arXiv:2611.30613v1 [cs.CV] 23 Nov 2026

CineAct：用于可交互分支式电影生成的交互状态智能体

Junyan Lin[1,7,*] Jianqi Ma[2,1,*,‡] ShujieDuan[3] Hanxin Zhu[4] ZhihaoHu[2] Jinming Liu[5]Shuai Li[6,1] Jun Zhang[2,†] Xin Jin[7,†] ChangWenChen[1,†] Weidong Zhang[2]1香港理工大学 [2]腾讯 [3]马来西亚厦门大学 [4]中国科学技术大学 a

[5]上海交通大学 [6]西北工业大学 [7]宁波东方理工大学*贡献相同。‡项目负责人。†通讯作者。

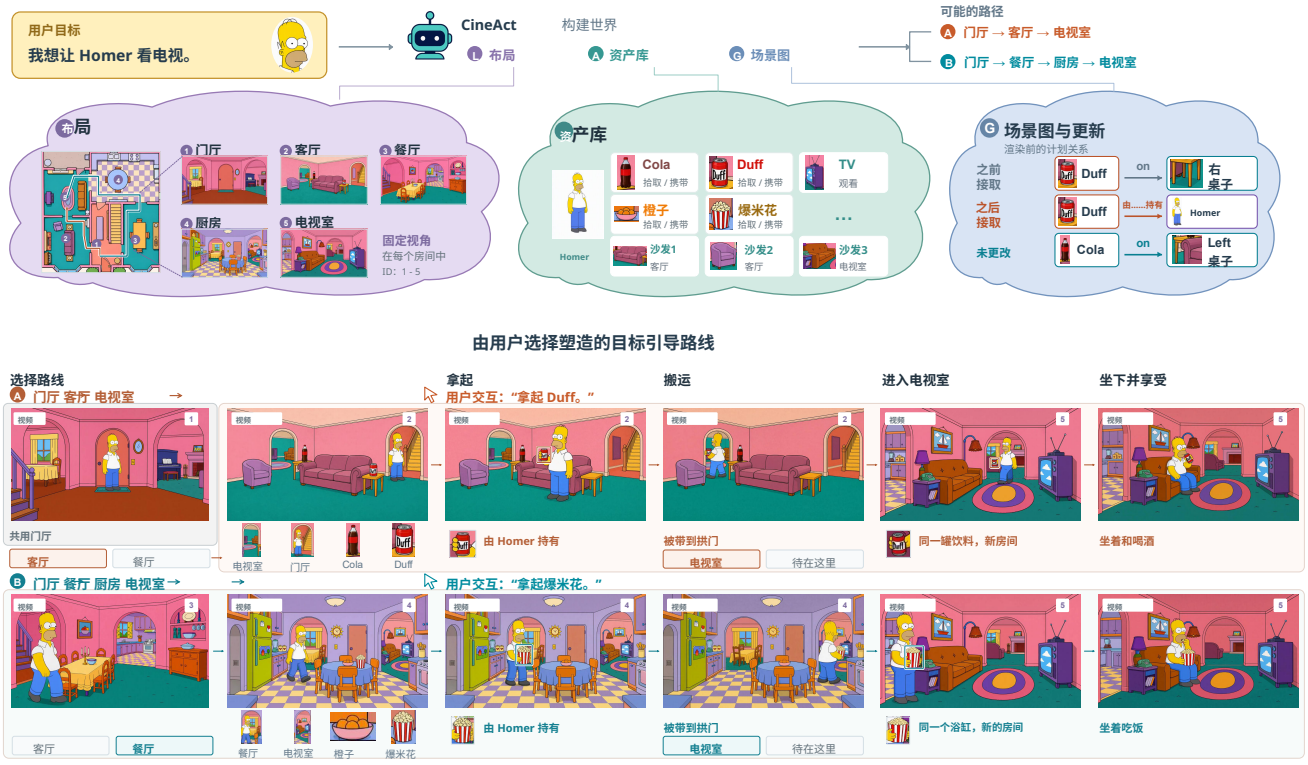


图 1. **CineAct 的目标锚定分支。** 给定一个共同目标（"我想看电视"），CineAct 会初始化空间布局（L）、资产库（{v6）和场景图（G），以管理每个回合的路线选择、实体身份和状态变化。无论 Homer 是从客厅拿 Duff（路线 A），还是从厨房拿爆米花（路线B），选定的行动都会在屏幕上执行，并在房间切换后持续保持，直到他坐进电视室。

摘要

像 *MiniMax-H3* 这样的开放式视频生成器每次只能生成几秒到十几秒的短片段。为合成长序列，现有方法要么依赖遵循预设脚本的智能体系统，要么从前一个片段的最后一帧开始自回归地展开，从而丢失屏外状态，并在多个回合中累积幻觉。因此，用户仍无法明确提出一个目标，

选择如何实现这一目标，并观察该选择在后续片段中持续存在。我们研究以目标为基础的交互式视频生成：一个目标允许多种有效方法，用户选择的方法必须在画面中执行，其后果也必须在各个片段之间持续存在。为此，我们提出 *CineAct*，这是一种状态优先智能体框架，其中语言模型智能体负责推理，而冻结的图像和视频生成器则纯粹充当渲染器。在交互之前，架构师将

世界初始化为空间布局，其中包含摄像机、资产库和场景图；在每个回合中，系统推断计划状态变化，利用图像生成器生成目标结束帧并进行严格验证，再通过 *FL2V* 模型合成中间运动。我们进一步引入了一个多空间基准测试，涵盖两个套件：长时程步骤执行和开放意图分支可接受性。实验表明，与仅像素基线相比，显式的世界状态建模提升了多步展开一致性（自动评估提高 *+11.4* 分，人工评估提高 *+23.1* 分），并使朝向共同目标的成对分支成功率提高了 *+30.8* 分。

1. 引言

开放式视频生成器，如 Wan [31], Hunyuan-Video [16], CogVideoX [37], 和 Open-Sora [41], , 以及 MiniMax-H 等模型3 [22], 在视觉质量和动作真实感方面取得了显著进展，但单次生成仍局限于数秒至十几秒的短片段。为突破这一限制并构建完整的长篇叙事，MovieAgent 等智能导演系统 [34], ViMax [12], 和 MVAgent [17], 结合基于节点的多模态创意画布，将高级脚本或关键帧串联成连贯的多镜头视频。与此同时，Genie 等交互式视频模型 [2, 26], GameNGen [30], 和 WorldPlay [27] 支持用户通过响应式控制，实时探索生成环境。近期研究进一步探索了将这两条路线结合起来：智能体根据任务目标规划下一步动作，视频模型对其进行渲染，系统则检查画面中是否已完成该动作 [8, 15, 36]。

然而，用户如何与这些系统交互，以及他们可以选择多少条路径，仍受到严格限制。智能导演和创意画布围绕预设脚本组织镜头；其审核和重新采样 [12, 17]将故事板限定为单一版本，而将它们改造成交互界面，也只是把固定工作流程拆分成一连串点击。交互式视频模型支持实时摄像机导航 [11, 27], 和指令遵循 [28, 42], , 或支持对象级轨迹和动作持久性 [10, 35], , 但它们执行步骤级控制，却不提供通往目标的其他任务级路径。相反，目标驱动的规划系统会朝任务目标推进 [8, 15, 36],, 但规划器会自主决定动作序列，而不是让用户从多种不同的有效方法中进行选择。然而，在许多领域，选择方法本身就是交互的核心：在技能训练中尝试不同工具；在可玩的游戏中，将解法多样性视为参与度的关键指标；在具身评估中测试任务在不同方法下的稳健性；以及在交互式叙事中留下持久的痕迹。通往共同目标的多条有效路径，也为智能体提供了探索、比较动作后果并获得积累可验证经验以促进策略学习和递归自我改进 [7, 24, 32, 39]。因此，我们研究目标锚定的交互式视频生成，通过菜单、点击、拖动或文本支持用户选择通往共同目标的路线，在屏幕上直观执行这些操作，并在不同回合之间保留其后果（图1）。

尽管语言智能体可以轻松提出多样的替代分支，但现有视频生成器缺乏可靠执行这些分支所需的物理基础。核心瓶颈在于，当前流程仅根据前一个片段的最后一帧自回归地展开，使整个世界状态都隐含在像素中。仅依赖像素会造成三个盲点：系统缺乏明确的前置条件，无法判断哪些动作可行并确保目标仍可达；当实体离开视野或出现外观相似的干扰物时，无法保持对象身份和状态；同时也没有几何参照来对齐不同空间之间的门口方向和镜头切换。因此，轻微的幻觉会在多回合过程中不断累积：被移除的物品重新出现，状态在重访时恢复，角色走错门，从而抹去用户选择之间的差异。

为克服这些盲点，我们提出 CineAct，这是一种将高层世界演化与低层像素合成解耦的状态优先智能体框架（图2）：明确的世界状态决定什么存在以及什么发生变化，而冻结的图像和视频模型则负责呈现其外观和运动方式。针对上述三个挑战，CineAct维护包含摄像机、资产库和场景图的空间布局，以约束跨空间转换、视觉身份和目标可达性。在每个回合中，智能体根据该状态对用户输入进行落地，推导出转换规范，合成并验证目标结束帧以锁定终点，然后在摄像机几何规则下对 FL2V 模型进行条件控制，渲染出运动，其最后一帧会刷新 $\mathcal{A}_t$ 并为下一回合提供初始条件。

为系统地评估这项任务，我们提出了一个多空间基准测试，涵盖长时程的逐步执行和分支可接受性。我们从动作、世界状态和空间维度开展了受控消融实验、跨骨干网络比较、反事实门控探测、配对分支测试以及界面研究。在多步 rollout 中，无论是在自动评估（+11.4 分）还是盲测人工评分（+23.1 分）方面，CineAct 的整体一致性都显著优于仅像素基线；消融实验证实，布局、资产和场景图分别带来了独特的提升。此外，配对分支评估证实，CineAct 能够对通往共同目标的多样化用户选择路线进行门控、执行并加以保留（配对分支成功率比基线高 +30.8 分）。

