arXiv:2602.03838v1 [cs.HC] 3 Feb 2026

PrevizWhiz：结合粗略三维场景与二维视频以引导生成式视频预可视化

Erzhen Hu
eh2qs@virginia.edu
Autodesk Research 与弗吉尼亚大学 多伦多，安大略省，加拿大

Frederik Brudy
frederik.brudy@autodesk.com
Autodesk Research 多伦多，安大略省，加拿大

David Ledo
david.ledo@autodesk.com
Autodesk Research 多伦多，安大略省，加拿大

George Fitzmaurice
george.fitzmaurice@autodesk.com
Autodesk Research 多伦多，安大略省，加拿大

Fraser Anderson
fraser.anderson@autodesk.com
Autodesk Research 多伦多，安大略省，加拿大

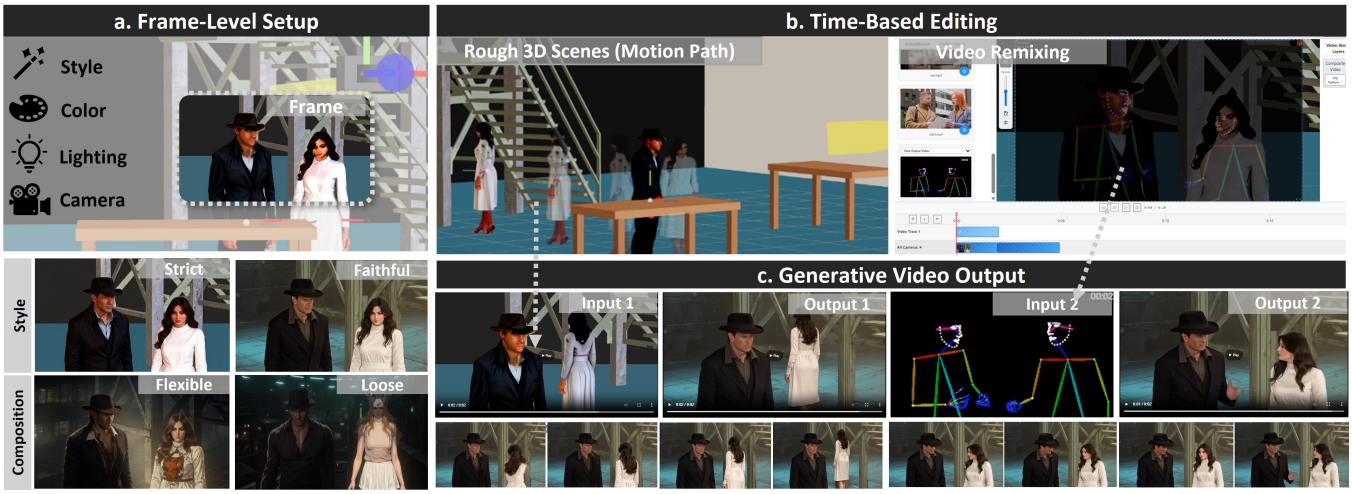

图1：PrevizWhiz支持一种创作工作流，将粗略三维场景和视频重混转化为生成式视频。（a1）用户从基础设置开始，在3D环境中捕获帧级图像，并可在其中定义色彩、灯光和摄像机设置。（a2）添加重新风格化图像作为输入，并可根据风格、色彩、灯光和空间构图与原始三维设置的接近程度，将相似度调整为严格、忠实、灵活或宽松。（b）随后，用户可以通过在时间线上为元素设置动画，或导入外部视频来定义动作，从而执行基于时间的编辑；编辑结果将作为视频生成模型的输入导出；（c）最后，我们展示使用（c1）3D场景或（c2）视频库作为生成式视频模型指导的视频输出结果。

摘要

在前期制作中，电影制作人和三维动画专家必须迅速将创意制作成原型，以便在全面制作前探索电影的可能性；然而，传统方法需要在效率和表现力之间进行权衡。手绘故事板通常缺乏复杂电影摄影所需的空间精度，而三维预可视化则要求具备专业知识以及高质量的绑定资产。为弥合这一差距，我们提出了PrevizWhiz，这是一套将粗略三维场景与生成式图像相结合、利用以及视频模型，创建风格化视频预览。该工作流程将可调节相似度的逐帧图像重风格化、通过运动路径或外部视频输入进行基于时间的编辑，以及进一步优化为高保真视频片段整合在一起。一项针对电影制作人的研究表明，我们的系统降低了电影制作人的技术门槛，加快了创意迭代，并有效弥合了沟通鸿沟，同时也揭示了人工智能辅助电影制作在连续性、作者身份和伦理考量方面面临的挑战。

本作品采用 CreativeCommons Attribution4.0 InternationalLicense 许可。
CHI '26，西班牙巴塞罗那© 2026 版权所有归所有者/作者所有。ACM ISBN 979-8-4007-2278-3/2026/04 https://doi.org/10.1145/3772318.3790534

CCS 概念

• **以人为中心的计算 → 交互式系统与工具**；

关键词
预演，生成式人工智能，电影制作，人机协作，创意支持

ACM参考文献格式：
Erzhen Hu、Frederik Brudy、David Ledo、George Fitzmaurice 和 Fraser Anderson。2026。PrevizWhiz：结合粗略三维场景与二维视频，引导生成式视频预可视化。载于《*2026*年*CHI*计算机系统中的人为因素会议论文集（*CHI '26*）》，*2026*年*4*月*13*日至*17*日，西班牙巴塞罗那。ACM，美国纽约州纽约市，共21页。https: //doi.org/10.1145/3772318.3790534

1 引言

预演（预可视化）是电影制作中的核心实践，使导演和创意团队能够在制作前探索场景的视觉与叙事结构 [54]。通过创建早期可视化，电影制作人可以测试摄影机角度、动作调度、节奏和情绪节拍等构想，而无需承担搭建全尺寸布景、聘请演员或制作精细资产的费用。除了作为创意草图工具，预可视化也可作为一种协作媒介，帮助导演、电影摄影师、制作设计师及其他利益相关者围绕共同愿景达成共识 [2, 4, 18]。

尽管十分重要，现有方法仍迫使电影制作人在速度、保真度和控制力之间取舍。故事板和情绪板制作快捷、表现力强，有助于尽早探索和传达创作意图 [16, 44]。然而，它们是静态的，空间与时间表征能力有限：无法充分表现动作或时序，因此难以直观呈现复杂的镜头或镜头序列。另一方面，3D预可视化工具能让电影制作人编排场景、尝试摄影机调度，并确保镜头之间的连续性 [12, 34, 54]。然而，这些工具需要高保真度的三维资产、角色绑定和动画专业技能 [40]。许多现有的3D预可视化工具也无法传达情绪节拍和细微动作等细腻之处。

生成式人工智能的最新进展可以通过直接从文本提示生成图像或视频来加速预演，使电影制作人能够快速生成具有引人注目的视觉风格的输出 [17]。然而，它们也带来了新的挑战。文生图和文生视频模型往往难以保持时间一致性，使得跨帧的连贯动作颇具挑战 [36]。它们还缺乏空间定位：物体和摄影机的精确放置、动作调度以及连续性都难以控制。因此，当前使用生成式人工智能的方法有可能产生外观高度精致、却与电影制作人预期结构脱节的结果。电影制作人需要一种轻量且灵活的方法，将三维工具的空间定位与生成式视频工具丰富的表现力结合起来。

我们提出 PrevizWhiz，这一系统使电影制作人能够快速探索并可视化其镜头，方法是将用于时序与空间结构的粗略的3D场景搭建、用于细致角色运动的二维视频参考，以及由图像和文本引导的生成式风格化结合起来。电影制作人首先布置粗略的三维代理模型，以确立道具位置、角色动作以及摄像机路径（图1a）。他们可以对三维场景中的帧重新风格化，以试验不同的美学风格，范围从对其构图的严格遵循到宽松的重新诠释（图1a）。

最后，PrevizWhiz 允许电影制作人指定三个级别的动作保真度：（1）基于三维调度的粗略运动；（2）将基于三维场面调度的运动与重新风格化后的帧相结合的风格化动画；（3）借助二维参考视频增强风格化动画、以呈现细致角色运动的控制视频动画（图1b）。这些场景构图的帧以及基于时间的元素可以引导最终输出的视频生成（图1c），从而塑造风格、灯光、构图与动作，在三维调度的结构一致性与二维生成工具的表现力之间取得平衡，为电影创建预演。

我们的贡献包括：（1）PrevizWhiz，一个结合粗略的三维场面调度、帧风格化与细粒度动画控制的系统，以实现轻量而富有表现力的预演；以及（2）一项针对电影制作人与三维艺术家的用户研究结果，展示了该系统如何在预演过程中实现快速构思，并探究他们对前期制作中生成式工具的看法。

2 相关工作

我们的工作建立在预演实践、用于预可视化的人工智能与生成式工具，以及用于风格迁移和引导式视频生成的生成式方法等既有研究之上。

2.1 电影制作中的前期制作与预可视化工具

预演技术有助于电影制作人在投入全面制作之前，对镜头、动作调度、节奏和情绪进行尝试 [54]。故事板和情绪板等传统方法快速且富有表现力，但仍是静态的 [16, 44], ，在空间或时间保真度方面有所局限。

相比之下，3D预可视化工具 [34] 能够实现精准的构图和动作调度，但需要大量专业知识、资产准备、时间和预算，因此通常用于预算和规模较大的电影制作团队 [54]。测试视频 [11, 26] 通过模仿现场拍摄降低了门槛，但严重依赖实体空间，导致迭代过程繁琐（例如，每个摄影机角度都要重新拍摄动作）。因此，在预可视化方面，电影制作人必须在速度、成本、表现力和保真度之间权衡。我们的系统从这些实践中汲取灵感，结合了粗略草图和故事板的速度、三维工具的细节，以及控制视频（如在线视频和测试视频素材）的表现力。

预可视化还可作为跨学科沟通工具，连接导演、电影摄影师、美术、声音和动画团队 [2, 4, 18]。SyncSketch [1] 和 Autodesk Flow Capture [2] 等工具支持对图像和视频进行标注以便协作，但人们仍需要外观更“成品化”、同时支持快速迭代的预可视化。

我们的工作建立在这类预可视化研究与商业工具的基础上，但我们使用3D环境并非将其视为完成度高、保真度高的素材，而是将其作为跨学科讨论和快速迭代的共享粗略环境。通过将这些环境与生成式二维视频输出相结合，我们实现了轻量却富有表现力的预可视化，既支持粗略状态，也支持精修状态，并在无需深厚三维专业知识的情况下协调构图、灯光、风格和表演。

[1] https://syncsketch.com/ [2] https://www.moxion.io/

2.2 面向预可视化的人工智能与生成式工具

研究人员探索了基于人工智能的工具，以支持超越手绘草图的故事构思与可视化。早期系统，如 Schematic Storyboarding [14]和动态故事板 [40]，能够根据剧本生成静态和动态帧，并导入布景。较新的方法则利用数据检索和生成模型。例如，ScriptViz [39] 根据剧本属性（例如地点、时段、角色）从电影数据集中检索图像，让电影制作人能够即时将剧本可视化。生成式方法也被用于为图像重新设定风格，例如 CineVision [53] 将剧本创作与检索到的电影参考素材相结合，以支持导演与电影摄影师协作。Previs Pro[3] 等商业工具可基于3D场景采集静态故事板，但风格调整能力有限。与之不同，我们的工作强调由用户创作的3D场景，而非由数据库驱动的检索。随后，这些场景可以被重新设定风格，生成精修后的输出，在保留创作意图的同时实现灵活、快速的迭代。

商业和研究工具通过融入视频，拓展了故事板的功能。例如，DigitalFish[4] 和 Jetset[5] 可将实拍素材与3D环境实时合成，而体积捕捉 [20] 则允许演员在虚拟背景中表演。CollageVis [26] 通过分割和重新合成支持2.5D视频合成；RADiCAL[6] 则提供高保真度的动作捕捉和3D预可视化，但需要繁复的素材管线。这些系统要么是静态的（故事板／情绪板），要么依赖数据库（检索），要么成本高昂（体积捕捉／三维资产）。相比之下，PrevizWhiz将粗略三维布景与二维视频参考相结合，并集成生成式视频模型，从而为预可视化生成兼具表现力与结构性的输出。值得注意的是，以往的研究尚未系统地将生成式视频模型 [32] 集成到富有表现力的预可视化工作流程中。

2.3 生成式方法：风格迁移与引导式生成

2.3.1 人工智能在电影制作工作流程中的应用。近期研究表明，人工智能带来的新工作流程正在推动电影制作发生转变：Freeman 等人 [11] 提出在制作过程中进行现场平板电脑编辑；Halperin 等人 [17] 考察了业余创作者如何在电影制作课程中使用生成式人工智能；Chung 等人 [7] 则推出了用于现场直播的人工智能编辑助手。Anderson和Niu [1] 分析了YouTube教程，发现生成式人工智能的使用大多发生在后期制作阶段（例如重风格化、视觉特效、升频）。我们的系统则将生成式视频引入更早的前期制作阶段，使导演和电影摄影师能够在后期制作前进行迭代。

2.3.2 图像生成：风格迁移与基于流的方法。近期方法探索了多种从剧本到视觉内容生成的策略。一些近期方法，例如生成对抗网络 [22, 27, 59]和变分自编码器，尝试直接根据剧本生成视觉场景，并结合脚本分析与图像检索，使文本描述与视觉素材相匹配。虽然这类系统在受限领域中效果良好，但 [39] 往往严重依赖预定义模板、大规模电影数据集或视觉创作提示，这限制了其创作新颖或风格独特视觉内容的灵活性。

相比之下，文本到图像模型（例如Stable Diffusion [41], DALL·E [38], Imagen [42]）能够直接根据文本提示生成全新的图像，提供了大得多的创意自由度。然而，这类模型往往缺乏精细控制能力，也难以在不同场景之间保持结构一致性。风格迁移方法解决的则是相反的问题：在根据目标风格改变图像外观的同时，保留输入图像的空间构图（例如动漫、黑色电影）[52]。不过，这类方法的适用范围较窄，因为它们难以在给定输入之外重新组合或创造新元素。FlowEdit [29] 等基于流的编辑技术超越了传统风格迁移：它们在保持结构对应关系的同时，将特征从源分布迁移至目标分布。这使得更灵活的转换成为可能，能够兼顾内容保留与风格变化。结合ControlNet [57], 等空间条件控制技术，此类方法既能实现与3D场景的可控相似性，也支持选择性重风格化。

在这些进展的基础上，PrevizWhiz 将生成式编辑与保留三维布景空间构图这一结构性益处相结合。通过 Flow Edit 和 ControlNet 方法，PrevizWhiz 能够按电影制作人的需求，对色彩、灯光和风格进行可控调整，帮助他们实现创作愿景。电影制作人还可以决定场景中的哪些方面（如空间构图、动作、色彩或灯光）沿用 3D环境，哪些则通过生成模型重新构想。

