

# K-LoRA: Unlocking Training-Free Fusion of Any Subject and Style LoRAs

欧阳子恒 李震<sup>†</sup> 侯淇彬  
南开大学计算机学院 VCIP 实验室  
[{zihengouyang666, zhenli1031}@gmail.com](mailto:{zihengouyang666, zhenli1031}@gmail.com)

项目主页: <https://k-lora.github.io/K-LoRA.io/>

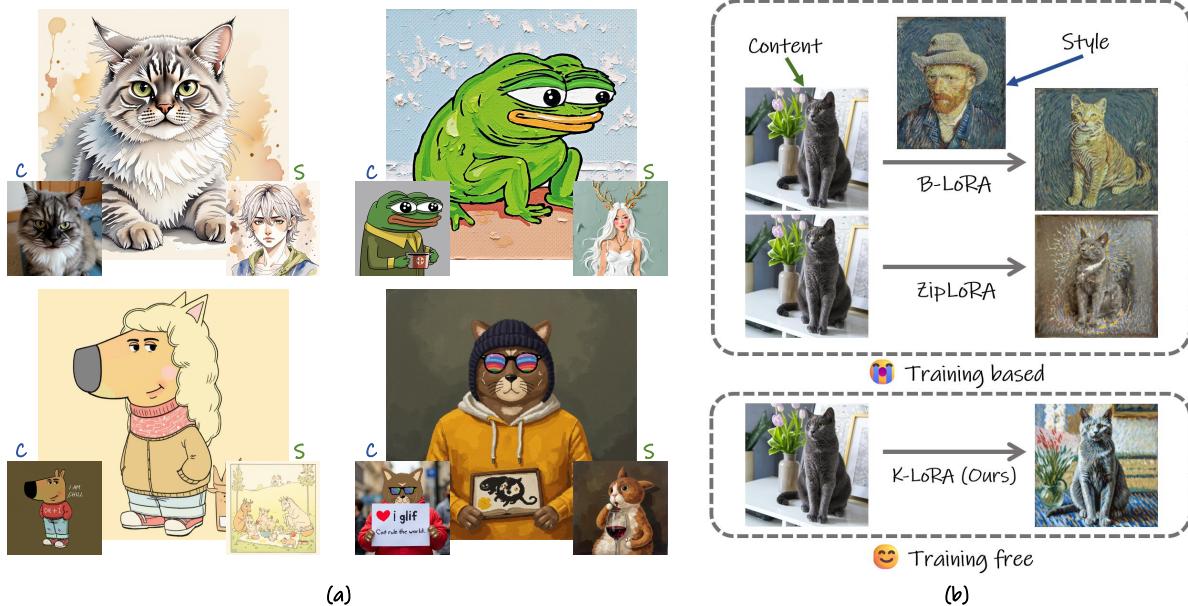


图 1. 图例. (a) 展示了 K-LoRA 使用 FLUX [3] 的卓越生成性能，其中左侧为对象参考，右侧为风格参考，中间为所生成的图像。(b) 比较了 K-LoRA 与现有的最先进方法 B-LoRA [8] 和 ZipLoRA [26]，这些方法改变了原始权重矩阵或未充分利用网络结构，容易丢失风格或内容信息。K-LoRA 增强了每个 LoRA 矩阵捕获的信息，从而实现更优越的融合效果且无需额外训练。

## Abstract

近期的研究已经探索了如何结合不同的 LoRA 模型，以共同生成学习到的风格和内容。然而，现有的方法要么无法同时有效地保留原始主体和风格，要么需要额外的训练。在本文中，我们认为 LoRA 的内在属性可以有效地指导扩散模型融合学习到的主体和风格。基于这一洞见，我们提出了一种简单而有效的免训练 LoRA 融合方法——K-LoRA。在每个注意力层中，K-LoRA 会比较待融合的各个 LoRA 中的 Top-K 元素，从而决

定选择哪个 LoRA 以实现最优融合。该选择机制确保了在融合过程中，主体和风格中最具代表性的特征都得以保留，从而有效地平衡了二者的贡献。实验表明，K-LoRA 能够有效整合原始 LoRA 学到的主体和风格信息，在定性和定量结果上均优于当前最先进的需要训练的方法。

## 1. 引言

个性化和风格化是计算机视觉领域两项成熟的任务，多年来一直是活跃的研究领域 [4, 6, 9, 13, 17, 24,

<sup>†</sup> 通讯作者。

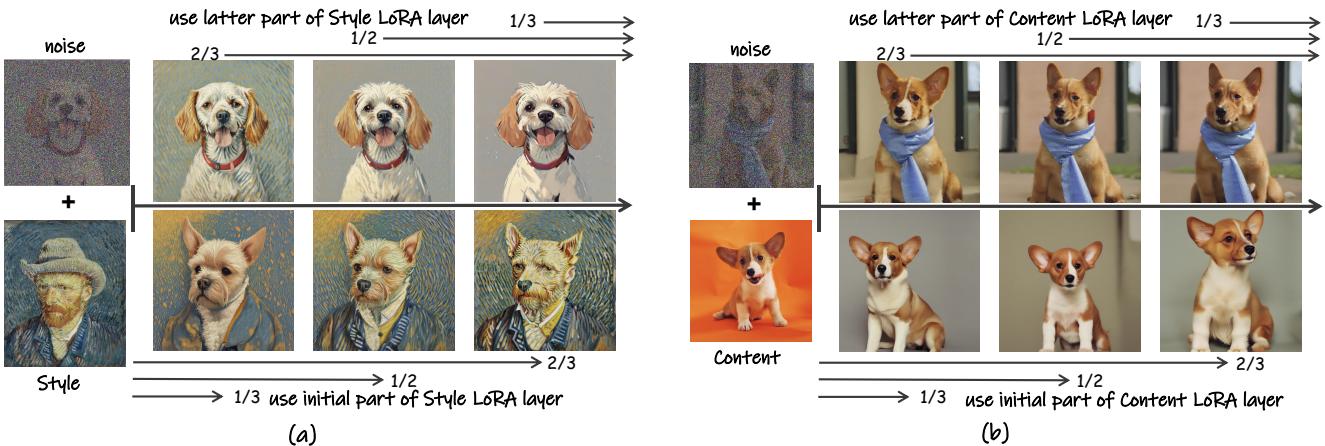


图 2. 研究发现的可视化结果。 (a) 仅使用内容 LoRA 进行微调的结果。 (b) 仅使用风格 LoRA 的结果。在这些实验中，我们测试了在初始和后期时间步中添加 LoRA 层的差异。

28, 34, 37, 42]。这些任务中的主要挑战在于，在通常由文本或视觉输入引导的情况下，保持独特的内容或修改图像的风格。在此背景下，“内容”指的是图像中的物体和结构，而“风格”则包含颜色、纹理和图案等视觉属性。由于风格定义的主观性以及风格与内容之间强烈的相互依赖性，操作图像风格尤为困难，这使得有效解耦这些元素变得复杂。

近期的技术，如 LoRA [12]，因其在图像合成中实现高效微调的能力而受到越来越多的关注。尽管风格和对象是分开训练的，但 LoRA 为图像生成任务中解耦风格与内容的问题提供了有效的解决方案，它通过独立于内容特征来训练风格特征，从而在控制风格迁移方面表现出色。随着利用 LoRA 的个性化应用日益普及，大量研究致力于通过融合 LoRA 权重来合并对象和风格 [25]。这些方法旨在允许用户通过可变系数来调整每个 LoRA 的贡献比例。还有一些方法，例如 ZipLoRA [26]，尝试训练一个融合比例向量来平衡不同的 LoRA。最近，一些方法提出将 LoRA 周期性地集成到模型中 [41]。此外，B-LoRA [8] 技术仅微调两个注意力模块以促进风格迁移。

在使用这些方法的实验中，我们发现了两个关键问题，如 Fig. 1(b) 所示。首先，生成的图像中风格细节常常丢失，且对象特征的保持不一致。其次，需要手动调整某些超参数和种子，或者需要进行额外的训练。针对第一个问题，我们进行了大量实验并观察到，在元素级别上合并两个 LoRA 的注意力层可能导致风格细节和纹理的平滑化，甚至丢失对象特征。鉴于元素级别

的融合可能导致次优结果，我们通过实验选择性地移除某些元素来保持良好性能。针对第二个问题，受近期研究 [20, 29, 35] 核心思想的启发，我们根据扩散时间步将 LoRA 的注意力层整合到模型中，以评估其对性能的影响。通过这种方法，我们得出了关键结论。(i) 仅需有限数量的扩散预测步骤就足以保留原始效果，如 Fig. 2 所示。(ii) 在应用 LoRA 时，初始的扩散步骤负责重建对象并捕捉较大的纹理细节，而后续步骤则专注于增强和细化对象以及风格纹理的更精细细节。

基于这些发现，我们提出了 K-LoRA，该方法同时解决了我们在实验中发现的两个问题，如 Fig. 1(a) 所示。它利用我们的第一个洞见，在注意力层的每个前向传播过程中引入一个 Top-K 选择过程，以在每个位置识别最合适的注意力组件。此外，我们在选择过程中应用了一个缩放因子，利用我们的第二个洞见来强调风格和内容在整个扩散过程中扮演的不同角色。

我们的方法可以有效解决上述问题，确保在面对具有挑战性的内容和风格组合时，融合后的 LoRA 能够同时捕捉主体和风格特征。这带来了稳定的生成结果，并显著提升了融合 LoRA 的性能。此外，我们的方法用户友好，无需额外训练。我们的贡献总结如下：

- 我们提出了 K-LoRA，一种简单而有效的优化技术，可以无缝地融合内容和风格 LoRA，从而能够为任何主题生成任何期望的风格，同时保留复杂的细节。
- 我们的方法用户友好，无需重新训练，可直接应用于现有的 LoRA 权重。它在各种图像风格化任务中表现出卓越的性能，超越了现有方法。

