arXiv:2602.15539v1 [cs.CV] 17 Feb 2026

主体与风格 LoRA 的动态免训练融合

Qinglong Cao[1,2] Yuntian Chen[2*] Chao Ma[1] Xiaokang Yang[1]上海交通大学 [2]宁波东方理工大学

*通讯作者

摘要

近期研究探索了组合多个 *LoRA*，以同时生成用户指定的主体和风格。然而，现有大多数方法采用静态统计启发式方法融合 *LoRA* 权重，这偏离了 *LoRA* 通过学习自适应特征调整来发挥作用的初衷，也忽略了采样输入的随机性。为解决这一问题，我们提出了一种贯穿整个生成过程的动态免训练融合框架。在前向传播过程中，我们在每个应用*LoRA*的层上，动态计算基础模型的原始特征与主体 *LoRA* 和风格 *LoRA* 分别生成的特征之间的 *KL* 散度，并自适应地选择最合适的权重进行融合。在反向去噪阶段，我们进一步通过动态应用基于 *CLIP* 和 *DINO* 分数等目标指标得出的基于梯度的校正，优化生成轨迹，从而提供持续的语义和风格引导。通过在整个扩散时间线上整合这两种互补机制——特征层面的选择与指标引导的潜变量调整——我们的方法无需重新训练，即可动态实现连贯的主体-风格合成。在多种主体与风格组合上的大量实验表明，无论是定性还是定量评估，我们的方法均持续优于最先进的 *LoRA*融合方法。

1. 引言

扩散模型在广泛的生成任务中展现了卓越的性能 [7, 14, 20,25, 26, 40, 45]。其中，个性化图像生成 [30, 34, 41] 日益受到关注，因为它要求模型合成高质量图像，以体现用户指定的内容或风格。这里，内容指语义结构和主体身份，而风格则体现色彩、纹理和图案等视觉属性。尽管在以内容或风格单独为条件生成图像方面已取得显著进展，但要生成能够忠实融合特定主体和特定风格的图像，仍然是一个充满挑战且尚未解决的问题。

近年来，低秩适配（LoRA） [18] 已成为一种广受欢迎且用途广泛的参数高效微调技术，尤其适用于个性化生成任务。利用 LoRA 的模块化特性，近期研究探索了融合独立微调的 LoRA，以联合生成特定风格的特定主体。例如，ZipLoRA [32] 提出利用系数向量，在每个应用LoRA的层中融合内容和风格 LoRA。与之不同，B-LoRA [11]研究了不同 LoRA 层的影响，并发现修改两个不同的 LoRA 层能够有效控制生成图像的内容和风格。此外，K-LoRA 着眼于 LoRA 权重的内在特征， [28] 通过比较权重中的 Top-K 元素来选择每一层的 LoRA。尽管如图1 (a) 所示，这些方法在 LoRA 融合方面展现出良好的性能，但其核心策略仍基于 LoRA 权重的统计特性，这偏离了 LoRA 的初衷——学习额外特征以适应不同功能。这种偏离表明，有效融合的关键在于微调特征本身，而非仅仅是 LoRA 权重。更重要的是，以往的静态融合方法忽略了生成过程中采样潜变量输入的随机性，限制了其适应能力，导致性能欠佳。

受此启发，如图1 （b）所示，我们提出了一种贯穿整个扩散过程的动态免训练融合框架。不同于以往将 LoRA融合视为静态权重混合的方法，我们将融合重新诠释为一种与扩散模型生成动态相一致的、动态的表示感知决策过程。具体而言，先前的方法 [28]认为，LoRA权重的绝对值表明了其在扩散过程中的重要性。相比之下，我们提出，LoRA引起的特征变化是其影响更直接且更关键的指标。在前向传播期间的每个应用LoRA的层中，我们分别计算风格 Lo-

*通讯作者。

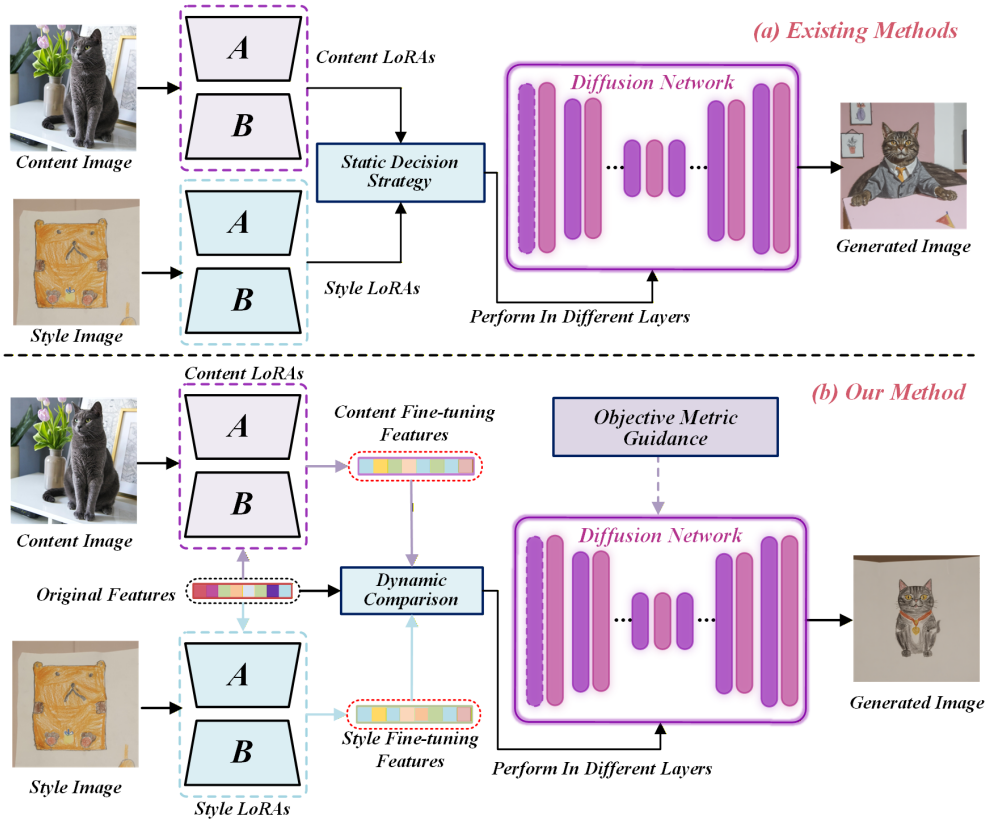

图 1. (a) 现有方法直接依赖 LoRA 权重的属性来实现融合。(b) 我们的方法在前向传播中整合特征级选择，并在反向过程中进行潜变量级精炼，从而实现动态免训练 LoRA 融合。

RA和内容LoRA产生的微调特征。为了更好地量化特征变化的程度，我们利用微调特征与原始特征之间的Kullback-Leibler（KL散度），并根据分布变化的幅度确定每一层更适合使用哪一个 LoRA。这样一来，尽管采样输入各不相同，每个应用LoRA的层都能自适应地保留最具代表性的内容和风格特征。

至关重要的是，这种逐层选择仅构成了我们动态融合策略的前半部分。为了确保全局连贯性和高保真合成，我们进一步引入了一种由指标驱动的精炼机制，使其在整个反向去噪过程中动态运行。具体而言，我们首先分别使用主体与风格LoRA生成两张参考图像，作为语义和风格锚点。在去噪过程中，我们在每个时间步使用 CLIP 等客观指标，将中间预测与这些参考图像进行比较 [29] 以及 DINO [5] 分数，分别量化内容保真度和风格一致性。随后，我们利用这些客观指标计算综合引导信号，并将其梯度用于引导潜在轨迹，使其朝着更符合预期主体—风格组合的区域移动。这种无需训练的逐步校正能够持续进行精炼，无需任何额外监督。

通过将前向传播中的特征级选择与反向过程中的潜变量级精炼相结合，我们的方法有效融合了主体与风格LoRA，无需训练即可生成高质量图像，同时保留主体保真度和风格准确性。本文的主要贡献总结如下:

