改变什么，保留什么：用于视频重构的经验演化智能体

Shiying Zhang[1], * Yuxia Zhu[2], * Huaze Xue[3], *Shengqi Dang[1] Yizhen Wu[1] Yijie Li[4]
[1]同济大学 [2]北京大学 [3]清华大学[4]上海财经大学计算与经济跨学科研究教育部重点实验室

*这些作者对本研究的贡献相同。

arXiv:2611.32196v2 [cs.CV] 25 Nov 2026

摘要

视频重构旨在将现有视频中的实体和环境转化为全新的视觉世界，同时保留其核心动作与交互。然而，要实现连贯的转化仍然面临挑战，因为协同编辑可能会破坏原有的事件结构。本文提出 *RecastAgent*，这是一种在模型权重冻结的条件下运行、由经验引导的源视频条件化视频重构智能体框架。我们的框架采用经验引导规划器，在定性视觉评估的指导下选择编辑提示词和输入视频状态，使智能体能够在已有进展的基础上继续推进，或回退以保留源视频结构。在各项任务中，经验积累模块会将已完成的执行轨迹提炼为带条件且与证据关联的原则，记录编辑策略及其适用范围。此外，我们构建了一个基准，其中包含 *67* 个源视频，以及涵盖六个创意类别的 *134* 项成对规格任务，用于评估。实验表明，在 *Bernini* 和 *Ditto* 上，*RecastAgent* 的表现优于单次处理基线和智能体式重写基线。此外，累积的经验还能提升模型在留出测试视频上的表现，并以零样本方式迁移至 *FiVE-Bench* 和 *VE-Bench* 等外部基准。

1. 引言

视频广泛应用于叙事、广告和创意媒介。创作者常通过剪辑和视觉变换，重新利用现有素材来传达新的叙事。然而，看似简单的创意请求——例如在截然不同的艺术媒介或世界中重新构想一个标志性动作——实际实现起来颇具挑战，因为这需要在保持时间一致性的同时，协调改变多个视觉元素。我们将这项任务称为视频重构，旨在重建并重新诠释现有素材中的实体和环境，将其置于一个新的视频世界中，同时保留其底层事件结构，
包括关键动作、交互和时空关系。从本质上说，视频重构改变的是事件展开的世界，而不是事件本身。

以往的研究主要从两个方向探索视频编辑。基于扩散模型的编辑器和指令驱动的编辑器能够实现多样的视觉变换 [1, 5, 9, 28, 33, 38],，而基于智能体的方法则通过规划和视觉反馈，将用户意图转化为编辑操作 [7, 16, 21, 22, 29, 42]。然而，在复杂的视频重构中，这些方法仍面临挑战，因为这需要在保留原始动作、交互和空间关系的同时，连贯地变换多个实体及其环境。这些要求彼此紧密关联：变换后的实体不仅必须符合目标世界，还必须与原有运动及周围环境保持一致。因此，现有方法可能会产生不完整的变换、错误的实体—属性关联，或对原始动作和交互造成意外改变。因此，在保留源视频底层结构的同时实现连贯的世界变换，仍然是一项挑战。

一个颇具前景的方向是构建精确的多模态提示——结合文本和视觉参考——以指定目标变换，同时保留原始动作与交互。通过视觉反馈进行迭代式精炼，可以借助测试时扩展进一步改善结果；然而，这一过程需要大量反复试错，而获得的洞见通常仍局限于单个任务。这会导致重复探索，也促使我们将编辑轨迹提炼为可复用的知识，以支持未来的重新构图决策。

为此，我们提出了**RecastAgent**，这是一种免训练智能体框架，可在迭代的视频重构轨迹中积累并重用经验。它包含三个核心组件：（1）一个经验引导规划器，



图 1。智能体视频创作与源条件重制。VideoWeaver [37] 组合视频生成工具。Direct 和适配后的 Aurora [42] 分别只执行一次编辑；PE Loop 在返回源视频的同时修订提示词。RecastAgent 在独立 Critic 的指导下选择视频状态和提示词，并参考冻结的设计参考与经验。分支图为示意图；图示示例未运行 PE。所有模型权重均保持冻结。

它利用视觉参考和先验经验来制定提示词，并在每个步骤中选择输入视频状态（源视频或中间结果）；（2）一个视觉评审器，用于根据用户请求评估输出，以指导优化和最终选择；（3）一个经验积累模块，将已完成轨迹提炼为可复用策略——包括适用条件和视觉证据——并存储在外部库中，以指导未见过的任务。

为评估 RecastAgent 的有效性，并区分任务内规划与跨任务经验累积的贡献，我们构建了一个视频重制基准，涵盖六个创意类别中的 67 个源视频，作为现有视频生成与编辑基准的补充 [3, 8, 14, 20, 32]。使用两种视频编辑器 Bernini [33]和 Ditto [1], 开展的实验表明，与现有基线方法相比，RecastAgent 取得了更优的结果。此外，累积经验能够提升测试集上的性能，并实现对 FiVE-Bench [14] 和 VE-Bench [32]的零样本泛化。

我们的主要贡献总结如下：

• 我们将视频重构 定义为一项结构化的视频适配任务，并提出 RecastAgent——一种智能体框架，通过规划和视觉反馈协调迭代编辑决策。

•我们提出了一种经验积累机制，将已完成的编辑轨迹提炼为可复用、基于证据的知识，从而在无需模型训练的情况下实现跨任务改进。

• 我们建立了一个视频重构基准，包含成对的创意指令和受控评估协议。大量实验表明，我们的框架有效，且累积经验能够迁移至未见过的视频和基准。

2. 相关工作

视频编辑与智能体创作。 </style}单次且无需训练的适配方法支持视频编辑修改 [11, 38]；反演、注意力操控和特征传播可提升对应度 [2, 5, 18, 28]。时间约束 [4, 10, 15] 以及结构或关键帧条件[12, 25, 40, 44] 有助于保留内容。统一编辑器包括 VACE、Ditto 和 Bernini [1, 9, 33]。我们将编辑器作为冻结的黑盒工具使用，而不修改其权重。Videodirectorgpt 和 Videostudio 可规划多场景生成 [16, 22]；Genmac 和 Vista 则运用协作与反馈 [7, 21]。Rewriting Video、Aurora 和 Editduet 涉及视频重创作或工具辅助编辑 [29, 35, 42]。此外，RecastAgent 还考虑已保存的视频状态和跨任务建议。

自我改进智能体。 React、Self-refine 和 CRITIC 研究推理与改进 [6, 24, 41]。Reflexion 和 ExpeL 保留以往经验 [30, 45], , 而生成式智能体和 MemGPT 维护持久记忆 [26, 27]。Voyager、SkillWeaver 和 Agentic Context Engineering 积累可执行或上下文知识

[34, 43, 46]。视觉领域的方法包括 VideoWeaver、CineForge 和 Video-RSI [17, 23, 37]；PAST-Bench 和 MetaSkill-Evolve 探究经验驱动的改进 [36, 39]。我们的可复用单元是带有轨迹证据的条件式编辑建议，有别于针对当前任务的 Critic 评审反馈。

基准。视频生成基准衡量感知、语义、时间和构图质量 [8,13, 19, 20, 31, 47]。IVEBench、FiVE-Bench 和 VE-Bench 评估指令引导编辑 [3, 14, 32]。我们的配对规范基准分别评估请求完成度、所需的源对应性以及非请求损伤。