## 2. 相关工作

**用于定制化任务的扩散模型。**在用于定制化任务的扩散模型 [23] 领域，定制化指的是模型学习并解释用户提供的新定义的过程。诸如 Textual Inversion [1, 29, 40]、DreamBooth [24] 和 Custom Diffusion [16] 等技术，通过基于令牌 (token) 的优化，使模型仅用少量图像就能捕捉目标概念。具体来说，Textual Inversion 通过微调嵌入 (embeddings) 来重构目标，DreamBooth 使用不常见的特定类别术语来扩展对象类别，而 Custom Diffusion 则专注于微调扩散模型中的交叉注意力层 (cross-attention layers) 以学习新概念。此外，也存在一些在推理时无需训练的方法 [2, 27, 31, 32]，但它们利用预训练模块的策略在处理某些特定任务时可能表现不佳。LoRA [12] 及其变体 [11, 15, 22, 39, 43, 43, 44] 因其在微调大型模型和提供高质量结果方面的出色能力而广为人知，使其成为从业者的优选。

**图像生成中的 LoRA 组合。**在图像生成领域，关于 LoRA 组合的研究主要沿着两个方向推进：多对象集成以及内容与风格的融合。在对象集成方面，研究主要集中于让模型能够融合封装在多个 LoRA 中的不同对象概念 [7, 10, 14, 18, 36]。通过微调主题 LoRA，这些模型可以吸收各种新概念，并利用掩码 (masking) 技术来管理对象布局。在内容-风格融合方面，一些工作，如 MergingLoRA [25]、Mixture-of-Subspaces [30] 和 ZipLoRA [26]，提出了通过超参数调整和学习融合矩阵来合并预训练 LoRA 权重层的方法。然而，这些方法可能会面临概念稀释、精细细节模糊以及特定的训练要求等挑战。近期，B-LoRA [8] 识别出注意力模块在生成过程中的不同作用，通过仅训练两个核心注意力模块，实现了 LoRA 内部的对象-风格解耦。此外，LoRA Composition [41] 采用对模型的 LoRA 模块进行循环更新的方式，允许多个 LoRA 协同引导模型，从而实现了多样的跨概念融合。尽管取得了这些进展，现有方法仍然面临着控制精度不足、对象风格丢失以及高昂的训练要求等挑战。

## 3. 方法

### 3.1. 预备知识

LoRA 是一种最初为适配大规模语言模型而设计的有效方法。LoRA 的核心前提是，在微调大型模型并与基线模型比较时，参数更新矩阵  $\Delta W \in \mathbb{R}^{m \times n}$  通常被发现包含许多微小或接近于零的元素，从而呈现出低秩结构。这一特性使得  $\Delta W$  可以被分解为两个低秩矩阵  $B \in \mathbb{R}^{m \times r}$  和  $A \in \mathbb{R}^{r \times n}$ ，其中  $r$  代表  $\Delta W$  的本征秩，并且假设  $r \ll \min(m, n)$ 。这一特点使我们能够冻结基础权重  $W_0$ ，仅训练矩阵  $B$  和  $A$  来替代  $\Delta W$ ，从而以  $\Delta W = BA$  的形式实现高效的参数化。最终， $\Delta W$  被加到原始模型的基础权重上以进行微调。更新后的权重可以表示为  $W_0 + \Delta W$ 。

在我们的工作中，我们采用了与 ZipLoRA [26] 中相同的符号表示。设  $D$  为一个基础扩散模型， $W_0$  表示需要通过 LoRA 层更新的预训练权重。基础模型  $D$  可以通过简单地将一个额外训练的 LoRA 权重集  $\Delta W_x$  添加到模型权重中来适配特定概念，从而得到  $D' = W_0 + \Delta W_x$ 。给定与基础模型  $D$  相关联的两个独立训练的 LoRA 权重集  $\Delta W_c$  和  $\Delta W_s$ ，我们的目标是充分利用这两个 LoRA 集的权重并实现它们的有效融合。为此，我们提出了一种名为 K-LoRA 的方法，用于无缝地组合这两个 LoRA 权重集，表示为

$$\Delta W_x = K(\Delta W_c, \Delta W_s),$$

其中  $K$  代表我们的方法，该方法可以高效地整合内容 LoRA 和风格 LoRA 的贡献。

接下来，我们将详细解释所提出的方法。我们的方法基于两个发现。(i) 在扩散步骤中，每步仅对一部分层应用 LoRA，可以达到与对所有层应用 LoRA 相媲美的效果；(ii) 在较早的扩散步骤中使用主体 LoRA 倾向于生成更好的主体信息，而在较晚的步骤中使用风格 LoRA 则更有效地生成风格和细节，且不影响内容的构建。

### 3.2. K-LoRA

[26] 中曾指出，在使用 LoRA 进行微调时，采用较小的关键元素集合可以达到与原始方法相同的生成效果。然而，作者并未提供相关实验来解释这一现象在图像生成领域的应用。我们首先尝试利用这种方法，采

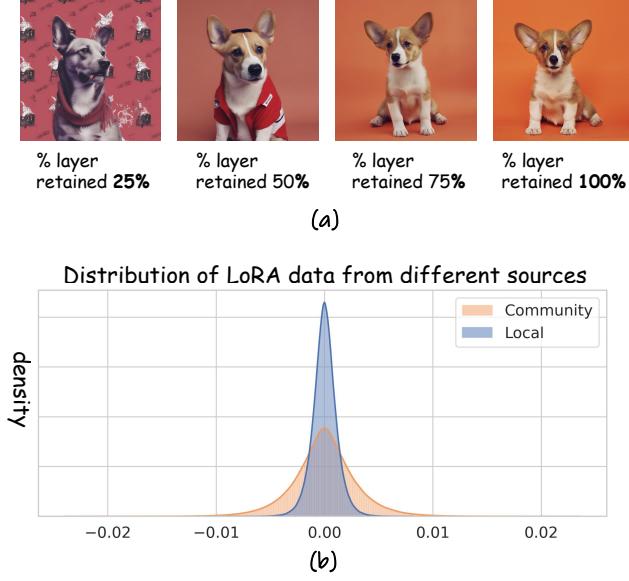


图 3. 实验可视化结果。 (a) 根据特定比例随机加载部分 LoRA 注意力层生成的图像。 (b) 来自不同来源的 LoRA 数据分布可视化：一个在本地训练，另一个从社区仓库下载。

用与 Magmax [19] 类似的方式，将值较小的元素置为零。我们发现，通过这种方式修改矩阵元素得到的结果与 [26, 30] 产生的结果相似，因为模型无法正确解释其先前学习到的概念，导致图像生成质量不佳。

鉴于直接修改注意力元素的复杂性和局限性，一个问题随之产生：我们能否在去噪过程中利用 LoRA 矩阵的稀疏特性？其目的是找到一种替代方法，能够在不修改原始 LoRA 权重的情况下，为每个步骤或层确定一个良好的权重选择方法和精确的 LoRA 定位。基于 Multi-LoRA Composition [41]，我们随机地将内容 LoRA 注意力层应用于扩散步骤中，通过使用  $x\%$  的注意力层来影响对象，以观察生成结果。如 Fig. 3(a) 所示，我们发现当  $x > 50$  时，其结果与原始模型几乎没有区别。然而，当  $x < 25$  时，模型维持其原始个性化概念的能力显著下降。

受近期研究 [20, 29, 35] 的启发，我们进一步扩展了 Fig. 2 中的上述实验，并发现，在较早的时间步应用风格 LoRA 会对原始对象的重构产生显著影响，而在较晚的时间步应用则能在不影响原始对象的情况下保留风格信息。此外，我们观察到，对于内容 LoRA，在较早的时间步应用比在较晚的时间步应用能产生明显更好的结果。

以上分析启发我们通过为每个注意力层自适应地选择 LoRA 模块来实现生成对象与风格的融合。根据发现 (i)，选择策略应保留整体的对象和风格信息。此外，根据发现 (ii)，生成过程应通过恰当地安排对象和风格组件来实现。即在早期的扩散步骤中，模型应更侧重于对象重构，同时引入风格纹理；而在后期的步骤中，则最好用细微的对象细节来完善风格。因此，我们提出了 K-LoRA 方法，如 Fig. 4 所示，该方法能够自适应地选择合适的 LoRA 层来融合学习到的主体和风格。

首先，我们对 LoRA 层中的每个元素取绝对值，以判断特定值是否在生成过程中起重要作用，

$$\Delta W'_c = |\Delta W_c|, \quad (1)$$

$$\Delta W'_s = |\Delta W_s|, \quad (2)$$

其中  $W_c$  和  $W_s$  分别表示内容和风格 LoRA 的权重。由于一小部分主导元素即可达到原始的生成效果，而数据分布（见 Fig. 3(b)）显示较小的元素占据了大量位置，这会影响重要元素的选择，因此我们使用少量最大值元素来代表每个层的重要性。

具体来说，我们分别从  $\Delta W'_c$  和  $\Delta W'_s$  中选取值最大的前  $K$  个元素。通过累加这些 Top-K 元素，我们评估这两个矩阵在给定注意力层的重要性：

$$S_c = \sum_{i \in \text{Top-K}(\Delta W'_c)} \Delta W'_{c,i}, \quad (3)$$

$$S_s = \sum_{j \in \text{Top-K}(\Delta W'_s)} \Delta W'_{s,j}, \quad (4)$$

其中 Top-K 返回最大  $K$  个值的索引。对于  $K$  的选择，我们注意到 LoRA 训练过程中的秩数在一定程度上反映了矩阵内包含的信息量。因此，我们对  $K$  的选择与每个 LoRA 的秩相关联：

$$K = r_c \cdot r_s, \quad (5)$$

其中  $r_c$  和  $r_s$  分别代表内容和风格 LoRA 层的秩。该公式使我们能够通过比较两个总和来确定一个注意力层内合适的权重

$$C(S_c, S_s) = \begin{cases} \Delta W_c, & \text{if } S_c \geq S_s \\ \Delta W_s, & \text{otherwise} \end{cases} \quad (6)$$

