

# Implicit Style-Content Separation using B-LoRA

## 摘要

图像风格化涉及在保留图像的基本物体、结构和概念（内容）的同时，操作图像的视觉外观和纹理（风格）。风格和内容的分离对于独立操作图像的风格至关重要，确保最终结果和谐且视觉上令人愉悦。实现这种分离需要深入理解图像的视觉和语义特征，通常需要训练专门的模型或进行复杂的优化。在本文中，我们介绍了 B-LoRA，一种利用 LoRA（低秩适配）方法隐式分离单一图像的风格和内容组件，从而促进各种图像风格化任务。通过分析结合 LoRA 的 SDXL 架构，我们发现联合学习两个特定块（称为 B-LoRA）的 LoRA 权重可以实现风格-内容分离，而单独训练每个 B-LoRA 无法实现这一目标。将训练过程集中到仅两个块，并进行风格和内容的分离，可以显著提高风格操作，并克服通常与模型微调相关的过拟合问题。一旦训练完成，这两个 B-LoRA 可以作为独立组件用于各种图像风格化任务，包括图像风格迁移、基于文本的图像风格化、一致风格生成和风格-内容混合。

**关键词：**风格迁移；扩散模型；LoRA

## 1 引言

风格迁移是视觉艺术和计算机视觉领域的一个重要技术，它将一张图像的内容与另一张图像的风格相结合，生成具有独特美学吸引力的风格化图像。图 1 展示了风格迁移的一个例子。早期的风格迁移方法通常基于非真实感渲染（Non-Photorealistic Rendering, NPR）技术 [5, 23]。这些方法往往依赖于手工设计的特征或先验知识，导致其不同风格间的泛化能力较差。随着神经网络在视觉任务中的成功应用，神经风格迁移（Neural Style Transfer, NST）利用卷积神经网络来提取图像的内容和风格特征 [4]，在自动生成风格化图像方面取得了显著进展。然而，传统的 NST 方法通常依赖 CNN 来捕捉低级图像特征，仍然缺乏对更复杂结构细节和纹理的高质量重现能力。此外，许多 NST 方法依赖逐层优化，这种方法计算复杂度较高，且在生成更高分辨率风格化图像时往往难以稳定，限制了 NST 在实际应用中的广泛采用。

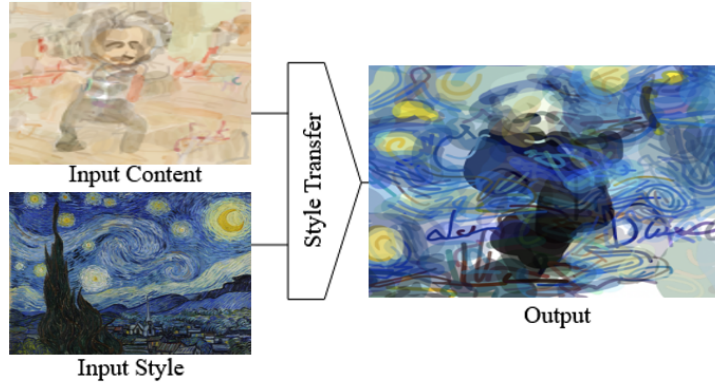


图 1. 将“星空”的风格转移到给定的照片上

近年来，扩散模型作为一种基于概率的生成模型，在计算机视觉领域引起了广泛关注 [9]。扩散模型通过逐步去噪的过程生成图像，展示了在生成精细和复杂纹理方面的强大潜力。扩散模型采用噪声逐步添加和去噪的过程，在每一步生成图像并学习更精细的局部特征分布。在风格迁移应用中，扩散模型逐渐生成的特性有助于捕捉更丰富的纹理细节和局部结构，并能平衡内容和风格信息，为风格迁移领域开辟了新的研究方向。利用这些特性，扩散模型逐渐成为风格迁移领域的重要方法。

文章在最近提出的 Stable Diffusion XL (SDXL) [16] 上使用 LoRA [10]，这是一种以强大的风格学习能力著称的文本到图像扩散模型。通过对 SDXL 中各个层及其对自适应过程的影响进行详细分析后发现：可以使用两个特定的 Transformer 块来分离输入图像的风格和内容，并在生成的图像中独立控制它们。为澄清起见，本文中将一个“块”定义为由 10 个连续的注意力层组成的序列。

