LAMP: 面向少样本视频生成的运动模式学习器

Ruiqi Wu[1, 3 *] Liangyu Chen[3] Tong Yang[3] Chunle Guo[2, 1 †] Chongyi Li[2, 1] Xiangyu Zhang[3]
[1]VCIP, CS, Nankai University [2]NKIARI, Shenzhen Futian [3]MEGVII Technology
wuruiqi@mail.nankai.edu.cn, {chenliangyu, yangtong, zhangxiangyu}@megvii.com
{guochunle, lichongyi}@nankai.edu.cn

图 1. **我们的文本生成视频结果**. 每组帧下方列出了运动提示/视频提示。我们的 LAMP 在各种运动场景中都能有效运行。生成的视频表现出显著的时间一致性，并且与视频提示接近。此外，LAMP 的两个优势在所展示的结果中得以体现。(1) 提出的运动-内容解耦流程使我们能够利用 SD-XL 的强大能力生成高度详细的内容。(2) 由于我们创新的调优方法，扩散模型出色的语义泛化能力得以保留（例如在未见过的漫画风格中强加微笑的动作）。

Abstract

在本文中，我们提出了一种小样本文本生成视频的框架，***LAMP***，该方法使得文本生成图像的扩散模型能够在单个 *GPU* 上通过 *8* 至 *16* 个视频 ***L****earn* ***A*** *specific* ***M****otion* ***P****attern*（学习特定的运动模式）。与现有方法

* 本工作为武睿祺于旷视科技实习期间完成。

不同的是，现有方法需要大量的训练资源或学习与模板视频精确对齐的运动，而我们的方法在生成自由度和模型训练的资源成本之间取得了平衡。具体而言，我们设计了一个运动-内容解耦的方法，使用现成的文本生成图像模型进行内容生成，使得我们调整过的视频扩散模型主要集中于运动学习。成熟的文本生成图像技术能够提供视觉效果令人满意且多样化的生成条件，这极大地提高了视频质量和生成自由度。为了捕捉时间维度的特征，我们将预训练的文本生成图像模型的二维卷积层扩展为我们创新的时空运动学习层，并将注意力模块修改为时间级别。此外，我们还开发了一种有效的推理方法——共享噪声采样，这可以在不增加计算成本的情况下提高视频的稳定性。我们的方法还可以灵活应用于其他任务，例如真实图像动画和视频编辑。大量实验表明，*LAMP* 能够在有限的数据上有效学习运动模式并生成高质量的视频。代码和模型可在 *https://rq-wu.github.io/projects/LAMP* 获取。

1. Introduction

近年来，生成模型，特别是基于扩散的模型 [13, 38, 39]，在通过文本提示生成图像方面取得了显著的成就，即文本生成图像（T2I）[6, 10, 17, 21, 27, 29, 32, 33, 35]。基于 T2I 领域的技术基础，扩散模型为文本生成视频（T2V）领域带来了繁荣发展。一些最新的研究 [2, 11, 14, 36, 41, 50] 试图通过训练基于扩散的 T2V 模型，利用数百万的文本-视频对来实现开放域的 T2V 生成，如图 2(b) 所示。这些方法促进了对视频与文本提示之间关系的深入理解。然而，庞大的标注数据需求和沉重的训练负担对大多数研究者来说是难以承受的，这限制了该研究方向的发展。另一种基于模板的方法 [5, 7, 26, 30, 43, 45, 46] 涉及使用视频模板，并在保持原始运动的同时使用扩散模型操控内容，如图 2(a) 所示。尽管这些方法成本较低，尤其是提出了单样本 [43] 甚至零样本 [7, 30, 46] 算法，但使用给定的视频模板显著限制了生成的自由度。此外，最近的一些工作 [15, 19, 22, 42] 对 T2I 扩散模型进行了修改，以在不进行训练的情况下生成一致的视频。然而，零样本方式下将文本-图像领域的知识转移到文本-视频领域具有挑战性，导致这些方法生成的帧看起来相似，但运动随机。

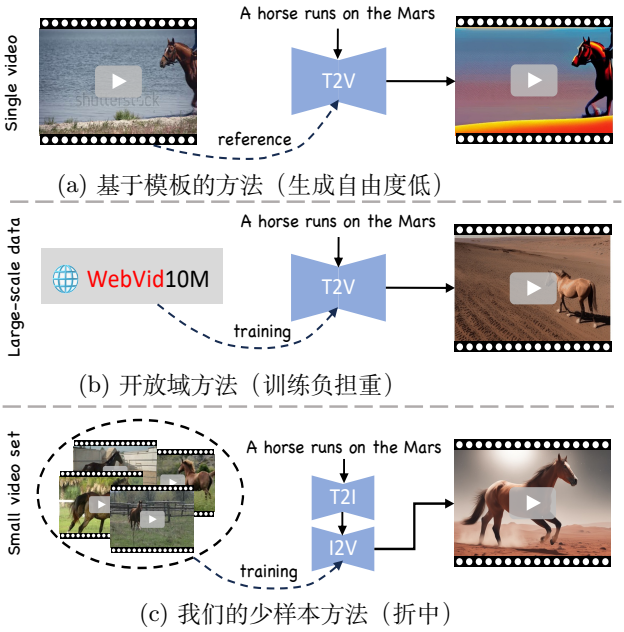


图 2. 基于模板的方法、开放域方法与我们的少样本方法的比较。我们使用一个小型视频集而不是单个视频模板和大规模数据集来学习该集中的通用运动模式。我们的少量样本方法可以在生成自由度（相对模板方法）和训练负担（相对于开放域方法）之间实现折中。

在使模型理解运动的同时，必须在训练负担和生成自由度之间取得平衡。由于预训练的 T2I 扩散模型在提示的指导下具有良好的语义理解能力，因此使其理解提示与运动之间的对应关系并生成多样化的视频，实际上需要的数据非常少。

本文尝试为 T2V 任务探索一种新的少样本方法。这种新方法旨在调整 T2I 扩散模型，从一个小型视频集学习一个通用的运动模式，如图 2 所示。

在少量样本的方式下，将 T2I 模型调整为 T2V 模型时，需要解决两个问题。(1) 由于数据量有限，存在过拟合视频集中内容的风险。如果生成的视频与视频集过于相似，这将破坏我们核心目标之一，即生成自由度。(2) T2I 扩散模型的基本操作只作用于空间维度，限制了其捕捉视频中时间信息的能力。

本文提出了一种用于少样本 T2V 生成的基础方法，命名为 ***LAMP***，以解决上述两个问题。我们对第一个问题的解决方案是提出的**运动-内容解耦方法**。该方法将 T2V 任务分解为两个子任务：由预训练的 T2I 模型生成首帧，并通过视频扩散模型预测后续帧。提出的方法无缝地将首帧作为条件，而无需对视频扩散模型进行任何额外的修改（如更改输入的数据结构或添加

新的交叉注意力层）。具体来说，在训练期间，我们保留输入视频的首帧，添加噪声，并仅对后续帧施加损失。由于首帧提供了视频的大部分内容，我们的模型可以专注于后续帧与首帧之间的关系，即运动模式而非内容。在推理时，首帧由预训练的 T2I 模型（如 SD-XL [29]）生成。我们观察到，高质量的首帧通过提出的流程可以提升视频生成的表现。有了首帧提供的参考，我们的模型基于 Stable Diffusion v1.4 (SD v1.4) [33]，能够在整个视频中保留由 SD-XL 生成的高质量内容。针对第二个问题，我们设计了**时间-空间运动学习层**，以同时捕捉时间和空间维度的特征。由于提出的流程要求基于首帧预测后续帧，我们基于视频预测任务 [18, 25] 对基本操作进行了修改，具体将在 Sec. 3.4 中介绍。与之前的工作类似 [22, 43]，我们修改了注意力层，以在帧之间建立有效的通信。此外，我们在推理期间采用了**共享噪声采样策略**，它从共享噪声构建每帧的原始噪声。该策略显著提高了生成视频的质量和稳定性，而计算成本几乎可以忽略不计。

