StoryDiffusion：用于长距离图像与视频生成的一致性自注意力

周宇鹏[1]* 周大权[2§†] 程明明[1,3] 冯佳时[2] 侯淇彬[1,3§†]

[1] VCIP & TMCC, CS, 南开大学 [2] ByteDance Inc. [3] NKIARI, 福田, 深圳
https://StoryDiffusion.github.io

Abstract

对于近期基于Diffusion的生成模型来说，在一系列生成图像中保持内容一致性，尤其是包含主体和复杂细节的图像，是一个重大挑战。 在本文中，我们提出了一种简单但有效的自注意力机制，称为一致性自注意力，它能够提升生成图像之间的一致性。 它可以以零样本的方式用于增强预训练的基于Diffusion的文本到图像模型。 基于内容一致的图像，我们进一步展示了我们的方法可以通过引入一个语义空间时序运动预测模块，称为语义运动预测器，来扩展到长距离视频生成。该模块被训练用于在语义空间中估计两幅给定图像之间的运动条件。 该模块将生成的图像序列转换为具有平滑过渡和内容一致性的稳定视频，尤其是在长视频生成的情况下，其稳定性优于仅基于潜在空间的模块。 通过融合这两个新颖的组件，我们的框架称为StoryDiffusion，能够以一致的图像或视频描述基于文本的故事，涵盖丰富多样的内容。 所提出的StoryDiffusion 包含了通过图像和视频呈现视觉故事的开创性探索，我们希望它能从架构改进的角度激发更多相关研究。

1 引言

通过大规模预训练和先进架构，扩散模型在生成高质量图像和视频方面表现出优于以往基于生成对抗网络（GAN）方法的性能 [5]。 然而，现有模型在生成用于讲述故事的主体一致（例如身份和服装一致的角色）图像和视频方面仍然面临挑战。 常用的 IP-Adapter [55] 以一张图像作为参考，可用于引导扩散过程生成与之相似的图像。 但由于其强烈的指导作用，文本提示对生成内容的可控性有所降低。 另一方面，近期最先进的身份保持方法，如 InstantID [47] 和 PhotoMaker [26]，侧重于身份的可控性，但服装和场景的一致性无法得到保障。 因此，本文旨在寻找一种方法，能够生成在身份和服装两方面均保持一致的角色图像和视频，同时最大化用户通过文本提示的可控性。

在图像（或在视频生成背景下的帧）一致性保持任务中，常见的方法是引入时序模块 [15, 4]。 然而，这类方法通常需要大量的计算资源和数据。与此不同，我们致力于探索一种轻量级的方法，在最小数据和计算成本下，甚至以零样本的方式实现一致性保持。

已有研究表明 [45, 19, 6]，自注意力是建模生成视觉内容整体结构的关键模块之一。 我们的主要动机是：可以利用共享的参考图像信息来引导自注意力计算，从而显著提升生成图像之间的一致性。 由于自注意力权重依赖于输入，因此可能不需要额外的模型训练或微调。

*在ByteDance Inc.实习期间。 §项目负责人。 †通讯作者。

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

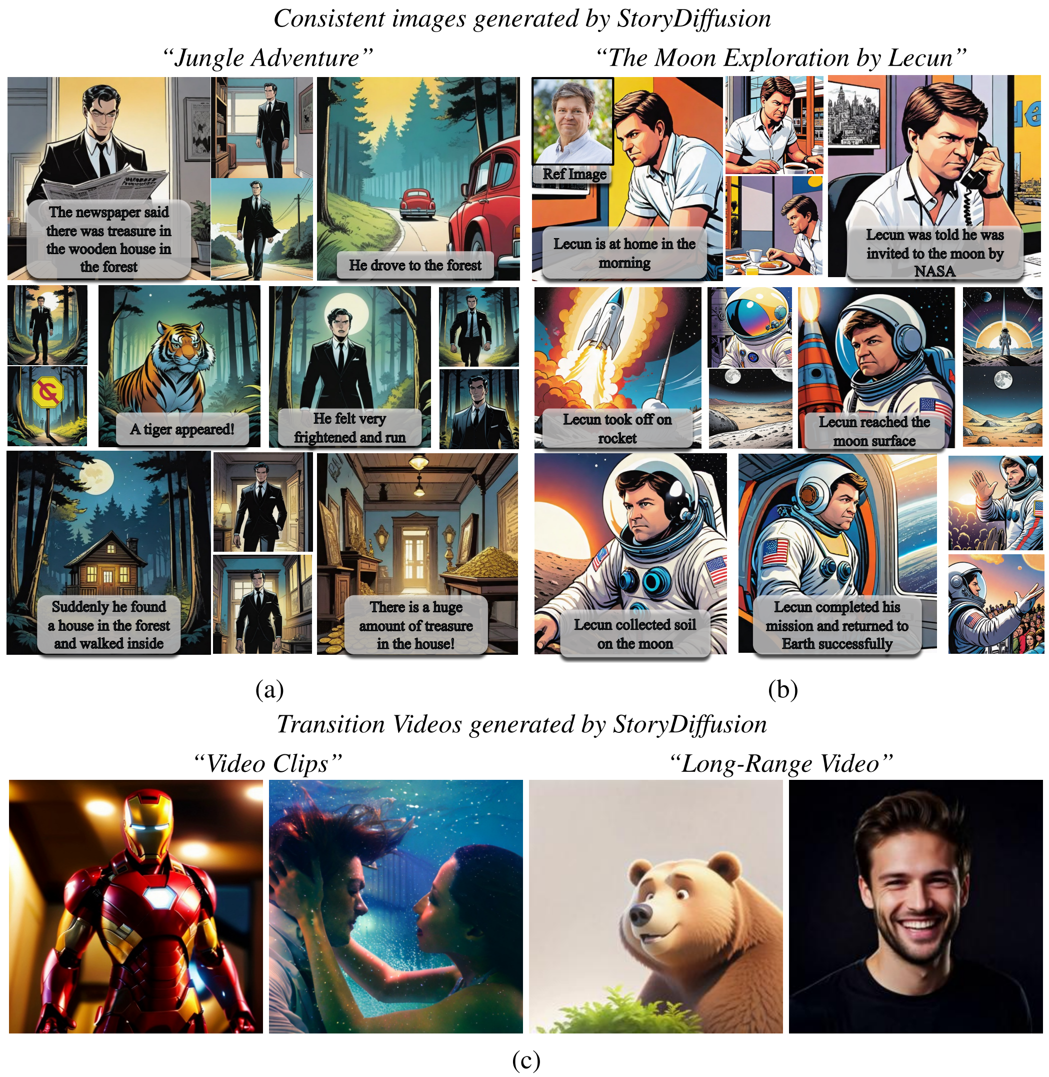

Consistent images generated by StoryDiffusion
"Jungle Adventure"
"The Moon Exploration by Lecun"
The newspaper said there was treasure in the wooden house in the forest
He drove to the forest
Ref Image
Lecun is at home in the morning
Lecun was told he was invited to the moon by NASA
A tiger appeared!
He felt very frightened and run
Lecun took off on rocket
Lecun reached the moon surface
Suddenly he found a house in the forest and walked inside
There is a huge amount of treasure in the house!
Lecun collected soil on the moon
Lecun completed his mission and returned to Earth successfully
(a)
(b)
Transition Videos generated by StoryDiffusion
"Video Clips"
"Long-Range Video"
(c)

Figure 1: 由我们的StoryDiffusion生成的图像和视频。 (a) 由StoryDiffusion生成的漫画，讲述一名男子在探索丛林时发现宝藏的故事。 (b) 由StoryDiffusion生成的漫画，描述Lecun登月探险的故事，采用了与Fig. 7(b)相同的参考图像控制方法 [26]。 (c) 由我们的StoryDiffusion生成的视频。点击图像即可播放视频。建议使用*Acrobat Reader*观看以获得最佳效果。 更多生成的视频可在上传的补充文件中查看。

基于这一思想，我们提出了一致性自注意力，它是我们StoryDiffusion的核心组件。该模块可以以零样本的方式插入到扩散模型主干中，用以替换原始的自注意力。

与标准自注意力仅在表示单张图像的token上进行操作不同，一致性自注意力在计算token相似度矩阵和进行token融合时，引入了从参考图像中采样的参考token。 这些采样的token与原始token使用同一组Q-K-V权重，因此无需额外训练。 如Fig. 1所示，使用一致性自注意力生成的图像在身份和服装两个方面都成功保持了一致性，这对于故事讲述至关重要。 直观来看，一致性自注意力在同一批次的图像之间建立关联，从而在身份和服装等方面生成角色一致的图像。 这使我们能够为故事讲述生成主体一致的图像。