因此，当输入一张单一图像时，文中方法联合优化了这两个特定 Transformer 块的 LoRA 权重，其目标是基于提供的文本提示重建给定图像。由于仅优化了这两个 Transformer 块的 LoRA 权重，因此将其称为“B-LoRAs”。关键在于，这些 B-LoRAs 仅针对一张图像进行训练，但成功地分离了图像的风格与内容，从而避免了常见 LoRA 技术中因过拟合导致的风格-内容混淆问题。

需要指出的是，最近有一些研究尝试将风格和内容的 LoRA 模型组合成统一模型 [21]。然而，该方法需要为每种风格-内容组合进行新的优化过程，这既耗时，又难以在风格变换与内容保留之间取得有效平衡。相比之下，B-LoRAs 可以轻松重新插入到预训练模型中，并与其他参考图像的学习块组合使用，无需进一步训练。

## 2 相关工作

### 2.1 图像风格迁移

风格迁移在图像处理和计算机视觉领域中扮演着重要角色。该领域已从早期的手动纹理合成 [30] 迅速发展到先进的神经风格迁移 (Neural Style Transfer, NST) [4, 12]，标志着从传统技术向现代深度学习方法的重大转变。生成对抗网络 (GANs) [6] 由于其卓越的图像生成

能力，被迅速应用于风格迁移任务 [32]，进一步推动了该领域的进步。随着扩散模型 [9] 的快速发展，图像风格迁移取得了显著的突破。

这些技术主要可分为两种方法：基于微调的方法和基于反演的方法。基于微调的方法 [11, 19, 26] 通过使用大量风格图像优化模型的部分或全部参数，从而将这些风格嵌入到模型的输出域中。相比之下，基于反演的方法 [28, 31] 利用风格图像或内容图像，将风格或内容概念嵌入到特定的词嵌入向量中，并通过包含这些词嵌入的提示词实现风格迁移。

上述基于扩散模型的方法通常需要风格图像来训练模型，导致优化过程相对缓慢。近期研究引入了跨图像注意力机制 [8]，提出了一种无需额外优化的风格迁移方法。

## 2.2 个性化文生图

基于文本引导的风格化图像生成已经得到了广泛研究，尤其是在文本到图像的扩散模型方面。当前的方法可分为三类：(1) 基于优化的方法 [3, 19, 22, 31]；(2) 非优化方法 [13, 17, 27]；(3) 无需训练的方法 [8, 25]。

具体而言，基于优化的方法在包含目标风格的少量参考图像上微调模型。例如，Textual Inversion [3] 通过优化与风格绑定的特殊标记的嵌入向量，并保持其他参数不变。Dream-Booth [19] 则微调 U-Net 的所有权重，并引入了先验保持损失，以避免遗忘先前学到的知识。

非优化方法在大规模风格化图像上微调 T2I 模型，以实现实时的风格化文本到图像生成。由于缺乏成对的（参考图像-样本）数据，这些方法不得不使用样本自身（或裁剪的局部区域）作为参考图像，这可能导致内容泄露。为了解决这一问题，最新的 DEADiff [17] 构建了两个成对数据集，并提出了一种内容-风格解耦机制和非重建学习方法，以实现实时风格迁移。

此外，还有无需训练的方法，可以在离线或在线环境下，在不进行参数调整的情况下完成风格化图像生成。StyleAligned [8] 在扩散过程中引入了带有 AdaIN（自适应实例归一化）调制的注意力共享操作，以保持与批次中第一张图像风格的一致性。而 InstantStyle [25] 则提出了通过在 U-Net 中向特定的风格模块注入参考图像特征，从而在不调整权重的情况下实现了高效的风格迁移。

## 2.3 LoRA 用于图像风格化

LoRA (Low-Rank Adaptation) 常用于图像风格化，通过微调模型以生成具有特定风格的图像。通常，LoRA 是在一组图像上进行训练的，然后与控制方法（如 ControlNet [29]）以及文本提示相结合，以控制生成图像的内容。尽管基于 LoRA 的方法在捕捉风格与内容方面表现出了显著能力，但完成该任务通常需要两个独立的 LoRA 模型，且缺乏一种直接且简单的方式将二者进行结合。一种常见的简单方法是直接对两个 LoRA 模型的权重进行插值 [20]，依赖于手动搜索所需的系数。其他方法 [7, 15] 提出了基于优化的策略，用于寻找此类组合的最优系数。然而，这些方法主要关注的是两个物体的组合，而非图像风格化任务。