3. RecastAgent

给定源视频 V_0 和请求 u，任务是在保留请求仍然要求的动作和关系的同时，构建一个全新的视觉世界。我们将其写作

$$\hat{V} = \mathcal{A}(V_0, u; \mathcal{K}, E, G_\theta), \tag{1}$$

其中 $\mathcal{K}$ 包含设计参考，E 是固定的经验素材库，而 G_θ 是冻结编辑器。两种外部知识源都不代表对当前生成结果的观察。图 2分别展示基础定位、在线编辑和仅开发阶段的累积。

3.1. 经验引导规划

在第 t 次尝试中，规划器选择输入 X_t 和提示词p_t:

$$(X_t, p_t) = \pi_\phi(V_0, u, M_t, H_{t-1}; \mathcal{K}, E). \tag{2}$$

规划器的权重 ϕ 已冻结。M_t 包含当前观察和工作笔记；H_{t-1} 记录较早的输入、提示词、输出、失败情况和 Critic 反馈。这些记录会在一次运行中不断更新；$\mathcal{K}$ 和 E 则不会。X_t 必须是源，或是在先前尝试中成功保存的候选项。首次尝试时，只有源可用。

基础定位与检索。规划器会区分请求进行的变换，与应保持可辨识的源内容角色、动作和关系。例如，将一位咖啡馆作家重新构想为海盗船长，会改变场景和外观，同时保留写作这一活动，除非请求允许改做其他活动。观察工具会返回原生帧、时间区间和空间裁剪。检索工具会读取设计参考和经验条目；记录工具会检查溯源信息并保留备注。这些动作不会消耗生成尝试。补充材料 A 展示了动作空间。设计参考描述目标世界中的外观和材料。经验条目则会记录先前的编辑建议、条件和支持性证据，供后续使用任务。检索到的建议可以为决策提供参考，但不能证明当前候选项一定成功。

3.2. 反馈驱动的迭代编辑

冻结的编辑器接收选定的输入和提示词:

$$V_t = G_\theta(X_t, p_t; \xi_t), \tag{3}$$

采样配置 ξ_t 由运行协议固定。成功的候选项可用于继续编辑；受损的候选项则可能促使系统从源重新开始，或返回较早的状态。提交失败会消耗预算，但不会生成可供选择的视频。

成功生成后，独立的视觉评审器会比较源、请求、选定的输入、提示词和候选项。它可以检查原生帧和裁剪区域，并就请求的更改、保留的关系和缺陷提供定性反馈。这是当前任务证据，而非先验经验或标量奖励。规划器会将其连同尝试溯源记录在历史记录中。属性绑定错误可能需要新的提示词；道具丢失或交互中断则可能需要更换输入状态。预算耗尽后，规划器会根据记录的证据选择一个成功保存的候选项。所选项不一定是最新生成的结果。中断和未交付的运行会单独报告（补充材料 C），不计入已完成的四候选运行。

3.3. 跨任务经验积累

对于开发任务 k，积累过程会审查其保存的轨迹 $\mathcal{T}_k$，检查检索到的建议是否适用，并提炼出带有证据链接的条件性贡献 Δ_k。每次轨迹记录和提炼过程都会读取相同的初始库 E_0，而不会读取其他并发任务的更新。贡献会汇总为

$$E_1 = \text{Consolidate}\left(E_0, \{\Delta_k : k \in \mathcal{I}_{\text{dev}}\}\right). \tag{4}$$

此处的 $\mathcal{I}_{\text{dev}}$ 仅包含开发任务。汇总会更新外部建议，而不会更新模型权重。积累后的素材库 E_1 会在测试前冻结；测试任务笔记和贡献不会合并到测试快照中。两个规划器配置都保留相同的初始精选知识，因此这并不是与空素材库进行比较。

4. 基准测试与评估

4.1. 配对规范与划分

我们的基准包含六个创意类别中的 67 个源视频（图3）。每个视频都配有一条简短请求，其中视觉设计选择保持开放；另有一条详细请求，具体说明环境、服饰、材料、灯光或交互，共计 134 个条件。正式冻结请求的平均长度分别为 15.2 和 62.7 个词（图5）；标点不计入



图 2. RecastAgent 工作流程。源观察结果和冻结参考资料输入规划器。规划器会选择输入和提示词，记录成功候选项，并接收独立的视觉批评。提交生成的总预算为四次，而不是每个动作四次。交付内容从已保存候选项中选出。只有已完成的开发轨迹才会提供经验贡献；在留出测试之前，整合库会被冻结。

词数，内部连字符和撇号仍视为词的一部分。每个输出均根据其对应的请求进行评审，而非替换后的提示词版本。相关片段组成 56 个源互斥组。40 个开发源（31 组、80 个条件）提供经验贡献；27 个留出源（25 组、54 个条件）则不提供。图4 展示了源 029 的规格轨迹。补充材料 A 展示了来自归档运行记录的不同匹配示例，而非因果性的提示词长度测试。

图 4 在两种规格中均保留了表演者、吉他和音乐动作；详细版本还增加了服装、庭院、装饰和灯光。仅有一个可辨认的动画表演者，并不能证明场景已完整呈现。因此，我们将完成度和对应度与视觉完成度分开评判，并在多个时刻进行评估。

4.2. 指标与评审者

指令诊断分别衡量显式请求完成度、必要的源对应性以及非请求所致的损坏。对应度取决于请求，而非对像素的逐一保留。详细请求还会增加实体—属性绑定和联合满足度；简短请求则会增加风格实现、场景特定设计和有效创造力。五轴质量以 0 至 4 分评定目标实现 （TR）、内容对应度（CP）、时间一致性 （TC）、视觉完整性（VI）和审美协调 （AC）。稳定但未转换的区域可以获得 TC 分，但不能获得 TR 分。完整锚点和有效分母规则见补充材料 D。

GLM-5.3-Flash 和 Claude-Sonnet-5.5 独立地

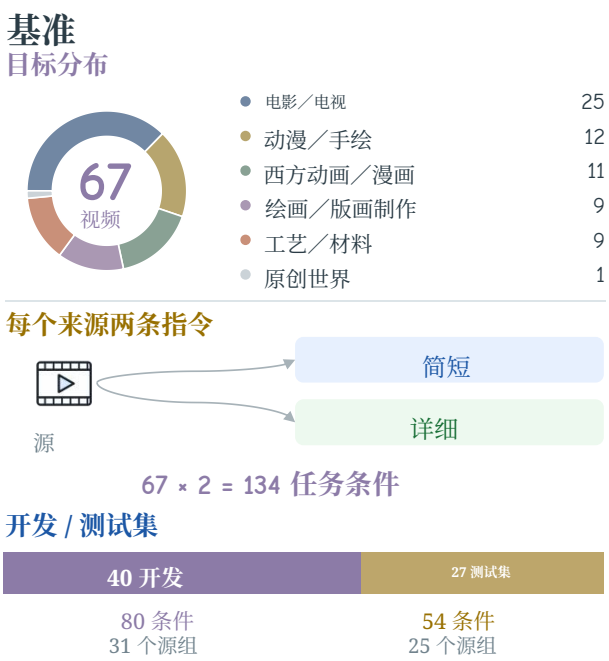

图3。67源基准。六个目标类别与配对的简短/详细指令构成134种条件。开发集和测试视频按相关来源组划分。

使用源派生帧计划和连续片段证据（补充材料B）评审匿名化输出。执行提示词、方法身份和在线批评者评审均被隐藏。成对任务完成度会进行两次评审，



图 4。开发源 029 的配对规范，目标是实现《寻梦环游记》风格的动画表演。详细请求增加了服装、庭院、横幅、暖色窗户和黄昏元素。该插图用于说明规范的细致程度；它并非来自经审计的 PE队列的选定输出比较，也不是经过验证的体验使用轨迹。补充图S2展示了来自归档运行记录的其他匹配示例。