我们的贡献主要有三点：

- **任务定义**。 我们将以目标为依据的交互式视频生成形式化为：要求系统支撑通往共同目标的多条用户选择的路线，同时

共同满足动作有效性、执行可见性和跨回合后果持久性。

• **状态优先智能体框架**。我们提出了 CineAct，将世界状态推理与像素合成解耦。通过维护明确的空间布局、动态更新的资产库和场景图，CineAct 借助预渲染目标帧验证锁定动作后端点，并在摄像机几何规则的约束下驱动 FL2V 生成。

•**基准测试与评估**。我们推出了一项多空间基准测试，涵盖长时程步骤执行，并从动作、世界状态和空间维度评估分支可接受性。受控消融实验、骨干网络研究和接口比较量化了每种表征的独特贡献。

2. 相关工作

多镜头视频与智能体导演。智能体导演系统统筹多个镜头，根据叙事脚本合成连贯的长篇视频。有些系统通过分层思维链模拟剧本创作、故事板制作和拍摄 [34]，另一些则借助视觉语言智能体持续监控并优化叙事连贯性 [12]。MVAgent [17] 维护连续性记忆，并从共享的镜头状态中采样候选后续内容；CineEvolve [40] 共同演进导演与评论者策略；VideoGen-Agent [18]则通过强化学习协调生成与验证工具。为支持对先前空间的非连续重访，ATAS [33] 将电影剧本编译为有界的镜头请求，并以明确的场景几何信息和事件记忆为支撑。尽管这些系统能够生成连贯的多镜头影片，但它们仍围绕预设脚本构建，而非围绕用户为实现任务目标所做的交互式选择组织。相比之下，我们要求生成的每个片段都能直接响应用户的在线选择，并支持多条可行路径。

可控交互式视频生成。交互式视频模型会在像素空间中生成视频帧，以响应控制信号或在线用户输入。早期工作通过无监督潜在动作 [2]，探索驱动的数据收集 [26],或智能体回放 [30]来合成交互式环境。随后出现的流式模型支持帧级键盘和鼠标控制 [11]，，通过从过去帧中动态重建上下文记忆来保持长程几何一致性 [27]，，或在游戏世界中遵循开放式指令提示 [28]。近期工作进一步发展到实体级操控：Incantation [42] 引入了逐实体语言控制和任务特定状态跟踪，WorldCraft [10] 将物体选择、二维拖动轨迹与摄像机导航相结合，以在离屏回访时保持对象放置位置，而 ActWorld [35]则通过动作感知记忆保留动作引发的状态，例如打开的门和拾取的物品。与此同时，CollabVR [15]和 WorldGuide [8] 等目标驱动系统将步骤级动作规划与视频执行验证相结合，以实现任务目标。然而，这些方法要么在执行步骤级控制时不检查多步骤目标可达性，要么依赖自动化规划器规定下一步，而无法为用户提供通往共同目标的其他同样有效的路径。

显式世界状态与生成式渲染。为抑制自回归展开中的幻觉累积，近期工作将状态逻辑和物理规则与神经像素合成解耦。PWM [14] 将语言编译为世界程序，并借助三维边界框和经训练的空间控制分支引导渲染；AgentGarten [7]为演化智能体构建代码驱动的世界；GameDirector [20] 与 Code World Model [6] 将战斗机制或可执行代码演化与视频渲染分离。在开放世界交互方面，Oneira [36] 将目标分解为对象状态表上按依赖关系排序的交互，并采用LoRA适配渲染，而 State World Models [4] 将类型化规则台账和目标条件与黏土场景相机渲染相结合。在单片段合成中，CausalMotion [43] 与 PhysPlan [23] 利用因果关键帧或对物理状态图的稀疏编辑来引导推理，SolarRenderer [9] 则通过骨干网络的原生参考路径注入结构化视频。尽管显式状态表示已相当成熟，现有框架仍用它们来执行预定义机制或单一分解计划，且其中若干方法需要训练任务专用适配器或控制分支 [6, 14, 20, 36]。我们则将冻结的图像和视频生成器与空间布局、更新后的资产库以及场景图相结合，以推导并验证通向共同目标的多条用户选择的路线，并利用预渲染目标帧验证在FL2V合成之前锁定终点。

3. 问题设定

我们将目标锚定的交互式视频生成表述为一种围绕共享任务目标组织的多轮可分支条件生成问题。令 g 为需要实现的全局目标（例如“在黄昏前准备好一个休息地点”），并从初始场景观测 I_1开始。在第 t回合，用户通过菜单选择、在当前帧上的点击或拖动轨迹，或自由文本，提供交互输入 $u_t \in \mathcal{U}$。系统根据当前帧 I_t 和交互历史，将 u_t 解析为语义动作 a_t，并合成视频片段 V_t；其实际最后一帧 $I_{t+1} = \text{Tail}(V_t)$作为第 $t+1$回合的起始帧。交互持续进行，直到完成 g，从而得到一条轨迹

$$\tau = (I_1, u_1, V_1, I_2, u_2, V_2, \ldots, u_K, V_K). \quad (1)$$

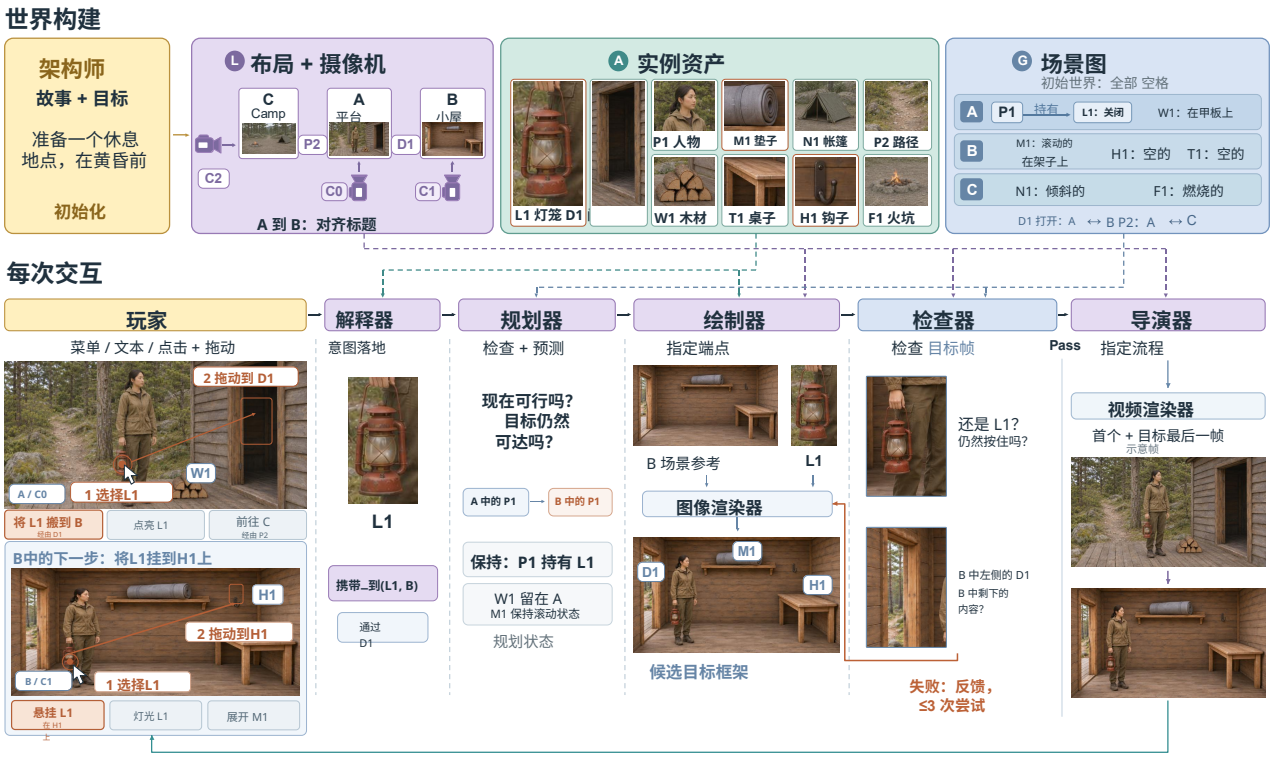


图 2。**CineAct 框架。顶部（世界构建）：**架构师根据故事目标，使用固定摄像机（L）初始化空间布局、由实例和场景参考组成的资产库（$\mathcal{A}$），以及多空间场景图（G）。**底部（每次交互）：**解释器将玩家输入落地，规划器在 G上检查前置条件和目标可达性，绘制器驱动图像渲染器合成候选目标结束帧，检查器将其与规划状态进行核验（最多三次尝试，并返回失败结果），随后导演器渲染转场视频，其尾部刷新 $\mathcal{A}$ 并为下一回合提供种子。

不同于单脚本视频生成，共享的初始状态或中间历史（$I_1, \ldots, I_t$）允许存在多条能够达到相同目标 g 的有效路线 $\{\tau^{(m)}\}$，同时在后续帧中留下路线特定后果。

一致性要求。对于要支持交互式游戏的生成轨迹 τ，它必须在各回合之间同时满足三项一致性标准 $t \in \{1, \ldots, K\}$:

（1）动作有效性和过程可见性。每个执行的动作 a_t 都必须满足当前场景的物理前置条件 $P(a_t)$，同时确保全局目标 g 在后续回合中仍然可达。此外，生成片段 V_t 必须直观呈现物理动作过程，而不能只是传送到最终状态:

$$V_t \models \Delta(a_t). \tag{2}$$

(2) 实体身份与状态持久性。被操纵的物体和角色必须保持其视觉身份，不得在不同回合中消失、重复出现或互换。新确立的持握或支撑状态，以及未受影响的场景属性 $K(a_t)$，都必须在后续片段中持续保持（当实体离开并重新进入视野时）：

$$V_k \models K(a_t), \quad \forall k \geq t. \tag{3}$$

（3）跨空间几何与摄像机连续性。当交互跨越多个连通空间或重访先前的空间时，角色离开源空间的方向、目标空间中入口所在侧以及摄像机视点转换必须在各片段间保持几何一致。标准的自回归展开仅以 I_t前一帧为条件，导致这三项约束都隐含在像素中。下文将介绍 CineAct，说明如何利用冻结渲染器显式建模并强制满足这些约束。