我们在多个运动场景下评估了 LAMP，通过使用 8 至 16 个视频在单个 GPU 上进行简单调整，提出的 LAMP 能够生成具有视频集通用运动模式的视频，并能够很好地泛化到未见过的风格和对象。（见图 1）。我们的主要贡献总结如下：

- 我们提出了一种新的少样本调整方法，用于 T2V 生成任务，旨在在生成自由度和训练成本之间取得平衡。
- 我们提出了 LAMP，作为少样本 T2V 的基础方法。该方法结合了提出的运动-内容解耦方法和时间层，可以通过简单的调整有效学习给定视频集中的运动模式。
- 大量实验证明，我们的 LAMP 在提示对齐、帧一致性和内容多样性方面表现出色。

2. Related Work

2.1. 文本生成图像扩散模型

近年来，扩散模型 [13, 24, 38, 39] 在文本生成图像任务中击败了 GANs [4, 8, 47]、VAEs [23, 37, 40] 和基于流的模型 [3, 9]，因其稳定的训练过程和出色的性能而备受关注。例如，GLIDE [27] 使用文本提示作为条件，并采用无分类器指导 [12] 来提升图像质量。DALLE-2 [32] 引入了预训练的 CLIP 模型 [31]，该模型广泛应用于多模态领域 [49]，用于对齐图像和文本的特征。Imagen [35] 将大语言模型的特征注入到扩散模型中，以更好地理解提示，并提出了从粗到细生成高分辨率图像的级联流程。为了减轻迭代去噪过程的计算负担，Rombach 等人提出了 LDM [33]，该模型使用自动编码器 [4, 23] 来减少图像的冗余。LDM 首先通过预训练的自动编码器将图像压缩到低维潜在空间，然后学习去噪有噪声的潜在数据。随着 LDM 的成功，许多技术变体 [28, 48] 被提出以进一步提升性能。最近，SD-XL [29] 被提出，它可以生成极具照片真实感且具备高分辨率细节的图像。在我们的工作中，SD-XL 用于生成首帧，SD-v1.4 则被修改用于预测后续帧。

2.2. 文本生成视频扩散模型

扩散模型在文本生成图像领域的繁荣展现了其在文本生成视频方面的潜力。主流的工作可以分为两类：开放域 T2V 生成和基于模板的方法。

开放域 T2V 生成。在早期阶段，ImagenVideo [14] 和 Make-A-Video [36] 在像素级别上学习 T2V。然而，由于像素空间中的高计算量，视频的长度和分辨率受到显著限制。随后，MagicVideo [50] 被提出，该方法在视频数据上训练了一个新的自动编码器。随着 LDMs [33] 在 T2I 领域的出现，MagicVideo 提升了 T2V 生成的计算效率。Blattmann 等人 [2] 提出了一种基于 LDMs 的 T2V 扩散模型，该模型在冻结的预训练层上增加了额外的 3D 卷积层。VideoComposer [41] 通过一个新的编码器，为 T2V 模型增加了多种条件，如草图和运动矢量。AnimateDiff [11] 训练了一组运动层，这些层可以应用于定制的 T2I 模型 [16, 34]，使其能够生成风格一致的视频。上述方法在 T2V 生成方面取得了显著的性能。然而，训练这些模型需要像 WebVid-10M [1] 和 HD-VILA-100M [44] 这样的海量数据，这对大多数研究人员来说是一个较为严重的障碍。此外，虽然一些零样本方法 [15, 19, 22] 已经被提出，但它们通常存在帧一致性不佳的问题。

基于模板的方法。基于模板的 T2V 生成旨在通过用户提示进行视频到视频的翻译，这也被称为视频编辑。Dreamix [26] 和 GEN-1 [5] 是基于模板方法的两个开创性工作，尽管它们的训练成本与开放域 T2V 方法相当。随后，Tune-A-Video [43] 提出了一种新的单样本方法，

利用 T2I 模型对原始视频进行过拟合，这可以在消费级 GPU 上实现。FateZero [30] 提出了一种无训练的方法，通过注入源视频的交叉注意力图并修改注意力层来实现。Rerender-A-Video [46] 和 TokenFlow [7] 通过整合先验和条件引导进一步提高了视频的一致性。与基于模板方法的目标不同，我们的少样本 T2V 方法旨在实现更高的生成自由度，而不是严格与模板视频的运动模式对齐。

3. Method

本节中，Sec. 3.1 和 Sec. 3.2 首先介绍了基础知识和最新的少样本方法。接下来，Sec. 3.3 详细描述了我们提出的运动-内容解耦方法。然后，Sec. 3.4 描述了我们如何将 T2I 扩散模型修改为 T2V 生成模型。最后，Sec. 3.5 介绍了我们的共享噪声采样策略以及推理时可以提高性能的一些技术。

3.1. 基础知识

在本节中，我们介绍基于扩散模型的基础知识。给定数据 $x_0 \in X$，可以定义一个马尔可夫链为：

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \quad (1)$$

其中 $t = 1, ..., T$，T 是总步数。β_t 是控制第 t 步噪声强度的系数。迭代添加噪声可以简化为：

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (2)$$

其中 $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_t)$。扩散模型通过最小化训练目标来学习数据集 X 的分布，训练目标可以写为：

$$\arg\min_{\theta} \mathbb{E}_{x_0,\epsilon\sim\mathcal{N}(0,I),t,c}[||\epsilon - \epsilon_\theta(x_t, t, c)||_2^2], \quad (3)$$

其中 $\epsilon_\theta(\cdot)$ 表示扩散模型的噪声预测函数，c 是条件，如文本提示。训练完成后，扩散模型可以通过逆转添加噪声的过程从噪声中生成数据。

然而，当扩散模型用于生成高分辨率图像时，计算成本将会变得很大。为了解决这一挑战，提出了用于 T2I 生成的潜在扩散模型（LDMs），采用自动编码器在隐空间中完成所有操作。它们生成低冗余的隐空间特征以实现有效计算，并通过解码器重建图像。LDMs 也被用于我们的方法中生成高分辨率视频。

3.2. 我们基于少样本的 T2V 生成设置

现有的 T2V 方法需要大规模数据进行训练或依赖模板视频获取低自由度的生成能力。为了使视频生成既便宜又灵活，我们提出了一种新的方法：少样本 T2V 生成。假设有一个视频集 $\mathbf{V} = \{\mathcal{V}_i | i \in [1, n]\}$，包含 n 个视频以及描述共同运动的训练提示 $\mathcal{P}_m$。所提出的新方法是在给定的视频集和运动提示上调优一个 T2I 模型。调优后的模型可以从与运动相关的提示 $\mathcal{P}$ 中生成具有与 $\mathbf{V}$ 类似运动模式的新视频 $\mathcal{V}'$。我们希望从少量的视频集中学习共同的运动模式，同时忽略内容。同时，由于数据量小，训练成本是可以承受的。基于所提出的方法，我们修改了预训练的 T2I 模型并提出了一个基础框架用于少样本 T2V 生成。

