arXiv:2602.01303v1 [cs.CV] 1 Feb 2026

ReDiStory：用于一致性视觉故事生成的区域解耦扩散

Ayushman Sarkar[1], *Zhenyu Yu*[2,*], *Chu Chen*[3], *Wei Tang*[3], *Kangning Cui*[3], *Mohd Yamani Idna Idris*[2]

[1] Birbhum工程技术学院， [2] 马来亚大学，[3] 香港城市大学， [4] 维克森林大学

摘要

生成连贯的视觉故事，需要在多张图像中保持主体身份，同时保留帧特定语义。近期免训练方法将身份与帧提示拼接为统一表示，但这往往会引入帧间语义干扰，削弱复杂故事中的身份保持。我们提出*ReDiStory*，这是一种免训练框架，通过推理时的提示嵌入重组来改进多帧故事生成。*ReDiStory* 显式地将文本嵌入分解为身份相关成分和帧特定成分，随后通过抑制帧间共享方向来去相关化帧嵌入，从而在不修改扩散参数或需要额外监督的情况下减少跨帧干扰。在扩散主干网络和推理设置相同的条件下，*Re-DiStory* 提升了身份一致性，同时保持提示保真度。在 ConsiStory+ 基准上的实验表明，与 1Prompt1Story 相比，该方法在多项身份一致性指标上均取得稳定提升。代码见：
https://github.com/YuZhenyuLindy/ReDiStory

索引词— 免训练生成、视觉叙事、主体一致性、提示嵌入

1. 引言

文本到图像（T2I）扩散模型 [1–3]能够根据文本描述合成高质量图像，并已成为可控图像生成[4–6]的标准基础模型。除单图生成外，视觉叙事旨在生成一系列图像，在不同场景和动作中反复描绘同一主体 [7]。这种能力对于插画故事书、动画预可视化和序列化内容创作至关重要 [7, 8]。

如图 1 所示，身份漂移的一个关键来源是提示表示中身份相关属性与帧特定指令之间的纠缠。核心挑战是在不牺牲帧特定语义的前提下，跨帧保持主体身份 。当各帧独立生成时，扩散模型往往会出现

身份漂移和属性不一致性。现有解决方案

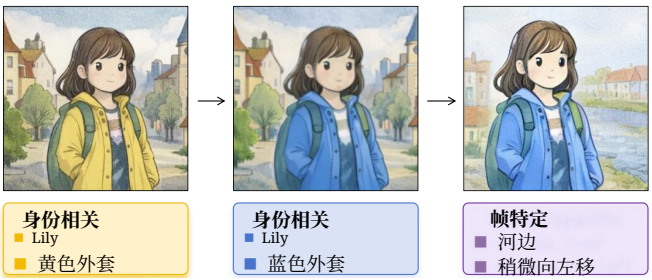


图 1：**动机**。身份相关成分（例如姓名和外观）可能会与帧特定指令（例如位置和姿势）纠缠在一起，导致不同帧中的身份发生意外变化。因此，我们将这两类成分分离，以实现更一致的多帧生成。

通过主体驱动微调 [9, 10]，（这需要重新训练并使用额外数据）来解决这一问题，或者采用无需训练的方法，在推理时间操控注意力或提示结构 [11, 12]。后者提供了一种轻量且与模型无关的替代方案，更适合实际的多帧生成。

在免训练方法中，1Prompt1Story [13]将身份提示与所有帧提示拼接成一个输入序列，利用文本编码器的上下文连贯性。尽管这种方法有效，但这种拼接策略也会引入帧间语义干扰：当帧提示被联合编码时，它们在嵌入空间中的表示可能会产生相关性；因此，帧特定语义可能会部分覆盖共享身份线索，导致主体外观在不同帧之间不一致。

我们提出了 *ReDiStory*，这是一种无需训练的提示嵌入重组框架，可缓解帧间语义干扰。*ReDiStory*仅在推理时对文本嵌入进行操作。它首先将提示嵌入分解为身份相关成分和帧特定成分，然后重新组织帧嵌入，以抑制帧间相关性，而无需修改扩散模型或引入额外监督。在扩散骨干网络和推理设置相同的情况下，*ReDiStory* 提升了身份一致性，同时保持提示保真度，取得了持续优于基线方法的效果。我们的贡献主要有三点：

