FairyGen：来自单一儿童绘制角色的故事卡通视频

JIAYI ZHENG and XIAODONG CUN*, GVC Lab, Great Bay University, China

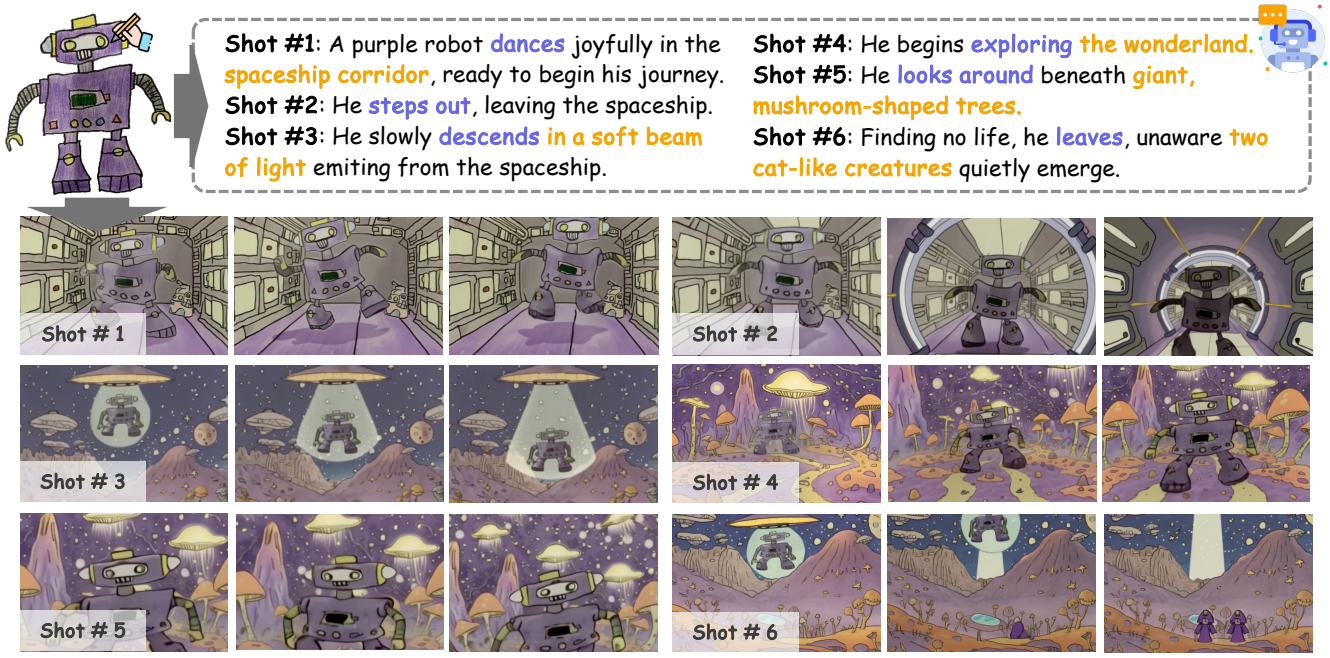

Fig. 1. **我们提出了 FairyGen，一个可视故事生成框架，可以从单个儿童绘制的角色生成多帧卡通视频，并且前景和背景之间的样式和动作保持一致。项目页面**：https://jayleejia.github.io/FairyGen/。

我们提出了一种自动系统**精灵生成器**，可以从单一儿童绘画中生成以故事驱动的卡通视频，同时忠实保留其独特的艺术风格。与以前主要关注角色一致性和基本动作的故事讲述方法不同，FairyGen 明确将角色建模与风格化背景生成分离，并融入电影镜头设计，支持表现力强且连贯的故事叙述。给定一个单独的角色草图，我们首先使用多模态大型语言模型 (MLLM) 生成包含镜头级描述的结构化的分镜脚本，这些描述指定了环境设置、角色动作和摄像机视角。为了确保视觉一致性，我们引入了一个风格传播适配器，捕捉角色的视觉风格并将其应用到背景中，在合成一致风格场景的同时忠实保留角色完整的视觉身份。一个射孔设计模块通过基于分镜脚本的帧裁剪和多视角合成为进一步增强视觉多样性和电影质量提供了支持。为了给故事添加动画，我们重构了一个字符的 3D 代理来导出物理上合理的动作序列，并使用这些序列对基于 MMDiT 的图像到视频扩散模型进行微调。我们还提出了一种两阶段的动作定制适配器：第一阶段从时间无序帧中学习外观特征，分离身份和运动；第二阶段利用冻结了身份权重的时间步移策略模型来建模时间动态性。训练完成后，FairyGen 直接渲染与分镜脚本一致的各种连贯视频场景。大量实验表明，我们的系统生成动画在风格上忠实、叙事结构化，并具有流畅自然的动作，突显了其在个性化和引人入胜的故事动画方面的潜力。

1 介绍

儿童经常通过抽象、风格化的绘画来表达生动的想象力，这些绘画以简单的卡通人物和富有表现力的视觉元素为特征。尽管缺乏照片级细节，这些插图传达了独特的艺术风格和情感意图。将这样的绘画转化为连贯的动画故事桥接了年轻创造力与富有表现力的故事

*Corresponding Author

讲述，提供了在教育、数字艺术疗法、个性化内容创作和互动娱乐中的潜在应用。故事情节可视化已成为计算机图形学的主要研究领域，最近生成视频模型的进步不断拓展其视野。此前的工作如 StoryGAN [Li et al. 2019], AR-LDM [Pan et al. 2024a] 和 Make-A-Story [Rahman et al. 2023] 提高了视觉真实度和语义连贯性，但受到风格多样性和训练数据限制的局限。更近期的由大型语言模型驱动的工作流如 TaleCrafter [Gong et al. 2023], DreamStory [He et al. 2024] 和 Animate-A-Story [He et al. 2023] 引入了模块化任务分解以获得更好的可控性，但往往受到角色不一致、叙事片段化和动作质量差的影响。基于扩散的视频模型，如 MEVG [Oh et al. 2024]、MovieDreamer [Zhao et al. 2024] 和 Vlogger [Zhuang et al. 2024] 改善了时间一致性及叙事流畅性，但仍然在跨镜头角色一致性、艺术风格保留以及复杂运动合成方面存在问题——这主要是由于依赖于现实世界数据的先验知识。这些限制在处理抽象的手绘角色时更加明显，特别是在单例场景中，输入样式与训练数据显著不同。为了解决这一问题，我们考虑采用分离设计生成卡通故事，明确将前景角色从背景合成中分开。使用符合物理约束的 3D 重建，我们在保持角色身份的同时使可能的动作生成得以实现。相应地，背景生成被处理为一种风格适应过程，将角色的视觉样式传播到场景元素 [Brooks and Efros 2022; Kulal et al. 2023; Pan et al. 2024b]。这种分离设计进一步扩展到了视频生成，首先从 3D 衍生序列中学习动作先验知识，然后使用大型预训练视频扩散模型对整个场景进行动画处理。此流程自然地保留了角色一致性，支持复杂运动，并实现了电影般的叙事方式。在此基础上，我们提出了 FairyGen，一个用于根据单个手绘角色生成动画故事视频的新框架。FairyGen 能够产生表现力的动作、风格一致的背景和电影级构图，而无需额外的训练数据。具体来说，给定一个手绘角色，首先使用 MLLM 生成描述动作、场景设置及镜头构图的结构化剧本。为了确保视觉一致性，我们提出了一种样式传播适配器，该适配器在学习其风格特征的同时保留了角色完整的视觉身份，并通过预训练的修复扩散模型将其传播到背景中。接下来，为使故事动画化，重建一个角色的 3D 代理并推导出物理上可能的动作序列，然后用于微调基于 MMDiT 的图像到视频扩散模型 [Wan et al. 2025]。为了实现稳健的运动合成，我们引入了两阶段动作定制适配器：第一阶段从时间打乱的帧中学习空间特征以消除时间偏差；第二阶段通过新颖的时间步长偏移策略（保持身份权重不变）捕获动态特性，确保动画平滑且连贯。广泛的实验表明，FairyGen 有效地生成个性化动画故事，这些故事在风格上一致、叙述连贯，并富含自然运动。总结来说，我们的主要贡献如下：