4. CineAct：用于交互式影片生成的状态优先智能体框架

图 2概述了 CineAct。该框架将高层推理与低层渲染解耦。推理任务，即意图解析、有效性检查、可达性搜索和目标帧验证，均由语言模型智能体处理物理渲染器 [9, 14]：它们负责生成

像素，这些像素基于经过验证的端点状态和几何约束生成，而渲染器不参与推断逻辑世界状态。

4.1. 架构师：世界初始化

在交互开始之前，语言模型智能体 **架构师** 围绕全局目标构建一个可交互的多空间世界 g，并设计多条有效路线（例如图 2上方：将灯笼 L1 带入小屋 B，或将柴火 W1 带到营地 C）。为确保多路径可玩性，系统会在初始化期间验证可达性，从而保证至少有两条实质上不同的动作序列——它们在工具、策略或空间路径上有所差异——能够从起点或共享的故事中段决策点抵达 g，且不会产生冗余的往返行程。随后，架构师将这个世界具体化为三种互补的表示 $S_0 =$（L、$\mathcal{A}_0$、G_0）（图 2上方）：

（1）空间布局与摄像机（$L = (\mathcal{R}、\mathcal{P}、\mathcal{C})$）。该布局将世界表示为一张二维轴对齐平面图，规定矩形空间边界 $r \in \mathcal{R}$ 以及共享边界的门户线段 $D \in \mathcal{P}$，并为每个空间配备一个确定性的固定摄像机 $\mathcal{C}$（放置在没有门户的墙面上，朝向一个出口）。几何投影会推导出每个门户在屏幕上的位置（左侧、右侧或远端墙面），以约束跨空间转换并验证门口位置。

（2）资产库（$\mathcal{A}_t$）。资产库将每个空间的视觉记忆（从初始渲染图初始化，并在每个回合结束后使用该空间最近生成的视频片段进行更新）与初始实体身份裁剪图$\mathcal{A}_0$（v）相结合；这些裁剪图均配有唯一标识符和稳定且与位置无关的外观描述。在每个步骤中，系统会检索目标空间记忆，以及动作参与者和画外角色的身份裁剪图，同时跳过 I_t 中已经可见的角色，以避免出现重复人物。

（3）场景图（$G_t =$（$\mathcal{V}_t$，$\mathcal{E}_t$））。场景图初始化为 G_0，用于追踪所有画面内和画外实体 $\mathcal{V}_t$。包含关系、支撑状态和持有关系会沿着层级所有权指针，结合可携带性、承载能力上限以及状态属性进行确定性推导$\mathcal{E}_t$（例如，打开/关闭、点亮/未点亮）。符号化转换引擎会在 G_t上检查前置条件和目标可达性，并在每个回合提供有效的动作建议。

4.2. 解释器和规划器：意图落地与状态推导

在第 t回合，系统根据 G_t提供可行的菜单建议，同时用户可以通过点击、拖动或自由文本 u_t 选择菜单选项或输入开放式内容。给定当前帧 I_t、资产库 $\mathcal{A}_t$以及 G_t，**解释器**将 u_t 映射到具体的场景实体，将其解析为结构化操作 a_t、预期的状态变化Δ_t以及不变性约束（图2，底部）。

规划器随后使用符号转换引擎，在 G_t 上验证计划内选择和分布外（菜单外）操作。如果 a_t 匹配一个预先验证的分支，它便直接推进；否则，规划器会检查 G_t 上的物理前置条件 P（a_t）、承载能力和局部性，推导出 $G_t^{\text{expected}} = G_t \oplus \Delta_t$，并运行图搜索，以验证全局目标 g 仍可从 G_t^{expected} 达成。物理上不可能、存在歧义或会导致 g 无法达成的输入，会在渲染前被拒绝。经过验证的动作会将计划场景图推进至 $G_{t+1} = G_t^{\text{expected}}$（与观测到的视频帧分开维护），并以确定性方式将差异计算（G_t，G_{t+1}）整理为明确的转换规范：

$$\text{Spec}_t = (\mathcal{V}_{\text{operated}}, \Delta_t, \mathcal{V}_{\text{kept}}, C_t), \tag{4}$$

其中详细说明了操作对象实例、修改后的关系、必须保持不变的实例，以及对于跨空间移动的相机转场指令 C_t。

4.3. 绘制器与检查器：经验证的目标帧生成

在首尾帧（FL2V）视频合成中，起始帧 I_t 是固定的，因此确保准确的目标结束帧 $I_t^\star$能够为渲染其中的运动提供可靠的边界锚点。给定转换规范，**绘制器**将规划好的变化、布局 L和资产库 $\mathcal{A}_t$ 整合为图像提示词，并调用图像编辑模型生成候选目标结束帧 $I_t^\star$，描绘预期的动作后状态，其条件包括当前帧 I_t、跨空间移动时的目标空间记忆，以及操作对象的身份裁剪图（例如图2中的场景 B 和灯笼 L1）。

在绘制器与导演器之间，**检查器**充当目标端点把关器，使用视觉语言模型在视频渲染前，逐项将每个候选目标结束帧判定为通过、失败或不明确：检查其中是否呈现已完成的动作、已操作的实例是否保持其身份，以及非目标对象是否未发生变化；对于跨空间移动，它还会在候选帧中定位门口，并将其屏幕侧边与 L的几何投影进行比对。具体而言，它会检查转换规范中的每一项，包括计划中的变化、手持物品和必须保留的实例（例如确认图 2 中 L1 仍是同一盏手持灯笼，且 D1 出现在 B 的左侧）。检查失败时，会将其视觉证据返回给绘制器进行重新生成（如图 2所示的反馈循环；每个步骤最多尝试三次）。验证完成后，$I_t^\star$ 即被锁定并传递给导演器。

4.4. 导演器：几何感知过程渲染

给定当前真实帧 I_t 和经过验证的帧 $I_t^\star$，**导演器**会合成两者之间的过渡。对于单个空间内的交互，导演器根据 I_t、$I_t^\star$ 和动作文本编写动作描述，并在一个固定机位镜头中渲染该动作。对于空间之间的移动，

空间，它读取 L 相机方向差异$\delta = (\phi_d - \phi_s)$ 模 $360°$ 源空间 r_s 与目标空间 r_d之间的值，并将其与门户几何相结合，以指定两个固定镜头中屏幕上的出口侧和入口侧，再通过一次剪辑将两个镜头连接起来。例如，当相邻空间具有相同的相机朝向（$\delta = 0°$ 见图 2）时，角色从右侧的门口退出后，自然会从左侧进入下一个空间。结合 和 $I_t^\star$，这些指令会对 FL2V 视频模型进行条件控制，使其渲染一段 10 秒的过渡片段 V_t。一旦 V_t 渲染完成，其实际的最后一帧 $I_{t+1} = \mathrm{Tail}(V_t)$ 既会作为下一回合的种子，也会更新 $\mathcal{A}_{t+1}$ 中的空间记忆，将其作为所访问空间的最新视觉观察（图2中指向下一回合面板的反馈箭头），从而在后续回合和重新访问中保留场景与物体的变化。

5. 基准测试

不同于现有的、分别评估片段级视觉质量 [13], 、规定的多回合指令遵循情况 [38], 、遮挡下消失并重新出现的物体恢复能力 [5], 或成对反事实干预 [3, 21], 的视频基准测试，我们构建了一个交互式基准测试，针对共同目标联合评估多空间执行一致性与预渲染分支可接受性。

5.1. 评估套件

该基准测试建立在由 Hy4 [29] 编写的 40 个自然长时程任务之上，涵盖多样的室内、室外和混合场景（分布于 1–4 个连通空间中，每类 10 个），由两个套件组成：**(1)多步交互式执行**评估这 40 个多回合情节中的展开一致性。**(2) 分支可接受性**评估其开场决策点的预渲染提案门控，涉及 571 个经事实核验、以玩家措辞表达的提案：237 条有效路线，以及 334 条伪装成可行方案的不可行反事实，其中涉及不存在的地点（$n = 80$）、不存在的物体（$n = 80$）或未满足的前置条件（$n = 174$）。

5.2. 评估指标

对于**多步交互式执行**，我们从**动作**、**世界状态**和**空间连续性**三个方面评估每一步。自动化 Hy4 评估对全部十个项目进行评分，而盲人人类评分者则标注四项核心指标：动作（动作完成度）、状态（物体持有状态和支撑状态）、进入（目的地和入口侧）以及退出（出口门口方向）。对于因摄像机画面丢失而遮挡目标的步骤，将其单独标记为不明确（第6.4节）；完整的十项列表及评分标准详见补充材料（补充表 S1）。

对于**分支可接受性**，我们主要在**开放**模式（可接受 /不可接受 / 不确定，其中不确定 计为失败）下进行评估，并报告有效提案的接受率（接受）以及不可行提案的拒绝率（拒绝）、其均值（平衡）以及不确定 比率。我们还加入**强制**二元控制，不允许弃权，以检查较高的拒绝率是否依赖于对有效提案一概拒绝这一取巧做法。

6. 实验

6.1. 实验设置

模型与骨干网络。所有高层推理角色（架构师、解释器、规划器、检查器以及任务编写模型）均由多模态语言模型 Hy4 [29]提供。对于生成式渲染，我们将两个图像生成器（Qwen-Image-2.1 [25] 和 FLUX.2 [klein] 4B [1]）分别与两个冻结的 FL2V 视频骨干网络（MiniMax-H 3 [22]和 LTX-2.5 [19]）进行交叉配对，以合成 10 秒的转场片段，并比较不同组合下的基线与 CineAct。第6.4节中的盲人人类评估和参考视频生成（Ref2V）接口比较均基于 Qwen-Image-2.1 + MiniMax-H3.