为了更有效地利用发现 (ii)，并让对象和风格在不同阶段发挥各自作用，同时确保从侧重对象到侧重风

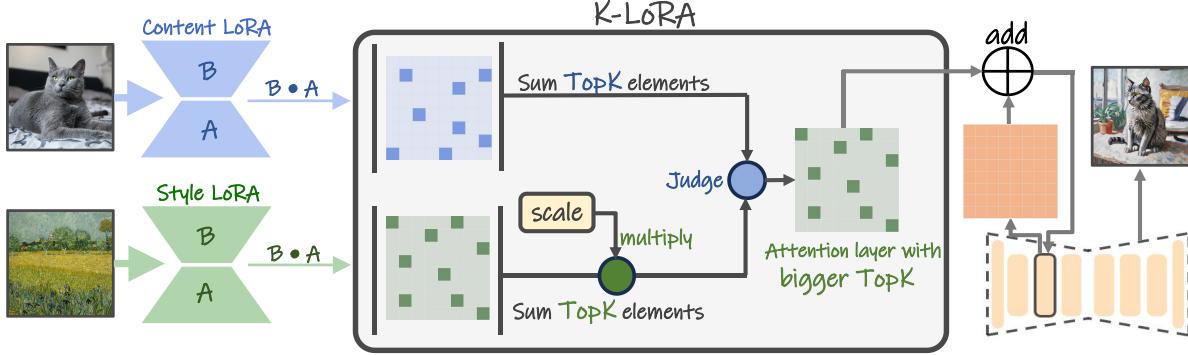


图 4. 所提出的 K-LoRA 方法概览。我们提出了 K-LoRA 方法，该方法利用 Top-K 函数根据矩阵元素的总和在每个前向层中选择重要的 LoRA 权重。这使我们能够同时保留风格细节和对象特征。

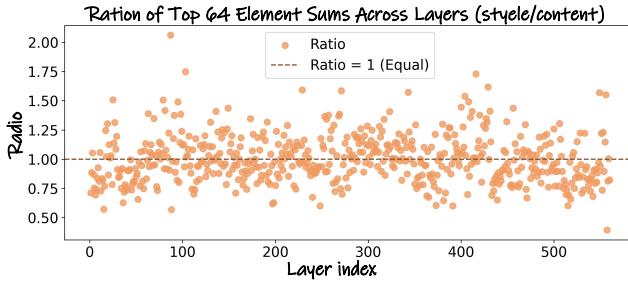


图 5. 比例可视化。该图反映了对 Top-K 元素求和后的比例结果，其中每个对应位置的比例差异都相当显著。

格的平滑过渡，我们引入了一个缩放因子  $S$ 。该因子  $S$  直接应用于 Top-K 选择过程，在生成的早期阶段增强对象内容，并在后期阶段逐渐强调风格

$$S = \alpha \cdot \frac{t_{now}}{t_{all}} + \beta, \quad (7)$$

其中  $t_{now}$  表示反向去噪过程中的当前步骤， $t_{all}$  是总步数，而  $\alpha, \beta$  是超参数。

为避免在使用来自不同来源的社区 LoRA 模型时出现过大的权重差异，这种差异可能导致 Top-K 选择在注意力分配上失效，我们引入了一个新因子  $\gamma$  来平衡两种权重

$$S' = \gamma \cdot S. \quad (8)$$

首先，我们计算每个层  $l$  内元素的绝对值之和，然后逐层累加这些和来计算  $\gamma$

$$\gamma = \frac{\sum_l \sum_i \Delta W'_{c_{l,i}}}{\sum_l \sum_j \Delta W'_{s_{l,j}}}. \quad (9)$$

$\gamma$  的引入解决了两个 LoRA 组件中元素之间显著的数值差异问题，如 Fig. 3(b) 所示。这一调整突显了

LoRA 层内的有用组件。引入  $\gamma$  后，每个层中内容和风格 LoRA 权重的比例关系如 Fig. 5 所示。可以观察到，在每个应用了 LoRA 的前向层中，主导成分总和的比例存在显著差异。这突显了每个层内不同 LoRA 权重的重要性，为选择提供了坚实的基础。

然后，我们将  $S'$  应用于风格 LoRA 并更新  $S_s$

$$S'_s = S_s \cdot S'. \quad (10)$$

通过引入  $S'$ ，我们可以在较早的时间步加强内容的影响，同时在较晚的时间步放大风格的主导地位。这一调整可以有效地利用发现 (ii)，优化对象和风格的选择，从而在图像生成过程中最大化它们的贡献。最终的 LoRA 权重可通过计算  $C(S_c, S'_s)$  获得。为了阐明，我们在算法 1 中展示了伪代码。

为更好地解释权重选择过程，我们在 Fig. 6 中展示了选择比例，其中对象和风格无缝地相互渗透和融合。前半部分主要侧重于对象，并融入了少量风格，而后半部分则主要强调风格，同时保留了对象的细微存在，这进一步证实了我们的关键发现。

## 4. 实验

### 4.1. 实验设置

**数据集.** 遵循 ZipLoRA [26] 的惯例，对于通过本地训练获得的 LoRA，我们从 DreamBooth [24] 数据集中选择了一组多样化的内容图像，每组包含 4-5 张关于特定主体的图像。在风格方面，我们选用了 StyleDrop [28] 作者之前提供的数据集，并加入了一些经典名作以及

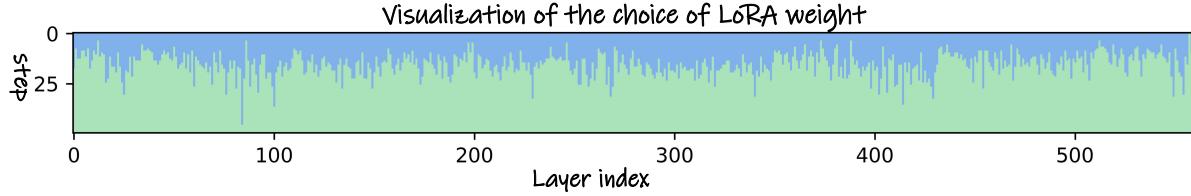


图 6. 生成过程中的 LoRA 选择。该图展示了每个前向层内的选择情况。纵轴代表总共 50 个扩散步骤，横轴表示 LoRA 层的数量。每个位置的颜色表示所选择的层。蓝色条对应对象，绿色条对应风格。

---

#### Algorithm 1 类 PyTorch 风格的伪代码。

---

```

# timestep: 当前时间步
# content_lora_weight, style_lora_weight: 输入权重
# alpha, beta, gamma: 缩放因子
# all_timesteps: 总时间步

# 根据秩设置k
k = rank * rank

# TopK 内容值的总和
abs_content_matrix = abs(content_lora_weight)
topk_content_values = topk(abs_content_matrix.fl(), k)
sum_topk_content = sum(topk_content_values)

# TopK 风格值的总和
abs_style_matrix = abs(style_lora_weight)
topk_style_values = topk(abs_style_matrix.fl(), k)
sum_topk_style = sum(topk_style_values)

# 计算并应用缩放因子
scale = alpha + beta * timestep / all_timesteps
scale = scale * gamma
sum_topk_style *= scale

# 比较并返回结果
if sum_topk_content >= sum_topk_style:
    return content_lora_weight
else:
    return style_lora_weight

fl: 展平;

```

---

一些现代创新风格。对于每种风格，我们仅使用单张图像进行训练。

**实验细节.** 我们使用 SDXL v1.0 基础模型和 FLUX 模型进行实验，并测试了 K-LoRA 在本地训练的 LoRA 和社区训练的 LoRA 上的性能。对于社区训练的 LoRA，我们使用 Hugging Face 上广泛可用的 LoRA 模型进行测试。对于本地训练的 LoRA，我们基于 ZipLoRA [26] 中概述的方法来获取一组风格和内容 LoRA。对于 Eqn. (7) 中提到的超参数，我们设置  $\alpha = 1.5$  和  $\beta = 0.5$ 。我们发现该配置在几乎所有情况下都行之有效，能够持续产生良好的生成结果。

#### 4.2. 实验结果

**定量比较.** 我们随机选择了 18 组物体和风格的组合，每组包含 10 张图像进行定量比较。我们使用 CLIP [21] 来衡量风格相似度，通过 CLIP 分数和 DINO 分数 [38] 来计算主体相似度。我们将我们的方法与社区流行的方法以及最先进的方法进行了比较，包括直接算术合并、联合训练、ZipLoRA [26] 和 B-LoRA [8]。结果如表 1 所示。可以观察到，与先前的方法相比，我们的方法显著提升了主体相似度指标，同时也取得了令人满意的风格相似度。

Method	Style Sim ↑	CLIP Score ↑	DINO Score ↑
Direct	48.9%	66.6%	43.0%
Joint	68.2%	57.5%	17.4%
B-LoRA [8]	58.0%	63.8%	30.6%
ZipLoRA [26]	60.4%	64.4%	35.7%
<b>K-LoRA (ours)</b>	<b>58.7%</b>	<b>69.4%</b>	<b>46.9%</b>