2.3.3 通过多模态方法驱动动作。在预可视化中，场景中的动作可通过三维动作库（例如Mixamo[7]）、文本到动作检索 [37], 、文本到动作生成模型 [15], 以及用于视频到动作制作流程的姿态估计方法 [10, 46, 51] 添加。尽管这些方法适用于原型制作，但在处理叙事所需的细微手势、表情和多角色动态方面仍有困难 [9, 13]。

生成式视频模型的最新进展为视频创作带来了新机遇，但在控制、连续性以及整合到制作工作流方面仍面临挑战。Stable Diffusion 等文本到图像模型 [41] 可借助空间引导（如 ControlNet [57]）实现灵活的造型设计，而 Stable Video Diffusion [3] 等扩展以及 Sora 和 Veo 3 等商业系统，则将这些能力拓展到了视频领域。然而，时间连贯性和精细控制能力方面的问题限制了它们在制作中的适用性。此外，尽管这些交互式系统和控制模态取得了令人瞩目的成果，但仍存在重大局限；尤其在复杂设计任务中，控制模态及其格式的直观性仍有待提升。

近期系统（如 WanFunControl [50] 和 VACE [25]）加入了多模态引导（例如骨架、深度和线稿），但由于生成视频的速度和时长有限，仍难以应用于结构化电影制作。借鉴这些近期的生成式视频方法，我们的方法将三维定义的全身

[3] https://www.previspro.com/ [4] https://www.digitalfish.com/onsight/ [5] https://lightcraft.pro/jetset/ [6] https://radicalmotion.com/

[7] https://www.mixamo.com/

动作与基于二维视频的手势／表情控制相结合，创建了一种混合视频引导方法，兼顾宏观层面的场面调度与微观层面的表现力。

3 PrevizWhiz

PrevizWhiz 是一个系统，电影制作人可以通过结合三维场景调度、利用生成式人工智能进行风格化，以及由视频参考驱动的精细角色运动，来创建预可视化。

3.1 设计依据

我们的设计依据围绕降低电影制作人快速创建预可视化的门槛展开，无需高保真度的三维资产或高级技术专长。在构思 PrevizWhiz 的过程中，并结合先前文献（第 2 节），我们确定了系统的三个关键方面：1）使用粗略三维场景调度空间环境；2）支持帧风格化，以生成更精细的预览；3）提供快速创建动作的方法，并支持不同的保真度层级。

R1. 通过粗略的三维场面调度支持场景设置。电影制作人使用预演来探索场景的时空结构。传统的预可视化工作流程往往依赖高保真度的三维资产和熟练的动画师，这会拖慢早期探索。不过，先前的研究表明，粗略或低保真输入有助于快速迭代，进而创作预可视化内容 [26, 40]。受到这些研究和低保真草图绘制实践的启发，我们使用粗略的3D环境为生成式输出提供参照，同时保持迭代过程轻量化。这些粗略场景无需精细制作的素材，即可编码时空线索（例如角色位置、摄影机调度），从而加快尝试和探索。

R2. 通过预览风格支持迭代与交流。预演不仅涉及动作调度和动作，也关乎如何向合作者和利益相关者传达基调与美学。有关故事板 [6], 、情绪板 [23, 45], 和预可视化 [53] 的先前研究强调了快速探索视觉风格的重要性。因此，我们允许电影制作人将从3D环境中捕获的帧风格化为不同效果。用户可以设置重新风格化图像对构图和风格的遵循程度是严格还是宽松，由此在粗略环境与构想中的电影美学之间搭起桥梁。这样一来，电影制作人不仅可以迭代调整理想的视觉效果，也能更清晰地与他人交流。

R3. 支持多个保真度层级。 预演所需的精确度、细化程度和控制程度各不相同。有时，粗略、块状的动作就足以测试节奏；而在其他情况下，则需要细粒度运动，例如详细的手势和情绪化面部表情。为此，我们的工作流程提供了三个控制层级：

(1) 基于三维场面调度的运动直接在三维场景中传达时序和定位。粗略的角色动作（即仅改变虚拟形象的位置而不添加任何动作）以及摄像机运动，均可通过调整其在空间中的位置和朝向，在三维环境中完成设定。

(2) 风格化运动 通过结合三维调度中的动作与定义目标视觉风格的重绘帧来生成视频。粗略的角色位移会转化为更加细致的动作，同时融入身体运动学，并应用一致的预期风格。

(3) 控制视频运动 借助外部视频参考（例如在线视频或测试视频素材），以细粒度运动增强风格化运动。这些参考内容叠加在三维调度之上，作为详细动作的模板，例如手势或面部表情。

3.2 界面概览

PrevizWhiz 是一个由多个具备独立功能的面板组成的 Web 应用程序（图2、图5）。场景调度面板（图2b-c）允许电影制作人配置场景的三维环境，并重新定位场景中的角色、摄像机和元素。该面板还包含一个可编辑的时间线，支持添加和修改关键帧，从而实现角色和摄像机的粗略动画。图像风格化与动画面板允许电影制作人使用文本描述和图像参考，为从三维场景回放中录制的镜头应用视觉处理（图2d）。最后，精细运动控制面板允许电影制作人添加和修改提供运动参考的视频，以进一步调整场景中角色的动作，例如为角色添加面部表情和手势（图5）。

3.3 场景调度与构图

场景调度与构图面板允许电影制作人定义（1）场景构图（图2b），例如使用基本形状、色彩和灯光设置场景的空间布局；以及（2）基于时间的元素（图2c），例如粗略的摄像机跟踪／动作调度，以及为元素创建运动路径。放置在场景中的摄像机用于录制片段，之后这些片段会由生成模型处理，以渲染出最终结果。

3.3.1 三维环境与时间线。

*3D*场景操控。对电影制作人而言，准确构图镜头至关重要，而 PrevizWhiz 提供了一个 3D环境，使他们能够在空间中放置物体。角色、物体、灯光和摄像机都可以放置在场景中，并使用标准的 3D操控工具进行移动（图2b1-2）。用户可以使用标准的平移／倾斜／环绕控制，在场景的主视口中导航，并通过摄像机预览从活动摄像机视角查看场景。

通过改变物体的灯光（图2b1）和着色（图2c4），用户可以在场景构图过程中更好地控制镜头的视觉效果和氛围。调整这些参数可以更真实地反映片场效果，或体现特定场景布光的实际条件。在后续的风格化阶段，用户可以通过指定系统对参考图像的遵循程度，控制着色和灯光的影响（图3和图4）。

3.3.2 摄像机与对象动画。时间线由三个主要轨道组成，用于区分场景中的关键物体：(1) 摄像机轨道（图2c1），用户可通过它创建和编辑用于构图和拍摄场景的摄像机；(2) 动画轨道（图2c2），用于显示角色或车辆等可移动对象的运动路径；以及 (3) 固定元素轨道

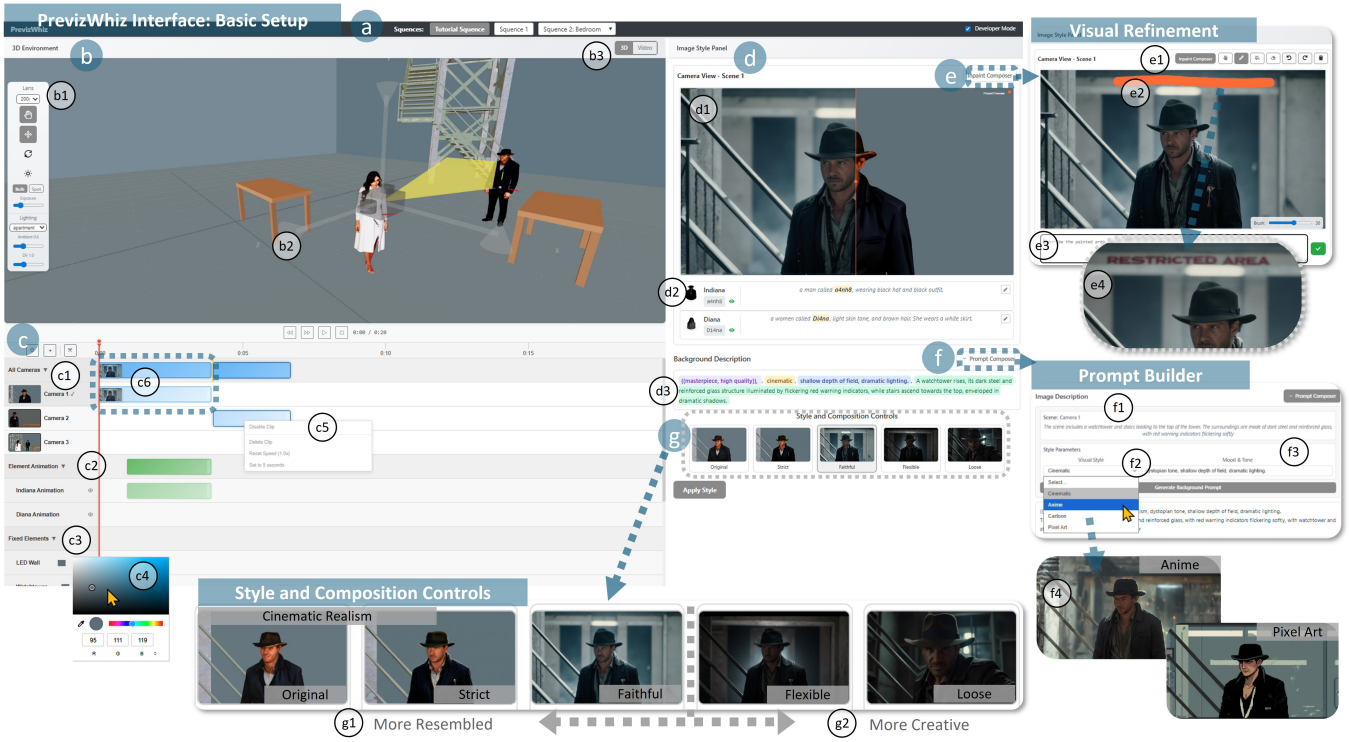

图2：PrevizWhiz场景调度与构图概览：(a) 用户可在此选择不同的序列和场景。(b) 3D环境面板用于设置摄像机(b1)、镜头和灯光，以及(b2)通过平移／倾斜／环绕控制探索三维场景。(c) 时间轴面板用于对摄影机、虚拟形象和元素运动进行动作调度，包括(c1)摄影机轨道、(c2)元素动画、(c3)固定/可移动元素、(c4)颜色调整和(c5)片段编辑；(c6)附有重新风格化图像的片段。(d) 图像风格面板，可在其中比较(d1)三维输入与二维输出，提示词输入包括(d2)自定义角色描述，以及(d3)输入或编写的背景提示词。(e) 图像修补编辑工具，包含(e1)用于识别(e2)可编辑区域的画笔工具、(e3)描述目标细节的文本提示，有助于获得(e4)已应用的细节，例如绘制模板印字。(f) 提示词编写器，包含(f1)基本背景场景描述、(f2)视觉风格选项和(f3)情绪/基调设置。(g) 相似度控制，用于在生成输出中平衡相似度与创造力。该场景的完整视频和场景输出见附录。

（图2c3），用于表示场景中不会移动、但其参数可能随时间变化的静态组件（例如色彩、灯光属性）。这些轨道建立了清晰的场景元素层级，使用户能够轻松地在不同物体和摄像机之间导航。

预设摄像机以缩略图列表的形式显示（图2c1），用户可以直接切换视图。这些预设摄像机代表电影制作中常用的镜头和参数，帮助用户快速取景并了解场景。用户可以通过 + 按钮添加其他摄像机或可运动元素，并将它们的动作记录在相应轨道中。

摄像机与物体运动。 可通过直接在三维场景中记录运动路径来为可移动元素制作动画，既可以通过键盘输入（WASD + Q/E），也可以通过拖动元素并使用3D平移操纵器。这些粗略轨迹（如行走、跑步或跳跃）随后会以片段的形式显示在元素动画轨道（图2c2）上。这些粗略动作提供了结构化的运动路径，是生成式视频模型的关键输入之一，可确保最终输出同时体现空间连续性和时间连续性。

用户还可以通过指定位置、旋转和镜头的起始与结束关键帧来定义摄影机动作，插值会自动处理，与传统的三维编辑工作流类似。多条运动路径可以在主摄像机轨道（图2c1）中组合并编排，从而形成由摄影机和元素运动构成的镜头序列。例如，摄影机可以从角色扣纽扣双手的中景移至面部特写，同时主体向前迈步，以突出缓慢行走动作。

3.4 图像造型设计与动作

位于界面右侧的**图像风格化与运动面板**可优化每个摄像机镜头中场景的视觉外观。界面会向用户显示选定的摄像机视图，用户可以使用一个滑块（图2d1）在原始三维图像输入与风格化后的输出之间连续切换。

镜头的风格以及元素的高层次动作可以通过界面中的文本指定。

类型与风格。 用户可以指定类型和风格，并设定所需的情绪与基调，以定义目标输出。类型（如爱情、反乌托邦、恐怖）主要塑造故事主题和叙事模式，而风格（如动漫、黑色电影、纪录片）则决定呈现故事时采用的美学与视觉方式。即使采用相同的3D构图，风格也会对场景产生显著影响（图4）：电影感风格（a）与卡通风格（b）的效果会大不相同。

对色彩与光照的可控遵循。 定义类型和风格后，用户可以选择 (i) 保留原始（粗略）渲染，(ii) 使用提示驱动的风格重塑（图2d3），或 (iii) 仅宽松地遵循空间构图，从而允许生成结果更具创意。

PrevizWhiz 采用 FlowEdit 方法，在源（3D环境）与目标分布（输出图像）之间进行映射，从而提供四个相似度等级（图 3）。其底层设计意图是体现电影制作人将镜头从粗略布局逐步打磨至精致美学的过程。

- **严格**：坚决保留原始构图与配色方案。
- **忠实**：保留大部分原始配色方案，同时允许适度的创意变化。
- **灵活**：主要受文本提示引导，与原始色彩和光照有所不同，但保持空间构图。
- **宽松**：既偏离原始色彩风格，也偏离空间构图。

这四个等级的设计既参考了迭代实验，也借鉴了以往关于可控生成式图像的研究（例如，类似ControlNet的空间条件控制和潜空间混合 [8, 29, 57]），但呈现给用户时则是直观的艺术控制选项。例如，更改背景色（图 3a/b）会生成近似于所需基调的输出。相比之下，“宽松”模式仅使用 ControlNet [57]，通过深度引导条件控制而不进行潜空间混合，从而支持完全由提示词驱动的修改，例如电影感光照或反乌托邦色调（图 3Loose）。

此选项决定输出在多大程度上保留原始3D输入的色彩与灯光。也可以直接在三维场景中施加其他约束，例如修改基础资产颜色，或尝试不同的灯光效果，如均衡照明与对比照明（图4）。