•我们提出了一种动态免训练 LoRA 融合框架，将范式从静态的权重级启发式方法转变为贯穿整个生成过程、能够适应输入的表征感知决策。

• 一种基于 KL 散度的策略根据特征扰动幅度，在每一层自适应地选择信息量最大的 LoRA。此外，客观指标分数还可在去噪过程中提供度量引导细化，从而提升语义和风格保真度。

•我们的方法完全免训练、即插即用，并在多个基准测试中的多种主体–风格组合上展现出卓越性能，且无需重新训练或额外监督。

2. 相关工作

用于定制生成的扩散模型。随着扩散模型的快速发展 [4, 9, 15–17, 25, 42],

许多研究者提出了多种方法，用于微调大规模扩散模型以实现定制生成，旨在根据语言描述生成用户指定的主体或风格图像。例如，文本反演 [12]侧重于优化单个词嵌入，以捕捉独特且多样的概念。DreamBooth [30] 设计包含唯一标识符的文本提示，从而更有效地生成具有所需主体的图像。CustomDiffusion [23]微调扩散模型中的交叉注意力层，以同时学习多个概念。此外，一些方法 [2, 33, 37, 38] 无需额外训练即可实现定制生成，但通常仅针对特定的单一任务。近年来，LoRA [18] 和 StyleDrop [34] 等参数高效微调技术因能够通过低秩适配微调模型而日益受到青睐，因此在定制生成方面颇具吸引力。

用于图像生成的 LoRA 组合。随着 LoRA 应用日益普及， [44, 46, 47],许多关于 LoRA 组合的研究相继提出。一些方法 [8, 13, 19, 39] 专注于融合多个物体 LoRA，使扩散模型能够生成各种新概念，并通过遮罩策略替换这些物体。与此同时，还有一些先进方法致力于内容-风格LoRA融合。例如，Mixture-of-Subspaces [36] 设计了可学习的混合器权重，以融合各种 LoRA；ZipLoRA [32]利用不同层中的合并向量，线性组合主体与风格LoRA；B-LoRA [11] 研究了不同 LoRA 层的影响，并发现修改两个不同的层能够有效控制生成图像的内容与风格；K-LoRA [28] 则通过比较不同 LoRA 权重中的 Top-K 元素，为每一层选择合适的 LoRA。尽管这些方法展现出良好的性能，但它们直接依赖于 LoRA 权重的特性。LoRA 的初衷是学习额外特征，以适应多种任务。以往方法忽略了输入随机性。因此，基于微调特征的动态融合方法可能更有效。为此，我们提出一种动态免训练 LoRA融合范式，能够根据输入自适应地做出表征感知决策。

3. 预备知识

扩散模型。 扩散模型 [1, 6, 22, 31] 在各种生成任务中都展现出了令人瞩目的性能。它们主要由前向加噪过程和反向去噪过程构成。在前向过程中，原始图像通过逐步添加噪声，逐渐转化为高斯噪声。在反向过程中，以配对文本提示为条件，扩散网络（通常为 U-Net）从随机采样的纯噪声开始，逐步对含噪输入进行去噪。推理时，训练好的扩散网络根据给定的文本输入实现文生图。

LoRA。 低秩适配（LoRA [18]）是一种轻量级微调技术，最初为大型语言模型和扩散模型开发。LoRA 并不更新完整的参数矩阵 W_0，而是利用了这样一个观察：更新量 $\Delta W \in \mathbb{R}^{m\times n}$ 通常位于低维子空间中。具体而言，将 $\Delta W = BA$ 分解为 $B \in \mathbb{R}^{m\times r}$ 和 $A \in \mathbb{R}^{r\times n}$，其中 $r \ll \min(m, n)$，并且仅学习 A，B，而 W_0 保持固定。因此，经过调优的模型使用权重 $W_0 + BA$。在我们的设定中，令 D 为一个扩散模型，其基础权重为 W_0。为了捕捉一个新概念，我们训练一对 LoRA ΔW_x，使适配模型为 $D_x = W_0 + \Delta W_x$。

动态表征引导。 LoRA 的主要目标是学习细粒度的特征调整，从而引导扩散模型表现出特定行为。为了利用这一能力，我们首先生成由内容 LoRA 和风格 LoRA 产生的修改后特征。尽管以往工作 [28] 使用 LoRA 权重更新的静态幅度作为衡量其影响的代理指标，但我们认为，由于输入是随机采样的，特征分布的实际动态变化能够提供更直接且更易解释的影响度量。因此，我们通过原始特征与 LoRA 修改后特征之间的 Kullback–Leibler（KL）散度来量化这种变化，并利用它在每个层中选择信息量最大的分支进行融合。

此外，CLIP 和 DINO 分数等客观指标为评估 LoRA融合的质量提供了有效方法。分数越高，表示与预期语义或风格的匹配度越高，因此可作为综合引导信号来引导扩散网络。基于 CLIP 指标的引导分数 R 计算如下：

$$R(\hat{x}_0) = 1 - S_{\text{CLIP}}(x_{\text{ref}}, \hat{x}_0), \tag{1}$$

其中， S_{CLIP} 计算 CLIP 相似度分数， x_{ref} 是参考图像，而 $\hat{x}_0$ 是步骤 t处预测的原始图像。

引导分数充当残差，可视为虚拟观测值 $\hat{}$ [21]，其对应的 $R = 0$ 和虚拟似然如下：

$$p(\hat{R} = 0 \mid x_t) = \mathcal{N}(0 \mid R(x_0), \sigma_r^2 I), \tag{2}$$

其中， x_t 是时间步 t处的中间结果， σ_r 是预定义常数，用于控制虚拟观测的约束强度。尽管残差在生成过程中会被最小化，但并不能保证其达到零。为将这一引导融入扩散过程，我们应用贝叶斯规则：

$$p(x_t \mid \hat{R} = 0) = \frac{p(x_t)\, p(\hat{R} = 0 \mid x_t)}{p(\hat{R} = 0)}. \tag{3}$$

对数似然关于 x_t的梯度为：

$$\nabla_{x_t} \log p(x_t \mid \hat{R}=0) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(\hat{R}=0 \mid x_t), \tag{4}$$

其中，第一项是扩散模型预测的标准得分函数 [35]。对于第二项，将式2 代入可得：

$$\nabla_{x_t} \log p(\hat{R}=0 \mid x_t) = -\frac{1}{\sigma_r^2} \nabla_{x_t} \|R(x_0)\|_2^2 \approx -\frac{1}{\sigma_r^2} \nabla_{x_t} \|R(\hat{x}_0)\|_2^2, \tag{5}$$

其中，我们用 $\hat{x}_0$近似 x_0。实际上，由于引导分数介于 $[0,1]$, 之间，我们将其简化为：

$$\nabla_{x_t} \log p(\hat{R}=0 \mid x_t) \propto -\nabla_{x_t} R(\hat{x}_0). \tag{6}$$

因此，为了实现引导，对去噪步作如下修改：

$$x_{t-1} = x_{t-1}^{\text{ori}} - m \nabla_{x_t} R(\hat{x}_0), \tag{7}$$

其中，x_{t-1}^{ori} 是去噪过程中第 t 步的原始输出，$\hat{x}_0$ 是第 t 步预测的原始图像，而 m是预定义的缩放因子。

4. 方法

如图 2所示，我们的框架通过动态融合内容 LoRA 和风格 LoRA，无需任何额外训练，即可引导基础扩散模型 D 生成指定风格的指定主体。设基础模型 D 由层 W_0^i 的预训练权重 i组成。将具有权重更新 L_x的 LoRA $\{\Delta W_x^i\}$ 应用于该模型，可得到适配模型：

$$D_{L_x} = D \odot L_x = W_0 + \Delta W_x. \tag{8}$$

在实验设置中，给定内容 LoRA 权重 $L_c\{\Delta W_c^i\}$、风格 LoRA 权重 $L_s\{\Delta W_s^i\}$ 和基础模型 D。我们不采用启发式方法合并 LoRA 权重，而是在整个扩散过程中进行动态的表征感知融合：在每个应用了 LoRA 的层中，选择能引起更显著特征扰动的分支；在去噪过程中，则利用客观指标分数优化潜在轨迹。这种两阶段的动态免训练策略，使主体与风格能够在一次前向传播中实现连贯融合。