- 我们提出了一种新颖的故事视频生成框架，该框架能够从单一的儿童绘制角色图像中合成风格一致、叙事连贯且时间平滑的动画。
- 我们提出了一种新颖的样式传播适配器，该适配器从角色插图中学习，并生成与之兼容样式的背景场景，同时保留特定于角色的视觉和语义特征。
- 我们证明，在图像到视频的生成中移动扩散时间步显著增强了模型学习自然流畅运动的能力。

2 相关工作

故事生成。从文本描述生成视觉故事包含许多具有挑战性的问题，包括视频整体风格的一致性和主题的一致性，等.。早期的工作直接基于一个精心准备的小数据集利用基于 GAN 的框架 [Li 2022; Li et al. 2019] 或变换器 [Chen et al. 2022; Maharana et al. 2022] 来直接生成故事视频。最近，图像和视频扩散模型 [Ho et al. 2020; Rombach et al. 2022a] 以及大型语言模型 [DeepSeek 2025; OpenAI 2022; Qwen 2025] 为规划、生成和动画问题带来了通用的生成先验。SEED-Story [Yang et al. 2024a] 直接训练了大型语言模型以实现故事的一致性和长篇可视化。至于基于扩散的方法，TaleCrafter [Gong et al. 2023] 提出了一种多阶段框架，从定制的文本到图像模型生成视觉故事，并结合传统的相机移动，然后通过 AutoStory [Wang et al. 2024] 扩展了更多的条件。StoryDiffusion [Zhou et al. 2024b] 提出了一个定制

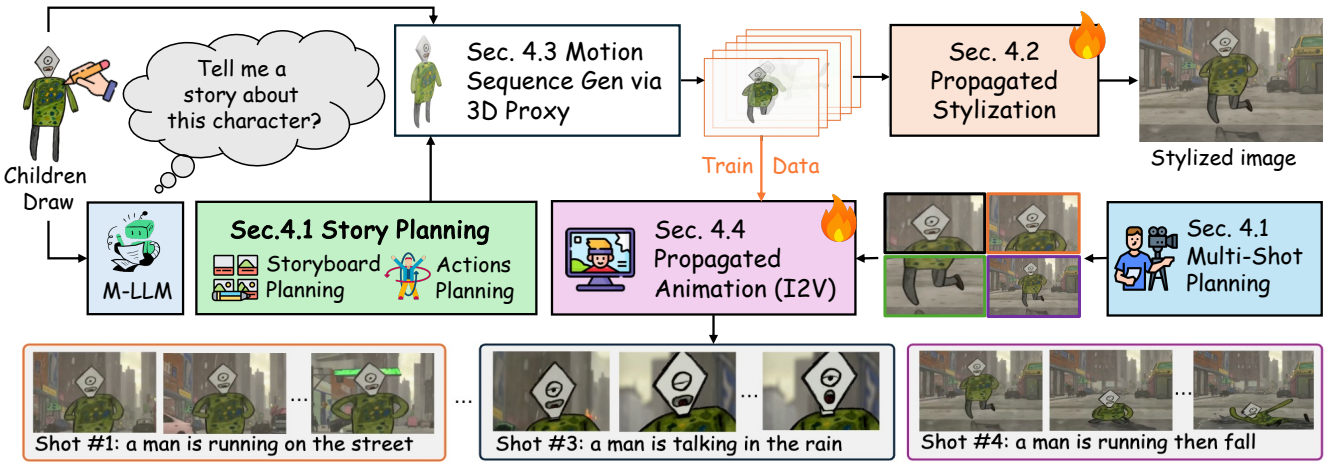

Fig. 2. **整个 *FairyGen* 的管道流程.**

的多帧注意力层以实现一致的身份生成，类似于 [Liu et al. 2024c]。故事代理 [Hu et al. 2024] 涉及将大型语言模型作为讲故事的代理的多个阶段，等. 。然而，所有这些方法 [Su et al. 2023; Tao et al. 2024; Zhang et al. 2025b; Zhu and Tang 2025] 主要集中在为角色创建自定义身份管道上，然后可以通过预训练的视频扩散模型生成视频。我们认为这种流水线仍然难以定制与相似风格对应的儿童绘画图像，并且直接利用生成先验来产生合适的动作是困难的。不同的是，为了避免前景角色的复杂运动，我们引入了一个 3D 代理来解决复杂的运动和角色一致性问题。

定制生成。在大型生成模型的时代，对文本到图像/视频的模型 [Chen et al. 2023, 2024; Kong et al. 2024; Wan et al. 2025; Yang et al. 2024b] 进行微调以实现定制用途是自然且实用的方法。我们的方法也与定制方法有着类似的灵感，因为我们需要生成定制化的背景风格和动作。对于主题定制而言，早期的工作 [Gal et al. 2022; Hu et al. 2021; Ruiz et al. 2023] 主要集中在通过额外的参数高效训练来实现单个主题的定制，还包括多个主题 [Kumari et al. 2023; Yuan et al. 2023]，以及视频 [Jiang et al. 2024]。至于同时对主题和动作进行定制。DreamVideo [Wei et al. 2024] , CustomTTT [Bi et al. 2024] , MotionDirector [Zhao et al. 2025] 训练不同的 LoRAs [Hu et al. 2021] 以实现外观和动作的定制。DynamicConcept [Abdal et al. 2025] 训练两阶段的 LoRAs [Hu et al. 2021] 以生成自定义动作。风格定制与主体定制背后的理念相似。例如，B-LoRA [Frenkel et al. 2025] 通过分层特定学习来分离主体和风格。StyleDrop [Sohn et al. 2023] 通过迭代训练增加风格化效果。然而，如何在给定前景角色的儿童绘画背景生成以及图像到视频框架中利用这些方法仍不清楚。

