arXiv:2602.01306v1 [cs.CV] 1 Feb 2026

DECORSTORY：格拉姆–施密特提示嵌入去相关，实现一致的叙事

Ayushman Sarkar[1]*,Zhenyu Yu*[2,*]*, Mohd Yamani Idna Idris*[2]

[1] 比尔布姆工程技术学院， [2] 马来亚大学ayush

mansarkar123@gmail.com; yuzhenyuyxl@foxmail.com; yamani@um.edu.my

摘要

在文本到图像故事生成中，保持不同帧之间的视觉和语义一致性是一项关键挑战。现有的免训练方法（如一提示词一故事）会将所有提示词串联成一个序列，这往往会造成严重的嵌入相关性，进而导致颜色泄漏、背景混合和身份漂移。我们提出了*DeCorStory*，一个免训练的推理时框架，旨在明确减少帧间语义干扰。*DeCorStory* 通过格拉姆–施密特提示嵌入去相关来正交化帧级语义，随后采用奇异值重加权来强化提示词特定信息，并利用身份保持交叉注意力，在扩散过程中稳定角色身份。该方法无需修改模型或进行微调，且可无缝集成到现有扩散流水线中。实验表明，该方法在提示词–图像对齐、身份一致性和视觉多样性方面均取得稳定提升，在免训练基线方法中达到最先进的性能。代码见：
https://github.com/YuZhenyuLindy/DeCorStory

索引术语— 文生图、免训练、提示词嵌入去相关、身份一致性、叙事

1. 引言

在动画、视觉故事讲述和视频合成等叙事生成任务中，如何在多帧之间保持一致性，仍然是文生图（T2I）扩散模型面临的一项根本挑战 [1]。连贯的故事要求模型在不同场景中保留主体的身份、外观和风格，同时还能灵活适应文本描述的背景和动作变化。近期基于扩散的生成方法 [2] 和身份保持文生图方法 [3] 取得了长足进展，显著提升了视觉质量和可控性，使文本驱动的故事讲述日益可行。诸如一提示词一故事（1Prompt1Story） [4] 这样的免训练方法，通过将所有帧提示词拼接为单个输入序列，进一步降低了计算开销，从而利用大型语言–视觉模型隐含的上下文一致性。

然而，这种简单的拼接策略也带来了新的退化因素。由于多个帧通常描述同一主体在不同情境下的状态，它们在语言空间中的词元嵌入往往高度相关，导致帧间出现语义干扰 [5]。一个场景中的视觉属性可能渗入另一个场景，造成背景混杂或细节漂移。随着帧数增加或场景多样性增大，这种相关性问题会愈发严重，限制现有免训练方法在复杂或长篇叙事中的可扩展性 [1]。由于以往方法完全依赖模型隐含的上下文理解，因此缺乏明确的帧语义解耦机制；当场景差异显著时，生成结果便会缺乏一致性。

为解决这一局限，我们提出*DeCorStory*，一种用于一致多帧生成的免训练框架。在提示拼接之后，*DeCorStory* 应用格拉姆-施密特提示词嵌入去相关，以显式正交化帧级嵌入，从而在保持共享身份子空间的同时减少帧间语义重叠。在此基础上，我们分别集成奇异值重加权（SVR）和身份保持交叉注意力（IPCA） [4] ，以强化帧特定语义并增强主体一致性。整个过程完全在推理时运行，无需任何额外训练或微调，并可无缝集成到依赖强大基础视觉编码器（如DINOv2 [6]）的现有扩散流水线中。我们的主要**贡献** 如下：

- **帧嵌入相关性分析：** 我们识别并分析拼接提示中帧级嵌入的相关性，揭示其是语义干扰背后的关键因素。
- **DeCorStory框架：** 我们提出了一种轻量级的基于格拉姆–施密特的去相关模块，可直接集成到扩散模型中，无需对其进行修改或重新训练。
- **统一推理管线：** 我们将嵌入去相关、SVR 和 IPCA 整合为一个连贯的免训练管线，实现保留身份且针对特定帧的多帧生成。

2. 相关工作

2.1. 一致性文生图

个性化文生图中的主体一致性和身份一致性已得到广泛研究。基于训练的方法，例如文本反演 [7] 和 DreamBooth [8]„ 通过学习主体特定嵌入或微调扩散模型，来保持不同生成结果中的身份一致性。后续研究引入了架构扩展或适配器，包括 BLIP-Diffusion、IP-Adapter 及相关设计 [9], ，以便更好地将身份线索注入扩散模型。尽管这些方法能够实现较强的一致性，但通常需要额外训练或修改模型，因此与推理时方法相比，灵活性有所受限。

2.2. 无需训练的一致性叙事

近期研究开始探索无需训练的多帧叙事一致性。ConsiStory 及相关方法在去噪过程中通过注意力操纵传播身份信息 [1, 10]。而 1Prompt1Story [4] 则将所有帧提示词拼接为一个序列，并将一致性归因于上下文连贯性，再通过奇异值重加权和身份保持交叉注意力进一步增强这种一致性。我们的工作遵循这一无需训练的范式，但针对一个被忽视的问题：提示拼接会使帧级嵌入之间产生强相关性，导致语义在帧间泄漏。

2.3. 提示词与嵌入操纵

大量研究探讨如何操纵文本嵌入或注意力，在无需重新训练的情况下引导扩散模型，包括提示词重加权、Prompt-to-Prompt 编辑和即插即用控制 [11]。其他研究则分析嵌入结构，或有选择地抑制不需要的概念 [12]。*DeCorStory*属于这一研究方向，通过格拉姆-施密特正交化显式去相关帧级提示词嵌入，并有选择地增强帧特定语义，从而为一致性多帧叙事提供一种轻量级的嵌入空间解决方案。

