T2VTree：面向智能体辅助的思维转视频创作的以用户为中心的可视分析

郑卓云[1], [2*] 董宇[1*] 梁高荣[1,2] 李冠[1,2] 单桂华[1,2,3] 程诗宇[1]田栋[1,2] 周建龙[4] 梁杰[4]

arXiv:2602.08368v1 [cs.MM] 9 Feb 2026

摘要

生成模型已大幅扩展了视频生成能力，但实际的从构想到视频的创作仍是一个多阶段、多模态且决策密集的过程。然而，现有工具要么将中间决策隐藏在反复重跑之后，要么暴露操作者级别的工作流，使探索轨迹难以管理、比较和复用。我们提出*T2VTree*，一种面向智能体辅助的思维转视频创作的以用户为中心的可视分析方法。*T2VTree* 将创作过程表示为树形可视化。树中的每个节点都将可编辑规范（意图、引用的输入、工作流选择、提示词和参数）与由此产生的多模态输出绑定在一起，使细化、分支和溯源检查可直接操作。为减轻决定下一步做什么的负担，一组协作智能体将步骤级意图转化为可执行计划，该计划在执行前保持可见且可由用户编辑。我们进一步实现了一个可视分析系统，它将分支式创作与就地预览和用于汇聚式组装的拼接相结合，从而能够在不离开创作上下文的情况下完成端到端的多场景创作。我们通过两项多场景案例研究和一项对比用户研究展示了*T2VTreeVA*，说明*T2VTree*可视化与可编辑的智能体规划如何在真实创作工作流中支持可靠的细化、局部比较和实际复用。T2VTree 可在以下地址获取：https://github.com/tezuka0210/T2VTree。

关键词： 生成式视频创作、可视分析、人–AI协作、多模态交互

1 引言

生成式视频模型的最新进展大幅拓展了视频生成能力，使用户能够根据文本描述或视觉输入生成短片。尽管这些技术在视觉保真度和运动连贯性方面取得了显著进步，但创作有意义的视频内容仍是一个复杂的创作过程。创作者很少会一开始就拥有对目标视频完整且可执行的规格说明。相反，他们通常从一个不断演变的心理目标出发，并通过交互逐步将其外化、构建并完善，最终才能获得令人满意的结果。

在实际工作流中，视频创作本质上是一个多阶段、多模态的过程。实用的流程管线通常从建立视觉锚点开始，例如生成或选择参考图像、首帧或关键帧。随后，创作者以这些锚点为条件生成运动和场景演化，并反复调整提示词、约束和参数，以修正不理想的运动、构图漂移或风格不一致。通常还需要进行中间操作，包括对图像或帧进行局部编辑、区域级细化，以及超分辨率放大或时间平滑等质量增强步骤。除这些操作外，有意义的视频内容通常还需要音频素材，包括背景音乐、音效、旁白和时间对齐。最后，创作者通过裁剪、拼接和同步，将多个片段和音轨组装起来，形成连贯的结果。这个流程管线通常涉及多个模型和工具，因此必须协调各阶段的决策，并保留可重复使用的中间结果。

重要的是，即便是详尽的提示词也无法保证输出符合创作者的意图。错位通常只有在检查中间结果之后才会显现，这迫使创作者诊断偏差，并逐步引入约束以引导后续生成。随着细化不断累积，创作者会重新审视上游选择、进行分支以探索备选方案、比较相互竞争的变体，并保留成功的局部决策以供复用。因此，思维到视频创作是一个探索性且决策密集的过程。然而，现有系统在以有用的抽象层级表示演化意图、借助显式中间结果追踪多阶段决策，以及以可检查且可复用的方式管理分支探索方面，仅提供有限支持。

现有的生成式视频创作工具通常遵循两种使用模式。一种是端到端商业平台（例如 Sora 和可灵），它们提供精简的交互：创作者可以快速根据文本或图像生成片段，但中间决策和变体往往难以在多次迭代中检查和复用。另一种是构建于开源模型和插件生态之上的基于节点的工作流工具（例如 ComfyUI），它们暴露可配置的算子图，并允许创作者集成专门操作和社区工作流，但需要大量技术专长，且往往使创作轨迹分散在图、参数和外部文件之中。在这两类工具中，非线性探索在创作者实际工作的层面上仍然难以管理。创作者缺乏一种结构化方式，使演化意图、中间状态及其推导关系随时间保持可见且可复用。这种过程可见性的缺失，使得难以有把握地迭代细化想法、理解备选结果如何与先前决策相关联，以及在扩展到更多场景时可靠地复用有效决策。

从视觉角度来看，思维到视频创作可以视为一个人机协同探索问题。创作者提出并修改候选方向，将中间结果作为决策依据，并在不确定性下反复做出选择。要解决这一问题，需要通过可视化表示和交互将意图外显、呈现决策溯源，并将分支感知比较和轨迹感知复用作为核心创作操作。

本文提出 *T2VTree*，一种以用户为中心的视觉分析方法，用于思维到视频的创作。*T2VTree* 将创作过程外化为树状表示，其中每个节点都记录一个持久化的创作状态。一个状态将

*单位：[1] 中国科学院计算机网络信息中心，中国北京；[2] 中国科学院大学，中国北京；[3] 杭州高等研究院，中国科学院大学，杭州，中国；[4] 悉尼科技大学，澳大利亚新南威尔士州悉尼。电子邮箱：{zyzheng, dongyu, chengshiyu, tian- dong, sgh}@cnic.cn；lianggaorong25@mails.ucas.ac.cn；liguan@sccas.cn；{Jianlong.Zhou, Jie.Liang}@uts.edu.au。郑卓云和董宇对本研究贡献相同。单桂华为通讯作者。

步骤输入（例如意图、引用的素材、工作流选择、提示词和关键参数）与跨模态（图像、视频和音频）的相应输出共同置于一处，使细化与探索得以累积，而不会覆盖先前的工作。这种树形表示使分支变得明确且可检查，让创作者能够回溯，在共享上下文下比较备选方案，并在扩展到其他场景时复用成功的部分配置。

为了让创作更加以用户为中心，并减少手动组装工作流的需求，*T2VTree*引入了智能体辅助规划，其操作层级对应创作者意图中的创作操作，而非算子图构建。创作者用自然语言说明目标，也可以提供视觉参考；随后，协作智能体会选择合适的工作流模块，并生成用于执行的草稿提示词和关键参数，将这些输入转化为可执行的后续步骤。*T2VTree*不会将这种自动化视为黑箱，而是将智能体生成的计划呈现为绑定到节点且可编辑的结果，方便创作者就地审阅、修改并复用相关决策。

结合决策轨迹可视化、上下文内的多模态预览和基于拼接的组装，这一设计支持创作者在整个创作闭环中由人引导地进行探索，使其能够迭代完善意图、分支探索并比较不同方案，最终无需离开创作上下文即可完成一段多场景视频。

本工作的贡献如下：

- 我们报告了一项针对经验丰富的创作者开展的形成性研究，并提炼出支持思维到视频创作的可视分析设计需求，包括意图外化、可追溯的决策溯源，以及分支感知比较与复用。
- 我们提出了*T2VTree*，这是一种基于树的视觉表示，用于记录将步骤规格与多模态输出相连接的持久化状态，从而支持在多场景工作流中进行检查、分支、局部比较和复用。
- 我们实现了*T2VTreeVA*，这是一种可视分析系统，集成了可编辑的智能体辅助规划、分支感知创作、上下文中的多模态预览，以及基于拼接的组装；并通过多场景使用情境和针对创作者开展的定性用户研究对其进行评估。

2 相关工作

2.1 生成式多模态创作工作流

近年来，文生视频和图生视频生成技术的进步，推动了学术界和工业界中多种生成式视频创作流程管线的发展 [1,3,20,34]。

在实践中，创作者会使用由文生图模型生成或从现有素材中提取的视觉锚点——例如参考图像、首帧或关键帧——[19,22,38,42],并通过迭代细化提示词、参数或参考素材，以缓解构图漂移和风格不一致 [23]。因此，图像生成既是视频合成的前提，也是支持迭代决策的中间表示 [15]。

生成式视频创作本质上是探索性的，需要反复试错：创作者经常重新生成内容，对中间图像或帧进行局部细化或质量增强 [25,30,46],，并在后续阶段复用细化后的结果 [47]。通过设置起始帧/结束帧约束、指定摄像机运动，或进行插值与平滑处理，可以进一步提升时间连贯性 [23]。创作目标在一开始时很少能完全明确，而是在持续检查与比较的过程中逐步厘清 [5]。

