

Mix-of-show论文复现

摘要

本文复现了Mix-of-show论文的实验结果。公共的大规模文本到图像扩散模型，如stable diffusion，已经得到了社会的广泛关注。可以使用低秩适配(low-rank adaptation, LoRA)轻松地为新概念定制这些模型。然而，利用多个概念LoRA来联合支持多个定制概念是一个挑战。Mix-of-show将这种场景称为分散的多概念定制，其中涉及单客户机概念调优和中心节点概念融合。在Mix-of-show中，提出了一个新框架，该框架解决了分散的多概念定制的挑战，包括由现有的单客户端LoRA调优和模型融合过程中的身份丢失导致的概念冲突。Mix-of>Show采用嵌入分解的LoRA (EDLoRA)进行单客户端调优和中心节点的梯度融合，保留了单个概念的域内本质，理论上支持无限概念融合。此外，引入了区域可控采样，它扩展了空间可控采样(如ControlNet和twi - adapter)，以解决多概念采样中的属性绑定和缺失对象问题。复现结果表明，Mix-of>Show能够以高保真度组合多个定制概念，包括角色、对象和场景。在此基础上，我对Mix-of-show的LoRA融合部分进行了改动，提出了一种低代价地、有效地合并独立多个LoRA的方法。在实验结果中可以看到在风格化中有着不错的效果，但是在多人物概念融合任务中效果不是很理想。

关键词: diffusion; LoRA; control

1 引言

深度生成模型的出现极大推动了计算机视觉领域的发展，为图像、视频和三维场景的生成、编辑和重建提供了能力。早期的深度生成模型包括生成式对抗网络（Generative Adversarial Networks, GAN）^[1]、变分自编码器（Variational AutoEncoder, VAE）^[2]等，并在图像生成领域表现出了一定的潜力。在之前，GAN因其高质量的样本生成能力而占据了深度生成模型的统治地位，并不断发展出了一系列能够进行实际运用的模型，例如StyleGAN、StyleGAN2等等，这些模型在生成多样化且高质量的图像上表现出了极大的优势。但受限于GAN的数学原理，早期的深度生成模型普遍具有稳定性较差、训练难度大、容易出现模型崩溃现象等缺点。为了解决上述问题，近年来，在GAN的基础上诞生了扩散模型（Diffusion Models, DMs）^[3]。在训练方式上，DMs相较于GAN有着较大的差别，通过对原始数据进行加噪使其接近高斯分布，再利用神经网络去除图像中的噪声以实现图像生成，另外，神经网络能够学习到原始图像的数据分布，因此能够实现对原始图像数据做小的修改而实现更加多样化的图像生成。这样的训练方式使得由DMs生成的样本具有很强的真实性，当前最先进的图像生成技术也受到了扩散模型的强烈影响，取得了令人惊叹的效果。扩散模型在样本生成的稳定性和多样性上相较于GAN表现出了更明显的优势，打破了GAN在挑战性领域的长期统治地位，在计算机视觉领域中表现出了比早期深度生成模型更高的潜力，因此，DMs也成为了目前为止最先进的深度生成模型。DMs最早可以追溯到由Sohl-Dickstein等人提出的扩散概率模型（Diffusion Probabilistic Model, DPM），但受到当时硬件条件的限制，该方法没能得到广泛的运用。扩散模型涉及两个互关联的过程，分别是前向过程与反向过程。前向过程将数据分布转换为更简单的先验分布，例如高斯分布；相对应的反向过程，利用经过训练的神经网络，通过模拟普通或随机微分方程实现前向过程逆过程。相比于GAN采用生成器（Generator）与判别器（Discriminator）相互对抗的训练方式，DMs的训练过程更加稳定，不易出现模型崩溃的现象。

现如今，由于其强大的图像生成能力，出现了一些开源的文本到图像扩散模型，如Stable diffusion，允许社区用户通过收集个性化概念图像并使用低秩自适应(low-rank adaptation, LoRA)对其进行微调来创建自定义模型。这些定制的LoRA模型通过细致的数据选择、预处理和超参数微调，为特定概念实现了无与伦比的质量。虽然现有的概念LoRA可作为预训练模型的即插即用插件，但