表 1. 定量比较。不同方法对齐结果的比较。

**定性比较.** 为确保公平评估，此阶段的所有实验均使用 SD 进行，结果如 Fig. 7 所示。在未经大量参数调整或种子点选择的情况下，当融合比例直接设置为 1:2 时，合并 LoRA 的方法 [25] 难以保留物体的原始形状、颜色和风格特征。B-LoRA [8] 主要捕捉原始图像中物体的颜色和外观，常常导致颜色过拟合，使得在生成图像中难以辨认原始物体。在 ZipLoRA [26] 和联合训练方法中，虽然融合了某些风格纹理，但模型倾向于关注风格的背景元素而非捕捉风格本身，导致成功率较低。相比之下，我们的方法解决了这些局限性，能够在各种种子点变化下稳定地生成更高质量的输出图像。此外，我们的方法无需额外的训练或参数微调。

我们向用户呈现了一组随机选择的 22 个结果进行比较评估。每组结果包括来自 ZipLoRA、B-LoRA 和

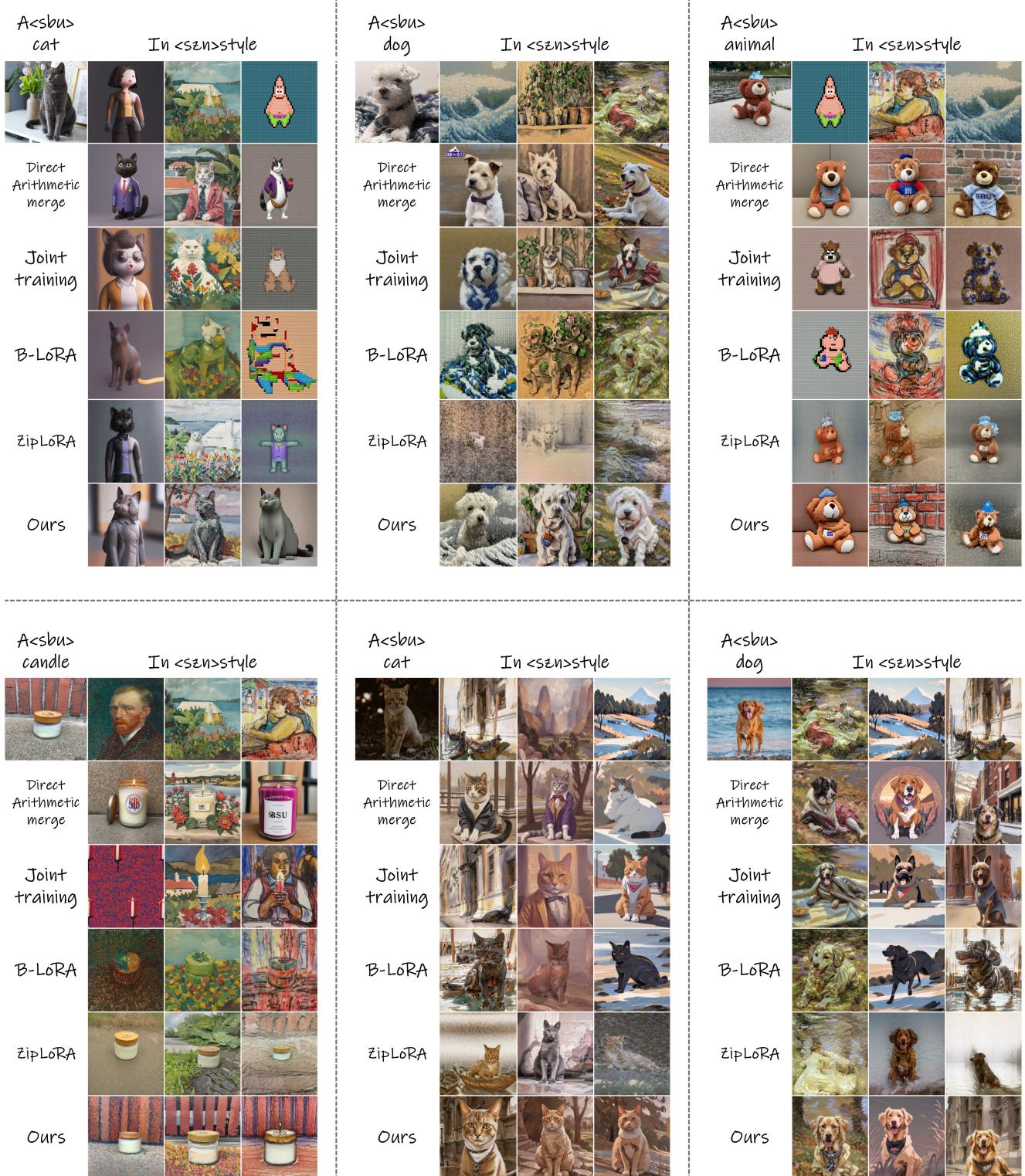


图 7. 定性比较。我们展示了由 K-LoRA 和其他对比方法生成的图像。K-LoRA 通常能够实现对象与风格的无缝融合，有效保持保真度并防止失真。

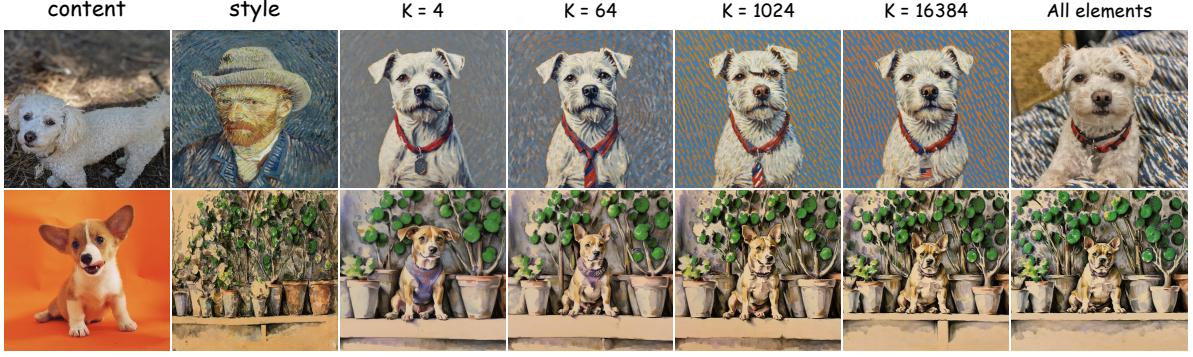


图 8.  $K$  的选择。评估 K-LoRA 中不同  $K$  值的影响。

我们方法的输出，以及用于训练主体和风格的参考图像。我们要求用户选出哪种方法最好地保留了风格和主体。结果如表 2 所示，表明我们的方法最受青睐。此外，我们也咨询了 GPT-4o 进行类似评估。我们的方法在 GPT-4o 的评估中显示出显著优势，进一步反映了我们方法的优越性。

Method	User Preference	GPT-4o Feedback
ZipLoRA [26]	29.2%	5.6%
B-LoRA [8]	18.1%	11.1%
Ours	52.7%	83.3%

表 2. 用户研究结果与 GPT-4o 反馈。

### 4.3. 消融分析

**Top-K 选择.** 我们进行了两组实验来验证 Top-K 选择方法的有效性：固定选择和随机选择。发现 (ii) 提示了一种直接的方法：如果缩放因子大于 1，则选择内容 LoRA；否则，选择风格 LoRA。我们将这种方法称为“固定选择”，它可作为测试 Top-K 选择方法消融实验的一个有效基线。它也可以被看作是多 LoRA 组合 [41] 的扩展和改进，该方法在某些场景下已显示出有前景的结果。然而，在特定的风格 LoRA 条件下，该方法可能导致物体模糊或内容外观的改变，如 Fig. 9 所示。

为确保我们的模块在指定的前向层配置中表现一致，而不是依赖于任意配置，我们使用随机种子进行了一项名为“随机选择”的对照实验。在此设置中，模型使用一个随机数，以  $1/3$  的概率选择内容注意力，以  $2/3$  的概率选择风格注意力。如 Fig. 9 所示，在这些随机选择条件下，生成的图像通常只保留了单一的物体

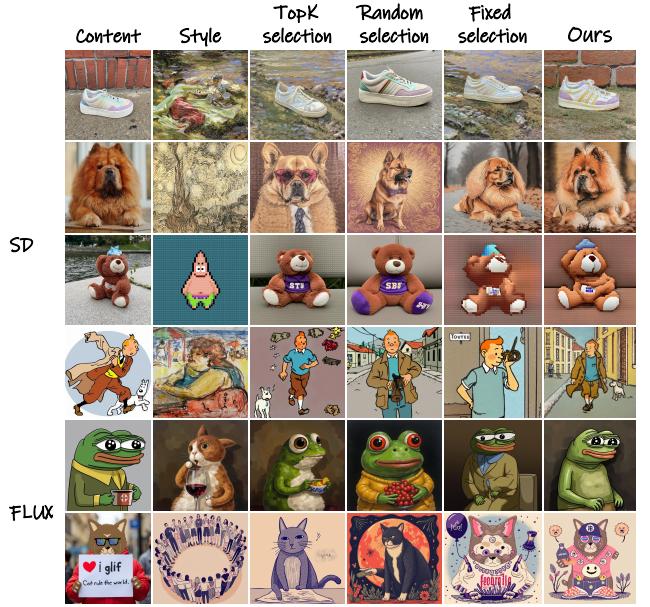


图 9. Top-K 选择与缩放因子的消融实验。我们使用五组图像比较了不同方法。上方四行代表 SD 的结果，而下方几行则展示了 FLUX 的结果，其中包括本地训练的 LoRA 和社区训练的 LoRA。

特征或风格特征，或者两者都未能保持。这一结果进一步验证了我们的发现 (ii)，突显了物体和风格组件在扩散过程的早期和晚期时间步中分别扮演的不同角色。

此外，我们评估了不同  $K$  值选择对生成图像的影响，如 Fig. 8 所示。在 Top-K 方法中，我们系统地改变了  $K$  的值。我们的观察表明，当  $K$  相对较小时，风格和物体的特征都不够突出。随着  $K$  的增加，这个问题逐渐改善。然而，如果  $K$  过大，风格可能无法保留，且物体的形状可能会发生显著扭曲。