除了视觉内容，音频通常会独立生成 [10,27]，并在后处理中同步 [6],，而对跨模态语义或时间对齐的支持有限。

尽管取得了这些进展，大多数生成式视频工具仍以反复的“生成–检查–重试”循环为中心 [2,35],，将中间结果和放弃的备选方案视为转瞬即逝的产物，迫使创作者依赖记忆或外部文件来管理分支和参数变化，增加了认知负担 [33,37]。

基于节点或以算子为中心的创作界面，将生成、编辑和处理步骤表示为可执行工作流中的可组合节点，从而提升了可控性，[4,13],为专业用户提供细粒度控制 [7,8,26]。然而，这类界面仍然侧重于执行和计算拓扑，带来了较高的认知负担 [43]。由于缺乏语义层面的引导，用户必须管理复杂的节点语义、参数耦合、执行顺序，并维持图像、视频和音频之间的一致性。

工作流图通常只编码数据流依赖，因此多模态创作过程被分散在不同的子图中。用户必须推断各个输出之间的对应关系，而不断演变的意图、决策依据和放弃的替代方案很少被明确记录 [12,29,31]。这种碎片化增加了认知和管理负担，限制了对以决策为驱动的分支探索的支持。总而言之，尽管现有工作流系统能够有效呈现执行结构，但在将生成式视频创作表示为一种以决策演变、分支探索和证据积累为核心的创作过程方面，仍存在局限。

2.2 人机协作与创意可视分析

在人机协作和可视分析领域，研究者长期以来一直在探讨可视化如何支持用户在复杂且不确定的任务中进行推理、探索和决策 [14]。大量研究表明，创意生成与设计活动本质上是非线性且发散的，其特点是分支探索，而非单路径优化 [9,11,21,44,45]。然而，大多数传统工具依赖线性时间排序历史，使得早期探索状态难以重新访问、比较或复用，进而限制了并行探索和反思性迭代。

近期的可视分析研究探讨了如何通过可视化交互历史来支持生成式创作中的反思 [8,28]。PromptThis [16] 通过可视化提示词–图像关系，帮助用户回顾提示词历史；PromptAid [24] 则通过协调可视化和扰动策略，支持交互式提示词优化。在更广泛的创意场景中，The Joy of Co-Painting [29]和Sel3DCraft [41] 展示了如何利用基于来源和结构感知的表示，支持图像与三维生成工作流中的可追溯探索。总的来说，这些研究表明，历史与分支的结构化表示对于降低试错成本、支持迭代探索以及增强用户的掌控感至关重要。

尽管取得了这些进展，现有工作主要针对数据分析、文本生成或静态图像与三维创作，对于多阶段、多模态视频生成在流程管理方面的需求支持有限。关于视频可视分析的既有研究强调了对时间结构复杂且具有多模态特征的视频数据进行推理所面临的挑战 [17,39],这类任务通常会带来更高的计算成本、更长的延迟以及显著的跨阶段依赖。类似地，这种固有的复杂性也加剧了生成式视频工作流中的试错成本 [32]。由于缺乏系统性的版本管理、分支和决策追踪机制，创作者往往不得不手动记录参数和中间结果，从而增加认知负担和操作负担。

随着参与与规划与执行的智能体出现，这一挑战变得更加突出。如何将智能体生成的建议、自动化决策过程及相应的生成式证据整合到可解释、可追溯的视觉结构中，仍是一个有待解决的问题 [18,36,40]。

受这些不足的启发，*T2VTree* 将生成式视频创作建模为一种以决策和证据为核心的人机协同可视分析过程，而非仅仅关注执行步骤或最终输出。

3 形成性研究与设计需求

为确保*T2VTree*反映真实的创作实践，而非理想化的一次性生成流水线，我们开展了一项形成性研究，以了解创作者如何使用现有生成式工具制作短视频、他们会反复执行哪些步骤，以及哪些环节容易出现问题。我们的目标并非穷尽所有制作行为，而是提炼出一组稳定且常见的创作操作，将其作为一流的界面操作予以支持，同时识别促使我们进行设计的关键交互缺口。

3.1 形成性研究设计

我们招募了五名（P1–P5）有 AI 辅助内容创作经验的参与者：其中两人熟悉基于节点的工作流（例如 ComfyUI），另外三人主要使用专有的端到端平台。我们请每位参与者围绕一个他们可能实际创作的主题，制作一段多场景短视频（3–5 个场景），例如文化遗产介绍、旅行情绪短片或氛围感微故事。我们仅提供了粗略的高层目标（主题和基调），要求参与者逐步明确意图、探索备选方案并完善中间输出。

参与者以开放的方式使用自己偏好的工具，让意图在探索过程中不断演变。每次会议时长约为 60 分钟，包括观察任务执行过程和进行半结构化访谈。在征得同意后，我们录制了屏幕活动并收集了笔记，随后通过迭代式定性编码和作者讨论，总结反复出现的模式与流程中断。

3.2 观察到的工作流：常见创作操作

无论使用何种工具或具备何种专业水平，参与者都逐渐形成了相似的创作模式。他们并非一步生成完整视频，而是通过反复执行具有可识别输入和输出（例如文本意图、参考素材、约束和生成结果）的操作来构建视频。

首先，创作者通过生成或选择参考图像来确立视觉锚点，以固定构图、风格和氛围。其次，他们在投入运动生成之前，对图像进行迭代细化，调整提示词并执行局部编辑、结构保持细化和质量增强等中间操作。第三，创作者基于选定的锚点生成运动，并通过重新生成、约束调整、镜头控制或帧插值，反复修正时序问题。第四，创作者通常会并行制作音频素材，之后再使其与场景时序对齐。最后，他们通过裁剪、拼接和同步来组装选定的视频片段和音频；这一过程往往需要导出到外部工具，增加了工作开销，也打断了上下文。

3.3 关键挑战

虽然整体操作序列保持一致，但参与者反复遇到以下问题：难以决定下一步该做什么、管理备选方案，以及协调跨模态工具。基于这些环节，我们确定了三个反复出现的挑战。

C1. 将意图转化为下一步可执行步骤并不容易。即使参与者能够清晰地描述详细提示词，输出结果也常常偏离预期的构图、风格或运动效果。现有工具对如何解读这些偏差并将其转化为具体的纠正操作，支持有限。随着约束不断增加，创作者缺乏指导，不知道应修改锚点、引入控制项，还是重构工作流，因而只能通过试错探索。

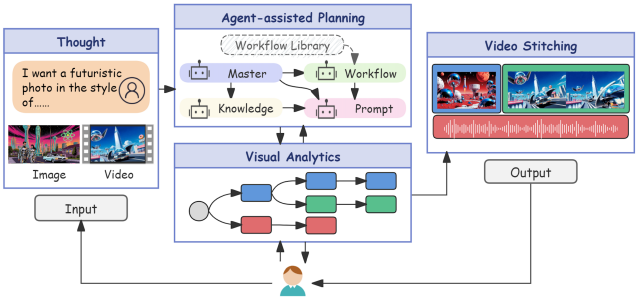

图 1：视频拼接过程中的以用户为中心的创作循环。自然语言意图被转化为可编辑计划，执行后生成可视分析结果，并经过整理形成视频输出，从而支持在不同状态之间进行分支与修订式迭代。

C2. 探索是非线性的，但轨迹难以管理。参与者经常生成变体、比较备选方案，并回溯到先前的结果。然而，现有系统要么隐藏中间步骤，要么展示无法充分体现探索逻辑的执行图，导致变体难以关联、比较和复用。

C3. 流程管线在工具和模态之间是割裂的。创作需要在图像优化、视频生成、音频创作和组装工具之间不断切换。反复进行素材转移增加了操作负担，也造成了上下文丢失。

3.4 设计需求

基于这些挑战，我们为以用户为中心的创作方式提出了四项需求，以支持将从构想到视频的创作作为一种迭代式、多阶段的探索。

R1. 为意图到操作的转换提供下一步依据（回应 C1）系统应帮助创作者在检查中间结果后，将演化意图转化为具体的下一步创作操作。这一转换应以当前上下文（场景意图、当前状态、可用素材）为依据，并生成可供创作者在执行前修改的可执行草案，从而减少选择纠正操作时临时尝试和反复试错的情况。

R2. 具有明确输入输出边界的动作级创作单元（回应 C1、C3）创作过程应围绕创作者可识别的操作（例如，设定锚点、优化图像、在约束条件下生成运动、生成音频、汇总结果）来组织，并明确输入输出边界。这种抽象使创作者能够专注于想要实现的目标，而非构建算子图，同时在底层工作流或模型发生变化时仍保持稳定。