消融设置。 为单独评估每种显式表示，我们在统一的跨空间房间记忆策略下测试四种逐步累加的设置（首次进入时使用初始房间渲染图，再次进入时使用自身的最后一帧视频）： 基线（直接滚动生成）仅接收当前帧和用户指令；+布局 增加空间布局 L 和摄像机转场规则； +资产 进一步增加实体身份裁剪图和稳定的外观描述 $\mathcal{A}_t$；*CineAct* （完整框架）增加场景图 G_t ，并结合符号状态推导、确定性的状态变化检查清单以及目标帧检查。对于不包含 G_t的设置，前三种设置会绕过符号规划，直接根据动作指令以及各自允许使用的布局或资产输入合成帧。所有对比设置均使用相同的冻结骨干网络、动作序列、初始帧、随机种子和逐步计算预算（最多尝试生成三个目标帧，并生成一次视频），并从各自生成的视频片段自回归地滚动生成，以单独评估显式世界建模的作用。

6.2. 多步执行

表 1 和图3 总结了所有四种骨干模型组合在多步交互式执行上的逐步消融实验（Hy4 评分；完整的 10 项细分见补充表 S3），以及 10 名评分者对 Qwen-Image-2 进行的盲人人工评分。1 +MiniMax-H3:

- **布局与摄像机（+布局与基线）：** 添加空间布局 L 和摄像机转场规则可提升不同骨干网络的空间一致性和总体一致性，有效解决跨空间迷失方向问题。
- **资产库（+资产与 +布局）：** 引入更新后的场景和实例参考后，在 Qwen-Image-2.1 下进一步提升了 $\mathcal{A}_t$ 动作和世界状态,

表1。图像和视频骨干网络组合上的自动化 Hy4 评估（FLUX.2+ H3 对基线与 CineAct 进行评估；10项细分结果见补充表 S3）。

图像	视频	设置	动作	世界	间隔	全部
Qwen2.1 [25]	H3 [22]	基线	.630	.760		.807
		+布局	.647	.734		.876
		+资产	.670	.780		.864
		CineAct	**.767**	**.878**		**.877**
FLUX.2 [1]	H3 [22]	基线	.517	.547		.823
		CineAct	**.600**	**.596**		**.889**
Qwen2.1 [25]	LTX-2.5 [19]	基线	.586	.677		.698
		+布局	.586	.679		**.746**
		+资产	.657	.747		.698
		CineAct	**.662**	**.827**		.712
FLUX.2 [1]	LTX-2.5 [19]	基线	.471	.502		**.751**
		+布局	.557	.541		.679
		+资产	.508	.535		.642
		CineAct	**.602**	**.687**		.686

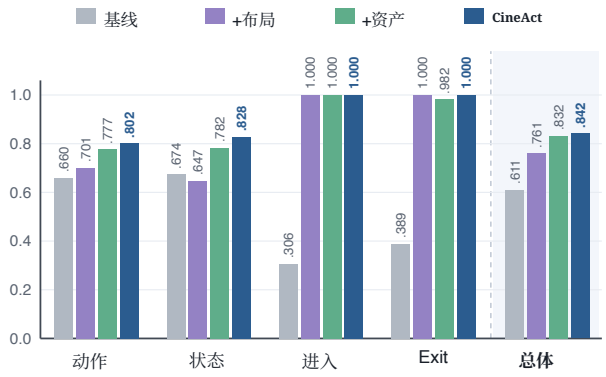


图3。**由10名评分者对10个样品故事进行盲人人类评估**，使用 Qwen-Image-2.1 + MiniMax-H3（FL2V）。

而轻量级 FLUX.2 [klein] 4B [1] 在完整的 Cine-Act 强力恢复之前，表现出轻微的多参考编辑方差。

• **场景图与验证（CineAct 与基线）：** 添加 G_t、转场规范和目标帧检查后，在全部四种骨干网络配对中实现了最高的一致性，使动作、世界状态及总体得分均超过基线（在 Qwen2.1 + H3 上总体得分为 .862，提升了+11.4 个百分点），展现出强大的跨骨干网络泛化能力。

• **盲人人类评估（图3）：** 10 名评分者对 Qwen-Image-2 上抽取的 10 个故事进行了盲评。1 +MiniMax-H3（40 个匿名故事—设定单元、196 个片段、784 项条目评分），印证了总体得分从 .611 单调提升至 .842（+23.1 分），且在每项指标上均取得最佳得分（动作 .802、状态 .828、En-

表2. 40 个故事中的分支可接受性（237 个有效 + 334 无效 = 571 提案）。开放 将不确定 计为失败；强制 不允许弃权。

Mode	设置	接受 ↑	拒绝 ↑	平衡 ↑	不确定 ↓
Open	基线	.080	.213	.146	80.0%
	+布局	.089	.287	.188	78.1%
	+资产	.498	.722	.610	29.8%
	CineAct	**.810**	**.964**	**.887**	**4.0%**
强制	基线	.262	.970	.616	–
	+布局	.650	.895	.772	–
	+资产	.738	.934	.836	–
	CineAct	**.827**	**.982**	**.905**	–

表3。开放模式下的不可行提案漏检率（未拒绝 ↓）。每行的最佳结果以粗体显示；带下划线的数值表示特定层的下降。

不可行类别	基线。	+布局	+资产	CineAct
不存在的地点（$n = 80$）	.95	.24	.05	**.00**
不存在的物体（$n = 80$）	.80	.91	.30	**.06**
前置条件失效（$n = 174$）	.71	.84	.37	**.04**

尝试/退出 1.000）。

图4（a–c）展示了四房间浇水任务中的这一逐步进展：给定（a）中的设置和路线，基线会走错中央拱门，前往水室，而 +布局则从正确的西门进入（b）；经过多次房间转换后，+布局会丢下物品，而+资产会将两个水袋松散地放在篮子外面；CineAct 则会将两个水袋都妥善装在篮子里（c）。

6.3. 分支可接受性

交互式视频生成不仅要求遵循单一固定脚本：给定一个共同目标，系统还必须接纳玩家提出的多种分支，在允许有效替代路径的同时，过滤掉违反世界约束的请求。在全部 40 个故事的开场决策点中，我们评估了使用 Hy4 编写的 571 条由玩家措辞的提案（其中 237 条是有效的目标达成方案，334 条是看似合理但实际不可行的反事实）。每种消融设置仅依据开场帧、任务目标及其允许使用的先验信息（$L, \mathcal{A}_0$ 包含所有空格的初始视图和实体裁剪图，或 G_0）来判断提案是否可接受，而不生成视频。

如表 2和 3所示，仅像素基线在 不确定模式下对 80.0% 的提案选择弃权（尤其是在多空间场景中），仅接受 .080 的有效分支（.146 的平衡准确率）。即使在强制控制下，基线也只是通过拒绝几乎所有输入来达到 .970 的拒绝率，阻断了 73.8% 的有效路线（接受率为 .262）。在三个反事实类

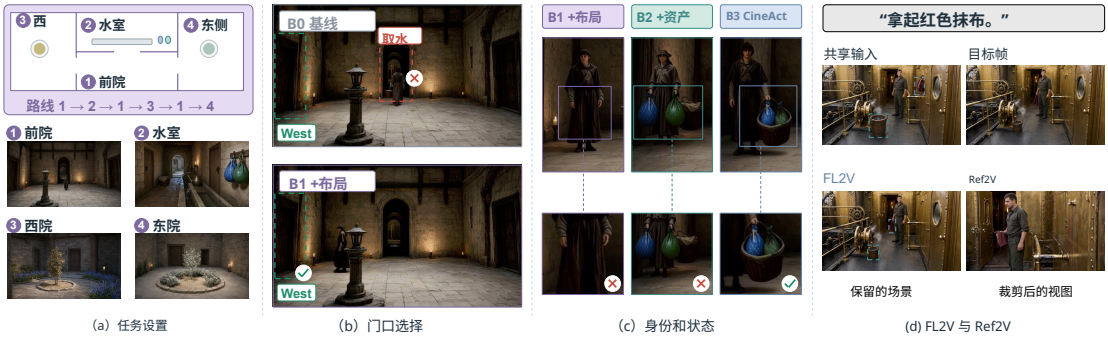


图 4. **定性比较。(a–c)** 多步执行：布局可实现正确的门道路径规划 (b)，而完整的 CineAct 可保留身份和包含关系 (c)。**(d)** FL2V 与 Ref2V：经验证的目标帧 保留了房间视图，而*Ref2V* 则裁剪了场景。

表 4. 对 13 个多路径故事进行成对分支执行的盲人人类评估（每种方法 26 条轨迹）。

指标	基线	CineAct	Δ
单路径目标完成率	.769	**1.000**	+.231
双路径目标完成率	.615	**1.000**	+.385
路径特定后果	.769	**1.000**	+.231
无交叉泄漏或回滚	.846	**.962**	+.115
两者均成功且不同（n=13）	.615	**.923**	+.308
两者均成功且不同（2–4 个空间）	.500	**.900**	+.400

别（表 3；其中，漏检率统计将不可行提案回答为 可接受或不确定的情况）中，逐步加入显式表示可逐类解决错误：+布局显著减少了不存在的地点漏检；+资产大幅减少了不存在的物体漏检，并将总体弃权率降至 29.8%；加入 G_0（CineAct）后，前置条件失效漏检进一步下降，使弃权率降至 4.0%，同时在开放（.887 平衡）和强制（.905 平衡）两种模式下均实现较高的接受率和拒绝率。为评估配对执行，评分者在不知方法的情况下，对 13 个多路径故事（每种方法 26 条轨迹，涵盖 1–4 个空间；表 4）按每条路径及双路径目标完成率、路径特定后果保持，以及零交叉泄漏或回滚进行评分（后两项合计计为两者均成功且不同）。从相同的初始状态出发，CineAct 实现了 1.000 的双路径目标完成率（相比 .615），以及 .923 的两者均成功且不同（在 2–4 个空间上为 .900，n=10，相比 .500）。

6.4. 接口分析：FL2V 与 Ref2V

我们将 FL2V 流程与参考到视频（Ref2V）进行比较。Ref2V 在相同的 10 个故事上，直接以 I_t 和资产裁剪图为条件输入 MiniMax-H3。对于可通过视觉判定的项目，在包含 10 个项目的 Hy4 评估中，Ref2V 达到了与 FL2V 相当的基线到 CineAct 增益（.713→.875，而 FL2V 为 .733→.856）。然而，Ref2V 在 26.0%（51/196）的物体持有/支撑（状态）步骤在人工评估中无法判定不明确（FL2V 为 0.0%），这对于多回合交互而言不可接受。具体而言，如图 4(d) 所示，从共享输入开始，FL2V 通过经过验证的 目标帧来锚定房间的取景，因此其输出保留了画面中的抹布、桶和机械设备（保留的场景）；相比之下，*Ref2V* 偏移至特写镜头，将桶和周围房间裁出了画面（裁剪后的视图），导致用户无法在后续回合中查看或选择这些物体（补充表S4）。