• 我们指出，帧间语义干扰 是多帧生成中身份漂移的一个关键原因。

• 我们提出*ReDiStory*，这是一种推理时间提示嵌入重组框架，无需重新训练即可将身份与帧特定语义解耦。

• 我们在相同设置下使用 ConsiStory+ 基准测试评估 *ReDiStory* ，结果表明，它在提升身份一致性的同时不会降低提示保真度。

2. 相关工作

视觉叙事要求在允许场景发生大幅帧特定变化的同时，保持主体身份的一致性。现有方法大致可按其强制保持一致性的方式分为三类：（i）调整模型或注入额外的身份线索（身份一致性强，但前提条件更多）；（ii）保持模型不变，并在推理时协调多帧生成（轻量，但容易出现漂移）；（iii）依赖提示设计，其中拼接可能无意间将帧语义耦合在一起，导致身份线索被覆盖。我们回顾这三类方法，以阐明我们的研究重点：仅在提示层面、无需训练地缓解跨帧语义干扰。

2.1. 主体一致的文本到图像生成。

在多张生成图像之间保持主体一致性，已在文本到图像生成中得到广泛研究 [3, 6]。基于训练的个性化方法，如 DreamBooth [9] 和文本反演 [10], ，利用额外数据将扩散模型适配到特定主体。这些方法通常能提供较强的身份保持，但需要逐主体优化，并且当故事涉及多样的场景、姿势或属性时，可能会降低灵活性。近期方法，包括IP-Adapter [14] 和PhotoMaker [15]„ 使用参考图像来引导身份定制，但仍依赖辅助输入或模型适配。

2.2. 免训练多帧叙事。

为避免重新训练，免训练方法仅在推理时间运行，因此非常适合使用现成的扩散模型进行多帧叙事。ConsiS-tory [11] 通过批量生成间的共享注意力传播主体信息，而 StoryDiffusion [16]则引入一致性自注意力，以稳定长程叙事生成。与我们的方法最为接近的是 1Prompt1Story [13] ，它将身份与所有帧提示拼接成单个输入序列，并通过嵌入重加权和身份保持注意力增强一致性，同时保持扩散模型不变。

算法1*ReDiStory*：提示嵌入重组

需要： 身份提示 P_{id}，帧提示 $\{P_n\}_{n=1}^{N}$
需要： 文本编码器 $\mathcal{E}$，扩散模型 $\mathcal{G}_\theta$
对于 $n = 1$到 N **执行**
  $E_n \leftarrow \mathcal{E}(P_{id} \oplus P_n)$
  $E^{id} \leftarrow$切片$_{id}(E_n)$，　$E_n^f \leftarrow$ 切片$_f(E_n)$
结束循环
循环遍历 $n = 1$至 N **执行**
  $\tilde{E}_n^f \leftarrow E_n^f - \frac{1}{N-1}\sum_{m\neq n}$ 投影$_{E_m^f}(E_n^f)$
  $\tilde{E}_n \leftarrow [E^{id}; \tilde{E}_n^f]$
结束循环
循环 $n = 1$到 N **执行**
  $I_n \sim$样本$(\mathcal{G}_\theta \mid \quad \tilde{E}_n)$
结束循环
返回 $\{I_n\}_{n=1}^{N}$

2.3. 提示结构与语义干扰。

除了模型或注意力层面的修改，近期研究还指出，提示结构会显著影响生成行为 [12]。基于拼接的提示可以利用语言编码器的上下文连贯性，但也可能使嵌入空间中的帧表示相互关联。在多帧设置中，这种关联可能导致帧间语义干扰：帧特定描述会部分覆盖共享身份属性，并导致不同帧之间出现身份漂移 [11, 13]。*ReDiStory* 针对这类与提示相关的问题，在推理时间重新组织帧级提示嵌入，无需显式状态表示、额外监督或修改底层扩散模型，即可减轻帧间语义干扰。

3. 方法

本节介绍 *ReDiStory*。我们的动机来自这样一个观察：联合编码身份与帧提示，可能会将持久的身份线索与帧特定指令纠缠在一起，进而导致跨帧语义干扰和身份漂移。我们的核心思路是显式分离身份相关嵌入与帧特定嵌入，然后从帧特定部分中移除各帧共享的方向。这种推理时重组无需额外监督或参数更新。算法1概述了整体流程。