R3. 支持非线性探索、比较和复用的结构化轨迹（回应 C2）系统应将非线性探索过程外化为显式轨迹，保留中间状态及其推导关系。创作者应能够回溯，在共享上下文下比较备选方案，并复用成功的局部决策和素材，而无需从头重建工作流。

R4. 跨阶段的跨模态连续性与最终收敛。（回应 C3）系统应在单一工作区中保持图像、视频、音频和组装阶段之间的连续性，方法是维持一致的状态上下文，并支持操作之间明确的素材交接。这可以减少工具切换和上下文丢失，并支持从探索性迭代收敛为连贯的多场景输出。

4 代理辅助下的创作工作流

为满足设计需求（R1–R4），我们将思维到视频创作组织为一个以用户为中心的循环（图 1），该循环结合了（i）创作者意图输入，（ii）智能体辅助的下一步规划，（iii）

基于持久化状态的视觉检查与修订，以及（iv）通过拼接实现的跨模态收敛。代理辅助主要通过将创作者的意图转化为可编辑操作草案，来满足下一步依据（R1）；界面则将创作者可识别的操作（R2）付诸实践，并保留显式轨迹，以支持分支、比较和复用（R3）。通过在同一工作区中整合图像、视频、音频和素材组装，并支持操作之间显式的素材交接，该循环也旨在实现跨模态连续性和最终收敛（R4）。以下小节将介绍我们如何定义创作操作、将规划呈现为可编辑的决策，以及持久化创作状态以实现可追溯性和复用。

4.1 创作者可识别的创作操作

为了支持具有清晰输入输出边界的动作级创作单元（R2），我们将形成性研究中发现的重复出现的创作意图，具体化为一组创作者可识别的创作操作。这些操作包括确立视觉锚点、优化视觉资产、在约束下生成运动、制作音频，以及通过裁剪、拼接和同步来组合选定的素材。每项操作都采用一致的界面结构呈现，并构成一个稳定的进度单元；即使底层使用不同的模型或工作流实现，这些单元对创作者而言仍具有实际意义。

每项操作都由输入和输出定义，并与创作者在实际工作中理解进度的方式相契合。例如，创作者可以提供文本，并可选择附上参考图像，以获取锚点图像；优化锚点图像以提升视觉质量；结合锚点图像（或一对图像）与运动约束来生成视频片段；或者根据脚本制作音轨。通过在操作层面保持界面一致，这种设计有助于迭代决策：创作者可以选择强化锚点、修改约束或切换操作类型，而无需被迫理解底层的执行结构。

4.2 代理辅助的规划作为创作决策

实践中的一个核心难点，是检查中间结果后决定下一步可执行步骤。为解决意图到操作的转换中的下一步依据问题（R1），我们引入代理辅助，在执行前生成一份可编辑计划。

如图 1 所示，给定创作者输入的步骤信息（包括简短的意图描述和可选的视觉参考），系统会通过检索场景根节点捕获的场景级意图，以及活动分支上的当前路径上下文来构建上下文。随后，四个智能体协作，将意图转换为具体的下一步操作。主代理整合上下文，并提出合适的操作类别。知识代理会酌情提供简要说明，以澄清可能影响生成选择的实体或概念。根据操作类别和可用输入，工作流代理会选择兼容且可执行的工作流模块（由工作流 ID 标识）。最后，提示代理生成模态感知的草案规范，其中包括提示文本以及与所选工作流相匹配的关键可编辑参数。

重要的是，该计划并非以不透明自动化的方式执行。它会作为绑定到节点、可供检查的结果呈现给创作者，并可就地修订，包括覆盖建议的操作、编辑提示词以及调整参数。执行仅在审阅之后才会触发，从而在保留用户控制的同时，使意图到操作的转换成为可见且可修订的步骤。

4.3 面向可追溯性与复用的持久化创作状态

迭代创作本质上是非线性的：创作者会通过分支探索备选方案、回溯到较早的稳定结果，并在不同步骤和场景之间复用中间素材。为支持比较和复用所需的结构感知轨迹（R3），我们将每个步骤视为持久化的创作状态，而非短暂的执行事件。

创作状态记录创作者的意图文本、引用的输入、所选操作和工作流标识符、智能体生成的计划、创作者对提示词和参数所做的编辑，以及由此产生的多模态输出（完成图 1 中的一个迭代循环）。每个新步骤都会作为其扩展步骤的子步骤存储，从而形成明确的状态分支结构。这种表示方式通过保留决策上下文来支持可追溯性，并允许创作者沿用有效的中间素材和配置，而无需从头重建工作流，从而支持复用。与上下文内预览和基于拼接的收敛相结合，持久化状态也有助于维持各阶段之间的跨模态连续性（R4）。

5 T2VTree的视觉设计

在下一步依据（R1）、操作级创作（R2）、结构感知轨迹（R3）和跨模态连续性（R4）等需求的指导下，*T2VTree* 的视觉设计将想法到视频创作视为创作者在迭代过程中可以看见、检查以及修订 的对象。界面并非只展示最终输出，而是将每一步明确呈现为持久状态：创作者请求了什么、智能体提出了什么、执行了什么，以及产出了什么。这使得渐进式规范、分支感知探索以及在不丢失上下文的情况下进行复用成为可能。

5.1 设计原理与备选方案

我们的表示设计经过了三次迭代，每次迭代都由创作过程中观察到的具体问题所推动。我们以一项结构引导的图像生成任务为例（图 2）：创作者从输入图像中提取结构控制信号（例如 Canny边缘图），然后以输入图像和提取出的控制信号为条件，生成受约束的结果。

备选方案 A：简化 ComfyUI 风格的工作流图。我们最初采用了受 ComfyUI 启发的表示方式，因为基于节点的工作流对经验丰富的创作者而言十分熟悉，并且能提供细粒度控制。为减少杂乱，我们将底层操作符归并为更高层级的模块（例如图 2 备选方案 A 中的 提取结构 和 受约束生成 ）。虽然这让单个工作流更易于阅读，但创作进度仍依赖于图编辑和重新运行。创作者通过重新连接模块或调整参数来生成变体，而产生的轨迹则分散在多个图状态和导出文件中。因此，比较备选方案和复用已有进展都需要手动记录，因为图并未将每个步骤视为创作推理中一个持久的单元。

方案 B：将步骤拆分为输入输出对。 随后，我们尝试了一种更以结果为导向的视图，将每个步骤折叠为一个输入输出对（图 2 方案 B）。这种抽象突出了哪些输入导向了哪些结果，但在控制密集型创作中便不再适用。在该示例中，控制信号源自一张输入图像，并与另一张图像和文本共同对生成过程施加条件。一旦引入此类多输入依赖和跨步骤复用，结构便会成为一个链接不断累积的有向无环图，从而降低可读性，并使分支探索难以跟踪。

方案 C（最终方案）：以树组织的动作边界节点。为确保探索过程清晰易读，同时支持复用，我们将全局组织结构与内部依赖关系解耦。我们保留树作为全局骨架，并将每个节点重新定义为一个自包含的创作操作（图 2，方案 C）。每个节点都将可编辑规范（意图/提示词、参数、引用的输入和选定的工作流模块）与其输出放在一起，因此节点本身成为检查、修订和比较的最小单元。在此设计中，子节点表示检查父节点结果后做出的后续决策，兄弟节点则表示相同上下文下的备选方案。节点内部通过引用先前资产来体现复用，从而避免使用显式依赖边，以免视图退化为图。

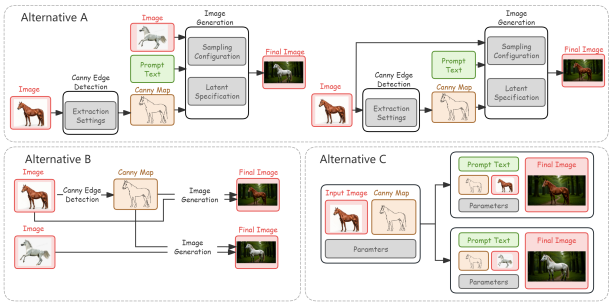

图 2：结构引导图像生成的三种创作备选方案，阐释了 *T2VTree* 中基于树的表示为何必要。

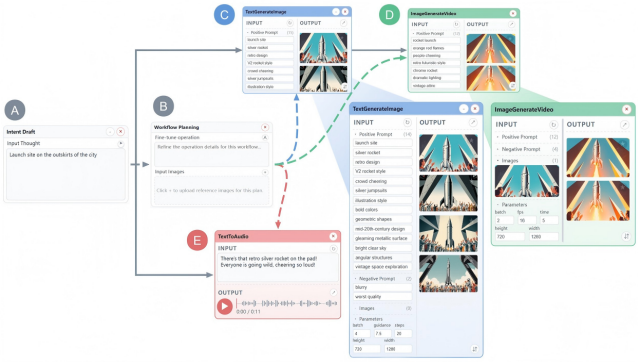

图 3：*T2VTree* 的视觉设计。以意图为先的创作步骤会具体呈现为一个 工作流规划 节点，在智能体规划后可转换为可执行的模态颜色编码节点。实线箭头表示生成后用户可见的创作轨迹，虚线箭头则表示系统侧从规划到特定模态状态的转换。