7. 局限性

尽管 CineAct 提升了多回合一致性和分支可靠性，但仍存在三项权衡。首先，跳过逐镜头三维重建和微调，使流程保持轻量且一致，但也将每个房间限制为固定摄像机，并采用剪辑转场，而非连续的三维相机运动。其次，顺序推理、迭代验证和 10 秒 FL2V 合成带来的是基于回合的延迟，而非实时流式延迟。第三，房间和实体由 Architect 预先固定；动态添加新房间、片段后的状态重新同步以及流式骨干网络，则留待未来工作。

8. 结论

我们提出了目标锚定的交互式视频生成，并介绍了 CineAct——一种状态优先智能体框架，将高层世界状态推理与低层像素渲染解耦。通过维护包含摄像机的显式空间布局、动态刷新的资产库和场景图，并通过预渲染目标帧验证锁定动作后端点，CineAct 实现了可靠的多轮执行，并能在预渲染阶段对开放式分支提案进行门控。跨越多步执行、反事实分支可接受性和界面比较的实验，验证了每种显式表示所发挥的专门作用。

参考文献

[1] Black Forest Labs。FLUX.2 [klein] 4B。模型发布于 Hugging Face，2026年。6, 7, 11, 13[2] Jake Bruce、Michael Dennis、Ashley Edwards、Jack Parker-Holder、Yuge Shi、Edward Hughes、Matthew Lai、Aditi Mavalankar、Richie Steigerwald、Chris Apps、Yusuf Aytar、Sarah Bechtle、Feryal Behbahani、Stephanie Chan、Nicolas Heess、Lucy Gonzalez、Simon Osindero、Sherjil Ozair、Scott Reed、Jingwei Zhang、Konrad Zolna、Jeff Clune、Nando de Freitas、Satinder Singh 和 Tim Rocktäschel。Genie：生成式交互环境。在第 41 届国际机器学习会议论文集，第 4603–4623 页，2024年。2, 3[3] Kunlin Cai、Rui Song、Jinghuai Zhang、Kaiyuan Zhang、Pranav Bodapati、Alicia Yu、Fnu Suya、Mohammad Rostami、Jiaqi Ma 和 Yuan Tian。What-If World：具身场景中通用世界模型的因果基准。arXiv 预印本 arXiv:2605.27589，2026年。6[4] Yan Ki Chan、Yanran Zhang、Wenzhao Zheng、Xiaofeng Wang、Zheng Zhu、Lei Chen、Jie Zhou 和 Jiwen Lu。State World Models：执行规则，生成像素。arXiv 预印本 arXiv:2610.26811，2026年。3[5] Haoyu Chen、Kaichen Zhou、Hang Hua、Kaile Zhang、Jingwen Qian、Wufei Ma、Haonan Chen、Chunjiang Liu、Yizhou Zhao、Xiaoyuan Wang、Weiyue Li、Alan L. Yuille、Paul Pu Liang 和 Yilun Du。MemoBench：动态变化环境中的世界建模基准测试。在欧洲计算机视觉会议（ECCV）论文集，2026年。6[6] Yiwen Chen、Guosheng Lin 和 Chi Zhang。Code World Model：作为世界大脑的编码智能体。arXiv 预印本 arXiv:2608.25927，2026年。3[7]Jiawei Chi、Shangchen Miao、Zhiyuan Shi、Kailu Wu、Hanyang Wang、Weiliang Chen、Qiyu Dai、Chengyang Li、Jinshan Ren、Jun Gao、Mingsheng Long、Yueqi Duan、Jiangran Lyu、Jialong Wu 和 Fangfu Liu。Agent-Garten：面向演化智能体的代码世界。arXiv 预印本 arXiv:2610.12374，2026年。2, 3[8]Ankan Deria、Komal Kumar、Hisham Cholakkal、Fahad Shahbaz Khan 和 Salman Khan。WorldGuide：用于程序化任务执行的目标导向视频世界模型。arXiv 预印本 arXiv:2610.12459，2026年。2, 3[9] Guian Fang、Shengju Qian、Dongyue Lu、Junchao Huang、Lutao Jiang、Xin Wang、Yiren Song、Li Jiang 和 Mike Zheng Shou。SolarRenderer：少步生成式参考渲染。arXiv 预印本 arXiv:2610 28630，2026年 3, 4 . . [10] Bohai Gu、Taiyi Wu、Yueyang Yuan、Jian Liu、Xiaocheng Lu、Dazhao Du、Jie Zhang、Jinxiang Lai、Shuai Yang、Xiaotong Zhao、Alan Zhao 和 Song Guo。WorldCraft：从摄像机导航到交互式视频世界模型中的物体操控。arXiv 预印本 arXiv:2605.25077，2026年。2, 3[11] Xianglong He、Chunli Peng、Zexiang Liu、Boyang Wang、Yifan Zhang、Qi Cui、Fei Kang、Biao Jiang、Mengyin An、Yangyang Ren、Baixin Xu、Hao-Xiang Guo、Kaixiong Gong、Cyrus Wu、Wei Li、Xuchen Song、Yang Liu、Eric Li 和 Yahui Zhou。Matrix-Game 2.0：一种开源、实时流式交互世界模型。arXiv预印本 arXiv:2508.13009，2025年。2，3[12] Lingxuan Huang、Sizhe He、Hengji Zhou、Liqiang Nie、Lianghao Xia 和 Chao Huang。ViMax：智能体视频生成。arXiv预印本 arXiv:2606.07649，2026年。2，3[13] Ziqi Huang、Yinan He、Jiashuo Yu、Fan Zhang、Chenyang Si、Yuming Jiang、Yuanhan Zhang、Tianxing Wu、Qingyang Jin、Nattapol Chanpaisit、Yaohui Wang、Xinyuan Chen、Limin Wang、Dahua Lin、Yu Qiao 和 Ziwei Liu。VBench：视频生成模型综合基准测试套件。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第21807–21818页，2024年。6[14] Zheng-Hui Huang、Guixu Lin、Jiacheng Lin、Yi-Chuan Huang、Ruihan Yu、Muyao Niu、Siqi Yang、Yu-Lun Liu、Yung-Yu Chuang、Kaipeng Zhang 和 Zhixiang Wang。可编程世界模型。arXiv预印本 arXiv:2609 10540，2026年。3，4。[15] Joowon Kim、Seungho Shin、Joonhyung Park 和 Eunho Yang。CollabVR：利用视觉语言模型与视频生成模型进行协作式视频推理。载于神经信息处理系统进展（NeurIPS），2026年。2，3[16] Weijie Kong、Qi Tian、Zijian Zhang、Rox Min、Zuozhuo Dai、Jin Zhou、Jiangfeng Xiong、Xin Li、Bo Wu、Jianwei Zhang 等。HunyuanVideo：大规模视频生成模型的系统性框架。arXiv预印本 arXiv:2412 03603，2024年。2。[17] Xiangyu Kong、Wenjie Zhou、Fengping Tian、Lihua Fang、Haoqin Sun、Chenyang Lyu、Longyue Wang 和 Weihua Luo。MVAgent：通过一致条件构建与镜头级策略优化实现多智能体视频生成。arXiv预印本 arXiv:2609.30609，2026年。2，3[18] Binxu Li、Haoyi Duan、Yuhui Zhang、Yaohui Zhang、Zihao Lin、Kaituo Feng、Hui Yuan、Suozhi Huang、Xiangyi Li、Yu Li、Chunyuan Li、Mengdi Wang 和 Shilong Liu。VideoGen-Agent：强化视频生成智能体。arXiv预印本 arXiv:2609.24997，2026年。3[19] Lightricks。LTX-2.5。发布于 Hugging Face 的模型，2026年。6，7，11，13[20] Zijun Lin、Zhiyang Deng、Yuzhe Wu、Bihan Wen 和 Yeying Jin。GameDirector：将游戏玩法逻辑与渲染解耦，实现玩家可配置的游戏世界模型。arXiv预印本 arXiv:2609.25652，2026年。3[21] Yu Ma、Hongli Shi 和 Xinran Xu。Twin Rollouts：交互式视频世界模型中的噪声耦合反事实分支。arXiv预印本 arXiv:2608.08982，2026年。6[22] MiniMax。MiniMax-H3。发布于 Hugging Face 的模型，2026年。2，6，7，11[23] Minh-Loi Nguyen、Xuan-Vu Le、Trung-Nghia Le、Tam V. Nguyen、Minh-Triet Tran 和 Thanh-Toan Do。PhysPlan：基于物理状态的推理与图引导优化，实现物理合理的视频生成。arXiv预印本 arXiv:2609.13006，2026年。3[24] Yuxiao Qu、Tianjun Zhang、Naman Garg 和 Aviral Kumar。递归内省：教会语言模型智能体如何自我改进。载于神经信息处理系统进展（NeurIPS），2024年。2