3 预备知识

潜在扩散模型。当前大多数生成模型基于潜在扩散模型 [Rombach et al. 2022b]。以图像扩散模型为例，它包含一个预训练的 VAE 来编码/解码图像到潜在空间。然后，扩散模型旨在训练一个去噪网络通过简单的 MSE 损失去除添加的单步噪声。形式上，给定图像 I 及其对应的潜码 $z = \mathcal{E}(I)$，我们首先向潜码 z 添加 t 步的噪声 ϵ，其中我们在 z_t 处训练一个去噪网络来通过以下方式去除添加的噪声：

$$\mathcal{L} = \|\epsilon - \epsilon_\theta(z_t, c, t)\|_2, \tag{1}$$

其中 c 是条件信号，通常是预训练文本编码器 [Radford et al. 2021; Raffel et al. 2020] 的文本特征。训练后，可以通过多步采样过程和 VAE 解码从噪声生成图像。

***LoRA* 用于定制**。LoRA [Hu et al. 2021] 首次提出用于高效训练语言模型。给定任何预训练模型的权重 $W \in R^{m\times n}$，LoRA 仅更新两个低秩矩阵 $A \in R^{m\times l}$ 和 $B \in R^{l\times n}$ 的参数，其中 $l < m$ 和 $l < n$，以高效地将已训练的知识适应到学习领域中，这可以定义为：$y = Wx + ABx$，其中 x 和 y 分别是输入向量和新的加权向量。在扩散模型中，LoRA 是一种常用的技术，用于通过定制数据集调整 ϵ_θ 的权重。

4 方法

给定一个单个儿童手绘字符图像 I，背景为空白，我们的目标是生成一个完全风格化的长卡通视频 V，该视频作为由多个不同镜头组成的故事连续展开。生成的视频应保持角色一致性的同时，实现复杂的动作、连贯的场景和电影叙事。如图 2 所示，我们提出一个多阶段管道来实现这一目标。首先，我们采用一种 MLLM [OpenAI 2022] 从给定的角色草稿中推断出结构化的故事板，构成时间和空间镜头规划的基础（第 4.1 节）。接下来，在输入风格图像的基础上，我们使用定制的样式传播模块合成风格一致的背景，将角色的美学特征传递到周围环境中，引入节奏感和多样化的时空线索对于电影叙事至关重要 (第 4.2 节)。然后，我们从 2D 角色图像重构一个 3D 代理，并通过绑定和重新定向导出物理上合理的运动序列（第 4.3 节）。这些运动序列随后被用作训练数据以微调基于 MMDiT 的图像到视频扩散模型网络，在此过程中，我们利用两阶段 LoRA 训练方案来分离空间身份和时间动作特征（第 4.4 节）。总体而言，该流程保留了角色身份，提升了运动保真度，并增强了叙事紧张感。我们在下一节中详细讨论每一部分。

4.1 故事和镜头规划来自单一角色

为了从单个人物草稿实现叙事驱动的视频合成，我们引入了一个故事和镜头规划模块，将叙述分解为电影镜头描述。与以前专注于低级帧插值的工作不同，我们的方法定义了一个清晰的故事板来指导动作合成、镜头构图和叙事节奏。通过将动画扎根于故事板中，我们确保了与讲故事惯例一致的连贯空间构图和时间进程。

如图 3 所示，我们的系统始于 故事板规划，通过分层结构组织叙述：全局叙事概览和详细的镜头级故事板。全局叙事概述了角色的出场、背景环境以及主要事件的高层次抽象。此外，镜头级故事板详细描述了每个镜头的背景、角色动作和摄像机配置（例如.、镜头类型、视角、焦点区域）。为了操作化故事板的组件，我们引入了两个模块：行动规划和多镜头规划。在行动规划阶段，使用 LLM 提取与行动相关的关键词，并用于从 3D 动画平台检索匹配的动作。这些动作随后通过绑定和重定向适应输入的角色。在多镜头规划阶段，镜头类型和焦点区域描述指导生成边界框（通过 LLM），以裁剪合成背景。同时，为了确保多视角一致性，使用 3D 代理渲染不同的角色视角（参见第 4.3 节），并通过视图条件综合应用确保背景的连贯生成（参见第 4.2 节）。

4.2 风格一致的场景生成来自角色

一个没有背景的角色图像不足以表达故事情节。为了在视觉上支持叙事，我们旨在生成既与故事线情境相符又与前景角色手绘风格一致的场景。这种一致性在卡通故事视频中尤为重要，因为视觉统一性可以增强连续性、沉浸感和情感共鸣。一个关键挑战是确保生成的背景忠实反映原始手绘角色的艺术风格，如笔触纹理、色彩搭配和线条密度。与传统的将样式从参考背景转移过来的方法不同，我们的方法是从角色向背景传播样式，要求背景继承前景的视觉属性。

为了达到目标，我们从一个预训练的文本到图像扩散模型即.，SDXL [Podell et al. 2023] 开始，并使用基于传播的定制策略对其进行适应。如图 4 所示，在训练过程中，我们仅对前景标记进行定制以学习艺术风格，而在推理时，则利用基于 SDXL 的 BrushNet [Ju et al. 2024] 适配器根据学到的样式修复背景，该样式通过适配器传播。值得注意的是，在推理阶段只将适配器应用于背景标记以保持目标化的风格化。具体来说，我们的传播适配器实现为一个低秩适配器 [Hu et al. 2021;

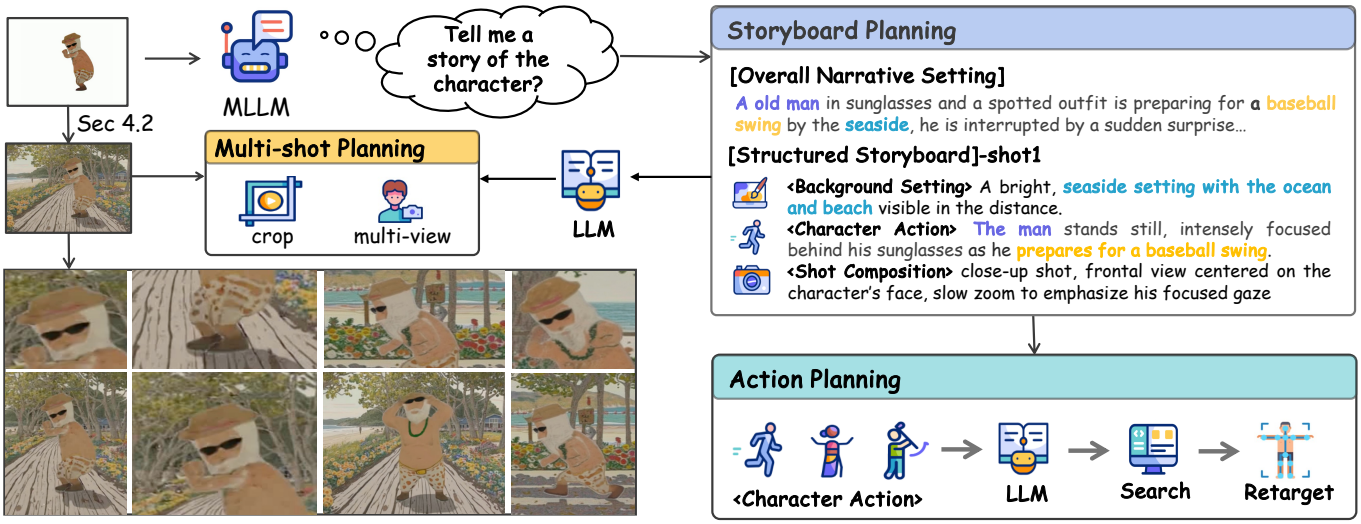

Fig. 3. **故事板生成的管道**。我们首先使用 M-LLM 规划整个故事，并构建包含场景、事件、角色动作、背景和镜头的分镜头脚本。然后，我们采用不同的镜头裁剪风格化的图像并生成最终的镜头图像。

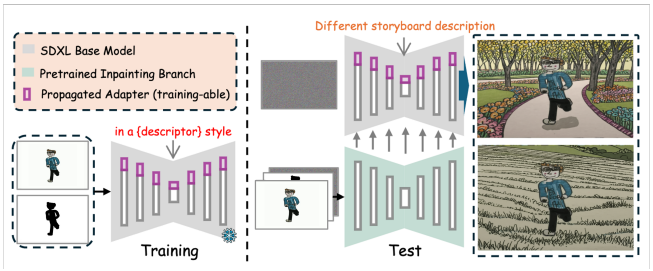

Fig. 4. **风格一致的场景生成**.