最近，Shah 等人提出了 ZipLoRA [21]，通过学习 LoRA 矩阵列的混合系数，将分别针对风格和内容训练的两个 LoRA 模型合并为一个新的“压缩”LoRA。这项研究与我们的工作密切相关，因为我们也尝试混合在不同图像上训练的 LoRA 权重，以实现图像风格化。然而，ZipLoRA 需要为每个新的内容与风格组合进行额外的优化阶段，从而限制了 LoRA 权重的灵活重用性，而灵活性正是 LoRA 的核心优势之一。相比之下，我们的方法允许直接重用已学

习的风格与内容 LoRA 权重，无需额外训练，从而提高了效率与适用性。此外，我们展示了我们的隐式方法在处理具有挑战性的风格与内容时更加鲁棒。

### 3 本文方法

#### 3.1 本文方法概述

文章将 LoRA 与 Stable Diffusion XL (SDXL) [16]，发现可以通过两个特定的 transformer block 将图像的风格和内容分离，并独立控制它们。每个 transformer block 由 10 个连续的注意力层组成，优化这两个块的 LoRA 权重后，能够基于文本提示重建图像，同时避免了常见 LoRA 方法中的过拟合问题。文章的 B-LoRA 方法无需额外的训练或微调即可进行风格迁移、文本引导的风格操作和一致的风格条件图像生成，如图 2。此外，B-LoRA 可以与其他参考图像的 transformer block 结合，无需重新训练，提供了比现有方法更高效、简便的风格化操作。



图 2. B-LoRA 的隐式分离

#### 3.2 SDXL 架构探索

本文使用了升级版的 Stable Diffusion XL (SDXL) [16]，它是 Stable Diffusion [18] 的改进版，采用潜在扩散模型 (LDM) 架构，扩散过程应用于预训练图像自编码器的潜在空间。SDXL 的 UNet 骨干网络比 Stable Diffusion 大三倍，具体架构如图 3，包含 70 个注意力层，分为 11 个 transformer block。每个块的层数不同，前两个和最后三个块分别为 4 和 6 层，中间六个块为 10 层。此外，SDXL 对文本条件生成进行了扩展，通过 OpenCLIP ViT-bigG 和 CLIP ViTL 两次编码文本提示，合并后的嵌入向量用于网络的交叉注意力层。



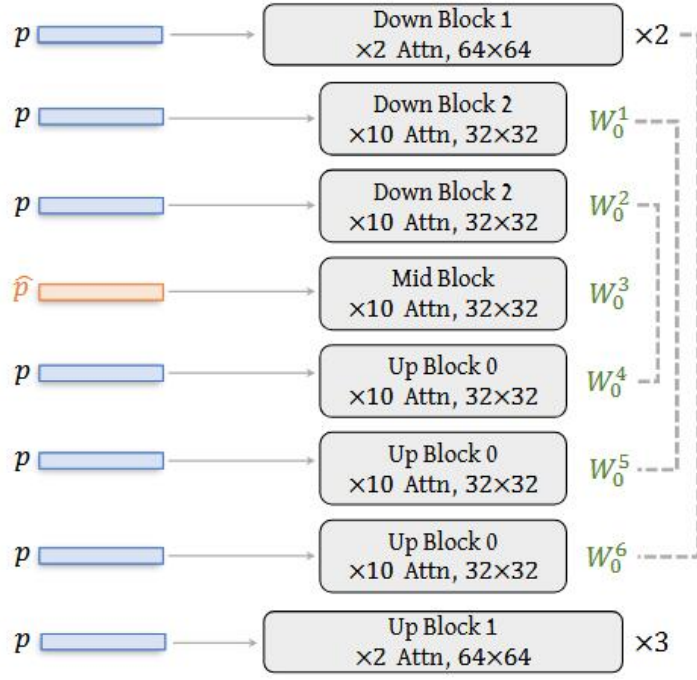


图 3. SDXL 的架构分析

文章研究了 SDXL 中不同层次对生成图像的影响，采用了类似 Voynov 等人 [24] 的方法。文章定义了两组随机文本提示：Pcontent（描述物体）和 Pstyle（描述颜色），通过将不同的文本提示注入 SDXL 中的某个 transformer block 的交叉注意力层，并检查生成图像与提示之间的相似度，如图 4 所示。结果显示，W2 和 W4 主导了生成图像的内容，而 W5 主导了颜色。