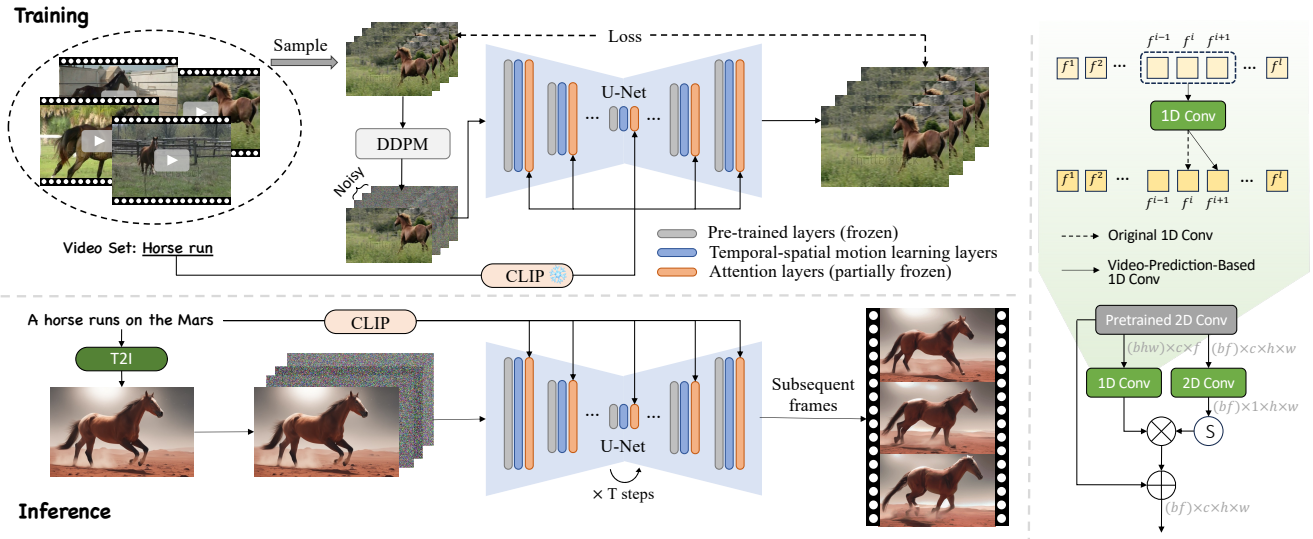
3.3. 运动-内容解耦管道

由于少样本调优过程中数据有限，存在过拟合小数据集内容的风险，可能会影响生成自由度。为了让我们的模型更专注于运动，我们提出了运动-内容解耦方法来解耦运动和内容。该方法如图 3 (a) 所示。根据我们的观察，首帧包含了短视频的大部分内容。因此，自然地可以使用首帧作为条件，使模型更关注运动。因此，T2V 生成任务被转化为首帧 T2I 生成和后续帧预测。一些之前的工作 [5, 41] 也使用了首帧作为条件。它们将其与输入噪声连接或添加一个特定的编码器将特征注入网络。然而，在少样本方法中使用这些方法是具有挑战性的，因为有限的数据使得通过对 T2I 模型的重大修改来促进模型训练几乎不可能。相比之下，提出的运动-内容解耦管道可以通过微小的参数变化实现类似效果，详细见 Sec. 3.4。

具体来说，设 $\mathcal{V} = \{f^i | i = 1, ..., l\}$ 为一个包含 l 帧的视频，将其编码为潜在空间：$\mathcal{Z}_0 = \{z^i | i = 1, ..., l\}$。在训练模型时，我们保留 z^1 的原始信号，并对 $\{z^2, ..., z^l\}$ 添加噪声。

损失函数与方程 (3) 一致，并且仅施加在第 2 帧到第 l 帧上。经过训练后，模型能够根据首帧生成带有视频集运动模式的视频。在推理过程中，使用强大的 SD-XL [29] 提供首帧 $\hat{f}^1$，该帧被解码为 $\hat{z}^1$。然后，序列 $[\hat{z}^1, \epsilon^2, ..., \epsilon^l]$（其中 ϵ 为随机噪声）被输入模型用于整个视频的生成。每一步中，我们保留首帧的深层特征，并对后续帧进行去噪。

所提出的方法有效地避免了学习视频集的内容，从



(a) 整体框架　　(b) 时空运动学习层

图 3. (a) 展示了 **LAMP 的框架**。LAMP 从一个小型视频集学习运动模式，使生成的视频带有已学习的运动模式。此方法在视频生成中实现了训练资源与生成自由度的平衡。我们将文本生成视频任务转化为首帧生成和后续帧预测，即解耦视频的内容和运动。在训练过程中，除首帧外，我们对所有帧添加噪声并计算损失函数。此外，仅调整新添加的层和自注意力模块的查询线性层的参数。在推理阶段，我们使用 T2I 模型生成首帧。调优后的模型仅负责在用户提示的引导下，对后续帧的隐特征进行去噪。(b) 展示了 **时空运动学习层** 的细节。基于视频预测的 1D 卷积层利用前两帧的特征，而不是相邻帧。

而可以在有限数据上训练模型。另一个优势在于 SD-XL 生成的内容质量高，为视频生成提供了良好的参考。这种方法使我们能够利用已有的 T2I 技术优势。运动-内容解耦方法显著提高了提示对齐性能和生成多样性。此外，该方法在应用中的灵活性也颇具吸引力，如实际图像动画和视频编辑，详见 Sec. 5。

然而，原始 T2I 模型将帧视为独立样本。因此，首帧的特征无法用于建立帧间的时间关系并生成视频。接下来的部分将介绍我们如何使模型在时间维度上发挥作用。

3.4. 将 T2I 模型适配视频生成

时空运动学习层。为了使 T2I 模型能够提取时间特征，我们将预训练的 2D 卷积层扩展为提出的时空运动学习层。如图 3(b) 所示，提出的层由两个分支组成。假设输入视频的隐特征表示为形状为 $b \times c \times f \times h \times w$ 的 5D 张量。在时间分支中，张量被重塑为 $bhw \times c \times f$ 并输入 1D 卷积层。然而，由于 1D 卷积核一次只能作用于一个空间坐标，因此无法考虑关键的空间特征。因此，增加了一个输出通道为 1 的 2D 卷积层以及 Sigmoid 函数，作为空间特征的补充。在空间分支中，输入特征被重塑为 $bf \times c \times h \times w$。

考虑到我们的运动-内容解耦方法需要视频扩散模型在第二步中根据给定的首帧预测后续帧，这与视频预测 [18, 25] 类似，我们以视频预测的方式设计了 1D 卷积层。当卷积核滑过帧的特征 $\{f^{i-1}, f^i, f^{i+1}\}$ 时，我们的**基于视频预测的 1D 卷积**生成帧 f^{i+1} 的特征，而不是原始版本中的 f^i。因此，我们可以利用前两帧而不是两帧相邻的特征来预测后续帧，即在基本操作中实现有效的视频预测。此外，为了避免新添加的层污染预训练 T2I 模型的生成能力，所有参数均按 ControlNet [48] 中的方式初始化为零。

注意力层。我们还修改了注意力层以确保一致性。对于自注意力层，所有键和值特征均来自首帧，表示为：

$$\text{Attention}(Q^i, K^1, V^1) = \text{Softmax}(\frac{Q^i (K^1)^T}{\sqrt{d}}) V^1, \quad (4)$$