Liu et al. 2024b]。特别地，我们发现 DoRA [Liu et al. 2024b] 在捕捉样式细节方面表现出更好的性能。

形式上，设 x 表示完整的图像标记，m 是二值前景掩码。传播适配器 $PA(\cdot)$ 在训练过程中按以下方式更新模型：

$$y = Wx + PA(x \cdot m), \tag{2}$$

其中 W 是 SDXL 的原始权重，而 y 是自定义特征输出。

在推理过程中，我们选择用于图像风格化的背景区域，这可以公式化为：

$$y = Wx + PA(x \cdot (1 - m)). \tag{3}$$

总结而言，我们的方法在训练过程中学习角色的视觉风格，并在推理过程中有选择地将其传播到背景中。这种简单而有效的策略确保生成的场景与角色在风格上保持统一，支持连贯且视觉沉浸感强的动画。

4.3 通过 3D 代理生成字符视频序列

先前的工作学习使用定制的图像生成方法 [Ruiz et al. 2023] 生成身份一致的视频，然后使用图像到视频扩散模型进行叙事 [Gong et al. 2023]。然而，受限于图像到视频扩散模型的生成先验，生成身份一致且连贯的动作本质上是具有挑战性的，因为前景人物动作非常复杂。

不同地，我们从传统的计算机图形管线中汲取灵感，这些管线通过中间的 3D 表示对角色动作进行细粒度控制。具体来说，根据绘制自旋向上图 [Zhou et al. 2024a]，我们采用基于 3D 代理的动作建模方法，该方法从单一 2D 草图重构出角色的基础 3D 几何形状。此代理允许我们将骨骼绑定和动作重定向技术应用于其中，使得能够将复杂的动作序列转移到角色上同时保持结构保真度和视觉一致性。通过融入这种显式的动作结构，我们为生成语义上有意义且视觉连贯的动画序列提供了坚实基础。更多细节请参阅原论文。

4.4 动画通过运动定制

为了从背景合成的帧生成可动画化的视频镜头，我们利用了一个图像到视频扩散模型。然而，现有的图像到视频扩散模型 [Blattmann et al. 2023; Kong et al. 2024; Wan et al. 2025] 在为风格化或拟人角色生成复杂动作时遇到困难，导致身份不一致和时间闪烁。此外，类似 ControlNet [Zhang et al. 2023] 的视频控制模型 (例如. ，姿势引导和深度引导) 在非人类角色上的泛化能力有限，并且由于过于僵硬的约束条件常常产生不自然或与场景脱节的动作。不同的是，我们利用从 3D 重建中提取的角色动作序列作为训练数据来微调视频扩散模型。我们的主要挑战是实现镜头级别的动画，其中只有特定的身体部位（例如. 、头部或腿部）被赋予动画效果。在这种部分动作下的直接训练通常无法保持帧间的外观一致性。此外，现有的视频扩散模型需要大量的训练迭代来学习复杂的运动模式，即使对于单个序列也是如此。为了解决这些问题，我们采用了一种受 DynamicConcept [Abdal et al. 2025] 启发的两阶段训练策略，明确地将空间外观学习与时间动作学习分离，如图 5 所示。在第一阶段，模型通过对时间打乱的帧进行训练来学习身份特征，而不考虑时间相关性。在第二阶段，冻结身份适配器，并引入一个单独的动作适配器。然后使用新颖的时间步移位策略对按时间顺序排列的帧进行训练，以有效捕捉时间动态。我们的策略不是精确复制动作，而是学会生成能够适应各种背景和叙事环境的动作序列。在推理过程中，学习到的动作与背景合成场景组合以产生连贯且风格化的动画。

令 W 表示视频扩散模型的基础权重，令 $A_{\text{id}}, B_{\text{id}}$ 表示低秩身份适配器矩阵即. ，LoRA [Hu et al. 2021]。身份适应特征的计算公式为：

$$y = Wx + A_{\text{id}}B_{\text{id}}x, \tag{4}$$

其中 x 是输入特征。为了防止模型在身份训练过程中无意中学习到时间模式，我们将 dropout 应用于 B_{id}：$B_{\text{id}} = B_{\text{id}} \odot M_p$，其中 M_p 是一个二进制掩码，其 dropout 概率为 p。在第二阶段，我们固定了身份适配器，并

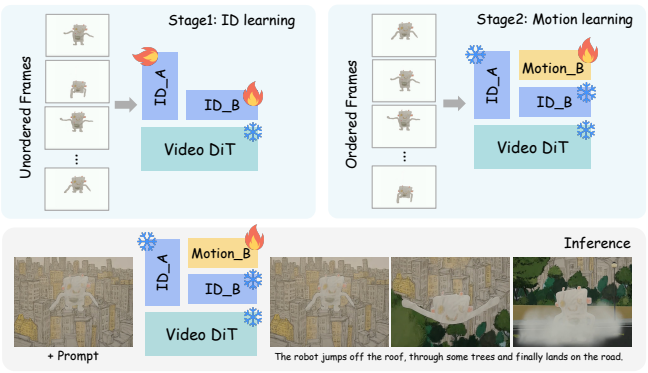

Fig. 5. **两阶段运动传动策略。我们首先使用无序帧来学习字符空间特征，而不带时间偏置。然后，在身份 LoRA 冻结的情况下，从连续的视频帧中学习运动残差。**

引入了一个特定于运动的适配器 B_{motion}，应用于连续帧。运动被建模为身份表示之上的残差变形：

$$y = Wx + A_{\text{id}}B_{\text{id}}x + A_{\text{id}}B_{\text{motion}}x. \tag{5}$$

还应用了 dropout 到 B_{motion} 上以稳定训练并防止过拟合，确保 A_{id} 在两个训练阶段中保持稳定的共享基础。

虽然这种两阶段训练有效地将运动与外观分离，但我们通过一种新颖的时间步长偏移采样策略进一步改进了运动建模，我们将其识别为捕捉图像到视频运动定制中现实且连贯的角色动态的关键。标准的扩散训练均匀地抽样时间步长 $t \in \{1, \ldots, T\}$，强调干净和噪声帧同等重要。然而，我们认为偏向于在扩散过程后期的更嘈杂的时间步长进行训练，会使模型依赖全局结构而不是低层次像素线索。我们使用高斯采样随后进行 sigmoid 变换来实现这种偏见：

$$t = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad z \sim \mathcal{N}(\mu, \sigma^2) \tag{6}$$