这一演进既保留了创作者熟悉的决策边界，又逐步抽象掉操作层面的复杂性，最终形成了 *T2VTree* 所采用的树形表示。

5.2 节点设计

我们将每个节点设计为一种持久的创作状态，使每个步骤都能就地理解。如图 3 所示，每个状态将以下内容并置：（i）用户的意图输入；（ii）用于执行该步骤的可编辑规范（例如提示词草稿、参数和可选参考素材）；以及（iii）生成的多模态产物。这种并置让修订和比较都能在局部完成（R4）：创作者可以调整步骤中的关键内容，而无需从外部来源重新构建上下文。

为了在树不断扩展时仍保持清晰易读，我们采用轻量级编码。非媒体状态（例如意图捕获和规划）以中性色调呈现，而媒体生成状态则根据输出模态进行颜色编码（初始设定为图像用蓝色、视频用绿色、音频用红色）。这有助于快速识别意图是在何处指定的、规划决策是在何处做出的，以及生成了哪些素材。

意图草稿。 意图草稿 节点将场景级意图记录为可编辑的语义锚点。创作者可以在此输入叙事角色、风格和情绪。意图确定后，保存按钮会锁定该语义锚点，使其成为稳定的参照点，供创作者在创建分支、回溯或修改先前决策时使用。

将工作流规划作为意图到操作的桥梁。 一个独特的设计是工作流规划 节点（图 3B），它将一个步骤定义为创作者接下来打算实现的目标。创作者可以提交简短请求，也可以通过上传 按钮附加视觉参考（例如图像锚点）。随后，发送 按钮会将上下文发送给协作智能体，由它们生成具体的规划方案，其中包括建议的创作操作、提示词草稿和关键可编辑控件。重要的是，这些规划结果会显示在节点中，以便创作者在提交执行之前进行检查和修改。

从规划到模态的转换。 规划状态不会停留为临时占位节点；选定操作后，它会转化为 特定模态的节点（图 3 中的虚线箭头）。根据规划的操作，规划节点可以变为图像 节点（蓝色）、视频 节点（绿色）或音频 节点（红色）。这种设计使树轨迹由可操作、可比较的状态构成，同时清晰呈现从意图到操作的转换。

模态节点及原位修订。 模态节点采用一致的布局，将输入（可编辑规范）与输出（生成候选项）分开。对于图像生成（图 3C），创作者可以修订提示词草稿、调整参数，并可选择提供参考图像，以细化锚点或进行受约束的细化。对于视频生成（图 3D），节点会记录所选图像锚点和时间控制，并与生成的视频片段一同呈现，使运动相关决策在不同备选方案之间始终可供检查和比较。对于音频生成（图 3E），节点将旁白／脚本提示词与生成的音频相结合，以便在多场景创作过程中修订和复用音频决策。在不同模态中，将规范和结果保留在同一节点内，既支持迭代调优，也保留了决策上下文，便于后续比较和复用。节点提供轻量级的迭代控制项，包括生成、用于详细检查的垂直调整大小，以及用于最终构图的添加到拼接。

5.3 树结构设计

树为迭代创作提供了紧凑的全局组织结构。该结构并非依赖明确的边含义，而是通过层级结构来理解：深度反映创作方向上的渐进式细化，而兄弟节点则代表共享上下文下的并行尝试和备选方案。所有分支均始于一个初始化 根节点，它代表项目初始状态。

父子关系：渐进式细化。 父子关系表示当前创作方向逐步深入。创作者查看中间结果时，会引入额外约束、更改操作类型，或将现有输出延伸至下一阶段。这会产生一个更深层的节点，该节点仍以先前的上下文为基础，使工作流能够被理解为一个渐进式规范化过程。实际上，深度与常见的创作轨迹相吻合：从粗略意图出发，建立视觉锚点，对其进行细化，然后将其延伸为基于时间或构图的结果。

兄弟关系：并行路径与备选方案。 兄弟节点表示从流程中的同一点探索的并行方向。它们将备选方案外化为一等状态，使创作者能够比较相互竞争的变体，同时保留先前结果（R3）。兄弟节点不受限于某一种模态或操作类型：并行尝试可能在提示词、参考素材、参数乃至所选创作操作上有所不同，但仍可进行比较，因为每个节点都将已指定的内容与生成的结果并置呈现。

面向多场景创作的项目根节点分支。 多场景创作需要同时处理多个场景方向。因此，我们以单个初始化 根节点为中心组织工作区，创作者可以从该根节点分出多个相互独立的顶层场景方向路径。每个场景都以意图草稿节点作为场景锚点，后续的细化、生成步骤和素材衍生则沿着该场景的分支逐步展开。这种组织方式避免将无关的迭代交错在同一条线性轨迹中，同时让所有场景方向都在一个工作区中清晰可见，便于跨场景导航。

通过折叠和剪枝管理树。 随着探索的推进，局部邻域可能迅速扩展。为了让

T2VTree 保持清晰易读，可通过折叠 按钮折叠分支，以概括已完成的方向；需要时仍可访问内部状态。可通过删除 按钮移除不需要的尝试（及其后代），让创作者能够剪除价值较低的分支，而不影响其他方向。结合能够回收屏幕空间的自动布局调整，这些操作使工作区保持易于浏览，同时忠实记录细化与探索的过程。

6 T2VTREE 可视分析系统

T2VTree 可视分析系统将第 4 节和第 5 节中的创作抽象落实为一个用于思维到视频创作的交互式工作区。它将创作意图和持久化创作状态映射到协同视图中，以支持检查、修订、分支探索和拼接输出。

6.1 控制面板

控制面板（图 4A）提供节点和场景共享的全局设置，涵盖项目范围、智能体上下文默认值以及树形布局/编码。

项目管理。 顶部（a1）支持创建、删除和切换项目；每个项目都维护一棵独立的树。

全局上下文默认值。 中部（a2）用于配置基础智能体模型和会话级上下文（风格、情绪、调色板），还可选择添加背景文本或参考材料。这些设置会自动注入新建节点的智能体辅助规划中，同时仍可由节点级意图、提示词或参数在本地覆盖。

布局与模态编码。底部（a3）可调整全局节点间距（水平/垂直），并提供模态颜色映射（图像/视频/音频）供自定义，确保 T2VTree 视图中的编码保持一致。

6.2 T2VTree视图

T2VTree视图（图4B）是用于检查、分支和修订创作状态的主要工作区。它在无限画布上呈现树，以支持大规模探索。

节点卡片作为持久化创作状态。每张卡片对应一个持久化创作状态，将步骤级意图、可编辑输入（提示词、参数、参考素材）以及生成的多模态输出集中呈现。这样无需从外部工具重建状态，即可在上下文中检查和修订。

直接操作与分支。选择节点后，该节点将设为源状态，并在其下生成一个 添加工作流 占位符，用于指定下一步意图。经过智能体辅助规划后，占位符会被替换为具体的媒体节点，以呈现所选的创作操作。这种交互明确体现了分支语义：子节点表示细化/扩展，兄弟节点则表示在相同父上下文下的备选决策。节点可以折叠或展开，以管理视觉复杂度；这样既能保留全局概览，也能在需要时查看详细历史记录。

集成式多模态预览。 输出会在节点卡片中预览：图像和视频以缩略图形式显示，视频循环播放以便快速浏览，双击可打开更高分辨率的预览。音频节点会显示波形并支持原地播放，让用户无需离开树即可快速评估。