3.1. 问题表述

我们研究免训练设置下的多帧视觉故事生成任务。给定用于描述主体的身份提示 P_{id}，以及一组用于指定帧级变化的 N 帧提示$\{P_n\}_{n=1}^{N}$ ，目标是生成一系列图像 $\{I_n\}_{n=1}^{N}$，使主体

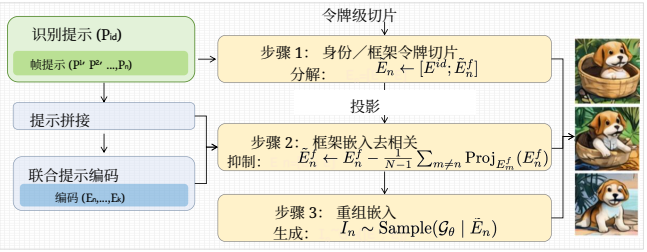


图 2：ReDiStory 流程。身份与帧提示经过联合编码，在令牌级别进行分解，并通过基于投影的去相关进行重新组织，从而在标准扩散采样之前减少帧间语义干扰。

身份在各帧中保持一致，同时每张图像都忠实地遵循对应的帧提示。

设 $\mathcal{G}_\theta$ 表示一个参数为 θ 的预训练文生图扩散模型。在免训练设置下，帧在推理时间独立生成：

$$I_n \sim \mathcal{G}_\theta(P_{id}, P_n), \quad n = 1, \ldots, N. \tag{1}$$

在免训练多帧生成中，一个关键挑战在于：在不修改 θ 或引入额外监督的情况下，保持跨帧的身份相关语义。已有研究表明，朴素的提示拼接往往会导致 帧间语义干扰，即帧特定内容抑制或扭曲共享身份属性。*ReDiStory* 通过在提示嵌入被扩散模型处理之前重新组织它们来解决这一问题。

3.2. 提示嵌入表示

我们首先介绍提示嵌入在 *ReDiStory* 中的表示与分解方式。遵循 1Prompt1Story [13]，我们将身份提示 P_{id} 与帧提示 P_n 拼接起来，并使用文本编码器 $\mathcal{E}(\cdot)$ 对其进行编码：

$$E_n = \mathcal{E}(P_{id} \oplus P_n) \in \mathbb{R}^{L_n \times d}, \tag{2}$$

其中，L_n 表示词元长度，d表示嵌入维度。由于 P_{id} 和 P_n 是联合编码的，身份线索可能会与共享嵌入空间中的帧特定指令纠缠在一起。

*ReDiStory*通过将 E_n划分为身份相关部分和帧特定部分，明确呈现了这一结构：

$$E_n = \left[E^{id}; E_n^f\right], \tag{3}$$

其中，$E^{id} \in \mathbb{R}^{L_{id} \times d}$ 对应P_{id} 的词元嵌入（在各帧间共享），$E_n^f \in \mathbb{R}^{L_n^f \times d}$ 对应 P_n的词元嵌入。这种分解使我们能够对共享身份和帧级内容进行非对称处理，这构成了所提出的重组策略的基础。我们将此提取操作记为 $E^{id} = \mathrm{Slice}_{id}(E_n)$和$E_n^f = \mathrm{Slice}^f(E_n)$，它们分别选取与身份词元和帧词元对应的嵌入行。由于身份提示在各帧间相同，因此可以从任意 E_n中一致地提取 E^{id}。

3.3. 提示嵌入重组

为减轻帧间语义干扰，*ReDiStory* 通过移除各帧之间共享的方向，重新组织帧特定嵌入。设 $\{E_n^f\}_{n=1}^N$ 表示帧特定嵌入矩阵的集合。对于每一帧 n，我们将重组后的嵌入 $\tilde{E}_n^f$ 计算为

$$\tilde{E}_n^f = E_n^f - \frac{1}{N-1}\sum_{m\neq n} \mathrm{Proj}_{E_m^f}(E_n^f), \tag{4}$$