3. 方法

DeCorStory 是一种免训练推理时框架，通过在冻结的扩散骨干网络上进行提示拼接、格拉姆–施密特去相关和身份感知重加权，减少帧间语义干扰（图 1），无需对模型进行微调。

3.1. 提示拼接

遵循 1Prompt1Story，我们首先构建一个长提示词，将身份描述与所有特定于帧的提示词拼接起来。这样一来，模型便能在各帧之间共享一致的上下文，同时保留不同的场景信息。具体而言，身份提示概括了全局主体特征（例如外观、风格），而每个帧提示词则描述该帧独有的局部属性，例如姿势、背景或事件。我们将身份提示 P_0 与 N帧提示词 $\{P_1, \ldots, P_N\}$ 拼接起来，并将其编码为

$$C = \tau_\xi([P_0; P_1; \ldots; P_N]), \tag{1}$$

$$C = [c_{\mathrm{SOT}}, c_{P_0}, c_{P_1}, \ldots, c_{P_N}, c_{\mathrm{EOT}}] \in \mathbb{R}^{M \times D}, \tag{2}$$

其中，$\tau_\xi(\cdot)$表示文本编码器，M 是词元总数，D 是嵌入维度。这里，c_{SOT}和 c_{EOT} 分别是特殊的文本起始标记和文本结束标记，而每个 c_{P_k} 对应于联合提示上下文中帧 k 的描述（可能包含多个词元）的嵌入。我们将帧级词元嵌入收集为

$$X = \begin{bmatrix} c_{P_1} \\ c_{P_2} \\ \vdots \\ c_{P_N} \end{bmatrix} \in \mathbb{R}^{N \times D}, \quad c_{P_k} \in \mathbb{R}^{1 \times D}. \tag{3}$$

该矩阵 X 可视为共享语义空间中所有帧提示词的紧凑表示。然而，由于所有 c_{P_k} 都与同一个身份提示 P_0联合编码，它们往往高度相关，这可能导致生成过程中不同帧的内容相互“渗入”。下一阶段将明确解决这一问题。

3.2. 格拉姆–施密特嵌入去相关

为了缓解帧间相关性，我们执行逐行˜格拉姆–施密特正交化，作用于 X，以得到 X:

$$\tilde{c}_{P_1} = \frac{c_{P_1}}{\|c_{P_1}\|_2}, \tag{4}$$

$$\tilde{c}_{P_k} = \frac{c_{P_k} - \sum_{i=1}^{k-1} \langle c_{P_k}, \tilde{c}_{P_i} \rangle \, \tilde{c}_{P_i}}{\left\| c_{P_k} - \sum_{i=1}^{k-1} \langle c_{P_k}, \tilde{c}_{P_i} \rangle \, \tilde{c}_{P_i} \right\|_2}, \quad k = 2, \ldots, N, \tag{5}$$

其中，内积 $\langle a, b \rangle = a\, b^\top$ 用于行向量。从几何上看，我们将每个 c_{P_k} 视为 D维嵌入空间中的向量，并构造一个张成同一子空间、但帧方向两两正交的标准正交基。随后，我们构造去相关词元矩阵

$$\tilde{C} = [\, c_{\mathrm{SOT}}, c_{P_0}, \tilde{c}_{P_1}, \ldots, \tilde{c}_{P_N}, c_{\mathrm{EOT}} \,]. \tag{6}$$

与 C 相比，身份嵌入 c_{P_0} 和特殊词元保持不变，只有帧级嵌入被旋转至彼此正交的方向。这保留了每个提示的整体语义内容，同时减少了不同帧在共享嵌入空间中的重叠。

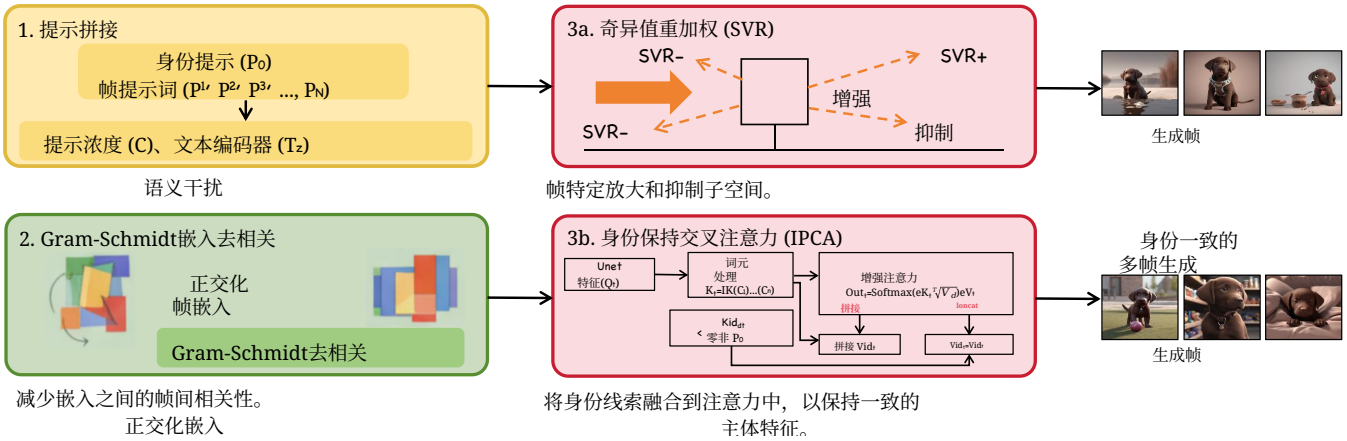


图 1：*DeCorStory* 流程概览。该方法由提示拼接、格拉姆–施密特嵌入去相关，以及用于身份一致的多帧生成的推理时模块（SVR 和 IPCA）组成。身份提示在所有帧之间共享，而经过去相关和重新加权的帧级嵌入则提供清晰且针对特定帧的条件化。