其中 μ 控制采样偏倚而 σ 控制方差。通过将 μ 设置得更接近 T，我们构建了一个偏向后期的采样分布，增加了抽样高噪声训练步骤的概率。这种偏向后期的采样促使模型在具有挑战性的条件下学习稳健的运动表示。实证观察表明，该策略产生更加平滑和时间上一致的运动轨迹，特别是在包含复杂角色互动的长序列中。

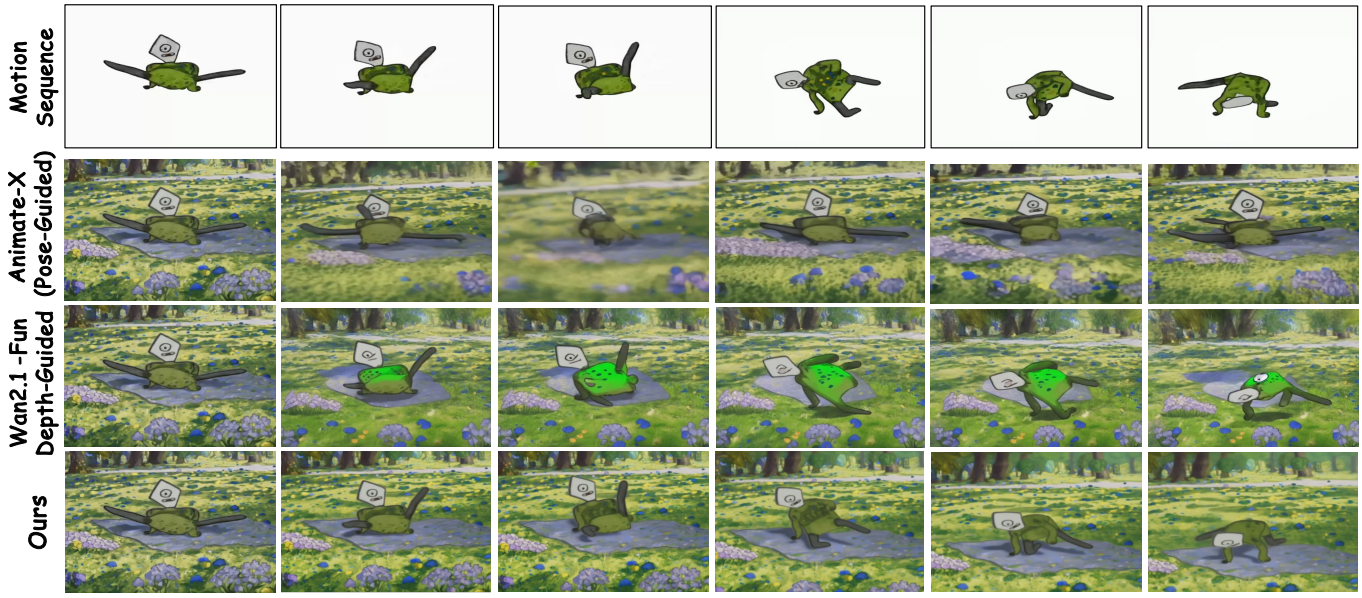

Fig. 6. **与运动定制比较**. 我们将提出的动作定制方法与基于深度的图像到视频方法使用 Wan2.1 [Wan et al. 2025]、基于姿态的图像到视频角色动画方法即. 、Animate-X [Tan et al. 2024] 进行了比较，提出的方法对于这种复杂动作显示出了与原始动作序列非常相似的结果。

5 实验

5.1 实现细节

对于实验和与其他方法的比较，我们使用了 Animated-Drawings 数据集 [Smith et al. 2023] 作为我们的儿童绘制角色。我们生成了 24 张不同风格的图像和 12 个不同的动作视频来进行风格和动作对比。所有实验都是基于一台 NVIDIA L20 GPU 进行的，学习风格需要 120 分钟，定制动作需要 180 分钟。对于风格化处理，我们没有像 DreamBooth 那样依赖人工标识符（例如.，“一种 [v] 风格”），而是发现使用描述性语言提示（例如.，“一种孩子般的梦幻风格”）能更好地与诸如纹理、笔触方向和线条质量等细粒度的风格属性保持一致。为了实现这一自动化，我们采用 GPT-4 [OpenAI 2022] 生成参考图像风格的文本描述，这些描述随后被用作类似于 StyleDrop [Sohn et al. 2023] 的训练提示的一部分。训练完成后，我们可以使用文本生成背景风格和动作。

对于风格评估，我们计算生成图像和源图像的 CLIP [Radford et al. 2021] 距离作为风格对齐得分，在此过程中我们也计算生成图像与相应的 CLIP 文本特征之间的 CLIP 距离。至于运动评估，我们从 VBench [Huang et al. 2024; Liu et al. 2024a] 中选择了两个指标，包括运动平滑度和主体一致性。

Fig. 7. **与风格化方法比较**。我们将我们的方法与不同的风格化方法在风格定制上进行了比较。

Methods	数值比较		用户研究	
	Style Align	Text Align	Style Quality	Visual Quality
B-LoRA	0.5060	**0.2829**	0.0267	**0.3429**
Instant Style	0.5468	0.2368	0.3403	0.0517
DreamBooth	0.6371	0.2819	0.0965	0.2803
Ours	**0.6580**	0.2702	**0.5365**	0.3251

表 1. **风格比较与其它方法相比**。

Methods	数值比较		用户研究	
	Motion Smooth.	Subject Consist.	Motion Realness	Visual Quality
Animate-X	0.974	0.908	0.106	0.023
Wan2.1-Fun	0.977	0.842	0.114	0.106
Ours	**0.987**	**0.955**	**0.780**	**0.871**

表 2. **运动比较**。**与其他方法相比**。

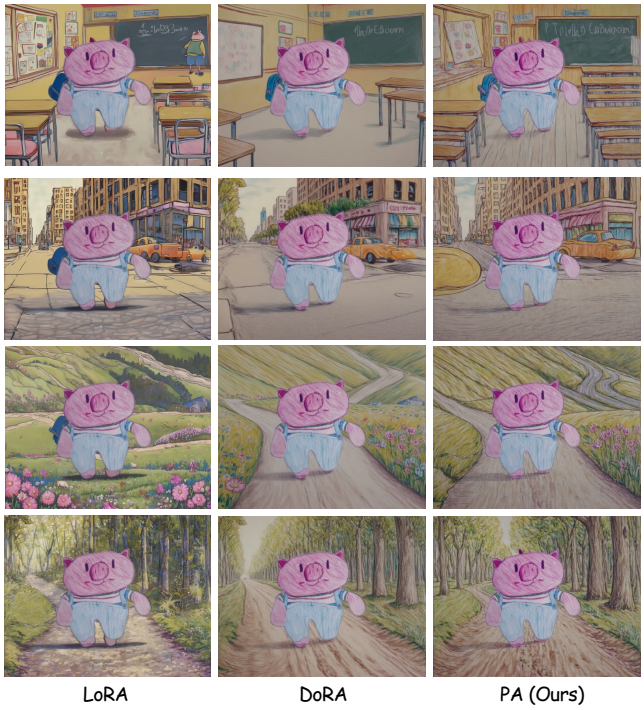

Fig. 8. **风格定制的消融研究**。**与基线方法** LoRA [Hu et al. 2021] **和** DoRA [Liu et al. 2024b] **相比，所提出的方法可以成功地将前景风格传播到具有不同提示的背景中**。**建议放大查看**。