其中，$\mathrm{Proj}_A(B)$将 B 投影到嵌入空间中由 A张成的子空间（例如，A的列空间），并逐行应用于 B中的词元嵌入。直观而言，这种减法会抑制跨帧共享方向，促使 $\tilde{E}_n^f$ 保留帧特定信息，而不是向其他帧偏移。

随后，我们将第 n 帧的重组提示嵌入构造为

$$\tilde{E}_n = \left[E^{id}; \tilde{E}_n^f\right]. \tag{5}$$

该过程具有确定性，完全在推理时执行，不涉及任何额外的目标函数、梯度或参数更新。

3.4. 生成与计算复杂度

˜给定 E_n，我们使用与基础模型相同的扩散采样器生成每一帧。设 $z_t^{(n)}$ 表示扩散步骤 t时第n帧的潜变量，其中 $t = T, \ldots, 1$，T为去噪步数的总数。在重组后的˜文本嵌入 E_n的条件下，采样器按以下方式迭代：

$$z_{t-1}^{(n)} \sim p_\theta\left(z_{t-1}^{(n)} \mid z_t^{(n)}, \tilde{E}_n\right), \quad t = T, \ldots, 1, \tag{6}$$

其中，p_θ 表示由 θ参数化的扩散模型条件去噪转移（即采样核）。从扩散模型的角度来看，*ReDiStory* 不会改变架构或推理过程，而是提供重组后的提示嵌入，以减少跨帧语义干扰。

就计算成本而言，每个故事只需执行一次重组，其开销会随帧数呈二次增长。设 d表示嵌入维度，N 表示帧数；最坏情况下的成本为 $\mathcal{O}(N^2 d)$，具体还要乘以取决于提示长度（即帧提示词元数量）的因子。与扩散推理成本相比，这一开销可以忽略不计，并且 *ReDiStory* 无需反向传播或额外的模型评估。