对于任意给定的故事文本，我们首先将其划分为多个提示语，每个提示语对应一张独立的图像。 随后我们的方法能够生成高度一致的图像序列，有效地讲述完整的故事。 为了支持长篇故事的生成，我们还将一致性自注意力与滑动窗口机制结合，应用于已生成的一致性图像序列中。该设计消除了峰值内存对输入文本长度的依赖，使得长故事生成成为可

能。 为了将生成的故事帧流畅地转化为视频，我们进一步提出了语义运动预测器，该模块能够在语义空间中预测两幅图像之间的过渡。我们的实验证明，在语义空间中进行运动预测，比在图像潜在空间中预测更稳定。结合预训练的运动模块 [13]，语义运动预测器能够生成平滑自然的视频帧，整体表现显著优于近期的条件视频生成方法，如SEINE [7]和SparseCtrl [12]。

我们的主要贡献总结如下：

- 我们提出了一种无需训练且可热插拔的注意力模块，称为一致性自注意力。该模块能够在保持高度文本可控性的同时，维持故事图像序列中角色的一致性。
- 我们提出了一种新的运动预测模块，称为语义运动预测器，可在语义空间中预测两张图像之间的过渡。与现有流行的图像条件的方法（如SEINE [7]和 SparseCtrl [12]）相比，该模块能够生成更加稳定的长距离视频帧，并且易于扩展至分钟级的视频生成。
- 我们展示了结合一致性自注意力与语义运动预测器后的方法能够基于预设的文本故事生成长图像序列或视频，并通过文本提示指定运动内容。我们将这一新框架称为StoryDiffusion。

2 相关工作

2.1 可控的文本到图像扩散生成

作为扩散模型应用中的一个重要子领域 [40, 17, 36, 38, 27, 52, 41]，文本到图像生成 [37, 33, 34]近年来受到了广泛关注。 此外，为了增强文本到图像生成的可控性，已经涌现出大量相关方法。 其中，ControlNet [57] 和T2I-Adapter [31]引入了控制条件，如深度图、姿态图或素描图，用以指导图像生成过程。 MaskDiffusion [61]和StructureDiffusion [9]则侧重于提升文本的可控性。 还有一些工作 [30, 28]专注于控制生成图像的布局。

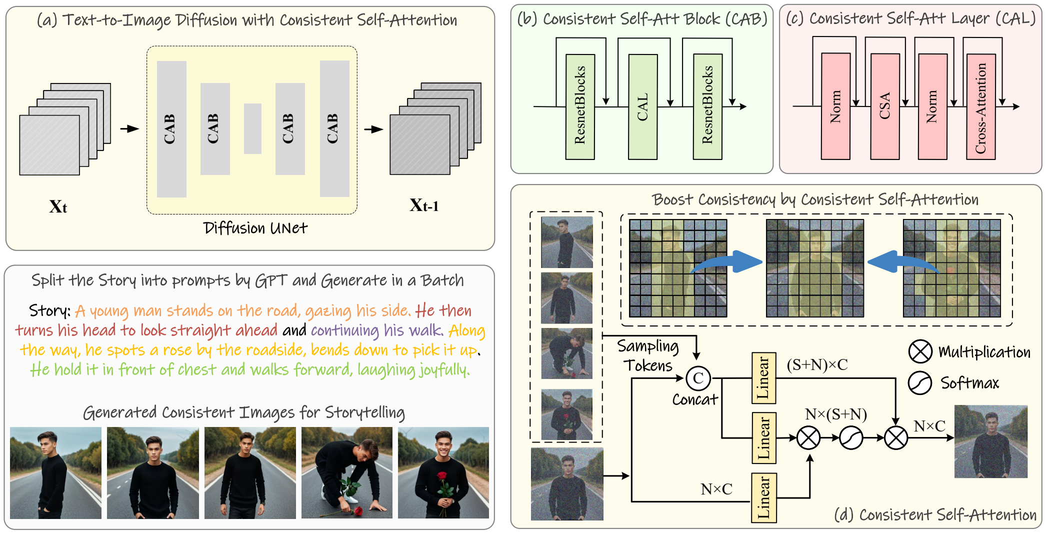
身份保持，即期望生成具有指定身份的图像，也是一个热门研究方向。 根据是否需要在测试阶段进行微调，这些工作可分为两大类。 第一类仅需使用给定图像对模型的部分参数进行微调，例如Textual Inversion [10]、DreamBooth [39]和Custom Diffusion [25]。 另一类方法以IPAdapter [55]和PhotoMaker [26] 为代表，利用在大规模数据集上预训练的模型，直接使用给定图像来控制图像生成。 与上述两类方法不同，我们关注的是在多张图像中保持主体一致性，以便讲述连贯的故事。 我们的一致性自注意力无需训练且可插拔，能够在同一批次的图像之间建立关联，从而生成多张主体一致的图像。

2.2 视频生成

由于扩散模型在图像生成领域的成功 [37, 17]，视频生成领域的探索 [13, 23, 42, 49, 54, 56]也日益受到关注。 VDM [15]是最早将图像扩散模型中的二维U-Net扩展为三维U-Net，实现基于文本的视频生成的工作之一。 后续的工作，如MagicVideo [60]和 Mindscope [46]引入了一维时序注意力机制，通过基于潜在扩散模型的方法降低计算量。 继Imagen之后，Imagen Video [16]采用了级联采样流程，通过多个阶段生成视频。

除了传统的端到端文本到视频（T2V）生成之外，基于其他条件的视频生成也是一个重要方向。 这类方法通过引入辅助控制条件生成视频，例如深度图 [12, 14]、姿态图 [53, 21, 48, 29]、RGB 图像 [3, 7, 32]，或其他引导运动视频 [59, 51]。

我们的视频生成方法聚焦于过渡视频生成，旨在根据给定的起始帧和结束帧生成连贯的视频。 典型相关工作包括SEINE [7]和SparseCtrl [12]。 SEINE在训练时随机遮蔽视频序列，作为视频扩散模型的初始输入，从而使模型能够预测两帧之间的过渡。 SparseCtrl则引入稀疏控制网络，利用稀疏的控制数据合成每一帧对应的控制信息，以此引导视频生成。 然而，


(a) Text-to-Image Diffusion with Consistent Self-Attention
CAB
CAB
CAB
CAB
Xt
Xt-1
Diffusion UNet
(b) Consistent Self-Att Block (CAB)
ResnetBlocks
CAL
ResnetBlocks
(c) Consistent Self-Att Layer (CAL)
Norm
CSA
Norm
Cross-Attention
Boost Consistency by Consistent Self-Attention
Sampling Tokens
Concat
Linear
(S+N)×C
N×(S+N)
N×C
Linear
N×C
Linear
Multiplication
Softmax
(d) Consistent Self-Attention
Split the Story into prompts by GPT and Generate in a Batch
Generated Consistent Images for Storytelling

Story: A young man stands on the road, gazing his side. He then turns his head to look straight ahead and continuing his walk. Along the way, he spots a rose by the roadside, bends down to pick it up. He hold it in front of chest and walks forward, laughing joyfully.

Figure 2: StoryDiffusion生成主体一致图像的流程。为了生成用于讲述故事的主体一致图像，我们将一致性自注意力集成到预训练的文本到图像扩散模型中。我们将故事文本拆分为多个提示语，并使用这些提示语批量生成图像。一致性自注意力在批量中的多张图像之间建立联系，以保证主体的一致性。

上述过渡视频生成方法仅依赖于图像潜在空间中的时序网络来预测中间内容。因此，这些方法在处理复杂过渡（例如角色的大规模移动）时表现较差。我们的StoryDiffusion旨在在图像的语义空间中进行运动预测，以实现更优的性能，并能应对更大幅度的运动变化，这一点将在实验部分进行展示。

3 方法

我们的方法可分为两个阶段，如Fig. 2和Fig. 3所示。第一阶段，StoryDiffusion利用一致性自注意力以无需训练的方式生成主体一致的图像。这些一致的图像既可直接用于故事讲述，也可作为第二阶段的输入。第二阶段，StoryDiffusion基于这些一致的图像生成连贯的过渡视频。

3.1 无需训练的一致性图像生成

解决上述问题的关键在于如何维持一批图像中角色的一致性。这意味着我们需要在生成过程中建立批量图像之间的联系。以往的图像和视频编辑方法 [6, 50]通过DDIM反演并在注意力计算中插入额外的键和值 [40]来保持相似性。不同于现有应用于单张图像或高度相似视频片段的方法 [24, 20, 11]，我们的目标是生成一组角色保持一致，但每张图像表现不同场景和动作的图像，用于动漫制作或故事板创作。因此，我们希望在一批图像中共享中间token，通过自注意力计算实现相互作用，从而保持一致性。我们通过在注意力计算前随机采样图像批次中的部分像素来获得这些中间token，实现了即插即用的功能，无需额外训练。我们将该操作命名为一致性自注意力，并将其插入到现有图像生成模型的U-Net架构中原始自注意力的位置，同时复用原有的自注意力权重。