我们采用数值稳定的修正 Gram–Schmidt 过程，其排序在实践中的影响可以忽略不计；鉴于帧数较少，其计算成本也微不足道。此操作促使帧级嵌入近似正交，从而在扩散推理前减少帧间语义纠缠。

3.3. 与 SVR 和 IPCA 集成

对于每个目标帧 j，SVR 会增强 P_j，同时抑制其他帧，随后通过 IPCA 强化身份。这种条件化通过特定帧的文本矩阵在采样过程中即时应用，无需更改架构或额外训练。

3.3.1. 奇异值重加权（*SVR*）。

令 表达集合为 $X^{\text{exp}} = [\tilde{c}_{P_j};\ c_{\text{EOT}}] \in \mathbb{R}^{2\times D}$。在其列空间上计算奇异值分解（等价地，对 $(X^{\text{exp}})^\top$ 计算）：

$$(X^{\text{exp}})^\top = U\,\Sigma\,V^\top, \tag{7}$$

$$\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r),\ \sigma_1 \geq \cdots \geq \sigma_r \geq 0. \tag{8}$$

这里，奇异向量捕捉由当前帧和文本结束嵌入共同支持的主要语义方向。因此，对奇异值重新加权，可以强调或抑制整个语义子空间，而非单个坐标。SVR+放大奇异值：

$$\hat{\sigma}_\ell = \beta\, e^{\alpha\,\sigma_\ell}\,\sigma_\ell, \quad \forall \ell, \tag{9}$$

得到 $\hat{\Sigma} = \text{diag}(\hat{\sigma}_\ell)$ 以及更新后的词元

$$\hat{X}^{\text{exp}} = \left(U\,\hat{\Sigma}\,V^\top\right)^\top = [\,\hat{c}_{P_j};\,\hat{c}_{\text{EOT}}\,]. \tag{10}$$

这些超参数控制对主导奇异方向的增强程度：数值越大，对最显著语义的强调就越强

帧 j。对于每个抑制 帧 $k \neq j$，构造 $X_k^{\text{sup}} = [\tilde{c}_{P_k};\ \hat{c}_{\text{EOT}}]$ 并以类似方式计算ˆ奇异值分解，以获得 Σ_k。SVR− 会衰减奇异值：

$$\tilde{\sigma}_{k,\ell} = \beta'\, e^{-\alpha'\,\hat{\sigma}_{k,\ell}}\,\hat{\sigma}_{k,\ell}, \quad \forall \ell, \tag{11}$$

从而得到 $\tilde{X}_k^{\text{sup}} = [\,\tilde{c}_{P_k}^{\downarrow};\ \tilde{c}_{\text{EOT}}\,]$。组装逐帧条件化文本矩阵：

$$\tilde{C}^{(j)} = [\,c_{\text{SOT}},\, c_{P_0},\, \tilde{c}_{P_1}^{\downarrow}, \ldots, \hat{c}_{P_j}, \ldots, \tilde{c}_{P_N}^{\downarrow},\, \tilde{c}_{\text{EOT}}\,]. \tag{12}$$

直观而言，SVR+ 强调当前帧的语义子空间，而 SVR− 抑制其他帧的影响，从而得到更清晰的帧特定条件化。由于 SVR 作用于词元矩阵的低秩分解，因此无需直接编辑原始嵌入，就能以紧凑且易于解释的方式调节帧语义。

3.3.2. 身份保持交叉注意力 *(IPCA)*。

在去噪步骤 t，交叉注意力使用来自 UNet 特征的查询 Q_t，以及来自 $\tilde{C}^{(j)}$ 的键/值：

$$K_t = \ell_K(\tilde{C}^{(j)}), \quad V_t = \ell_V(\tilde{C}^{(j)}). \tag{13}$$

令 K_t^{id}，$V_{\text{身份}t}$ 为副本，其中所有非身份（非 P_0）词元特征均置零。我们将其拼接

$$\widetilde{K}_t = \text{Concat}(K_t,\, K_t^{\text{id}}), \qquad \widetilde{V}_t = \text{Concat}(V_t,\, V_t^{\text{id}}), \tag{14}$$

并计算增强注意力

$$\widetilde{A}_t = \text{softmax}\Big(\frac{Q_t\,\widetilde{K}_t^\top}{\sqrt{d}}\Big), \qquad \text{Out}_t = \widetilde{A}_t\,\widetilde{V}_t, \tag{15}$$