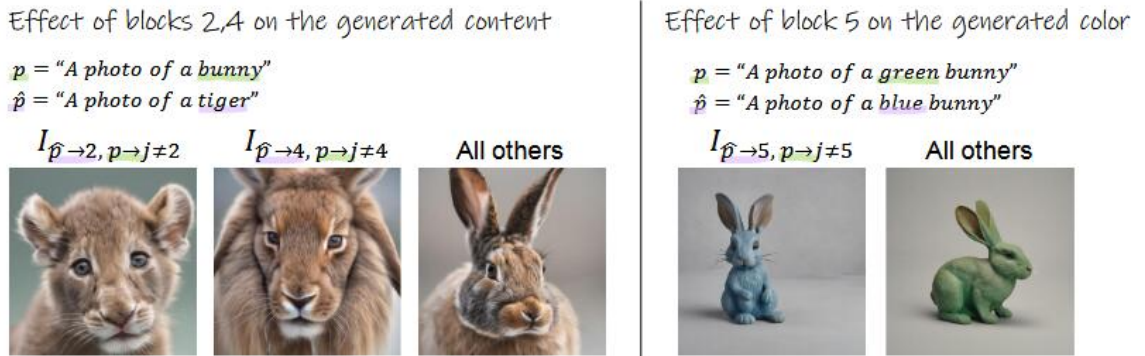


图 4. 对生成图像进行不同文本注入

### 3.3 基于 LoRA 的 B-LoRA 分离

接下来的目标是检查 SDXL 中定位的层是否能有效捕捉给定输入图像 I 的内容和风格，并通过 LoRA [10] 微调模型生成图像的变体。文章冻结基础预训练模型的权重  $W_0$ ，并优化每个变换器块的残差矩阵  $W_i$ 。在两个实验中，如图 5，分别优化 W2, W5 和 W4, W5，因为上面发现 W2 和 W4 主导内容，W5 主导颜色。文章使用通用提示 “A [v]” 避免过度引导风格或内容捕捉。结果表明，优化 W4, W5 能够最佳重建图像并捕捉其内容。此外，使用 UNet 较深层 W4 有助于保留图像细节。文中称这种训练方案为 B-LoRA，能够减少 70% 的存储需求。

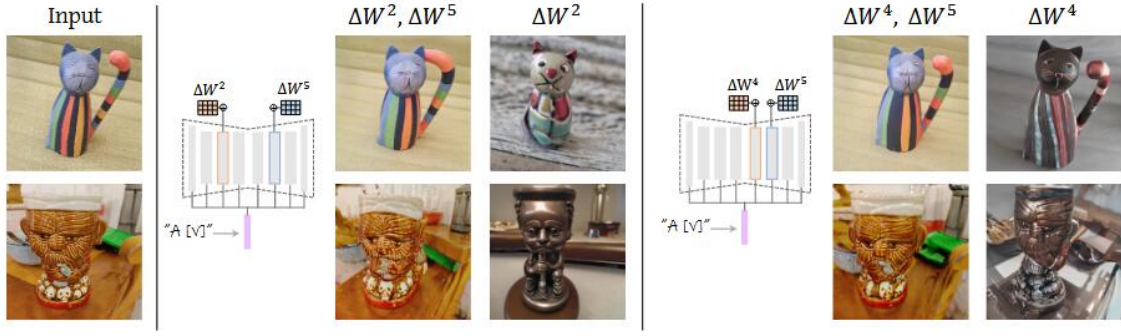


图 5. W2 和 W4 的对比分析试验

### 3.4 用于图像风格化的 B-LoRA

结合上述分析的见解，B-LoRA 训练方法定义为：给定输入图像  $I$ ，我们仅微调 LoRA 权重  $W_4$  和  $W_5$ ，目标是根据通用文本提示“A [v]”重建图像。除了提高效率外，文章发现仅训练这两层可以实现隐式的风格-内容分解，其中  $W_4$  捕捉内容， $W_5$  捕捉风格。一旦找到这些更新矩阵，就可以通过更新预训练 SDXL 模型的相应块权重，轻松地将其应用于风格操作任务，如下文所述并在图 6 中展示。