3.4.1 视频风格面板：重风格化简单的三维运动。从3D视图切换到视频选项卡（图2b3）后，界面会从3D编辑器转变为视频预览，摄像机和元素的所有子轨道都会隐藏，动画导出后则以视频轨道的形式显示。这一变化表明，粗略的空间布局和场景调度已固定，接下来将重点转向对录制视频进行重风格化。

右侧的图像风格面板将转变为**视频风格面板**（图5c），支持视频专用提示词输入。

回放时间线时，视频输出面板中将显示3D动画预览（图5c）。不过，当用户选择某个视频片段时，系统会在面板中将其突出显示为专属预览，并可单独以视频格式回放（图5c1）。这让用户能够流畅地在查看完整时间线和聚焦单个片段进行重风格化之间切换。

如果片段附有风格化图像，视频预览下方也会显示该图像及其相关描述（图5c2），为重风格化提供结合视觉与文本的参考。如果没有附加图像，该区域只会显示“尚无图像”，并保持空白。视频文本描述（图5c3）显示在界面底部，用户可在视频提示词输入栏中扩展或修改此描述。

此外，相似度-创意度控制项（图5c4）允许用户设定动作是遵循已定义的动作，还是完全依据视频描述生成。确认后，该片段便会提交进行视频生成。在处理期间（约一分钟），片段状态会显示进度条，同时界面仍可交互——用户可以编辑其他片段、调整风格化提示词，或准备更多摄影机和虚拟形象动作。

处理完成后，生成的视频会显示在视频风格面板中，并与时间线中的对应片段同步。用户可以在粗略输入动画和生成的视频输出之间切换，以便直接比较并进行迭代优化。实际操作中，用户首先通过粗略动作调度定义片段，然后对其进行视觉重风格化，最后将其批量导出至生成式视频模型进行最终合成。

通过结合粗略3D输入与重新风格化控制项，并允许灵活遵循原始设置，图像风格化面板为用户在预演过程中探索视觉设计选择提供了一套轻量而强大的工作流程。

3.5 精细动作控制

完成场景设置和重风格化后，用户可以评估结果，并在需要时添加更多精细动作控制。如果风格化阶段生成的动作与预期动作不够吻合，或者用户希望呈现特定的动作、视线方向或手势，就可能需要这样做。

3.5.1 使用视频游乐场和视频混剪编辑器进行高级视频重混。 一些更复杂的场景可能同时包含空间和时间元素，例如并行动作：前景是三个人在桌子前交谈的构图，同时有人从前景移动到背景 去拿一个物体。虽然用户可以在三维场景中预先安排这些动作，但录制自己的视频来展示动作可能更有效。这类场景可能需要更精细的控制，而不是像文本这样更具创造力的粗略控制。

视频游乐场。 为了超越基本动作调度（其中动作主要由从三维场景记录的基于时间的元素定义），用户可以在视频游乐场面板中导入外部视频（在线视频或自拍视频素材）作为参考（图5a）。其中角色的动作



图3：启用 LoRA 视觉风格时，图像生成对源色彩的遵循示例：(a) 电影感；(b) 三维卡通。每行展示四种相似度等级（严格、忠实、灵活和宽松），依次降低对原始三维输入（第0列）的遵循程度。每列上方显示的 FE 值指 FlowEdit 参数，CN 值指 ControlNet 参数，二者共同控制生成图像遵循源三维帧的程度。

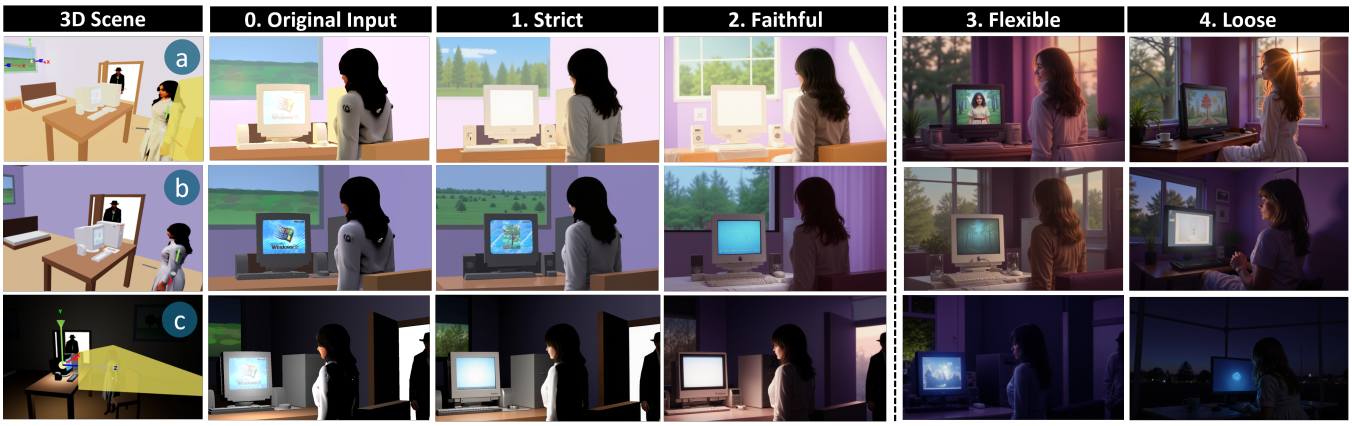

图4：具有不同灯光 变化的场景 (a) 晴天 (b) 黎明/清晨 (c) 暗室，展示四种相似度等级 ls.

视频将被自动提取为三维骨架，并可用于驱动最终视频的生成。

视频游乐场（图5a）展示了视频库，用户可以在其中导入在线视频或自拍视频素材，对其进行裁剪（图5a2），并将其处理为动作（骨架）视频（图5a3-4）。处理后的视频可以拖放到时间线上，并显示为视频轨道（图5a4-5）。

VideoRemix Preview. 处理后的视频将显示在 Video Remix Preview（图5b）中，此处现在可用作视频图层编辑器（图5b1-3），用户可以调整视频大小和位置（图5b1）、将其拆分为多个视频（以处理多个人物的情况，图5b2），并调整这些单独处理后的视频的大小和位置，使其与三维场景中定义的构图位置相匹配（图5b3），然后再将拆分的视频合成为一个引导视频。该视频也可以导入视频重风格面板（图5c1），用于视频生成流程，其方式类似于三维场景原始摄像机输出的结果。此外，这种细粒度的动作（例如，多个人物进行对话，图5b3）可以与三维场景中最初定义的动作（例如，John 在图5c1中移动到后方）混合重组，如图5所示。结合3D环境和二维视频源的能力，我们可以利用3D环境勾勒空间运动，同时还可以使用二维视频控制单个人物的动作，从而提供更细粒度的运动构图。

虽然我们使用一小组具有代表性的场景和风格来展示工作流程，但附录中还提供了更多示例，展示该方法在不同视觉风格、光照条件、场景类型和交互复杂度下的泛化能力。

3.6 PrevizWhiz 操作演示

Alice 是一位年轻的导演，她想为一部短片制作预览，以便向他人提案，争取资金，从而完成影片制作。她希望获得能传达电影叙事和基调的高质量渲染图，但她没有时间学习复杂的三维软件，也没有预算外包预演制作。于是，她转向 PrevizWhiz，梳理一些想法，并生成

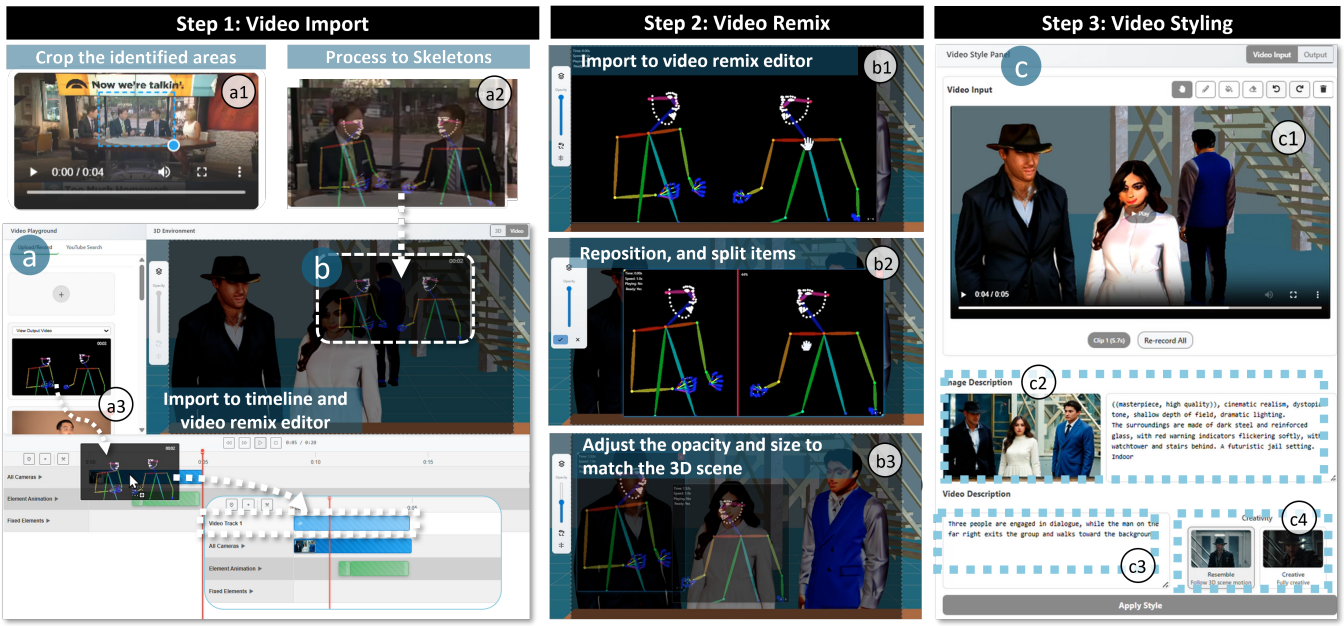

图5：以三人互动示例演示视频风格界面：(a) 视频导入面板：用户可以 (a1) 导入在线或实时视频素材并对其进行裁剪，以及 (a2) 将其处理为骨架视频。这些骨架随后可以 (a3) 拖放到时间线上，并在此 (a5) 创建一条新的视频轨道，以引导场景中的角色动作。(b) 视频混剪编辑器：提供用于处理、对齐和优化已处理视频图层与三维场景中角色位置（例如朝向、手势）的工具。用户可以 (b1) 调整片段大小并重新定位，(b2) 将其分割，以及 (b3) 排列分割片段以匹配角色比例，然后再将其重新合成为引导视频。(c) 视频风格面板：包括 (c1) 已处理的外部视频输入；(c2) 来自图像风格面板的图像描述和参考；(c3) 视频描述字段，用户可在其中指定额外的动作细节；以及 (c4) 相似度–创造力控制，其中相似模式遵循空间定义的运动，而创意模式则基于文本提示生成输出。

第一场景的视觉内容：一名黑客正试图访问一个安全系统，这时一名共谋者走近她，告诉她时间不多了。

在这个场景中，主角正在一间光线昏暗的房间里，坐在发光的电脑前工作（图 6a）。因此，Alice 使用场景调度面板，将黑客和共谋者这两个角色以及电脑添加到场景中。她将角色和道具移至初始位置。她尝试了不同的灯光和色彩试验（图 4a-c）来营造理想的氛围，最终调暗房间，并在电脑屏幕上增添微光，营造出预期的视觉氛围（图 6a）。Alice 在场景中添加了两台摄像机：一台从黑客肩后取景，对准屏幕；另一台则将两个角色一同纳入画面，展现他们交谈的场景。为了给场景添加基础动画，Alice 选中共谋者并录制一条运动路径，使用 WASD键操控角色沿轨迹走向黑客。随后，她选中过肩镜头并添加一个关键帧，让镜头跟随共谋者走向黑客（图 6a-b1）。

完成基本动作调度后，Alice 转到图像风格化与运动面板（图2d-g）。她尝试了不同的风格重塑选项，并选择“忠实”遵循程度，让系统可以更大程度地偏离三维调度。她还指定了一个大致的动作方向：一名男子走向一名正在台式电脑上打字的女子，并与她交谈。经过几次迭代后，Alice意识到她喜欢这种风格，但动作看起来不够可信。

Alice 转到精细运动控制面板（图5），尝试将生成的视频与外部拍摄素材结合，以更准确地为角色制作动画。她找到一段包含讨论场景的在线视频，并将其导入 PrevizWhiz（图 6b2）。PrevizWhiz 将其处理为基于骨架的视频引导，她随后可以将其叠加到视频上作为引导。处理后的视频包含手势和面部表情细节，使她不仅能为单个角色制作动画，还能捕捉角色之间的群体动态（图5b）。Alice 还可以对齐视频，并将共谋者的三维运动路径（图5b）与展示黑客和共谋者之间互动的二维视频引导相结合，从而创建更复杂的交互动态。

调整对齐方式并生成新视频后，Alice 对最终结果感到满意（例如，图6c2）。通过这三个步骤，Alice 得以搭建场景，并确信自己能够将其拍摄出来；生成的视频片段不仅动作和表情可信，还传达了 Alice 希望在电影提案中表达的基调和情绪。生成的视频与原始三维输入的完整关键帧对比见附录（图12）。

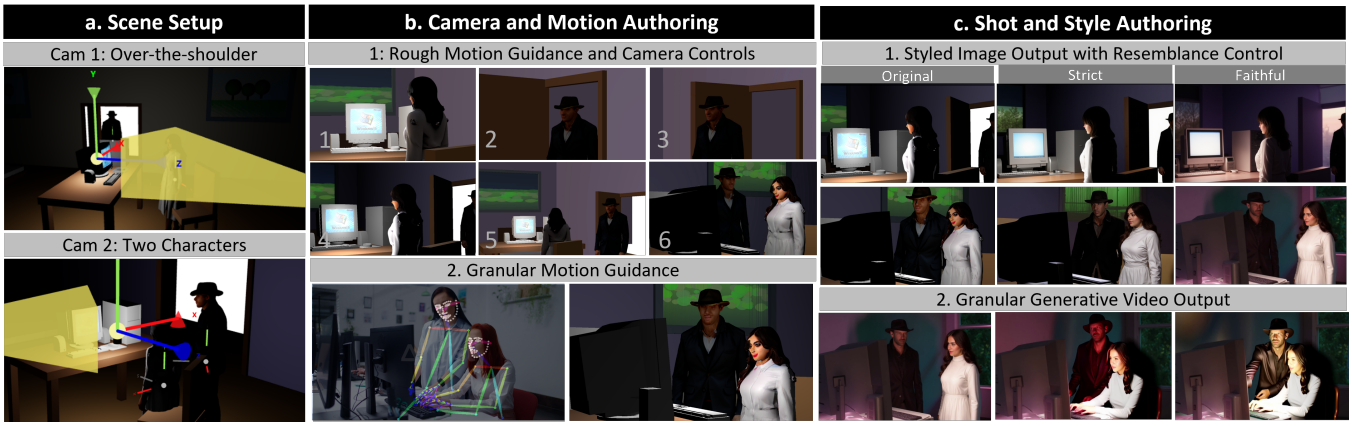

图 6：PrevizWhiz 操作示例：(a) 场景设置。导演在三维调度面板中放置两个角色，调整道具、灯光和色彩，并配置互补的摄影机角度。(b) 摄影机与动作创作。导演先定义粗略的动作引导和摄影机布置（1–6），再利用控制视频参考和骨架对齐应用细粒度动作引导，以完善手势、身体姿势和交互时序。(c) 生成镜头与风格创作。图像风格化面板通过相似度等级（原始、严格、忠实）预览灯光预设，以探索视觉基调。细粒度生成式视频输出进一步将动画化三维调度与外部动作影像融合，增强最终序列中的写实感、灯光连续性和角色互动。

