arXiv:2304.01186v2 [cs.CV] 3 Jan 2024

Follow Your Pose：利用无姿态视频进行姿态引导的文本转视频生成

马越[1*†] 何映清[2*] Xiaodong Cun[3] Xintao Wang[3] 陈思然[4]李修[1‡] 陈启峰[2‡]

[1] 清华大学深圳国际研究生院，清华大学，中国深圳 [2] 香港科技大学，中国香港[3] 腾讯 AI 实验室，中国深圳[4] 中国科学院深圳先进技术研究院，中国深圳{y-ma21, lixiu}@mails.tsinghua.edu.cn, yhebm@connect.ust.hk, {vinthony,xintao.alpha}@gmail.com Chensiran17@mails.ucas.ac.cn, cqf@ust.hk

摘要

生成可文本编辑、姿态可控的角色视频，对于创作各种数字人有着迫切需求。然而，由于缺乏包含成对视频姿态描述的综合数据集，以及视频生成先验模型，这项任务一直受到限制。在这项工作中，我们设计了一种新颖的两阶段训练方案，利用易于获取的数据集（即图像姿态对和无姿态视频）以及预训练文本到图像（T2I）模型，生成姿态可控的角色视频。具体而言，在第一阶段，仅使用姿态图像对进行可控的文本到图像生成。我们训练一个零初始化卷积编码器来编码姿态信息。在第二阶段，我们通过加入可学习的时间自注意力和改造后的跨帧自注意力模块，利用无姿态视频数据集对上述网络的动作进行微调。得益于这些新设计，我们的方法成功生成了连续且姿态可控的角色视频，同时保留了预训练 T2I 模型的概念生成和构图能力。代码和模型可在 https://follow-your-pose.github.io/ 获取。

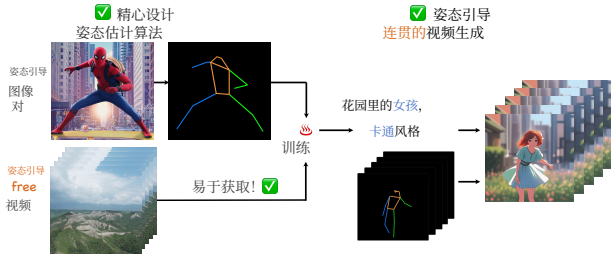


图1：训练方案的高层次概览。从公开的大规模带字幕图像数据集（Schuhmann 等人，2021）中，很容易获取包含不同角色的图像。我们可以利用现成的姿态估计算法，获得准确的姿态-图像对。然而，目前缺乏包含不同角色的带字幕视频数据集。因此，我们利用容易获取的野外视频来学习时间连贯性。通过这种分离式学习策略，我们的可调模块还能学习合成具有特定姿态运动、创意十足且多样的视频，同时保持真实的时间连贯性。

1 引言

近期，文本到图像（T2I）合成技术取得了长足进展（Rombach 等人，2022年；Ho 等，2022a年；Ma、Jia 和 Zhou，2023年），展现出令人瞩目的性能与创造力，使文本描述能够转化为视觉效果出众的作品。这类人工智能生成系统在多个方面表现出色，包括渲染真实且多样的外观（Mou 等人（2023）；Zhang 和 Agrawala，2023年；Ma 等人，2022年；Gao 等人，2023年）、强大的编辑能力（Hertz 等人，2022年；Qi 等人，2023年；Ma 等人，2023c年）、构图（Avrahami、Fried 和 Lischinski，2022年）以及泛化能力... ...

性，并创作满足人们想象的全新艺术作品（Mou 等人（2023）；Zhang 和 Agrawala，2023）。然而，在文本到视频生成领域，由于高质量视频数据集和视频生成先验模型的限制，其应用仍然有限。

在这项工作中，我们旨在根据文本描述生成高质量的角色视频，并通过给定的控制信号（即人体骨架）控制角色姿态。为此，我们设计了一种高效的训练方案，利用易于获取的图像数据集和预训练文本到图像模型，实现可控的文本到视频生成。具体而言，给定图像-姿态对和无姿态视频，我们基于预训练文本到图像模型（即稳定扩散（Rombach 等人，2022）），设计了一种新颖的两阶段训练策略，并对各模块进行了精心调优。第一阶段，我们通过额外的多层姿态编码器，为预训练文本到图像模型添加控制能力。第二阶段，我们使用无姿态视频训练模型，以学习时间一致性（例如连贯背景）并避免闪烁。为使预训练文本到图像模型适用于视频输入，我们进行了若干关键改动。首先，我们添加了

*贡献相同。

†在腾讯 AI 实验室实习期间完成的工作。

‡通讯作者。版权所有 © 2024，人工智能促进协会 (www.aaai.org)。保留所有权利。



图 2：姿态引导的文本到视频生成。我们提出了一种高效的训练方案，使预训练文本到图像模型（即稳定扩散（Rombach 等人，2022））能够以极少的数据生成姿态可控角色视频。得益于所提出的方法，我们能够生成各种高清姿态可控角色视频，使其与姿态序列和文本提示的语义高度一致。

额外的时间自注意力层（Singer 等人，2022；Wu 等人（2022）；Qi 等人，2023年；Ma 等人，2023a,b；He 等人，2023年；Li 等人，2023年）到稳定扩散网络中。其次，受近期单样本视频生成模型（Wu 等人（2022））的启发，我们将注意力重塑为跨帧注意力，以提升内容一致性。随后，我们针对文本到视频生成任务对视频模型进行微调，但不使用姿态控制。在第二阶段的训练过程中，我们只更新与时间一致性相关的参数，即新引入的时间自注意力层和跨帧注意力块，同时固定其他参数。由于姿态编码器确保了姿态控制，而视频数据集和精心设计的时间模块使模型能够学习时间信息，经过两个阶段的训练后，我们的方法能够生成姿态可控、外观可通过文本编辑的角色视频。