并交换呈现顺序。我们仅针对 n 顺序一致配对报告胜率、平局和负例。其余比较均视为未决，而非平局。组加权得分界限会保留未决和缺失的结果。五个维度的平均值先在每个来源内对有效的简短评分和详细评分取平均，再对有效来源赋予相同权重。未知评分不纳入计算，绝不记为零。

5. 实验

我们的实验旨在回答三个主要研究问题：

- **RQ1（智能体工作流与单次处理及仅提示词迭代的比较）：** 在冻结编辑器和匹配的生成预算下，与单次重写和仅源提示词迭代相比，结合设计先验与视觉批评的轨迹分支是否能改善视频重构？
- **RQ2（约束满足与开放式设计自由）：** 在保留所需源对应关系的同时，RecastAgent 在详细指令下执行细粒度指定编辑，以及在信息不充分的简短指令下合成连贯视觉设计的效果如何？
- **RQ3（跨任务自我改进与零样本迁移）：** 从 Agent 自身开发轨迹中提炼的经验，是否会因果性地改善其在未见测试视频上的编辑决策，并能否零样本泛化至外部公开基准？

5.1. 实验设置

冻结编辑器与对比基线。 我们使用两个不同的开源视频到视频编辑器，在完全相同的冻结检查点和采样计划下（表 S1、附录A），对所有134个基准条件中的所有方法进行评估：Bernini [33]和 Ditto [1]，。我们比较五种配置：(1) 直接 ($B = 1$)，将原始用户请求和源视频直接输入编辑器；(2)*Aurora*

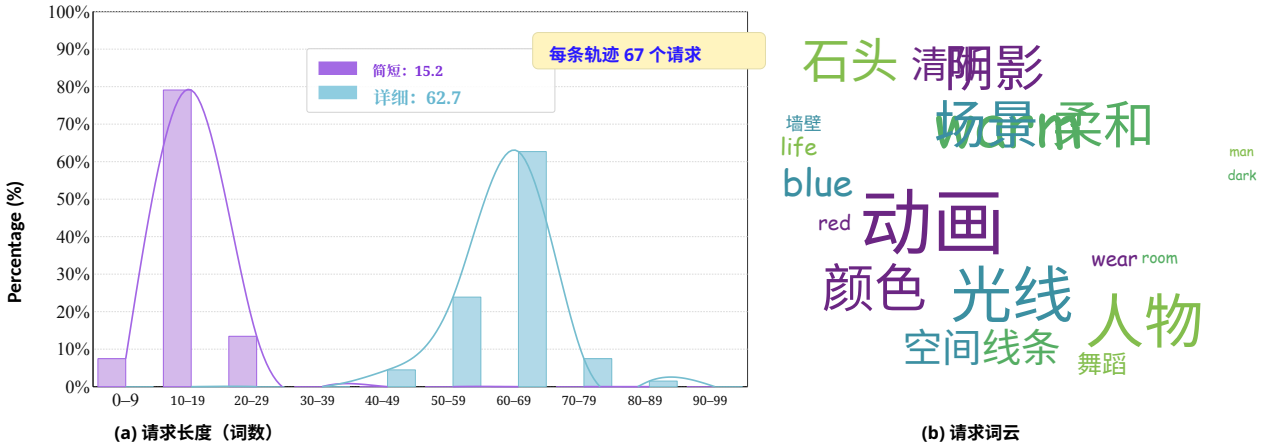

图5. 134 个冻结基准请求的描述性统计。每个长度分布在每个 10 词区间内包含 67 个请求；曲线仅用于直观呈现这些频率。词云总结了两条轨道中移除停用词后的实词频率。它描述的是请求词汇，而非输出质量。

Agent [42] ($B = 1$)，一种开源多模态编辑智能体，使用其官方 Qwen3-VL-8B 规划器和发布的 LoRA 适配器，将源视频和用户简述改写为结构化编辑提示；(3)*PE Loop* ($B = 4$)，一种多轮提示工程基线方法，在四轮生成过程中利用源帧和独立的 Critic 评审反馈迭代优化提示词，同时始终重新进行编辑V_0，随后使用视觉语言模型选择器；以及 (4) 规划器 (E_0)和 (5) 规划器 (E_1) ($B = 4$)，运行带有设计先验的 RecastAgent $\mathcal{K}$、动态状态分支，以及初始经验 E_0 或自累积的经验快照E_1。

骨干模型与预算控制。 为确保多轮方法之间的公平比较，在线规划、提示增强（在 PE Loop 中）、视觉批评、候选选择和任务后反思均使用具有高推理力度的 GPT-6-Luna，而 Aurora Agent 使用其官方经过微调的 Qwen3-VL-8B-Instruct 检查点与 LoRA。多轮方法（$B = 4$）在生成轮次 $t = 1$、.....、4 中使用固定随机种子 42–45，单次处理方法（$B = 1$）使用随机种子 42。生成预算 B 是唯一的在线停止准则。每条开发轨迹的任务后反思最多持续 4 小时、进行 64 次模型调用并使用 600 万词元（最多两次尝试）。

评估方案。 为回答 RQ1，我们在全部 134 个条件下比较 Planner (E_0)、Aurora Agent 和 PE Loop，并在 54 个留出测试条件下比较 Planner (E_1) 与两种基线方法。为回答 RQ2，我们分析简短简述和详细需求说明中的指令轨道诊断结果。为回答 RQ3，我们在相同随机种子和模型权重下，在 54 个留出测试条件中直接比较 Planner (E_1) 和 Planner (E_0)，追踪检索到的经验条目如何具体影响分支和提示决策，并使用 Bernini 在两个标准公开基准 FiVE-Bench 上评估零样本迁移 [14] 和 VE-Bench[32] （表2）。盲式成对任务完成度是我们的主要指标，并辅以五个维度的整体质量评分。

5.2. 工作流程质量与优化（RQ1）

在两种编辑器和两位评审者的评估中，我们完整的 Agent 工作流程显著优于单次指令改写（图6，表1）。规划器（E_0）对比 Aurora，在两种编辑器上的胜率约为 80%–90%；在留出测试划分上，规划器（E_1）也保持了这一优异表现。

任务内推理时优化。 迭代规划和多次尝试选择始终优于 First 代输出。对于 E_0 和 E_1，在大多数已解决案例中，最终选定的交付都胜过初始尝试（表1(a)）。关键在于，规划器会积极利用历史记录，而非默认选择最后一次迭代的结果：在两种编辑器的已完成运行中，超过 50% 的情况下会选择较早的候选项（补充材料 C），这验证了灵活保留候选项的必要性。

5.3. 指令诊断（RQ2）

在复杂简述下审查 Bernini Direct 基线会揭示出鲜明的对比：虽然详细需求说明中的单元素完成率达到中等水平，但所有指定约束的联合满足度却骤降至接近于零。这暴露了单次编辑在处理多属性绑定时的根本局限。（详细诊断——

表1。RecastAgent 与基线方法在 Bernini 和 Ditto 上的视频重构结果，评估基于 67 个源视频和 134 个编辑条件。（a）盲式成对任务完成度，报告每组比较中第一种方法的胜平负率。（b）五个维度的质量评分（0–4），基于所有条件下的有效评分按源视频进行宏平均。在每个编辑器内，粗体和下划线数值分别标示各列中的最佳和次佳评分。