[25] Qwen团队。Qwen-Image-2.1。发布于 Hugging Face 的模型，2026年。6，7，11[26]Nedko Savov、Naser Kazemi、Mohammad Mahdi、Danda Pani Paudel、Xi Wang 和 Luc Van Gool。探索驱动的生成式交互环境。载于 IEEE/CVF计算机视觉与模式识别会议论文集，2025年。2，3[27] Wenqiang Sun、Haiyu Zhang、Haoyuan Wang、Junta Wu、Zehan Wang、Zhenwei Wang、Yunhong Wang、Jun Zhang、Tengfei Wang 和 Chunchao Guo。WorldPlay：迈向实时交互式世界建模中的长期几何一致性。arXiv预印本 arXiv:2512 14614，2025年。2，3。[28] Junshu Tang、Jiacheng Liu、Jiaqi Li、Longhuang Wu、Haoyu Yang、Penghao Zhao、Siruis Gong、Xiang Yuan、Shuai Shao、Linfeng Zhang 和 Qinglin Lu。Hunyuan-GameCraft-2：指令遵循交互式游戏世界模型。arXiv预印本 arXiv:2511.23429，2025年。2，3[29] 腾讯Hy团队。Hy4，2026年。6，11[30] Dani Valevski、Yaniv Leviathan、Moab Arar 和 Shlomi Fruchter。扩散模型是实时游戏引擎。载于国际学习表征会议，2025年。2，3[31] Wan团队。Wan：开放且先进的大规模视频生成模型。arXiv预印本 arXiv:2503.20314，2025年。2[32] Guanzhi Wang、Yuqi Xie、Yunfan Jiang、Ajay Mandlekar、Chaowei Xiao、Yuke Zhu、Linxi Fan 和 Anima Anandkumar。Voyager：基于大语言模型的开放式具身智能体。《机器学习研究汇刊》（TMLR），2024年。2[33] Weijie Wang、Haomin Zhang、Ye Li、Akide Liu、Zeyu Zhang、Haoyu Zhao、Yefei He、Donny Y. Chen、Feng Chen 和 Bohan Zhuang。任时任地：以世界状态为中心的智能体视频生成框架。arXiv预印本 arXiv:2610.20571，2026年。3[34] Weijia Wu、Zeyu Zhu 和 Mike Zheng Shou。通过多智能体思维链规划实现自动化电影生成。arXiv预印本 arXiv:2503.07314，2025年。2，3[35] Zhexiao Xiong、Yizhi Song、Hao Kang、Qing Yan、Liming Jiang、Jenson Yang、Zhoujie Fu、Stathi Fotiadis、Angtian Wang、Zichuan Liu、Bo Liu、Yiding Yang、Xin Lu 和 Nathan Jacobs。ActWorld：通过动作感知记忆，从可探索世界模型走向交互式世界模型。arXiv预印本 arXiv:2606.17730，2026年。2，3[36] Xindi Yang、Baolu Li、Liam Lee、Zhenfei Yin、Songxin Zhang、Zhuoyang Song、Xu Jia、Jianfei Cai、Tien-Tsin Wong、Bingyi Jing 和 Mengyue Yang。Oneira：从开放式生成走向视频世界模型中的开放世界交互。arXiv预印本 arXiv:2610.01614，2026年。2，3[37] Zhuoyi Yang、Jiayan Teng、Wendi Zheng、Ming Ding、Shiyu Huang、Jiazheng Xu、Yuanming Yang、Wenyi Hong、Xiaohan Zhang、Guanyu Feng 等。CogVideoX：采用专家Transformer的文本到视频扩散模型。载于国际学习表征会议（ICLR），2025年。2[38] Kaining Ying、Hengrui Hu、Siyu Ren、Jiamu Li、Fengjiao Chen、Ziwen Wang、Xuezhi Cao、Xunliang Cai 和 Henghui Ding。WBench：用于交互式视频世界模型评估的综合性多回合基准。 *arXiv* 预印本 *arXiv:2605.25874*，2026年。 6
[39] Eric Zelikman、Eliana Lorch、Lester Mackey 和 Adam Tau-man Kalai。自教优化器（STOP）：递归式自我改进代码生成。在 语言建模会议（*COLM*），2024年。 2[40] Jingzhe Zhang、Peixuan Zhang、Genying Li、Bing Ma、Shuchen Weng、Si Li 和 Boxin Shi。CineEvolve：用于电影级多镜头视频生成的智能体自进化框架。 *arXiv* 预印本 *arXiv:2610 29590*，2026年3。。[41] Zangwei Zheng、Xiangyu Peng、Tianji Yang、Chenhui Shen、Shenggui Li、Hongxin Liu、Yukun Zhou、Tianyi Li 和 Yang You。Open-Sora：让高效视频制作惠及所有人。 *arXiv* 预印本 *arXiv:2412.20404*，2024年。 2[42] Shangwen Zhu、Qianyu Peng、Zhao Pu、Zhilei Shu、Xian-grui Ke、Zhaohu Xing、Zizhao Tong、Zeqing Wang、Xinyu Cui、Zian Zheng、Huangji Wang、Jian Zhao、Yeying Jin、Fan Cheng 和 Ruili Feng。Incantation：自然语言作为多实体视频世界模型的动作接口。 *arXiv* 预印本 *arXiv:2605.18601*，2026年。 2，3[43] Sihan Zhuang、Xinyuan Chen、Tianfan Xue 和 Yaohui Wang。CausalMotion：将结构化物理推理作为免训练视频生成的关键帧与轨迹引导。 *arXiv* 预印本 *arXiv:2606.14317*，2026 年。 3

CineAct：交互状态智能体驱动的可交互分支式电影生成

补充材料

骨干网络规格

所有推理智能体和自动评估器均运行于多模态语言模型 Hy4 [29]。对于图像编辑，我们使用 Qwen-Image-2。1 [25]以及蒸馏版 FLUX。2 [klein] 4B [1]。对于视频合成，MiniMax-H 的 FL2V 和 Ref2V 接口均3 [22] 使用 8 步蒸馏变体，而 LTX-2。5 [19] 以完整 BF16、未经蒸馏的两阶段模式运行（30 步）。

系统表示与执行流程

这里详细介绍第 4 节中的三种表示，并总结算法1中的一次执行回合。图 S6展示了三层露台幼苗任务的一个具体示例：面板 (a)–(d) 分别给出了布局 L、初始视角、场景图G_1和资产裁剪图 $\mathcal{A}_1$，而面板 (e)–(f) 则追踪了两次展开。在每个回合中，规划器会检查G_t的前置条件，仅检索目标视角和画外实体裁剪图（第 4–5 行），在渲染前验证目标帧 V_t （第 7–12 行），并将 $\mathrm{Tail}(V_t)$ 仅写回已访问的露台（第 13 行）。

•空间布局与摄像机（L）：空间是不重叠的二维轴对齐矩形，共用墙面上设有门户，每个空间由一台固定摄像机观察（65°水平视场角，摄像机放置在没有门户的墙面上，并朝向一个出口）。几何投影可推导出每个门口在屏幕上的方位（左侧、右侧或远墙），用于向绘制器/导演器提供提示并由检查器进行验证。

• 资产库（$\mathcal{A}_t$）：将每个空间的视觉记忆与初始化时提取的静态实体身份裁剪图相结合。只有实际视频片段 $I_{t+1} = \mathrm{Tail}(V_t)$（绝不会使用中间目标图像 $I_t^\star$）会更新已访问空间在 $\mathcal{A}_{t+1}$ 中的记忆；条件控制使用目标空间的最新记忆，以及所操作实体和画外实体的裁剪图，并跳过 I_t中已经可见的演员。

• 场景图（G_t）与符号引擎：节点存储位置/所有权指针（在……中、在……上、持有）、可携带性、承载容量和状态属性。在渲染之前，规划器通过广度优先搜索检查前置条件和目标可达性，推进独立的规划图G_{t+1}，并将差异计算（G_t，G_{t+1}）以确定性方式写入 Spec$_t$。

评估指标与人工–自动化一致性

表 S1 列出了多步交互式执行中使用的全部十个逐步问题，图S1–S2 展示了四种消融设置下院子浇水任务的全部六个步骤。对于四个共同项目（动作、状态、进入、退出）而言，Qwen-Image-2。1 +MiniMax-H3、Hy4 与盲人人类评分的二元一致性为 73.0%（三等级精确一致性为 64.4%），并且遵循与论文正文表 1 和图 3 相同的消融顺序。表S2 将这项盲人人类研究扩展到相同 10 个故事上的两种 LTX-2.5 配对。CineAct 在全部三种配对中均排名第一（.842、.748 和 .614； +23.1、 +18.0 和 +17.6 分，高于基础线），与表S3中的自动化跨骨干网络结果一致。

表 S1. 多步交互式执行中三个一致性维度的逐步评估项目。自动化 Hy4 评估会对全部十个项目进行评分； ⋆ 标示了用于快速人工标注的四个代表性核心项目（动作、状态、进入、退出）。

ID	Dim.	评估问题
A1⋆	动作	所请求的动作是否清晰可见地执行并保持?
W1	世界状态	操作对象是否为准确的指定实例?
W2	世界状态	实体是否没有复制或消失?
W3⋆	世界状态	物体是否被拿着或支撑在表面是否正确?
W4	世界状态	请求的状态属性是否按预期发生了变化?
W5	世界状态	被移除或带走的物体是否保持缺失状态?
S1⋆	空间	从目标空间到达乐彩票是否从正确的入口侧进入?
S2⋆	空间	朝正确的门口方向离开源空间?
S3	空间	镜头切换是否仅发生在跨空间步骤上?
S4	空间	房间布局和相对尺度是否保持一致?

接口比较：Target-Frame FL2V 与 Ref2V

表S4 给出了同一组 10 个故事中，主论文图 4(d) 背后完整的人类与 Hy4 数量。当某一步骤可以被清晰评分时，我们的表示方法能让两种界面都获得提升：在 Hy4 下，Ref2V 从基线中的 .488/.621/.900/.713 提升至 CineAct 中的 .725/.866/.959/.875（动作/世界/空间/总体）。瓶颈在于构图稳定性。在每个设置的 49 个状态步骤中，**FL2V 有 0/196 个不明确项目（0.0%），而 Ref2V 从基线到 Cine-**

算法1： CineAct 在第 t 回合的单轮执行

输入： 当前帧 I_t、布局 L、资产 $\mathcal{A}_t$、场景图 G_t、用户输入 u_t

$a_t \leftarrow$ 解释器$(u_t, I_t, \mathcal{A}_t, G_t)$ // 将输入定位到实体和门户ID

若 ¬检查前置条件(G_t, a_t) **或** ¬目标可达性检查$(G_t \oplus \Delta_t, g)$ **则**

  返回 拒绝(u_t) // 像素合成前的预渲染门控

$G_{t+1} \leftarrow G_t \oplus \Delta_t$; $\text{Spec}_t \leftarrow$ 差异计算$(G_t,\ G_{t+1},\ L)$ // 规划图 +确定性差异

$R_t \leftarrow$ RetrieveRefs$(\mathcal{A}_t,\ I_t,$ 规范$_t)$ // 目标空间记忆 +不可见裁剪区域

fb $\leftarrow \emptyset$

对 尝试 $= 1 \ldots 3$ **执行**

  $I_t^\star \leftarrow$ 绘制器$(I_t,\ R_t,$ 规格$_t,$ fb)