其中 d 是键/查询维度。原始键和值编码了混合语义，而仅身份分支则为前景身份线索预留了注意力容量。

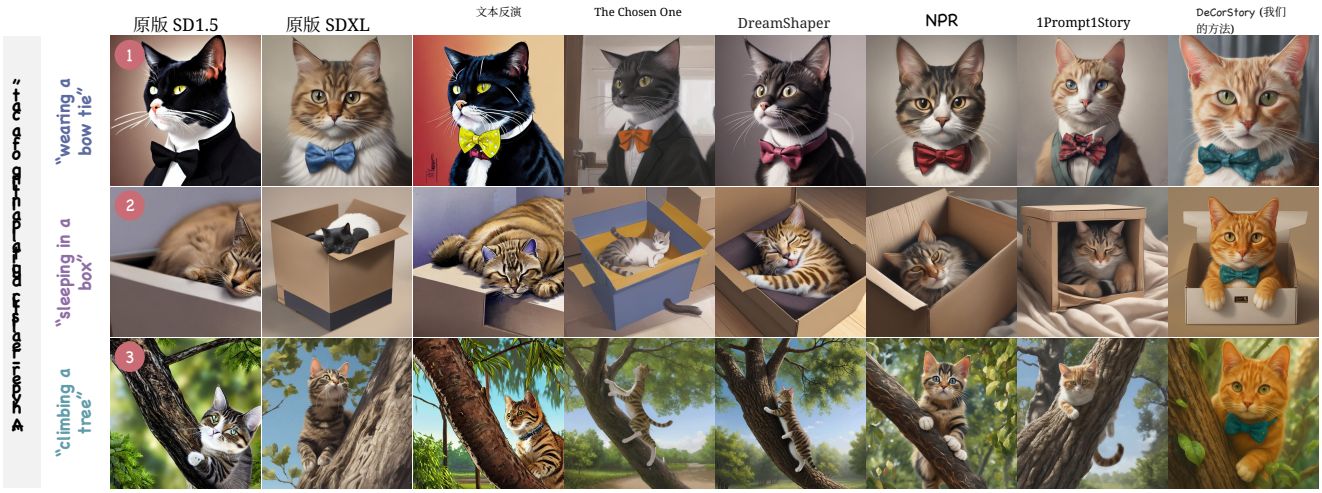


图2：定性比较。不同文本到图像生成方法在代表性提示词上的视觉比较。*DeCorStory*能够生成更加一致的全局与局部细节，更清晰地保持角色身份，并在各个场景之间实现更好的叙事连贯性。更多示例请见附录图A1和A2。

注意力输出 Out，融合了身份线索和帧特定语义，在每个去噪步骤中与去相关且经 SVR 调制的提示词一同应用时，可实现稳定的跨帧一致性。

中和。遵循 Fu 等人 [19], 的方法，我们使用 CarveKit [20]去除背景，并将其替换为随机噪声，以确保相似度评估仅关注主体身份，而非上下文伪影。

4. 实验

4.1. 实现细节

数据集描述。 在实验中，我们采用 Con- siStory+ 数据集 [4], ，这是原始 ConsiStory 基准 [13]的扩展版本。它涵盖了更多主体、场景描述和视觉风格，适用于评估多帧生成和叙事生成任务。该数据集包含 200 组提示词集，约有 1500 个生成帧，我们据此评估提示对齐和身份一致性。

实验设置。 我们将*DeCorStory*与多种具有代表性的文生图一致性生成方法进行比较，包括 1Prompt1Story [4], BLIP-Diffusion [14], 文本反演（TI） [15], IP-Adapter [16], PhotoMaker [3], The Chosen One [17], ConsiStory [13], 和 StoryDiffusion [10]。所有基线方法均采用其官方配置或公开发布的实现进行评估，以确保公平性。

评估指标。 我们采用三种互补指标评估方法。(1) 提示对齐。我们计算每个生成帧与其对应提示之间的平均 CLIP 分数 [18] ，记为 CLIP-T，以衡量文本与图像之间的语义对齐。(2) 身份一致性。我们采用 DreamSim [19] 和 CLIP-I [18] 来量化帧间的视觉一致性。Drea mSim 反映感知相似度，而 CLIP-I 则通过计算 CLIP 图像嵌入之间的余弦相似度得出。(3) 背景

4.2. 比较

定性比较。 图 2 展示了两组三帧故事的定性结果，它们具有不同的风格和动作。原版 SD1.5/SDXL 以及现有的个性化或叙事基线方法，通常会在不同帧之间出现身份漂移、风格不一致以及语义元素缺失，例如外观、配饰或背景细节发生变化。相比之下，*DeCorStory* 能够生成一个易于识别的单一角色，使其在所有帧中保持一致，同时准确遵循叙事动作。视觉属性、配饰和渲染风格均得到保留，背景语境也随故事连贯地演变，展现出相较于其他方法更优的身份保持、风格连贯性和故事层面的忠实度。

定量比较。 表 1 比较了 *DeCorStory*与 ConsiStory+ 上基于训练和免训练的基线方法。无需任何额外训练或微调，*DeCorStory* 便取得了与基于训练的方法相当或更优的性能，并在免训练方法中排名第一，获得最高的 CLIP-T（0.9001）和 CLIP-I（0.9134），同时取得最低的 DreamSim（0.1922）。与包括 ConsiStory、StoryDiffusion 和 1Prompt1Story 在内的免训练叙事基线方法相比，*DeCorStory*在所有指标上均持续提升，证实显式的嵌入去相关能够有效抑制帧间语义干扰。值得注意的是，相较于 1Prompt1Story 所取得的提升，证明了其收益并非仅来自简单的提示词拼