6.3 视频拼接视图

视频拼接视图（图 4C）通过多轨时间线编辑器，支持从发散性探索过渡到汇聚式组装。

候选项收集。 创作者可以从任意分支中收集图像、视频片段和音轨，并放入临时收集区 (c1)。候选项保留模态颜色编码，以保持与树的一致性。

拼接拖放。 将已收集素材拖到时间线轨道 (c2, c3) 上并重新排序，即可形成线性叙事。时间线提供基本的对齐与排序控件，用于同步视觉内容和音频。

端到端可追溯性。 已组装片段仍与其源自的创作状态保持关联，使创作者能够跳转回对应节点，修改提示词/参数，重新生成素材，并在保留来源信息的同时更新拼接结果。

7 评估

我们通过两个案例研究和一项用户研究，展示*T2VTreeVA* 如何支持在真实的多阶段场景中进行端到端的思维到视频创作。

7.1 案例1：从构思到结构化视频创作

我们制作了一段关于唐三彩釉陶骆驼的短视频，以展示*T2VTreeVA* 如何支持多场景、多模态创作，将模糊构思逐步转化为拼接完成的最终视频。图4展示了案例1的创作轨迹：这是一棵从初始化根节点生长出的树，其中每个节点都封装了相应的步骤级规范和生成输出。

场景级设置与意图草拟。我们首先创建了一个新项目并配置全局生成偏好（图4A），包括模型和粗略的风格/情绪预设（a2），以及模态颜色映射（a3）。我们还在全局参考素材区域中输入背景信息，包括朝代及其象征意义，从而确立了以唐代中西文明交流为主题基调的创作方向。随后，我们从初始化节点扩展出三个分支，创建了三个场景（图4B，“场景1–3”）。对于每个场景，我们首先创建一张意图卡片，描述期望的内容和约束（例如历史基调和场景主题），作为生成前可编辑的入口。

智能体辅助工作流规划与提示词实例化。对于描绘唐代长安骆驼的主要场景，我们在规划节点上调用了代理辅助。系统返回了 (i) 一份可编辑的提示词草稿，以及 (ii) 一个与意图相匹配的已选工作流模块（例如，结构引导图像生成）。这些规划输出会存储在节点状态中，并以工作流样式卡片的形式呈现，使我们能够在执行前修改提示词/参数。此步骤免去了手动组装底层算子图的需要，同时在轨迹中明确记录了所选工作流和提示词。

以可追溯分支进行迭代式图像优化。从初始锚点图像出发，我们首先查看了生成的四种风格变体，并选定淡水彩素描风格作为既定的视觉基准。如图4B所示，其他优化尝试会作为同级图像节点记录在同一个父节点下。父节点保留优化过程中未发生变化的内容，而每个同级节点则记录一种不同的修订结果。

在此场景中，我们创建了一个分支，直接应用放大操作（“Refine & Upscale”），以快速评估选定风格的视觉质量。与此同时，另一个分支执行了两项遗产风格创作中常用的局部精修：通过添加背景细节来丰富场景背景（“Enrich street contex”），以及移除冗余元素（“Remove the base”）。随后，我们将放大操作（“Refine & Upscale”）作为派生子节点应用，以提升选定锚点图像的质量，然后再进行运动生成。

图像到视频交接和视频端校正。选定最终锚点后，我们生成了一个视频片段，并将其表示为由蓝色图像节点派生的绿色节点（图4B，“生成视频”）。当第一个片段呈现出时间结果时，我们创建了一个

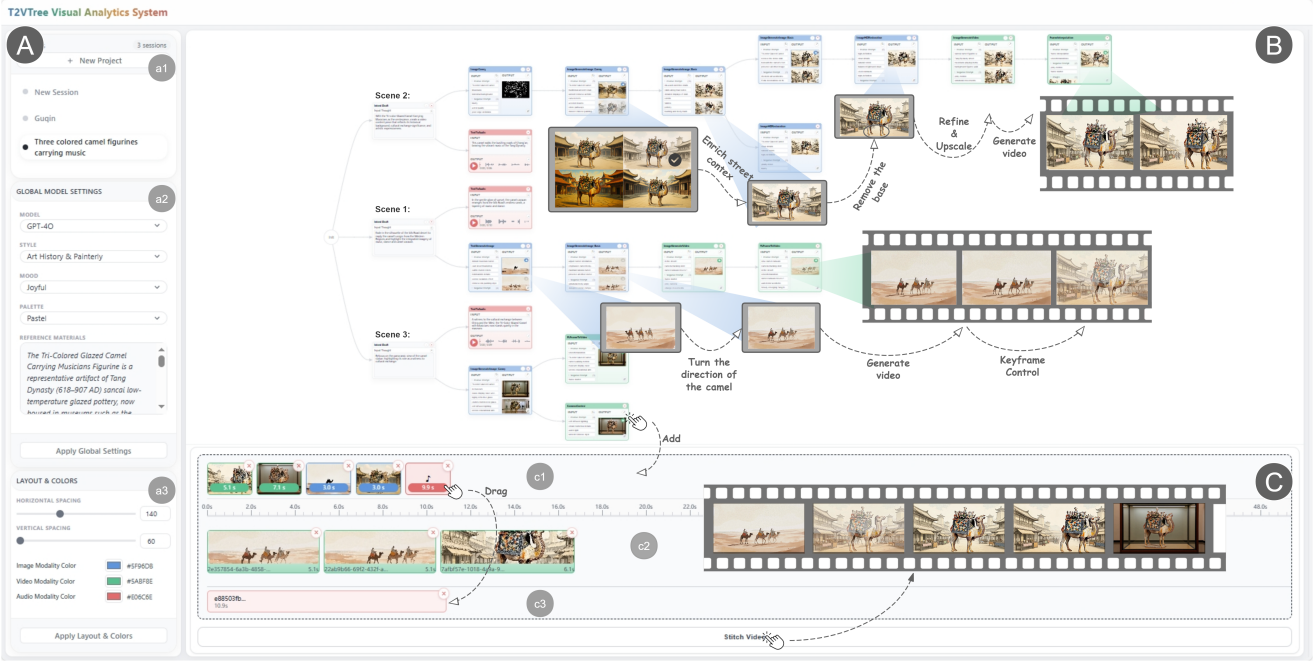

图4：*T2VTreeVA* 界面展示了案例1的创作过程。(A) 控制面板，用于输入全局历史背景和风格设置。(B) T2VTree视图，用于呈现分支溯源轨迹，其中节点代表生成的图像/视频/音频，兄弟分支则展示了不同的创作流程和细化结果。(C) 视频拼接视图，将从不同发散分支中选取的片段组合成最终的纪录片时间线。

派生视频节点，通过帧插值进行处理，同时将原始片段保留在父状态中，以便回退。

多场景构建与转场。 为了拓展叙事，我们为第二个场景创建了一条平行分支：一头骆驼驮着音乐家穿行于沙漠中，同时保持既定的水彩风格。发现运动方向与叙事相悖（由西向东）后，我们创建了一个校正兄弟节点来调整方向。决定将该场景作为开场后，我们采用首尾帧条件控制——使用骆驼视频片段的末帧和街景的首帧——生成一个转场视频节点，平滑衔接这两个不同的分支。

音频生成与端到端拼接。 与此同时，我们生成了一条音频旁白轨道，并将其作为红色节点关联到同一项目上下文。为制作最终的纪录片，我们将选定的视频片段（来自不同的场景分支）和音轨添加到拼接面板中（图4C）。拼接视图提供了一个多轨时间线缓冲区，创作者可将选定片段拖入轨道（c1–c3）并导出一个完整视频，同时每个片段都可以追溯到树中对应的源节点。

7.2 案例2：通过渐进式细化进行多场景创作

第二个案例着重展示如何对并行创作的多个场景进行渐进式细化：创作者反复修正、完善并复用各项决策，同时保持长程叙事连贯。我们制作了一段关于中和琴的文化遗产视频（中和琴是与朱常涝王子相关联的明朝乐器）。图5将这一过程呈现为从 Init 根节点延伸出的多个场景分支：在每个场景中，子节点记录逐步完善决策的过程，兄弟节点则记录在共同父节点下进行的并行尝试，以便局部比较。

无需重新开始的渐进式细化。 我们从参考材料和最初的全局偏好“写实且具有摄影质感”开始。对于开场场景，我们提供了一个粗略的创作意图（“营造宁静的氛围，包含树木、河流和月亮”），智能体辅助规划据此建议了若干提示语短语。确认这些建议后，系统在单个节点中生成了多张候选图像供我们选择。选中一张符合预期基调的图像后，我们发现其中有一个与历史背景冲突的时代错置物件（一根电线杆）。我们没有从头重新生成场景，而是提出了自然语言校正请求，要求移除该物件，同时保留构图。这次修订被记录为一个子节点，而并行尝试则作为同一父节点下的兄弟节点呈现，使校正过程保持局部化且可审查。之后，当编辑后的输出变得过于模糊时，我们将超分辨率放大作为进一步的子节点，同时保留放大前的结果作为其父状态。