4.1. 前向传播过程中的特征级选择

先前的工作将 LoRA 权重更新的绝对值用作其在扩散模型中重要性的代理指标 [28]。然而，LoRA 的核心功能是诱导特征级调整，而不仅仅是改变权重幅值，且静态方法忽略了采样输入的随机性。受这一原则启发，我们明确考察内容与风格LoRA的权重更新 ΔW_c^i和 ΔW_s^i 如何影响基础网络的第 i 层。具体而言，我们将这些更新应用于基础权重 W_0^i，并计算相应的微调特征图：

$$\hat{F}_c^{i+1} = \left(W_0^i + \Delta W_c^i\right) F_i, \tag{9}$$

$$\hat{F}_s^{i+1} = \left(W_0^i + \Delta W_s^i\right) F_i, \tag{10}$$

其中，F_i 表示层 i 的原始特征。

为量化这些修改的影响，我们计算每个微调后的特征分布与原始特征分布 F_{i+1} 之间的 KL 散度：

$$d_c^i = \mathrm{KL}\left(\hat{F}_c^{i+1} \,\|\, F_{i+1}\right), \tag{11}$$

$$d_s^i = \mathrm{KL}\left(\hat{F}_s^{i+1} \,\|\, F_{i+1}\right). \tag{12}$$

随后，我们比较 d_c^i 和 d_s^i，以确定哪种调整会引起更显著的特征变化：

$$F^{i+1} = \begin{cases} \hat{F}_c^{i+1}, & \text{if } d_c^i \geq d_s^i, \\ \hat{F}_s^{i+1}, & \text{otherwise.} \end{cases} \tag{13}$$

通过在每一层执行这一选择，我们保留影响最大的内容或风格信息，从而实现主体与风格 LoRA 的动态免训练融合。

与静态且与输入无关的基于权重的融合策略相比，我们的基于特征的方法能够动态适应输入：随着提示词的变化，特征分布也会随之变化，进而改变融合决策。这种输入条件机制使我们的方法能够灵活处理具有多样化且不断变化需求的生成任务。

4.2. 去噪阶段的潜变量级精炼

如预备知识部分所述，CLIP 和 DINO 分数等客观指标能够有效评估 LoRA 融合的质量，分数越高表示与目标内容或风格的对齐程度越好。为获得引导分数，我们首先利用内容 LoRA ΔW_c 和风格 LoRA ΔW_s，结合内容描述 l_c 和风格描述 l_s，生成新的参考内容图像 I_c^{ref}和参考风格图像 I_s^{ref}：

$$I_c^{\text{ref}} = (W_0 + \Delta W_c)[l_c] \tag{14}$$

$$I_s^{\text{ref}} = (W_0 + \Delta W_s)[l_s] \tag{15}$$

给定第 t 步去噪后的原始输出 x_{t-1}^{ori} 和第 t 步预测的原始图像 $\hat{x}_0$，我们计算 CLIP 和 DINO 分数 [32]，以评估预测性能。这些分数通过从生成图像和参考图像中提取嵌入并计算余弦相似度得到：

$$S_{\text{CLIP}}^{\text{content}} = \mathrm{Sim}_{\cos}\left(E_{\text{CLIP}}(I_c^{\text{ref}}), E_{\text{CLIP}}(\hat{x}_0)\right) \tag{16}$$

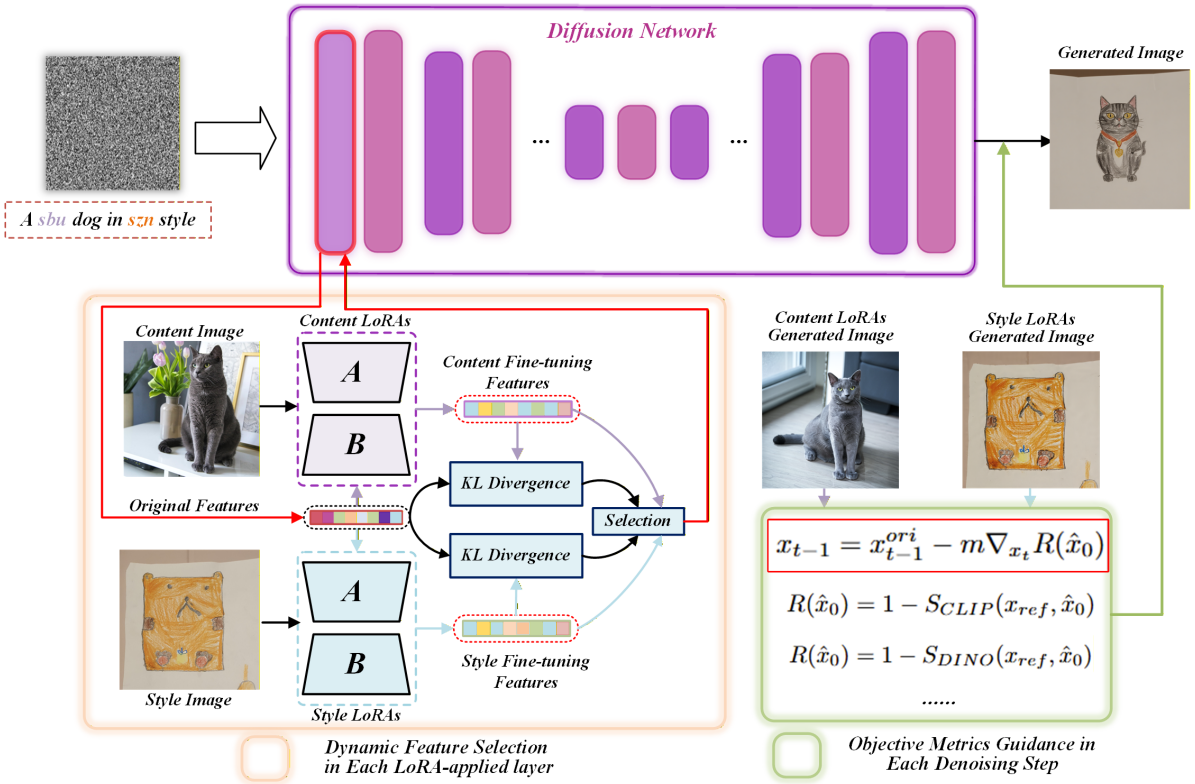

图2. 我们方法的概述。通过基于表示扰动进行动态特征选择，并在整个去噪过程中应用基于度量的细化，我们的框架实现了主体与风格LoRA的免训练融合。

$$S_{\text{CLIP}}^{\text{style}} = \text{Sim}_{\cos}\big(E_{\text{CLIP}}(I_s^{\text{ref}}), E_{\text{CLIP}}(\hat{x}_0)\big) \tag{17}$$

$$S_{\text{DINO}}^{\text{style}} = \text{Sim}_{\cos}\big(E_{\text{DINO}}(I_s^{\text{ref}}), E_{\text{DINO}}(\hat{x}_0)\big) \tag{18}$$

其中 E_{CLIP} 和 E_{DINO} 分别表示 CLIP [29] 和 DINO 的图像编码器 [5], 。为计算最终引导分数，我们对这三项指标赋予相同权重：

$$R(\hat{x}_0) = 1 - \frac{S_{\text{CLIP}}^{\text{content}} + S_{\text{CLIP}}^{\text{style}} + S_{\text{DINO}}^{\text{style}}}{3} \tag{19}$$

结合引导分数，并通过经验确定缩放因子$m = 10$，我们按如下方式引导扩散过程：

$$x_{t-1} = x_{t-1}^{\text{ori}} - m\nabla_{x_t} R(\hat{x}_0) \tag{20}$$

这种免训练细化会在每个去噪步动态应用，提供持续、客观的反馈，引导生成结果同时趋向主体保真度和风格准确性——无需重新训练或额外监督。

总之，我们的方法通过在整个扩散过程中运作，实现动态免训练的主体–风格LoRA融合：它基于KL 散度在前向传播中执行动态特征选择，以在每一层保留最具信息量的表示，并在去噪期间应用指标引导的潜变量精炼，使用客观指标反馈来确保全局语义与风格一致性。这种双阶段、输入自适应策略无需重新训练，却能够对独立训练的LoRA实现高保真、即插即用的组合。