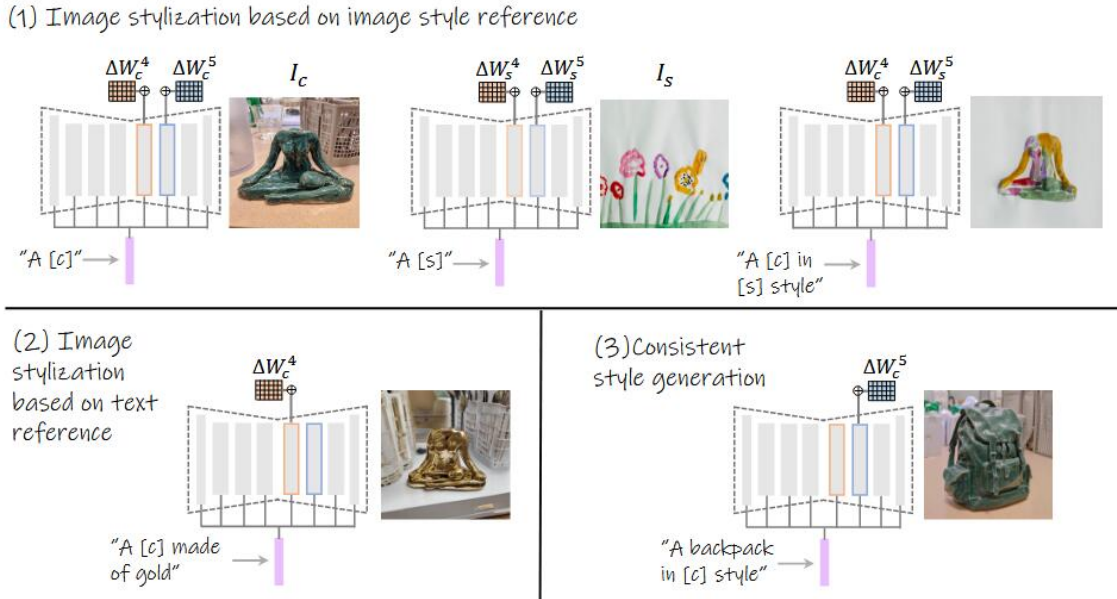


图 6. 用于图像风格化的 B-LoRA

## 4 复现细节

### 4.1 与已有开源代码对比

此篇文章的代码在 GitHub 上进行了开源(<https://github.com/yardenfren1996/B-LoRA>)，在已经开源的代码上，我首先复现了 B-LoRA，并发现对于一些抽象概念的迁移生成效果较差。经过分析，我认为这个问题可能与 B-LoRA 所使用的基础模型——SDXL 有关。在 SDXL 模型的训练过程中，抽象概念的表达和理解可能没有得到充分的训练，缺乏足够的先验知识或训练样本，这导致它在迁移生成时未能有效捕捉这些抽象概念的语义特征。

为提升抽象概念的生成能力，我采取了两种策略：一是对 SDXL 模型进行领域适应微调训练，即通过增加包含抽象概念的图像数据集，使模型学习这些概念的视觉特征，从而增强其理解与生成抽象概念的能力；二是使用一个具备更强抽象概念生成能力的新模型，对于方法二我首先是使用 SD3 进行测试，并最终选择 Black Forest Labs 发布的 FLUX.1 模型，该模型采用 MM-DiT (Diffusion Transformer) [2] 架构，替代了 SDXL 原有的 U-Net 架构，在抽象概念生成方面表现更为出色，能够更精准地分离风格与内容，展现出更高的灵活性与生成质量。

## 4.2 实验环境搭建

复现工作的环境为：Ubuntu 20.04.6 LTS、55GB 内存、Quadro RTX 6000(24G) 显卡。根据作者的 Readme (<https://github.com/yardenfren1996/B-LoRA/blob/main/README.md>) 进行虚拟环境搭建及依赖安装。复现项目的模型权重在 hugging face 下载，SDXL 模型权重下载地址：<https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0/tree/main>。FLUX.1 模型权重下载地址：<https://huggingface.co/black-forest-labs/FLUX.1-dev>，实测 FLUX.1 在单卡 24G 显存进行训练时会出现显存不够的情况，所以使用的是 A100 (40G) 进行训练。

## 4.3 界面分析与使用说明

训练 B-LoRAs

要针对给定的输入图像训练 B-LoRA，请运行下面脚本：

```
1 accelerate launch train_dreambooth_b-lora_sd-xl.py \
2   --pretrained_model_name_or_path="stabilityai/stable-diffusion-xl-base
   -1.0" \
3   --instance_data_dir="<path/to/example_images>" \
4   --output_dir="<path/to/output_dir>" \
5   --instance_prompt="<prompt>" \
6   --resolution=1024 \
7   --rank=64 \
8   --train_batch_size=1 \
9   --learning_rate=5e-5 \
10  --lr_scheduler="constant" \
11  --lr_warmup_steps=0 \
12  --max_train_steps=1000 \
13  --checkpointing_steps=500 \
14  --seed="0" \
15  --gradient_checkpointing \
16  --use_8bit_adam \
17  --mixed_precision="fp16"
```