总体而言，我们提出的方法经过精心设计，能够通过姿态序列和文本描述灵活控制视频生成。此外，我们的模型继承了预训练文本到图像模型强大的概念生成与组合能力。因此，它能够生成多样化的角色、背景、吸引人的风格以及多个角色。尽管我们的重点是姿态控制，但正如 T2I adapter（Mou 等人（2023））和 ControlNet（Zhang 和 Agrawala，2023）等同期工作所展示的那样，我们的方法可以轻松扩展到其他控制模态（即深度、边缘、草图，例如）。

总而言之，我们的贡献如下：

- 我们提出了姿态可控的文本生成视频这一新任务，并引入 LAION-Pose 数据集，以促进姿态-内容对齐的学习。
- 为克服角色多样性不足这一挑战，我们将任务分解为两个子问题：姿态对齐和时间连贯性。随后，我们提出一种两阶段训练机制，通过在不同数据集上精心调整不同的参数集来实现。
- 与多种基线方法进行的大量实验表明，我们的方法在生成质量、文本-视频对齐、姿态-视频对齐和时间连贯性方面均表现出优势。

2 相关工作

2.1 文本到视频生成

由于涉及复杂且高维的结构特性，合成自然视频是一项颇具挑战性的任务。早期工作聚焦于生成对抗网络（GAN（Goodfellow 等人，2020年）），这种网络通过双人极小极大博弈，从高斯分布中生成样本。然而，基于 GAN 的框架训练不稳定，也难以对大规模数据集进行建模。得益于大语言模型（Radford 等人，2021年）和 Transformer（Vaswani 等人，2017年）的能力，近期有更多工作根据文本描述生成视频。例如，GODIVA（Wu 等人，2021年a）将 VQ-VAE（van den Oord、Vinyals 和 Kavukcuoglu，2018年）扩展到文本到视频生成，通过将文本词元映射为视频词元来实现。NUWA（Wu 等人，2021年b）提出了一种可用于文本到图像和视频生成任务的自回归框架。CogVideo（Hong 等人，2022年）通过引入时间注意力模块和预训练文本到图像模型，将 CogView-2（Ding 等人，2022年）扩展至文本到视频生成，从而提升了视频生成质量。随着基于扩散的方法日益普及，视频扩散模型（VDM）（Ho 等，2022b）采用一种因式分解的

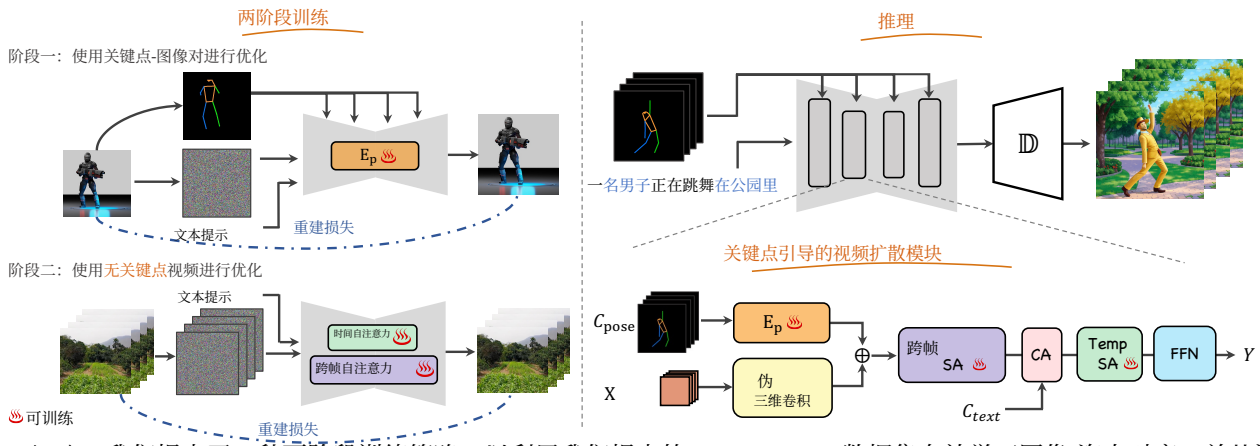


图3：概述。我们提出了一种两阶段训练策略，以利用我们提出的 LAION-Pose 数据集有效学习图像-姿态对齐，并从没有姿态标注的自然视频中学习时间连贯性。在第一阶段训练中，只有姿态编码器 E_p 可训练，以学习姿态控制。在第二阶段训练中，只有时间模块可训练，包括时间自注意力（SA）和跨帧自注意力。推理时，输入一段描述目标角色及其相应外观的文本，以及一段表示动作的姿态序列，即可生成时间连贯的视频。预训练稳定扩散模型的大部分参数均被冻结，包括伪三维卷积层、交叉注意力（CA）和前馈网络（FFN）模块。

时空 U-Net，直接在像素上执行扩散过程。Imagen Video（Ho 等，2022a）通过采用级联扩散模型和 v-预测参数化来增强 VDM。与 CogVideo 类似，Make-A-Video（Singer 等人，2022）将基于扩散的文本到图像生成模型（Rombach 等人，2022）的重大进展扩展到了文本到视频任务。与 Make-A-Video（Singer 等人，2022）类似，我们通过在现有图像模型中引入时间连接，将图像合成扩散模型扩展到视频生成。不同的是，我们以一种能够实现可控生成的方式，将动作与外观解耦。

