视频补全技术的十字路口：传统光流传播与生成式扩散模型的深度解析

在影视后期制作、短视频编辑以及计算机视觉研究中，视频补全（**Video Inpainting**）是一项至关重要的技术。它的主要目标通常分为两类：物体移除（**Object Removal, OR**）——把画面中不想要的物体抹掉，仿佛它从未存在过；以及背景恢复（**Background Restoration, BR**）——把缺失或破损的视频区域连贯地修补回来。

这项任务的核心挑战可以用八个字概括：空间合理，时间一致。为了达成这个目标，当前的学术界和工业界演化出了两条截然不同的技术路线：以 ProPainter 为代表的传统光流补全，以及以 DiffuEraser 为代表的生成式（扩散模型）补全。

一、传统光流补全：勤恳的“像素搬运工”

以 ProPainter 为代表的传统方法，其核心机制是**“搬运与传播”**。

在视频的时空连续性中，一个物体在当前帧挡住的背景，很可能在上一帧或下一帧中是暴露出来的。传统模型利用光流（Optical Flow）技术，在时间轴上精准寻找这些“曾经露面”的真实像素，并将它们对齐、传播到当前帧的缺失区域。

- 优势（高上限）：当“时间可见性”高时（即被遮挡的像素在其他帧中大量存在，例如相机跟随运动、物体快速移开），这种方法几乎完美。因为填补进去的都是真实的原始像素，所以画面无比逼真，且时序上极其连贯。
- 劣势（清晰的下限）：它极其依赖“时间可见性”。如果一个背景区域从头到尾都被遮挡（极低可见性），模型在其他帧里找不到可以搬运的像素，就会彻底抓瞎。此时，它只能无奈地将遮挡物边缘的像素向内拉伸，导致修复区域出现严重的模糊伪影。

二、生成式扩散模型：充满不可控创造力的“画师”

随着扩散模型（Diffusion Models）的崛起，以 DiffuEraser 为代表的生成式视频补全应运而生。它的核心机制是**“无中生有”**——不再苦苦寻找真实像素，而是依靠模型庞大的先验知识，根据上下文重新“画”出缺失区域的像素。

然而，尽管生成式模型在图像领域大放异彩，但在视频修复任务中，它却面临着几个核心的系统性错配：

1. 文本提示（**Prompt**）的语义困境与实操痛点：Prompt 作为条件信号天然是全局性、粗粒度的（比如“一条阳光明媚的公园小路”），它难以精确控制局部被遮挡区域所需的纹理延伸或光照渐变。
   工程实战经验：在实际的代码部署中，我们通常可以通过 CLI 命令行来控制是否加入 Prompt 进行监督生成，这在 OR 和 BR 任务中都适用。但大量实操表明：对于物体移除（

OR）任务而言，不加 **Prompt** 往往比加 **Prompt** 效果更好。> 究其原因，加入了文本引导后，模型获得的信号过于宽泛，不仅无法对局部修复提供有效指导，反而容易引发不连贯的幻觉；而如果不加文本引导（Guidance scale 设为零，即空 Prompt），让扩散模型单纯依赖 Mask 周围的上下文特征进行无条件去噪，反而能更自然地填补缺失区域，融合度更高。

2. 掩膜（**Mask**）的双刃剑效应：Mask 本身作为一个独立通道送入网络，携带了强烈的物体轮廓先验。特别是在物体移除（OR）任务中，Mask 就是目标物体的精确轮廓。扩散模型会隐式地“感知”到这个特征，倾向于在区域内生成一个与原物体形状一致的伪影。你想让物体消失，模型却顺着 Mask 画了个“幽灵”留在原地。
3. 时空异质性的感知盲区：扩散模型缺乏对时序维度的感知能力。在时间可见性高的区域（如快速运动物体露出的背景），扩散模型的强行介入反而破坏了光流本已恢复的精确像素；而在可见性极低的区域，又因缺乏语义引导无法生成合理内容。这种**“该保守时过度创造，该创造时能力不足”**的矛盾，本质上源于现有框架对时空特性的无感。

三、寻找破局点：生成式模型的“最佳工作区间”

既然生成式模型有这么多痛点，它在什么场景下才能发挥真正的价值呢？答案是：物体缓慢运动 **+** 相机静止 的 **OR** 任务。

这个特定的场景恰好避开了 DiffuEraser 的所有弱点，并精准命中了 ProPainter 的死穴：

- **ProPainter** 的溃败：相机静止且物体移动慢，意味着同一块背景被连续遮挡了几十帧。ProPainter 找不到源像素，只能生成模糊的废片。
- **DiffuEraser** 的成功密码：
    - **Mask** 无干扰：缓慢移动让 Mask 形状在帧间平滑过渡，边缘背景连续性高，模型不容易被轮廓误导。
    - 完美的**“低频先验 (Priori)”**：ProPainter 在这种场景下虽然效果差、画面模糊，但保留了大致正确的颜色分布和低频结构。扩散模型恰好最擅长在正确的低频基础上补充高频细节和纹理，这比从纯噪声零基础生成（无先验）要稳定得多。
    - 时序压力骤降：固定机位（相机静止）说明这是一个固定视角的场景，背景（往往是草地、墙壁、水面等）本身在帧间几乎不变。扩散模型生成这类重复纹理的能力极强，且不需要精确的语义理解，大大降低了其最头疼的“时序不一致（闪烁）”问题。

结语

目前的视频补全技术并没有完美的“银弹”。传统光流模型是**“稳健的搬运工”，上限高且下限清晰；而生成式模型则是“不可控的画师”**，创造力是一把双刃剑（这也是为什么在工程实操中，往往需要通过去掉 Prompt 来压制其不可控的幻觉）。目前，生成式模型最舒服的工作区间，依然是建立在“低复杂度纹理 + 有模糊但方向正确的先验 + 低时序一致性压力”的极窄窗口内。未来的技术演进，必将是如何更优雅地将光流的时序稳定性与扩散模型的生成能力深度耦合。