这将优化内容和风格的 B-LoRA 权重并将其存储在 output\_dir 中。需要替换 instance\_data\_dir、output\_dir、instance\_prompt 的参数。

## 风格迁移图像生成

(1) 对于基于参考图像的图像风格迁移，请运行：

```
1 python inference.py --prompt="A <c> in <s> style" --content_B_LoRA="<path  
/to/content_B-LoRA>" --style_B_LoRA="<path/to/style_B-LoRA>" --  
output_path="<path/to/output_dir>"
```

这将生成具有第一个 B-LoRA 的内容和第二个 B-LoRA 的风格的新图像。注意，需要根据优化提示替换提示中的 c 和 s

(2) 对于基于文本的图像风格化，运行：

```
1 python inference.py --prompt="A <c> made of gold" --content_B_LoRA="<path  
/to/content_B-LoRA>" --output_path="<path/to/output_dir>"
```

这将生成具有给定 B-LoRA 内容和文本提示指定样式的新图像。

(3) 为了生成一致的风格，请运行：

```
1 python inference.py --prompt="A backpack in <s> style" --style_B_LoRA="<  
path/to/style_B-LoRA>" --output_path="<path/to/output_dir>"
```

这将生成具有指定内容和给定 B-LoRA 风格的新图像。

## 4.4 创新点

在复现 B-LoRA 模型时，我发现对于一些抽象概念的迁移生成效果较差。例如，复现的结果图7中，关于“Hope”、“Database”和“Education”等抽象概念的生成结果并不理想。经过分析，我认为这个问题可能与 B-LoRA 所使用的基础模型——SDXL 有关。在 SDXL 模型的训练过程中，抽象概念的表示和理解可能没有得到充分的训练。具体而言，SDXL 系列模型在处理抽象概念时，缺乏足够的先验知识或训练样本，这导致它在迁移生成时未能有效捕捉这些抽象概念的语义特征。因此，尽管 SDXL 在处理具体、形象的对象时表现较好，但在抽象概念的生成方面，其能力仍然存在一定局限性。

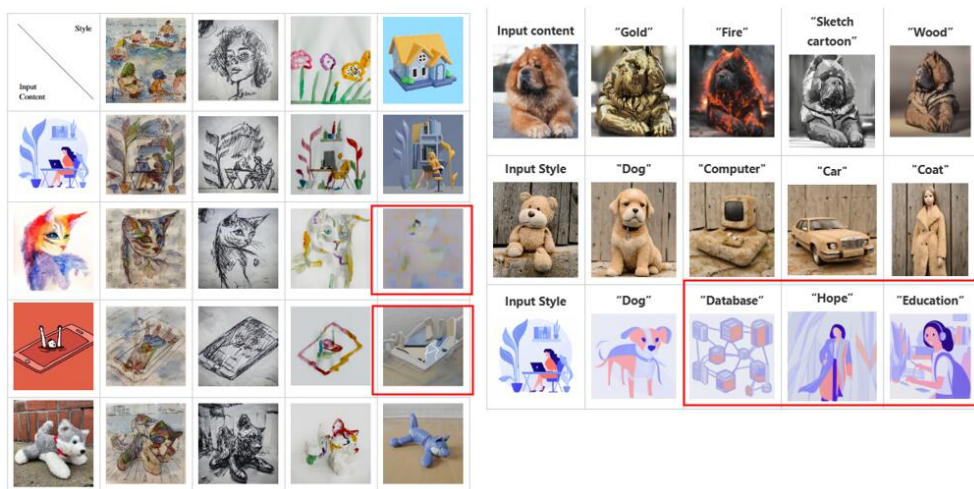


图 7. B-LoRA 的复现结果



对于抽象概念的生成，我们可以采取两种策略来解决当前模型在这方面的不足。第一种方法是领域适应，即通过对包含抽象概念的图像进行微调训练，使 SDXL 模型能够学习和识别这些抽象概念的特征。这种方法可以通过增加具有抽象性质的训练数据来增强模型对抽象概念的理解，从而提升生成效果。

第二种方法是寻找一个具备更强抽象概念生成能力的新模型。经过研究，我发现 Black Forest Labs 发布的 FLUX.1 模型是一个替代方案。与传统的 SDXL 模型不同，FLUX.1 将 SDXL 原有的 U-Net 架构替换为了 MM-DiT (Diffusion Transformer) [2] 架构，如图8。这一创新使得 FLUX.1 在生成抽象概念时表现出色，能够更好地理解和生成抽象的视觉特征。因此，FLUX.1 模型不仅克服了 SDXL 在抽象概念生成方面的局限性，而且在风格和内容分离的任务中也展现了更高的灵活性和精准度。