2.2 姿态到视频生成

近年来，基于驱动信号（例如关键点）生成逼真人物视频已得到广泛研究，尤其是在基于 GAN 的模型（Mirza 和 Osindero，2014年）在条件视频合成方面取得前所未有的成功之后。Vid2vid（Wang 等人，2018年）提出了一种视频合成方法，利用条件生成对抗网络，结合光流和时间一致性约束以及多个判别器，生成既逼真又多样的姿态视频序列。Chan 等人（2019）也提出了类似的工作，通过关键点将源人物的动作迁移到目标人物身上，对两段舞蹈视频进行过拟合。然而，这些方法需要针对每一对视频分别训练。为了解决这一问题，fewshot-vid2vid（Wang 等人，2019年）将 vid2vid 扩展到少样本设置中，使新图像能够直接通过学习得到的模型或经少样本适配后进行动画生成。LiquidwarpingGAN（Liu 等人，2019年）提出了一种统一框架，利用液态变形和对抗学习来实现人体动作模仿。FOMM（Siarohin 等人，2019年）设计了一种一阶动作表示方法，能够驱动任意单幅图像，使其通过良好解耦的运动场实现动画生成。FRAA（Siarohin 等人，2021）通过一种针对可动角色的新型动作表示与合成框架扩展了 FOMM，该框架使用姿态依赖的运动嵌入。尽管这些方法能够生成或驱动与原始数据分布相似的图像，但在通过文本有效编辑视频方面仍存在不足。

2.3 可控扩散模型

由于根据文本描述生成视频或图像存在歧义，人们越来越关注条件图像合成。早期工作直接通过图像编辑方法生成可控结果（Meng 等人，2021）。近年来，已有多种条件策略被提出（Mou 等人（2023）；Zhang 和 Agrawala，2023）。这些方法会训练一个额外的编码器或适配器，将提取的信息（例如，深度图、关键点和分割图）映射为受控特征。然而，由于这些方法仍然是图像模型，将其应用于视频时，结果会出现闪烁问题。

3 方法

我们的文本引导姿态可控视频生成任务旨在根据姿态序列和文本中的外观描述生成逼真的视频。我们的方法基于预训练文本到图像模型，即潜在扩散模型（Rombach 等人，2022），并对其进行了若干修改以适应我们的任务。本节首先在第 3.1 节简要介绍潜在扩散模型，然后在第 3.2 节详细说明我们的姿态引导的可控文本到视频生成网络。

3.1 预备知识：潜在扩散模型

潜在扩散模型（LDM）（Rombach 等人，2022年）是一种扩散模型，用于对图像潜在空间的分布进行建模，

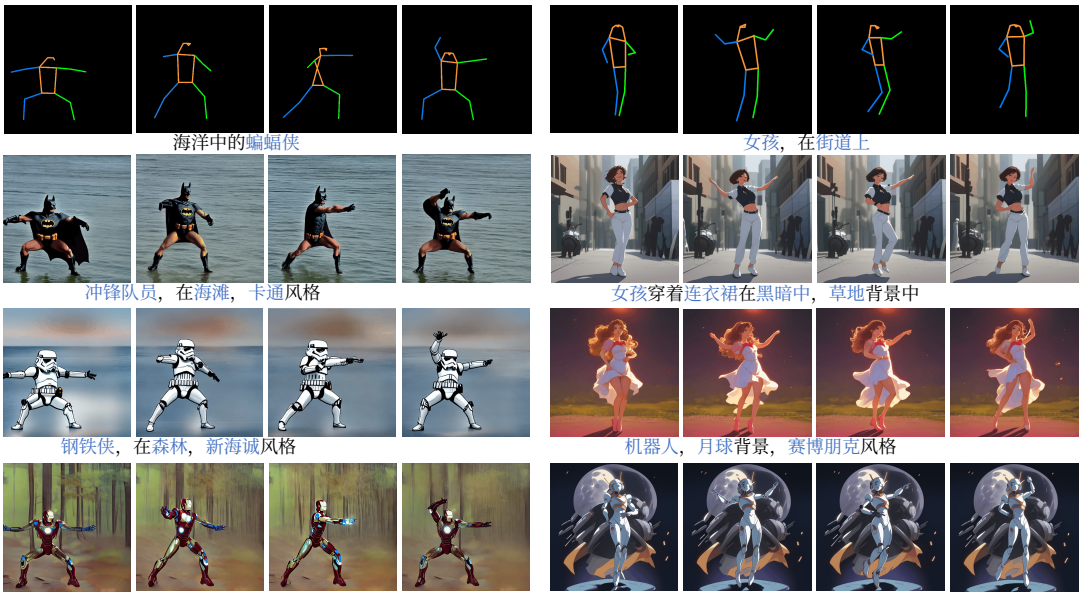


图 4：关于各种姿态序列和文本提示的结果。我们的方法能够遵循姿态引导和文本提示的语义，生成具有高内容多样性和时间连贯性的视频。

并且最近展现出卓越的图像合成性能。它由两个模型组成：一个自编码器和一个扩散模型。自编码器通过编码器 $\mathcal{E}$ 和解码器 $\mathcal{D}$学习重建图像。编码器首先将图像$\mathbf{x}$ 投影到一个较低维的潜变量：$\mathbf{z} = \mathcal{E}(\mathbf{x})$，解码器则从该潜变量重建原始图像：$\tilde{\mathbf{x}} = \mathcal{D}(\mathbf{z})$。

扩散模型通过 DDPM（Ho、Jain 和 Abbeel，2020年）学习图像潜变量 $\mathbf{z}_0 \sim p_{data}(\mathbf{z}_0)$ 的分布，并在潜在空间中生成新的样本。生成过程是一个包含T 个时间步的逐步反向去噪过程，从纯高斯噪声 $\mathbf{z}_t$ 开始，最终得到新样本 $\mathbf{z}_0$：

$$p_\theta(\mathbf{z}_{0:T}) := p(\mathbf{z}_T) \prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t), \tag{1}$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) := \mathcal{N}(\mathbf{z}_{t_1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)), \tag{2}$$