方法	风格相似度 ↑	CLIP分数 ↑	DINO分数 ↑
Direct	48.9%	66.6%	43.0%
B-LoRA [11]	58.0%	63.8%	30.6%
ZipLoRA [32]	60.4%	64.4%	35.7%
K-LoRA [28]	58.7%	69.4%	46.9%
Ours	63.0%	78.5%	43.3%

表1。对齐结果比较。Direct 表示直接算术合并。

5. 实验

我们在先前方法（包括 K-LoRA、ZipLoRA 和 B-LoRA）所确立的实验设置下评估所提出的 LoRA融合方法。具体而言，我们的方法同时应用于 Stable Diffusion XL v1.0 基础模型和 FLUX 模型。

数据集。 在训练局部LoRA时，我们遵循先前工作 [28, 32]的惯例。为训练内容LoRA，我们从DreamBooth

方法	用户偏好	GPT-4o 反馈	Qwen2.5-VL 反馈
ZipLoRA [32]	13.80%	20.13%	3.40%
B-LoRA [11]	21.89%	11.67%	9.11%
K-LoRA [28]	11.11%	12.56%	21.82%
Ours	53.20%	55.64%	65.67%

表2. 用户研究结果、GPT-4o 与 Qwen2.5-VL 反馈的性能比较。

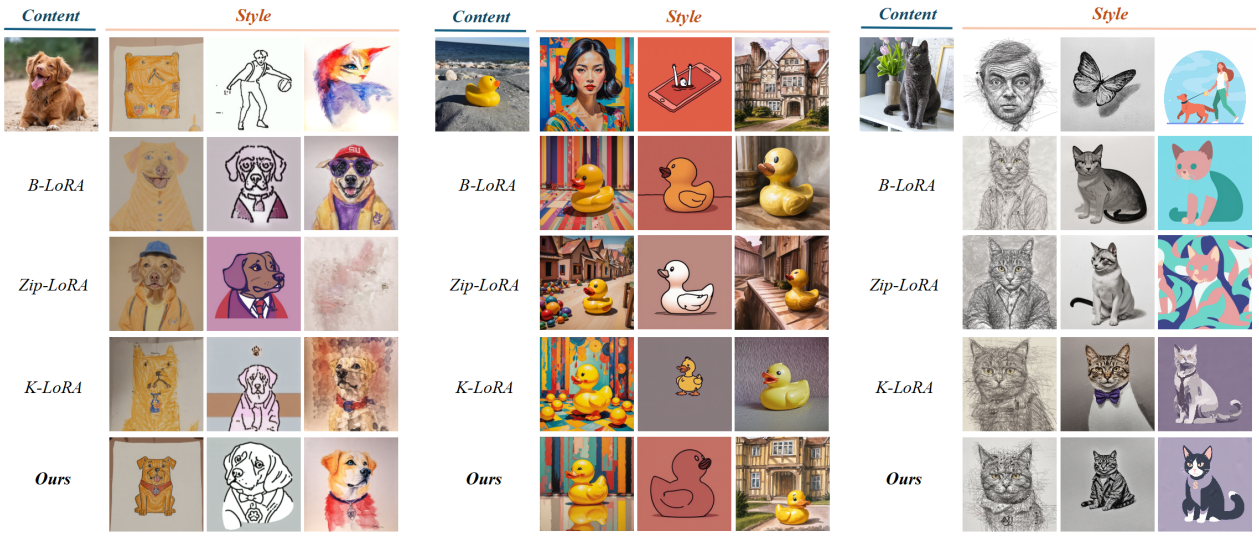

图3. **定性比较**。我们展示了由我们的方法以及所比较的先进生成方法生成的图像。通过在整个生成过程中做出输入自适应、表征感知决策，我们的方法有效实现了主体与风格LoRA的免训练融合。

数据集 [30], 中选取多样化的图像集，其中每个实例由4–5张图像表示。对于风格 LoRA，我们采用StyleDrop作者 [34], 引入的数据集，其中包含从古典艺术到现代创意风格的丰富风格范例。每个风格LoRA均使用单张参考图像进行训练。

实现细节。为获得用于实验的局部LoRA，我们采用 K-LoRA [28] 策略，使用秩设为 64 的低秩适配对 SDXL v1.0 基础模型进行微调。风格与内容两方面的 LoRA权重均使用 Adam 优化器优化，共 1000 步，批大小为 1，学习率为 5e-5。对于 FLUX 模型，我们使用从 HuggingFace 获取的公开可用、训练良好的社区 LoRA权重。FLUX 上的相应实验结果见补充材料。

5.1. 实验结果

5.1.1. 定量比较

为了客观评估我们无需训练的生成方法的性能，我们采用以往研究中常用的指标，包括风格相似度、CLIP分数和DINO分数，以评估生成图像的质量。

。按照以往的方法，我们随机选取了30组不同的内容-风格配对，每组包含10张图像，用于定量比较。具体而言，CLIP [29]用于评估风格对齐程度（风格相似度）和内容保留（*CLIP*分数），而 DINO [43] 则通过*DINO*分数衡量内容一致性。表 1 详细比较了我们的方法与现有最先进方法的表现。我们的方法在风格相似度（63.0%）和*CLIP*分数（78.5%）两项指标上均取得最佳表现。值得注意的是，与最强基线方法相比，我们的方法在*CLIP*分数上提升了9.1%。尽管我们的方法在*DINO*分数上并非表现最佳，但仍位列第二，展现出在风格与内容保真度之间的良好整体平衡。这些结果验证了我们动态融合方法的有效性。作为补充验证，我们在补充材料中报告了采用其他指标进行的进一步评估，这些结果同样证明了所提方法的有效性。

5.1.2. 用户研究和基于多模态大语言模型的评估

为进一步评估生成图像的感知质量（超越传统指标），我们开展了一项全面的

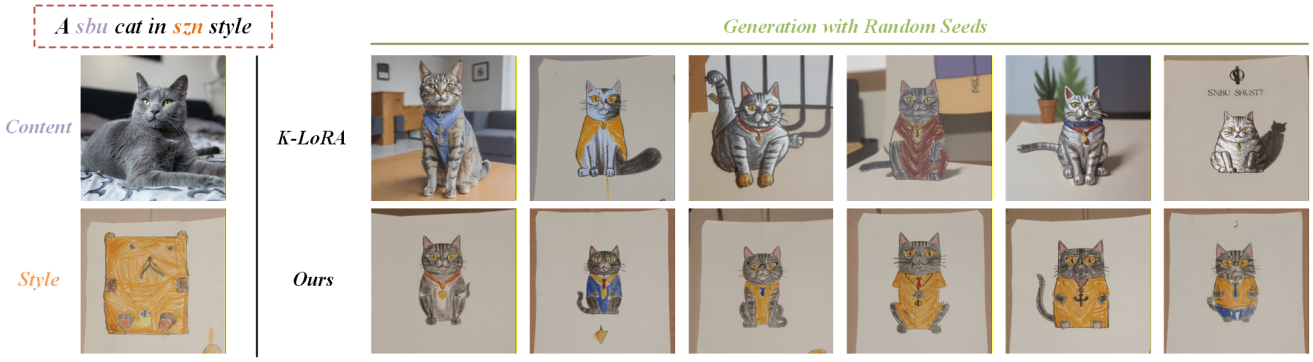

图4. **鲁棒性分析** . 我们展示了使用随机种子由我们的方法和K-LoRA生成的图像，以分析鲁棒性 .