**缩放因子.** 为了评估缩放因子的有效性，我们移除了

它，并仅关注原始的 Top-K 方法。在第一个实验中，如 Fig. 9 所示，我们的分析表明，虽然单独使用 Top-K 在某些条件下可以产生令人满意的结果，但扩大实验范围后会发现物体失真和风格丢失的情况。为了进一步评估缩放因子中  $\gamma$  的重要性，我们测试了两个来源不同、元素总和差异显著的 LoRA 模型的性能。如 Fig. 9 的最底一行所示，很明显 Top-K 选择未能准确捕捉风格，而固定选择中物体和风格的融合效果与我们的方法相比明显较弱。我们还尝试了另一种缩放方式，详细过程见补充材料（第 D 节）。

总之，移除这两个模块会导致生成性能下降，这突显了它们对模型整体有效性的关键贡献。

## 5. 结论

在本文中，我们介绍了 K-LoRA，该方法可以无缝地融合独立训练的风格 LoRA 模型和主体 LoRA 模型。K-LoRA 能够在精确地对物体进行微调的同时，保留原始风格的复杂细节。我们的方法通过 Top-K 选择和缩放因子，在每个扩散步骤中有效地利用了物体和风格 LoRA 的贡献，最大限度地利用了原始权重，并实现了精确的风格融合，而无需重新训练或手动调整超参数。

## References

- [1] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 42(6):1–10, 2023. 3
- [2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 3
- [3] black-forest labs. Flux.1. <https://github.com/black-forest-labs/flux>, 2024. 1
- [4] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1897–1906, 2017. 1
- [5] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer. *arXiv e-prints*, art. arXiv:2312.09008, 2023. 12, 18
- [6] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2719–2727, 2020. 1
- [7] Jiahua Dong, Wenqi Liang, Hongliu Li, Duzhen Zhang, Meng Cao, Henghui Ding, Salman Khan, and Fahad Shahbaz Khan. How to continually adapt text-to-image diffusion models for flexible customization? *arXiv preprint arXiv:2410.17594*, 2024. 3
- [8] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using blora. *arXiv preprint arXiv:2403.14572*, 2024. 1, 2, 3, 6, 8
- [9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 1
- [10] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [11] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024. 3
- [12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3
- [13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 1
- [14] Jiaxiu Jiang, Yabo Zhang, Kailai Feng, Xiaohe Wu, and Wangmeng Zuo. Mc<sup>2</sup>: Multi-concept guidance for customized multi-concept generation. *arXiv preprint arXiv:2404.05268*, 2024. 3
- [15] Dawid J Kopiczko, Tijmen Blankevoort, and Yuki M

- Asano. Vera: Vector-based random matrix adaptation. *arXiv preprint arXiv:2310.11454*, 2023. 3
- [16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3
- [17] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1
- [18] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023. 3
- [19] Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzcinski, and Sebastian Cygert. Magmax: Leveraging model merging for seamless continual learning. *arXiv preprint arXiv:2407.06322*, 2024. 4
- [20] Or Patashnik, Daniel Garabi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23051–23061, 2023. 2, 4
- [21] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6
- [22] Pengjie Ren, Chengshun Shi, Shiguang Wu, Mengqi Zhang, Zhaochun Ren, Maarten Rijke, Zhumin Chen, and Jiahuan Pei. Melora: Mini-ensemble low-rank adapters for parameter-efficient fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3052–3064, 2024. 3
- [23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 3, 5, 12
- [25] Simo Ryu. Merging loras. <https://github.com/cloneofsimo/lora>, 2023. 2, 3, 6, 12
- [26] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2024. 1, 2, 3, 4, 5, 6, 8
- [27] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 3
- [28] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 2, 5, 12
- [29] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2, 3, 4
- [30] Taiqiang Wu, Jiahao Wang, Zhe Zhao, and Ngai Wong. Mixture-of-subspaces in low-rank adaptation. *arXiv preprint arXiv:2406.11909*, 2024. 3, 4
- [31] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédéric Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20, 2024. 3
- [32] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 3

- [33] Yu xin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42:1 – 14, 2023. 12
- [34] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 2
- [35] Youcan Xu, Zhen Wang, Jun Xiao, Wei Liu, and Long Chen. Freetuner: Any subject in any style with training-free diffusion. *arXiv preprint arXiv:2405.14201*, 2024. 2, 4
- [36] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2403.11627*, 2024. 3
- [37] 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 arXiv:2308.06721*, 2023. 2
- [38] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 6
- [39] Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, and Bo Li. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning. *arXiv preprint arXiv:2308.03303*, 2023. 3
- [40] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023. 3
- [41] Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-lora composition for image generation. *arXiv preprint arXiv:2402.16843*, 2024. 2, 3, 4, 8, 12, 19
- [42] Donghao Zhou, Jiancheng Huang, Jinbin Bai, Jiaze Wang, Hao Chen, Guangyong Chen, Xiaowei Hu, and Pheng-Ann Heng. Magictailor: Component-controllable personalization in text-to-image diffusion models. *arXiv preprint arXiv:2410.13370*, 2024. 2
- [43] Hongyun Zhou, Xiangyu Lu, Wang Xu, Conghui Zhu, and Tiejun Zhao. Lora-drop: Efficient lora parameter pruning based on output evaluation. *arXiv preprint arXiv:2402.07721*, 2024. 3
- [44] Bojia Zi, Xianbiao Qi, Lingzhi Wang, Jianan Wang, Kam-Fai Wong, and Lei Zhang. Delta-lora: Fine-tuning high-rank parameters with the delta of low-rank matrices. *arXiv preprint arXiv:2309.02411*, 2023. 3

## 补充材料

本补充材料的结构安排如下：

1. 我们首先在第 A 节中，通过在广泛的数据集和不同模型的社区 LoRA 上评估我们的结果，来验证我们方法的有效性。
2. 我们在第 B 节中将我们的方法与其他方法进行了比较。
3. 我们在第 C 节中评估了复杂提示词对模型性能的影响。
4. 我们在第 D 节中实验了一种新的尺度，并测试了其对比效果。
5. 我们在第 E 节中，结合使用社区 LoRA 与本地 LoRA 进行了综合性能评估，并通过全面测试检验了随机种子对模型性能的影响。
6. 我们在第 F 节中测试了尺度因子中不同参数的选择。

### A. 视觉效果

我们使用了来自 StyleDrop [28] 和 DreamBooth [24] 的数据集，并在 Stable Diffusion (SD) 上进行了实验，如 Fig. 13 和 Fig. 14 所示。此外，我们还在 FLUX 上评估了我们的方法，使用了来自 Hugging Face 的 LoRA 模型，效果如 Fig. 11 和 Fig. 12 所示。通过系统性地结合这些对象 LoRA 和风格 LoRA，我们获得了一系列图像。这些结果表明，我们的方法能够有效地无缝融合对象与风格，并产生一致且高质量的视觉输出。

### B. 补充比较

我们增加了与 StyleID [5] 的比较，如 Fig. 15 所示。可以观察到，StyleID [5] 在保留纹理质量的同时，有效地实现了风格迁移。然而，其生成的对象可能会有些许模糊，或者生成的风格不够鲜明。此外，与我们的方法相比，他们的方法基于原始图像的固定布局，可能无法很好地泛化到不同的背景和动作。

### C. 提示词控制

我们进行实验以评估我们的方法是否能通过调整提示词 (prompt) 来修改对象的动作、周围环境或引入新元素。如 Fig. 18 和 Fig. 19 所示，在修改提示词后，

我们的方法有效地保留了原始对象的特征和风格属性，同时也无缝地融入了新的元素或场景细节。

### D. 新的尺度

在本文的主体部分，我们使用了由 Eqn. (7) 给出的尺度，具体如下：

$$S = \alpha \cdot \frac{t_{now}}{t_{all}} + \beta. \quad (11)$$

受 [33] 的启发，我们还引入了另一种尺度因子：

$$S^* = \left( \alpha' \cdot \frac{t_{now}}{t_{all}} + \beta' \right) \% \alpha. \quad (12)$$

在此等式中，我们设置  $\alpha' = 1.5$  和  $\beta' = 1.3$ ，这意味着在生成过程的初始阶段，风格信息在一定程度上被增强了，从而使得模型能够从风格 LoRA 中捕捉到特定的块信息。下面的 Fig. 10 展示了两种尺度之间的主要区别。

对于  $S^*$  的结果，由于风格信息在扩散的早期阶段得到了增强，生成的图像会从风格 LoRA 中捕捉到背景和颜色块信息。然而，这种方法会导致模型对风格 LoRA 中的纹理和笔触信息的学习效果减弱。这是一种权衡，用户可以根据自己的偏好选择不同的尺度因子。



图 10. 不同尺度因子的结果。K-LoRA 在不同尺度因子下的相应生成结果，其中每个对象-风格对都随机选择了两个种子进行生成。

### E. 鲁棒性分析

我们评估了来自不同来源的 LoRA 模型，其中对象 LoRA 来自社区，而风格 LoRA 则在本地进行训练。我们还比较了 DirectMerge [25]、Multi-LoRA composition [41] 以及我们提出的固定选择 (Fixed Selection) 方法。如 Fig. 16 所示，我们的方法在学习对象和风格特征方面均表现出卓越的性能，优于其他方法。此外，我们通过选择随机种子来测试我们方法的鲁棒性，以

评估其稳定性。如 Fig. 17 中呈现的结果所示，我们的方法在广泛的种子选择范围内始终能实现稳定的融合，从而确保了可靠的集成效果。

## F. 补充消融实验

在正文中，我们采用了一个包含两个超参数  $\alpha$  和  $\beta$  的缩放因子。具体来说，我们将  $\alpha$  设置为 1.5， $\beta$  设置为 0.5，从而使得对象和风格能够在不同位置施加不同程度的影响。为了验证所选参数的适切性，我们计算了 18 组随机选择的生成图像与其对应的原始对象/风格参考之间的 CLIP 相似度分数。下表中显示的结果是 CLIP 相似度分数的总和。