表1：定量比较。最佳和次佳结果分别以**粗体**和下划线标示。原版 SD1.5 和原版 SDXL 作为参考展示，不纳入此次比较。

方法	基础模型	无需训练	CLIP-T↑	CLIP-I↑	DreamSim↓	步数	内存（GB）↓	推理时间（秒）↓
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
BLIP-Diffusion	SD1.5	✗	0.7607	0.8863	0.2830	26	**7.75**	**1.9284**
文本反演	SD1.5	✗	0.8378	0.8229	0.4268	40	32.94	282.507
The Chosen One	SDXL	✗	0.7614	0.7831	0.4929	35	10.93	11.2073
PhotoMaker	SDXL	✗	0.8651	0.8465	0.3996	50	23.79	18.0259
IP-Adapter	SDXL	✗	0.8458	**0.9429**	**0.1462**	30	19.39	13.4594
ConsiStory	SDXL	✓	0.8769	0.8737	0.3188	50	34.55	34.5894
StoryDiffusion	SDXL	✓	0.8877	0.8755	0.3212	50	45.61	25.6928
朴素提示词重加权（NPR）	SDXL	✓	0.8411	0.8916	0.2548	50	16.04	17.2413
1Prompt1Story	SDXL	✓	0.8942	0.9117	0.1993	50	18.70	23.2088
DeCorStory（我们的方法）	SDXL	✓	**0.9001**	0.9134	0.1922	50	18.72	23.3521

表2：对 100 名参与者开展用户研究，根据人类偏好选出最佳的一致性文生图生成方法。

方法	原版 SD1.5	原版 SDXL	文本反演	The Chosen One	DreamShaper	NPR	1Prompt1Story	*DeCorStory*
(%)↑	2	4	8	10	16	18	19	**23**

表3：消融研究。我们评估了各个组件的贡献，包括 GS、SVR$^+$/SVR$^-$ 和 IPCA。

方法	CLIP-T↑	CLIP-I↑	DreamSim↓
不使用 GS	0.8874	0.9021	0.2375
不使用 SVR	0.8896	0.9058	0.2312
不使用 IPCA	0.8912	0.9084	0.2259
仅使用 GS	0.8825	0.8993	0.2418
仅 SVR	0.8859	0.9046	0.2356
GS + SVR	0.8926	0.9091	0.2237
GS + IPCA	0.8918	0.9102	0.2241
完整方法（GS + SVR +IPCA）	**0.9001**	**0.9134**	**0.1922**

接，而是在推理时几乎不会增加额外的计算开销。

4.3. 用户研究

我们通过用户研究评估了八种方法的感知质量与一致性，包括原始扩散基线、个性化方法和无需训练的故事生成模型。遵循 ConsiStory+，我们使用30组提示词，每组包含四帧来生成图像序列，并由100名参与者选出在身份一致性、提示对齐和视觉多样性之间取得最佳平衡的序列。如表2所示，*DeCorStory* 获得了最高的偏好率，优于个性化方法以及 NPR 和 1Prompt1Story 等无需训练的基线方法。这一结果与我们的定量评估一致，表明显式的嵌入去相关结合 SVR 和 IPCA，能够生成在人类看来更加连贯且忠实于语义的图像序列。

4.4. 消融研究

表3报告了分析 GS、SVR 和 IPCA 各自贡献的消融研究。移除 GS 会导致 CLIP-T 和 DreamSim 明显下降，表明帧间语义干扰加剧；而禁用 SVR 则会削弱共享语义与帧特定语义之间的平衡，从而降低身份一致性。移除 IPCA 会进一步损害跨帧身份保持，尤其是在外观变化的情况下。GS 与 SVR 结合已能实现稳定的场景级语义，但完整配置（GS + SVR +IPCA）在所有指标上均取得最佳表现（CLIP-T = 0.9001，CLIP-I = 0.9134，DreamSim = 0.1922）。这些结果表明，嵌入去相关、语义重加权和身份感知注意力相辅相成，使*DeCorStory* 能够实现出色的提示对齐和身份一致性。

5. 结论

我们提出了 *DeCorStory*，这是一个免训练框架，通过减少帧级提示词之间的语义干扰，增强多帧叙事。我们的方法结合了 Gram–Schmidt 提示词嵌入去相关（GS）、奇异值重加权（SVR）和身份保持交叉注意力（IPCA），将相关的提示词嵌入解耦，增强帧特定语义，并在扩散过程中保持身份稳定。*DeCorStory* 完全在推理时运行，无需重新训练模型。与现有免训练基线方法相比，它在提示词–图像对齐、身份一致性和叙事连贯性方面表现更优，凸显了嵌入去相关作为文本到图像故事生成有效原则的价值。

6. 参考文献

[1] Yoad Tewel、Omri Kaduri、Rinon Gal、Yoni Kasten、Lior Wolf、Gal Chechik 和 Yuval Atzmon，《免训练一致性文生图》，*ACM* 图形学汇刊，第 43 卷，第 4 期，第 1–18 页，2024 年。

[2] Robin Rombach、Andreas Blattmann、Dominik Lorenz 等，《利用潜在扩散模型进行高分辨率图像合成》，发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第10674–10685页。

[3] Zhen Li、Mingdeng Cao、Xintao Wang 等，《PhotoMaker：通过堆叠身份嵌入定制逼真人像照片》，发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第8640–8650页。

[4] Tao Liu、Kai Wang、Senmao Li、Joost van de Weijer、Fahad Shahbaz Khan、Shiqi Yang、Yaxing Wang、Jian Yang 和 Ming-Ming Cheng，《一提示词一故事：使用单一提示词实现免费的一致性文生图》，发表于第*13*届国际学习表征会议，2025年。