3.7 实现

前端 Web 界面使用 React+Vite 和 React Three Fiber 构建。后端服务器使用 Node.js 构建。服务器实现了一条视频处理流水线，可接收视频，使用 FFmpeg 对其进行裁剪和处理，然后将其发送至 ComfyUI 后端[8]以作进一步处理。ComfyUI 工作流程包含一组 JSON 文件，用于生成处理后的视频（深度、骨架、边界框、轮廓）、合成工作流（用于合成视频，例如深度视频、骨架视频），以及图像到图像工作流（使用 ControlNet [57],扩散模型 [8],、Flux [30, 31],和 Flow Edit [29]），还有图像/文本/视频到视频模型（Wan 2.1 [50], Wan Fun Control [49], VACE [25]）。为提高创作效率，我们实现了提示词编写器（图2f1-3），帮助用户在无需提示词工程专业知识的情况下编写有效提示词。它为所选摄影机画面的视觉风格、情绪/基调以及背景环境描述提供了结构化输入字段，并通过调用 Ope-nAI GPT 服务[9]将其扩展为详细的自然语言提示词。

3.7.1 元素。对于教程和示例场景，场景元素由程序化几何体和导入的三维模型共同组成。家具（桌子、椅子、柜子和床）由立方体等基本几何体构成，并使用标准材质和适当的色彩进行着色。更具体的物件（旧镜子、台式电脑和瞭望塔）是从 SketchFab 下载的 GLB 模型[10]。此外，我们场景中的三维角色是使用混元3D [47] 根据二维全身角色图像生成的人工智能图像（图 7b1）。由于我们的大多数参考图像都是半身肖像，我们首先为每个角色制作了全身版本使用 Flux Kontext [31], ，然后将这些合成的全身图像用作混元3D的输入。

3.7.2 角色与风格低秩适应（*LoRA*）。为了增强扩散模型等大型预训练模型的能力，LoRA [21] 通过为现有模型微调训练新的权重，使其能够掌握更专业的知识或风格，从而在 PrevizWhiz 中实现角色一致性和不同的风格化效果。

角色 *LoRA*。例如，从一张捕捉到三维场景中 3 个角色的图像开始，系统会调整该摄影机图像的尺寸，并使用 Florence-2 [55] 的边界框检测功能定位出三个“人物”实例。每个 Florence 边界框都会发送给 SAM2 [28]，将粗略的立方体转换为每个角色像素级精确的蒙版；随后对蒙版进行柔和扩展（扩展 12–18 px，模糊 3.5–6），并将其合成，以便从背景图像中分离每个角色，确保重新风格化时边缘不会留下接缝。接着将所有角色蒙版合成为一体，并对结果进行反转，以获得干净的背景蒙版。

分割后的角色各区域（图 7:b2）随后分别接受各自的条件引导：通过 CLIP 钩子应用带有身份LoRA的角色专属提示词，而背景提示词（图 7:b3）则单独进行引导（例如，营造一个由钢铁和玻璃构成、带有微妙红色警示标志的电影感反乌托邦监狱）。将角色与背景图像分离后，便可形成统一的复合条件引导：在各自的蒙版内，角色由其专属提示词/LoRA控制；其他区域的环境则由背景提示词驱动。

风格*LoRA* 仅通过文本描述来引导风格，效果可能不够一致。因此，图像风格面板中选定的视觉风格也通过动漫、卡通、像素艺术和写实风格的 LoRA 进行引导。随着更多 LoRA 经过训练并发布，

[8]https://github.com/comfyanonymous/ComfyUI [9] https://openai.com/index/openai-api/ [10] https://sketchfab.com/feed

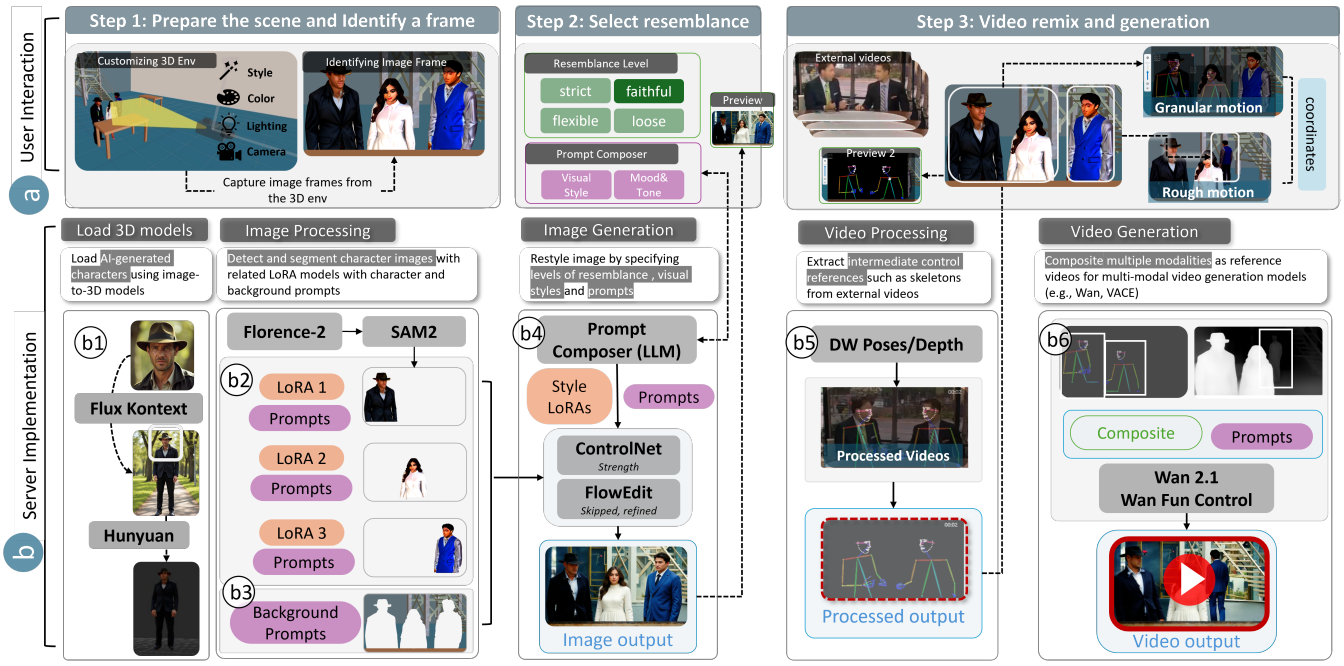

图 7：PrevizWhiz 创作与生成流程概览。上行展示面向用户的交互：(1) 准备三维场景，(2) 选择相似度等级，以及 (3) 导入外部视频以进行动作优化。下行展示相应的服务器端实现：(a) 加载三维模型，(b) 使用 LoRA 处理角色区域，(c) 生成重新风格化图像，(d) 提取姿态/深度参考，以及 (e) 组合多模态输入，使用 Wan/Wan-Fun Control 进行视频生成。

未来可以将它们添加到系统中，从而提供更大的灵活性。

*3.7.3 FlowEdit 与 ControlNet 参数。*生成图像的生成过程（图 7：b4）共运行 20 步。在 20 步生成过程中，通过调整跳过的步数和 ControlNet 强度，创建了四个相似度等级：(1) 严格：跳过 5 步，ControlNet 强度为 0.7。它强烈保留原始图像的色彩、风格和空间结构；(2) 忠实：跳过 1 步，ControlNet 强度为 0.7。该等级同样保留原始图像的色彩、风格和结构，但保留程度略低于严格。(3) 灵活：跳过 0 步，ControlNet 强度为 0.7。它保持原始空间构图，但生成新的色彩和风格。(4) 宽松：跳过 0 步，ControlNet 强度降低至 0.3。它在色彩、风格和空间结构上偏离原始图像，提供最大的创意自由度。这四个相似度等级是作者基于实验选择的，旨在实现直观的艺术控制，同时也参考了先前的工作 [29, 57]。它们具有代表性，用于展示四个遵循程度之间的差异，但当用户选择不同的步数时可以调整，同时也取决于原始三维支架的粗糙程度。

*3.7.4 用于视频生成的骨架提取与重组流程。*为从外部视频中提取动作（图 7:b5），PrevizWhiz 会逐段处理每个片段，为每一帧生成骨架关键点与深度图。用户可以裁剪、重新定位，并在视频混剪编辑器中缩放处理后的骨架视频，使角色位置与三维场景的空间布局对齐。调整完成后，系统将骨架与深度镜头序列（图 7:b6）重新合成为条件视频，随后将其作为多模态输入提供给生成式视频模型。这使得最终输出既能遵循三维调度，也能遵循从参考视频素材中提取的细粒度手势。

*3.7.5 硬件配置。*图像与视频生成模型在一台独立的 RTX 5090、24GB 显存、64GB 内存计算机上运行，并通过局域网以 HTTP POST 访问。我们在用户研究中使用了 1.3B VACE 和 WAN 模型，因为它性能更高，能够更快地生成图像。

4 评估

我们开展了一项研究，以了解电影制作人如何使用 PrevizWhiz 生成预演内容。该研究旨在收集电影制作人在使用系统不同组件和功能时的定量与定性观察及反馈。本研究已通过我们的机构伦理审查流程审核并获得批准。

4.1 参与者

我们通过定向招募、社区渠道和滚雪球抽样招募了10名参与者（5名女性、4名男性、1名非二元性别者），年龄在23至42岁之间。其中8人为电影制作人和创意专业人士

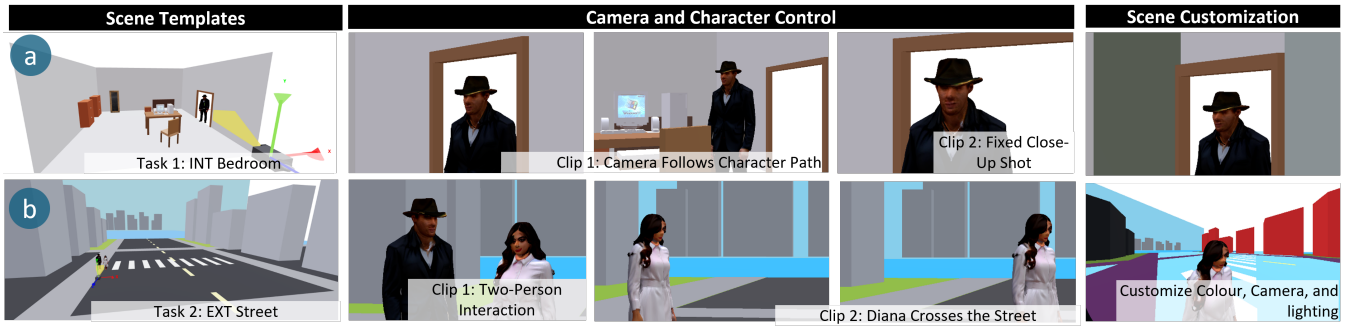

图 8：研究场景与任务（a）使用内景卧室的任务 1；（b）使用外景街道的任务 2

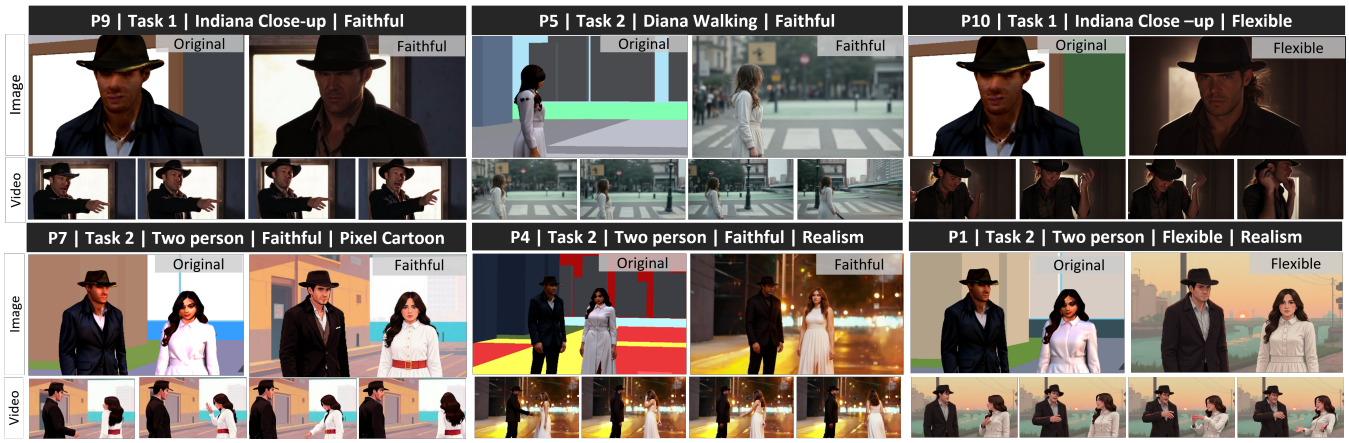

图 9：用户研究中生成的图像与视频。用户尝试了不同的文本提示词、视觉风格和引导视频，并调整色彩、灯光、摄像机位置和镜头，以获得最终视频输出。

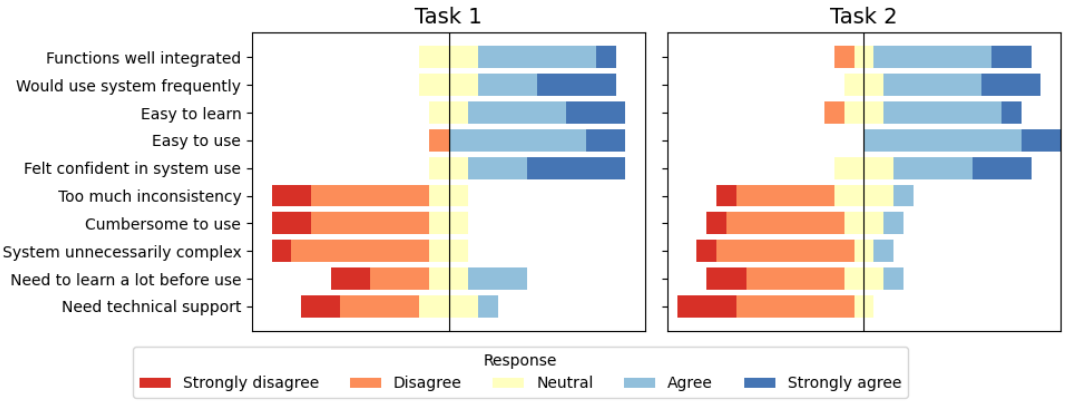

图 10：系统可用性量表问卷回答概览

（电影摄影师、导演、广告导演、技术导演以及服装/造型专家，其中3人还拥有使用三维工具的经验），其余2人为曾在电影制作行业工作过的3D/动画专家。

参与者拥有1至15年的从业经验（$Mean = 7$，$SD = 4.3$），从职业生涯早期的从业者到经验丰富的专业人士均有涵盖（详情见附录表1）。参与者的背景及相关技术的使用情况各不相同。六人表示目前正在使用三维制作流程，其余人员接触三维技术的经验有限。七位参与者曾