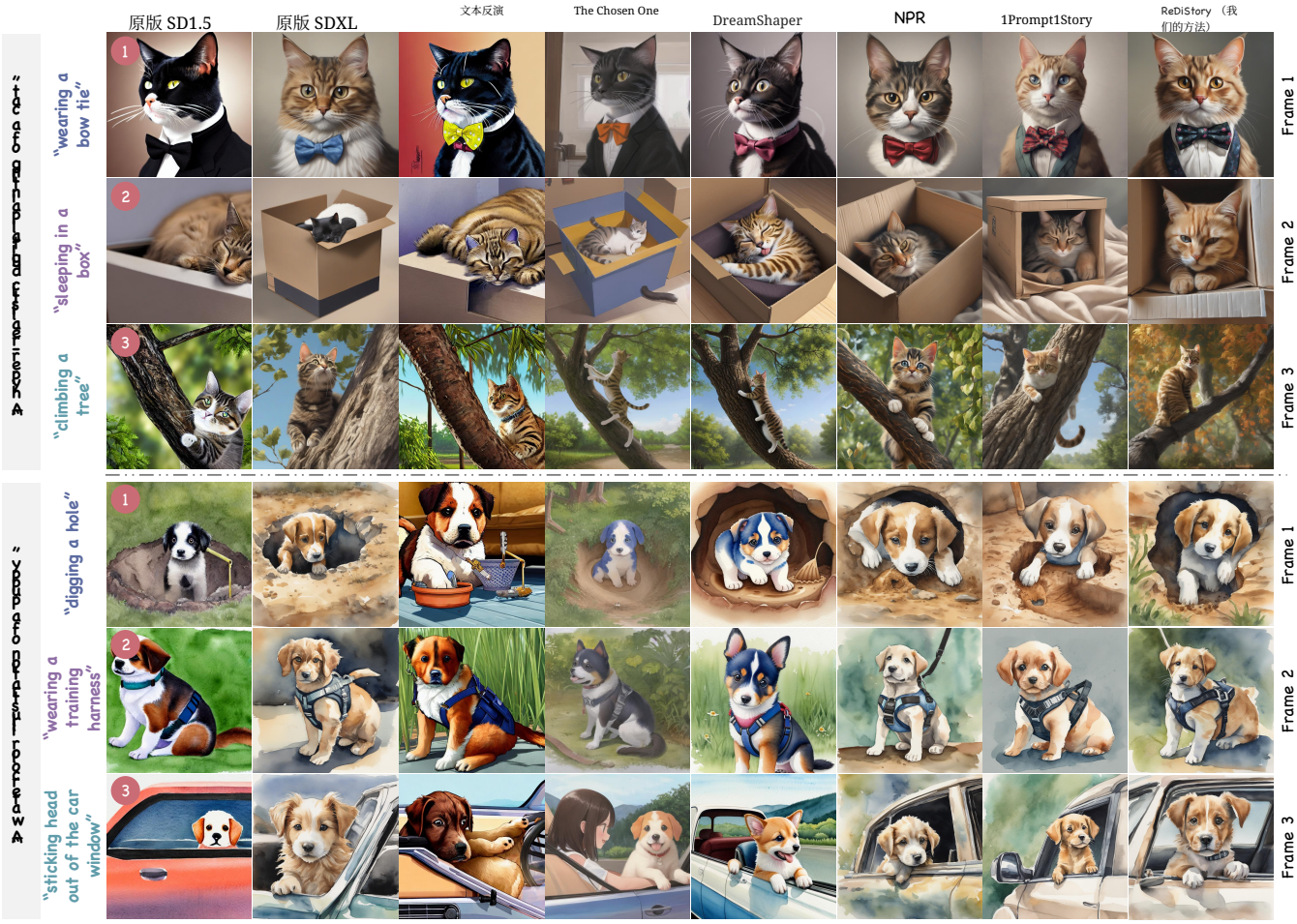


图3：多帧生成的定性比较。在身份与帧提示完全相同的情况下，*ReDiStory* 能够在场景和姿势变化时保持主体特定外观。

4. 实验

4.1. 实验设置

4.1.1. 数据集

我们在 ConsiStory+基准测试 [13],上评估 *ReDiStory*，该基准测试是用于免训练多帧故事讲述的标准测试平台。该数据集包含 50 个主体，涵盖人物、动物、物体和虚构角色等多种类别。对于每个主体，5 个帧提示分别描述不同的场景、动作或视角，共计 250 个多帧测试用例。所有方法均在相同的提示词设置下进行评估，以确保公平比较。

4.1.2. 基线方法

我们将*DeCorStory* 与几种具有代表性的文本到图像一致性生成方法进行比较，包括 1Prompt1Story[13], BLIP-Diffusion [17], 文本反演 (TI) [9], IP- Adapter [14], PhotoMaker [15], The Chosen One [18], Con- siStory [11], 和 StoryDiffusion [16]。所有基线方法均经过评估——评估时采用其官方配置或公开发布的实现，以确保公平性。

4.1.3. 实现细节

所有实验均使用由 diffusers 库实现的 SDXL [19] 。所有方法均采用 50 个去噪步数和 7.5 的无分类器引导尺度。除非另有说明，我们固定随机种子以确保结果可复现。

4.1.4. 评估指标

我们从两个方面评估 *ReDiStory*： 提示保真度和身份一致性。提示词对齐度通过 CLIP-T 衡量，计算每张生成图像与其对应帧提示之间 CLIPScore [20] 的平均值。身份一致性采用 CLIP-I 进行评估，其定义为 CLIP 图像嵌入之间的余弦相似度；同时还使用 DreamSim [21], 这一与人类判断相关的感知相似度指标。按照 DreamSim 协议，图像

表1：在 ConsiStory+ 上的定量比较。**粗体**和下划线分别表示免训练多帧方法中的最佳和次佳结果。原版 SD1.5 和原版 SDXL 作为模型参考展示，不参与排名。BLIP-Diffusion 和文本反演基于 SD1.5；其余方法均使用 SDXL。尽管数值提升幅度有限，ReDiStory 在各项指标上始终提升身份一致性。