（a）任务完成度

评审者	编辑器	比较	胜率（%）	平（%）	负（%）	n
GLM	Bernini	初始/Aurora（全部）	84.6	0.0	15.4	104
		累积/Aurora（测试集）	83.7	0.0	16.3	43
		初始最终/首次（全部）	56.7	22.2	21.1	90
		累积最终/首次（全部）	54.4	20.0	25.6	90
	Ditto	初始/PE（全部）	37.3	10.2	52.5	59
		初始/Aurora（全部）	89.6	1.3	9.1	77
		累积/PE（测试集）	34.6	11.5	53.8	26
		累积/Aurora（测试集）	85.7	0.0	14.3	21
		初始最终/首次（全部）	47.2	33.3	19.4	72
		累积最终/首次（全部）	61.4	27.1	11.4	70
		PE 循环/直接（全部）	81.2	5.8	13.0	69
Claude	Bernini	初始/Aurora（全部）	79.8	0.0	20.2	114
		累积/Aurora（测试集）	84.8	0.0	15.2	46
		初始最终/首次（全部）	63.5	17.3	19.2	104
		累积最终/首次（全部）	53.8	20.2	26.0	104
	Ditto	初始/PE（全部）	48.2	0.0	51.8	85
		初始/Aurora（全部）	83.5	0.0	16.5	109
		累积/PE（测试集）	42.3	0.0	57.7	26
		累积/Aurora（测试集）	85.4	0.0	14.6	41
		初始最终/首次（全部）	57.1	19.8	23.1	91
		累积最终/首次（全部）	57.0	22.6	20.4	93
		PE 循环/直接（全部）	74.5	0.0	25.5	94

（b）五轴评分

		GLM					Claude				
编辑器	方法	TR	CP	TC	VI	AC	TR	CP	TC	VI	AC
Bernini	直接	2.520	3.020	<u>3.130</u>	<u>2.860</u>	2.770	2.178	2.748	2.797	2.365	2.309
	Aurora Agent	2.671	3.032	3.052	2.661	2.895	2.425	2.510	2.705	2.231	2.510
	PE Loop†	2.975	**3.260**	3.117	**2.900**	3.044	2.602	**3.002**	2.796	**2.454**	2.685
	规划器初始值	**3.165**	<u>3.226</u>	**3.205**	2.798	<u>3.152</u>	<u>2.827</u>	<u>2.787</u>	**2.844**	<u>2.406</u>	<u>2.808</u>
	规划器累积值	<u>3.137</u>	3.170	3.099	2.798	**3.202**	**2.884**	2.739	<u>2.828</u>	2.376	**2.827**
Ditto	直接	3.080	3.116	3.340	<u>3.049</u>	3.376	2.924	2.778	2.978	2.496	3.022
	Aurora Agent	2.980	3.037	3.303	2.867	3.340	2.796	2.739	2.947	2.478	2.969
	PE Loop	3.314	3.232	<u>3.367</u>	2.950	3.427	3.103	<u>3.015</u>	**3.015**	**2.600**	**3.145**
	规划器初始值	<u>3.354</u>	**3.288**	3.329	3.016	**3.462**	**3.164**	**3.016**	2.967	2.547	<u>3.133</u>
	规划器累积值	**3.360**	<u>3.279</u>	**3.458**	**3.106**	<u>3.445</u>	<u>3.115</u>	3.012	<u>2.995</u>	<u>2.565</u>	3.131

n 统计顺序一致配对；测试集 表示留出条件。Final/First 比较从四次尝试中选出的输出与首次生成的结果。† Bernini PE 循环使用混合评估队列。

断和基线瓶颈已整理于补充材料 D 中）。

5.4. 留出经验与迁移（RQ3）

在保留的测试条件下，将规划器（E_1）与规划器（E_0）进行比较，结果显示改进效果取决于评审者。GLM 更偏好累积库（胜率为 60%–67%），而 Claude 对两种配置的评价几乎相同。尽管定量识别界限仍具有包容性，但定性轨迹追踪证实，检索到的经验能够有效指导结构决策——例如映射中心元素、统一整体风格设计，以及选择更优的早期候选项（补充材料E）。

零样本外部迁移。 在外部基准（FiVE-

表2. 在 FiVE-Bench [14] 和 VE-Bench [32]的分层36案例子集上使用 Bernini 进行外部评估。累积库在不进行基准特定更新的情况下重用。峰值信噪比衡量未编辑区域的分贝值；FiVE-Acc 和胜率以百分比表示。胜率将各方法与直接方法进行比较。粗体/下划线标示每列中的最佳/次佳结果。这些比较的是完整配置，而非孤立经验效应。

	FiVE-Bench [14]			VE-Bench [32]	
方法	峰值信噪比 ↑	CLIP-T ↑	FiVE-Acc ↑	问答得分 ↑	胜率（%） ↑
直接	24.61	26.95	<u>22.1</u>	<u>0.067</u>	–
Aurora Agent [42]	**25.80**	<u>27.27</u>	18.6	0.037	<u>42.9</u>
规划器累积值	<u>25.51</u>	**27.68**	**26.4**	**0.199**	**57.1**

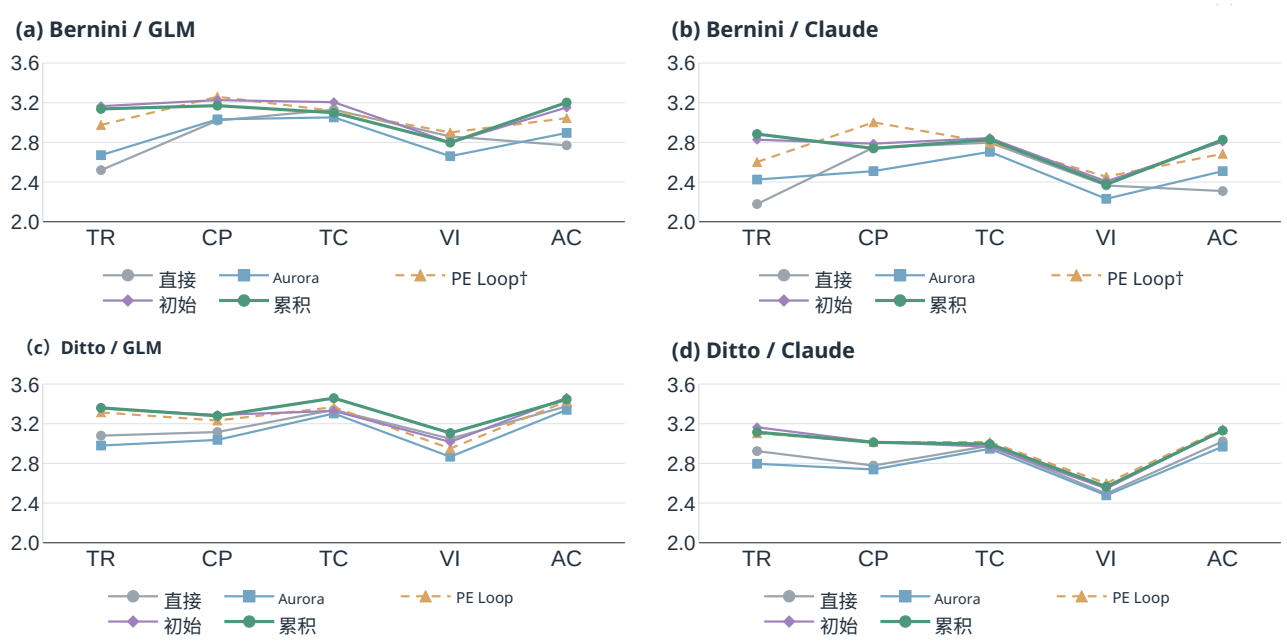

图6. 测得的五轴质量，复现表1(b) 中的全部 100 个数值。评分采用 0–4 分；每个面板均明确缩放至 2.0–3.6。初始和累积分别表示两个规划器库。Bernini PE† 保留了历史模型/输入分层。全队列的描述性均值并不能证明留出集上的经验增益或因果经验增益。