形式上，给定一批图像特征$\mathcal{I} \in \mathbb{R}^{B\times N\times C}$，其中$B$、$N$和$C$分别表示批量大小、每张图像中的token数量和通道数，我们定义一个函数 Attention(X_k, X_q, X_v)用于计算自注意力。X_q、X_k和X_v分别代表注意力计算中的查询（query）、键（key）和值（value）。原始的自注意力是在批次中每个图像特征I_i内独立执行的。特征I_i被投影为Q_i、K_i、V_i并

输入注意力函数，得到：

$$O_i = \text{Attention}\left(Q_i, K_i, V_i\right). \tag{1}$$

为了在批次内的图像之间建立交互以保持主体一致性，我们的一致性自注意力会从批次中其他图像特征采样一些tokenS_i：

$$S_i = \text{RandSample}\left(I_1, I_2, .., I_{i-1}, I_{i+1}, ..., I_{B-1}, I_B\right), \tag{2}$$

其中，RandSample 表示随机采样函数。 采样后，我们将采样的令牌S_i与图像特征I_i配对，形成新的token集合P_i。 然后，我们对P_i进行线性投影，生成用于一致性自注意力的新键K_{Pi}和新值V_{Pi}。 这里，原始的查询Q_i保持不变。 最后，我们按如下方式计算自注意力：

$$O_i = \text{Attention}\left(Q_i, K_{Pi}, V_{Pi}\right). \tag{3}$$

给定配对的令牌，我们的方法在一批图像之间执行自注意力，促进不同图像特征之间的交互。 这种交互有助于模型在生成过程中实现角色、面貌和服饰的一致性收敛。 尽管方法简单且无需训练，我们的一致性自注意力能高效地生成主体一致的图像，这一点将在实验部分详细展示。 这些图像作为插图，用于讲述如Fig. 2所示的复杂故事。

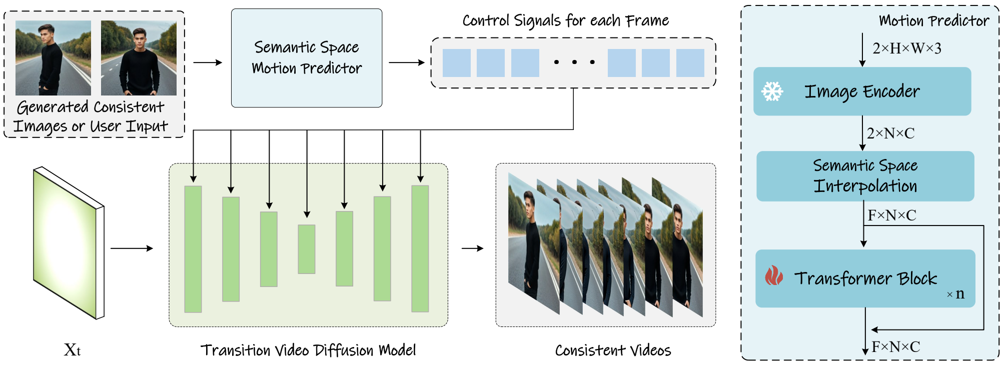

Generated Consistent Images or User Input
Semantic Space Motion Predictor
Control Signals for each Frame
Motion Predictor
2×H×W×3
Image Encoder
2×N×C
Semantic Space Interpolation
F×N×C
Transformer Block
×n
F×N×C
Xt
Transition Video Diffusion Model
Consistent Videos

Figure 3: 我们方法生成过渡视频的流程，用于获得主体一致的图像，如Sec. 3.1所述。 为了有效建模角色的大幅运动，我们将条件图像编码到图像语义空间以编码空间信息，并预测过渡嵌入。 随后，利用视频生成模型对这些预测的嵌入进行解码，嵌入作为交叉注意力中的控制信号，引导每一帧的生成。

3.2 用于视频生成的语义运动预测器

如Fig. 3所示，生成的角色一致图像序列可以通过在每对相邻图像之间插入过渡帧，进一步细化为视频。这可以视为一个已知起始帧和结束帧条件下的视频生成任务。 然而，我们通过实验观察到，近期的方法如SparseCtrl [12]和 SEINE [7]在两幅图像差异较大时，无法稳定地将这两幅条件图像连接起来。 我们认为，这一局限性源于它们仅依赖于时序模块来预测中间帧，而这可能不足以处理图像对之间的大的状态的差异。 时序模块在每个空间位置的像素上独立操作，因此在推断中间帧时可能没有充分考虑空间信息，导致难以建模长距离且具物理意义的运动。

为了解决该问题，我们提出了语义运动预测器，它将图像编码到图像语义空间中，以捕捉空间信息，从而实现从给定的起始帧和结束帧进行更精准的运动预测。 更具体地说，在我们的语义运动预测器中，我们首先使用一个函数E，将RGB图像映射到图像语义空间中的向量，以编码空间信息。 我们没有直接使用线性层作为E，而是利用预训练的CLIP图像编码器作为E，借助其零样本能力以提升性能。通过E，给定的起始帧F_s和结束帧F_e被压缩为图像语义空间向量K_s和K_e。

$$K_s, K_e = E\left(F_s, F_e\right). \tag{4}$$

随后，在图像语义空间中，我们训练了一个基于Transformer的结构预测器，用于预测每一帧的中间帧。该预测器首先对起始帧和结束帧的语义向量K_s和K_e进行线性插值，扩展成序列 $K_1, K_2, ..., K_L$，其中 L 表示所需的视频长度。接着，将序列$K_1, K_2, ..., K_L$输入到一系列Transformer块B中，以预测过渡帧：

$$P_1, P_2, ..., P_l = B\left(K_1, K_2, ..., K_l\right). \tag{5}$$

接下来，我们需要将这些在图像语义空间中预测的帧解码为最终的过渡视频。受到图像提示方法 [55]的启发，我们将这些图像语义嵌入$P_1, P_2, ..., P_L$作为控制信号，利用视频扩散模型作为解码器，从而发挥视频扩散模型的生成能力。同时，我们还插入额外的线性层，将这些嵌入投影为键和值，参与到U-Net的跨注意力计算中。

形式上，在扩散过程中，对于每一帧视频特征V_i，我们将文本嵌入T与预测的图像语义嵌入P_i进行拼接。跨注意力计算如下：

$$V_i = \text{CrossAttention}\left(V_i, \text{concat}(T, P_i), \text{concat}(T, P_i)\right). \tag{6}$$

与之前的视频生成方法类似，我们通过计算预测的L帧过渡视频$O = (O_1, O_2, ..., O_L)$与对应的L帧真实视频$G = (G_1, G_2, ..., G_L)$之间的均方误差（MSE）损失来优化模型：

$$Loss = \text{MSE}\left(G, O\right). \tag{7}$$

通过将图像编码到包含空间位置信息的图像语义空间中，我们的语义运动预测器能更好地建模运动信息，从而生成具有大幅运动的流畅过渡视频。显著的改进效果可见于Fig. 1和Fig. 6中的结果与对比。

4 实验

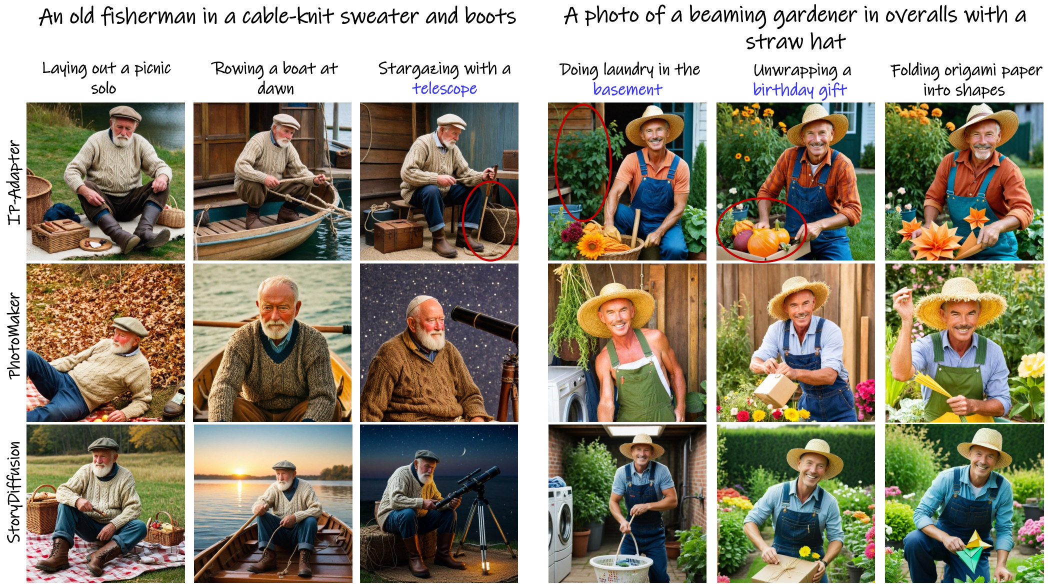
4.1 实现细节

关于生成保持主体一致性的图像，由于我们的方法具有无需训练和可插拔的特性，故在Stable Diffusion XL [34]和Stable Diffusion 1.5 [37]两个版本上均实现了我们的方案。为与对比模型保持一致，我们在Stable-XL模型及其相同的预训练权重上进行了对比实验。所有对比模型均采用50步DDIM采样[43]，CFG得分（classifier-free guidance score）[18]统一设置为5。