5.2 与其他方法的比较

由于此任务之前没有基线，我们首先将我们的方法与从文本描述生成多事件视频的方法进行比较，即., MEVG [Oh et al. 2024] 和 Vlogger [Zhuang et al. 2024]，以及最先进的少样本主题驱动的视频生成方法，即., DreamVideo [Wei et al. 2024]。如图 10 和图 11 所示，所提出的方法在保持角色外观一致性、动作平滑性和风格保留方面表现出更好的效果。相比之下，多事件生成模型难以产生连贯的叙述和一致的视觉风格，而以主题驱动的方法则无法稳健地保存身份和动作，并且经常会产生不现实或不一致的背景。通过利用 3D 代理作为运动先验和定制化的风格传播适配器，我们的方法实现了更加连贯、忠实于风格并且在时间上平滑的视频生成结果。

除了视觉比较，我们还进行了详细的定量评估。对于风格对比，我们评估了风格一致性以及与伴随文本描述的相关性。如表 1 所示，我们的方法在主观和客观的风格化指标上都优于先前的方法。至于运动质量，我们将之与两种视频角色动画方法进行比较：一种是基于姿态引导的方法 Animate-X，它使用人类运动视频作为精确关键点检测的参考；另一种是基于深度引导的方法，利用从 3D 重建的角色运动序列中提取的深度序列。我们在表 2 中将运动质量与之前的基线进行比较，结果显示提出的方法显著优于其他方法。

我们进一步在风格化图像和生成视频上进行主题实验，以评估所提出方法的有效性。如表 2 和 表 1 所示，我们邀请了 24 名用户来评估 24 组风格化图像集，每组包含 4 种不同的方法，并需要从两个方面进行评估。对于运动部分，我们使用了 3 种不同方法的 12 个视频集，且用户需要从两个方面进行评价。最后，我们获得了 3360 条意见。如表中所示，用户一致认为我们的方法在风格对齐、动作真实感和视觉连贯性上更优。如表 1 所示，我们的方法在风格相似度上得分最高，超过了 B-LoRA、InstantStyle 和 DreamBooth。虽然我们的视觉印象评分（0.3251）略低于 B-LoRA（0.3429），但我们将此归因于 B-LoRA 的写实输出可能比儿童风格卡通图像更具有视觉吸引力。对于视频结果（表 2），

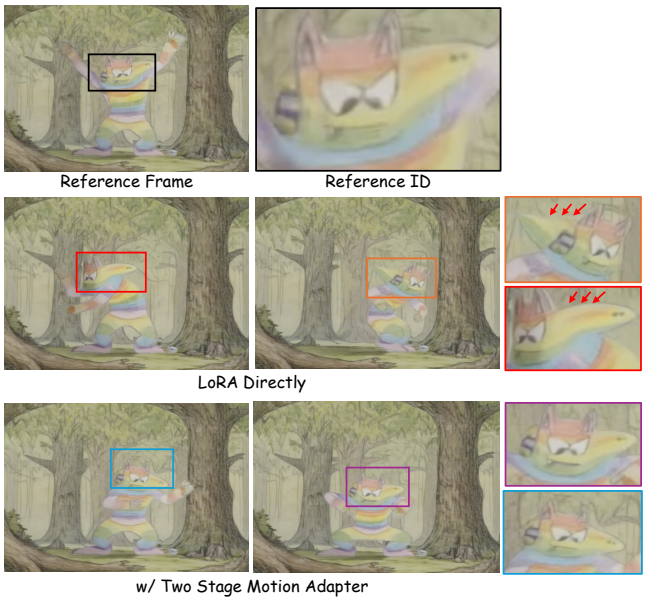

Fig. 9. **两阶段运动适配器的消融研究。我们在提出的图像到视频生成中的动作定制中移除了两阶段适配器。这里，第一阶段的训练提高了身份相似性。**

我们的方法相较于现有方法表现出明显的优势，突显了其生成时间上一致且风格忠实动画的能力。

5.3 消融研究

5.3.1 **传播样式适配器**. 我们首先评估所提出的风格传播适配器的有效性。如图 8 所示，给定前景角色的图像，DoRA 自定义方法比原始 LoRA 方法显示出更好的风格传播效果。此外，所提出的传播适配器能够成功学习到更详细地保留背景条纹的更好风格。

5.3.2 **运动适配器**. 我们提出的运动适配器采用了两阶段定制策略来进行动作定制。这里我们给出了每个阶段的有效性。如图 9 所示，直接在图像到视频扩散模型上训练 LoRA 会导致不自然的身份生成，而两阶段的运动适配器则表现更好。

5.3.3 **运动学习的时间移位**. 我们的关键技术之一是在步长采样中利用时间步移位以获得更好的运动效果。我们在图 12 中展示了不同采样步骤的结果。如图所示，直接使用均匀采样可能无法学习到与原始样本相似的运动，而 $\mu = 6$ 提供了比均匀采样和其他超参数更好的结果。

5.4 限制与未来工作

我们仅展示了单个字符的结果。然而，我们的方法很容易扩展到多个具有多个 3D 代理的主体。前景角色（或动物）可能并不总是能被 3D 代理正确重建，我们认为更先进的绑定方法 [Zhang et al. 2025a] 将有助于更好地生成前景的动作。视频扩散模型的生成先验可能并不总能准确地生成稳定且可动画化的背景。如图 13 所示，提出的方法生成了静态背景图像。我们将尝试不同的图像到视频扩散模型 [Kong et al. 2024] 并包括更多的相机运动 [Bai et al. 2025] 以提高动作的真实性。

6 结论

我们提出了 FairyGen，这是一个新的故事可视化框架，将叙事分解为前景角色动作和环境动态，使得能够以统一的风格和运动进行分层建模。我们首先使用多模态大型语言模型（M-LLM）规划故事情节，然后采用样式传播适配器生成与叙述背景一致的样化背景。为了定制动作，我们引入了动作感知适配器和时间步采样策略以灵活控制角色动态。与几个基线方法相比，我们的方法在样化背景生成和动作定制方面实现了高质量的结果，展示了优越的适应性和视觉连贯性。

References

Rameen Abdal, Or Patashnik, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. 2025. Dynamic Concepts Personalization from Single Videos. arXiv:2502.14844 [cs.GR]

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. 2025. ReCamMaster: Camera-Controlled Generative Rendering from A Single Video. arXiv:2503.11647 [cs.CV] https://arxiv.org/abs/2503.11647

Xiuli Bi, Jian Lu, Bo Liu, Xiaodong Cun, Yong Zhang, WeiSheng Li, and Bin Xiao. 2024. CustomTTT: Motion and Appearance Customized Video Generation via Test-Time Training. *arXiv preprint arXiv:2412.15646* (2024).