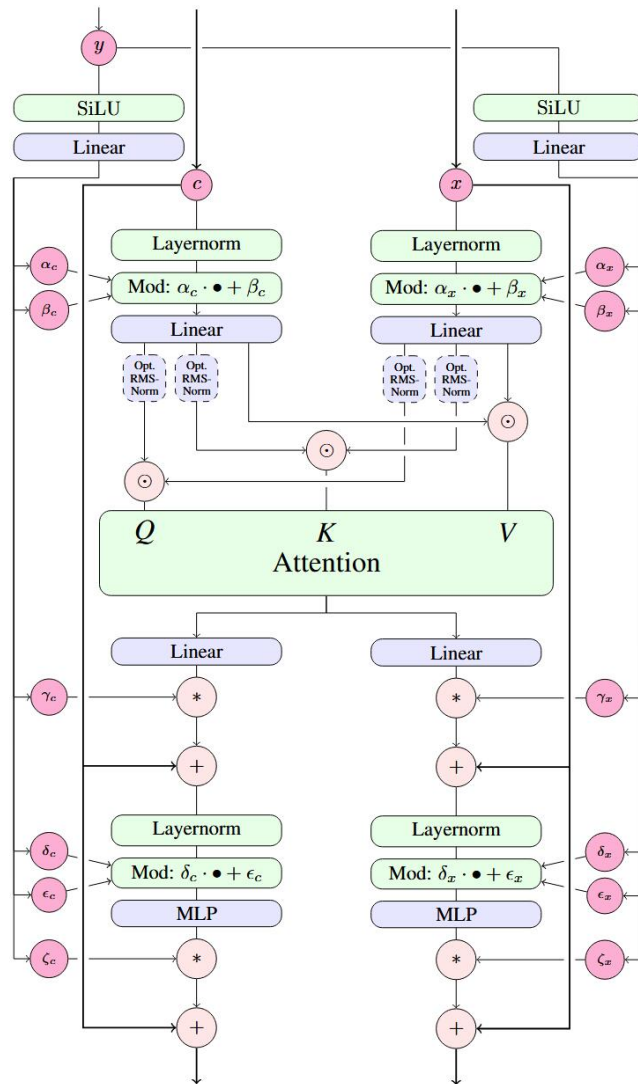


图 8. 一个 MM-DiT 块

同时在复现过程中，我选择了一组具有特殊性质的数据集进行实验。该数据集的特殊之处在于它融合了高度抽象的风格，而这种抽象风格与图像的内容相互交织，形成了一种复杂的表现形式。经过实验发现目前的所有方法都存在的对于深纠缠的内容风格不好迁移的通病。

## 5 实验结果分析

在第一种方法中，我使用 LoRA 对模型进行微调时，发现尽管 SDXL 模型能够生成大致的结果，但在细节还原方面存在一定的不足。具体来说，一些较为精细的细节未能得到有效的恢复，尤其是图9中红框标出的区域。在这些区域，模型未能准确地捕捉和再现图像的细微特征，导致生成结果缺乏应有的精确度和清晰度。这表明，尽管通过 LoRA 微调可以提升模型的整体生成效果，但在处理细节和高精度要求的任务时，SDXL 的能力仍然存在局限性，可能需要进一步的优化或调整。



图 9. LoRA 的微调训练

之后，我在使用以 DiT [14] 架构为基础而扩展的两个模型——SD3 [2] 和 FLUX.1 上进行了实验，生成了包含抽象概念的图像。结果显示，两个模型在风格统一性方面表现较好，能够有效地保持风格的一致性。如图10为 SD3 的生成结果，图11为 FLUX.1 的生成结果。同时，它们也能够生成一定程度的抽象概念图像，展现出较强的抽象概念捕捉能力。这表明，DiT 架构在处理抽象概念生成任务时具有一定优势，能够平衡风格的保持与抽象概念的表达。然而，尽管效果较为理想，仍有进一步优化的空间，尤其是在更复杂的抽象概念生成上，可能需要更多的训练或模型调整以提升其表现。



图 10. SD3 的生成结果

Ground Truth	Inference	Ground Truth	Inference
			
			
			