马尔可夫链通过预定义的噪声调度 $\beta_1, \ldots, \beta_T$进行逐步的正向加噪过程：

$$q(\mathbf{z}_{1:T}|\mathbf{z}_0) := \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1}), \tag{3}$$

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) := \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}). \tag{4}$$

在每个时间步，会从对角高斯分布中抽取随机噪声 $\boldsymbol{\epsilon}$，并训练时间条件去噪模型 θ，使其预测每个时间步中添加的噪声，使用简单的均方误差：

$$\mathcal{L}_{\text{simple}}(\theta) := \|\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t) - \boldsymbol{\epsilon}\|_2^2, \tag{5}$$

3.2 姿态引导的文本到视频生成

由于各种数据集中符合条件的视频-姿势对十分稀缺，我们决定将时间与控制条件解耦，而我们的模型从图像中学习姿态控制能力，并从视频中学习时间一致性。因此，我们分两个不同的阶段训练模型。如图 3 所示，我们的方法包含两个目的不同的训练阶段。下面，我们先简要介绍潜在扩散的去噪网络，然后详细说明各个阶段。

基础模型架构。广泛使用的图像合成扩散模型（Rombach 等人，2022）采用 U-Net（Ronneberger、Fischer 和 Brox（2015））进行去噪。U-Net 是一种多阶段神经网络架构，包含先进行空间下采样、再进行上采样的过程。U-Net（Ronneberger、Fischer 和 Brox（2015））的每个阶段都由多个注意力块和二维卷积残差块层组成。注意力块由空间自注意力、交叉注意力和前馈网络（FFN）构成。空间自注意力用于计算潜变量表示中各位置之间的相似相关性，而交叉注意力则考虑潜变量与条件输入（例如文本）之间的对应关系。

训练阶段 1：姿态可控的文本到图像生成。在此阶段，我们训练姿态可控的文本到图像模型。然而，现有方法没有用于姿态生成的姿态-图像-描述对。因此，我们使用 MMpose（Contributors 2020年）从 LAION（Schuhmann 等人，2021年）中收集人体骨架图像，仅保留检测到的关键点超过 50% 的图像。最终，我们构建了一个名为 LAION-Pose 的图像-文本-姿态数据集。该数据集包含多种类似人类的角色，以及各式各样的背景语境。为了将姿态条件融入预训练的文本到图像去噪模型，我们固定原始去噪 U-Net 的所有参数，并提出一种简单且轻量的方法，将多个三维卷积层用作

姿态编码器，并将其插入 U-Net 的每个模块中。具体而言，我们在每个模块中使用该卷积层，从输入姿势序列中提取姿态特征。随后，将姿态特征下采样至不同分辨率（即 $64 \times 64, 32 \times 32, 16 \times 16, 8 \times 8$）。如图 3 所示，我们通过残差连接将这些额外的控制信息注入预训练 U-Net，即将特征添加到 U-Net 模型的每一层中。这种残差注入方式的优点是能够保留预训练扩散模型的生成能力，只需更新少量参数即可添加额外的姿态控制信息。此外，如图 5 所示，只需将残差设为全零，即可轻松移除条件信息，恢复原始扩散模型。同时，我们还尝试了将姿态条件融入扩散模型的其他方式，例如将其与输入视频沿通道维度拼接，但实验结果表明其性能不如我们的方法。我们使用与原始稳定扩散相同的简单噪声重建损失来调整注入的参数。

训练阶段 2：通过无姿态视频生成视频。然而，第一阶段模型能够生成姿态相似的视频，但背景并不一致。因此，我们在第一阶段模型的基础上，使用无姿态视频数据集 HDVLIA（Xue 等，2022）进一步微调模型。该数据集包含连续的野外视频-文本对。为生成时间一致的视频，我们利用文生图模型的生成先验来进行姿态到视频生成。遵循先前的视频扩散模型基线方法（Ho 等，2022a；He 等，2022；Ho 等，2022b），我们将预训练 U-Net 的第一个卷积层扩展为 $1 \times 3 \times 3$ 个卷积核，作为用于视频输入的伪三维卷积层，并添加时间自注意力以进行时间建模（Carreira 和 Zisserman，2017）。为进一步保持时间一致性，我们利用帧间的跨帧自注意力（自注意力）（Wu 等人（2022））。不同的是，我们只需在 DDIM 的前一次采样过程中重复使用每个时间步的噪声，即可生成更长的视频序列。具体而言，假设每次采样 T 帧，首次采样后，我们将最后 $\frac{T}{2}$ 帧的噪声作为先验知识加入下一轮循环。请注意，在整个去噪过程中，
在每个时间步向先验知识中添加噪声，以确保生成的长视频具有时间一致性。如图 3 所示，所提出的方法仅对交叉帧自注意力和时间自注意力进行微调，用于视频生成。

讨论。得益于上述两阶段训练，以及为每个阶段精心设计的可调模块，所提出的方法能够利用易于获取的数据集生成连续姿态可控视频，例如人像与姿态图像对以及随机视频。正如同期的条件图像生成研究（Zhang 和 Agrawala，2023；Mou 等人（2023））所示，我们的方法也可应用于其他相关的条件视频生成任务。