方法	CLIP-T↑	CLIP-I↑	DreamSim↓	步数	显存（GB）↓	推理时间（秒）↓
原版参考模型（不纳入排名）						
原版 SD1.5	0.8353	0.7474	0.5873	50	4.73	2.4657
原版 SDXL	0.9074	0.8165	0.5292	50	16.04	13.0890
基于训练的方法 / 参考引导方法（作为参考项报告）						
BLIP-Diffusion	0.7607	0.8863	0.2830	26	**7.75**	**1.9284**
文本反演	0.8378	0.8229	0.4268	40	32.94	282.507
The Chosen One	0.7614	0.7831	0.4929	35	10.93	11.2073
PhotoMaker	0.8651	0.8465	0.3996	50	23.79	18.0259
IP-Adapter	0.8458	**0.9429**	**0.1462**	30	19.39	13.4594
免训练多帧方法						
ConsiStory	0.8769	0.8737	0.3188	50	34.55	34.5894
StoryDiffusion	0.8877	0.8755	0.3212	50	45.61	25.6928
朴素提示词重加权（NPR）	0.8411	0.8916	0.2548	50	**16.04**	**17.2413**
1Prompt1Story	0.8942	0.9117	0.1993	50	18.70	23.2088
ReDiStory（我们的方法）	**0.8966**	**0.9149**	**0.1952**	50	18.89	23.6413

使用 CarveKit [22] 去除背景，以便专注于评估主体身份。

4.2. 比较

4.2.1. 定量结果

表1总结了性能和效率。在免训练方法中，*ReDiStory* 在身份一致性方面较最强基线 1Prompt1Story 有所提升（尽管数值提升幅度不大）：CLIP-I 更高（0.9149 对 0.9117），DreamSim 更低（0.1952 对 0.1993），同时提示保真度（CLIP-T）相当。这一提升伴随着轻微的效率折衷：显存从 18.70 GB 增至 18.89 GB，推理时间从 23.21 s 增至 23.64 s，表明提示词级重组仅需付出少量额外成本，便可提升一致性。

4.2.2. 定性结果

图3展示了具有代表性的示例。对于同一主体，随着帧变化增大，1Prompt1Story 可能会出现逐渐的身份漂移（例如，外观细节发生细微变化）。*ReDiStory*能够更好地保留主体在不同帧中的特征，同时仍能体现帧级提示词。这表明，提升一致性并不会以牺牲视觉多样性或提示词遵循度为代价。

4.3. 消融研究

表2分析了提示嵌入重组的效果。移除重组会导致身份一致性下降，表现为 CLIP-I 降低、Dream-Sim 升高。仅重组身份相关嵌入可带来一定改善，而完整的 *ReDiStory* 模型

表2：在 ConsiStory+ 基准测试上的消融研究。逐步加入嵌入重组可提升身份一致性（更高的 CLIP-I、更低的 DreamSim），同时保持提示保真度（CLIP-T）稳定。

方法	CLIP-T↑	CLIP-I↑	DreamSim↓
无重组	0.8921	0.9078	0.2014
仅身份重组	0.8943	0.9112	0.1986
ReDiStory（完整）	**0.8966**	**0.9149**	**0.1952**

在所有指标上均取得最佳结果。这些发现支持了我们的假设：解耦身份相关语义与帧特定语义对于缓解帧间语义干扰至关重要；所提出的重组方法能够提升一致性，同时不损害提示保真度。

5. 结论

我们提出了 *ReDiStory*，一种用于多帧视觉故事生成的免训练框架，通过推理时提示嵌入重组来提升主体一致性。该框架通过分离身份相关和帧特定嵌入，并抑制跨帧共享方向，缓解跨帧语义干扰，且无需修改扩散主干网络。在相同设置下，这种提示级干预始终能够提升身份保持能力，同时只带来极低的计算开销。虽然提升幅度有限，但结果表明，在嵌入层面减少跨帧耦合是一种有效且轻量的方法，也为未来研究如何增强其在更长序列和更具挑战性的外观变化中的效果指明了方向。

6. 参考文献

[1] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer，“使用潜在扩散模型进行高分辨率图像合成”，载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，2022 年，第 10684–10695 页。

[2]C. Saharia、W. Chan、S. Saxena、L. Li、J. Whang 等，“具备深度语言理解能力的照片级真实感文本到图像扩散模型”，载于神经信息处理系统进展（*NeurIPS*），2022 年，第 35 卷，第 36479–36494 页。

[3] Z. Yu、M. IDRIS 和 P. Wang，“可视化不断变化的地球：通过卫星影像实现环境叙事民主化的创意人工智能框架”，载于第三十九届神经信息处理系统年会创意人工智能专题：人类，2025 年。