上标 $i \in \{1, ..., l\}$ 表示特征来自第 i 帧。结合所提出的方法，重新设计的自注意力层使得后续帧能够参考由首帧建立的条件。此外，事实证明，即使不进行调优，这种修改也能有效保持主要物体的完整性 [22]。此外，按照 [43] 中的修改，我们还加入了时间注意力层，这些层在时间维度上执行自注意力。

3.5. 推理中的共享噪声采样

在推理过程中，我们提出了一种简单而有效的共享噪声采样策略，以进一步提高生成视频的质量。具体来说，我们首先采样一个共享噪声 $\epsilon^s \sim \mathcal{N}(0, I)$。然后，采样一个与基础噪声具有相同分布的噪声序列 $[\epsilon^2, ..., \epsilon^l]$。在我们的采样策略中，用于第 i 帧生成的原始噪声 ϵ^i 被更新为：

$$\epsilon^i = \alpha\epsilon^s + (1-\alpha)\epsilon^i, \tag{5}$$

其中 α 是控制共享程度的系数。根据经验，在实验中我们设置 $\alpha = 0.2$。这种方法确保了各帧之间噪声水平的一致性，最终表现为生成视频的一致性。直观上，这种方法符合先验知识，即视频的每一帧都有一定的相似性。从数学上讲，噪声方差的减少可以缩小隐空间的动态范围，有助于生成过程的稳定性。此外，AdaIN [20] 技术在隐空间中以及像素级的直方图匹配被用于后处理。我们推理策略的有效性在 Sec. 4.3 中得到了验证。

4. Experiments

4.1. 实验细节

在我们的实验中，我们生成了分辨率为 320×512 且包含 16 帧的视频。我们使用 SD-XL [29] 进行计算成本较低的第一帧生成，并使用相对轻量的 SD-v1.4 [33] 进行计算成本较高的后续帧预测，从而平衡了两个阶段的推理成本。对于训练阶段，我们使用了一组自己收集的视频，每次迭代中随机采样一个 16 帧的片段。所有帧都被调整为 320×512 的分辨率。我们仅调整了新添加的层和自注意力块中的查询线性层的参数，学习率设为 3.0×10^{-5}。所有实验均在一块 A100 GPU 上实现，训练大约需要 15 GB vRAM，推理大约需要 6 GB vRAM。

4.2. 对比实验

我们针对 8 种运动模式训练了我们的 LAMP，包括直升机（刚体运动）、瀑布（流体运动）、雨水和烟花（粒子运动）、马跑动（动物运动）、鸟飞翔（多体运动）、转身微笑（人类情感）以及弹吉他（人类运动）。我们为每种运动设计了 6 个提示语以构建包含 48 个视频的评估集。选择了三个公开可用的方法作为我们的对比方法，分别是大规模预训练的 AnimateDiff [11]、基于单样本的视频编辑方法 Tune-A-Video [43] 和基于零样本的 Text2Video-Zero [22]。我们在多种主流方法下考虑了有代表性的工作，进而有效地反映了我们少样本学习方法的优势。特别地，对于每种运动模式，我们从对应的视频集中随机选择一个视频作为模板来训练 Tune-A-Video [43]。比较在客观和主观两个方面进行。

表 1. **与评估的文本到视频方法的定量比较**。$*$：表示在我们的数据上微调。†：表示使用 SDv1.4/1.5 进行第一帧生成或作为骨干网络。

方法	对齐性 ↑	连续性 ↑	多样性 ↓
Tune-A-Video [43]	27.22	94.87	84.72
T2V-Zero [22]	26.94	91.47	73.01
AnimateDiff [11]	<u>28.88</u>	97.81	<u>73.47</u>
AnimateDiff$*$	28.85	**98.56**	78.81
LAMP (Ours)	**31.35**	<u>98.31</u>	**71.65**
AnimateDiff †	<u>28.54</u>	96.03	<u>75.11</u>
AnimateDiff $*^\dagger$	27.65	**97.82**	81.91
LAMP (Ours) †	**29.99**	<u>97.15</u>	**73.65**

定量结果我们在文本对齐、帧一致性和生成多样性方面评估了我们的 LAMP 与基线方法。使用了客观指标和用户调研进行全面评估。

客观指标为了衡量视频的文本对齐性，我们取每帧的 CLIP 得分的平均值 [31]。按照 [43]，我们还用所有帧对之间的 CLIP 图像嵌入的平均余弦相似度来表示帧一致性。由于生成自由度是我们的核心目标之一，我们还在定量评估中包括了生成多样性。我们使用所有帧的平均 CLIP 图像嵌入来表示一个视频。对于每个运动模式对象，我们随后计算并平均所有视频对之间的余弦距离。得分越低表示相似度越低，即多样性越好。AnimateDiff 和我们的 LAMP 都可以使用更好的生成模型作为骨干来提升性能。因此，我们在 LAMP 的第一帧生成阶段使用 SD-v1.4，并将 AnimateDiff 的骨干网络从 RealisticVision 替换为基础模型 SDv1.5。此外，我们还在自己的数据上微调了 AnimateDiff，以进一步展示所提出的运动-内容解耦方法的有效性。表 1 展示了 LAMP 与基线方法的定量结果。在所有三项评估标准上，我们的方法相对于其他基线方法取得了最好的表现。微调的 AnimateDiff 在视觉一致性方面表现最好，但由于过拟合，生成多样性表现非常差。