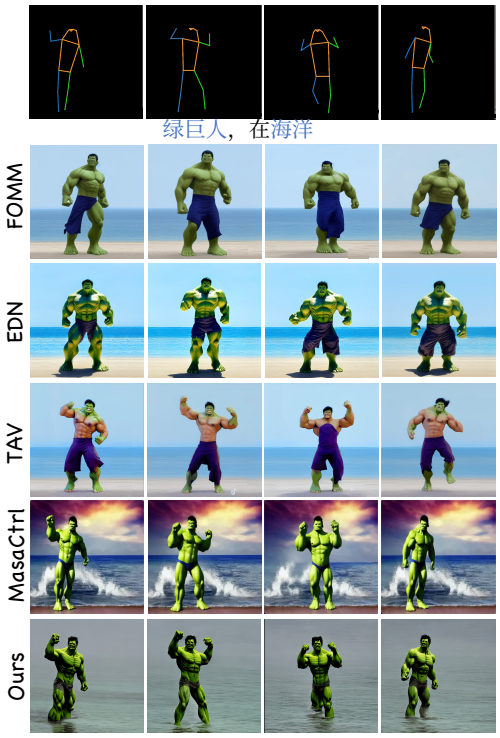


图 5：各评估方法的定性比较。对于 FOMM（Siarohin 等人（2019））和 EDN（Chan 等人（2019）），我们使用源视频驱动单张图像来生成结果。对于 TAV（Wu 等人（2022）），它从源视频中学习动作。在 MasaCtrl（Cao 等人（2023））中，我们利用 T2I 适配器有效控制动作。我们的方法能够在生成的视频片段中实现一致的背景和外观。

4 实验

4.1 实现细节

我们基于 Stable Diffusion 的官方代码库（Rombach 等人，2022）[1] 、公开可用的 14 亿参数文到图像模型[2] 以及 CivitAI 上的若干 LoRA 模型[3]实现了我们的方法。我们冻结图像自动编码器，分别将每个视频帧编码为潜在表示。我们首先在 Laion-Pose 上训练模型 10 万步，随后在 HDVILA（Xue 等，2022）上训练 5 万步。从输入视频中采样连续八帧，分辨率为 512×512，用于时间一致性学习。训练使用 8 块 NVIDIA Tesla 40G-A100 GPU，可在两天内完成。推理时，我们在姿态引导的文生视频生成实验中使用 DDIM采样器（Song、Meng 和 Ermon，2020），并采用无分类器引导（Ho 和 Salimans，2022）。

[1] https://github.com/CompVis/stable-diffusion. [2] https://huggingface.co/CompVis/stable-diffusion-v1-4. [3] https://civitai.com/

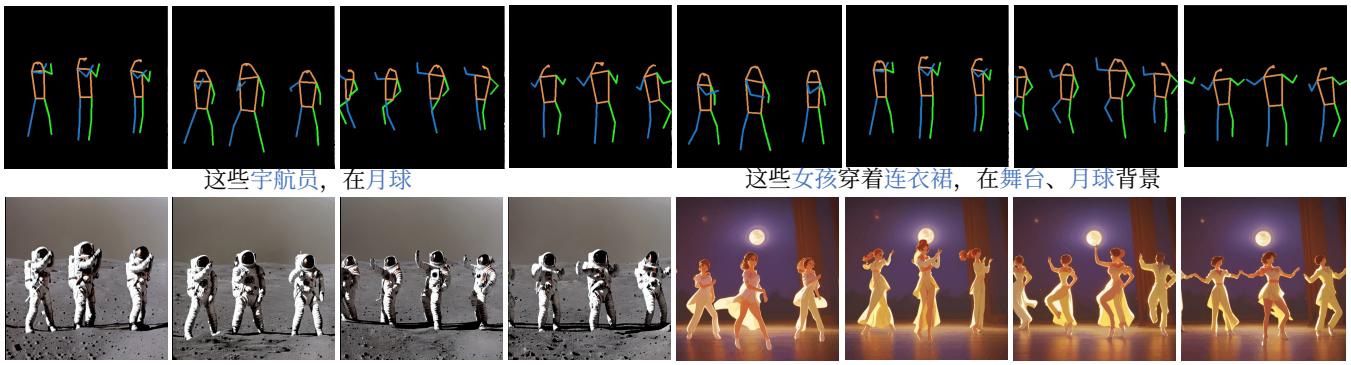


图4：关于多个骨架的结果。给定一个包含多个人物姿势的序列，我们的方法在这种情况下也能很好地泛化。此外，如第二行所示，可以通过输入的文本描述（例如“月球上的宇航员”）指定角色身份。借助这一应用，我们可以将无关角色引入同一个视频，同时指定背景语境。

4.2 应用

我们展示了该方法在生成具有新颖视觉概念（如角色、背景、风格等）且姿态由文本提示引导的视频方面的若干应用。这里，我们展示了两个不同的姿势引导训练示例及其通过编辑后的提示词生成的变体。更多示例见补充材料。

多样化的角色生成。我们方法的一项应用是在生成的视频中编辑角色，使其呈现多样的外观。从图4中可以看出，通过更改输入文本提示中的相应描述，我们的方法能够生成定制角色的视频。例如，给定一个输入姿势序列，我们可以通过改变输入文本提示，在保持动作相同的情况下生成“蝙蝠侠”“冲锋队员”和“钢铁侠”等角色。不同视频帧中的背景和角色外观都能成功保持一致。

多样化背景生成。我们的方法还能够生成自然视频背景（即，角色所在的地点），同时保持角色的姿势和时间信息一致。例如，我们可以合成“海洋”和“海滩”背景（图4左侧第2行和第3行）。与此同时，我们还能够生成“月球”和“草地”等特殊背景（图4右侧第3行和第4行）。生成的视频保持了较高的背景一致性，并能有效匹配姿势和文本提示。