Bench [14] 和 VE-Bench [32]）上对完整的规划器（E_1）流程进行评估，且不针对特定基准进行调优，展现出强大的零样本泛化能力。规划器（）在文本对齐、编辑准确度和视觉质量指标方面均优于单次处理方法和基线替代方案（表2），证实了该智能体框架广泛的可迁移性。

6. 讨论与局限

工作流程与经验增益。 将规划器（E_0）与基线方法进行比较，评估的是完整的智能体工作流程；而对比 E_1 与 E_0，则是在固定权重和随机种子的条件下，隔离出自蒸馏经验带来的因果增益。

单批次与持续演进。 我们的冻结快照（$E_0 \to E_1$）可避免测试时泄漏和顺序依赖。要扩展为持续演进，需要进行记忆整合，以精简冗余启发式规则并解决冲突规则。

基准规模与校准。 每种条件仅使用一条轨迹。尽管采用交换顺序的视觉语言模型评判可缓解位置偏差，但人类感知校准以及源/示例归因仍是未来工作的重要方向。

7. 结论

我们提出了 RecastAgent，这是一种用于源条件视频重构的经验演化智能体，在冻结权重的情况下，将任务内轨迹分支与跨任务知识蒸馏融为一体。具体而言，RecastAgent 在定性视觉批评的引导下，联合选择候选视频状态和提示词，以便在基于已有的部分进展

与为保持结构忠实度而回溯之间取得平衡，同时通过反思已完成轨迹，蒸馏出有条件且与证据关联的原则。我们在涵盖 Bernini 和 Ditto 的 67 个源（134 种条件）成对规范基准上进行评估，并在 FiVE-Bench 和 VE-Bench 上开展零样本测试。结果显示，RecastAgent 显著优于单次处理和智能体式改写基线方法，证明定性编辑经验能够因果性地改善模型对未见视频所作的决策。

大语言模型使用声明

ChatGPT（OpenAI）用于语法纠正、措辞润色，以及检查手稿的一致性和参考资料。作为实验工作流程的组成部分，GPT-6-Luna 用于规划、提示词润色、视觉批评、候选选择和任务后反思。Aurora 基线使用 Qwen3-VL-8B-Instruct 及其发布的 LoRA 适配器进行提示词重写。Bernini 和随 Ditto 发布的视频编辑模型用于生成实验视频输出。GLM-5.3-Flash 和 Claude-Sonnet-5.5 作为自动评估器。这些使用方式及其配置在方法和实验部分中进行了说明。作者设计并监督了本研究，审查并核实了最终手稿及所报告的分析结果，并对工作的准确性、归属、解释和可复现性承担全部责任。

参考资料

[1] Qingyan Bai、Qiuyu Wang、Hao Ouyang、Yue Yu、Hanlin Wang、Wen Wang、Ka Leong Cheng、Shuailei Ma、Yanhong Zeng、Zichen Liu 等。使用高质量合成数据集扩展基于指令的视频编辑。在 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 37971–37981 页，2026年。 1， 2， 5[2] Duygu Ceylan、Chun-Hao P Huang 和 Niloy J Mitra。Pix2video：使用图像扩散进行视频编辑。在 *2023*年 *IEEE/CVF*国际计算机视觉会议（*ICCV*），第 23149–23160 页。IEEE，2023年。 2[3] Yinan Chen、Jiangning Zhang、Teng Hu、Yuxiang Zeng、Zhu- cun Xue、Qingdong He、Chengjie Wang、Yong Liu、Xiaobin Hu 和 Shuicheng Yan。IVEBench：用于评估指令引导视频编辑的现代基准测试套件。在国际学习表征会议，第 137363–137383 页，2026年。 2， 3[4] Yuren Cong 、 Mengmeng Xu 、 Christian Simon 、 Shoufa Chen、Jiawei Ren、Yanping Xie、Juan-Manuel Perez-Rua、Bodo Rosenhahn、Tao Xiang 和 Sen He。Flatten：用于一致性文本到视频编辑的光流引导注意力。在 国际学习表征会议，第 52086–52106 页，2024年。 2[5] Michal Geyer、Omer Bar Tal、Shai Bagon 和 Tali Dekel。TokenFlow：用于一致性视频编辑的一致性扩散特征。在 国际学习表征会议，第 1608–1620 页，2024年。 1， 2

[6] Zhibin Gou、Zhihong Shao、Yeyun Gong、Yujiu Yang、Nan Duan、Weizhu Chen 等。Critic：大语言模型可以通过工具交互式批评进行自我纠错。载于国际表示学习会议，第57734–57811页，2024年。 2[7] Kaiyi Huang、Yukun Huang、Xuefei Ning、Zinan Lin、Yu Wang 和 Xihui Liu。Genmac：通过多智能体协作进行组合式文本到视频生成。载于AAAI人工智能会议论文集，第5049–5057页，2026年。 1, 2[8] Ziqi Huang、Yinan He、Jiashuo Yu、Fan Zhang、Chenyang Si、Yuming Jiang、Yuanhan Zhang、Tianxing Wu、Qingyang Jin、Nattapol Chanpaisit 等。Vbench：视频生成模型的综合基准套件。载于2024年IEEE/CVF计算机视觉与模式识别会议（CVPR），第21807–21818页。IEEE，2024年。 2, 3[9] Zeyinzi Jiang 、 Zhen Han 、 Chaojie Mao 、 Jingfeng Zhang、Yulin Pan 和 Yu Liu。Vace：一体化视频创作与编辑。载于2025年IEEE/CVF国际计算机视觉会议（ICCV），第17191–17202页。IEEE，2025年。 1, 2[10] Ozgur Kara、Bariscan Kurtkaya、Hidir Yesiltepe、James M Rehg 和 Pinar Yanardag。Rave：使用随机噪声洗牌，通过扩散模型实现快速且一致的视频编辑。载于2024年IEEE/CVF计算机视觉与模式识别会议（CVPR），第6507–6516页。IEEE，2024年。 2[11] Levon Khachatryan、Andranik Movsisyan、Vahram Tade-vosyan、Roberto Henschel、Zhangyang Wang、Shant Navasardyan 和 Humphrey Shi。Text2video-zero：文本到图像扩散模型是零样本视频生成器。载于2023年IEEE/CVF国际计算机视觉会议（ICCV），第15908–15918页。IEEE，2023年。 2[12] Max Ku、Cong Wei、Weiming Ren、Harry Yang 和 Wenhu Chen。Anyv2v：适用于任意视频到视频编辑任务的免调优框架。arXiv预印本 arXiv:2403.14468，2024年。 2[13] Baiqi Li、Zhiqiu Lin、Deepak Pathak、Jiayao Li、Yixin Fei、Kewen Wu、Xide Xia、Pengchuan Zhang、Graham Neubig 和 Deva Ramanan。评估和改进组合式文本到视觉生成。载于2024年IEEE/CVF计算机视觉与模式识别会议研讨会（CVPRW），第5290–5301页。IEEE，2024年。 3[14] Minghan Li、Chenxi Xie、Yichen Wu、Lei Zhang 和 Mengyu Wang。Five：用于评估新兴扩散模型和整流流模型的细粒度视频编辑基准。arXiv预印本 arXiv:2503.13684，2025年。 2, 3, 6, 8[15] Xirui Li、Chao Ma、Xiaokang Yang 和 Ming-Hsuan Yang。Vidtome：用于零样本视频编辑的视频令牌合并。载于2024年IEEE/CVF计算机视觉与模式识别会议（CVPR），第7486–7495页。IEEE，2024年。 2[16] Han Lin、Abhay Zala、Jaemin Cho 和 Mohit Bansal。Videodirectorgpt：通过大语言模型引导的规划实现一致的多场景视频生成。arXiv预印本 arXiv:2309 15091，2023年 1, 2 . . [17] Junxiang Liu、Lin Wang、Haiyu Shi、Hongxu Ma、Xiaoyu Yang、Chunjie Chen、Xiaoxiao Xu、Kaiqiao Zhan、Boao Wang、Shuizhou Shi 等。Cineforge：用于长时程视频生成的自我改进智能体。arXiv预印本 arXiv:2608.29621，2026年。 3