以固定的结构锚点细化展览属性。 对于博物馆场景，我们引入了一张古琴的参考照片，并将目标描述为：在将古琴置于展览环境中的同时，忠实呈现乐器本身。与案例7.1一样，我们使用 ControlNet-Canny 引导来确保几何结构在多次迭代中保持稳定（“参考结构”），因此后续细化主要影响展览属性，而非乐器结构。

基于固定结构锚点，我们在同一层级创建了多个并行分支，以探索未充分指定的属性。与强约束分支（例如白墙、白色照明、白色展示台）不同，我们引入了开放式生成分支，尽量减少人工干预，以鼓励多样性并避免过早确定单一方向。选定偏好的方向（例如昏暗背景前的白色展示台）后，我们通过明确添加“玻璃展示柜”约束来收紧请求，并将其记录为子节点，以便在轨迹中清楚呈现这一决策变化。随后，我们通过相机变焦工作流生成了一段简短的特写片段，并将其记录为一个绿色

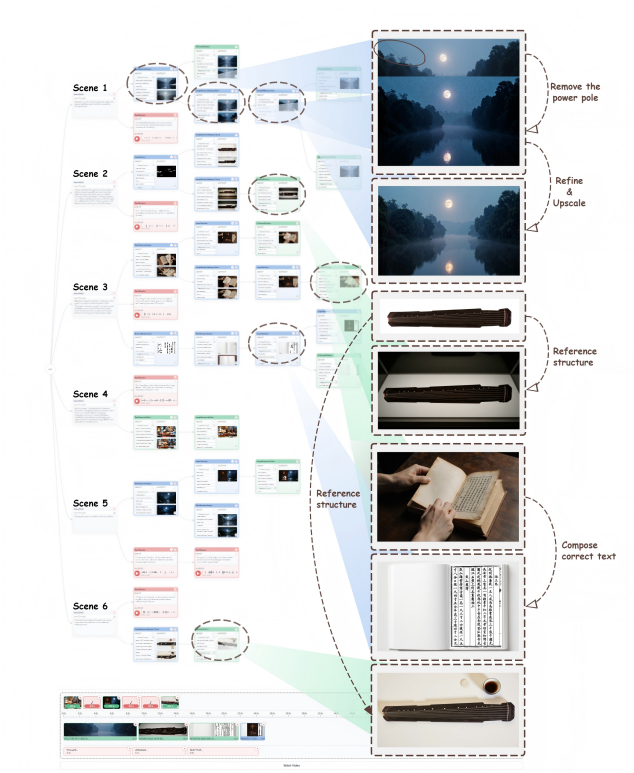

图 5：在 *T2VTree* 中可视化的案例 2 创作轨迹。左：以 Init 为根节点的完整多场景树。右：四个关键片段的放大视图（以虚线连接），展示了系统对渐进式细化的支持。

视频节点，该节点源自选定的蓝色图像节点，使图生视频交接易于核验和回溯。

具有可追溯输入的分层合成。 第三个场景需要呈现历史文献，但直接生成的文本难以确保清晰可读。因此，我们采用分层构图策略，将内容准备与视觉构框分开。我们先清理一张真实历史文本的截图（去除背景并清理水印），然后生成一张打开书本的承载图像，最后将清理后的文本和“中和”题字叠加到书本上（“合成正确文本”）。每个阶段都作为一个子节点记录下来，因此合成结果始终可以追溯到具体输入，并且可以通过替换单个步骤进行修改，而不必从头重做整个场景。为了使场景之间平滑衔接，我们生成了一个可控制起止帧的过渡片段，并以选定的上游结果为锚点，而不是覆盖先前的尝试。

作为局部决策的回溯与复用。 在呈现古琴底面镌刻的诗句时，我们经历了一次非线性探索。我们首先要求生成克制的背景（“月光、河水和露珠”），随后收紧要求，突出“露珠”。优化后的结果虽然更加丰富，视觉上却也变得杂乱。借助图5中可见的决策轨迹，我们回溯到较早的父状态，并尝试另一种方向：先将诗句文本叠加到较早、更加简洁的背景上，再将其转换为视频。不同方向保留为同一父节点下的兄弟节点，因此无需舍弃任何尝试即可进行比较。

我们可以复用部分进展。例如，我们将先前生成的结构控制选作新节点中的引用输入，以请求进行现代极简风格重新渲染，随后应用逆时针摄像机旋转工作流来完成叙事。由于复用是在节点内部通过引用输入来表达的而不是通过添加新边来实现，因此随着图5中的整体轨迹不断扩展，它仍然清晰易读。

最终视频输出。 最后，我们将来自多个场景分支的选定片段（以及适用时的配套音频）整理到拼接视图中进行构图。由于每个选定片段始终关联着其源节点，我们能够在组装过程中核实其来源，并在需要时修订上游决策，而无需重建中间步骤。

7.3 用户研究

我们开展了一项小规模比较研究，以考察与基于节点的工作流基线（ComfyUI）相比，*T2VTreeVA* 是否能够降低协调开销，并更好地支持思维到视频创作中的分支感知探索与复用。

参与者与研究设计。 我们招募了六名参与者（P6–P11），他们此前均有使用 ComfyUI 进行基于扩散模型的图像与视频生成的经验；此前无人使用过 *T2VTreeVA* 。参与者被平均分配至两个被试间实验条件：ComfyUI（对照组）和*T2VTreeVA* （实验组）。

任务与流程。 参与者围绕开放式主题“未来城市”创作一段五场景短视频。他们可以自由地迭代、创建分支并复用中间结果，同时保持多场景成果的连贯性。为捕捉创作早期阶段的行为，我们还记录了参与者完成前三个场景所用的时间。每次实验约持续 90 分钟，包括 20 分钟的引导式教程和一段简短的热身环节（不计入测量），随后进行任务执行和任务后访谈。我们将任务时间定义为从开始创作场景到成功完成最终导出的时长。在任务过程中，我们要求参与者描述其工作时的意图与决策；同时记录屏幕交互和音频评论。

基线与环境。 为避免基线设置不切实际，我们为 ComfyUI 参与者提供了一个经过筛选的工作流模板库，其中涵盖了与*T2VTreeVA*支持的主要创作操作相同的内容。参与者无需从头构建算子图，但需要选择模板、连接输入、调整参数，并随着约束变化调整工作流。两种条件使用相同的后端服务器进行媒体生成，该服务器配备 NVIDIA H100 GPU（80 GB）；操作所用的工作站配备 Intel Xeon Silver 4110 CPU（2.10GHz）、64 GB RAM 和 NVIDIA Quadro P5000 GPU（16 GB），以确保两种条件下的执行延迟具有可比性。两种条件下的最终组装方式有所不同：*T2VTreeVA* 支持系统内拼接和导出，而 ComfyUI 参与者则在 ComfyUI 之外按照惯常工作流组装并导出片段。

测量指标。 我们记录了完成三个场景的时间（T_3）以及最终导出所需的时间（T_5）。为区分模型延迟以外的时间差异，我们记录了系统等待时间总计T_{wait} （从发出请求到获得可预览结果的累计时间），并计算活跃时间$T_{\text{active}} = T_5 - T_{\text{wait}}$，将其作为交互与协调工作量的代理指标。为单独评估最终组装，我们记录了组装/导出时间T_{assemble} （从进入最终选择与组装阶段到导出的时间），并计算迭代创作时间$T_{\text{author}} = T_5 - T_{\text{assemble}}$。我们还记录了生成调用次数$N_{\text{calls}}$ （包括后处理）以及保留变体数量N_{variants} （明确保留以供比较或复用的候选项）。

参与者还需针对所使用的工具回答三个李克特五级量表问题（1=非常不同意， 5=非常同意）：

L1：下一步清晰度。“这个工具帮助我将初步想法转化为明确的下一步行动。”

L2：迭代控制。“我可以在保持过程可控的同时，逐步改进我的创作内容。”

L3：端到端收敛。“我可以高效地确定最终的多场景视频并将其导出。”

结果。 综合来看，各条件下的总等待时间T_{wait} 相近，而端到端完成时间的差异

表1：ComfyUI 与 *T2VTreeVA* 在创作效率和探索行为方面的定量比较。

指标（平均值）	ComfyUI	T2VTreeVA
完成时间		
到达3个场景的时间T_3 （分钟）	34.5	**31.3**
到最终导出的时间T_5 （分钟）	68.2	**52.6**
组装与导出时间T_{assemble} （分钟）	10.5	**1.8**
系统等待与用户投入		
总等待时间T_{wait} （分钟）	30.4	**27.5**
活跃时间T_{active} （分钟）	37.8	**25.1**
探索行为		
生成调用（次数）N_{calls}	32.3	**20.3**
保留变体（次数）N_{variants}	4.3	**8.3**