关于一致性视频的生成，我们的方法基于Stable Diffusion 1.5预训练模型实现，并引入了一个预训练的时序模块 [13]以支持视频生成。所有对比模型均采用7.5的CFG得分和50步DDIM采样。遵循之前的方法 [12, 7]，我们使用 Webvid10M [2]数据集训练过渡视频模型。训练过渡视频模型时，我们以AnimateDiff V2的运动模块 [13]作为时序模块的初始权重，并对其进行微调。学习率设为1e-4，在8块A100 GPU上训练了10万次迭代。为了将条件图像编码到图像语义空间，我们采用了预训练的OpenCLIP ViT-H-14模型 [35, 8]。我们的语义运动预测器包含8层Transformer，隐藏维度为1024，注意力头数为12。

4.2 一致性图像生成的对比

我们主要通过与两种最新的身份保持方法IP-Adapter [55]和PhotoMaker [55]进行对比，来评估我们方法在主体一致性图像生成方面的效果。为测试模型性能，我们使用GPT-4生成了二十条人物描述提示词和一百条活动描述提示词来描述特定活动。人物提示词的格式为“[形容词] [群体或职业] [穿着服饰]”，活动提示词的格式为“[动作] [地点或物体]”。我们将人物提示词与活动提示词组合，得到测试用的提示词组。对于每个测试用例，我们使用三种对比方法生成一组图像，这些图像描述同一个人在进行不同活动的情景，以测试模型在一致性方面的表现。由于IP-Adapter和PhotoMaker需要额外输入一张图像来控制生成图像的


An old fisherman in a cable-knit sweater and boots
Laying out a picnic solo
Rowing a boat at dawn
Stargazing with a telescope
A photo of a beaming gardener in overalls with a straw hat
Doing laundry in the basement
Unwrapping a birthday gift
Folding origami paper into shapes
IP-Adapter
PhotoMaker
StoryDiffusion

	Text Controllability	Face Consistency	Attire Cohesion
IP-Adapter	✗	✓	✓
PhotoMaker	✓	✓	✗
StoryDiffusion	✓	✓	✓

	Text Controllability	Face Consistency	Attire Cohesion
IP-Adapter	✗	✓	✓
PhotoMaker	✓	✓	✗
StoryDiffusion	✓	✓	✓

Figure 4: 与近期方法在一致性图像生成上的对比。

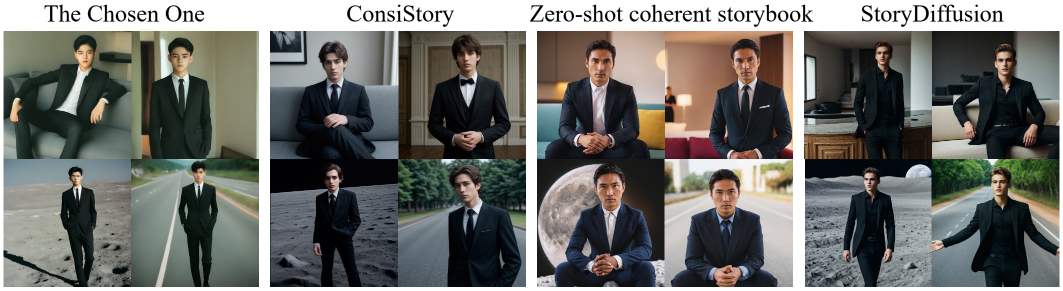


Figure 5: 我们的StoryDiffusion与近期的故事书生成方法的额外比较，包括The Chosen One [1]、ConsiStory [44]和Zero-shot coherent storybook [22]。

身份，我们首先生成一张角色图像作为控制图像。我们进行了定性和定量对比，从多个角度全面评估这些方法在一致性图像生成任务上的表现。

定性对比。 定性结果如Fig. 4所示。 我们的StoryDiffusion能够生成高度一致的图像，而其他方法（如IP-Adapter和PhotoMaker）则可能出现服装不一致或文本可控性减弱的问题。 在第一个示例中，IP-Adapter方法生成的图像中缺失了文本提示词“Stargazing with a telescope”中的 “telescope”。PhotoMaker虽然生成的图像符合文本提示，但三张图像中的服装存在明显差异。相比之下，我们的StoryDiffusion在第三行生成的图像在人脸和服装方面表现出高度一致性，同时具备更好的文本可控性。 在最后一个示例“A focused gamer wearing oversized headphones”中，IP-Adapter在第二张图像中缺失了“dog”，在第三张图像中缺失了“cards”。PhotoMaker生成的图像未能保持服装一致性。而我们的StoryDiffusion仍然能够生成主体一致的图像，保持相同的面部特征和服装，并符合提示语中的描述。为了进一步验证我们方法的有效性，我们在Fig. 5中将我们的方法与当前或近期的故事书生成方法进行了比较，包括The Chosen One [1]、ConsiStory [44]和Zero-shot Coherent Storybook [22]。我们的方法不仅在效果上优于这些方法，还具备更高的灵活性和更快的推

Table 1: 一致性图像生成的定量对比。我们的StoryDiffusion即使在无训练的情况下，也在文本相似度和主体相似度上取得了更好的表现。

Metric	IP-Adapter [55]	Photo Maker [26]	StoryDiffusion (ours)
Text-Image Similarity	0.6129	0.6541	**0.6586**
Character Similarity	0.8802	0.8924	**0.8950**

理速度。 相比之下，The Chosen One需要对每个样本进行耗时的LoRA自训练；Zero-shot Coherent Storybook采用两阶段流程，首先生成图像，然后通过迭代式身份一致性注入进行嵌入；ConsiStory则在扩散过程中需迭代计算分割掩码以保持一致性。

定量比较。 我们进行了定量比较，结果展示在Tab. 1中。我们评估了两个指标：第一个是文本-图像相似度，计算文本提示与对应图像之间的CLIP分数；第二个是角色相似度，通过对角色图像使用背景去除方法RMBG-1.4后计算CLIP分数来衡量。我们的StoryDiffusion在这两个定量指标上均取得了最佳表现，展示了本方法在保持角色一致性的同时，能更好地符合提示描述的稳健性。

4.3 过渡视频生成的对比

在过渡视频生成方面，我们与两种最先进的方法SparseCtrl [12]和 SEINE [7]进行了性能对比。 我们随机抽取了约1000个视频作为测试数据集。 对于给定的起始帧和结束帧，我们使用这三种对比模型预测过渡视频的中间帧，以评估它们的表现。

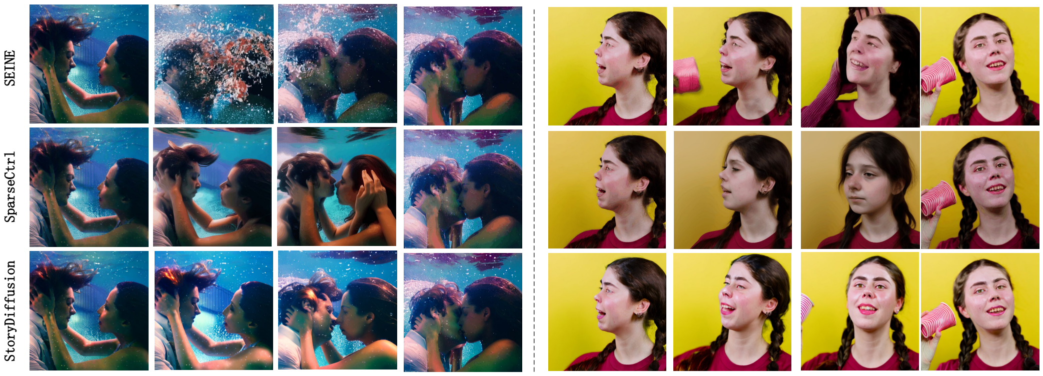


Figure 6: 与近期最先进的方法的过渡视频生成对比。

定性对比。我们进行了过渡视频生成的定性对比， 结果如Fig. 6所示。 我们的StoryDiffusion显著优于SEINE [7]和SparseCtrl [12]，生成的过渡视频流畅且有物理合理性。 第一个例子中，两人在水下亲吻，SEINE生成的中间帧出现损坏，且直接跳转到最终帧。SparseCtrl的结果连续性稍好，但中间帧仍包含损坏的图像，出现了大量多余的手部。相比之下，我们的StoryDiffusion成功生成了运动平滑且中间帧无损坏的视频。 第二个例子中，SEINE生成的中间帧手臂部分损坏；而SparseCtrl无法保持外观一致性。我们的StoryDiffusion生成了连续性极佳且一致性良好的视频。 最后一个例子中，我们生成的视频遵循物理空间关系，而SEINE和SparseCtrl仅在过渡中改变了外观。 更多视觉示例可见于Sec. A。