$\beta \setminus \alpha$	1.0	1.5	2.0
0.25	125.3%	126.7%	127.0%
0.50	126.5%	<b>128.1%</b>	126.2%
0.75	124.5%	125.8%	125.3%

我们可以看到， $\alpha$  和  $\beta$  的最优设置分别为 1.5 和 0.5。根据我们的实验，该权重配置几乎能满足所有的内容-风格对，用户无需进行额外调整。



图 11. 使用 FLUX 的补充生成结果。每个位置的图像对应其上方的对象和左侧的风格，展示了使用我们的方法应用不同 LoRA 生成的结果。

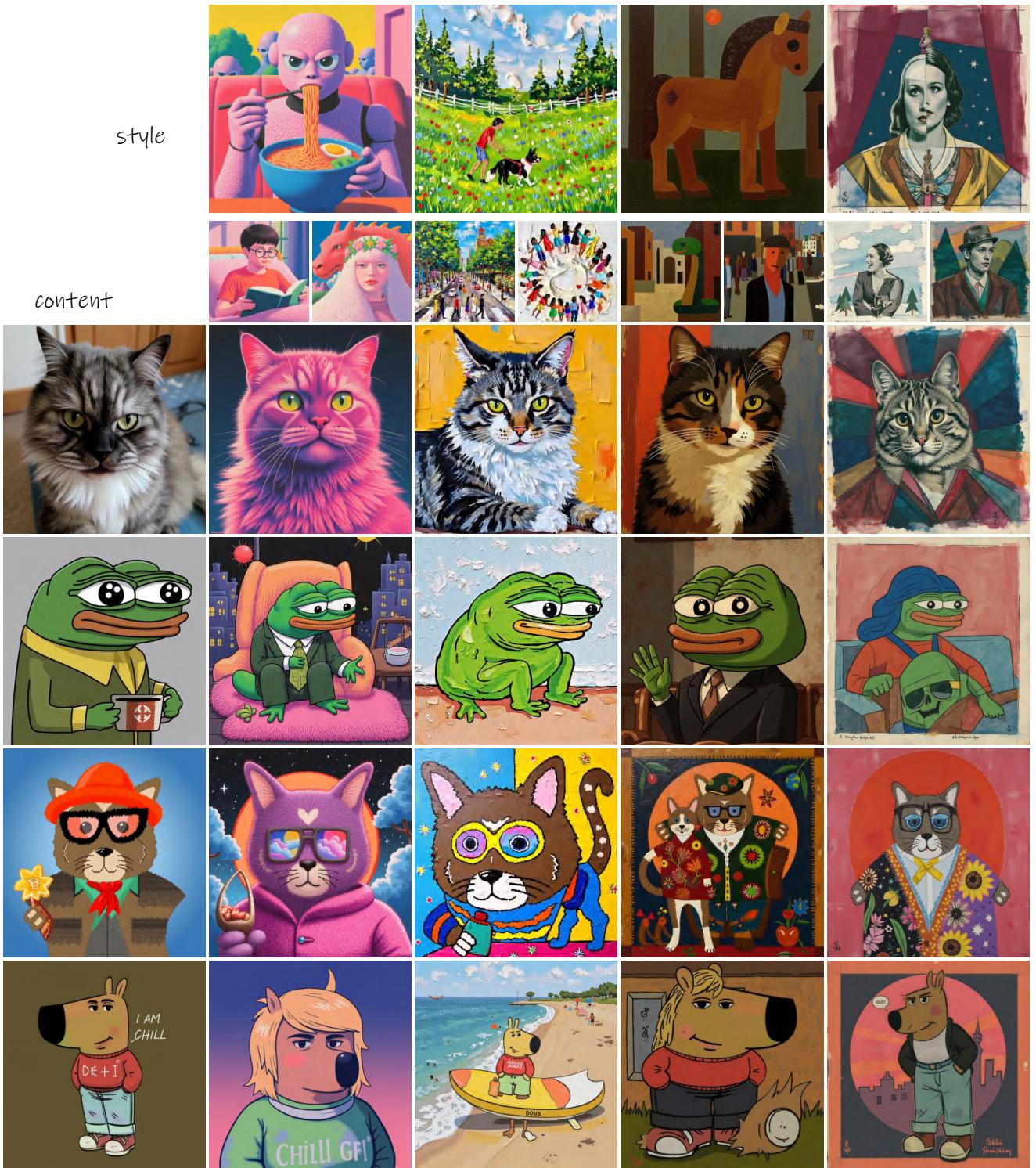


图 12. 使用 FLUX 的补充生成结果。每个位置的图像对应其上方的对象和左侧的风格，展示了使用我们的方法应用不同 LoRA 生成的结果。



图 13. 使用 SD 的补充生成结果。每个位置的图像对应其上方的对象和左侧的风格，展示了使用我们的方法应用不同 LoRA 生成的结果。