[5] Mehdi Cherti、RomainBeaumont、RossWightman等，“对比语言-图像学习的可复现缩放定律”，载于*2023 IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），2023年，第2818–2829页。

[6] Maxime Oquab、Timothee Darcet、Theo Moutakanni等，“Dinov2：无监督地学习稳健视觉特征”，2023年。

[7] Rinon Gal、Yuval Alaluf、Yuval Atzmon、Or Patashnik、Amit Haim Bermano、Gal Chechik和Daniel Cohen- Or，“一图一词：使用文本反演实现个性化文生图”，载于第*11*届国际学习表征会议论文集，2022年。

[8] Nataniel Ruiz、Yuanzhen Li、Varun Jampani、Yael Pritch、Michael Rubinstein和Kfir Aberman，“DreamBooth：面向主体驱动生成微调文本到图像扩散模型”，*arXiv*预印本 *arXiv:2208.12242*，2023年。

[9] Hu Ye、Jun Zhang、Sibo Liu、XiaoHan 和 Wei Yang，《IP-Adapter：适用于文本到图像扩散模型的文本兼容图像提示适配器》，*arXiv* 预印本 *arXiv:2308.06721*，2023。

[10] Yupeng Zhou、Daquan Zhou、Ming-Ming Cheng、Jiashi Feng 和 Qibin Hou，《Storydiffusion：用于长程图像与视频生成的一致性自注意力》，神经信息处理系统进展，第 37 卷，第 110315–110340 页，2024 年。

[11] Ron Mokady、Amir Hertz、Kfir Aberman、Yael Pritch、Shai Peleg 和 Daniel Cohen-Or，《用于文本驱动的图像到图像转换的即插即用扩散特征》，*arXiv* 预印本 *arXiv:2211.12572*，2023。

[12] Alvin Heng 和 Harold Soh，《选择性遗忘：一种在深度生成模型中实现遗忘的持续学习方法》，发表于神经信息处理系统进展，2024 年，第 36 卷。

[13] Yoad Tewel、Omri Kaduri、Rinon Gal 等，《无需训练的一致性文生图》，*ACM* 图形学汇刊（*TOG*），第 43 卷第 4 期，第 1–18 页，2024 年。

[14] Dongxu Li、Junnan Li 和 Steven 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”，神经信息处理系统进展，第 36 卷，第 30146–30166 页，2023。

[15] Rinon Gal、Yuval Alaluf、Yuval Atzmon、Or Patashnik、Amit Haim Bermano、Gal Chechik 和 Daniel Cohen-Or，“一张图胜过千言：使用文本反演个性化文本到图像生成”，发表于第十一届国际学习表征会议。

[16] HuYe、Jun Zhang、Sibo Liu、Xiao Han 和WeiYang，“IP-Adapter：适用于文本到图像扩散模型的文本兼容图像提示适配器”，*arXiv* 预印本 *arXiv:2308.06721*，2023。

[17] Omri Avrahami、Amir Hertz、Yael Vinker、Moab Arar、Shlomi Fruchter、Ohad Fried、Daniel Cohen-Or 和 Dani Lischinski，“The Chosen One：文本到图像扩散模型中的一致性角色”，发表于*ACM SIGGRAPH 2024* 会议论文集，2024，第 1–12 页。

[18] Jack Hessel、Ari Holtzman、Maxwell Forbes、Ronan Le Bras 和 Yejin Choi，“CLIPScore：一种用于图像描述的无参考评估指标”，发表于*2021*年自然语言处理实证方法会议论文集，2021年，第 7514–7528 页。

[19] Stephanie Fu、Netanel Tamir、Shobhita Sundaram、Lucy Chai、Richard Zhang、Tali Dekel 和 Phillip Isola，“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。