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

Tim Brooks and Alexei A Efros. 2022. Hallucinating pose-compatible scenes. In *European Conference on Computer Vision*. Springer, 510–528.

Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama, and Nanyun Peng. 2022. Character-centric story visualization via visual planning and token alignment. *arXiv preprint arXiv:2210.08465* (2022).

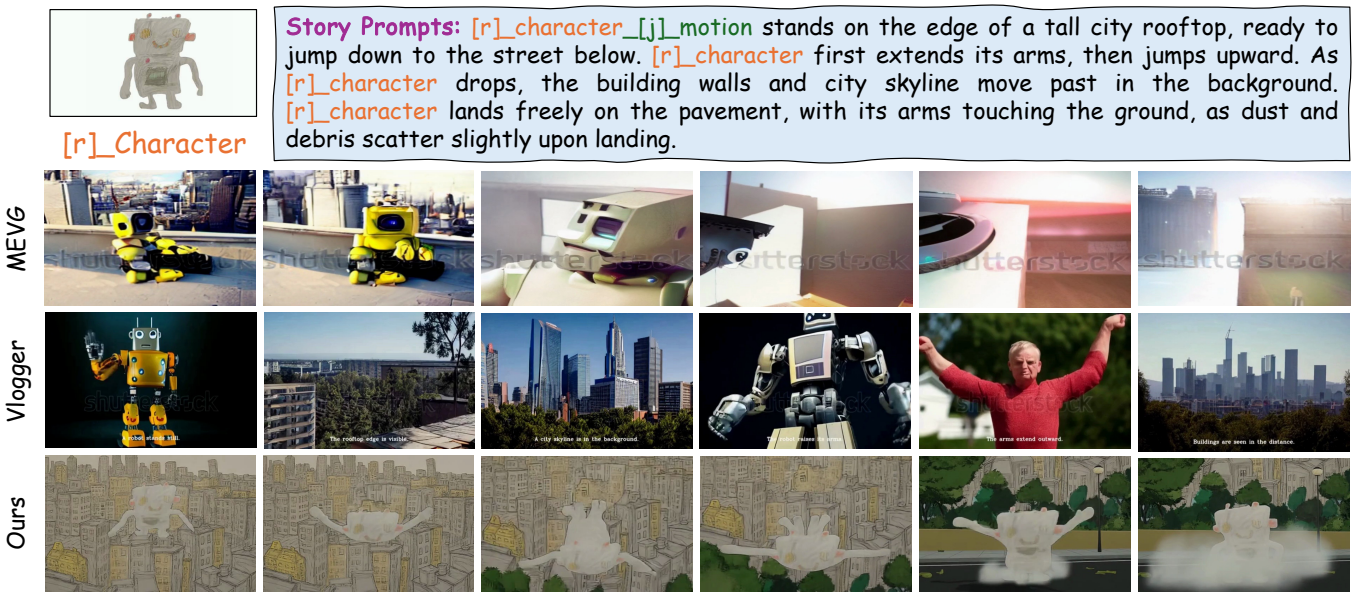

Fig. 10. **多事件视频生成的比较**。我们的方法将前景和背景建模分开，这有利于较长且多事件视频的生成。这里，我们使用相同的故事情节提示来生成视频，在此情况下，提出的方法显示出与文本描述一致的结果。

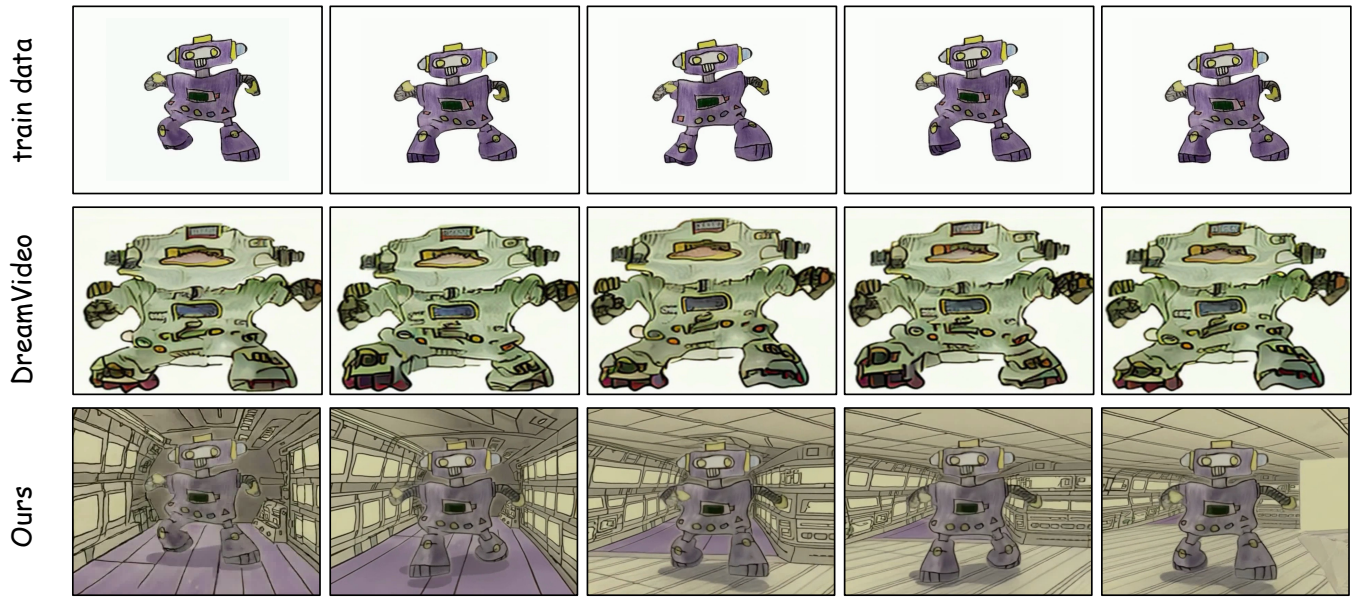

Fig. 11. **外观和运动定制方法的比较**。我们将我们的方法与最先进的外观和运动定制方法即. 、DreamVideo [Wei et al. 2024] 进行了比较，所提出的方法在风格化、动作以及整体质量方面显示出了明显更好的结果。

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. 2023. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512* (2023).

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7310–7320.

DeepSeek. 2025. DeepSeek-R1: A Reasoning Model. https://deepseek.com/.