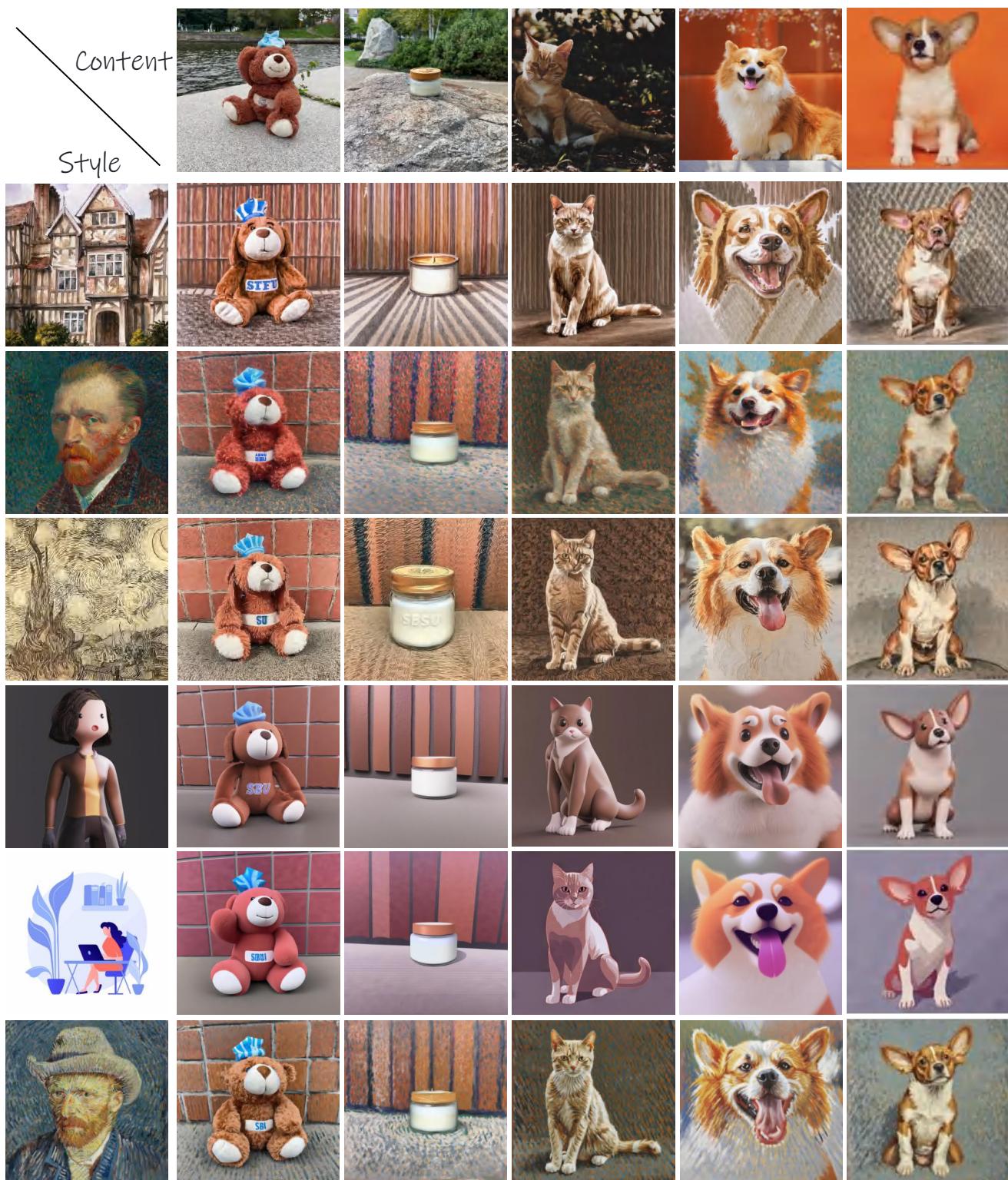


图 14. 使用 SD 的补充生成结果。每个位置的图像对应其上方的对象和左侧的风格，展示了使用我们的方法应用不同 LoRA 生成的结果。

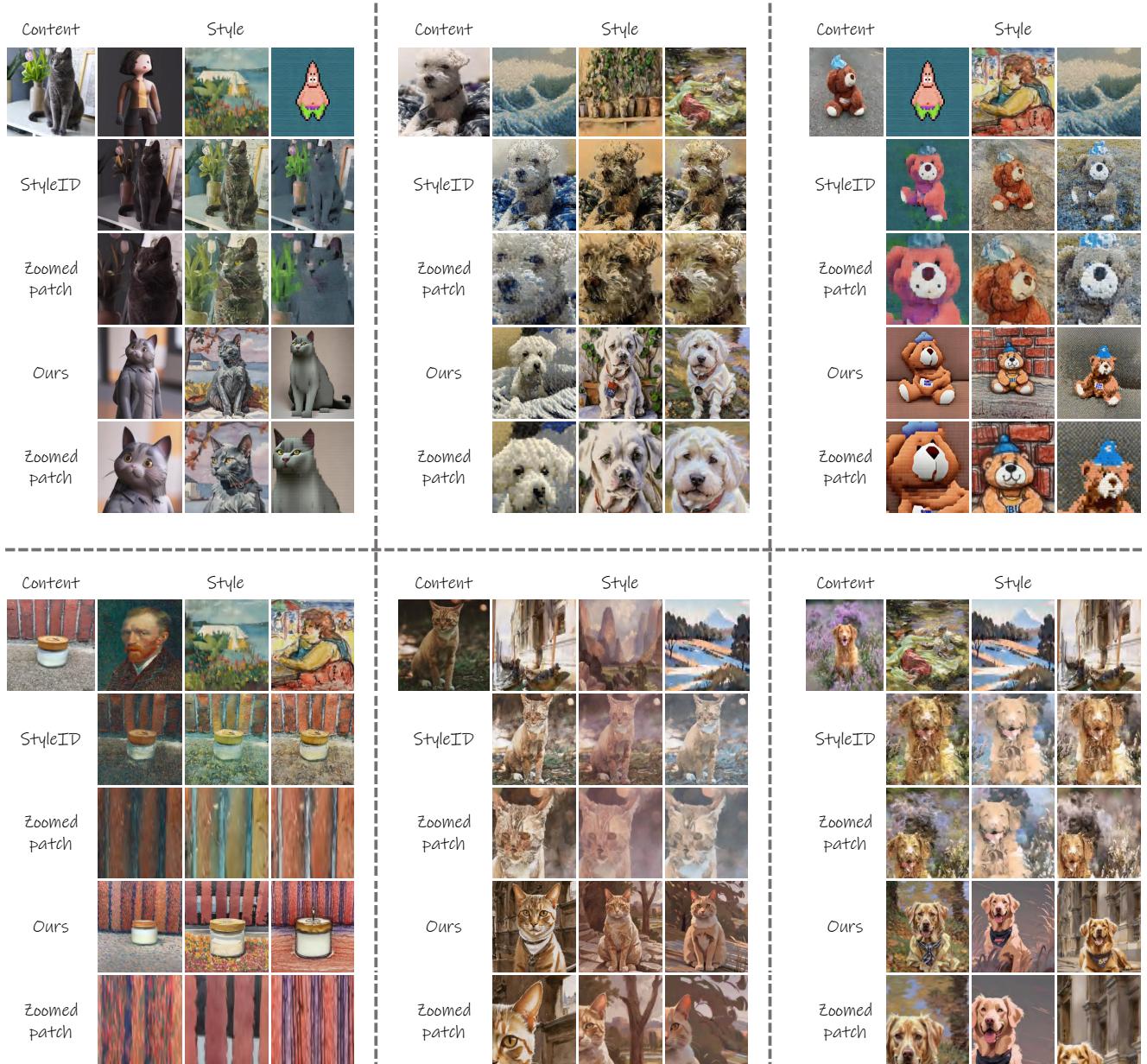


图 15. 补充比较。我们比较了 StyleID [5] 方法，并截取输出图像中的局部放大图块来观察细节纹理信息和风格特征。在每个块中，第二行和第三行分别展示了 StyleID 的结果及其对应的局部放大图块，而随后的两行则展示了我们方法的结果及相关的局部放大图块。

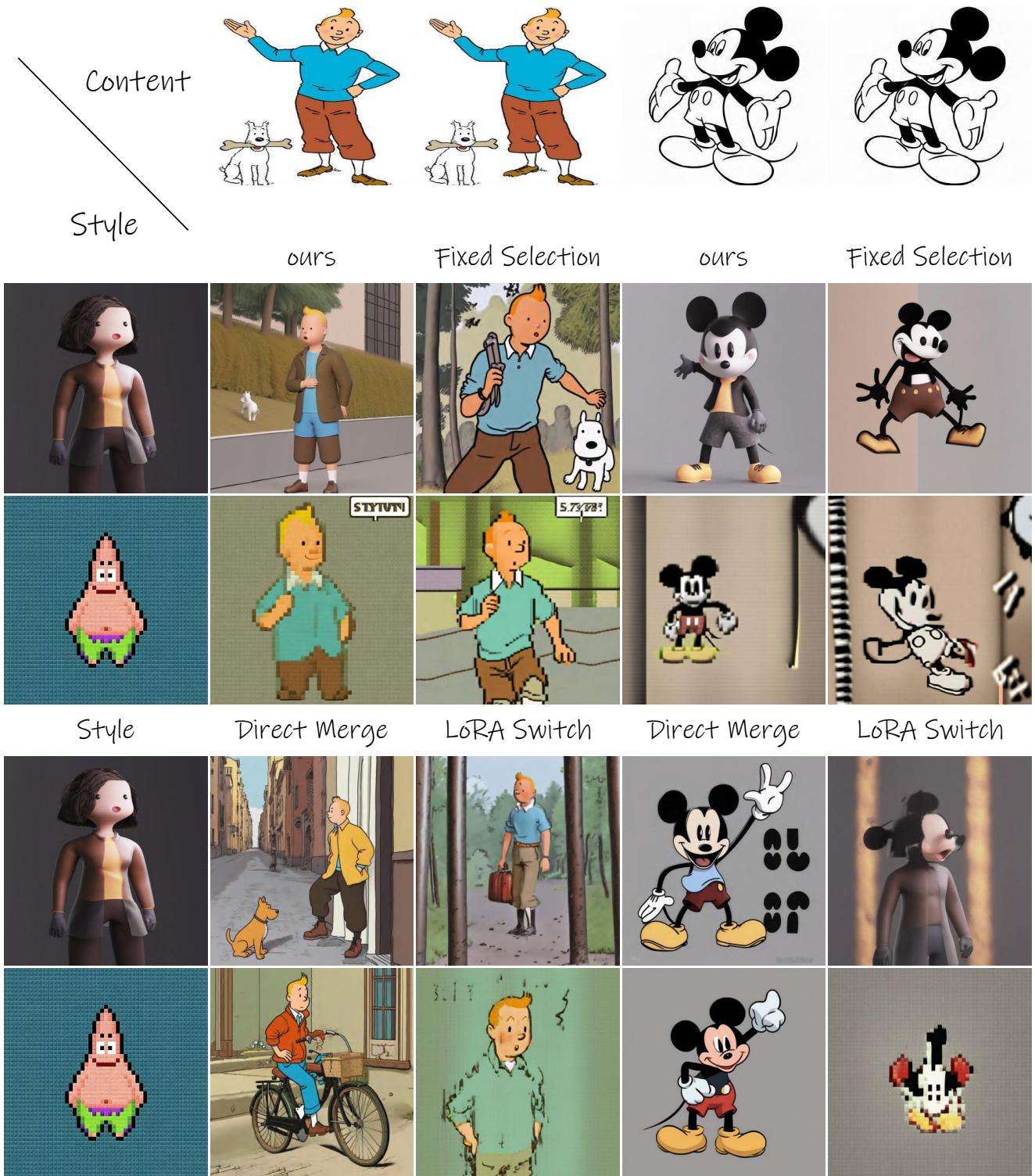


图 16. 鲁棒性验证。我们使用社区 LoRA 和本地训练的 LoRA，与正文中提出的固定选择方法、作为基线比较的直接算术合并 LoRA 方法，以及 Multi-LoRA Composition [41] 方法进行比较，以验证方法的泛化性和鲁棒性。