[4] G. Gaur、M. Zolfaghari 和 T. Brox，“Storysync：通过区域协调实现文本到图像生成中的免训练主体一致性”，载于*IEEE/CVF* 国际计算机视觉会议论文集（*ICCV*），2025 年。

[5] A. Ma 等，《Lay2story：利用布局引导的故事视觉化与空间一致性扩散》，载于*IEEE/CVF* 国际计算机视觉会议（*ICCV*）论文集，2025年。

[6] Z. Yu、M. IDRIS、P. Wang 和 R. Qureshi，《Cotextor：通过分层解耦与深度感知融合实现免训练模块化多语言文本编辑》，载于第三十九届神经信息处理系统年会创意人工智能专题：人类，2025年。

[7]S. Yang、Y. Wang、K. Chen 等，《Seed-story：利用大语言模型进行多模态长篇故事生成》，载于*IEEE/CVF* 国际计算机视觉会议（*ICCV*）研讨会论文集，2024年。

[8]L. Hu、X. Gao、P. Zhang、K. Sun、B. Zhang 和 L. Bo，《Animate Anyone：用于角色动画的一致且可控的图像到视频合成》，载于*IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集，2024年，第8153–8163页。

[9] N. Ruiz、Y. Li、V. Jampani、Y. Pritch、M. Rubinstein 和 K. Aberman，《Dreambooth：微调用于主体驱动生成的文本到图像扩散模型》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2023年，页码 22500–22510。

[10] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano 等，“一张图像胜过一个词：使用文本反演个性化文本到图像生成”，载于国际学习表征会议论文集（*ICLR*），2023年。

[11] Yoad Tewel、Omri Kaduri、Rinon Gal、Yoni Kasten、Lior Wolf、Gal Chechik 和 Yuval Atzmon，“免训练一致性文本到图像生成”，*ACM* 图形学汇刊（*TOG*），第43卷，第4期，第1–18页，2024年。[12] A. Hertz、A. Voynov、S. Fruchter 等，“利用交叉注意力控制进行提示词到提示词的图像编辑”，载于国际学习表征会议论文集（*ICLR*），2023年。[13] T. Liu、K. Wang、S. Li、J. van de Weijer、F. S. Khan 等，“One-prompt-one-story：使用单个提示词实现免费的一致性文本到图像生成”，载于国际学习表征会议论文集（*ICLR*），2025年。

[14]H. Ye、J. Zhang、S. Liu、X. Han 和 W. Yang，《IP-Adapter：兼容文本的图像提示适配器，适用于文本到图像扩散模型》，*arXiv*预印本 *arXiv:2308.06721*，2023年。

[15] Z. Li、M. Cao、X. Wang 等，《PhotoMaker：通过堆叠式 ID 嵌入定制逼真人像照片》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2024年，页码 8640–8650。

[16] Y.Zhou、D.Zhou、M.Cheng、J. Feng 和 Q. Hou，“Sto- rydiffusion：用于长程图像和视频生成的一致性自注意力”，发表于神经信息处理系统进展（*NeurIPS*），2024年。

[17] D. Li、J. Li 和 S. Hoi，“Blip-diffusion：用于可控文本到图像生成与编辑的预训练主体表征”，发表于神经信息处理系统进展（*NeurIPS*），2023年，第36卷，第30146–30166页。

[18] O. Avrahami、A. Hertz、Y. Vinker、M. Arar、S. Fruchter 等，“The chosen one：文本到图像扩散模型中的一致性角色”，发表于*ACM SIGGRAPH* 会议论文，2024年。

[19] D. Podell、Z. English、K. Lacey、A. Blattmann、T.Dock- horn 等人，“SDXL：改进用于高分辨率图像合成的潜变量扩散模型”，*arXiv* 预印本 *arXiv:2307.01952*，2023年。

[20] J.Hessel、A. Holtzman outdoor、M. Forbes 等人，“Clip-score：一种用于图像字幕生成的无参考评估指标”，发表于自然语言处理经验方法会议论文集（*EMNLP*），2021年，第7514–7528页。