定量比较。 继之前的工作 [12, 58]，我们使用包括LPIPS-*f*、LPIPS-*a*、CLIPSIM-*f*和CLIPSIM-*a*在内的四个定量指标，将我们的方法与SEINE和SparseCtrl进行比较，如Tab. 2所示。 其中，LPIPS-*f*和CLIPSIM-*f*测量第一帧与其他帧之间的相似度，反映视频的整体连续性。 LPIPS-*a*和CLIPSIM-*a*测量相邻帧之间的平均相似度，反映帧与帧之间的

Table 2: 与最先进的过渡视频生成方法的定量比较。

Methods	LPIPS-*f* (↓)	LPIPS-*a* (↓)	CLIPSIM-*f* (↑)	CLIPSIM-*a* (↑)	FVD (↓)	FID(↓)
SEINE	0.4332	0.2220	0.9259	0.9736	321	140
SparseCtrl	0.4913	0.1768	0.9032	0.9756	429	181
Ours	**0.3794**	**0.1635**	**0.9606**	**0.9870**	**271**	**109**

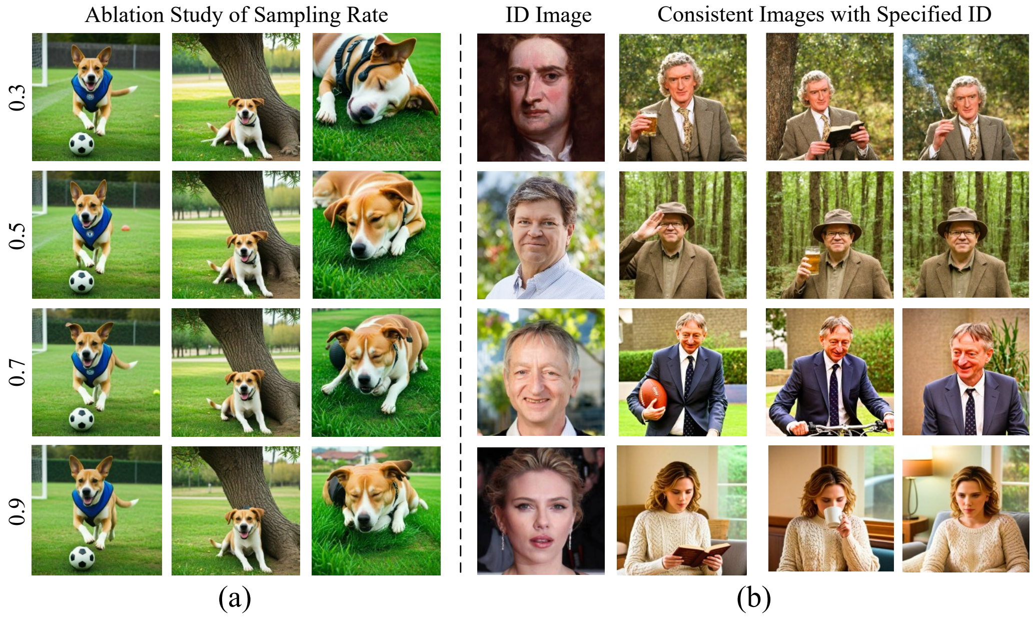

Ablation Study of Sampling Rate
ID Image
Consistent Images with Specified ID
0.3
0.5
0.7
0.9
(a)
(b)

Figure 7: 消融实验。(a) 评估了在一致性自注意力中不同采样率的影响。 (b) 我们探索了引入外部控制ID来控制角色生成。我们的StoryDiffusion能够生成与ID图像一致的连贯图像。

连续性。 此外，我们还计算了 FVD 和 FID 来评估生成质量。 我们的模型在所有四个定量指标上均优于另外两种方法。 这些定量实验结果证明了我们的方法在生成一致且平滑过渡视频方面的强大性能。

4.4 消融实验

Table 3: 对随机采样和网格采样中不同随机采样率的消融实验。

Sampling Method	Rand 0.3	Rand 0.5	Rand 0.7	Grid 0.5
Character Similarity	86.39%	88.37%	89.26%	**89.29%**
CLIP Score	**57.14%**	57.11%	56.96%	56.53%

用户指定**ID**生成。 我们进行了一个消融实验，以测试使用用户指定ID生成一致图像的性能。由于我们的一致性自注意力是可插拔且无需训练的，我们将其与PhotoMaker结合，利用图像来控制角色，实现一致的图像生成。结果如Fig. 7所示。 在ID图像的控制下，我们的StoryDiffusion依然能够生成与给定控制ID相符的一致图像，这充分体现了我们方法的可扩展性和即插即用能力。

一致性自注意力的采样率。我们的一致性自注意力从同一批次的其他图像中采样token，将其融合到自注意力机制中的键和值中。 随机采样的初衷是保持一致性的同时，避免过多结构信息的干扰，从而防止文本控制力减弱，并保持姿态的多样性。 我们进行了消融研究以

Table 4: 关于主体一致图像生成和过渡视频生成的用户研究。

Consistent Images Generation	IP-Adapter	PhotoMaker	StoryDiffusion (ours)
User Preference	20.8 %	10.9 %	**68.3 %**
Transition Video Generation	SEINE	SparseCtrl	StoryDiffusion (ours)
User Preference	5.9 %	9.6 %	**84.5 %**

寻找最佳采样率（结果见Fig. 7）。采样率为0.3时无法维持主体一致性，如Fig. 7左侧第三列图像所示，而更高的采样率则能保持一致性。定量分析表明，较高的采样率可能导致图像间过度关联，降低文本控制效果；而较低的采样率则会削弱角色一致性。我们还与网格采样方法进行了定量比较（见Tab. 3）。虽然网格采样更好地保持了角色一致性，但它牺牲了文本提示的可控性，因此我们可调的采样比例在两者之间实现了良好的平衡。在实际应用中，我们将采样率设置为0.5，既保证了一致性，又对扩散过程的影响最小。

4.5 用户研究

我们开展了一项包含79名参与者的用户研究。每位用户需回答50个问题，以评估我们主体一致图像生成方法和过渡视频生成方法的效果。在主体一致图像生成方面，我们与近期的最先进的方法IP-Adapter和PhotoMaker进行了比较；在过渡视频生成方面，则与近期的最先进的方法SparseCtrl和SEINE进行了对比。为了保证公平性，结果的展示顺序是随机的，且用户不知道每个结果对应的是哪种生成模型。如表Tab. 4所示，无论是主体一致图像生成还是过渡视频生成，我们的模型均展现出压倒性的优势。

5 总结

在本文中，我们提出了StoryDiffusion，一种能够以无训练方式生成一致图像用于讲故事，并将这些一致图像平滑过渡为视频的新方法。我们的一致性自注意力在多张图像之间建立联系，有效生成具有一致人脸和服装的图像。我们进一步提出了语义运动预测器，用于将这些图像过渡为视频，从而更好地叙述故事。我们希望我们的StoryDiffusion能够为未来的可控图像和视频生成研究带来启发。

更广泛的影响。我们的StoryDiffusion能够生成高质量且角色一致的图片和视频。当然，与以往的图像和视频生成方法类似，我们的方法也可能面临一些伦理问题。生成的人物肖像和视频可能被不当使用，例如用于制造虚假信息。我们强烈希望相关技术的使用能够明确责任，加强法律和技术监管，促进其规范和正确的应用。

致谢。本工作得到国家自然科学基金（编号：62225604）和中央高校基本科研业务费（南开大学，070-63223049）的资助。我们还感谢“科创甬江2035”关键技术突破计划项目（2024Z120）的支持。

References

[1] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 conference papers*, 2024. 7

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 6, 19

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint*, 2023. 3

[4] 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 *CVPR*, 2023. 1

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 1, 4

[7] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint*, 2023. 3, 5, 6, 8, 17, 23

[8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 6, 19

[9] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*, 2023. 3

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint*, 2022. 3

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

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint*, 2023. 3, 5, 6, 8, 17, 23