[20] Nikita Selin，“CarveKit：自动化高质量背景移除框架”，2023年。

A. 附录

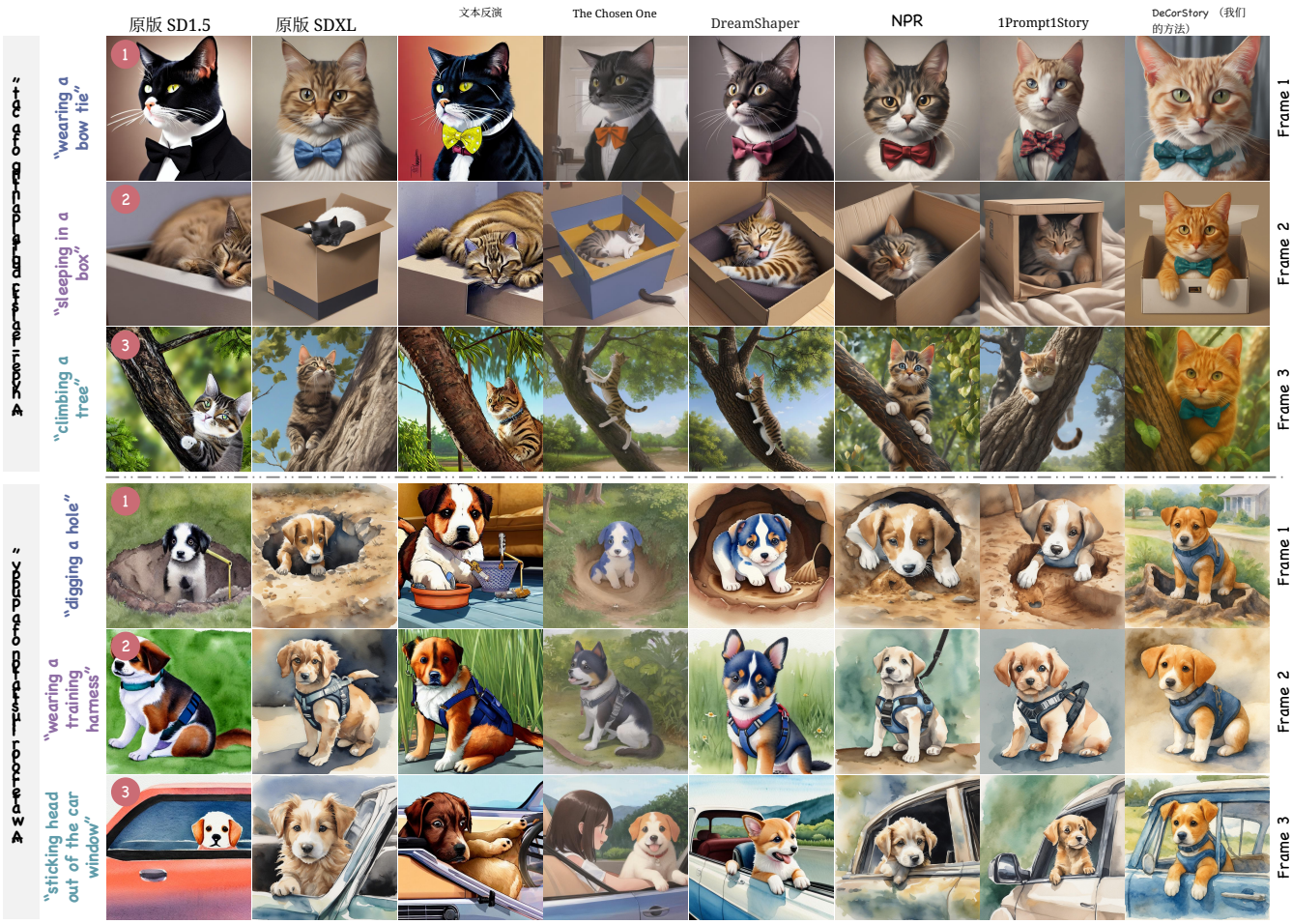


图 A1：定性比较结果。

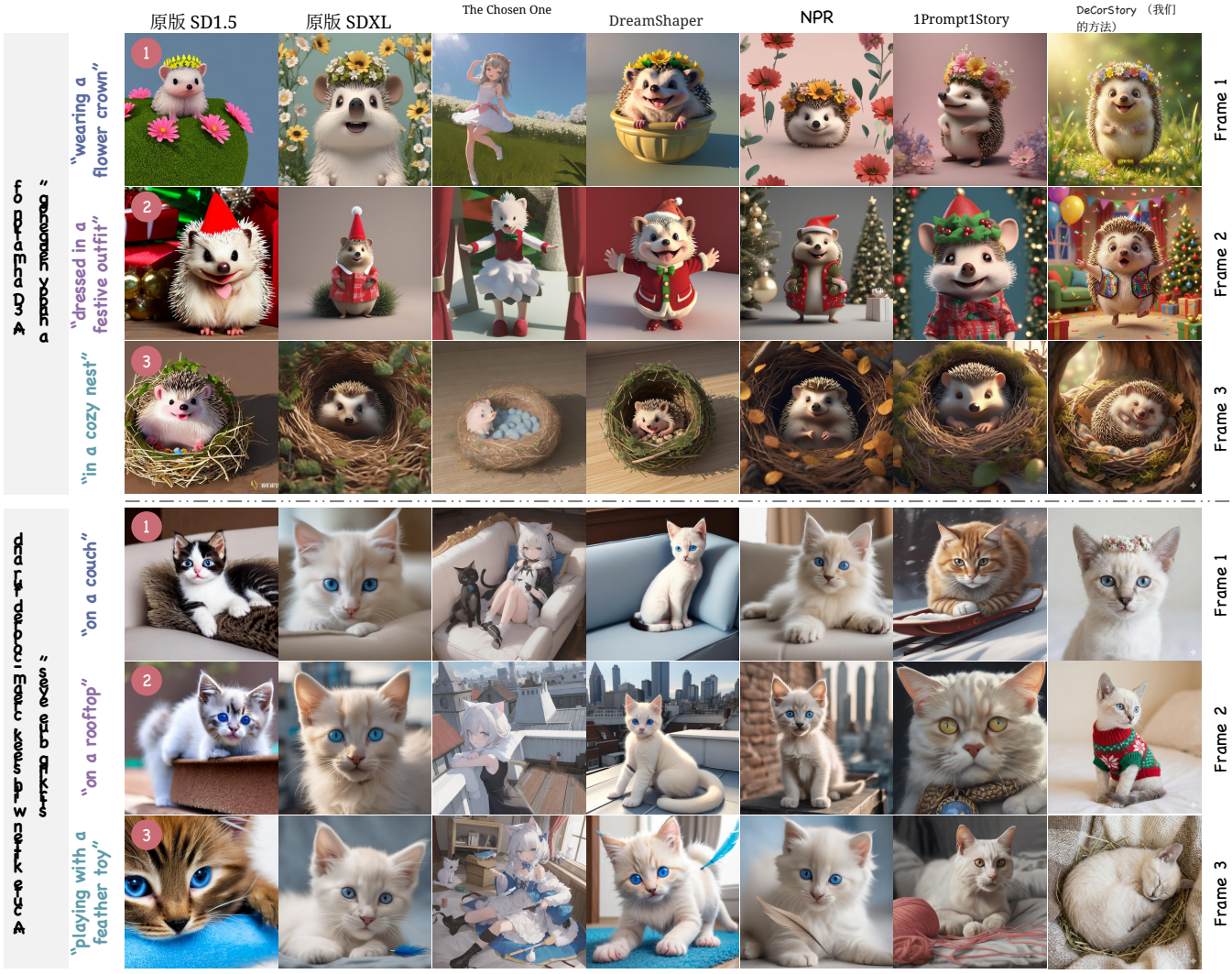

原版 SD1.5
原版 SDXL
The Chosen One
DreamShaper
NPR
1Prompt1Story
DeCorStory （我们的方法）
"wearing a flower crown"
"dressed in a festive outfit"
"in a cozy nest"
"on a couch"
"on a rooftop"
"playing with a feather toy"
Frame 1
Frame 2
Frame 3

图 A2：定性比较结果。

算法 1*DeCorStory*推理流水线

输入： 提示词 $\{P_0, P_1, \ldots, P_N\}$；编码器 τ_ξ；UNet ϵ_θ；解码器 D；扩散步数 T；SVR 参数（α，β，α'，β'）

输出： 生成帧 $\{I_j\}_{j=1}^{N}$

阶段 1：提示词编码与拼接

$C \leftarrow \tau_\xi([P_0;\ P_1;\ \ldots;\ P_N])$

$= [c_{\mathrm{SOT}}, c_{P_0}, c_{P_1}, \ldots, c_{P_N}\ \ c_{\mathrm{EOT}}]$

阶段 2：格拉姆–施密特去相关

$\tilde{c}_{P_1} \leftarrow c_{P_1} / \|c_{P_1}\|_2$

对 $k = 2$**到** N **循环**

$$\tilde{c}_{P_k} \leftarrow \frac{c_{P_k} - \sum_{i<k} \langle c_{P_k}, \tilde{c}_{P_i} \rangle \tilde{c}_{P_i}}{\left\| c_{P_k} - \sum_{i<k} \langle c_{P_k}, \tilde{c}_{P_i} \rangle \tilde{c}_{P_i} \right\|_2}$$

结束循环

$\tilde{C} \leftarrow [c_{\mathrm{SOT}}, c_{P_0}, \tilde{c}_{P_1}, \ldots, \tilde{c}_{P_N}, c_{\mathrm{EOT}}]$

对 $j = 1$**到** N **循环**

阶段 3：针对帧 j进行SVR重加权

$X^{\mathrm{exp}} \leftarrow [\tilde{c}_{P_j};\ c_{\mathrm{EOT}}]$; $(X^{\mathrm{exp}})^\top = U \Sigma V^\top$

$\hat{\Sigma} \leftarrow \mathrm{diag}(\hat{\sigma}_\ell)$ 其中 $\hat{\sigma}_\ell = \beta e^{\alpha \sigma_\ell} \sigma_\ell$

