已经很出色，但仍然存在诸多挑战。通过上述策略的实施，我们有理由相信，未来Sora能够在保持创新的同时，克服当前面临的局限性，展现出更加强大和广泛的应用潜力。

作者：金色传说大聪明

我可聪明了！

原文：https://mp.weixin.qq.com/s/KUnXlDlg-Rs_6D5RFpQbnQ