[18] Shaoteng Liu、Yuechen Zhang、Wenbo Li、Zhe Lin 和 Jiaya Jia。Video-p2p：使用交叉注意力控制进行视频编辑。载于2024年IEEE/CVF计算机视觉与模式识别会议（CVPR），第8599–8608页。IEEE，2024年。 2[19] Yuanxin Liu、Lei Li、Shuhuai Ren、Rundong Gao、Shicheng Li、Sishuo Chen、Xu Sun 和 Lu Hou。Fetv：开放域文本到视频生成的细粒度评估基准。载于神经信息处理系统进展，36:62352–62387，2023年。 3[20] Yaofang Liu、Xiaodong Cun、Xuebo Liu、Xintao Wang、Yong Zhang、Haoxin Chen、Yang Liu、Tieyong Zeng、Raymond Chan 和 Ying Shan。Evalcrafter：大型视频生成模型的基准测试与评估。载于2024年IEEE/CVF计算机视觉与模式识别会议（CVPR），第22139–22149页。IEEE，2024年。 2, 3[21] Do Xuan Long、Xingchen Wan、Hootan Nakhost、Chen-Yu Lee、Tomas Pfister 和 Sercan ¨ O Arik。Vista：测试时自我改进的视频生成智能体。载于IEEE/CVF计算机视觉与模式识别会议论文集，第6021–6032页，2026年。 1, 2[22] Fuchen Long、Zhaofan Qiu、Ting Yao 和 Tao Mei。Videostudio：生成内容一致的多场景视频。载于欧洲计算机视觉会议，第468–485页。施普林格，2024年。 1, 2[23] Bingjun Luo、Jialin Guo 和 Siqi Li。Video-RSI：通过工具框架演进实现视频理解智能体的递归式自我改进。arXiv预印本 arXiv:2609.37950，2026年。 3[24] Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hal-linan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang 等。Self-refine：通过自我反馈进行迭代式改进。载于神经信息处理系统进展，36:46534–46594，2023年。 2[25] Hao Ouyang、Qiuyu Wang、Yuxi Xiao、Qingyan Bai、Juntao Zhang、Kecheng Zheng、Xiaowei Zhou、Qifeng Chen 和 Yujun Shen。Codef：用于时间一致视频处理的内容变形场。载于IEEE/CVF计算机视觉与模式识别会议论文集，第8089–8099页，2024年。 2[26] Charles Packer、Sarah Wooders、Kevin Lin、Vivian Fang、Shishir G Patil、Ion Stoica 和 Joseph E Gonzalez。Memgpt：迈向作为操作系统的大语言模型。arXiv预印本 arXiv:2310.08560，2023年。 2[27] Joon Sung Park、Joseph O'Brien、Carrie Jun Cai、Mered-ith Ringel Morris、Percy Liang 和 Michael S Bernstein。生成式智能体：人类行为的交互式仿真体。载于第36届ACM用户界面软件与技术年会论文集，第1–22页，2023年。 2[28] Chenyang Qi、Xiaodong Cun、Yong Zhang、Chenyang Lei、Xintao Wang、Ying Shan 和 Qifeng Chen。Fatezero：融合注意力实现零样本文本视频编辑。载于2023年IEEE/CVF国际计算机视觉会议（ICCV），第15886–15896页。IEEE，2023年。 1, 2[29] Marcelo Sandoval-Casta˜neda、Bryan Russell、Josef Sivic、Gre-gory Shakhnarovich 和 Fabian Caba Heilbron。Editduet：用于视频非线性编辑的多智能体系统。载于计算机图形学与交互技术特别兴趣小组会议论文集，第1–11页，2025年。 1, 2[30] Noah Shinn、Federico Cassano、Ashwin Gopinath、Karthik Narasimhan 和 Shunyu Yao。Reflexion：通过语言强化学习实现语言智能体。神经信息处理系统进展，36：8634–8652，2023年。2[31] Kaiyue Sun、Kaiyi Huang、Xian Liu、Yue Wu、Zihan Xu、Zhenguo Li 和 Xihui Liu。T2V-CompBench：用于组合式文本到视频生成的综合基准。在2025年IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第8406–8416页。IEEE，2025年。3[32] Shangkun Sun、Xiaoyu Liang、Songlin Fan、Wenxu Gao 和 Wei Gao。VE-Bench：与主观评价对齐的文本驱动视频编辑质量评估基准测试套件。在AAAI人工智能会议论文集，第7105–7113页，2025年。2，3，6，8[33] Bernini团队、Chenchen Liu、Junyi Chen、Lei Li、Lu Chi、Mingzhen Sun、Zhuoying Li、Yi Fu、Ruoyu Guo、Yiheng Wu 等。Bernini：用于视频扩散的潜在语义规划。arXiv预印本 arXiv:2605.22344，2026年。1，2，5[34] Guanzhi Wang、Yuqi Xie、Yunfan Jiang、Ajay Mandlekar、Chaowei Xiao、Yuke Zhu、Linxi Fan 和 Anima Anandkumar。Voyager：一种基于大语言模型的开放式具身智能体。arXiv预印本 arXiv:2305.16291，2023年。3[35] Sitong Wang、Anh Truong、Lydia B Chilton 和 Dingzeyu Li。重写视频：以文本驱动重新创作视频素材。在第31届国际智能用户界面会议论文集，第1675–1686页，2026年。2[36] Zefeng Wang、Minxi Yan、Jinhe Bi、Sikuan Yan、Volker Tresp 和 Yunpu Ma。MetaSkill-Evolve：通过双时间尺度元技能演化递归式自我改进大语言模型智能体。arXiv预印本 arXiv:2607.05297，2026年。3[37] Jianhui Wei、Yan Zhang、Jie Tan、Hengchuan Zhu、Xiaotian Zhang、Ziyi Chen、Daoan Zhang、Wei Xu、Yeying Jin 和 Zuozhu Liu。VideoWeaver：评估并演进智能体长视频生成技能。arXiv预印本 arXiv:2606.08091，2026年。2，3[38] Jay Zhangjie Wu、Yixiao Ge、Xintao Wang、Stan Weixian Lei、Yuchao Gu、Yufei Shi、Wynne Hsu、Ying Shan、Xiaohu Qie 和 Mike Zheng Shou。Tune-A-Video：通过单次调优图像扩散模型实现文本到视频生成。在2023年IEEE/CVF国际计算机视觉会议（ICCV）论文集，第7589–7599页。IEEE，2023年。1，2[39] Shuhan Xue、Zixin Ding、Yichen Shen、Yinjie Wang、Zhenfei Yin、Yingcheng Wu、Yuxin Chen、Mengdi Wang 和 Ling Yang。PAST-Bench：个人智能体递归式自我改进基础能力的基准测试。arXiv预印本 arXiv:2608.04003，2026年。3[40] Shuai Yang、Yifan Zhou、Ziwei Liu 和 Chen Change Loy。Rerender a Video：零样本文本引导的视频到视频转换。在SIGGRAPH Asia 2023会议论文集，第1–11页，2023年。2[41] Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。React：在语言模型中协同推理与行动。arXiv预印本 arXiv:2210.03629，2022年。2[42] Yongsheng Yu、Ziyun Zeng、Zhiyuan Xiao、Zhenghong Zhou、Hang Hua、Wei Xiong 和 Jiebo Luo。Aurora：利用工具使用智能体实现统一视频编辑。arXiv预印本 arXiv:2605.18748，2026年。1，2，6，8