方法	风格相似度 ↑	CLIP分数 ↑	DINO分数 ↑
基线方法	60.3%	75.6%	40.1%
仅使用 FLS	59.5%	75.9%	43.7%
仅使用潜在层细化	66.1%	78.9%	40.7%
我们的方法（FLS+潜在层细化）	64.0%	79.1%	43.4%

表 3. 不同组件的消融研究，包括特征级选择（FLS）和潜变量级精炼（LLR）。

散度	风格相似度 ↑	CLIP分数 ↑	DINO分数 ↑
KL	59.5%	75.9%	43.7%
JS	59.3%	75.3%	43.1%
余弦相似度	58.9%	75.4%	43.3%
点积	58.4%	74.9%	43.8%

表 4. 特征级选择中选择标准的消融研究，包括 Kullback-Leibler（KL）散度、Jensen-Shannon（JS）散度、余弦相似度和点积。

用户研究，并采用基于多模态大语言模型（MLLM）的评估。如表2所示，我们收集了人类偏好，并从两个强大的多模态大语言模型 GPT-4o [27] 和 Qwen2.5-VL [3] 获得自动反馈。在用户研究中，参与者需要从四种竞争方法生成的输出结果中选出自己偏好的图像。我们提出的方法获得了压倒性的青睐，赢得了总票数的 53.20%，表现优于所有基线方法。同样，在基于大语言模型的评估中，我们提出的方法始终排名最高，获得 GPT-4o 55.64% 的偏好，并在 Qwen2.5-VL 的评选中取得更高的 65.67%。这些结果不仅证实了生成图像在人类偏好方面的良好质量，也凸显了我们的方法在生成风格鲜明且语义连贯、与多模态模型高度契合的输出结果方面的有效性。更多设置细节见补充材料。

5.1.3. 定性比较

为了直观评估不同 LoRA融合方法的性能，我们在图3中展示了定性比较。总体而言，我们的免训练LoRA融合方法呈现出更优的视觉质量，能够有效保留内容和风格信息。相比之下，大多数现有方法虽然能够较好地保留内容，却难以忠实地呈现目标风格。例如，在第二行第三列中，B-LoRA 正确识别了“狗”的内容，却错误地采用了与参考风格不一致的粉色。K-LoRA 也存在类似问题，见第四行第三列。

缩放因子 m	风格相似度 ↑	CLIP分数 ↑	DINO分数 ↑
1	62.1%	72.8%	32.4%
5	64.8%	74.1%	32.3%
10	64.0%	79.1%	43.4%
20	66.8%	76.8%	35.1%

表5。潜变量级精炼中缩放因子 m 的消融研究。

此外，在倒数第二行中，K-LoRA 生成的“猫”头部和身体风格不一致，说明其未能实现全局风格一致性。在第八行中，K-LoRA 也未能保留油画风格。有趣的是，尽管 Zip-LoRA 是唯一具有可学习参数的方法，其表现却相对较差。它往往无法捕捉所需的风格（例如第四列中的“狗”），或生成语义不准确的内容。这些定性结果进一步证明，我们的动态融合策略在整个生成过程中均有效，无需额外训练即可持续生成视觉上协调一致的输出结果。

5.2. 消融研究

为了全面分析动态免训练融合框架中各个组件的作用，我们从三个方面进行了大量消融研究：(1) 特征级选择和潜变量级精炼的效果，(2) 用于选择的特征差异度量，(3) 引导缩放因子的影响。对于每种设置，我们随机抽取 25 组主体–风格组合，以确保

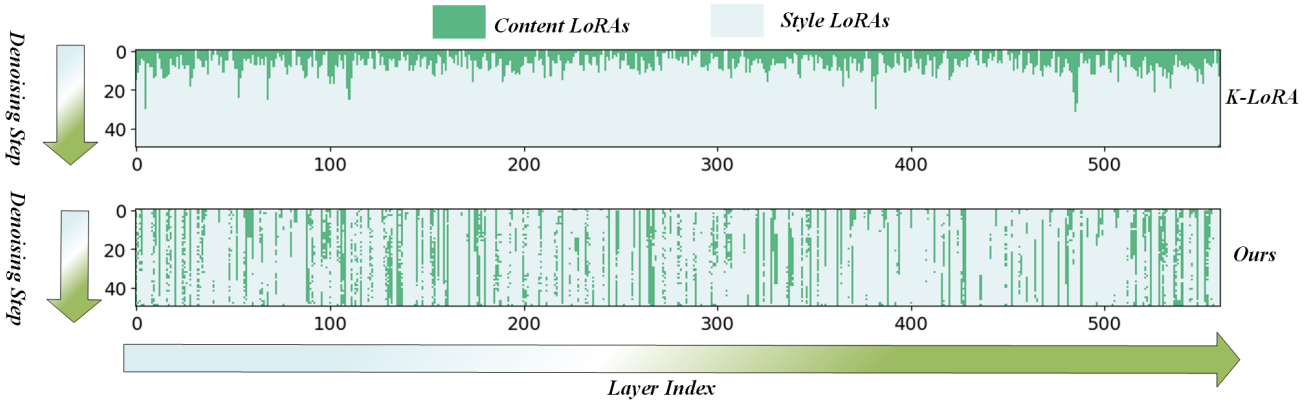

图 5. **生成过程中的 LoRA 选择**。纵轴表示 50 个扩散步，横轴表示不同的 LoRA 层。深绿色表示选择主体 LoRA，浅蓝色表示选择风格 LoRA。

鲁棒性。

特征级选择和潜变量级精炼的效果。 表3 展示了不同配置的性能。仅使用特征级选择（基于 KL 散度）即可使*DINO*分数 比基线方法提高 3.6%，表明内容一致性更好。相比之下，仅使用潜变量级精炼便显著提升了风格相似度 (66.1%) 和 *CLIP*分数 (78.9%)，体现出更强的语义和风格对齐效果。两种组件结合后，我们的方法取得了最佳的整体性能，尤其在*CLIP*分数 (79.1%) 和 *DINO*分数(43.4%)方面表现突出，证明它们在局部特征适配和全局生成引导方面具有互补作用。

选择标准的选择。我们评估了用于特征级选择的其他标准，包括 Kullback–Leibler (KL) 散度、Jensen–Shannon (JS) 散度、余弦相似度和点积（表4）。与基于相似度的指标相比，KL 和 JS 散度均取得了更优的性能。在*DINO* 分数方面，KL 散度略优于 JS（43.7% 对 43.1%），可能是因为它对非对称分布偏移更敏感，因此更适合检测内容表征中的显著特征扰动。

缩放因子的影响。如表5所示，我们还探究了平衡潜变量级精炼影响的缩放因子的敏感性。该因子设置过低（例如 1）会导致*CLIP*和*DINO* 分数不理想，而缩放值过高（例如 20）则会导致*DINO* 分数不稳定。取适中的值 10 可实现最佳权衡，使*CLIP*分数（79.1%）和*DINO* 分数均显著提升（43.4%），凸显了正确校准引导强度的重要性。

鲁棒性分析。为进一步评估我们提出的框架的鲁棒性，我们展示了不同随机种子下生成的结果，并在图中将其与 K- LoRA 进行比较。4然而，K-LoRA 表现出显著的变化——在场景布局方面，且无法始终保留目标风格；我们的方法则能在不同采样条件下同时保留语义内容和风格属性。这种在随机扰动下的稳定性，凸显了输入自适应、表征感知调制在实现连贯且可靠的图像生成方面的有效性。

动态特征选择的可视化。 图 5展示了我们的方法所采用的动态特征级选择机制。与仅依赖静态权重元素进行选择的 K-LoRA 不同，我们的方法根据输入相关特征进行输入条件选择，动态挑选更相关的 LoRA 权重。这种自适应策略能够更好地与输入语义对齐，从而促进更有效、连贯的风格与内容融合。此前定量和定性比较中观察到的优越性能，进一步验证了这种无需训练、适应输入的选择方案的优势。

6. 结论

本文提出一种动态免训练方法，通过在整个生成过程中进行输入自适应、表征感知的决策，融合主体与风格LoRA。具体而言，在前向传播过程中，我们的方法不会依赖静态的权重级启发式方法，而是根据特征扰动的幅度——通过KL散度量化——自适应地选择每一层的LoRA分支。此外，在去噪阶段，我们利用CLIP和DINO分数等客观指标，通过基于梯度的反馈动态优化去噪轨迹。这种双阶段策略能够协调地融合主体身份与艺术风格，无需重新训练或额外监督。在多个基准测试上进行的大量实验表明，无论是定性真实感还是定量指标，我们的方法都始终优于现有的LoRA融合方法。