用户调研我们进一步进行了一项用户调研，以主观评估我们的方法和三个基线方法。我们从评估集中随机

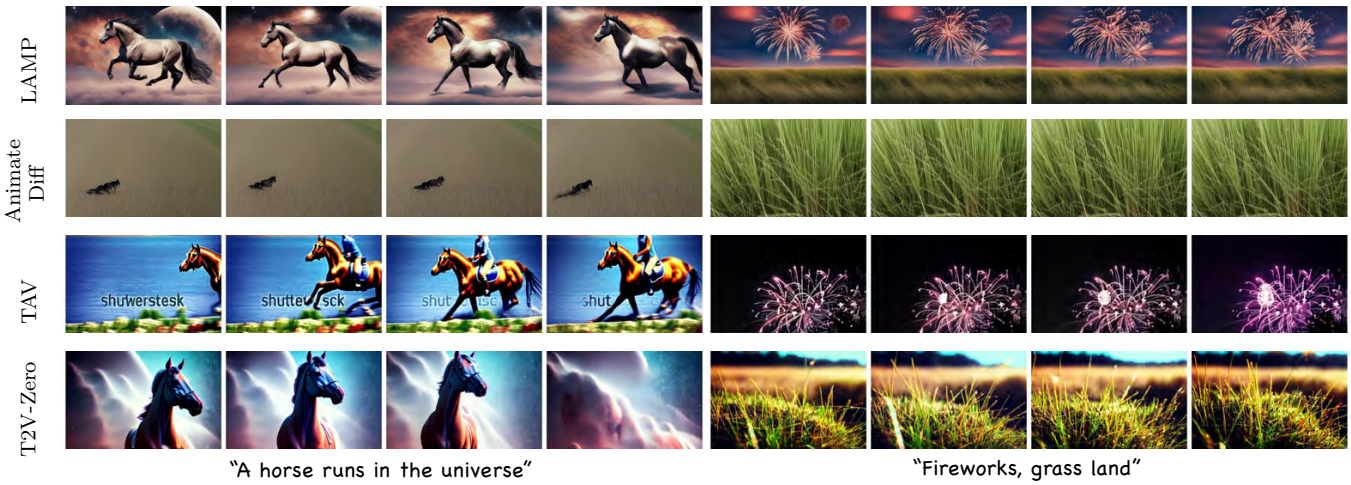


图 4. 提出的 LAMP 与三个方法的定性比较。**请放大查看细节**。

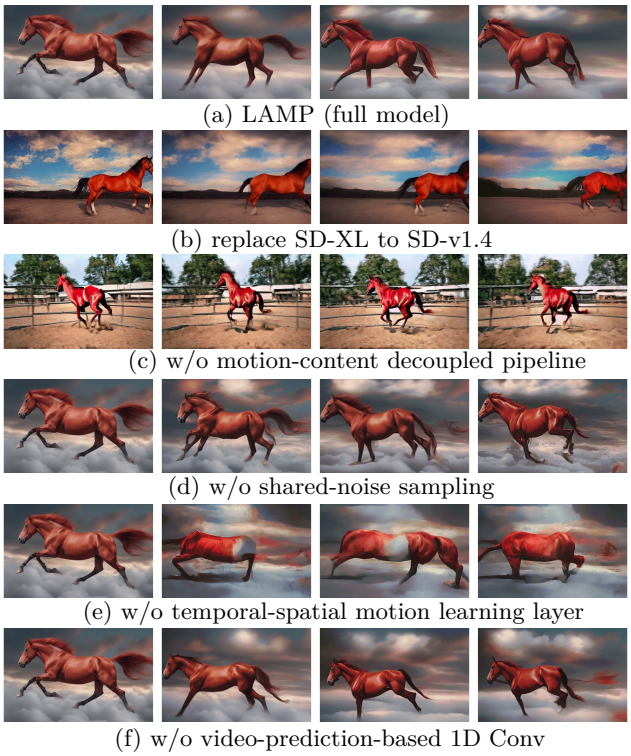


图 5. **消融实验结果**。给定提示语为“红色的马在天空中奔跑”。

选择了 24 个案例。在每个案例中，我们询问参与者：“您认为哪个视频的视觉质量更好，并且更符合提示语中的场景和运动?” 用户调研共收集了 70 份来自不同背景的参与者的反馈，包括该领域的专家和没有特定背景知识的个人。从统计上看，46.84% 的受访者更喜欢我们的方法，AnimateDiff [11] 得到 19.11% 的支持，Tune-A-Video [43] 得到 22.15% 的支持。然而，值得注

图 6. LAMP 将真实世界图像动画化的可视化结果。

意的是，Tune-A-Video 在不同的情况下会出现选择上的极端分化。当给定的视频模板的布局与提示语描述的场景相似时，加上其良好的帧一致性，大多数受访者会认可它。相反，其生成视频的文本对齐度较差，例如图 4 中展示的“烟花，草地”。此外，有 11.90% 的受访者选择了 Text2Video-Zero [22]。因此，我们的 LAMP 在受访者中获得了最高的认可率。我们的问卷可以通过链接 访问，您可以通过补充材料中的视觉示例识别出我们的结果。

定性结果。我们在图 4 中展示了我们的方法和三个基线方法的几个可视化示例。AnimateDiff [11] 在大规模数据上学习运动层，并将其插入个性化的 T2I 模型中，以生成具有特定风格和更好视觉质量的视频。然而，该方法无法与性能更好但异构的 T2I 模型相结合，导致其文本对齐能力受限，尽管其一致性和多样性令人满意。这种限制在“宇宙中奔跑的马”和“烟花，草地”等案例中尤为明显。Tune-A-Video (TAV) 只能生成具有相同运动模式的视频，有时提示语无法有效控制生

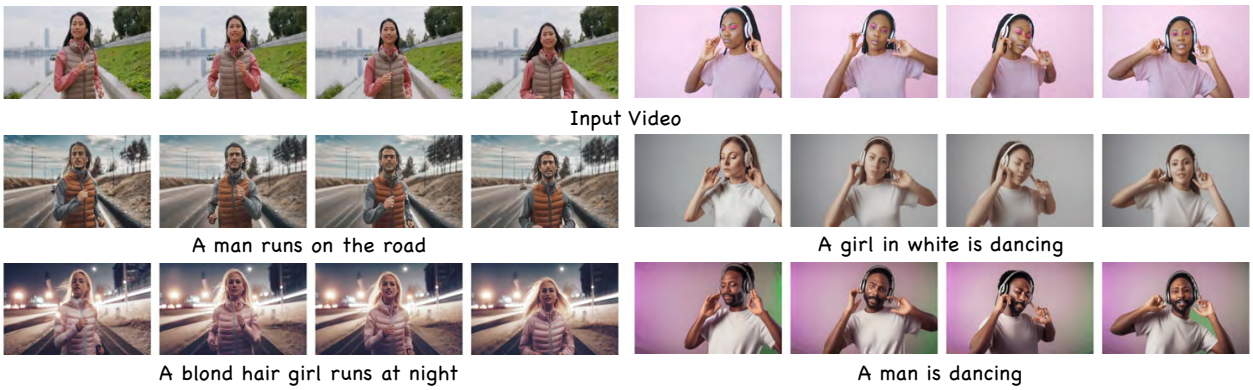


图 7. 我们的视频编辑应用的可视化结果。**请放大查看细节**。

成的视频内容，因为它在给定的视频上过度拟合。虽然 T2V-Zero 能生成视觉上令人满意的帧，但在生成具有有意义的运动模式的视频方面存在不足。相比之下，我们的 LAMP 利用所提出的运动学习层，达到了良好的一致性，并生成了具有适当运动模式的视频。此外，利用我们运动-内容解耦方法的优势，所提出的方法即使基于 SD-v1.4 的修改版，也能达到与最先进的 T2I 模型相媲美的视觉质量。图 1 和补充材料中提供了更多的可视化结果。我们的方法对运动的理解非常好，能够推广到多样化甚至未见过的场景和风格。

4.3. 消融实验

我们进行了消融实验，以证明每个提出的组件的有效性。如图 8(b) 所示，与完整模型相比，使用 SD-v1.4 生成第一帧会降低性能。将图 8(c) 与由完整模型生成的视频进行比较时，没有运动-内容解耦方法的模型会产生低质量的结果。特别是，视频中出现的不相关物体（如栅栏和泥土）表明内容对视频集的过拟合。此外，缺少共享噪声采样的模型可以生成相对一致的帧，但结果缺乏平滑性。当去掉时间-空间运动学习层时，模型无法有效捕捉复杂的运动模式，导致生成失败的结果。最后，当我们将基于视频预测的一维卷积改为原始版本的一维卷积时，视频的主要对象变得不一致。这些结果验证了每个关键模块对最终完整模型的重要贡献。

5. 更多应用

在本节中，我们提供了 LAMP 的更多应用场景，包括真实图像动画和视频编辑。

5.1. 真实图像动画

通过训练所提出的运动-内容解耦方法，我们的 LAMP 包含一个基于给定第一帧预测后续帧的网络。这使得由 T2I 模型生成的真实世界图像能够实现动画。因此，如果将这些图像放在第一帧上，我们的方法自然可以基于学习到的运动模式来为真实世界图像制作动画。图 6 显示了几个具有代表性的案例。即使面对复杂的真实场景，这一应用进一步展示了我们的泛化性能。