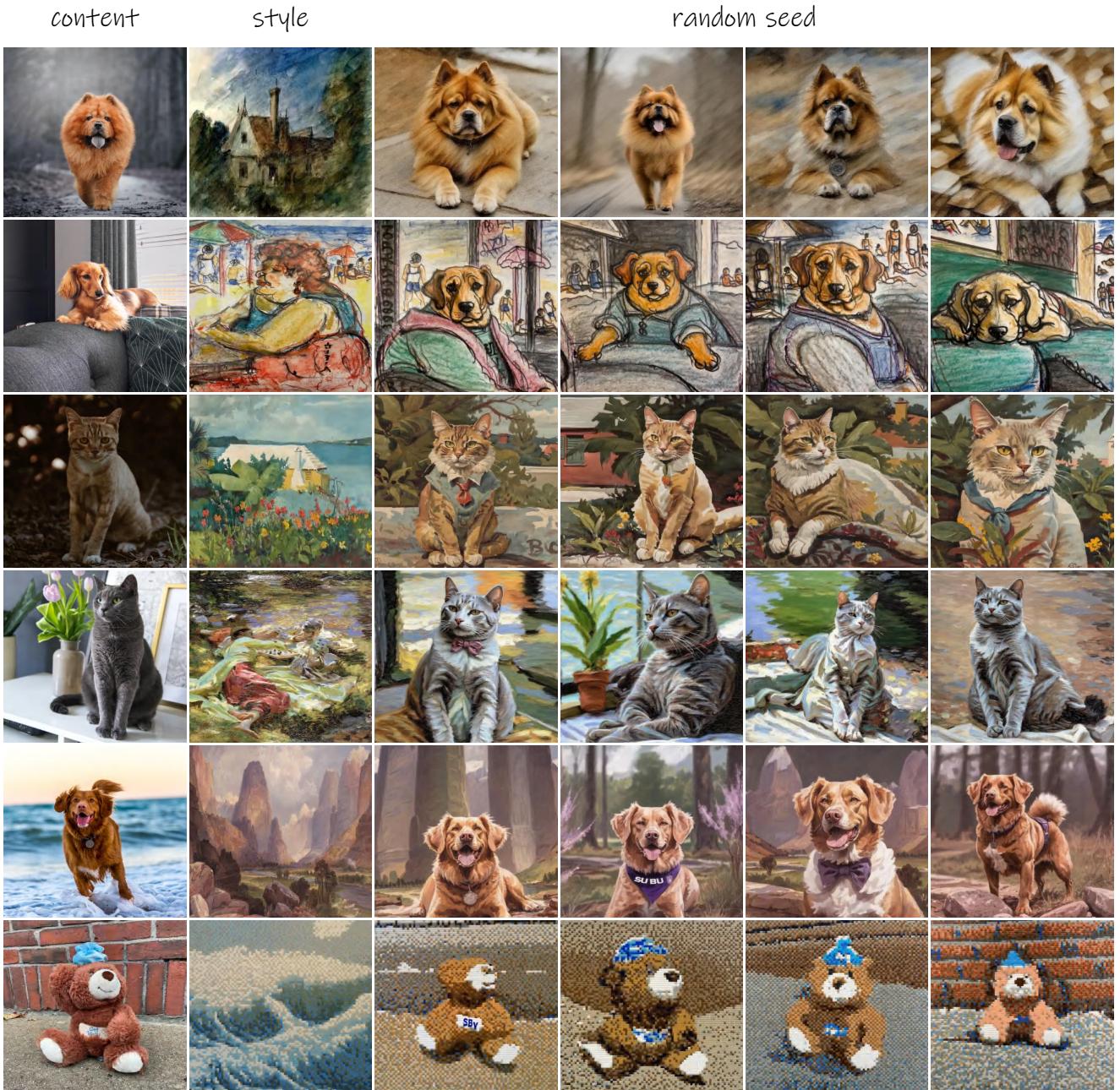


图 17. 鲁棒性验证。我们随机选择种子以进一步验证稳定性。

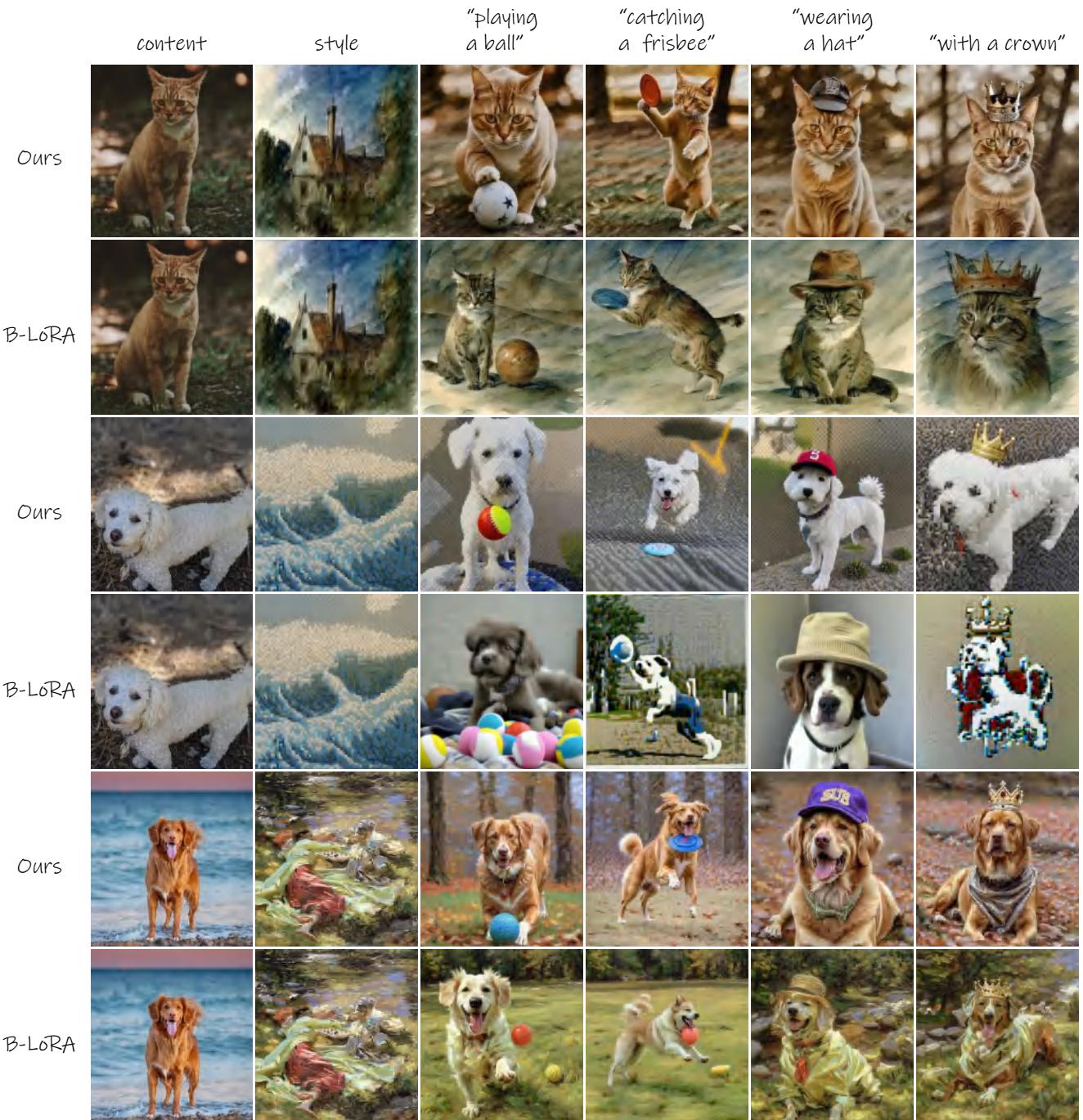


图 18. 提示词控制。我们引入了关于新场景、新动作和新对象的提示词，以验证我们方法对内容进行重新情境化并保持风格一致性的能力。

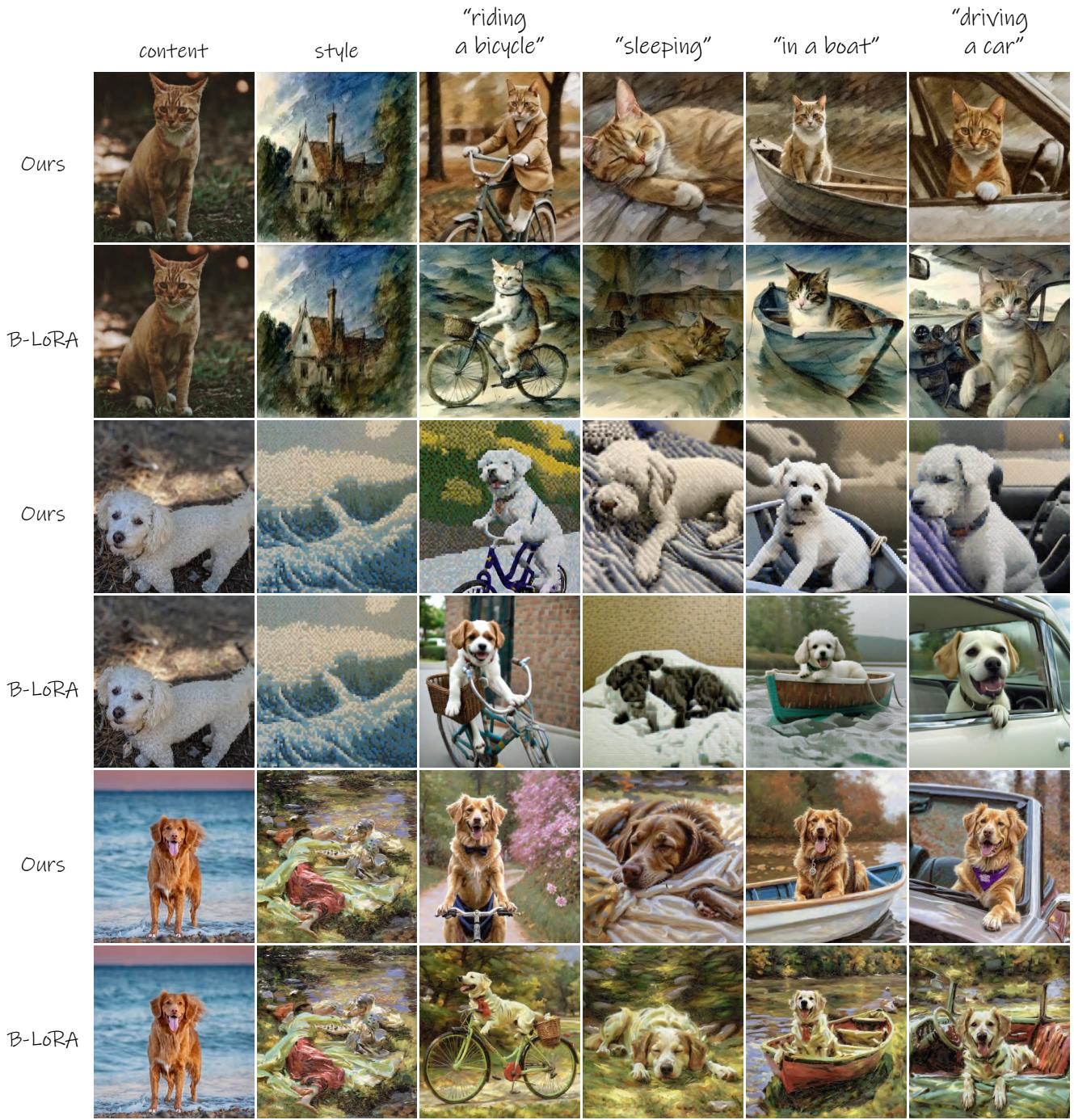


图 19. 提示词控制。我们引入了关于新场景、新动作和新对象的提示词，以验证我们方法对内容进行重新情境化并保持风格一致性的能力。