风格化角色视频生成。 得益于预训练文本到图像模型和LoRA（Hu等人，2022）的广泛知识，我们的方法能够生成具有多种出色风格的视频。只需在句子中添加风格描述（例如，卡通风格、新海诚风格或赛博朋克风格），我们的方法就能生成具有相应风格的视频，同时保持姿势和语义一致（图4左侧第2行和第3行，以及右侧第3行）。

多角色生成。我们发现，我们的方法甚至能够将多个骨架转换为视频，
（例如，图 6 第二行的宇航员和女孩）。我们注意到，生成的视频特征与输入姿势描述和文本提示非常吻合，表明我们的方法具有强大的泛化能力。

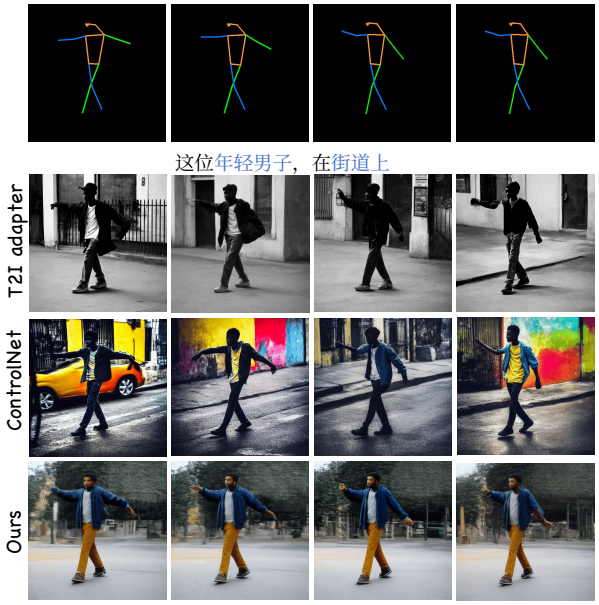


图 7：与 ***ControlNet***（***Zhang*** 和 ***Agrawala***，***2023***）和 ***T2I adapter***（***Mou*** 等人（***2023***））的比较。这两种方法都能将姿势转换为逼真的图像。然而，它们往往会生成外观各异的角色，并产生不同的背景。我们的方法则能在生成的视频片段中保持外观一致。

4.3 与基线方法的比较

我们将本文方法与两项同期工作进行了定量和定性比较，包括 Tune-A-Video（Wu 等人（2022））和 ControlNet（Mirza 和 Osindero，2014 年）。ControlNet 基于预训练的稳定扩散（He 等，2022），主要用于各种条件图像生成。它使用从互联网收集的包含 20 万个姿态-图像-描述对的大规模数据集进行了广泛训练。

定量结果。1）*CLIP* 分数：我们遵循（Ho 等，2022a），使用 CLIP 分数（Nguyen 等，2021；Park 等，2021）评估我们的方法，即评估视频-文本对齐情况。我们计算每一帧的 CLIP 分数，再对所有帧取平均值。最终的 CLIP 分数基于 1024 个视频样本计算。结果见表 1。我们的方法得分更高，表明其视频-文本对齐效果优于另外两种方法。

方法	CS	质量（%）	姿态准确率 (%)	帧一致性 (%)
FOMM（Siarohin 等人（2019））	22.93	0.8	11.7	81.25
Everybody Dance Now（Chan 等人（2019））	23.04	1.3	13.7	79.83
Tune-A-Video（Wu 等人（2022））	23.57	23.81	27.74	**93.78**
ControlNet（Zhang 和 Agrawala，2023）	22.31	6.69	33.23	54.35
T2I adapter（Mou 等人（2023））	22.42	8.27	33.47	53.86
MasaCtrl（Cao 等人（2023））	23.64	19.17	33.19	87.64
Ours	**24.09**	**39.96**	**34.92**	93.36

表 1：与相关工作的定量比较。CS、QU、PA 和 FC 分别表示 CLIP 分数、质量、姿态准确率和帧一致性。

2）质量：遵循 Make-A-Video（Singer 等，2022）的做法，我们对一个包含 32 个视频的测试集进行人工评估，以评价视频质量。具体而言，我们以随机顺序展示三个视频，并请评估者选出质量最高的视频。结果显示，评估者认为，与 Tune-A-Video 和 ControlNet 相比，我们的方法生成的视频质量更高。3）姿态准确度：我们将输入的姿态句子视为真实标注视频，并在 1024 个视频样本上评估骨架的平均精度。为确保公平比较，我们在处理 LAION-Pose 数据集收集流程和评估时采用相同的姿态检测器（Sun 等，2019）。结果表明，我们的模型取得了与 ControlNet 相当的性能。4）帧一致性：遵循（Esser 等，2023）的做法，我们报告以连续帧之间的 CLIP 余弦相似度衡量的帧一致性，结果见表 1。就时间一致性而言，我们的模型优于 ControlNet，这表明我们的时间设计确有必要。此外，我们的得分与 Tune-A-Video 相当。不过，Tune-A-Video 依赖输入视频，并且需要针对该输入视频过拟合，因此难以作为通用视频生成模型使用。

定性结果。我们首先在图 5 中，使用相同的姿态序列和文本提示，将我们的方法与其他四种方法进行比较。我们的方法在一致性和艺术性方面表现更佳。在类似的设置下，我们还在图 7 中将我们的方法与 ControlNet 和 T2I 适配器进行比较。显然可以发现，ControlNet 中存在不一致的背景（例如街道的颜色和男孩衬衫的颜色）。这种现象很常见，在图 7 的第 3 行中也可以观察到（墙壁颜色）。相比之下，我们的方法能够有效解决时间一致性问题，并学习到良好的帧间连贯性。