5.2. 视频编辑

在给定的训练集中仅包含一个视频片段的情况下，我们的方法只能学习特定的运动而不是运动模式。在这种特殊情况下，我们的方法有效地转变为一种视频编辑算法。训练过程与小样本方法中的情况类似。在推理过程中，我们基于 SD-XL [29] 采用 ControlNet [48] 并以 canny 边缘作为条件来编辑第一帧。同时使用 DDIM 反演 [43] 提供基础运动。类似于视频生成，当应用于视频编辑时，我们的方法也可以充分利用图像编辑技术。如图 7 中所示的可视化示例，我们的 LAMP 在保持良好帧一致性的同时生成了逼真的视频。

6. 结论

本文提出了一种新方法，即用于 T2V 生成的小样本微调，该方法从一个小视频集中学习通用的运动模式，以在训练负担和生成自由之间取得平衡。所提出的 LAMP 作为这种新方法的基础。在我们的方法中，我们将 T2V 任务转化为生成第一帧的 T2I 任务，并预测后续帧。这避免了在小样本微调期间对数据集内容的过拟合，同时利用了文本到图像技术的优势。此外，我们在网络架构和推理策略方面的新设计进一步提升了

T2V 生成的性能。大量实验表明了我们方法的有效性和泛化能力。我们相信，小样本微调方法提供了更优的权衡，并将有助于 T2V 领域更广泛地探索视频扩散训练所需数据的下限。

致谢。本研究工作由中国国家自然科学基金（62306153，62225604）资助。本研究所需的计算设备由南开大学超级计算中心（NKSC）支持。

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 3

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 3

[3] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[5] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2, 3, 4

[6] Deng-Ping Fan, Ziling Huang, Peng Zheng, Hong Liu, Xuebin Qin, and Luc Van Gool. Facial-sketch synthesis: a new challenge. *Machine Intelligence Research*, 19(4):257–287, 2022. 2

[7] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2, 4

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020. 3

[9] Matej Grcić, Ivan Grubišić, and Siniša Šegvić. Densely connected normalizing flows. *Advances in Neural Information Processing Systems*, 34:23968–23982, 2021. 3

[10] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. *arXiv preprint arXiv:2312.10113*, 2023. 2

[11] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 3, 6, 7

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 3

[15] Susung Hong, Junyoung Seo, Sunghwan Hong, Heeseong Shin, and Seungryong Kim. Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023. 2, 3

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[17] Guyue Hu, Bin He, and Hanwang Zhang. Compositional prompting video-language models to understand

procedure in instructional videos. *Machine Intelligence Research*, 20(2):249–262, 2023. 2

[18] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2023. 3, 5

[19] Hanzhuo Huang, Yufan Feng, and ChengShi LanXu JingyiYu SibeiYang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *arXiv preprint arXiv:2309.14494*, 2023. 2, 3

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 6, 12

[21] Ge-Peng Ji, Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Christos Sakaridis, and Luc Van Gool. Masked vision-language transformer in fashion. *Machine Intelligence Research*, 20(3):421–434, 2023. 2

[22] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2, 3, 5, 6, 7

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[24] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023. 3

[25] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016. 3, 5

[26] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 2, 3

[27] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 4, 6, 8, 12

[30] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 2, 4

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2, 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 6, 12

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 3

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic

text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[36] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3

[37] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 3

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 3

[40] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[41] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 2, 3, 4

[42] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. *arXiv preprint arXiv:2311.16813*, 2023. 2

[43] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 3, 5, 6, 7, 8

[44] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 3

[45] Shuzhou Yang, Chong Mou, Jiwen Yu, Yuhan Wang, Xiandong Meng, and Jian Zhang. Neural video fields editing. *arXiv preprint arXiv:2312.08882*, 2023. 2

[46] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 2, 4

[47] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 3

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 5, 8

[49] Xuying Zhang, Bo-Wen Yin, Yuming Chen, Zheng Lin, Yunheng Li, Qibin Hou, and Ming-Ming Cheng. Temo: Towards text-driven 3d stylization for multi-object meshes. *arXiv preprint arXiv:2312.04248*, 2023. 3

[50] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2, 3

Abstract

我们的补充材料提供了关于 *LAMP* 的更多细节和实验结果，具体如下：

- 我们介绍了 *LAMP* 的更多细节，尤其是推理阶段。
- 我们提供了更多消融实验和可视化结果。
- 我们提供了源代码和视频文件。

7. LAMP 的更多细节

在本节中，我们更具体地描述了我们方法的推理过程。我们首先使用文本到图像模型 $\mathcal{M}_I$（如 SD-XL [29]）生成第一帧。然后，将第一帧的隐特征与通过共享噪声采样获得的噪声进行拼接。在每一步中，只有后续帧的特征会通过基于 SD-v1.4 [33] 的视频扩散模型 $\mathcal{M}_V$ 进行更新。此外，采用 AdaIN [20] 确保外观一致性。需要注意的是，AdaIN 仅在前 30 步的后续帧中使用，因为在这些步中强制与噪声更接近的后续帧与第一帧保持一致会导致不期望的伪影。此外，在像素空间上采用直方图匹配来消除闪烁。推理过程的伪代码见算法 1，其中 $\mathcal{P}_I$ 和 $\mathcal{P}_V$ 分别为第一帧和整个视频的提示，t 为视频长度，$T = 50$ 表示 DDIM 反向的总步骤。

8. 实验

8.1. 更多消融实验

在本节中，我们给出了推理过程的更多消融结果，如图 8 所示。当我们在推理阶段移除 AdaIN 时，外观一致性会受到破坏。例如，在最后一帧中，马匹缺少前蹄。此外，直方图匹配可以有效恢复帧间的闪烁，如图 8(d) 所示。我们在经验上仅在 DDIM 反向的前 30 步中在后续帧之间使用 AdaIN，在最后 20 步中使后续帧与通过 AdaIN 的第一帧保持一致。我们尝试从一开始就在后续帧和第一帧之间使用 AdaIN。然而，在早期阶段强制与噪声接近的后续帧与第一帧保持一致会产生意外的伪影，如图 8(e) 所示。

8.2. 更多可视化结果

在我们的补充材料中，我们提供了 8 种运动模式的更多可视化结果，如图 9-16 所示。我们的 LAMP 可以生成多样且高质量的结果，并具有适当的运动。此外，还提供了视频文件。

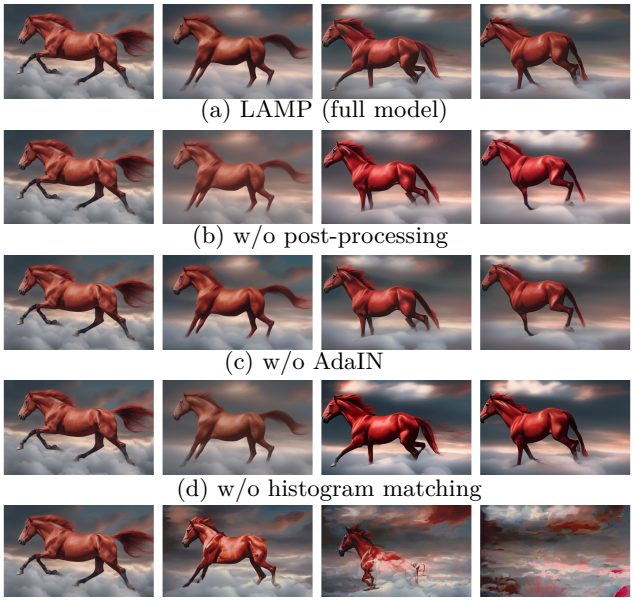
Algorithm 1 推理阶段伪代码

Require: $\mathcal{P}_I, \mathcal{P}_V, \mathcal{M}_I, \mathcal{M}_V, f \in \mathbb{N}, T \in \mathbb{N}$ and latent decoder $\mathcal{D}$