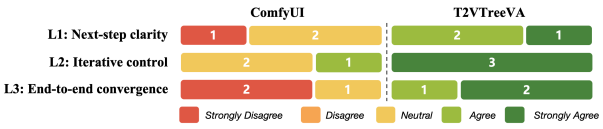

图 6：ComfyUI 与*T2VTreeVA*在 L1–L3 上的李克特五级量表比较结果。

（T_5）主要体现在T_{active} 和N_{calls}上，这表明使用*T2VTreeVA*可减少交互和重新运行带来的开销。在我们的研究中，ComfyUI 会话的N_{calls} 较高，这与参数调优和工作流适配过程中反复重新运行的情况一致；而*T2VTreeVA* 参与者更常保留中间状态，并保留明确的候选项以进行分支感知检查，这反映在较高的N_{variants}上。我们还发现，在 ComfyUI 条件下，T_{assemble} 对端到端耗时的影响不可忽视，这与需要在创作工作区之外完成最终选择和导出的情况一致。李克特量表回答进一步为定量趋势提供了背景信息（图6）。对于 L1，ComfyUI 的评分集中在中立或负面回答附近，而 *T2VTreeVA* 的回答则倾向于认同，表明从初始想法到可执行步骤的引导更加清晰。对于 L2，*T2VTreeVA* 条件下的回答始终偏向正面，而 ComfyUI 的回答则更加多样，这反映出管理迭代改进需要付出更多努力。对于 L3，ComfyUI 收到的评分以负面为主，而*T2VTreeVA* 的回答则集中在认同一侧，这与表1中观察到的组装时间T_{assemble} 缩短和主动投入T_{active} 降低相一致。综合来看，这些主观评估与测得的协调开销和探索行为差异相吻合，表明明确的决策轨迹和系统内组装支持能够提升创作者在思维到视频创作工作流中的感知掌控力，并帮助其收敛创作方向。

8 讨论与局限

8.1 可持续性考量

生成式创作系统的可持续性不仅取决于计算和经济成本，还取决于负责任的使用，以及在不断演进的模型生态系统中实现可复现的运行。

成本。 *T2VTree*仅在低频规划和提示词具体化阶段使用商业大语言模型 API，而不将其置于高频生成循环中。在当前实现中，典型的规划请求会消耗 2k–4k 个 token，与视频生成的成本相比，这一开销较小。所有图像、视频和音频生成均由开源模型执行。这降低了对持续 API 调用的依赖，并提升了可复现性和可扩展性。此外，这也支持系统的长期演进：随着生成器不断改进，*T2VTree*的基于状态的表示和工作流级规划仍然适用，因此系统可以接入更新的开源模型，而无需更改核心创作机制。

伦理。 为降低生成式创作中的伦理风险，*T2VTree*采用基于角色的智能体分配和精选提示模板，以约束下游生成请求。这些模板明确规定内容规则和特定情境限制，在生成前提供轻量级防护层。同时，规划输出始终可见且可编辑，创作者必须审核每个步骤，从而支持负责任的使用，而不依赖不透明自动化。

8.2 局限性与未来工作

细粒度控制有限且迭代开销较大。 在实践中，当生成控制项不足以表达期望的约束，或修正构图、运动或风格上的细微偏差时，创作者可能需要进行多次迭代。*T2VTreeVA* 通过保留中间状态并支持分支感知比较来应对这一问题，使创作者能够探索备选方案而不丢失上下文。然而，这种设计仍可能将未满足的控制需求转化为更多分支和节点增长，从而增加交互开销并降低创作流畅度。未来工作将探索更丰富的控制表示和交互支持，使各步骤中的可控因素更加明确，并减少不必要的迭代，例如结构化提示词编辑、约束导向控制，以及基于先前成功状态的轻量级引导。

对长篇复杂创作的可扩展性。 树表示适用于中小规模的探索，但长视频和复杂故事板可能会生成又深又宽的树，难以阅读和导航。尽管*T2VTreeVA* 支持折叠子树，未来仍需通过更高层次的摘要与组织方式提升可扩展性，例如将相似变体分组、突出重要路径，并提供多层级概览，同时在需要时保留对详细节点证据的访问。

9 结论

本文将思维到视频创作重新定义为一种以用户为中心的可视分析问题：创作者在其中迭代地将演化意图转换为具体操作，分支探索不同替代方案，并在多个模态之间逐步收敛。为支持这一过程，我们提出*T2VTree*，围绕操作级单元组织创作流程，为每个单元明确输入输出边界，并将每一步外化为持久的创作状态。状态将意图、参考资产、工作流选择、提示词和关键参数，与生成的多模态结果集中呈现，从而支持就地获取证据，同时保留树结构以实现可追溯性、比较和复用。

*T2VTreeVA*进一步将多个视图与智能体辅助规划相结合，根据当前场景意图和分支上下文生成可编辑计划。系统将建议的操作以及提示词/参数草案作为绑定于节点的工件呈现，供用户在执行前审阅和修改，从而将意图到操作的转换变为可控的创作决策，并进一步将其具体化为可随时间持续检查的特定模态状态。结合一体化的拼接工作区，创作者可以在不同场景之间迭代优化素材，并在不离开创作上下文的情况下组装出连贯的视频。

我们通过两项多场景文化遗产案例研究和一项对比用户研究对*T2VTreeVA* 进行了评估。在这些评估中，创作者能够逐步将想法完善为完整的多场景视频，利用树轨迹进行回溯、比较备选方案、复用中间结果并协调跨模态交接；总体而言，可视分析设计与智能体辅助规划降低了实际创作工作流中的协调开销。

致谢

本研究得到中国科学院—湖南省联合研究计划（重点技术研究

与数字文化遗产旗舰应用，编号：2024年JK4002）的资助。

参考素材

[1] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts 等。稳定视频扩散：将潜在视频扩散模型扩展至大型数据集。*arXiv* 预印本 *arXiv:2311.15127*，2023年。[2] S. Brade, B. Wang, M. Sousa, S. Oore 和 T. Grossman。Promptify：借助大语言模型，通过交互式提示词探索进行文生图。载于第*36*届 *ACM* 用户界面软件与技术年度研讨会论文集，第1–14页，2023年。[3] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman 等。视频生成模型作为世界模拟器。*OpenAI* 博客，1(8):1，2024年。[4] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva 和 H. T. Vo。VisTrails：可视化与数据管理的结合。载于*2006*年 *ACM SIGMOD* 数据管理国际会议论文集，第745–747页，2006年。[5] N. Cao, X. Qi, C. Chen 和 X. Yan。CODS：一种基于设计空间的计算设计理论模型。*arXiv* 预印本 *arXiv:2506.18455*，2025年。[6] M. Chion。视听：银幕上的声音。哥伦比亚大学出版社，2019年。[7] comfyanonymous。ComfyUI：一款功能强大、模块化的稳定扩散图形界面，采用图/节点界面。GitHub 仓库，2023年。[8] G. Croisdale, E. Huang, J. J. Y. Chung, A. Guo, X. Wang, A. Z. Henley 和 C. Omar。Deckflow：在多模态生成画布上进行迭代式规范定义。*arXiv* 预印本 *arXiv:2506.15873*，2025年。[9] N. Cross。设计思维：理解设计师如何思考与工作。布鲁姆斯伯里出版社，2023年。[10] W. Cui, D. Yu, X. Jiao, Z. Meng, G. Zhang, Q. Wang, S. Y. Guo 和 I. King。语音语言模型的最新进展：综述。载于第*63*届计算语言学协会年会论文集（第*1*卷13970，2025年 13943：长篇论文），第 – 页。[11] M. E. De Carlo 和 P. Li mone。创造力：培养发散思维。人工创造力：展望数字文化的未来，第97页，2025年。[12] E. Dritsas, M. Trigka, C. Troussas 和 P. Mylonas。多模态交互、界面与通信：综述。多模态技术与交互，9(1):6，2025年。[13] R. Du, N. Li, J. Jin, M. Carney, S. Miles, M. Kleiner, X. Yuan, Y. Zhang, A. Kulkarni, X. Liu 等。Rapsai：通过可视化编程加速多媒体应用的机器学习原型开发。载于*2023*年 *ACM*人机交互大会论文集，第1–23页，2023年。[14] K. Eberhard。可视化对判断与决策的影响：系统性文献综述。管理评论季刊，73(1):167–214，2023年。[15] R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Rambhatla, A. Shah, X. Yin, D. Parikh 和 I. Misra。Emu视频：通过显式图像条件控制分解文本生成视频。*arXiv* 预印本 *arXiv:2311.10709*，2023年。[16] Y. Guo, H. Shao, C. Liu, K. Xu 和 X. Yuan。PromptHIS：可视化文生图创作过程中的提示词编辑及其影响。*IEEE*《可视化与计算机图形学汇刊》，2024年。[17] J. He, X. Wang, K. K. Wong, X. Huang, C. Chen, Z. Chen, F. Wang, M. Zhu 和 H. Qu。VideoPro：一种用于交互式视频编程的可视分析方法。*IEEE*《可视化与计算机图形学汇刊》，30(1):87–97，2023年。[18] J. Heer。能动性加自动化：将人工智能融入交互系统的设计。美国国家科学院院刊，116(6):1844–1850，2019年。[19] R. Henschel, L. Khachatryan, H. Poghosyan, D. Hayrapetyan, V. Tadevosyan, Z. Wang, S. Navasardyan 和 H. Shi。StreamingT2V：从文本生成连贯、动态且可扩展的长视频。载于计算机视觉与模式识别会议论文集