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *ICLR*, 2024. 3, 6

[14] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint*, 2023. 3

[15] J Ho, T Salimans, A Gritsenko, W Chan, M Norouzi, and DJ Fleet. Video diffusion models. arxiv 2022. *arXiv preprint*, 2022. 1, 3

[16] 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*, 2022. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint*, 2022. 6

[19] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *ICCV*, 2023. 1

[20] Susung Hong, Junyoung Seo, Heeseong Shin, Sunghwan Hong, and Seungryong Kim. Direct2v: Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint*, 2023. 4

[21] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint*, 2023. 3

[22] Hyeonho Jeong, Gihyun Kwon, and Jong Chul Ye. Zero-shot generation of coherent storybook from plain text story using diffusion models. *arXiv preprint*, 2023. 7

[23] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. *arXiv preprint*, 2023. 3

[24] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 4

[25] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[26] 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*, 2023. 1, 2, 3, 8

[27] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 3

[28] Wan-Duo Kurt Ma, JP Lewis, W Bastiaan Kleijn, and Thomas Leung. Directed diffusion: Direct control of object placement through attention guidance. *arXiv preprint*, 2023. 3

[29] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint*, 2023. 3

[30] Jiafeng Mao and Xueting Wang. Training-free location-aware text-to-image synthesis. *arXiv preprint*, 2023. 3

[31] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint*, 2023. 3

[32] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, 2023. 3

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint*, 2022. 3

[34] 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*, 2023. 3, 6, 19

[35] 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 *ICML*, 2021. 6, 19

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, 2022. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 6

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015*, 2015. 3

[39] 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 *CVPR*, 2023. 3

[40] 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. *NeurIPS*, 2022. 3, 4

[41] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2023. 3

[42] 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*, 2022. 3

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint*, 2020. 6

[44] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 2024. 7

[45] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv preprint*, 2023. 1

[46] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint*, 2023. 3

[47] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint*, 2024. 1

[48] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint*, 2023. 3

[49] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint*, 2023. 3

[50] 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 *ICCV*, 2023. 4

[51] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint*, 2023. 3

[52] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. *arXiv preprint*, 2022. 3

[53] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint*, 2023. 3

[54] Mengjiao Yang, Yilun Du, Bo Dai, Dale Schuurmans, Joshua B Tenenbaum, and Pieter Abbeel. Probabilistic adaptation of text-to-video models. *arXiv preprint*, 2023. 3

[55] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint*, 2023. 1, 3, 6, 8, 19

[56] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint*, 2023. 3

[57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3, 19

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8

[59] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint*, 2023. 3

[60] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint*, 2023. 3

[61] Yupeng Zhou, Daquan Zhou, Zuo-Liang Zhu, Yaxing Wang, Qibin Hou, and Jiashi Feng. Maskdiffusion: Boosting text-to-image consistency with conditional mask. *arXiv preprint*, 2023. 3

附录总览

我们提供了附录内容的总览, 同时我们还上传了一个视频文件夹。

- 在Sec. A中，我们提供了额外的实验结果，包括生成的视频和漫画，以及图像和视频生成的更多对比。
- 在Sec. B中，我们列出了局限性及未来工作方向。
- 在Sec. C中，我们提供了所使用的数据集和预训练模型的许可信息。

A 额外实验结果

A.1 Comic results

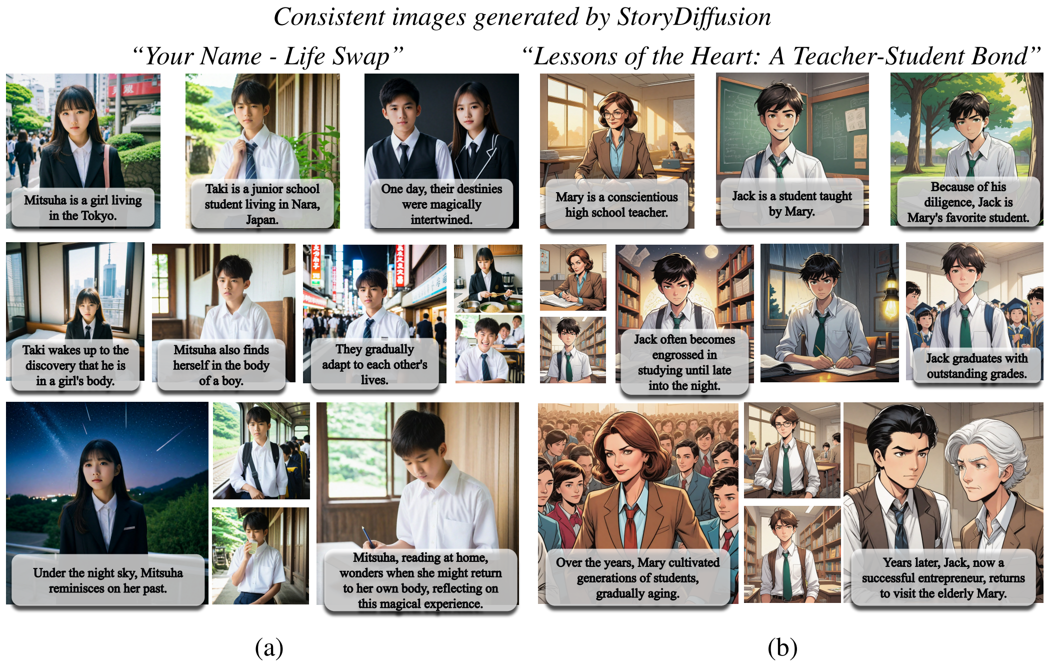
我们在上传的文件中展示了视频结果，并在Fig. 8中展示了更多漫画结果。这里补充呈现了额外的对比结果，以补充正文中的讨论，见Fig. 9和Fig. 10。

A.2 视频结果

我们还在补充材料中上传了视频文件，包含了由我们的方法生成的视频。上传的视频文件共计20个，这些视频基于我们的语义运动预测器，将由一致性自注意力生成的关键帧或来自视频的关键帧转化为连贯视频。所上传的视频分为两类：一类是较长且动作丰富的视频，另一类是较短且动作较少的视频。两类视频共同展示了模型对不同风格的处理能力。由于上传大小限制为100MB，我们对部分视频文件进行了压缩。