[21] S. Fu、N. Tamir、S. Sundaram、L. Chai、R. Zhang 等人，“Dreamsim：利用合成数据学习人类视觉相似性的新维度”，*arXiv* 预印本 *arXiv:2306.09344*，2023年。

[22]N. Selin，“Carvekit：自动化高质量背景移除框架”，https://github.com/OPHoperHPO/image-background-remove-tool，2023年。

A.1. 附录

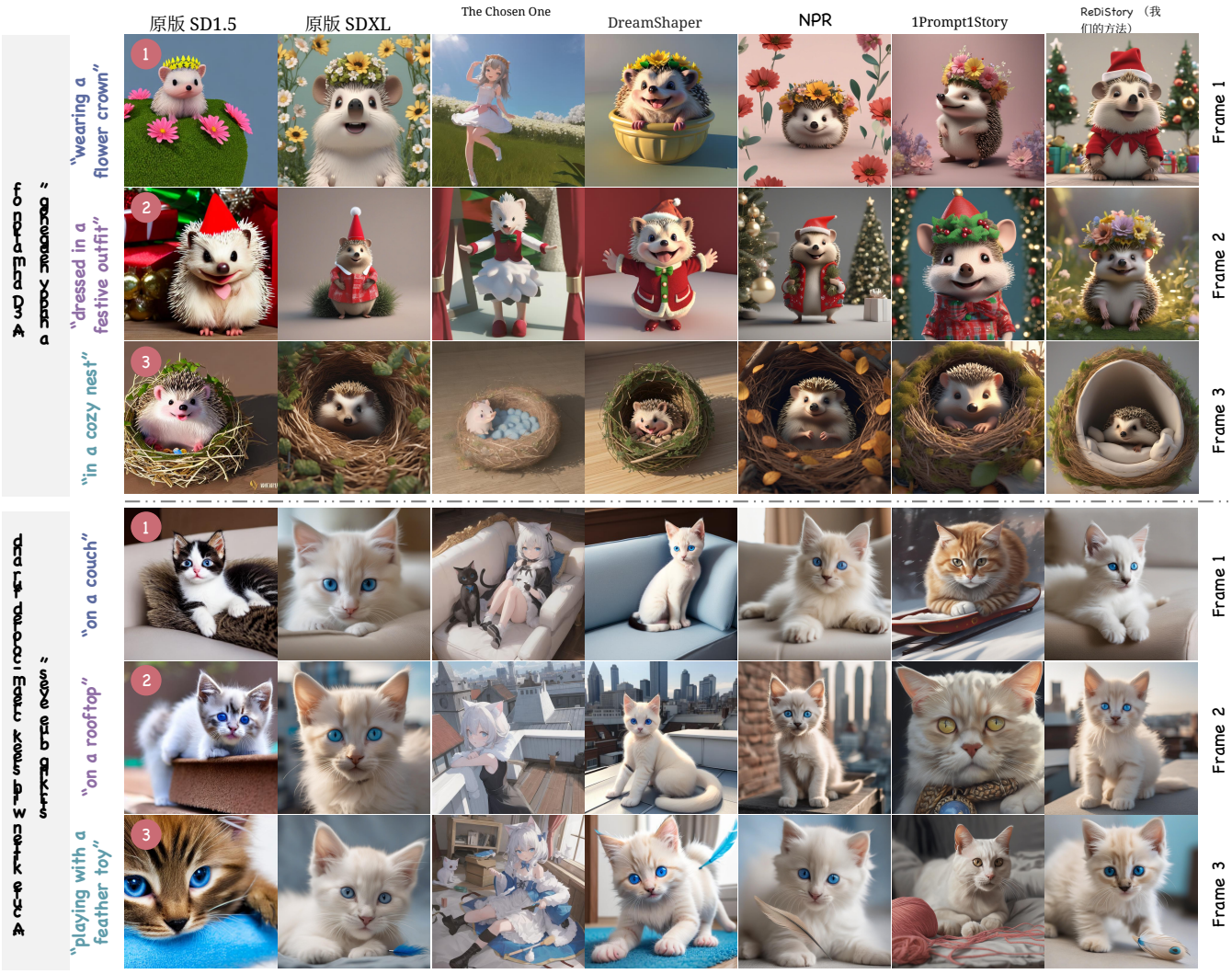


图 A.1：附加比较结果。