  $(v,$ fb$) \leftarrow$ 检查器$(I_t^\star,\ I_t,\ R_t,$ 规格$_t,\ L)$ // 合成候选目标结束帧 // 语义 + 门口侧 投影检查

  如果 $v =$ **通过，则中断**

$C_t \leftarrow$ CameraRule$(L,\ r_s \to r_d)$ // 1 个固定镜头（相同）或 2 个镜头 + 剪辑（交叉）

$V_t \leftarrow$ 导演器$(I_t,\ I_t^\star,$ 规格$_t,\ C_t)$ // Frozen FL2V 视频渲染

$I_{t+1} \leftarrow$Tail(V_t); $\mathcal{A}_{t+1} \leftarrow$ UpdateMemory$(\mathcal{A}_t,\ r_d,\ I_{t+1})$ // 仅刷新已访问空间记忆

返回$(V_t,\ I_{t+1},\ \mathcal{A}_{t+1},\ G_{t+1})$

表 S2. 多步交互式执行中，不同图像和视频骨干组合的盲人人类评估（故事等权平均值，三档 0/0.5/1 评分）

图像	视频	设置	动作状态进入	Exit	All
Qwen2.1 H3		基线 .660	.674 .306	.389	.611
		+布局 .701	.647 **1.000**	**1.000**.761	
		+资产 .777	.782**1.000**	.982	.832
		CineAct **.802**	**.828**	**1.000**	**1.000 .842**
Qwen2.1 LTX-2.5		基线 .613	.676 .310	.279	.568
		+布局 .569	.668 .626	.524	.627
		+资产 .548	.789 .821	**.914**.711	
		CineAct **.619**	**.808 .907**	.871**.748**	
FLUX.2 LTX-2.5		基线 .479	.449 .574	.148	.438
		+布局 .497	.395 .724	.638	.506
		+资产 .455	.434 .650	**.736**.503	
		CineAct **.565**	**.616 .829**	.679**.614**	

Act（51/196，即 26.0%）。如果没有经过验证的目标帧来锚定镜头（图 4(d)），Ref2V 会在接取过程中放大画面并裁掉桶和左侧机械设备，导致用户无法在下一回合检查或选择这些物体。

成对分支执行：逐任务得分和 rollout 示例

桌子 S5 详细列出主论文表 4 中的 13 个成对分支任务（共 52 次 rollout，每个任务有两条路径），并采用与表 1 和表 S3相同的逐步 10 项 Hy4 协议。对全部 13 个任务进行等权平均后，CineAct 在路径 1 上得分为 .884（动作 .802、世界 .903、空间 .889，而基线为 .765），在路径 2 上得分为 .871（.770/.896/.872，而基线为 .847）。在 10 个多空格任务（2–4 个空格）中，CineAct 在路径 1 上达到 .878，在路径 2 上达到 .838（而基线分别为 .740 和 .820）。图 S3–S7 展示了五个示例，涵盖 1–4 个空格：点亮祭坛、封好袋子、运送木炭、给梯田浇水，以及固定实验室支架。

分支可接受性：空间分层与反事实细分

表S6 将主论文表 2–3 中的 40 个故事分支可接受性基准按空间数量和反事实错误类型进行划分。我们针对每种空间数量评估 10 个故事；从 627 个提案中筛除 56 个含糊项目（8.9%）后，剩下 571 个经核验的查询。在开放设置下，纯像素基线在单空间查询中有 47.8% 的回答为不确定，在 2–4 个空间中这一比例为 87.1%–95.1%，因为当前帧中不存在屏幕外的房间。CineAct 在所有空间数量下都将弃权率保持在 2.7%–5.1%。在强制二选一检查中，基线拒绝了 97.0% 的无效提案，主要是因为几乎拒绝了所有提案（包括有效移动）；而 CineAct 同时保持了较高的有效接受率和无效拒绝率。

下面是表格中三类反事实类别的具体示例S6。（1）不存在的地点要求移动到未连接或不存在的房间（例如，从石厅前往 L 中不存在的上层阳台）；加入布局几何后，未被拒绝的错误率从95.0%降至23.8%。（2）不存在的物体指代 $\mathcal{A}$中缺失的物品（例如，用海绵擦拭甲板烟灰，而不是使用红色或蓝色抹布）；+资产将这些错误率从91.3%降至30.0%。（3）前置条件失效违反容量、位置或状态规则（例如，在装满两个水囊之前就向石盆中倒水），这需要 Cine-Act 的符号状态检查（4.0%，而 +资产为37.4%）。

表 S3。在多步交互式执行任务中，针对全部四种图像–视频骨干模型组合进行的详细 10 项 Hy4 评估。对于 FLUX.2 [Klein] 4B + MiniMax-H3，仅评估了基线和 CineAct。

			动作	世界状态						空间				
图像	视频	设置	A1*	W1	W2	W3*	W4	W5	平均	S1*	S2*	S3	S4	平均	总体
Qwen2.1	H3	基线	.630	.946	.835	.719	.390	.725	.760	.492	.596	.712	.922	.807	.748
		+布局	.647	.973	.850	.637	.460	.688	.734	**.932**	**.990**	.745	**.931**	.876	.777
		+资产	.670	.970	.825	.755	.489	.682	.780	.910	.924	.700	.916	.864	.791
		CineAct	**.767**	**.993**	**.870**	**.879**	**.652**	**.919**	**.878**	.925	.941	**.800**	.904	**.877**	**.862**
FLUX.2 [1]	H3	基线	.517	**.889**	**.619**	.473	.429	.362	.547	.649	.736	.711	.925	.823	.636
		CineAct	**.600**	.833	.503	**.642**	**.537**	**.598**	**.596**	**.804**	**.875**	**.817**	**.930**	**.889**	**.691**
Qwen2.1	LTX-2.5 [19]	基线	.586	.963	.704	.683	.377	.525	.677	.713	.619	.476	**.935**	.698	.675
		+布局	.586	.925	**.827**	.539	.533	.537	.679	.662	.883	**.605**	.888	**.746**	.699
		+资产	.657	.950	.684	.745	**.644**	.733	.747	**.917**	**1.000**	.447	.865	.698	.731
		CineAct	**.662**	**.971**	.813	**.827**	.512	**.925**	**.827**	.849	.821	.470	.888	.712	**.766**
FLUX.2 [1]	LTX-2.5 [19]	基线	.471	**.979**	.568	.415	.315	.400	.502	.457	.762	**.657**	**.936**	**.751**	.586
		+布局	.557	.852	**.633**	.429	.360	.558	.541	**.768**	.929	.453	.848	.679	.607
		+资产	.508	.882	.499	.544	.329	.433	.535	.721	.857	.425	.799	.642	.576
		CineAct	**.602**	.971	.556	**.695**	**.569**	**.725**	**.687**	.636	**.964**	.420	.855	.686	**.675**

表 S4。同一 10 个故事上的 Target-Frame FL2V 与 Ref2V 对比。FL2V 的人工评估结果与主论文图 3 一致；Ref2V 和 Hy4 报告故事等权平均值。

		可判定分数				**状态不明确**
轨迹界面		基线	+布局。	+助手。	CineAct	**比率 (%)**
人工 FL2V（我们的方法）		.611	.761	.832	.842	**0.0%**
人工 Ref2V		.538	.624	.781	.870	26.0%
Hy4	FL2V（我们的方法）	.733	.785	.739	.856	–
Hy4	Ref2V	.713	.771	.836	.875	–



图 S1。**水袋拾取与搬运**（庭院浇水，步骤 1–3，Qwen-Image-2.1 和 MiniMax-H3）。基线和 +布局都将两个水囊留在挂钩上。 +资产在接取时复制了这对水囊，并在返回时将水囊带到篮子外。CineAct 将两个水囊都装在篮子里搬运。插图放大了视频帧中的框选区域。



图 S2。**门选择和浇水**（庭院浇水，第 4–6 步）。基线在向西行进时穿过中央拱门。CineAct 选择左门，并用蓝色水袋给西侧树苗浇水（第 4 步插图）。在东院，水从基线的裸手中流出。借助 +布局，树苗周围的土壤会在 7.25 至 10.00 秒之间自行变湿。 +资产用蓝色水袋给东侧树苗浇水。CineAct 使用绿色水袋，并将蓝色水袋留在篮子里。第 6 步的剪辑使 CineAct 最终穿过门口的动作无法确认。

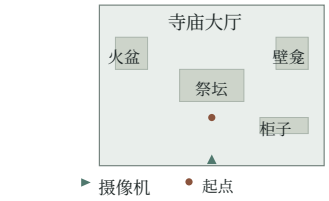
表 S5. 每任务 10 项 Hy4 13 个任务的成对分支执行分数，分别报告路径 1 和路径 2 下 r
基线和 CineAct（动作、Wo rld、空间、全部）。

	基线路径 1				基线路径 2				CineAct 路径 1				CineAct 路径 2			
Task	动作。	世界间隔。		All	Act.	世界间隔。		All	Act.	世界间隔。		All	Act.	世界间隔。		All
阳台花卉	.750	.921	1.000	**.919**	.833	.962	1.000	.955	.625	.861	.875	.833	.833	1.000	1.000	**.976**
祭坛蜡烛	.500	.875	.900	.833	.750	.821	1.000	.854	.700	.913	.900	**.882**	1.000	.929	1.000	**.958**
公寓包	.750	.722	1.000	.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**	1.000	1.000	1.000	1.000
图书馆卷轴	.556	.629	.741	.664	.600	.711	.786	.724	.722	.833	.933	**.861**	.600	.684	.821	.724
海滩浮标	.500	.364	.950	.620	1.000	1.000	.875	**.957**	.750	.821	.900	**.839**	1.000	.900	.875	.905
礁区海带	.571	.654	.667	.647	.875	.923	.625	.820	.786	.911	.850	**.873**	1.000	1.000	.889	**.964**
Sanctum 木炭	.643	.778	.857	.791	.500	.727	.861	.756	.857	.920	.905	**.906**	.700	.896	.825	**.847**
地窖钩	.667	.885	.500	**.750**	.500	.875	.875	.821	.667	.773	.625	.705	.500	.844	1.000	**.850**
锅炉煤	.800	.861	.833	.843	.750	.967	.917	**.919**	.900	.929	.917	**.921**	.750	.833	.833	.823
楼梯毯	.875	.923	.917	.914	.700	.667	.857	.743	1.000	1.000	.875	**.953**	.600	.925	.821	**.846**
梯田幼苗	.625	.560	.609	.589	.600	.809	.808	.782	.875	.950	.913	**.926**	.700	.900	.643	.782
甲板栏杆	.429	.900	.917	.837	.571	.740	.850	.760	.929	.919	1.000	**.941**	.571	.904	.842	**.837**
实验室支架	.556	.823	.708	.742	.875	.929	.913	**.915**	.611	.914	.865	**.852**	.750	.833	.792	.805
均值（2–4 个支架）	.622	.738	.770	.740	.697	.835	.837	.820	.810	.897	.878	**.878**	.717	.872	.834	**.838**
均值（全部，n=13）	.632	.761	.815	.765	.735	.856	.874	.847	.802	.903	.889	**.884**	.770	.896	.872	**.871**