，第 2568 页–2577，2025年。

[20] M. T. Jan、M. G. Al-Jassani、M. Nadar、E. M. Vunnava、V. Chakrapani、H. Ullah、A. Khan、S. A. Abbas 和 B. Furht。《文本生成视频生成器：全面综述》。《大数据杂志》，12(1):253，2025年。[21] J. Kolko。《设计思维的分歧性》。《*Interactions*》，25(3):28–34，2018年。[22] F.-L. Liu、H. Fu、X. Wang、W. Ye、P. Wan、D. Zhang 和 L. Gao。《Sketchvideo：基于草图的视频生成与编辑》。载于计算机视觉与模式识别会议论文集，第23379–23390页，2025年。[23] Y. Ma、K. Feng、Z. Hu、X. Wang、Y. Wang、M. Zheng、X. He、C. Zhu、H. Liu、Y. He 等。《可控视频生成：综述》。*arXiv*预印本 *arXiv:2507.16869*，2025年。[24] A. Mishra、B. Danzy、U. Soni、A. Arunkumar、J. Huang、B. C. Kwon 和 C. Bryan。《PromptAid：面向大语言模型的视觉提示词探索、扰动、测试与迭代》。*IEEE*《可视化与计算机图形学汇刊》，2025年。[25] C. Mou、X. Wang、L. Xie、Y. Wu、J. Zhang、Z. Qi 和 Y. Shan。《T2i-adapter：学习适配器以挖掘文本生成图像扩散模型更强的可控能力》。载于*AAAI*人工智能会议论文集，第38卷，第4296–4304页，2024年。[26] D.-H. Nguyen、H.-P. Huynh、M.-T. Tran 和 T.-N. Le。《Genflow：用于图像生成的交互式模块化系统》。*arXiv*预印本 *arXiv:2506.21369*，2025年。[27] Z. Peng、J. Yu、W. Wang、Y. Chang、Y. Sun、L. Dong、Y. Zhu、W. Xu、H. Bao、Z. Wang 等。《Vibevoice技术报告》。*arXiv*预印本 *arXiv:2508.19205*，2025年。[28] E. D. Ragan、A. Endert、J. Sanyal 和 J. Chen。《可视化与数据分析中的来源特征：来源类型与用途的组织框架》。*IEEE*《可视化与计算机图形学汇刊》，22(1):31–40，2015年。[29] J. Satkunarajan、S. Koch 和 K. Kurzhals。《The Joy of Co-Painting：面向可追踪图像生成工作流的创意人机协作》。载于*2025*年*IEEE*第*18*届太平洋可视化会议（*PacificVis*），第318–328页。IEEE，2025年。[30] Y. Shi、C. Xue、J. H. Liew、J. Pan、H. Yan、W. Zhang、V. Y. Tan 和 S. Bai。《Dragdiffusion：利用扩散模型实现交互式基于点的图像编辑》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第8839–8849页，2024年。[31] A. Simkute、L. Ta nkelevitch、V. Kewenig、A. E. Scott、A. Sellen 和 S. Rintel。《生成式人工智能的悖论：理解并缓解人机交互中的生产力损失》。《国际人机交互期刊》，41(5):2898–2919，2025年。[32] S. R. Solanki 和 D. K. Khublani。《从剧本到银幕：揭示文本生成视频》。载于《生成式人工智能：探索生成式人工智能的力量与潜力》，第81–112页。Springer，2024年。[33] S. Suh、B. Min、S. Palani 和 H. Xia。《Sensecape：借助大语言模型实现多层次探索与意义建构》。载于第*36*届*ACM*用户界面软件与技术研讨会论文集，第1–18页，2023年。[34] K. Team、J. Chen、Y. Ci、X. Du、Z. Feng、K. Gai、S. Guo、F. Han、J. He、K. He 等。《Kling-omni技术报告》。*arXiv*预印本 *arXiv:2512.16776*，2025年。[35] K. Trinh、S. Seidenberger、R. Wijewickrama、M. Jadliwala 和 A. Maiti。《一幅图胜过千条提示词？以人为主导的迭代式提示词优化在图像再生成任务中的有效性》。*arXiv*预印本 *arXiv:2504.20340*，2025年。[36] K. Venkatesh、C. Dunlop 和 P. Yanardag。《Crea：用于创意图像编辑与生成的协作式多智能体框架》。*arXiv*预印本 *arXiv:2504.05306*，2025年。[37] Y. Wang、S. Shen 和 B. Y. Lim。《Reprompt：自动编辑提示词，使人工智能生成艺术更精准地表达意图》。载于*2023*年*ACM*人机交互大会论文集，第1–29页，2023年。[38] W. Weng、R. Feng、Y. Wang、Q. Dai、C. Wang、D. Yin、Z. Zhao、K. Qiu、J. Bao、Y. Yuan 等。《Art-v：结合扩散模型的自回归文本生成视频》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第7395–7405页，2024年。[39] W. Xia、G. Sun、T. Li、B. Chang、J. Tang、G. Zhang 和 R. Liang。《视频可视化与可视分析：基于任务与应用的

驱动型研究》。*IEEE*《电路与系统视频技术汇刊》，34(11):11316–11339，2024年。[40] D. Xiang、W. Xu、K. Chu、T. Ding、Z. Shen、Y. Zeng、J. Su 和 W. Zhang。《Promptsculptor：基于多智能体的文生图提示词优化》。载于*2025*年自然语言处理经验方法会议：系统演示论文集，第774–786页，2025年。[41] N. Xiang、T. Liang、H. Huang、S. Jiang、H. Huang、Y. Huang、L. Chen、C. Wang 和 C. Li。《Sel3DCraft：用于用户友好型文本生成三维的交互式视觉提示》。*arXiv*预印本 *arXiv:2508.00428*，2025年。[42] J. Xiao、F. Cheng、L. Qi、L. Gui、Y. Zhao、S. Lin、J. Cen、Z. Ma、A. Yuille 和 L. Jiang。《Videoauteur：迈向长篇叙事视频生成》。载于*IEEE/CVF*国际计算机视觉会议论文集，第19163–19173页，2025年。[43] V. Yoghourdjian、Y. Yang、T. Dwyer、L. Lawrence、M. Wybrow 和 K. Marriott。《从认知负荷角度看网络可视化的可扩展性》。*IEEE*《可视化与计算机图形学汇刊》，27(2):1677–1687，2020年。[44] D. Yu、O. Ian、L. Jie、Y. Xiaoru 和 N. Q. Vinh。《时空空间中过程内比较的用户中心可视化探索工具》。《可视化期刊》，26(2):403–421，2023年。[45] C. Zhang、Y. Dong、Y. Wang、Y. Han、G. Shan 和 B. Tang。《Auragenome：一个大语言模型驱动框架，可即时生成可复用且可扩展的环形基因组可视化》。*IEEE*《计算机图形学与应用》，(01):1–14，2025年。[46] L. Zhang、A. Rao 和 M. Agrawala。《为文本生成图像扩散模型添加条件控制》。载于*IEEE/CVF*国际计算机视觉会议论文集，第3836–3847页，2023年。[47] Y. Zhang、J. Li 和 Y.-W. Tai。《Layercraft：通过思维链推理与分层对象集成增强文生图》。*arXiv*预印本 *arXiv:2504.00010*，2025年。