图 11. FLUX.1 的生成结果

我还使用 FLUX.1 模型对那组风格与内容高度交织的数据集进行了训练，该数据集的特殊之处在于它融合了高度抽象的风格，而对于这种抽象风格与图像的内容相互交织，形成了一种复杂的表现形式。在这种情况下，B-LoRA、StyTr2 [1] 和 DEADiff [17] 在进行风格与内容的隐式分离时，都未能有效地处理两者之间的交织关系，如图12。具体来说，风格的抽象特征与内容的语义信息紧密结合，导致 B-LoRA 在迁移过程中未能准确地提取并分离出风格与内容的独立特征。因此，生成的风格迁移结果更多地表现为对颜色的简单迁移，而非对风格和内容精确区分与再现。这表明，在面对这种特殊的风格与内容交织的情况时，现有的隐式分离方法可能存在一定的局限性。





















Content	Style	StyTr2	B-LoRA	DEADiff
				
				
				
				

图 12. 特殊风格迁移的对比实验

同时我使用 FLUX 模型对那组风格与内容高度交织的数据集进行了训练，结果如图13，结果表明生成的图像在一定程度上呈现出抽象化特征。然而，风格迁移的效果并不显著，虽然抽象概念得到了较好的表达，但风格的变化并未达到预期。这可能表明 FLUX 模型在处理这种复杂数据集时，风格迁移和内容分离之间的平衡仍需进一步优化。



图 13. FLUX.1 的特殊风格生成

## 6 总结与展望

文章提出了一种简单而有效的方法，用于分离单个输入图像的风格和内容。这种方法通过两个独立的 B-LoRA 分别对风格和内容进行编码，从而为各种图像风格化任务提供了高度的灵活性，可以独立使用这些编码。与现有主要关注风格提取的方法不同，文章采用了一种复合的风格-内容学习方法，使风格和内容的分离更加清晰，从而提高了风格化的保真度。

尽管本文的工作能够在多种复杂的输入图像上实现稳健的图像风格化，但仍存在一些局限性。首先，在风格-内容分离过程中，对象的颜色通常被包含在风格组件中。然而，在某些情况下，颜色在保持对象身份方面起着关键作用。因此，在仅对内容组件进行风格化时，结果可能无法很好地保留对象的身份，如图14(a)所示。其次，由于 s 仅使用单张参考图像，所学习的风格组件可能会包含背景元素，而不是仅仅聚焦于中心对象，如图14(b)所示。最后，尽管该方法在场景图像的风格化方面表现出色，但在包含大量元素的复杂场景中，可能难以准确捕捉场景结构，从而在一定程度上影响内容保留，如图14(c)所示。

关于未来的研究方向，一个可能的途径是进一步探索 LoRA 微调中的分离技术，以实现更具体的子组件分离，如结构、形状、颜色和纹理等，从而为用户提供更精细的控制能力。另一个未来的研究方向是利用我们方法的稳健性，将其扩展为结合来自多个不同对象的 LoRA 权重，或结合多种风格的能力。





图 14. 方法缺陷

## 参考文献

- [1] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022.
- [2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL <https://arxiv.org/abs/2403.03206>, 2.
- [3] 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 arXiv:2208.01618*, 2022.
- [4] Leon A Gatys. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.
- [5] Bruce Gooch and Amy Gooch. *Non-photorealistic rendering*. AK Peters/CRC Press, 2001.
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [7] 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.
- [8] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024.
- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

- [10] 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.
- [11] Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1085–1094, 2022.
- [12] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019.
- [13] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.
- [14] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.
- [15] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7964–7973, 2024.
- [16] 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 arXiv:2307.01952*, 2023.
- [17] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8693–8702, 2024.
- [18] 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.
- [19] 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*, pages 22500–22510, 2023.

- [20] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. *Low-rank adaptation for fast text-to-image diffusion fine-tuning*, 2023.
- [21] 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, 2025.
- [22] 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.
- [23] Thomas Strothotte and Stefan Schlechtweg. *Non-photorealistic computer graphics: modeling, rendering, and animation*. Morgan Kaufmann, 2002.
- [24] 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.
- [25] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024.
- [26] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023.
- [27] Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A single-pass lora-free model for stylized image generation. *arXiv preprint arXiv:2309.01770*, 2023.
- [28] Hu Ye, Jun Zhang, Sibio 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.
- [29] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- [30] Wei Zhang, Chen Cao, Shifeng Chen, Jianzhuang Liu, and Xiaoou Tang. Style transfer via image component analysis. *IEEE Transactions on multimedia*, 15(7):1594–1601, 2013.
- [31] 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.
- [32] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.