参考文献

[1] Tomer Amit、Tal Shaharbany、Eliya Nachmani 和 Lior Wolf。Segdiff：使用扩散概率模型进行图像分割。arXiv:2112.00390，2021年。3[2] Omri Avrahami、Kfir Aberman、Ohad Fried、Daniel Cohen-Or 和 Dani Lischinski。Break-a-scene：从单张图像中提取多个概念。载于 SIGGRAPH Asia 2023 会议论文集，第 1–12 页，2023年。3[3] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhaohai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu、Jiabo Ye、Xi Zhang、Tianbao Xie、Zesen Cheng、Hang Zhang、Zhibo Yang、Haiyang Xu 和 Junyang Lin。Qwen2.5-VL 技术报告。arXiv:2502.13923，2025年。7[4] Qinglong Cao、Ding Wang、Xirui Li、Yuntian Chen、Chao Ma 和 Xiaokang Yang。利用潜在物理现象知识训练视频扩散模型。arXiv:2411.11343，2024年。2[5] Mathilde Caron、Hugo Touvron、Ishan Misra、Herv´e J´egou、Julien Mairal、Piotr Bojanowski 和 Armand Joulin。自监督视觉变换器中的新兴特性。载于 IEEE/CVF 国际计算机视觉会议论文集，第 9650–9660 页，2021年。2，5[6] Shoufa Chen、Peize Sun、Yibing Song 和 Ping Luo。Diffusiondet：用于目标检测的扩散模型。载于 IEEE/CVF 国际计算机视觉会议论文集，2023年。3[7] Shoufa Chen、Mengmeng Xu、Jiawei Ren、Yuren Cong、Sen He、Yanping Xie、Animesh Sinha、Ping Luo、Tao Xiang 和 Juan-Manuel Perez-Rua。Gentron：用于图像和视频生成的扩散变换器。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 6441–6451 页，2024年。1[8] Jiahua Dong、Wenqi Liang、Hongliu Li、Duzhen Zhang、Meng Cao、Henghui Ding、Salman H Khan 和 Fahad Shahbaz Khan。如何持续适配文生图扩散模型以实现灵活定制?《神经信息处理系统进展》，37：130057–130083，2024年。3[9] Yuhao Dong、Zuyan Liu、Hai-Long Sun、Jingkang Yang、Winston Hu、Yongming Rao 和 Ziwei Liu。Insight-v：探索多模态大语言模型的长链视觉推理。载于计算机视觉与模式识别会议论文集，第 9062–9072 页，2025年。2[10] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。图像胜过千言万语：用于大规模图像识别的变换器。载于国际学习表征会议，2021年。1[11] Yarden Frenkel、Yael Vinker、Ariel Shamir 和 Daniel Cohen-Or。使用 B-LoRA 实现隐式风格-内容分离。载于欧洲计算机视觉会议，第 181–198 页。Springer，2024年。1，3，5，6，2[12] Rinon Gal、Yuval Alaluf、Yuval Atzmon、Or Patashnik、Amit H Bermano、Gal Chechik 和 Daniel Cohen-Or。一种

一张图胜过千言万语：使用文本反演实现个性化文生图生成。arXiv:2208.01618，2022年。3[13] Yuchao Gu、Xintao Wang、Jay Zhangjie Wu、Yujun Shi、Yunpeng Chen、Zihan Fan、Wuyou Xiao、Rui Zhao、Shuning Chang、Weijia Wu 等。Mix-of-show：用于扩散模型多概念定制的分布式低秩适配。神经信息处理系统进展，36:15890–15902，2023年。3[14] Agrim Gupta、Lijun Yu、Kihyuk Sohn、Xiuye Gu、Meera Hahn、李飞飞、Irfan Essa、Lu Jiang 和 Jos´e Lezama。使用扩散模型生成照片级真实感视频。载于《欧洲计算机视觉会议》，第393–411页。Springer，2024年。1[15] Hao He、Yinghao Xu、Yuwei Guo、Gordon Wetzstein、Bo Dai、Hongsheng Li 和 Ceyuan Yang。CameraCtrl：实现对文本到视频生成的相机控制。arXiv:2404.02101，2024年。2[16] Hao He、Ceyuan Yang、Shanchuan Lin、Yinghao Xu、Meng Wei、Liangke Gui、Qi Zhao、Gordon Wetzstein、Lu Jiang 和 Hongsheng Li。CameraCtrl II：通过相机控制视频扩散模型探索动态场景。arXiv:2503.10592，2025年。[17] Jonathan Ho、Tim Salimans、Alexey Gritsenko、William Chan、Mohammad Norouzi 和 David J Fleet。视频扩散模型。神经信息处理系统进展，35:8633–8646，2022年。2[18] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等。LoRA：大语言模型的低秩适配。国际学习表征会议，1(2):3，2022年。1，3[19] Jiaxiu Jiang、Yabo Zhang、Kailai Feng、Xiaohe Wu、Wenbo Li、Renjing Pei、Fan Li 和 Wangmeng Zuo。MC2：用于定制化多概念生成的多概念引导。arXiv:2404.05268，2024年。3[20] Zeyinzi Jiang、Chaojie Mao、Yulin Pan、Zhen Han 和 Jingfeng Zhang。Scedit：通过跳跃连接编辑实现高效、可控的图像扩散生成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第8995–9004页，2024年。1[21] Sebastian Kaltenbach 和 Phaedon-Stelios Koutsourelakis。将物理约束纳入用于动力系统粗粒化的深度概率机器学习框架。计算物理学杂志，419:109673，2020年。3[22]Amirhossein Kazerouni、Ehsan Khodapanah Aghdam、Moein Heidari、Reza Azad、Mohsen Fayyaz、Ilker Hacihaliloglu 和 Dorit Merhof。用于医学图像分析的扩散模型：综合综述。arXiv:2211.07804，2022年。3[23] Nupur Kumari、Bingliang Zhang、Richard Zhang、Eli Shechtman 和 Jun-Yan Zhu。文生图扩散的多概念定制。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第1931–1941页，2023年。3[24] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。BLIP-2：使用冻结图像编码器和大语言模型引导语言-图像预训练