使用过图像生成工具（例如，MidJourney、ComfyUI），五人探索过视频生成（例如，Runway、Veo、Kling），通常用于故事板、提案演示文稿、海报或探索性片段。六位参与者使用过3D预可视化工具（Previs Pro、CineTracer、Set aLight 3D），三人依赖传统方法（故事板、动态分镜、动作捕捉），一人表示从未使用过预可视化工具。他们的制作工作涵盖大型电影和广告（需要团队协作），也包括独立电影和小规模内容创作，体现了从高预算工作室制作流程到灵活务实的独立制作实践的多样性。参与者获得了135美元报酬，研究耗时90至120分钟。

4.2 流程与任务

参与者填写知情同意书并完成人口统计调查后，观看了一段介绍系统及其使用方法的教程，随后完成两项研究任务，使用该系统为室内和室外场景生成预可视化渲染图。完成任务后，参与者填写了一份调查问卷，并参加了半结构化访谈。实验使用的硬件设备与 §3.7.5 中所述的相同。

4.2.1 教程视频（*10-15* 分钟）。参与者观看了一段介绍系统的教程视频，实验人员则讲解了系统的功能和操作方式。教程中的场景是一名正在行走的人，与两项研究任务中使用的场景不同。在整个研究过程中，如果参与者在完成任务时需要帮助，实验人员也会随时解答疑问。

4.2.2 研究任务（*50–60* 分钟）。我们提供了两个任务和两个粗略场景，要求用户通过更改背景色彩或灯光来反复调整并创建自己的场景。场景中预设了摄像机，参与者可以调整其位置、角度和视场角，以创作自己的叙事。

任务 *1*（*20* 分钟）：单人场景。实验人员介绍了摄影机调度、角色动作、图像与视频生成流程等功能的交互方式，并要求参与者进行相应的具体操作。用户拿到了一段卧室场景的剧本（图 8a）：

> **内景：卧室**
> Indiana 站在卧室门旁，随后走进卧室。

根据这段剧本，参与者需要使用两个预设摄像机创建两个连续的视频片段。其中一个摄像机需要能够移动，以捕捉角色的运动路径（走进卧室）；另一个摄像机则固定拍摄特写镜头，捕捉人物的面部表情，并参考一组预先载入的外部视频片段。参与者还需要调整墙壁的色彩与灯光。

任务 *2*（*25* 分钟）：多人场景。与前一个任务类似，实验人员通过一个多人对话场景介绍了各项功能。此任务重点展示了其他功能，例如基于详细 LoRA 的角色描述和视频图层拆分。该任务使用的示例和场景是街景（图 8b），多人对话剧本如下：

> **外景. 街道**
> Diana在人行道上与Indiana聊天，随后Diana离开并穿过马路。

在此任务中，参与者需要创建两个片段：第一个片段中有两个角色在说话，第二个片段中其中一个角色离开并穿过街道。尽管关键场景元素为了各场次之间的可比性而被固定，该任务仍被设计为提供充分的创意自由。参与者可以修改街道场景的色彩、灯光和风格（包括街道、人行道、地面及周围建筑），并定义摄像机镜头和动作。此任务使我们能够考察多角色拍摄，其中使用外部视频源来为更复杂的互动制作动画，参与者需要编辑并重新混合引导视频。

4.3 数据收集与分析

每完成一项任务，参与者都会填写系统可用性量表（SUS）。所有任务结束后，参与者填写一份研究后问卷，回答有关系统功能和用户体验的问题，随后参加半结构化访谈，深入谈论他们的体验、对系统的印象、我们提供的创作层级以及潜在应用。所有文本、图像和视频输入与输出均被记录，整个过程也进行了视频和音频录制。

我们使用归纳式主题分析法分析定性数据 [5]。首先，我们对参与者的口头陈述和访谈回答进行开放式编码，以识别系统中反复出现的概念和使用模式。随后，我们反复修订编码并将其归纳为更高层级的类别；合著者也定期讨论新出现的解读，以确保分析具有一致性和严谨性。这一自下而上的过程确保主题直接源自参与者的体验。

5 研究发现

在分析中，我们将洞见归纳为三个主要主题：**(1)** 灵活高效的创作工作流；**(2)** 在控制与创意探索之间取得平衡；**(3)** 在制作前期使用人工智能的益处与顾虑。下面我们将总结每个主题。

5.1 灵活高效的创作工作流

总体而言，参与者认为 PrevizWhiz 易于学习、操作快捷，并且非常适合快速迭代；他们尤其肯定了在三维场景中进行直观的摄影机和角色动作调度，以及快速生成预览的功能。图 10 展示了 SUS 调查结果。研究后的调查问卷显示，参与者认为，记录摄影机和角色的运动路径来引导视频生成很有帮助（摄影机：$Median = 4$，$IQR = 1$；角色：$Md = 4$，$IQR = 0.5$），风格控制灵活（$Md = 4$，$IQR = 1$），而且用户感受到了充分的创意控制（$Md = 4$，$IQR = 0$）。他们还表示，生成的输出对他们来说合乎情理（$Md = 4$，$IQR = 1$）。与此同时，参与者对于生成内容是否符合自己的想象意见不一（$Md = 3$，$IQR = 0.5$），这凸显了快速生成与结果契合度之间的张力

> 与熟悉的工具相比，参与者认为其相对于传统方法具有速度优势（$Md = 5$，$IQR = 2$），且生成式流水线的输出质量相当（$Md = 4$，$IQR = 0.5$）。

几位参与者将 PrevizWhiz 与专业预可视化软件（例如 C4D、Cinetracker）更陡峭的学习曲线进行对比，指出它所需培训更少，却能产出可用的结果；一位参与者将其概括为“基本上就像剪辑软件的人工智能版本 *[...]* 让事情更方便”，从而减少了在 MidJourney 和 Runway 等多个工具之间来回切换的需要（P9）。

5.1.1 三维到二维工作流：在精确度与灵活性之间取得平衡的折中方案。参与者认为，三维到二维工作流是一种切实可行的折中方案：它比构建完整且高保真的三维制作流程（例如 Cine Tracer）更快捷、更省钱，同时又比文生视频或图生视频方法（例如 MidJourney、Runway）提供更多结构与控制。这使 PrevizWhiz 兼具高效性与可靠性，尤其有助于早期创意对齐。

几位参与者指出，这种平衡通过提前厘清想法，降低了高成本修改的风险。例如，P9 解释道：“这种呈现形式也有助于减少我们需要进行的修改次数，并降低整体沟通成本。”他们还强调了其在提案场景中的益处，因为客户的预期可能与专业人士的设想不同：“如果客户没有相关专业背景，他们的想象可能会完全不同。 *[...]* 使用这类软件*[...]* 不需要投入大量资金。 *[...]* 即使能节省约 *10%*，这种方法也值得采用。”

这些发现表明，PrevizWhiz 不仅加快了创作速度，还充当了电影创作者与技术专长各异的利益相关者之间的沟通桥梁。

5.1.2 快速迭代与重风格化实验。参与者重视从粗略脚手架快速过渡到精致成品的能力，并通过重风格化测试卡通、写实或电影感等多种美学风格（用户研究中的示例见图 9）。这种工作流程让他们能在短时间内生成多个方案，并灵活尝试不同的色彩、灯光和摄影机布置。

参与者认为这一过程既易于上手，又富有成效。P6 强调，它“非常容易上手，使用简单，而且能在相对较短的时间内生成大量方案”，而 P8 则指出，能够“先粗略、快速地形成一个想法……然后再迅速尝试不同的最终美学风格”很有帮助。

总体而言，这些发现凸显了从粗到精流程的价值：从低保真脚手架出发，快速探索构图，再逐步完善，获得更高质量的成品。PrevizWhiz 不仅是一种预可视化工具，也是一种创意草图工具，支持快速试验，而不会将用户束缚在僵化的制作管线中。

5.2 平衡控制与创意探索

参与者强调，PrevizWhiz 在多个维度（色彩、灯光、构图、动作和服装）上支持创造力，同时还能根据他们的专业角色和创作目标提供不同程度的可控性。

5.2.1 特定角色的创意优先事项。不同参与者对工作流程的不同方面赋予了优先级。例如，P5（从事造型与服装工作）重视尽早尝试灯光和道具，以确保服装配色与环境和演员相协调。相比之下，P7（从事动画工作）强调了进一步细化角色动作的重要性，这超出了关键帧或定格动画近似手法所能实现的效果。这些差异彰显了该工具的跨学科价值。

5.2.2 平衡可控性与创造力。 大多数参与者更喜欢在保留动作框架核心要素的同时，仍为风格变化留出空间的输出。十人中有七人更喜欢忠实 相似度设置，欣赏其在遵循与精致之间的平衡：P5 指出，“‘忠实’感觉非常可控，并且具有足够的预期风格”（P5）。而 P6 指出，“它保持对所有参数的控制 *[...]* 但仍为意外惊喜保留了可能性。”

三位参与者更倾向于更开放的模式。P3 认为 灵活 有助于“ 增强光影以营造更强的情绪，” 而 P8 看重它留下了 “ 让我的文本提示真正对最终结果产生美学影响的空间。” P9 更喜欢宽松，并将其描述为 “更真实，细节更丰富，更具电影感，也带来更多意想不到的结果。”

5.2.3 视频输出中相似模式与创意模式的偏好。视频生成中也出现了类似模式。七位参与者更偏好相似模式，认为其具有可预测性（“动作基本满足了用户的要求，没有多余内容”——P2）；而三位参与者选择了创意模式，认为它更具表现力，也更能引发思考（P3：“更有趣，也更发人深省”）。一些参与者指出，两种模式在不同情境下都很有用：创意模式可以引入意料之外的张力或动感，而相似模式则能确保与规划好的动作框架保持保真度（例如 P3、P9）。

5.2.4 动作框架：粗粒度与细粒度动作。参与者强调，根据目标选择不同动作保真度的价值。在早期动作调度阶段，行走路径或镜头摇摄等粗略轨迹通常已经“足够好”，可以用于测试节奏和镜头构图。几位参与者指出，这种控制层级感觉快速而轻量，降低了试验的门槛。

与此同时，参与者也很欣赏能够在需要时进一步细化为细粒度运动。参与者表示，详细的手势、视线或表情对于某些叙事时刻至关重要；与使用关键帧或定格动画的传统预可视化工具相比，拥有超越粗略动作调度的选项，使他们能够捕捉更加细微的动作。

5.2.5 跨模态考量中的不匹配。参与者还指出，在传达情感或意图时，不同模态有时无法相互匹配。例如，P7指出，动画通常依赖夸张的动作来传达目的和戏剧性：“他们会过度夸张动作，让它看起来更有目的性……甚至还会加入一些毫无意义的帧和极端动作，使整个动作显得更加有意图。”这揭示了真实感捕捉与风格化表现力之间的差距，而动画专业人士往往依赖后者。

另一项挑战是需要手动协调文本描述与3D环境中的视觉元素。P5 提议，系统可以通过自动检测不匹配之处来改进

例如，提示词指定了忧郁的基调，但搭建的场景仍保留明亮、高对比度的灯光。

综合来看，这些发现表明，PrevizWhiz 能够在不同动作保真度层级之间提供灵活性，同时支持大范围笔触探索和微观层面的表现性细节。它允许电影制作人根据任务、角色和前期制作阶段，在控制与创造力的连续谱上选择自己的位置——从精确控制到开放式生成变化。与此同时，不同模态之间偶尔出现的不匹配凸显了强化反馈与整合的必要性

5.3 人工智能在前期制作中的益处与担忧

参与者对于人工智能生成视频在预可视化和制作中的角色持有不同看法。尽管大多数人认为人工智能是一种支持现有工作流程的辅助工具，但也有人认为，在资源有限的小型制作团队中，它可以得到更广泛的应用。与此同时，参与者也对可控性、专业规范和协作动态表示担忧。

5.3.1 专业质感与对外沟通。参与者一致表示，经过人工智能重新设计风格的输出帮助他们以更专业的方式向非专业利益相关者（例如投资者、资助评审人）展示想法，减少了对利益相关者想象力的依赖。这与传统工具生成的粗糙三维或屏幕抓取输出形成对比。例如，P8 表示，经过润色的人工智能预览“可能会让我显得更专业”，而且与传统预可视化工具生成的静帧相比，在资助申请场景中可能表现更好，“因为要看三维人偶，*[……]* 需要发挥很多想象力”。P3 同样区分了不同受众：较为粗糙的预览对于内部团队来说已经足够，但外部受众往往需要更精细的呈现。

5.3.2 适应不同团队和工作流程。参与者认为，该系统广泛适用于不同制作风格和规模，也适用于“简单的视频或电影”（P3），同时也承认其适用性取决于团队和导演的偏好。导演的工作风格会显著影响预览的使用方式。有些导演更偏好严格对齐，要求“要求*[]* 拍摄素材与实际拍摄的内容完全一致。*[……]* 另一些导演则比较随意 *[……]* 这反映了导演的风格。”（*P3*）

一些参与者将定制化配置与预制场景库作了对比，并指出人工智能辅助前期可视化可以减少搭建开销：“例如，如果我们想要两个人在香港风格街道上打斗，我只要从三维制作人员提供的数百个场景中选一个，加载后再放置角色即可。但之后我就得根据场景已有的内容调整镜头设计”（P9）。借助人工智能，镜头或许不必从库中挑选，而是可以更轻松地进行定制。