[43] Qizheng Zhang、Changran Hu、Shubhangi Upasani、Boyuan Ma、Fenglu Hong、Vamsidhar Kamanuru、Jay Rainton、Chen Wu、Mengmeng Ji、Hanchen Li 等。智能体上下文工程：演进上下文以实现语言模型的自我改进。在国际学习表征会议论文集，第86069–86100页，2026年。3[44] Yabo Zhang、Yuxiang Wei、XIAOPENG ZHANG、Wangmeng Zuo、Qi Tian 等。ControlVideo：无需训练的可控文本到视频生成。在国际学习表征会议论文集，第54441–54461页，2024年。2[45] Andrew Zhao、Daniel Huang、Quentin Xu、Matthieu Lin、Yong-Jin Liu 和 Gao Huang。Expel：大语言模型智能体是经验学习者。在*AAAI*人工智能会议论文集，第19632–19642页，2024年。2[46] Boyuan Zheng、Michael Y Fatemi、Xiaolong Jin、Zora Zhiruo Wang、Apurva Gandhi、Yueqi Song、Yu Gu、Jayanth Srinivasa、Gaowen Liu、Graham Neubig 等。Skillweaver：Web智能体能够通过发现和磨练技能实现自我改进。*arXiv*预印本 *arXiv:2504.07079*，2025年。3[47] Dian Zheng、Ziqi Huang、Hongbo Liu、Kai Zou、Yinan He、Fan Zhang、Lulu Gu、Yuanhan Zhang、Jingwen He、Wei-Shi Zheng 等。Vbench-2.0：推进视频生成基准测试套件，以提升内在忠实度。*arXiv*预印本 *arXiv:2503.21755*，2025年。3

《**What to Change, What to Keep: Experience-Evolving Agents for Video Recasting**》

补充材料

A. 编辑工作流与匹配示例

重构指令指定要修改的内容；源视频则确立了编辑后应保留的动作和关系。67 个来源各有一条简短请求和一条详细请求。由此形成的 134 个条件涵盖 40 个开发源和 27 个留出源。相关片段始终划分在同一数据集中：开发集有 31 个相关来源组，测试集有 25 个。

每次编辑从何处开始。直接方法使用原始请求，Aurora Agent 会改写该请求；二者各生成一个视频。PE Loop 会根据 Critic 的评审反馈修改提示词，但四次编辑都从原始源视频开始。规划器也可以继续编辑中间版本。因此，其动作空间（图 S1）将两个选择区分开来：接下来编辑哪个视频，以及交付哪个已保存版本。每次提交生成都会消耗一次尝试，包括生成失败的情况。观察和检索不会改变生成预算。

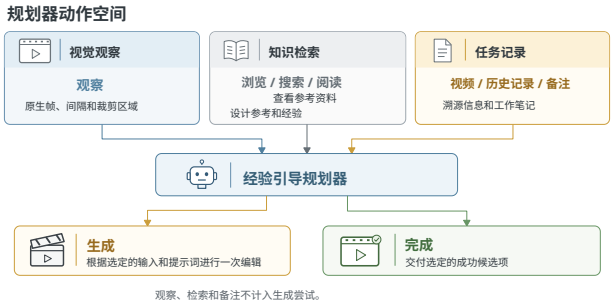

图 S1. 规划器可以检查候选项、检索建议，然后选择继续生成一次编辑或选取一个已保存候选项。

生成设置。多次尝试的方法使用随机种子 42–45；直接和 Aurora 使用种子 42。表S1 列出了两个冻结编辑器的设置。PE Loop 和规划器中的提示修订在调用编辑器之前进行。对于主表 † 1中标记为的 Bernini PE 行，117 个条件使用 Luna/high，17 个使用 Sol/high。在 134 次交付中，126 次使用 24 帧率下覆盖完整时长的 117 帧输入。一次使用了时序相同的源前缀，另有七次使用了原始帧率下长度为 81、97 或 101 帧的原生前缀。

阅读示例。 主图 4比较针对单个源的两个请求。图 S2 在每个案例中固定详细请求，并展示每种工作流程选出的输出。在素材库案例中，比较终端周围的阅读位置。在节庆活动案例中，比较舞者的服饰与绘制的周围环境。这些细节使我们能够查看每次编辑保留了什么，又改变了什么。

表 S1. 冻结渲染器设置。两个编辑器均禁用内置提示增强，并使用空的负面提示词。

场景	Bernini	Ditto
推理步数	40	50
引导	APG（v2v）	CFG 5.0
附加 LoRA	None	系数 1.0
网络精度	bfloat16	bfloat16
变分自编码器 / 分块	32位浮点数 / 默认	分块推理
文本容量	512 词元*	512 词元
最大边长	848 像素	848 像素

*包含官方前缀。Ditto 保留源宽高比，并将尺寸四舍五入为 16 的倍数。

B. 用于评估的视频证据

编辑过程中。 规划器首先查看源的六个概览帧，并接收每个新结果的六帧。Critic 会查看每个相关视频的六帧。两者都可以检查裁剪区域和连续帧（每个视频每页 2–16 帧，每次调用最多 32 帧，帧间有一帧重叠）。PE提示词修订从八个源帧开始；其选择器会接收每个视频的六个概览帧以及候选项评审。

交付后。 GLM-5.3-Flash 和 Claude-Sonnet-5.5 独立将匿名视频与原始请求和源进行比较。对于时长为 D 秒的源，概览会在最多 20，$\lceil 20D/6 \rceil$ 个均匀间隔的时间点进行采样。每个时间点对应最接近的原生帧；重复帧只显示一次。源还会定义连续区间，边界每隔 28 帧设置一次。每个输出都会以其原生帧率，在相同的时间区间内接受检查。超出较短输出时长的时间点会标记为缺失。裁剪区域保持原有宽高比，并缩放至最长边不超过 960 像素（JPEG 质量为 90）。评审者会分别审查每个区间，并结合这些观察与概览给出最终评分。时间一致性要求有有效的连续帧证据。

C. 交付与候选选择

最终交付哪个版本? 规划器运行完成后会返回其选定的候选项。如果编辑过程被中断，规划器会返回最后一个优先选择的候选项；如果尚未表达偏好，则返回最新生成的候选项。PE Loop 使用其最近一次完成的选择；如果没有，则回退到第一个已完成的候选项。未交付视频的运行会单独归类。

在两个编辑器中，规划器在超过半数正常完成的运行中选择了较早的版本（表



图 S2。机械档案馆（053）和手工制作的节庆（059）。每行比较视频时长中相同比例（5% 或 95%）处的源、PE Loop、 E_0 和 E_1 。

S2 ）。这一频率衡量的是编辑决策。主表 1(a) 通过将选定的视频与首次生成的视频进行比较，衡量其视觉结果。在所有八种编辑器–素材库–评审者组合中，选择结果的胜率都高于负率。对于 Ditto E_0，GLM 在 72 对已判定配对中记录了 47.2% 的胜率、 33.3% 的平局和 19.4% 的负率。