。载于《国际机器学习会议》，第19730–19742页。PMLR，2023年。1[25] Shiyin Lu、Yang Li、Yu Xia、Yuwei Hu、Shanshan Zhao、Yanqing Ma、Zhichao Wei、Yinglun Li、Lunhao Duan、Jianshan Zhao 等。Ovis2 11737，2025 5 2508 1，2。技术报告。arXiv:。。。[26] Chaofan Ma、Yuhuan Yang、Chen Ju、Fei Zhang、Jinxiang Liu、Yu Wang、Ya Zhang 和 Yanfeng Wang。DiffusionSeg：让扩散模型适用于无监督目标发现。arXiv:2303.09813，2023年。1[27] OpenAI 和微软。推出 GPT-4o：OpenAI 全新的旗舰多模态模型。Azure 博客，2024年。7[28] Ziheng Ouyang、Zhen Li 和 Qibin Hou。K-LoRA：解锁任意主体与风格LoRA的免训练融合。arXiv:2502.18461，2025年。1，3，4，5，6，2[29]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于《国际机器学习会议》，第8748–8763页。PMLR，2021年。2，5，6[30] Nataniel Ruiz、Yuanzhen Li、Varun Jampani、Yael Pritch、Michael Rubinstein 和 Kfir Aberman。DreamBooth：微调文生图扩散模型以实现主体驱动生成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第22500–22510页，2023年。1，3，6[31] Chitwan Saharia、William Chan、Huiwen Chang、Chris Lee、Jonathan Ho、Tim Salimans、David Fleet 和 Mohammad Norouzi。Palette：图像到图像扩散模型。载于 ACM SIGGRAPH 2022 会议论文集，2022年。3[32] Viraj Shah、Nataniel Ruiz、Forrester Cole、Erika Lu、Svetlana Lazebnik、Yuanzhen Li 和 Varun Jampani。ZipLoRA：通过有效融合LoRA实现任意主体与任意风格。载于《欧洲计算机视觉会议》，第422–438页。Springer，2024年。1，3，4，5，6，2[33] Jing Shi、Wei Xiong、Zhe Lin 和 Hyun Joon Jung。InstantBooth：无需测试时微调的个性化文生图生成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第8543–8552页，2024年。3[34] Kihyuk Sohn、Lu Jiang、Jarred Barber、Kimin Lee、Nataniel Ruiz、Dilip Krishnan、Huiwen Chang、Yuanzhen Li、Irfan Essa、Michael Rubinstein 等。StyleDrop：任意风格的文生图合成。神经信息处理系统进展，36:66860–66889，2023年。1，3，6[35] Yang Song、Jascha Sohl-Dickstein、Diederik P Kingma、Abhishek Kumar、Stefano Ermon 和 Ben Poole。通过随机微分方程进行基于分数的生成建模。arXiv:2011.13456，2020年。4[36] Taiqiang Wu、Jiahao Wang、Zhe Zhao 和 Ngai Wong。低秩适配中的子空间混合。arXiv:2406.11909，2024年。3[37] Guangxuan Xiao、Tianwei Yin、William T Freeman、Fr´edo Durand 和 Song Han。FastComposer：通过局部注意力实现免调优的多主体图像生成。国际计算机视觉杂志，第1–20页，2024年。3[38] Shaoan Xie、Zhifei Zhang、Zhe Lin、Tobias Hinz 和 Kun Zhang。Smartbrush：利用扩散模型进行文本和形状引导的物体修复。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第22428–22437页，2023年。3[39] Peng Xing、Haofan Wang、Yanpeng Sun、Qixun Wang、Xu Bai、Hao Ai、Renyuan Huang 和 Zechao Li。Csgo：文本到图像生成中的内容-风格组合。arXiv:2408.16766，2024年。3[40] Zhen Xing、Qijun Feng、Haoran Chen、Qi Dai、Han Hu、Hang Xu、Zuxuan Wu 和 Yu-Gang Jiang。视频扩散模型综述。《ACM计算调查》，57(2):1–42，2024年。1[41] Yiyan Xu、Wenjie Wang、Yang Zhang、Biao Tang、Peng Yan、Fuli Feng 和 Xiangnan He。利用大型多模态模型进行个性化图像生成。载于《2025年ACM Web会议论文集》，第264–274页，2025年。1[42] Fei Zhang、Tianfei Zhou、Boyang Li、Hao He、Chaofan Ma、Tianjiao Zhang、Jiangchao Yao、Ya Zhang 和 Yanfeng Wang。揭示弱开放词汇语义分割中的原型知识。《神经信息处理系统进展》，36:73652–73665，2023年。2[43] Hao Zhang、Feng Li、Shilong Liu、Lei Zhang、Hang Su、Jun Zhu、Lionel M Ni 和 Heung-Yeung Shum。DINO：通过改进去噪锚框实现端到端目标检测的DETR。arXiv:2203.03605，2022年。6[44] Longteng Zhang、Lin Zhang、Shaohuai Shi、Xiaowen Chu 和 Bo Li。LoRA-FA：用于大语言模型微调的内存高效低秩适配。arXiv:2308 03303，2023 3 。。[45] Guangcong Zheng、Xianpan Zhou、Xuewei Li、Zhongang Qi、Ying Shan 和 Xi Li。LayoutDiffusion：用于布局生成图像的可控扩散模型。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第22490–22499页，2023年。1[46] Hongyun Zhou、Xiangyu Lu、Wang Xu、Conghui Zhu、Tiejun Zhao 和 Muyun Yang。LoRA-Drop：基于输出评估的高效LoRA参数剪枝。arXiv:2402 07721，2024 3 。。[47] Bojia Zi、Xianbiao Qi、Lingzhi Wang、Jianan Wang、Kam-Fai Wong 和 Lei Zhang。Delta-LoRA：利用低秩矩阵的增量微调高秩参数。arXiv:2309.02411，2023年。3

主体与风格 LoRA 的动态免训练融合

补充材料

1. 使用额外指标进行评估

除了风格相似度、CLIP分数和DINO分数等常用指标外，我们还采用基于 ViT [10] 和 BLIP-2 [24] 的内容与风格分数，以更全面地评估所提出的方法。如表 1 所示，我们的方法在四项指标中的三项上取得了最先进的结果。仔细观察可发现，不同基线方法各有所长：ZipLoRA 在 BLIP-2 内容对齐方面表现突出，而 B-LoRA 在 BLIP-2 风格一致性方面表现优异。K-LoRA 在各项指标上的表现相对均衡。相比之下，我们提出的方法不仅提升了基于 ViT 和 BLIP-2 的分数，还更好地兼顾了内容保真度与风格保留。我们认为，这一优势源于所提出的软性推理时度量引导，它能够自适应地校准生成过程，在避免过拟合单一表征空间的同时，维持语义与风格的一致性。这些结果表明，我们的动态融合框架在不同类型的评估视角下具有更强的鲁棒性，也展现出其在更广泛的多模态生成场景中实现泛化的潜力。

2. 基于 FLUX 的其他实验结果

如主实验部分所述，为了更全面地展示我们提出的方法在 FLUX 框架下的卓越性能和泛化能力，我们进一步使用社区在 HuggingFace 上分享的公开、训练充分的 LoRA（低秩适配）权重，开展了广泛的定性评估。具体而言，我们选取了一组涵盖多种物体类别和艺术风格的 LoRA权重，系统地评估模型整合并合成复杂跨域表示的能力。图1和图2展示了由此生成的融合图像，呈现了丰富多样的组合；在这些图像中，物体语义与风格属性经由我们的方法共同编码并渲染。

我们的方法通过解耦并重组对象特定与风格特定的潜在表示来融入这些LoRA权重，其方式是基于表示扰动执行特征级选择，并在去噪过程中应用度量引导细化。这使得模型能够以可控而又灵活的方式对齐并合成视觉内容。生成的样本不仅高度忠实于目标对象的语义结构，而且与期望的风格高度一致，展现了模型保留两个输入域中关键属性的能力。此外，外观与内容的无缝融合证实了我们的框架在处理多样且未见组合方面的稳健性，凸显了其在要求风格泛化与组合创造力的真实世界生成任务中的潜在适用性。总体而言，这些视觉结果为该方法在广泛且具有挑战性的场景中生成连贯、高质量输出的有效性提供了有力证据。

3. 用户研究和多模态大语言模型评估的设置详情

按照 K-LoRA [28], 的评估方案，我们开展了一项用户研究：向参与者展示一张**主体参考图像**、一张**风格参考图像**，以及两张匿名输出结果——一张由**我们的方法**生成，另一张由随机选取的基线方法（Zi-pLoRA [32], B-LoRA [11], 或 K-LoRA [28]）生成。为减少呈现偏差，两张输出结果的呈现顺序在各试次中随机调整。参与者需要回答：“哪张图像在保留主体身份的同时，更好地体现了给定的艺术风格？”我们共收集了 43 位参与者的 1,290 份回答，每位参与者均评估了一组独特的 30 个试次。

除了人工评估外，我们还采用**GPT-4o**和 **Qwen2.5-VL**作为多模态大语言模型（MLLM）评判模型，以评估感知对齐情况。每次试验中，提示词均包含内容图像、风格参考图，以及来自 ZipLoRA、B-LoRA、K-LoRA 和我们提出的方法的四个匿名输出结果（顺序随机）。我们要求这些大语言模型选出在风格保真度与主体保留之间取得最佳平衡的图像。为确保评估稳健性，我们随机抽取了**100 组主体—风格配对**，并对每组配对进行了三次独立运行。最终得分通过对各次运行结果进行多数投票确定，这符合近期基于多模态大语言模型的评估研究中的常见做法。值得注意的是，用户研究反映了人类主观偏好，而基于多模态大语言模型的评估则能提供可扩展且可复现的判断结果，因此两种评估相辅相成、彼此印证。

方法	ViT 内容 ↑	ViT 风格 ↑	BLIP-2 内容 ↑	BLIP-2 风格 ↑
ZipLoRA [32]	23.2%	18.0%	48.2%	46.0%
B-LoRA [11]	29.6%	19.9%	42.0%	54.3%
K-LoRA [28]	32.8%	20.7%	49.4%	51.1%
Ours	33.7%	21.8%	49.5%	51.4%