$[\hat{c}_{P_j}; \hat{c}_{\mathrm{EOT}}] \leftarrow (U \hat{\Sigma} V^\top)^\top$

对每个 $k \neq j$ **执行**

$X_k^{\mathrm{sup}} \leftarrow [\tilde{c}_{P_k}; \hat{c}_{\mathrm{EOT}}]$; $(X_k^{\mathrm{sup}})^\top = U_k \quad \hat{\Sigma}_k V_k^\top$

$\tilde{\Sigma}_k \leftarrow \mathrm{diag}(\tilde{\sigma}_{k,\ell})$，其中 $\tilde{\sigma}_{k,\ell} = \beta' e^{-\alpha \quad ' \hat{\sigma}_{k,\ell} \hat{\sigma}_{k,\ell}}$

$[\tilde{c}_{P_k}^{\downarrow};\ \tilde{c}_{\mathrm{EOT}}] \leftarrow (U_k \tilde{\Sigma}_k V_k^\top)^\top$

结束循环

$\tilde{C}^{(j)} \leftarrow [c_{\mathrm{SOT}}, c_{P_0}, \tilde{c}_{P_1}^{\downarrow}, \ldots\ldots, e_{P_j}, \ldots\ldots, \tilde{c}_{P_N}^{\downarrow}\ \ \tilde{c}_{\mathrm{EOT}}]$

阶段 4：使用 IPCA 去噪

初始化 $z_T \sim \mathcal{N}(0, I)$

对于 $t = T$ **到** 1 **执行**

$K_t \leftarrow \ell_K(\tilde{C}^{(j)})$; $\quad V_t \leftarrow \ell_V(\tilde{C}^{(j)})$

$K_t^{\mathrm{id}} \leftarrow K_t$; $\quad V_t^{\mathrm{id}} \leftarrow V_t$

将所有非P_0 词元行置零于 K_t^{id}，$V_{\mathrm{id}t}$

$\widetilde{K}_t \leftarrow$ 拼接$(K_t,\ K_t^{\mathrm{id}})$; $\quad \widetilde{V}_t \leftarrow$ 拼接$(V_t,\ V_{\mathrm{id}t})$

输出$_t \leftarrow$softmax$(Q_t \widetilde{K}_t^\top / \sqrt{d})\, \widetilde{V}_t$

$z_{t-1} \leftarrow \epsilon_\theta(z_t,\ t,\ \tilde{C}^{(j)}$; 输出$_t)$

循环结束

$I_j \leftarrow D(z_0)$

循环结束

返回 $\{I_j\}_{j=1}^{N}$