表 S2。134 种条件下的规划器交付次数。Early 是正常四候选运行中返回 v_1、 v_2 或 v_3 的子集。

编辑器	素材库	正常	中断	缺失	提前
Bernini	E_0	129	2	3	66
	E_1	128	0	6	71
Ditto	E_0	130	1	3	68
	E_1	128	1	5	69

成对偏好。每位评审者都会对两种呈现顺序进行评分。意见一致时，结果为胜、平或负；正文中的胜率基于这些已判定配对计算。对于按组加权的评分，胜、平、负分别计为 1、0.5 和 0。单独交付计为 1；如果两种方法都未交付，则各计为 0.5。未决判定计入区间$[0, 1]$。我们先在每个相关来源组内取平均值，再对各组进行等权加权。

D. 质量评分与指令诊断

视频评分。 评审者使用表S3中的标准，以十分位为每个维度评分。3 表示存在轻微的次要缺陷；1 表示表面化或碎片化实现。若出现影响重大的核心失败，则受影响维度的评分上限为 2。我们对每个来源内可用的评分取平均值，然后对具有有效评分的来源进行等权重处理。

表 S3。五轴视频质量（0–4）的参考锚点。每个维度评估交付视频的一项独特属性；中间分数遵循附录 D中的规则。

Axis	4：完整	2：部分	0：缺失或失败
TR（目标真实-现）	所有核心请求元素都体现了目标世界或媒介，并贯穿整个视频，包括主体、道具、场景和光线。	目标清晰可辨；中心主体或场景仍保留源的风格，媒介则间歇出现，或变化主要是表面处理。	请求的目标不存在，或与事实相矛盾。
CP（内容对应性）	源内容角色、事件节拍、空间关系以及关键交互在目标世界中都有可识别的对应物。	场景仍然可以被识别，但某个核心角色、节拍或交互丢失、被替换，或变得难以理解。	缺少必需的源对应关系，或与之矛盾。
TC （时间一致性）	目标世界中的对应实体。事件推进连贯；媒介，身份、接触关系和运动保持到结尾的一致性。	尽管存在以下方面的重要中断，该事件仍然可以理解：媒介、身份、物体持续性、接触关系或焦点清晰度。	该序列不连贯或其核心发展脉络会崩溃。
VI（视觉完整性）	主体、物体、边界和接触都符合该媒介的特点，且所有焦点结构均完好无损。	严重缺陷影响焦点面部，手、道具、接触部位或边界，例如融化、融合、斑块或接缝。	关键结构崩塌且核心互动无法读取。
AC （审美协调）	构图、光线、色彩搭配、材料、主体和场景构成统一的设计，适合目标世界，并契合其风格。	当视觉系统相互冲突时，整合性会减弱，或出现未被请求的媒介混合，通用装饰或杂乱的层级结构。	视觉 组织 col-中央场景中的疏漏。
	性。	chy.	

表 S4。按适用性分组的 Bernini 直接指令诊断。单元格报告均值（有效条件数/67）。设计和美学评分采用 0–4 分制；完成度、对应度、绑定、满意度和损坏采用 0–1 分制。

(a) 两种请求轨道共有的指标

	简短请求		详细请求	
指标	GLM	Claude	GLM	Claude
编辑完成度	0.577 (65/67)	0.379 (66/67)	0.711 (66/67)	0.574 (65/67)
必要的对应关系	0.870 (54/67)	0.493 (67/67)	0.754 (63/67)	0.576（66/67）
未请求损坏标记	0.719（64/67）	1.000（49/67）	0.831（65/67）	1.000（38/67）
重大损伤标志	0.094 (64/67)	0.347 (49/67)	0.123 (65/67)	0.289 (38/67)
审美诊断均值	2.623 (66/67)	2.306 (67/67)	3.049 (65/67)	2.739 (66/67)

(b) 短请求设计

指标	GLM	Claude
风格实现	2.203 (66/67)	1.622 (67/67)
场景特定设计	2.288 (66/67)	1.639 (67/67)
有效创造力	2.050 (66/67)	1.524 (66/67)

(c) 详细请求约束

指标	GLM	Claude
绑定准确率	0.796 (62/67)	0.542 (66/67)
联合满足	0.015 (66/67)	0.000 (66/67)

完成编辑与满足整个请求。 对于详细请求，GLM 给出的 Bernini Direct 平均编辑完成度为 0.711，但联合满足度仅为 0.015（表 S4）。后者要求每项适用的编辑修改、绑定和源对应性均通过，且不得造成未请求的损坏。Claude 在全部 38 个具有有效损坏评分的详细输出中都标记出了此类损坏。

诊断分母。检查项评分为 1（通过）、0.5（部分通过）或 0（未通过）。对于总计为 s 项的 m 个已知检查项和 q 个未知检查项，覆盖率为 $m/(m+q)$，均值位于 $[s/(m+q),\ (s+q)/(m+q)]$。点数评分需要 $m>0$ 和 $q=0$。每个单元格给出其在 67 项中的有效计数。审美均值需要全部五个组成部分：叙事契合度、世界一致性、构图与光线、色彩与纹理，以及视觉层次。

E. 在留出任务上的经验迁移

什么发生了变化：E_0 与 E_1？ 两个规划器都从相同的模型权重开始。初始库 E_0 包含 21 条从开发试点中精选的条目。累积快照 E_1 纳入了任一配置在 54 个留出条件上接受评估之前，来自开发轨迹的贡献。因此，留出条件下的比较改变了规划器能够检索到哪些建议。



图 S3。留出Ditto任务069，短请求。对齐帧展示了圆圈舞及其宫殿对应物；轨迹链接将检索到的建议与 v_1的首次编辑和最终选择联系起来。

表 S5。留出累积/初始比较（E_1/E_0）：27 个来源、54 个条件和 25 个组。最终比较针对选定输出，First 比较首次生成结果。比率和 n 使用顺序一致配对；未解决的比较仍纳入组加权界限。

评审者	编辑器	阶段	胜率（%）	平（%）	负（%）	n
GLM	Bernini	最终	60.7	0.0	39.3	28
		First	54.1	5.4	40.5	37
	Ditto	最终	66.7	0.0	33.3	21
		First	42.1	5.3	52.6	19
Claude	Bernini	最终	51.4	0.0	48.6	37
		First	55.0	2.5	42.5	40
	Ditto	最终	48.5	0.0	51.5	33
		First	55.3	0.0	44.7	38

在所选结果中进行符合顺序的比较时，GLM 在两个编辑器上都更偏好 E_1，而 Claude 的偏好则接近各半（表 S5）。将未解决的结果纳入后，GLM（Bernini 和 Ditto）的组加权得分界限为$[0.35, 0.77]$ 和 $[0.30, 0.80]$，Claude 的则为 $[0.38, 0.63]$ 和 $[0.33, 0.64]$。每个区间都包含 0.5 的均等偏好分数。

从检索到的建议到宫殿场景。在 Ditto 任务 069 中，源舞者围绕中央显示屏跳舞。在首次编辑之前，E_1 检索了有关简短请求、电影世界和源对象对应物的建议。生成的提示词将显示屏映射为一根镀金柱，并将舞者装扮成宫殿宾客，同时保留他们围绕中心的位置。在尝试延续版本以及基于源内容进行新的编辑后，规划器返回这个初始版本并将其交付（图 S3）。