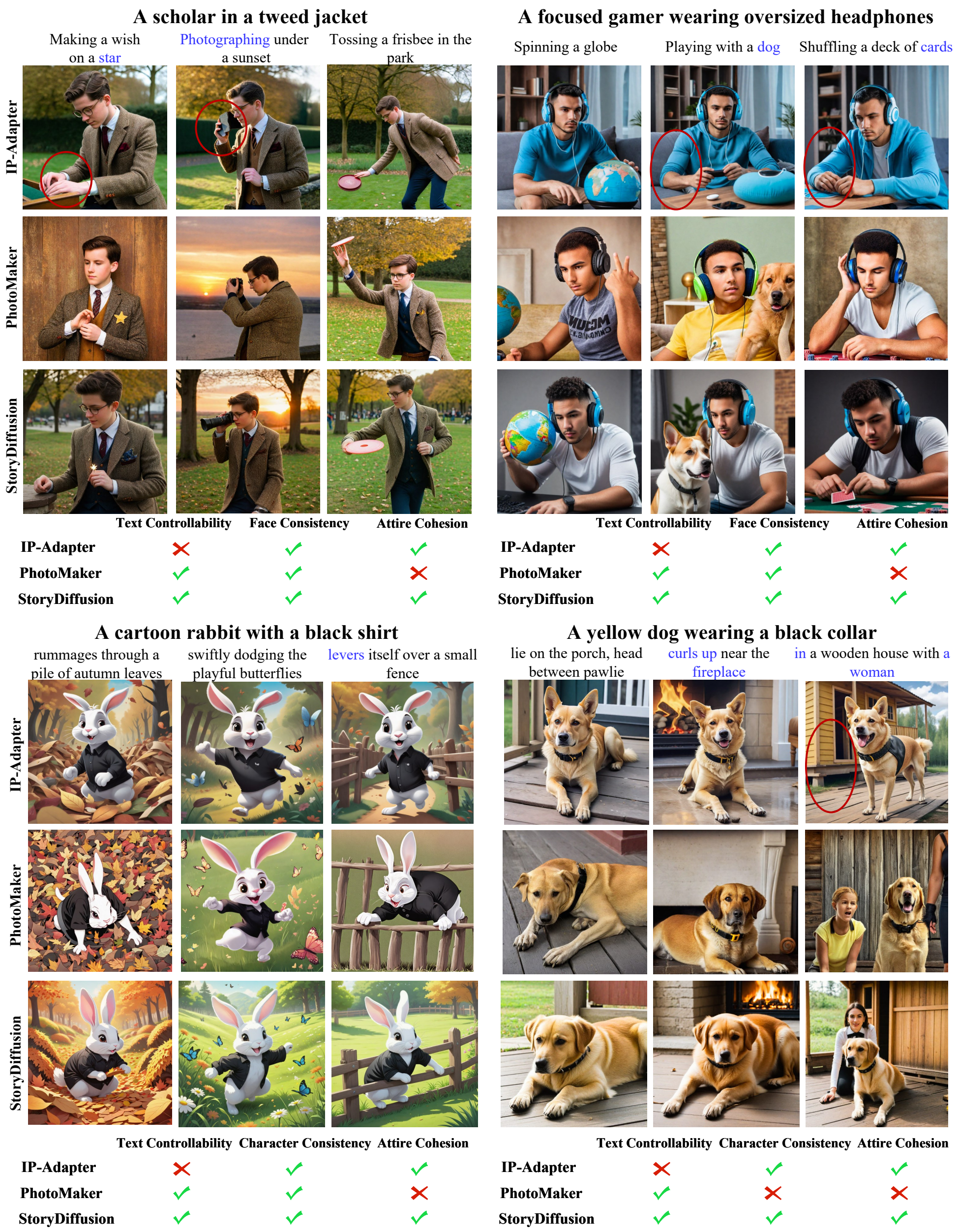
A.3 漫画结果

我们在Fig. 8中展示了更多漫画示例，讲述了两个引人入胜的故事，作为对Fig. 1 的补充。这些示例共同展示了我们方法在艺术创作中的实际应用价值。


Consistent images generated by StoryDiffusion
"Your Name - Life Swap"
Mitsuha is a girl living in the Tokyo.
Taki is a junior school student living in Nara, Japan.
One day, their destinies were magically intertwined.
Taki wakes up to the discovery that he is in a girl's body.
Mitsuha also finds herself in the body of a boy.
They gradually adapt to each other's lives.
Under the night sky, Mitsuha reminisces on her past.
Mitsuha, reading at home, wonders when she might return to her own body, reflecting on this magical experience.
"Lessons of the Heart: A Teacher-Student Bond"
Mary is a conscientious high school teacher.
Jack is a student taught by Mary.
Because of his diligence, Jack is Mary's favorite student.
Jack often becomes engrossed in studying until late into the night.
Jack graduates with outstanding grades.
Over the years, Mary cultivated generations of students, gradually aging.
Years later, Jack, now a successful entrepreneur, returns to visit the elderly Mary.

(a) (b)

Figure 8: 我们的StoryDiffusion生成的更多漫画结果。


A scholar in a tweed jacket
Making a wish on a star
Photographing under a sunset
Tossing a frisbee in the park
A focused gamer wearing oversized headphones
Spinning a globe
Playing with a dog
Shuffling a deck of cards
IP-Adapter
PhotoMaker
StoryDiffusion

	Text Controllability	Face Consistency	Attire Cohesion
IP-Adapter	✗	✓	✓
PhotoMaker	✓	✓	✗
StoryDiffusion	✓	✓	✓

	Text Controllability	Face Consistency	Attire Cohesion
IP-Adapter	✗	✓	✓
PhotoMaker	✓	✓	✗
StoryDiffusion	✓	✓	✓

A cartoon rabbit with a black shirt
rummages through a pile of autumn leaves
swiftly dodging the playful butterflies
levers itself over a small fence
A yellow dog wearing a black collar
lie on the porch, head between pawlie
curls up near the fireplace
in a wooden house with a woman
IP-Adapter
PhotoMaker
StoryDiffusion

	Text Controllability	Character Consistency	Attire Cohesion
IP-Adapter	✗	✓	✓
PhotoMaker	✓	✓	✗
StoryDiffusion	✓	✓	✓

	Text Controllability	Character Consistency	Attire Cohesion
IP-Adapter	✗	✓	✓
PhotoMaker	✓	✗	✗
StoryDiffusion	✓	✓	✓

Figure 9: 关于一致性图像生成的更多视觉对比。

A.4　一致性图像生成

我们在Fig. 9中展示了更多一致角色图像生成的结果。结果显示，一致性自注意力在文本一致性方面优于IP-Adapter。例如，在“A scholar in a tweed jacket”这一案例中，IP-Adapter 未能生成星星图案；在“A cartoon rabbit with a black shirt”中，IP-Adapter生成的图像未能正确表现翻越栅栏的动作；在“A yellow dog wearing a black collar”中，则未能生成女性角色且位置关系错误。PhotoMaker虽然未降低文本一致性，但无法保持生成角色的服装一致性，如前两个示例所示。此外，其在非人类角色上的表现有所下降，如后两个示例所示。与此不同，我们的一致性自注意力生成了文本一致性高且角色连贯性更强的结果。实验结果进一步验证了我们一致性自注意力的有效性。

A.5 过渡视频生成

我们在过渡视频生成方面，补充展示了与SparseCtrl [12]和 SEINE [7]的对比结果。如Fig. 10所示，相较于SparseCtrl和SEINE，我们的语义运动预测器能生成更连贯且平滑的中间帧，进一步彰显了语义运动预测器的优势。