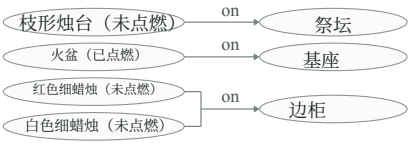
表 S6. 开放式评分标准下的空间分层分支可接受性（n=40 个故事，每层 10 个，共保留 571 个查询）及各类别未拒绝反事实率。

	查询计数				开放式弃答／不确定率（%）				未拒绝的无效项（占类别的%）			
层／类别	有效对象	地点	前置条件	基线	+布局	+助手	CineAct		基础。	+布局。	+助手。	CineAct
单一空间（n=10，总计=138）	60	20	20	38	47.8	53.6	23.9	**5.1**	–	–	–	–
2 个空格（n=10，总计=142）	60	20	20	42	88.7	82.4	36.6	**3.5**	–	–	–	–
3 个空格（n=10，总计=147）	56	20	20	51	87.1	87.8	27.9	**2.7**	–	–	–	–
4 个空格（n=10，总计=144）	61	20	20	43	95.1	87.5	30.6	**4.9**	–	–	–	–
不存在的物体（n=80）	–	80	–	–	–	–	–	–	80.0	91.3	30.0	**6.3**
不存在的地点 (n=80)	–	–	80	–	–	–	–	–	95.0	23.8	5.0	**0.0**
被破坏的前置条件 (n=174)	–	–	–	174	–	–	–	–	70.7	83.9	37.4	**4.0**
汇总总数 (n=40，总计=571)	237	80	80	174	80.0	78.1	29.8	**4.0**	78.7	71.3	27.8	**3.6**

目标：点燃寺庙祭坛上的蜡烛。



（a）空间布局

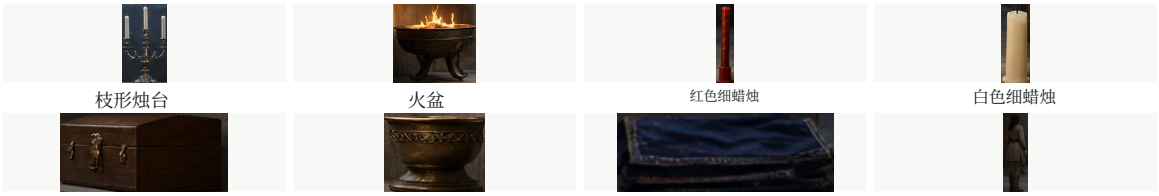


(c) 初始场景图（节选）



寺庙大厅

(b) 初始房间视图



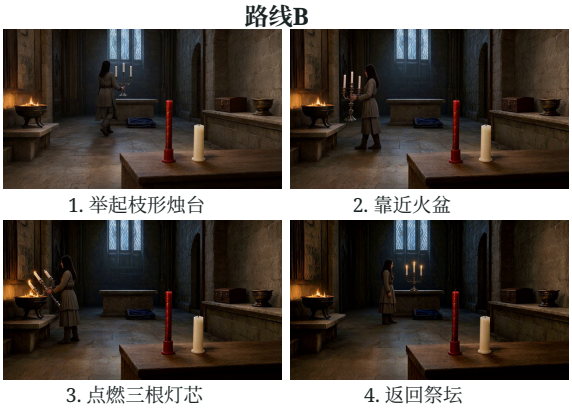
（d）已登记资产引用

路线 A



（e）用红色细蜡烛将火焰带到祭坛，然后归还细蜡烛。

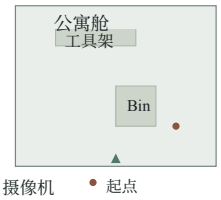
路线B



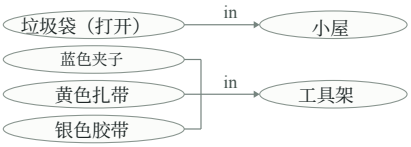
（f）将枝形烛台带到火盆处，然后将其带回祭坛。

图 S3。**点燃祭坛上的枝形烛台。**共享开场场景中有一个未点燃的枝形烛台和一个燃烧的火盆。路线 A 使用红色细蜡烛将火焰带到祭坛处。路线B 将枝形烛台带到火焰处。每条路线每回合展示一个代表性画面，每行内从左到右阅读。

目标：在公寓中封好垃圾袋。



（a）空间布局



(c) 初始场景图（摘录）



公寓舱

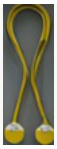
(b) 初始房间视图



垃圾袋



蓝色夹子



黄色扎带



银色胶带



角色

（d）已登记资产引用

路线 A



1. 拿起蓝色夹子



2. 夹紧袋口

（e）收拢袋口，并用蓝色夹子夹紧。

路线B



1. 取黄色扎带

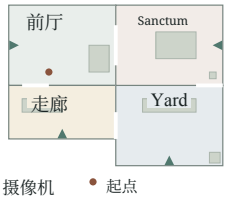


2. 连接磁性末端

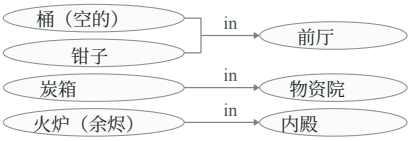
（f）将黄色扎带缠绕在袋颈处，并连接其磁性末端。

图 S4。**在公寓中封袋**。两条路径都从同一个打开的袋子和壁挂式工具架开始。它们所选用的袋颈紧固件不同。每条路径每回合显示一个代表性画面，每行内从左到右阅读。

目标：将木炭添入内殿的火炉中。



(a) 空间布局



(c) 初始场景图（节选）



(b) 初始房间视图



(d) 已登记资产引用



（e）用铜桶装木炭，然后挂起内殿门旁的空桶。

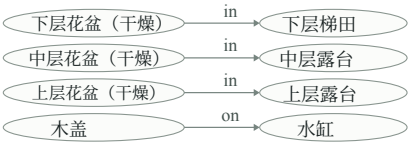
(f) 用钳子装满麻袋，然后将麻袋里的木炭倒入火炉，并将两件工具都归还到院子。

图 S5. **将木炭送至内殿**。两条路径都从同一处物资院取木炭。桶的路径要求将空桶留在内殿门口。麻袋的路径要求将麻袋和钳子归还到院子。每条路径按从左到右的顺序，在每行展示每回合的一幅代表性画面。

目标：为全部三层梯田上的幼苗盆取水。



（a）空间布局



（c）初始场景图（节选）



下层梯田



中层露台



上层露台

(b) 初始房间视图



（d）已登记资产引用

路线 A



（e）用水罐给三层梯田全部浇水，并将水罐放回下层梯田的篮子中。

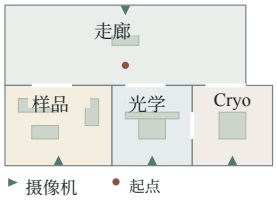
路线B



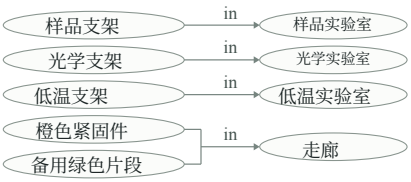
(f) 在每层梯田为蓝色水勺添满水，然后将它放在上方岩石上。

图 S6. **在三层梯田的幼苗盆浇水。** 水罐路线最终回到下层梯田，水罐和篮子也留在那里。水勺路线最终到达上层梯田，蓝色长柄勺将留在岩石上。每条路线都展示了每个回合的一幅代表性画面，每行内从左向右阅读。

目标：固定全部三个实验室中的设备支架。



（a）空间布局



(c) 初始场景图（摘录）



中央走廊　样品实验室



光学实验室　低温实验室

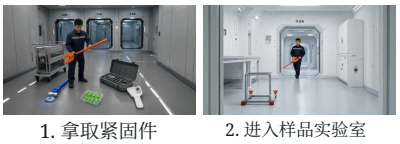
(b) 初始房间视图



橙色紧固件　备用绿色夹子　蓝色紧固件　白色紧固件　工具箱　技术员

（d）已登记资产引用

路线 A



1. 拿取紧固件　2. 进入样品实验室



3. 按下样品夹具　4. 返回走廊



5. 进入光学实验室　6. 按下光学夹具



7. 打开检修门　8. 进入低温技术区



9. 按下最终夹具

(e) 将现有的红色夹具压平，使其呈扁平状。橙色紧固件将现有的红色夹具压平，并将其留在最后一个支架处。

路线B



1. 拿取夹具盒　2. 进入样品实验室



3. 安装样品夹具　4. 返回走廊



5. 进入光学实验室　6. 安装光学夹具



7. 进入低温技术区　8. 安装最终夹具

（f）将凸起的红色夹具替换为绿色备用夹子，并将零件盒留在最后一个支架处。

图 S7。**在三个实验室中固定支架**。每个支架开始时都装有四个凸起的红色夹具。一条路线使用橙色紧固件压下现有夹具；另一条路线则将它们替换为绿色备用夹子。每条路线在每个回合展示一帧代表性画面，每行内从左到右阅读。