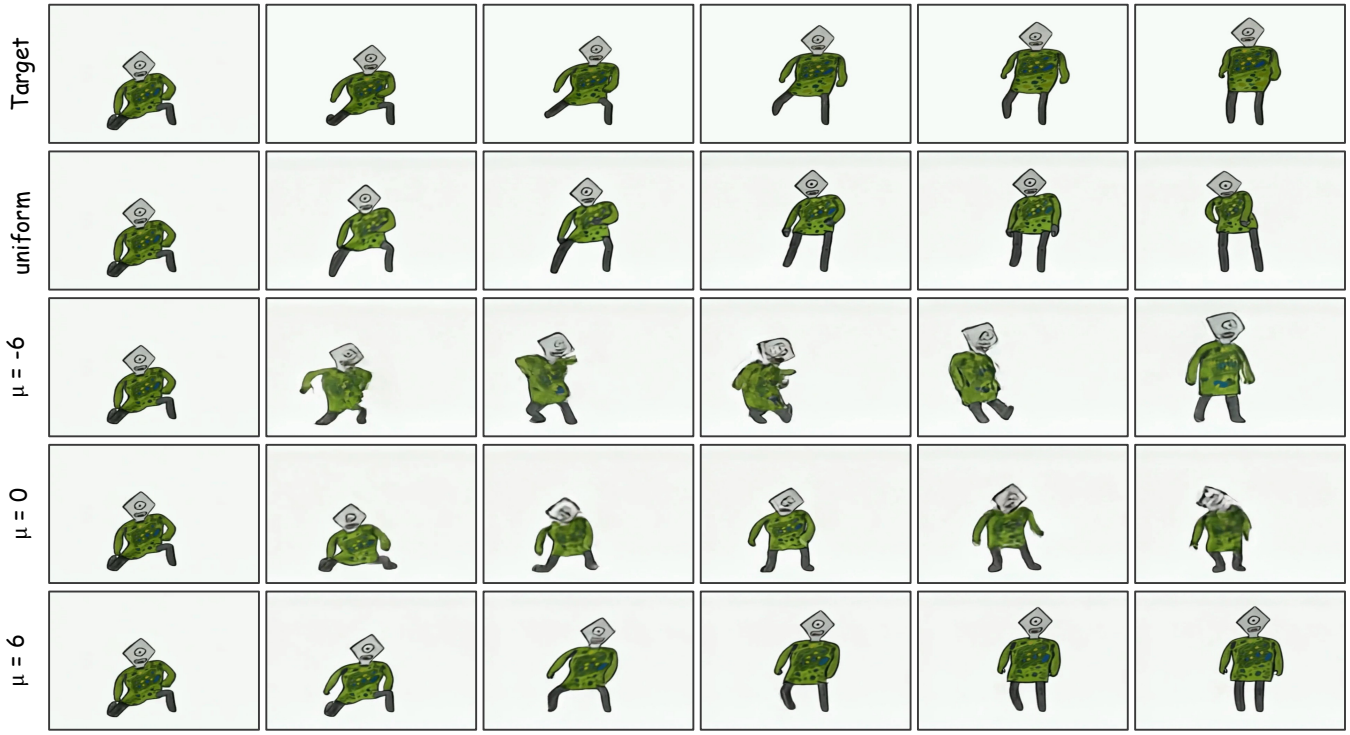

Fig. 12. **时间步移除研究**提出的运动定制中的时间步移位策略可以更好地表示运动。

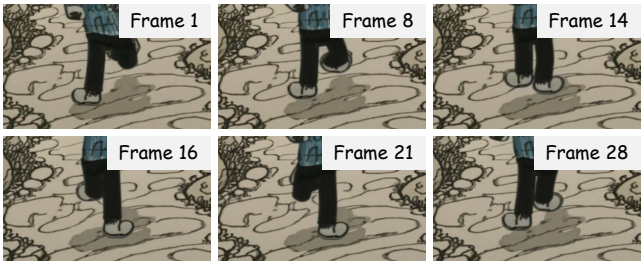

Fig. 13. **限制**。由于视频扩散模型的不可控生成先验，所提出的方法可能只会生成带有动画前景运动的静态背景 (例如. ，跑步)。

Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. 2025. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*. Springer, 181–198.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. TaleCrafter: Interactive Story Visualization with Multiple Characters. arXiv:2305.18247 [cs.CV]

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. 2024. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *arXiv preprint arXiv:2407.12899* (2024).

Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940* (2023).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. 2024. StoryAgent: Customized Storytelling Video Generation via Multi-Agent Collaboration. *arXiv preprint arXiv:2411.04925* (2024).

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. 2024. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition. 6689–6700.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. 2024. BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion. arXiv:2403.06976 [cs.CV]

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).

Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. 2023. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17089–17099.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.

Bowen Li. 2022. Word-level fine-grained story visualization. In *European conference on computer vision*. Springer, 347–362.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6329–6338.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. 2024c. Intelligent Grimm - Open-ended Visual Storytelling via Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6190–6200.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024b. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024a. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22139–22149.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European conference on computer vision*. Springer, 70–87.

Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. 2024. Mevg: Multi-event video generation with text-to-video models. In *European Conference on Computer Vision*. Springer, 401–418.

OpenAI. 2022. ChatGPT. https://openai.com/blog/chatgpt.

Boxiao Pan, Zhan Xu, Chun-Hao Huang, Krishna Kumar Singh, Yang Zhou, Leonidas J Guibas, and Jimei Yang. 2024b. Actanywhere: Subject-aware video background generation. *Advances in Neural Information Processing Systems* 37 (2024), 29754–29776.

Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. 2024a. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2920–2930.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

Qwen. 2025. Qwen Model. https://help.aliyun.com/product/170553.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2493–2502.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22500–22510.

Harrison Jesse Smith, Qingyuan Zheng, Yifei Li, Somya Jain, and Jessica K. Hodgins. 2023. A Method for Animating Children ' s Drawings of the Human Figure. *ACM Trans. Graph.* 42, 3, Article 32 (jun 2023), 15 pages. doi:10.1145/3592788

Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. 2023. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983* (2023).

Sitong Su, Litao Guo, Lianli Gao, Heng Tao Shen, and Jingkuan Song. 2023. Make-a-storyboard: A general framework for storyboard with disentangled and merged control. *arXiv preprint arXiv:2312.07549* (2023).

Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobin Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. 2024. Animate-X: Universal Character Image Animation with Enhanced Motion Representation. *arXiv preprint arXiv:2410.10306* (2024).

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. 2024. Storyimager: A unified and efficient framework for coherent story visualization and completion. In *European Conference on Computer Vision*. Springer, 479–495.

Team Wan, 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, and Ziyu Liu. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).

Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. 2024. AutoStory: Generating Diverse Storytelling Images with Minimal Human Efforts. *International Journal of Computer Vision* (2024), 1–22.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. 2024. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6537–6549.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. 2024a. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683* (2024).

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024b. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072* (2024).

Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. 2023. Inserting Anybody in Diffusion Models via Celeb Basis. *arXiv preprint arXiv:2306.00926* (2023).

Jinlu Zhang, Jiji Tang, Rongsheng Zhang, Tangjie Lv, and Xiaoshuai Sun. 2025b. Storyweaver: A unified world model for knowledge-enhanced story character customization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 9951–9959.

Jia-Peng Zhang, Cheng-Feng Pu, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. 2025a. One Model to Rig Them All: Diverse Skeleton Rigging with UniRig. *arXiv preprint arXiv:2504.12451* (2025).

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. 2024. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655* (2024).

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. 2025. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*. Springer, 273–290.

Jie Zhou, Chufeng Xiao, Miu-Ling Lam, and Hongbo Fu. 2024a. DrawingSpinUp: 3D Animation from Single Character Drawings. *SIGGRAPH Asia 2024 Conference Papers* (2024). https://api.semanticscholar.org/CorpusID:272654016

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. 2024b. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. *NeurIPS 2024* (2024).

Zhongyang Zhu and Jie Tang. 2025. Cogcartoon: towards practical story visualization. *International Journal of Computer Vision* 133, 4 (2025), 1808–1833.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. 2024. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8806–8817.