这些观察表明，尽管 PrevizWhiz 可以适配不同的工作流程，但其作用在很大程度上取决于团队结构和导演的创意方式，以及他们会在多大程度上采用人工智能。

5.3.3 可控性与意图对齐。一个反复出现的担忧是，AI输出能否可靠地符合剧本和创作意图。几位参与者认为，当前的制作流程可控性不足，因此将人工智能的应用定位为探索性或局部用途，而非端到端的解决方案。例如，P9指出：“缺点*[...]* 在于它不太可控 *[...]* 仍然需要时间来完善。*[...]* 我们不敢向客户承诺 *[...]*，只能说我们可能会用它来辅助。 *[...]* 如果它能适用于整个项目，那当然很好。但如果不行，它就只能作为参考。”此外，参与者还指出了跨模态错位（例如动作夸张与预期情绪不符、色调提示与灯光不匹配），这凸显了加强文本、风格和动作之间反馈与整合的必要性。

5.3.4 人工智能与电影制作中的创造性劳动。 关于人工智能在电影制作中的讨论，反复出现的一个主题是对岗位替代的担忧：参与者指出，生成式系统虽然能够加快预可视化和创意构思，但也引发了人们对这类工具可能如何重塑预算、角色和工作保障的担忧。P9担心客户对人工智能的看法可能会直接导致资源减少：“我们不太依赖人工智能进行预可视化的另一个原因与预算有关。如果客户知道可以使用人工智能，他们可能会削减预算。例如，一部电影原本可能有 x 的预算，但他们可能会想：‘既然你们用一两个人借助人工智能就能完成，能不能按 y 的预算做？’”在规模较大的制作团队中工作的 P3 也表达了类似的担忧。她指出，完全依赖人工智能生成内容（AIGC）可能会削弱既有的团队结构，将协作式创作变成一种节约成本的手段：“我们仍然倾向于采用基于团队的分工。完全利用 *AIGC* 是不可能的。如果真这么做，那就只是为了节约成本，并把其他一切都抹去。”她进一步强调了完全采用 AI驱动流程所带来的影响：“一个人就能完成……那我们就会失业。原本这些工作应该由我们来做。”

这些观点反映了更广泛的行业焦虑：人工智能或许能加快某些任务的完成速度，但也可能使专业角色整合为规模更小、专业化程度更低的团队，尤其是在协作已成常态的制片厂规模环境中。

然而，并非所有参与者都对这一转变持负面看法。P6从事规模较小的短剧制作，根据自己参与AI生成短剧的经验，提出了更乐观的观点。他将人工智能在制作中的作用分为几个层次：“首选是纯*AIGC*，也就是说，完全交由*AIGC*直接生成，不进行任何实拍。其次是先用三维生成，再进行转换，这与目前你们软件支持的工作流程类似。再往后，才是先实景拍摄，再转成三维，最后转换为*AIGC*。因为每增加一步，难度和复杂度都会上升，但与此同时，你能实现的上限也会提高。所以，核心原则是：如果人工智能单独无法处理某些内容，或者简单工具解决不了，就借助三维场景和真实实拍来应对更复杂的情况并完成制作。”对P6而言，人工智能驱动的电影制作并非威胁，而是一种机遇：它降低了独立创作者和小团队的门槛，使他们能够制作出受限于自身条件、原本无法完成的作品。他尤其看重PrevizWhiz利用粗略三维场景，在轻量级AI输出与更可控、更精致的结果之间搭建桥梁的作用——这表明存在一种混合路径：人工智能增强而非取代创造性劳动。

6 讨论

用户研究揭示了电影制作人如何与三维场景和生成式视频互动，并引出了关于人工智能的实用性和生成媒体保真度的新问题。

6.1 场景调度

参与者很喜欢能够在 3D环境 中定义色彩、灯光和摄影机角度，这使他们能够确定镜头的整体构图、基调和感觉。他们也很看重将直接在三维场景中创作的运动路径与从外部 2D视频输入中提取的细致动作相结合的灵活性，这符合 **R1** 关于实现灵活场景设置的要求。我们发现，导演很欣赏创建粗略场景的简便性，因为之后可以将其重新设计为精致的视频输出；不过，他们希望能更多地尝试图像和视频生成，通常他们会指派其他工作人员执行较为技术性的操作。

参与者很欣赏 PrevizWhiz 无需访问高保真电影数据集 [53] 或详细的 3D环境 [40]，而且他们可以使用现有角色和物体，以及广泛可用的视频来创建自己的场景。这启发了未来的工作方向：从现有场景中引入一些人工智能素材，或对现有场景进行去纹理化，将其作为起点，让用户据此定义粗略的色彩和灯光，从而打造自己的场景。

6.2 通过多模态意图激发创造力

我们以 PrevizWhiz 为原型，考察了用户如何借助粗略的三维场景设置、风格化预览和生成式视频进行动作生成。该工作流程使用户能够以不同程度的相似度重新设计环境，在忠于原始场景与受文本提示引导的创意性重新诠释之间取得平衡（**R2**）。参与者强调，将在 3D环境中定义的运动路径与外部二维视频中的参考动作相结合，既灵活又富有表现力，这与旨在支持多个保真度层级的早期设计依据相符（**R3**）。总体而言，参与者更倾向于在色彩、灯光和构图方面忠实于他们创建的3D场景。

一项核心挑战是更有效地关联视觉和文本输入，以实现更好的控制。与此同时，不可预测性也引发了担忧。创意视频输出有时被认为“太难以控制”，偏离预期的提示词，或导致情感、灯光和动作之间出现不匹配。例如，特写参考视频能有效传达面部情感，但在广角镜头中，这些表情可能会变得不够鲜明。参与者还指出，不同类型的作品对动作幅度的要求各不相同：自然主义戏剧偏好含蓄，而动画往往依靠夸张的手势来增强叙事清晰度。要解决这类不匹配，系统必须能够适应不同模态、检测不一致之处（例如，文本中定义的情绪与视觉内容中的色彩搭配不一致 [56]），并据此调整素材。

6.3 素材粗糙度与迭代成本

在当前原型中，可以纳入保真度层级各异的三维素材，从粗糙的人工智能生成物体到从外部来源下载的、更平滑的预定义几何体。这些素材主要用作环境中的代理对象。粗糙程度会直接影响细化过程：较平滑的几何体通常只需较少的迭代即可获得精致的输出，而较粗糙的素材，尤其是人工智能生成的角色，可能需要更全面的细化。例如，在我们的测试序列中，单次细化不足以达到理想效果，而经过 19 次迭代后，结果显著改善。这种差异凸显了素材质量对细化流程效率的影响。

虽然这些观察主要突出了使用 FlowEdit 进行迭代时的技术考量，但也为设计提供了指导。**粗糙度**不应仅被视为一种限制，也应被看作一种能够为早期构思提供灵活性的特征。设计师和电影制作人可以有策略地决定：何时粗糙代理对象已足以用于构图和实验，以及何时值得投入资源制作更平滑的素材，以减少后续的精修工作。

6.4 人工智能如何通过预可视化塑造协作式前期制作

除了个人创作方面的益处之外，参与者还反思了人工智能辅助的前期可视化所带来的更广泛机遇与局限。他们认为，从粗糙到精致的工作流程对于沟通与协作具有价值，尤其是在与缺乏技术或电影专业知识的客户或利益相关者分享想法时。将结构清晰度与风格润色相结合的早期预览，有助于避免在后续制作中出现代价高昂的误解。

综合来看，这些发现表明，人工智能在预可视化中的作用不仅是技术性的，也是协作性的 [48]。参与者强调，角色和职责会根据团队规模与制作规模而变化。在大型团队中，导演可能会将预可视化作为跨部门沟通的载体；而在较小的团队中，人工智能驱动的前期可视化则可以延伸至制作本身，成为传统制作流程的轻量级替代方案。

通过将粗略的3D环境与生成式重风格化和视频重混相结合，PrevizWhiz使电影制作人能够在不同角色之间协调创意意图，同时兼顾他们对控制与创造力的不同偏好。这体现了人工智能不仅有潜力塑造前期制作中的电影制作实践，也能影响更广泛的创意协作生态系统 [36]。

6.5 电影制作中生成式人工智能的伦理考量

多位参与者提出了一些问题，指向未来工作中的伦理考量 [24]（见§5.3.4）。尽管人工智能驱动的前期可视化实现了快速迭代和创意灵活性，参与者也指出了失实呈现、不可预测性和创作归属方面的风险。例如，在更具“创意”的模式下生成的结果有时会大幅偏离预期提示词，产生可能在早期决策过程中误导合作者或客户的动作或情感。

另一个相关问题涉及电影制作团队中的劳动分工。导演往往专注于精致完善的成品，将三维场景搭建委派给他人，而服装设计师、动画师和技术人员则注重色彩调色板或细粒度运动等细节。通过将多种创意职能集中起来

纳入单一的人工智能驱动工作流程，PrevizWhiz 这类系统可能会模糊传统的职责边界。这引发了一些问题：谁的专业技能会被取代？不同角色的贡献应如何归功？生成式预览是否有可能将不同部门的专业贡献过度简化或覆盖 [17, 19]。

因此，未来的工作不仅应改善不同模态之间的技术对齐，还应在生成式视频工作流程中嵌入透明度、归属标注和知情同意方面的保障措施。例如，PrevizWhiz 可以展示相似程度与创意偏离程度，让用户清楚了解输出内容何时偏离原始场景；也可以为外部视频输入提供来源标记，并协助开展协作迭代，避免歪曲创作意图。

6.6 局限与未来工作

我们的研究和原型存在若干局限。首先，当前的生成模型在可控性和连续性方面仍面临挑战。角色专属 LoRA 虽然提高了特写镜头的保真度，但在广角镜头或遮挡视图中往往会导致效果下降，而且扩展到多个角色会显著延长生成时间。训练此类 LoRA 还需要经过精心筛选的数据集和说明文字标注，而这些数据集和标注在不同风格和人口群体之间分布不均，带来了偏差风险。道具、灯光和服装的连续性也会中断，降低多镜头序列的连贯性；即使在更严格的遵循模式下，不同片段中的这些元素也常常发生变化。此外，生成式视频模型本身偶尔会在单个片段中引入不一致之处，例如添加意料之外的视觉元素，或改变不同帧之间的细节，从而进一步影响时间稳定性。当非角色物体对叙事或任务至关重要时（例如关键道具或场景内广告），可能还需要为这些物体训练专用 LoRA，以便在不同镜头中保持其身份和外观；然而，正如参与者所提到的，对于预可视化而言，这种保真度通常并非必需，但在实际制作场景中，物体外观必须保持一致，因此可能会变得重要。这些因素限制了多镜头序列的连贯性，也凸显了对更可靠的跨镜头保持机制的需求。随着人工智能和机器学习的发展（例如生成对抗网络和扩散模型），我们预计未来这些模型和计算能力将有所提升，从而能够生成更逼真的角色和更一致的服装 [33, 35, 58], ，并进一步改善年龄、性别和表情方面的表现，同时支持交互式定制面部特征、发型和衣着 [43]。未来的工作还可以通过渐进式预览或异步生成来降低延迟，以应对这些挑战。

其次，我们的观察来自时长相对较短的实验室环节（在两个场景中共 90–120 分钟），这限制了探索范围和对较长流程的观察。此外，模型延迟限制了快速分支探索：在我们的设置中，图像/视频生成通常每个视频片段需要 1 分钟，因此难以在一次环节中广泛进行参数扫描。我们没有纳入直接的基线比较，因为现有的预可视化或游戏引擎工具在目标、交互范式和保真度假设方面差异显著，使一对一比较难以解读，也容易受到建模等因素的混淆技能和资产的差异。相反，我们着重了解从业者如何使用一种融合了三维调度、生成式重风格化和运动条件控制的新型统一工作流程。这些发现有助于指导未来采用基线控制的研究，确定相似度控制、运动条件控制或跨镜头连续性等哪些组成部分对从业者最为重要，值得在比较评估和长期评估中单独考察。未来有必要在实地与创意团队开展更长期的部署，以了解人工智能驱动的预可视化在实践中如何与制作时间安排、预算结构以及不断演变的劳动分工相互作用。

7 结论

我们介绍了 PrevizWhiz，这是一个通过生成式人工智能将粗略的三维场景调度、细致的角色运动和视频风格化相结合的系统，旨在支持灵活、快速的预演。通过对电影制作人开展用户研究，我们发现，该系统支持轻量级场景设置、迭代优化，以及跨模态的表现性创作。研究结果表明，人工智能辅助的前期可视化能够拓展创作实践，降低独立创作者表达其创作意图的门槛。同时，延迟、一致性和被取代的担忧等问题，也凸显了未来进行审慎设计的必要性。

致谢

我们在开发 PrevizWhiz 的过程中使用了 Cursor[11] 。

参考文献

[1] Torin Anderson 和 Shuo Niu。2025。《制作人工智能增强视频：分析 YouTube 内容创作中的生成式人工智能应用案例》。载于《*CHI*计算系统中的人为因素会议扩展摘要集》（*CHI EA '25*）。计算机协会，美国纽约州纽约市，文章编号 388, 7 10 1145 3706599 3719991 页。doi: . / . Tom Bartindale、[2] Alia Sheikh、Nick Taylor、Peter Wright 和 Patrick Olivier。2012。《StoryCrate：面向现场电影制作的桌面故事板创作》。载于《*SIGCHI*计算系统中的人为因素会议论文集》（美国得克萨斯州奥斯汀）（*CHI '12*）。计算机协会，美国纽约州纽约市，169–178。doi:10.1145/2207676.2207700[3] Andreas Blattmann、Tim Dockhorn、Sumith Kulal、Daniel Mendelevitch、Maciej Kilian、Dominik Lorenz、Yam Levi、Zion English、Vikram Voleti、Adam Letts 等。2023。《Stable Video Diffusion：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）。[4] Bruce Block。2020。《视觉故事：构建电影、电视和数字媒体的视觉结构》。劳特利奇。[5] Virginia Braun 和 Victoria Clarke。2006。在心理学中运用主题分析。《心理学定性研究》3, 2（2006），77–101。[6] Shizhe Chen、Bei Liu、Jianlong Fu、Ruihua Song、Qin Jin、Pingping Lin、Xiaoyu Qi、Chunting Wang 和 Jin Zhou。2019。神经网络故事板生成器：用连贯图像序列呈现故事。载于《第*27*届*ACM*多媒体国际会议论文集》。2236–2244。[7] Joon Gi Chung、Soongi Hong、Junho Choi 和 Changhoon Oh。2025。了解在广播系统中部署基于人工智能的内容创作支持工具的动态——益处、挑战与发展方向。载于《*2025*年*CHI*计算系统中的人为因素会议论文集》（*CHI '25*）。计算机协会，美国纽约州纽约市，文章编号 27，共28页。doi:10.1145/3706598.3713532[8] Florinel-Alin Croitoru、Vlad Hondru、Radu Tudor Ionescu 和 Mubarak Shah。2023。视觉领域中的扩散模型：综述。《*IEEE*模式分析与机器智能汇刊》45, 9（2023），10850–10869。[9] Abhinav Dhall、Roland Goecke、Shreya Ghosh、Jyoti Joshi、Jesse Hoey 和 Tom Gedeon。2017。从个体到群体层面的情绪识别：EmotiW 5.0。载于《第*19*届*ACM*多模态交互国际会议论文集》。524–528。