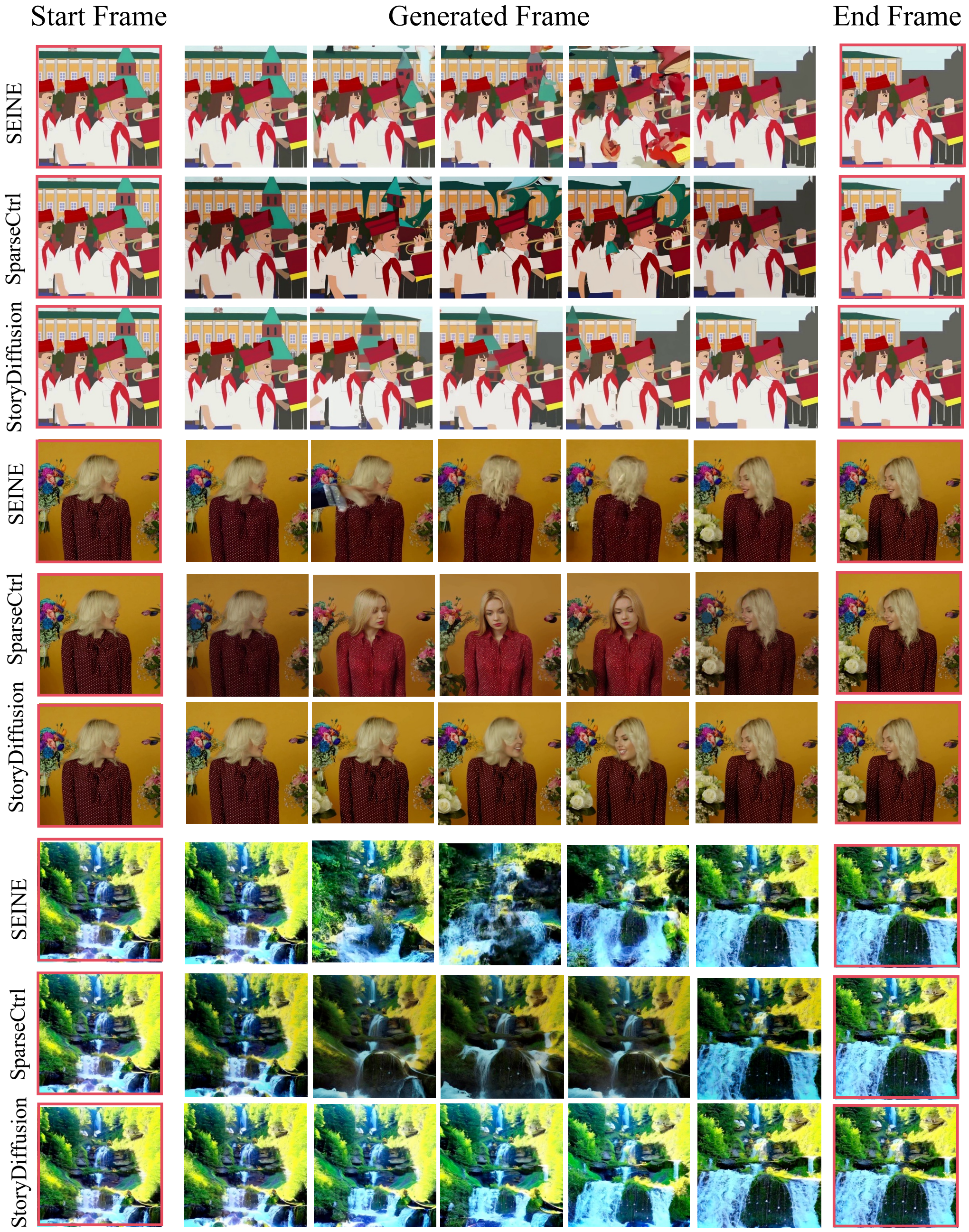
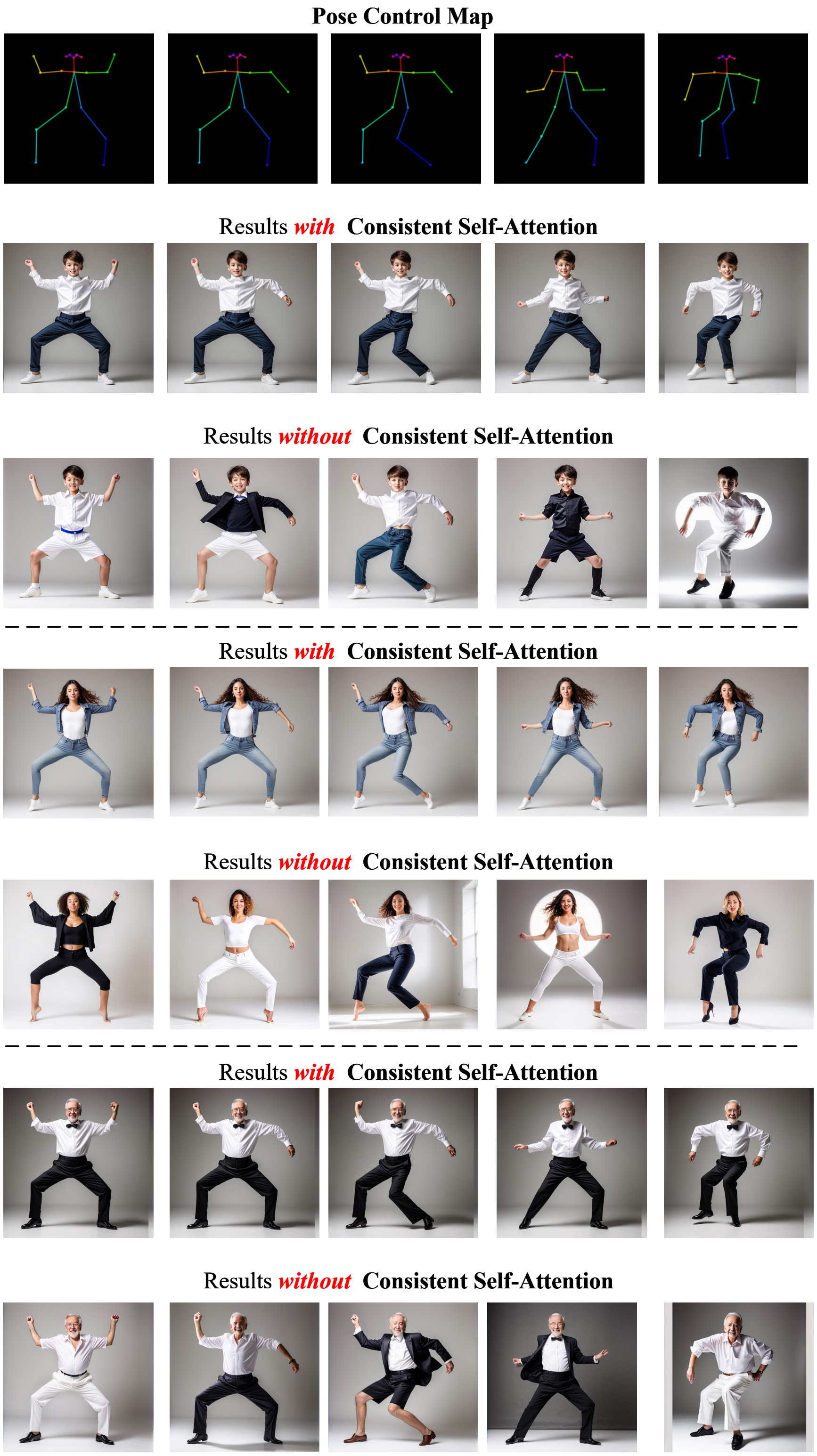


Figure 10: 过渡视频生成的额外视觉对比。红色框表示输入模型的帧。



Pose Control Map

Results *with* Consistent Self-Attention

Results *without* Consistent Self-Attention

Results *with* Consistent Self-Attention

Results *without* Consistent Self-Attention

Results *with* Consistent Self-Attention

Results *without* Consistent Self-Attention

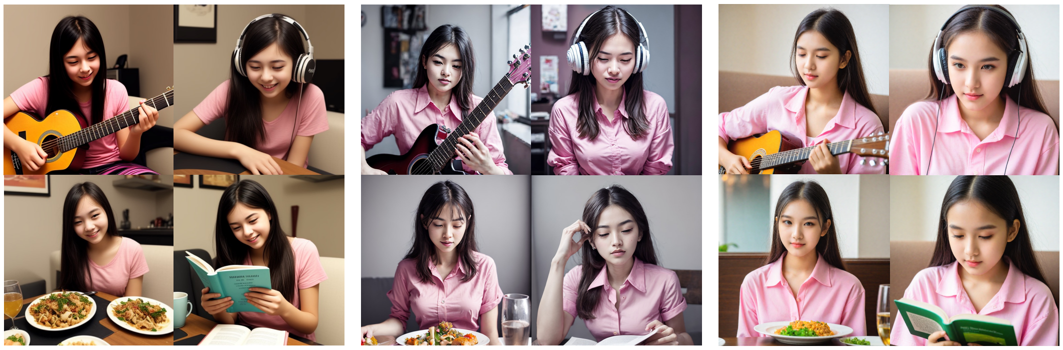
Figure 11: 我们将一致性自注意力与ControlNet结合后的生成结果。

A.6 结合ControlNet的一致性图像生成

鉴于我们的一致性自注意力无需训练且可插拔，我们进一步探索将其与 ControlNet [57]结合，以在生成主体一致图像时引入姿态控制。结合一致性自注意力与ControlNet的结果展示于图 Fig. 11。我们的方案同样能够在ControlNet的指导下生成主体一致的图像。

A.7 即插即用能力实验

我们还在SD 1.5和SD 2.1上实现了我们的StoryDiffusion，并将结果与SDXL进行了比较，见图 Fig. 12。实验表明，我们的方法在集成到不同模型时仍能保持良好性能，展示了其“即插即用”的特性。



StoryDiffusion based SD 1.5　StoryDiffusion based SD 2.1　StoryDiffusion based SDXL

Figure 12: 我们的StoryDiffusion在多款主流扩散模型上的可插拔能力，一致性自注意力在多个模型中均能良好运行。

B 局限性与未来工作

我们的方法存在的第一个局限是在主体一致图像生成方面。 类似于现有的最先进的方法 [55]，在一些细节服饰（如领带）上可能仍存在不一致的情况。 对此，我们的一致性自注意力可能需要更为详尽的文本提示，以保持图像间的一致性。 第二个局限出现在过渡视频生成上。虽然可以通过顺序连接生成的一致图像来生成较长的视频，但当两张图像差异较大时，拼接变得具有挑战性。 因此，由于缺乏全局信息交互，我们的方法尚未能完美支持超长视频的生成。 未来，我们将进一步探索长视频生成的技术。

C 数据集和模型许可

Webvid-10M: Webvid-10M [2]是一个大规模视频数据集，包含1000万段带有文本描述的视频片段，旨在用于训练和评估视频理解与生成任务的机器学习模型。

网址: www.robots.ox.ac.uk/~vgg/research/frozen-in-time/

Stable XL: Stable XL [34]是由Stability AI提供的基于扩散模型的文本生成图像模型，能够根据输入文本生成高质量图像。

许可: CreativeML Open RAIL++-M License. 网址: stability.ai/stable-image

OpenCLIP: OpenCLIP [8] 是OpenAI提出的CLIP模型 [35]（预训练对比语言-图像）的开源实现。

网址: github.com/mlfoundations/open_clip

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We confirm the main claims reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation is discussed in Sec. B.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We did not include theoretical results in our paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have described the model details in the implementation details in Sec. 4.1 and carefully described the experimental evaluation in the experimental chapter. The information we provide is sufficient and detailed for replication purposes.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We have made a detailed description of the implementation details to ensure that they are repeatable and use publicly available datasets. However, we intend to make our code publicly available following the paper's acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have carried out a detailed narration in the implementation details in Sec. 4.1

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Based on our experimental experience, the reproducibility of the experiments involved in this work is high, with results that are replicable and stable, rather than simply reporting the highest outcomes. Additionally, previous related work [7, 12] has also not reported error bars. We thus do not run the statistical significance test.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We state this detailed information of computer resources in Sec. 4.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We confirm that the research involved in the article complies with the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper have conducted a discussion of broader impacts at Sec. 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: At present, there is no relevant description we will set up safeguards when we release the model of StoryDiffusion.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We state dataset and model license in Sec. C.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce new assets in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.