▷ First frame generation
$x_T^1 \sim \mathcal{N}(0, I)$
$x_0^1 = \text{DDIM\_Backward}(x_T^1, T, \mathcal{P}_I, \mathcal{M}_I)$
▷ Shared-noise sampling
$x_b \sim \mathcal{N}(0, I)$
for $i = \{2, 3, ..., f\}$ **do**
  $x_T^i \sim \mathcal{N}(0, I)$
  $x_T^i \leftarrow 0.8 x_T^i + 0.2 x_b$
end for
▷ Video generation
$x_T^{1:f} \leftarrow \{x_0^1, x_T^2, x_T^3, ..., x_T^f\}$
for $t = T - 1, ..., 0$ **do**
  $x_t^{2:f} \leftarrow \mathcal{M}_V^{2:f}(x_{t+1}^{1:f}, t + 1, \mathcal{P}_V)$
  ▷ Post-processing on latent space
  if $t > 20$ **then**
    for $i = 3, 4, ..., f$ **do**
      $x_t^i \leftarrow \text{AdaIN}(x_t^i, x_t^2)$ ▷ Ensure consistency between predicted frames
    end for
  else
    for $i = 2, 3, ..., f$ **do**
      $x_t^i \leftarrow \text{AdaIN}(x_t^i, x_0^1)$ ▷ Ensure consistency with the first frame
    end for
  end if
end for
▷ Post-processing on pixel space
$I^{1:f} = \mathcal{D}(x_0^{1:f})$
for $i = 2, 3, ..., f$ **do**
  $I^i \leftarrow \text{Historgram\_Matching}(I^i, I^1)$
end for

9. 局限性与未来工作

在我们的实验中，我们观察到当方法尝试学习复杂运动时，失败案例的发生频率会增加。更有效的运动学习模块可能是解决此问题的可能方案。此外，我们发现前景物体的运动有时会影响背景的稳定性。我们认



(a) LAMP (full model)

(b) w/o post-processing

(c) w/o AdaIN

(d) w/o histogram matching

(e) AdaIN with the first frame in all steps

图 8. 更多消融结果

为，独立学习前景和背景的运动可能是一个有效的解决方案。我们将这些改进留待未来工作中研究。

10. 更广泛的影响

由于我们的工作是基于现有的文本到图像技术，它可能会带有预训练模型本身的缺陷。像其他生成模型一样，我们的模型在没有安全检查器的情况下，可能会生成不安全或有偏见的视频，造成潜在的伤害。在模型实际部署之前，彻底评估模型的潜在风险并过滤有害内容是非常重要的。