[11]https://cursor.com/

[10] Hao-Shu Fang、Shuqin Xie、Yu-Wing Tai 和 Cewu Lu。2017。RMPE：区域多人姿态估计。载于 IEEE 国际计算机视觉会议论文集。2334–2343。[11] Dustin Freeman、Stephanie Santosa、Fanny Chevalier、Ravin Balakrishnan 和 Karan Singh。2014。LACES：通过合成与编辑流媒体视频进行实时创作。载于《SIGCHI计算系统中的人为因素会议论文集》（加拿大安大略省多伦多）（CHI '14）。计算机协会，美国纽约州纽约市，1207–1216。doi:10.1145/2556288. 2557304[12] Quentin Galvane、I-Sheng Lin、Fernando Argelaguet、Tsai-Yen Li 和 Marc Christie。2019。虚拟现实作为电影预可视化的内容创作工具。载于2019年IEEE虚拟现实与3D用户界面会议（VR）。303–311。doi:10.1109/VR.2019.8798181[13] Rong Gao、Xin Liu、Bohao Xing、Zitong Yu、Bjorn W Schuller 和 Heikki Kälviäi- nen。2024。通过理解微手势实现无身份人工情感智能。arXiv预印本 arXiv:2405.13206（2024）。[14] Dan B Goldman、Brian Curless、David Salesin 和 Steven M Seitz。2006。用于视频可视化与编辑的示意故事板创作。《ACM图形学汇刊》（TOG）25, 3（2006），862–871。[15] Chuan Guo、Shihao Zou、Xinxin Zuo、Sen Wang、Wei Ji、Xingyu Li 和 Li Cheng。2022。根据文本生成多样且自然的3D人体动作。载于IEEE/CVF计算机视觉与模式识别会议论文集。5152–5161。[16] Mieke Haesen、Jan Meskens、Kris Luyten 和 Karin Coninx。2009。利用故事板支持多学科团队和设计早期阶段。载于《人机交互国际会议》。施普林格，616–623。[17] Brett A. Halperin、Diana Flores Ruíz 和 Daniela K. Rosner。2025。地下AI？通过业余电影制作批判性探讨生成式电影。载于《2025年CHI计算系统中的人为因素会议论文集》（CHI '25）。计算机协会，美国纽约州纽约市，文章编号 1141, 18 10 1145 3706598 3713342 页。doi: . / . John Hart。《故事板艺术：电影制作人的入门指南》。劳特利奇。[19] Marina Hassapopoulou。2024。《互动电影：媒体参与的模糊伦理》。第63卷。明尼苏达大学出版社。[20] Jonathan Heagerty、Sida Li、Eric Lee、Shuvra Bhattacharyya、Sujal Bista、Barbara Brawn、Brandon Y. Feng、Susmija Jabbireddy、Joseph JaJa、Hernisa Kacorri、David Li、Derek Yarnell、Matthias Zwicker 和 Amitabh Varshney。2024。HoloCamera：面向电影级虚拟现实应用的先进体积捕捉技术。《IEEE 可视化与计算机图形学汇刊》30, 5（2024），2767–2775。doi:10.1109/TVCG.2024.3372123[21] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等。2022。LoRA：大语言模型的低秩适应。ICLR 1, 2（2022），3。[22] Phillip Isola、Jun-Yan Zhu、Tinghui Zhou 和 Alexei A Efros。2017。利用条件对抗网络进行图像到图像的转换。载于 IEEE计算机视觉与模式识别会议论文集。1125–1134。[23] Alexander Ivanov、David Ledo、Tovi Grossman、George Fitzmaurice 和 Fraser Anderson。2022。MoodCubes：用于收集、发现和构想灵感素材的沉浸式空间。载于《2022年ACM交互系统设计会议论文集》（线上活动，澳大利亚）（DIS '22）。计算机协会，美国纽约州纽约市，189–203。doi:10.1145/3532106. 3533565[24] Harry H. Jiang、Lauren Brown、Jessica Cheng、Mehtab Khan、Abhishek Gupta、Deja Workman、Alex Hanna、Johnathan Flowers 和 Timnit Gebru。2023。人工智能艺术及其对艺术家的影响。载于《2023年AAAI/ACM人工智能、伦理与社会会议论文集》（加拿大魁北克省蒙特利尔）（AIES '23）。计算机协会，美国纽约州纽约市，363 374 10 1145 3600211 3604681 – 。doi: . / .[25] Zeyinzi Jiang、Zhen Han、Chaojie Mao、Jingfeng Zhang、Yulin Pan 和 Yu Liu。2025。VACE：一体化视频创作与编辑。arXiv预印本 arXiv:2503.07598（2025）。[26] Hye-Young Jo、Ryo Suzuki 和 Yoonji Kim。2024。CollageVis：利用视频拼贴进行独立电影制作的快速预可视化工具。载于《2024年CHI计算系统中的人为因素会议论文集》（美国夏威夷州檀香山）（CHI '24）。计算机协会，美国纽约州纽约市，文章编号 164, 16 10 1145 3613904 3642575 页。doi: . / . Tero Karras、Samuli Laine 和 Timo Aila。2019。一种用于生成对抗网络的基于风格的生成器架构。载于IEEE/CVF计算机视觉与模式识别会议论文集。4401–4410。[28] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C Berg、Wan-Yen Lo 等。2023。任意分割。载于 IEEE/CVF国际计算机视觉会议论文集。4015–4026。[29] Vladimir Kulikov、Matan Kleiner、Inbar Huberman-Spiegelglas 和 Tomer Michaeli。2024。FlowEdit：使用预训练流模型进行无反演文本编辑。arXiv预印本 arXiv:2412.08629（2024）。[30] Black Forest Labs。2024。FLUX。https://github.com/black-forest-labs/flux。[31] Black Forest Labs、Stephen Batifol、Andreas Blattmann、Frederic Boesel、Saksham Consul、Cyril Diagne、Tim Dockhorn、Jack English、Zion English、Patrick Esser、Sumith Kulal、Kyle Lacey、Yam Levi、Cheng Li、Dominik Lorenz、Jonas Müller、Dustin Podell、Robin Rombach、Harry Saini、Axel Sauer 和 Luke Smith。2025。FLUX.1 Kontext：用于潜在空间中上下文图像生成与编辑的流匹配。arXiv: 2506 15742 [ ] 2506 15742 。cs.GR [32] 2025 https://arxiv.org/abs/ 。David Ledo。 。生成式转描：关于生成式视频到视频实践的第一人称自传式探索。载于《2025年创造力与认知会议论文集》（C&C '25）。计算机协会，美国纽约州纽约市，931 948 10 1145 3698061 3726926 – 。doi: 。/ 。[33] 2019 Henry Melki、Ian Montgomery 和 Greg Maguire。 。关于生成可信照片级真实感电影角色的创作过程的研究。博士论文。阿尔斯特大学。[34] Thomas Muender、Thomas Fröhlich 和 Rainer Malaka。2018。赋能创意人士：用于预演的虚拟现实。载于《2018年CHI计算机系统中的人为因素会议扩展摘要》。1–6。[35] G Pai nguzhali、Viswanath Ananth、D Kavitha、R Sujatha 等。2025。人工智能与生成式人工智能如何改变时尚产业。载于《面向服务业商业分析与战略决策的生成式人工智能》。IGI Global Scientific Publishing，281–316。[36] Arjama Pal、Sakalya Mitra 和 D Lakshmi。2025。借助灯光、摄影与人工智能，照亮从剧本到银幕的道路。载于《以人工智能变革电影》。IGI Global Scientific Publishing，97–142。[37] Mathis Petrovich、Michael J Black 和 Gül Varol。2023。TMR：使用对比式3D人体动作合成进行文本到动作检索。载于《IEEE/CVF国际计算机视觉会议论文集》。9488–9497。[38] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。2021。零样本文本到图像生成。载于《国际机器学习会议》。PMLR，8821–8831。[39] Anyi Rao、Jean-Peïc Chou 和 Maneesh Agrawala。2024。ScriptViz：基于大型电影数据库辅助剧本创作的可视化工具。载于《第37届ACM用户界面软件与技术研讨会年会论文集》（美国宾夕法尼亚州匹兹堡）（UIST '24）。计算机协会，美国纽约州纽约市，文章21，13页。doi:10.1145/3654777.3676402[40] Anyi Rao、Xuekun Jiang、Yuwei Guo、Linning Xu、Lei Yang、Libiao Jin、Dahua Lin 和 Bo Dai。2023。用于视频制作的基于引擎的虚拟环境中的动态故事板生成。载于《ACM SIGGRAPH 2023 海报》。1–2。[41] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022。使用潜在扩散模型进行高分辨率图像合成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。10684–10695。[42] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily L Denton、Kamyar Ghasemipour、Raphael Gontijo Lopes、Burcu Karagol Ayan、Tim Salimans 等。2022。结合深度语言理解的照片级真实感文生图扩散模型。《神经信息处理系统进展》35（2022），36479–36494。[43] Tassneam M Samy、Beshoy I Asham、Salwa O Slim 和 Amr A Abohany。2025。借助 FITMI 革新在线购物：一种逼真的虚拟试穿解决方案。《神经计算与应用》37，8（2025），6125–6144。[44] Yang Shi、Nan Cao、Xiaojuan Ma、Siji Chen 和 Pei Liu。2020。EmoG：支持故事板创作中的情绪表情草绘。载于《2020年CHI计算机系统中的人为因素会议论文集》（美国夏威夷州檀香山）（CHI '20）。计算机协会，美国纽约州纽约市，1–12。doi:10.1145/3313831.3376520[45] Evgeny Stemasov、David Ledo、George Fitzmaurice 和 Fraser Anderson。2023。沉浸式采样：探索沉浸式媒体丰富空间中的采样，以开拓未来创作实践。载于《2023年ACM交互系统设计会议论文集》（美国宾夕法尼亚州匹兹堡）（DIS '23）。计算机协会，美国纽约州纽约市，212 229 10 1145 3563657 3596131 – 。doi: 。/ 。[46] 2021 Yu Sun、Qian Bao、Wu Liu、Yili Fu、Michael J Black 和 Tao Mei。 。单目单阶段多个人体回归。载于《IEEE/CVF国际计算机视觉会议论文集》。11179–11188。[47] 腾讯混元3D团队。2025。混元3D 2.5：以极致细节实现高保真3D资产生成。arXiv:2506.16504 [cs.CV] https://arxiv.org/abs/ 2506 16504 。Steve Thomas。 。纪录片制作中的协作与伦理——案例研究。《电影与电视研究新评论》[48] 2012 10，3（2012），332–343。[49] Wan团队、Ang Wang、Baole Ai、Bin Wen、Chaojie Mao、Chen-Wei Xie、Di Chen、Feiwu Yu、Haiming Zhao、Jianxiao Yang 等。2025。Wan：开放且先进的大规模视频生成模型。arXiv预印本 arXiv:2503.20314（2025）。[50] Wan团队、Ang Wang、Baole Ai、Bin Wen、Chaojie Mao、Chen-Wei Xie、Di Chen、Feiwu Yu、Haiming Zhao、Jianxiao Yang、Jianyuan Zeng、Jiayu Wang、Jingfeng Zhang、Jingren Zhou、Jinkai Wang、Jixuan Chen、Kai Zhu、Kang Zhao、Keyu Yan、Lianghua Huang、Mengyang Feng、Ningyi Zhang、Pandeng Li、Pingyu Wu、Ruihang Chu、Ruili Feng、Shiwei Zhang、Siyang Sun、Tao Fang、Tianxing Wang、Tianyi Gui、Tingyu Weng、Tong Shen、Wei Lin、Wei Wang、Wei Wang、Wenmeng Zhou、Wente Wang、Wenting Shen、Wenyuan Yu、Xianzhong Shi、Xiaoming Huang、Xin Xu、Yan Kou、Yangyu Lv、Yifei Li、Yijing Liu、Yiming Wang、Yingya Zhang、Yitong Huang、Yong Li、You Wu、Yu Liu、Yulin Pan、Yun Zheng、Yuntao Hong、Yupeng Shi、Yutong Feng、Zeyinzi Jiang、Zhen Han、Zhi-Fan Wu 和 Ziyu Liu。2025。Wan：开放且先进的大规模视频生成模型。arXiv

*arXiv*预印本：*2503.20314*（2025）。[51] Cheng Yao Wang、Qian Zhou、George Fitzmaurice 和 Fraser Anderson。2022。VideoPoseVR：利用在线视频创作虚拟现实角色动画。*Proc. ACM Hum.-Comput. Interact.* 6，ISS，文章575（2022年11月），20页。doi:10.1145/3567728[52] Zheng Wei、Jia Sun、Junxiang Liao、Lik-Hang Lee、Chan In Sio、Pan Hui、Huamin Qu、Wai Tong 和 Xian Xu。2025。照亮场景：虚拟环境与学习模式如何影响虚拟现实中的电影灯光掌握。*IEEE* 可视化与计算机图形学汇刊（2025）。[53] Zheng Wei、Hongtao Wu、Lvmin Zhang、Xian Xu、Yefeng Zheng、Pan Hui、Maneesh Agrawala、Huamin Qu 和 Anyi Rao。2025。CineVision：面向导演与电影摄影师协作的交互式预可视化故事板系统。载于《第*38*届*ACM*用户界面软件与技术研讨会年会论文集》（*UIST '25*）。ACM，釜山，大韩民国，1–18。doi:10.1145/3746059.3747793[54] Hock Hian Wong。2012。预演：协助电影制作人实现创作愿景。载于*SIGGRAPH Asia 2012* 课程（新加坡，新加坡）（*SA '12*）。计算机协会，美国纽约州纽约市，文章9，20页。doi:10.1145/2407783.2407792[55] Bin Xiao、Haiping Wu、Weijian Xu、Xiyang Dai、Houdong Hu、Yumao Lu、Michael Zeng、Ce Liu 和 Lu Yuan。2024。Florence-2：为多种视觉任务推进统一表征。载于《*IEEE/CVF*计算机视觉与模式识别会议论文集》。4818–4829。[56] Jingyuan Yang、Jiawei Feng 和 Hui Huang。2024。Emogen：利用文生图扩散模型生成情感图像内容。载于《*IEEE/CVF*计算机视觉与模式识别会议论文集》。6358–6368。[57] Lvmin Zhang、Anyi Rao 和 Maneesh Agrawala。2023。为文生图扩散模型添加条件控制。[58] Yanbo Zhang 和 Chuanlan Liu。2024。释放人工智能在时尚设计与电子商务应用中的潜力：以 Midjourney 为例。《理论与应用电子商务研究期刊》19，1（2024），654–670。[59] Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A Efros。2017。使用循环一致性对抗网络进行无配对图像到图像转换。载于《*IEEE*国际计算机视觉会议论文集》。2223–2232。