4.4 消融研究

残差姿态编码器的效果。我们对所提出的姿态编码器中的特征残差设计进行了消融实验。向扩散模型添加控制信息的一种自然方式，是将条件直接拼接到模型的噪声潜变量输入中（Rombach 等人，2022）。然而，我们发现，拼接方法的效果不如残差方法。图 8 第一行和第二行分别展示了通过残差方法和拼接方法注入姿态信息的生成结果。

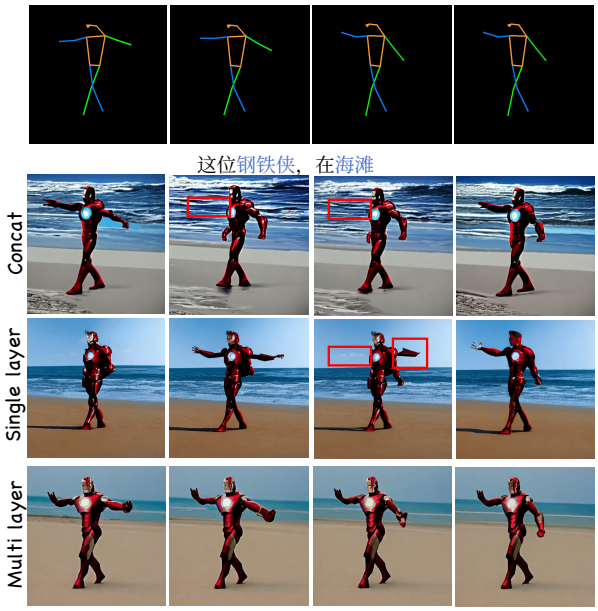


图 8：消融研究。包括条件注入机制（特征残差和逐通道拼接）以及带有条件信号的层数。

我们发现，以特征残差的方式添加额外控制信息，能够更多地保留预训练稳定扩散模型的生成能力。这是因为拼接机制需要重新训练第一个卷积层，以适配输入通道的数量，从而牺牲了预训练模型高质量图像合成的先验能力。

条件控制层数。我们还对注入控制信号的层数进行了消融，以实现更好的可控性（见图 8 第二行与第三行的对比）。结果表明，向更多层添加控制信号能够改善姿态帧对齐。具体而言，仅在单层中添加姿态会导致钢铁侠的目标手臂与生成的手臂不匹配。

5 结论

本文致力于解决生成可通过文本编辑且姿态可控的角色视频这一问题。为此，我们对预训练文本到图像模型进行改造和微调，利用其强大的语义编辑与组合能力。为此，我们设计了一种新颖的两阶段训练方案，充分利用大规模图像-姿态对和多样化的无姿态数据集。具体而言，在第一阶段训练中，我们使用姿态编码器将姿态信息注入网络结构，并从图像-姿态对中学习姿态可控的文本到图像生成。在第二阶段训练中，我们将图像模型扩展为三维网络，从无姿态视频中学习时间连贯性。借助多项全新设计，我们能够生成新颖且富有创意、时间连贯的视频，同时保留原始 T2I 模型的概念组合能力。

6 致谢

本项目获得中国国家重点研发计划资助，项目编号为2022ZD0161501；同时获得深圳市下一代交互媒体创新技术重点实验室资助（编号：ZDSYS20210623092001004）。

参考文献

Avrahami, O.; Fried, O.; and Lischinski, D. 2022. 混合潜在扩散. *arXiv* 预印本 *arXiv:2206.02779*. Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. MasaCtrl：无需调参的互自注意力控制，用于一致性图像合成与编辑. arXiv:2304.08465. Carreira, J.; and Zisserman, A. 2017. 动作识别，你往何处去？一种新模型与 Kinetics 数据集. In*CVPR*. Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. 此刻共舞. In*IEEE* 国际计算机视觉会议 *(ICCV)*. Contributors, M. 2020. OpenMMLab 姿态估计工具箱与基准. https://github.com/open-mmlab/ mmpose. Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. CogView2：通过层次化 Transformer 实现更快速、更优质的文本到图像生成. *arXiv* 预印本 *arXiv:2204.14217*, 2(3): 8. Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; and Germanidis, A. 2023. 通过扩散模型进行结构与内容引导的视频合成. *arXiv* 预印本 *arXiv:2302.03011*. Gao, K.; Bai, Y.; Gu, J.; Yang, Y.; and Xia, S.-T. 2023. 通过自适应拆分中毒数据集进行后门防御. In*IEEE/CVF* 计算机视觉与模式识别会议论文集, 4005–4014. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. 生成对抗网络. *ACM* 通讯, 63(11): 139–144. He, C.; Li, K.; Zhang, Y.; Zhang, Y.; Guo, Z.; Li, X.; Danell- jan, M.; and Yu, F. 2023. 策略性猎物造就敏锐捕食者：通过生成伪装目标增强伪装目标检测器. *arXiv* 预印本 *arXiv:2308.03166*. He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. 用于任意长度高保真视频生成的潜在视频扩散模型. *arXiv* 预印本 *arXiv:2211.13221*. Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. 使用交叉注意力控制进行提示到提示的图像编辑. *arXiv* 预印本 *arXiv:2208.01626*. Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen Video：通过扩散模型实现高清晰度视频生成. *arXiv* 预印本 *arXiv:2210.02303*. Ho, J.; Jain, A.; and Abbeel, P. 2020年. 去噪扩散概率模型. 神经信息处理系统进展, 33: 6840–6851. Ho, J.; and Salimans, T. 2022年. 无分类器扩散引导. *arXiv* 预印本 *arXiv:2207.12598*. Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. 视频扩散模型. *arXiv* 预印本 *arXiv:2204.03458*. Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022年. CogVideo：通过 Transformer 实现大规模文本到视频生成预训练. (3). Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022年. LoRA：大语言模型的低秩适配. 载于国际学习表征会议. Li, R.; Zhao, J.; Zhang, Y.; Su, M.; Ren, Z.; Zhang, H.; Tang, Y.; and Li, X. 2023年. FineDance：用于 3D 全身舞蹈生成的细粒度编舞数据集. 载于*IEEE/CVF*国际计算机视觉会议论文集, 10234–10243. Liu, W.; Piao, Z.; Min, J.; Luo, W.; Ma, L.; and Gao, S. 2019年. 液态变形生成对抗网络：用于人体动作模仿、外观迁移和新视角合成的统一框架. 载于*IEEE/CVF*国际计算机视觉会议论文集, 5904–5913. Ma, Y.; Cun, X.; He, Y.; Qi, C.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023a. MagicStick：通过控制手柄变换实现可控视频编辑. *arXiv* 预印本 *arXiv:2312.03047*. Ma, Y.; He, Y.; Cun, X.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023b. Follow Your Pose：利用无姿态视频进行姿态引导的文本到视频生成. *arXiv* 预印本 *arXiv:2304.01186*. Ma, Y.; Wang, Y.; Wu, Y.; Lyu, Z.; Chen, S.; Li, X.; and Qiao, Y. 2022年. 用于视频中人体动作推理的视觉知识图谱. 载于第*30*届 *ACM* 国际多媒体会议论文集, 4132–4141. Ma, Z.; Jia, G.; and Zhou, B. 2023年. AdapEdit：用于基于文本的连续性敏感图像编辑的时空引导自适应编辑算法. arXiv:2312.08019. Ma, Z.; zhihuan yu; Li, J.; and Zhou, B. 2023c. LMD：通过潜变量掩码扩散实现更快速的图像重建. arXiv:2312.07971. Meng, C.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021年. Sdedit：利用随机微分方程进行图像合成与编辑. *arXiv* 预印本 *arXiv:2108.01073*. Mirza, M.; and Osindero, S. 2014年. 条件生成对抗网络. *arXiv* 预印本 *arXiv:1411.1784*. Mou, C.; Wang, X.; Xie, L.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023年. T2I-Adapter：学习适配器以挖掘文本到图像扩散模型的更多可控能力. *arXiv* 预印本 *arXiv:2302.08453*. Nguyen, H.-L.; Kim, J.; Yeo, H.; Lee, D.; and Yoon, S.-E. 2021年. Clipscore：一种用于图像描述的无参考评估指标. 载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 3386–3395.