表1. 使用额外评估指标对对齐结果进行比较。

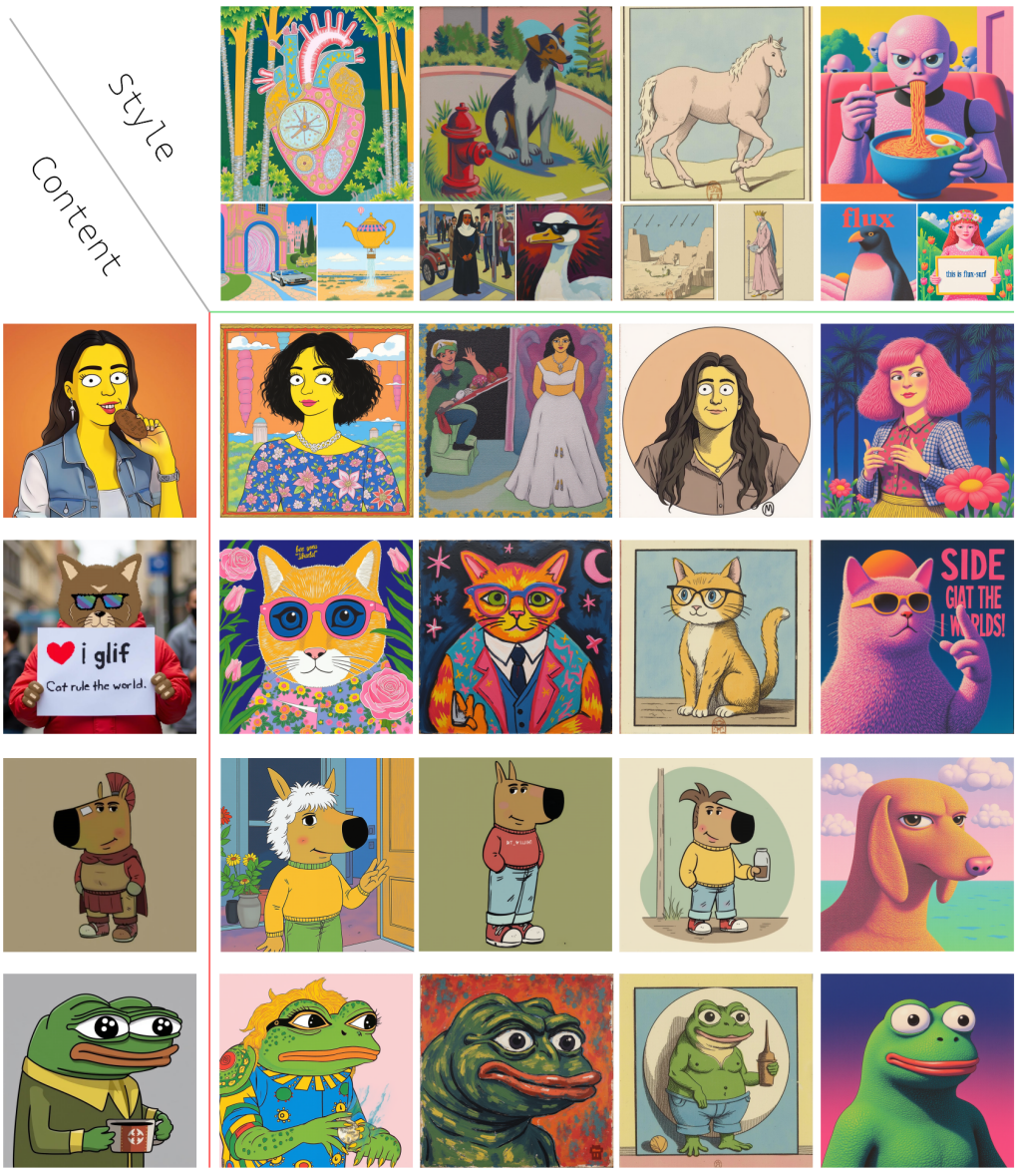

图 1. 使用 FLUX 生成的更多结果。每张图像对应上方标注的物体标签和左侧的风格参考。这些结果展示了通过我们提出的方法应用不同 LoRA 模块的效果。

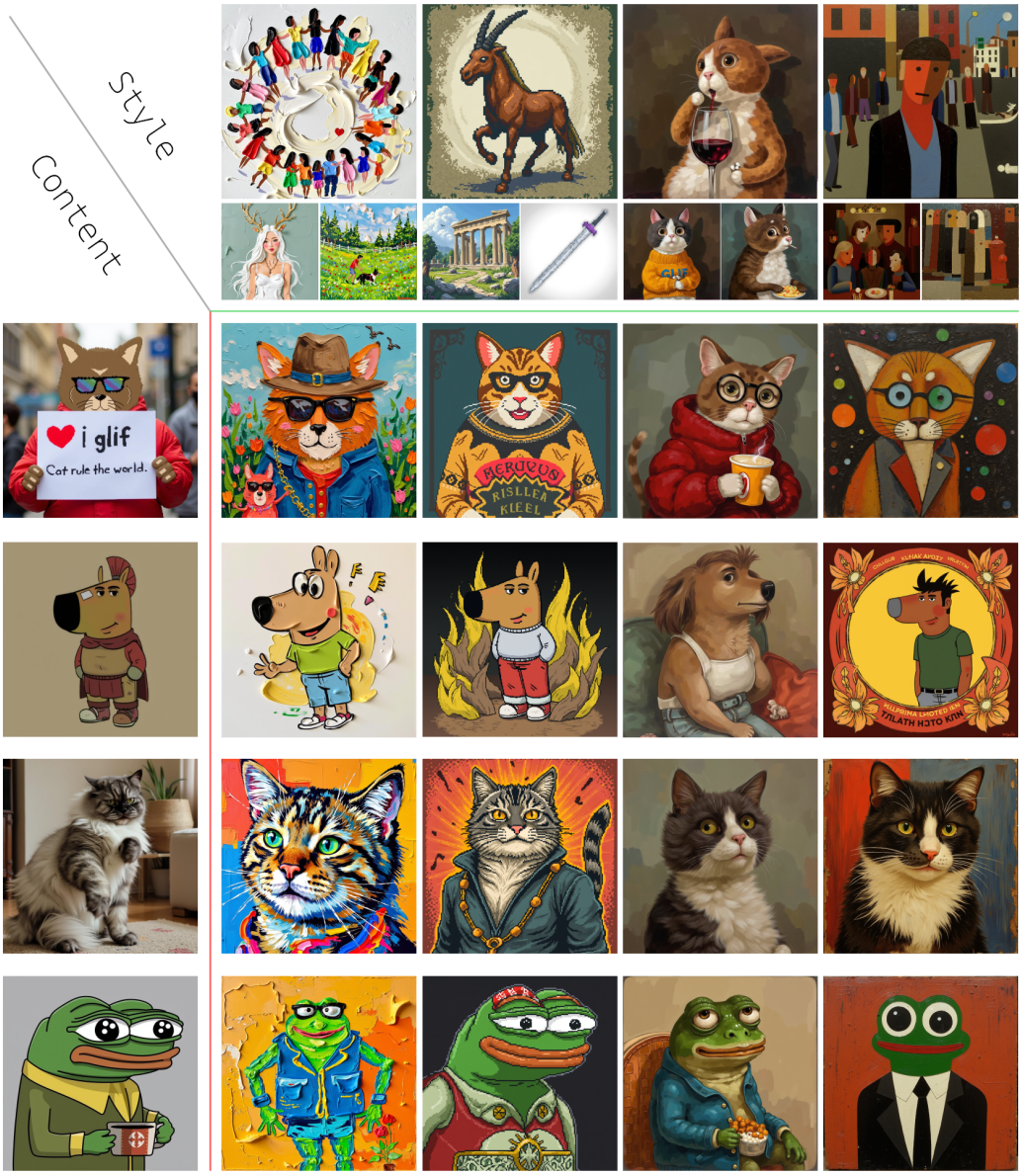

图2. 使用 FLUX 生成的更多结果。每张图像对应上方标示的物体标签和左侧的风格参考。这些结果展示了通过我们提出的方法应用不同 LoRA 模块所产生的效果。