A 附录

为补充正文中展示的场景，我们提供了使用 PrevizWhiz 生成的更多示例输出，以说明该工作流程可推广至预演中的更多镜头类型（小节 A.1）、电影类型和视觉风格（小节 A.2）。这些示例采用第 3–4 节所述的相同功能创建（三维调度、基于相似性的风格重塑，以及可选的视频驱动运动），并不代表额外的案例研究。我们提供这些示例，以展示系统丰富的表现范围。

A.1 镜头类型多样性

为展示其对多种摄影机运用方式的适用性，图 11 和图 12 展示了电影制作中常用的不同镜头类型的输出。这些图中示例的视频见补充材料。

例如，示例1（图11）展示了 PrevizWhiz 如何在单个场景中支持多种标准电影制作镜头和摄影机运动。序列以一个建立镜头（a）开场，画面中男子正走向桌子，并通过宽幅构图来界定空间和角色布置。接着转入跟踪推进镜头（b），摄影机向上移动，在男子扣上外套时展现他的面部。随后，中景（c）从正面呈现角色，同时摄影机略微向后移动，展现更多背景细节并保持空间连续性。最后，反应特写镜头（d）捕捉男子的表情，强调情感基调和角色焦点。这些阶段共同展示了 PrevizWhiz 如何处理不同的镜头景别、角色运动和摄像机路径，从而支持电影预演中常见的动态、叙事驱动型镜头呈现。

A.2 类型与风格多样性

我们的示例场景展现了电影感写实风格以及多种视觉风格（例如三维卡通、动漫和电影感风格），能够呈现角色的动作与互动。例如，黑客场景示例（图 12）体现了新黑色电影和科技惊悚片的惯例：昏暗的室内环境、来自屏幕的定向光、高对比度的紫色与蓝色，以及具有监视、间谍活动或网络犯罪叙事典型特征的紧张双人角色调度。相比之下，神秘男子示例（图 11）采用侦探或犯罪剧情片的视觉语言，呈现出低饱和度的工业色彩、风衣轮廓、富有情绪感的楼梯间照明，以及突出情感模糊性的电影式特写。除了从外部视频转换而来的打字、行走等细微动作，PrevizWhiz 还能够重现高能量动作，例如示例 3 中的武术打斗场面，并采用动漫风格（图 13）。这些示例表明，借助多模态图像到图像和图像到视频模型，即使构图仍以三维框架为基础，也能实现符合特定类型的电影摄影特质。

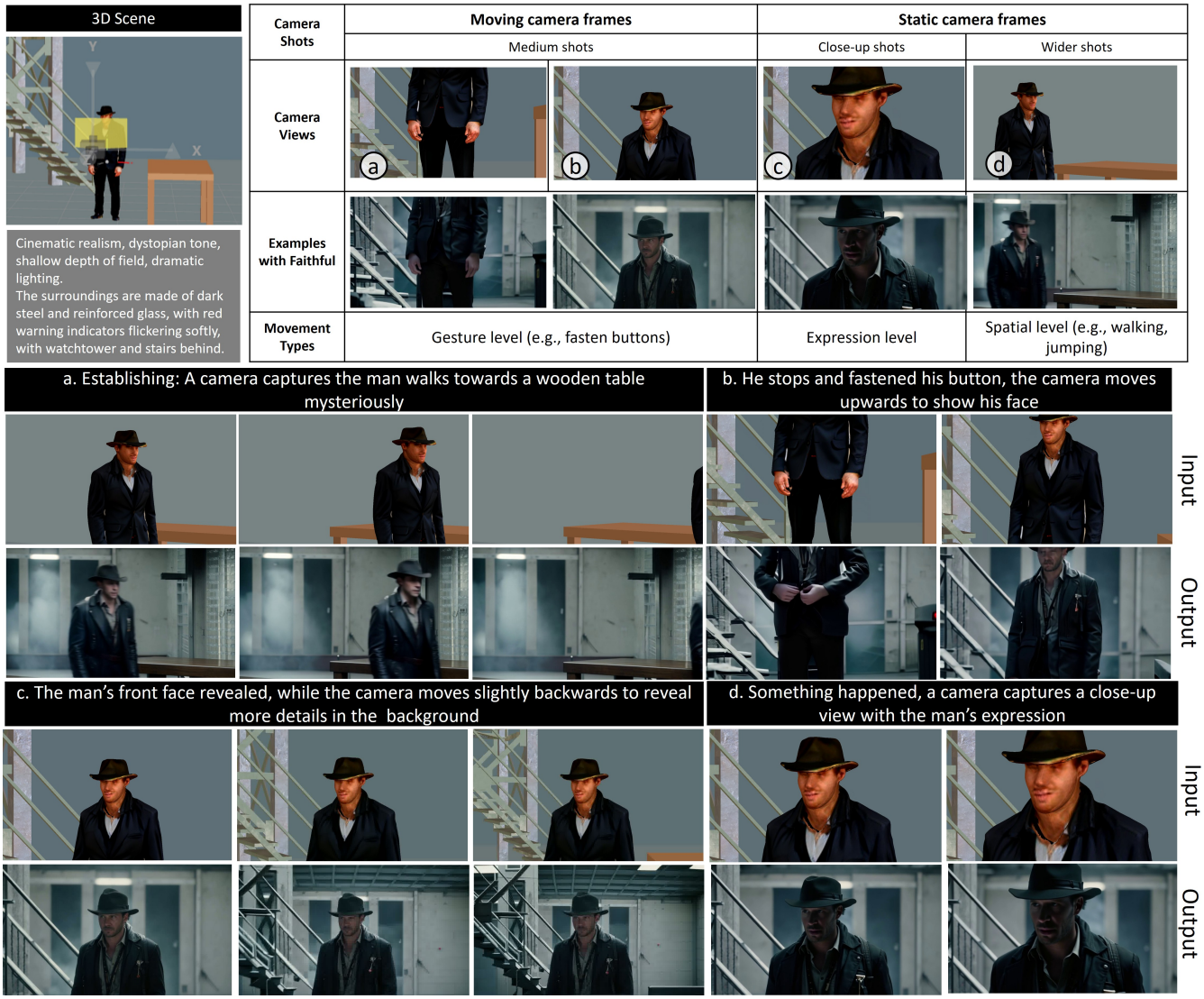

图 11：“神秘男子登场”示例的演示。本图展示了系统如何将一个简单的叙事时刻组织成两个具有不同镜头运动类型的镜头。镜头 1：(a–c) 侧重于呈现空间环境并揭示人物身份。(a) 建立镜头：摄影机观察男子走向一张木桌，引入环境并展现他神秘的存在。(b) 向上揭示：当他停下并扣好外套纽扣时，摄影机向上摇摄，逐渐展现他的面部。(c) 向后漂移：男子正面全貌进入画面时，摄影机轻轻向后移动，以呈现更多工业背景。镜头 2：(d) 转为特写镜头运动类型。(d) 情绪特写：在一个隐含的叙事节拍之后，摄影机切入特写，捕捉男子的表情，突出张力和人物情绪。

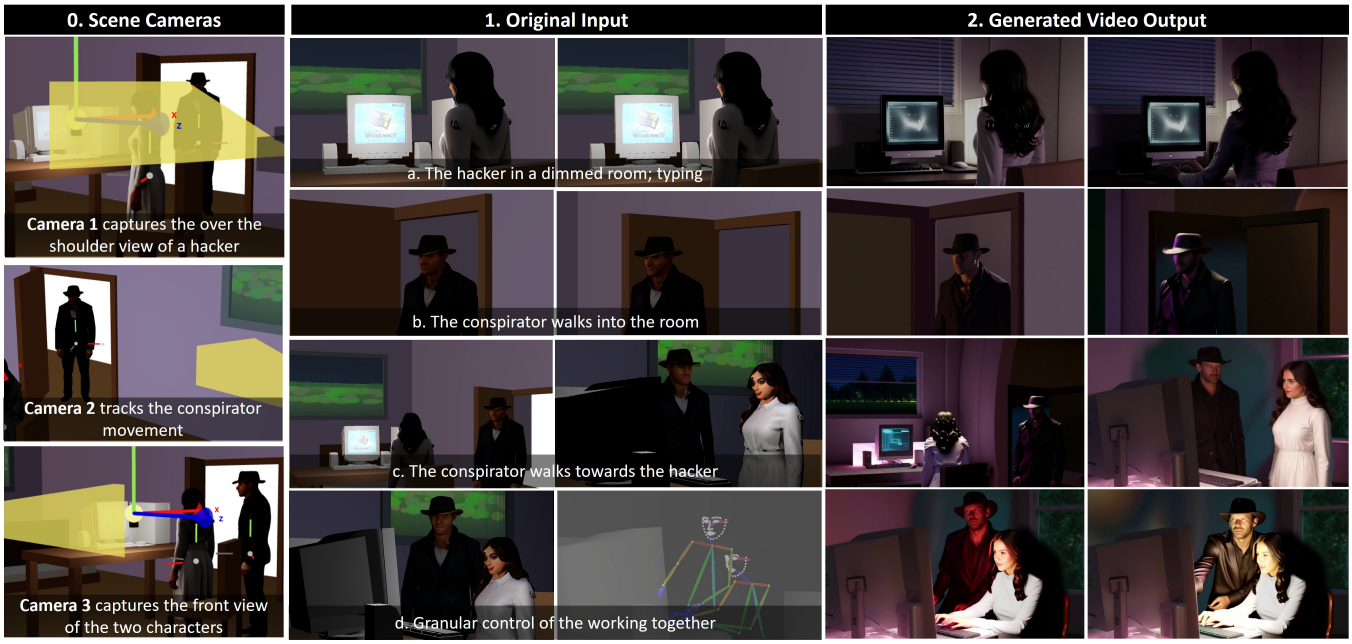

图 12：示例 2：“黑客场景”的原始输入 (1) 与生成输出 (2)。(a) 黑客独自在一间光线昏暗的房间里工作，主要由电脑屏幕的光照亮。她打字时，系统的三维调度面板定义了房间布局、灯光和角色布置。(b) 共谋者进入房间。他出现在门口并走进房间时，系统会更新灯光、遮挡和角色调度。(c) 共谋者走近黑客。随着角色之间的距离缩短，系统保持空间连续性和镜头连贯性，展现动作调度、灯光和视角调整如何在整个场景中支撑逐步营造张力的角色动作。

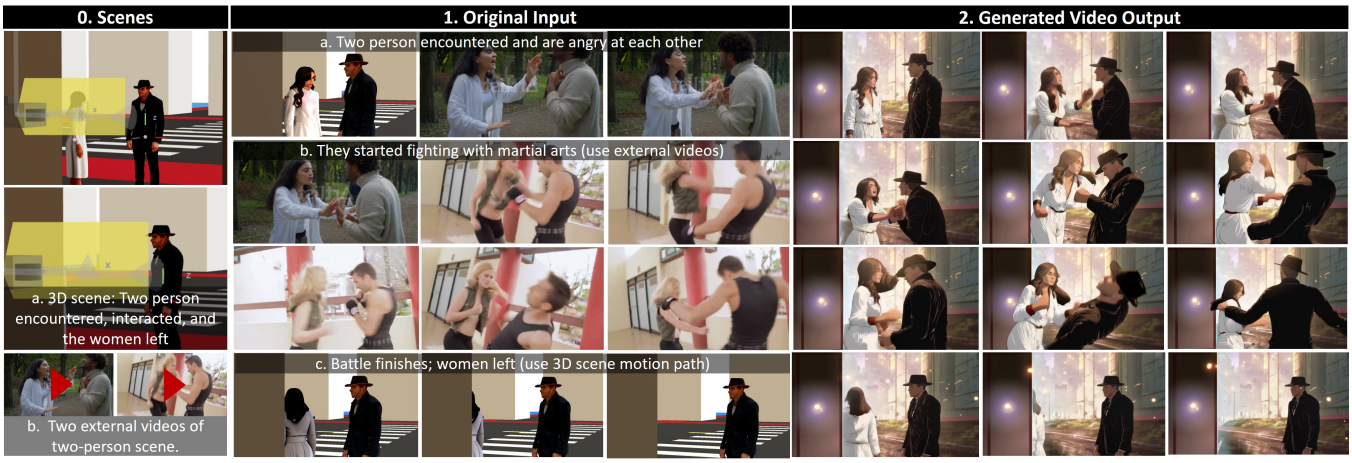

图 13：示例 3：使用 PrevizWhiz 生成的（户外）“打斗场景”：(1) 原始输入与 (2) 生成输出：(a) 两个角色在预设三维场景中对峙，营造张力并确立初始调度。(b) 将外部参考视频中的动作应用到角色身上，生成武术打斗序列，呈现充满活力、高强度的互动。(c) 打斗结束后，女性角色沿着三维定义的运动路径离开场景，展现引导动作与场景布局之间的流畅整合。

参与者ID	Age	性别	Role	从业年限	工作期间使用人工智能	描述：人工智能的使用	预可视化经验
1	24	女性	导演/编剧，独立影视作品	5	视频生成	将人工智能生成的视频用作角色的梦境/幻想片段	镜头设计和故事板用于拆解
2	41	Male	3D/动画专家，电影制作行业中	12	图像生成	类似 MidJourney 的图像生成	故事板
3	30	女性	广告导演（情节与纪录片）	10	N/A	N/A	指导工作人员进行测试，使用故事板或3D预可视化
4	30	Male	电影专业学生	1	图像/视频生成	为海报设计扩展图像；使用视频生成制作风景镜头	Cine Tracer
5	42	女性	造型与服装在电影制作中	15	N/A	N/A	N/A
6	32	Male	独立电影技术总监/电影摄影师/短片内容创作者	10	图像/视频生成	使用图像模型作为参考；使用视频模型（Runway、Veo、Kling）为独立/商业视频制作视频素材	Previs Pro、Cine Tracer，Shot Designer
7	29	非二元性别	艺术部门管理员 / 艺术部门制片助理/道具	6	图像生成	使用 ComfyUI 创建图像生成系统	拥有绘制故事板的经验以及制作动态分镜。手绘和使用 Harmony。
8	29	Male	电影摄影师，负责剧情片、纪录片，社交媒体内容创作者	7	图像/视频生成	在 MidJourney 上生成图像，尤其是出于以下意图为电影和剧集制作提案演示文稿；使用 Runway、Veo 2 进行视频创作（使用不多）。	Set a Light 3D、Cine Tracer
9	30	女性	品牌广告导演、短片	7	图像/视频生成	参与一个项目，旨在使用人工智能创作短剧。	聘请分镜师绘制故事板；或使用动作捕捉演员和用于制作预览的绿幕。
10	23	女性	技术绑定师和动画师	2	N/A	N/A	三维预演/布局，三维动画制作流程，广告、长篇电影

表 1：参与者人口统计信息