Park, D. H.; Azadi, S.; Liu, X.; Darrell, T.; and Rohrbach, A. 2021年. 组合式文本到图像合成基准. 载于神经信息处理系统第三十五届会议数据集与基准赛道（第一轮）. Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023年. FateZero：融合注意力以实现零样本文本视频编辑. arXiv:2303.09535. Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021年. 从自然语言监督中学习可迁移视觉模型. 载于国际机器学习会议, 8748–8763. PMLR. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Om- mer, B. 2022年. 利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. Ronneberger, O.; Fischer, P.; and Brox, T. 2015年. U-net：用于生物医学图像分割的卷积网络. 载于医学图像计算与计算机辅助干预——*MICCAI 2015*：第*18*届国际会议，德国慕尼黑，*2015*年*10*月*5*日至*9*日，论文集，第*III*部分*18*, 234–241. Springer. Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021年. Laion-400m：包含4亿个经 CLIP 筛选的图像-文本对的开放数据集. *arXiv* 预印本 *arXiv:2111.02114*. Siarohin, A.; Lathuili`ere, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019年. 用于图像动画的一阶运动模型. 神经信息处理系统进展, 32. Siarohin, A.; Woodford, O. J.; Ren, J.; Chai, M.; and Tulyakov, S. 2021年. 关节式动画的运动表示. 载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 13653–13662. Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022年. Make-a-Video：无需文本-视频数据的文本到视频生成.*arXiv* 预印本 *arXiv:2209.14792*, 2(3): 4, 7, 8. Song, J.; Meng, C.; and Ermon, S. 2020年. 去噪扩散隐式模型. *arXiv* 预印本 *arXiv:2010.02502*. Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019年. HRNet：用于人体姿态估计的深度高分辨率表示学习. 载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 5693–5703. van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2018年. 神经离散表示学习. 载于*NeurIPS*. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017年. 注意力机制就是你所需要的一切. 神经信息处理系统进展, 30. Wang, T.-C.; Liu, M.-Y.; Tao, A.; Liu, G.; Kautz, J.; and Catanzaro, B. 2019年. 少样本视频到视频合成. 载于神经信息处理系统会议（*NeurIPS*）. Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Liu, G.; Tao, A.; Kautz, J.; 和 Catanzaro, B. 2018年。《视频到视频合成》。载于神经信息处理系统进展（*NeurIPS*）。Wu, C.; Huang, L.; Zhang, Q.; Li, B.; Ji, L.; Yang, F.; Sapiro, G.; 和 Duan, N. 2021年a。GODIVA：根据自然语言描述生成开放领域视频。*arXiv* 预印本 *arXiv:2104.14806*。Wu, C.; Liang, J.; Ji, L.; Yang, F.; Fang, Y.; Jiang, D.; 和 Duan, N. 2021年b。N¨uwa：面向神经视觉世界创建的视觉合成预训练。Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Hsu, W.; Shan, Y.; Qie, X.; 和 Shou, M. Z. 2022年。Tune-A-Video：用于文本到视频生成的图像扩散模型单样本微调。*arXiv* 预印本 *arXiv:2212.11565*。Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; 和 Guo, B. 2022年。借助大规模视频转录稿推进高分辨率视频-语言表征。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，5036–5045。Zhang, L.; 和 Agrawala, M. 2023年。为文本到图像扩散模型添加条件控制。*arXiv* 预印本 *arXiv:2302.05543*。