图 9. 关于“马跑”运动的可视化结果。**视频文件可以在补充材料中找到**。

图 10. 关于“鸟飞”运动的可视化结果。**视频文件可以在补充材料中找到**。

图 11. 关于“直升机”运动的可视化结果。**视频文件可以在补充材料中找到**。

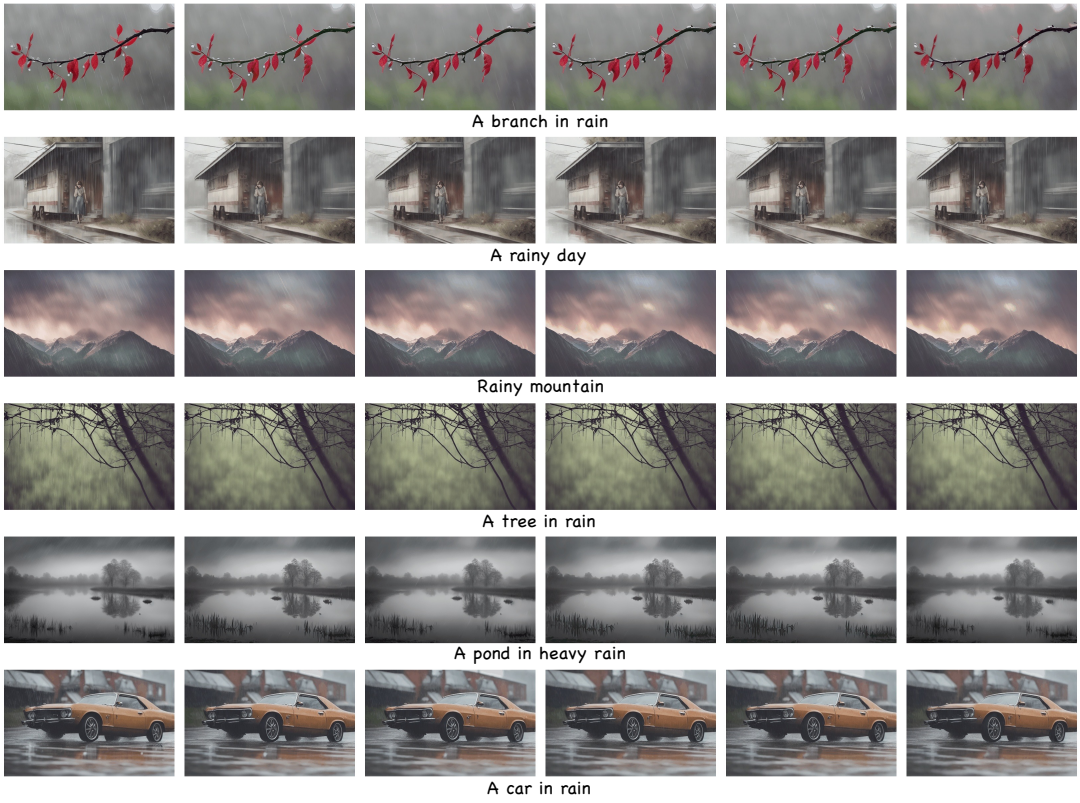


图 12. 关于“雨”运动的可视化结果。**视频文件可以在补充材料中找到**。

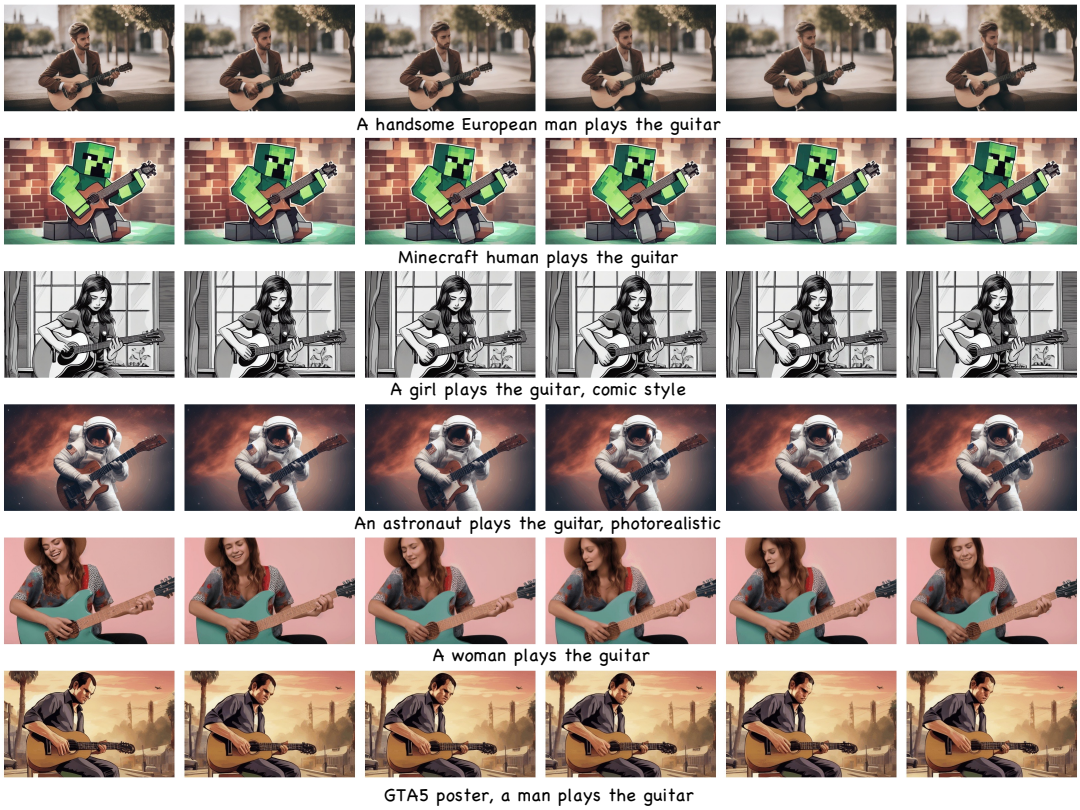


图 13. 关于“弹吉他”运动的可视化结果。**视频文件可以在补充材料中找到**。

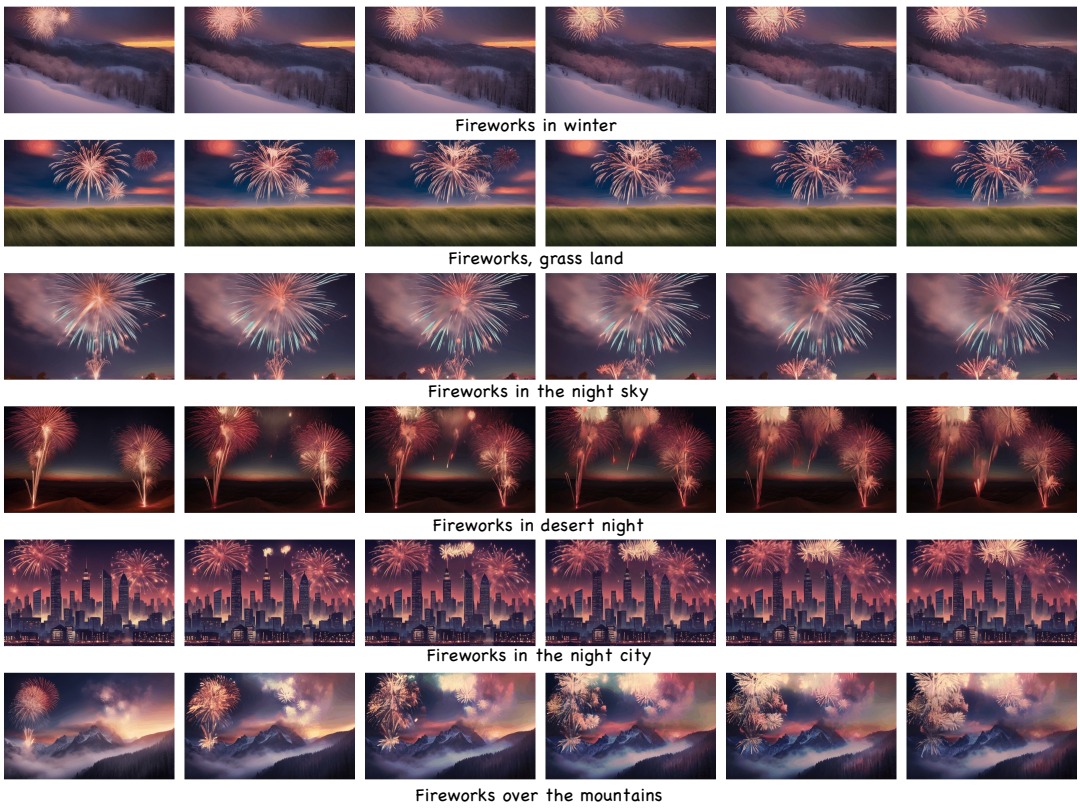


图 14. 关于“烟火”运动的可视化结果。**视频文件可以在补充材料中找到**。

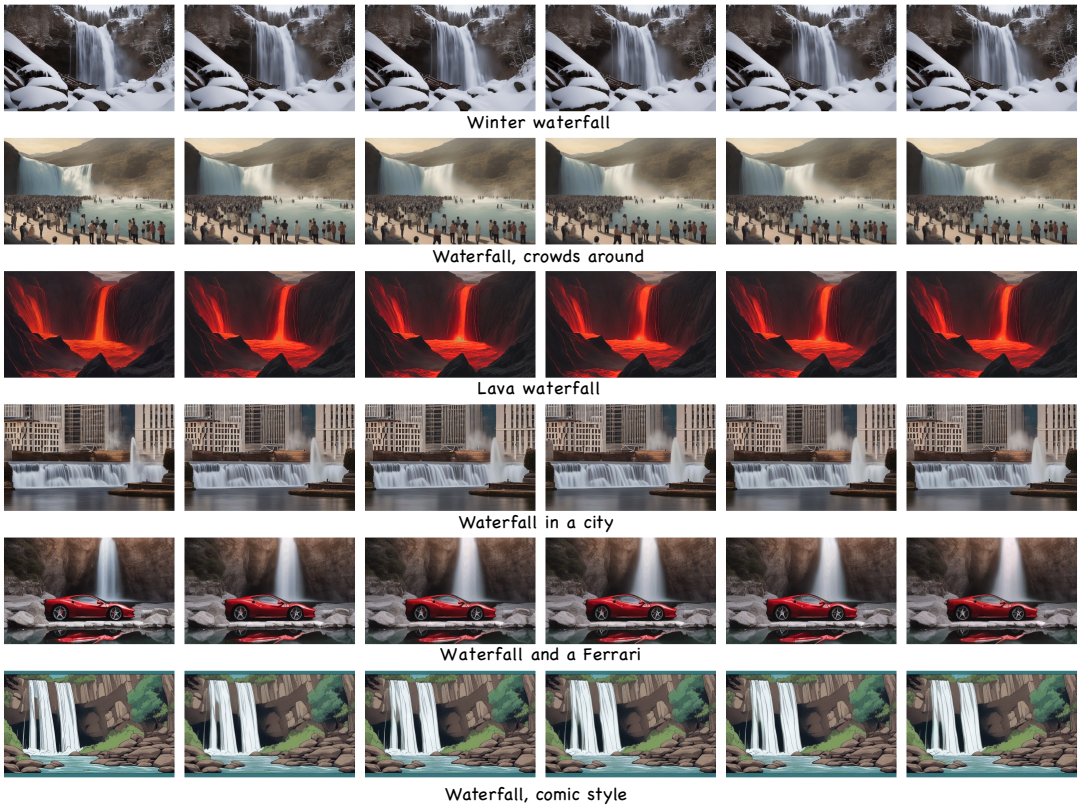


图 15. 关于“瀑布”运动的可视化结果。**视频文件可以在补充材料中找到**。

图 16. 关于“转头微笑”运动的可视化结果。**视频文件可以在补充材料中找到**。