在利用多个概念LoRA扩展预训练模型并实现这些概念的联合组合方面仍然存在挑战。场景作为分散的多概念定制。Mix-of-show对于这个问题提出了一种解决方法。如图1所示，它包括两个步骤：单客户机概念微调和中心节点概念融合。每个客户机在共享微调的LoRA模型的同时保留其私有概念数据。中心节点利用这些概念LoRA来更新预训练的模型，支持对这些定制概念进行联合采样。分散的多概念定制促进了社区在生产高质量概念LoRA方面的最大参与，并提供了重用和组合不同概念LoRA的灵活性。然而，mix-of-show的LoRA微调和权重融合技术无法很好的解决分散的多概念定制的挑战。其中，概念冲突的问题尤其严重。概念冲突的产生是因为当前的LoRA微调方法没有区分嵌入和LoRA权重的作用。

针对这个问题，提出了一种正交方法，旨在减少相似方向和的数量，同时保留原始LoRA的内容和样式生成属性，将产生更高质量的合并。提出的基于优化的方法，找到一组不相交的合并系数来混合两个LoRA。这确保了合并后的LoRA能够熟练地捕获主题和样式。优化过程是轻量级的，并且提高了具有挑战性的LoRA组合的合并性能，其中两个LoRA高度对齐。

2 相关工作

2.1 概念定制

概念定制旨在扩展预训练的扩散模型，以支持仅使用少量图像的个性化概念。有两种主要的概念微调方法：嵌入微调（例如，Textual Inversion^[4]和P+^[5]）和联合嵌入权重调整（例如，Dreambooth^[6]和Custom Diffusion^[7]）。此外，SD社区采用低阶适配器（low-rank adapter, LoRA^[8]）进行概念微调，是轻量级的，可以达到与全权重微调相当的保真度。尽管在单概念定制方面取得了重大进展，但多概念定制仍然是一个挑战。Custom Diffusion提出了对多个概念进行协同训练或对多个现有概念模型进行约束优化。在此之后，SVDiff^[9]引入了数据增强来防止协同训练多概念中的概念混合，cone^[10]发现3个概念神经元可以被添加来支持多个概念。然而，他们的方法通常仅限于融合2-3个语义上不同的概念。相比之下，Mix-of-Show可以在理论上组合无限的定制概念，包括那些在同一语义范畴内的概念。Instantbooth^[11]、ELITE^[12]和Jia^[13]等的研究探索了概念定制的另一个研究方向，重点是实现快速测试时间定制。这些方法包括在特定于所需类别的大规模数据集上预训练编码器。在推理过程中，当从训练的类别中获得一些具有代表性的概念图像时，编码器提取特征来补充预训练的扩散模型并支持自定义生成。然而，这些方法需要为每个类别训练一个单独的编码器，通常限于共同的类别（例如，人或猫）。这种限制阻碍了他们定制和编写更多样化和开放世界主题的能力。

2.2 分散学习

分散或联合学习的目的是在不共享数据的情况下跨不同客户端协作训练模型。联邦学习的实际算法是由FedAvg^[14]提出的。该方法简单地将每个客户模型的权重取平均值以获得最终模型。然而，Mix-of-show发现直接应用这种简单的加权平均对于融合不同概念的LoRA并不理想。为了改进fedavg，以前的工作要么集中在本地客户端培训，要么集中在全局服务器聚合。在此基础上，研究了分散多概念定制中单客户端微调和中心节点融合的优化设计。

2.3 可控多概念生成

仅使用文本提示直接生成多概念面临对象缺失和属性绑定等挑战。之前的方法，如Attend-and-Excite和Structure Diffusion，都试图解决这些问题，但问题仍然存在，限制了多概念生成的有效性。最近的作品，如ControlNet和twi - adapter，引入了空间控制（如keypose和sketch），实现了更精确的构图，解决了多概念生成中缺少对象的问题。然而，属性绑定仍然是一个挑战。在我们的工作中，Mix-of-show中通过区域可控采样来解决这一挑战。

3 本文方法

3.1 本文方法概述

在本节中，将在3.2简要介绍文本到图像扩散模型和概念定制的背景知识。然后，我们在第3.3节中介绍分散的多概念定制的任务制定，然后在第3.4节和第3.5节中详细描述mix-of-show的方法。

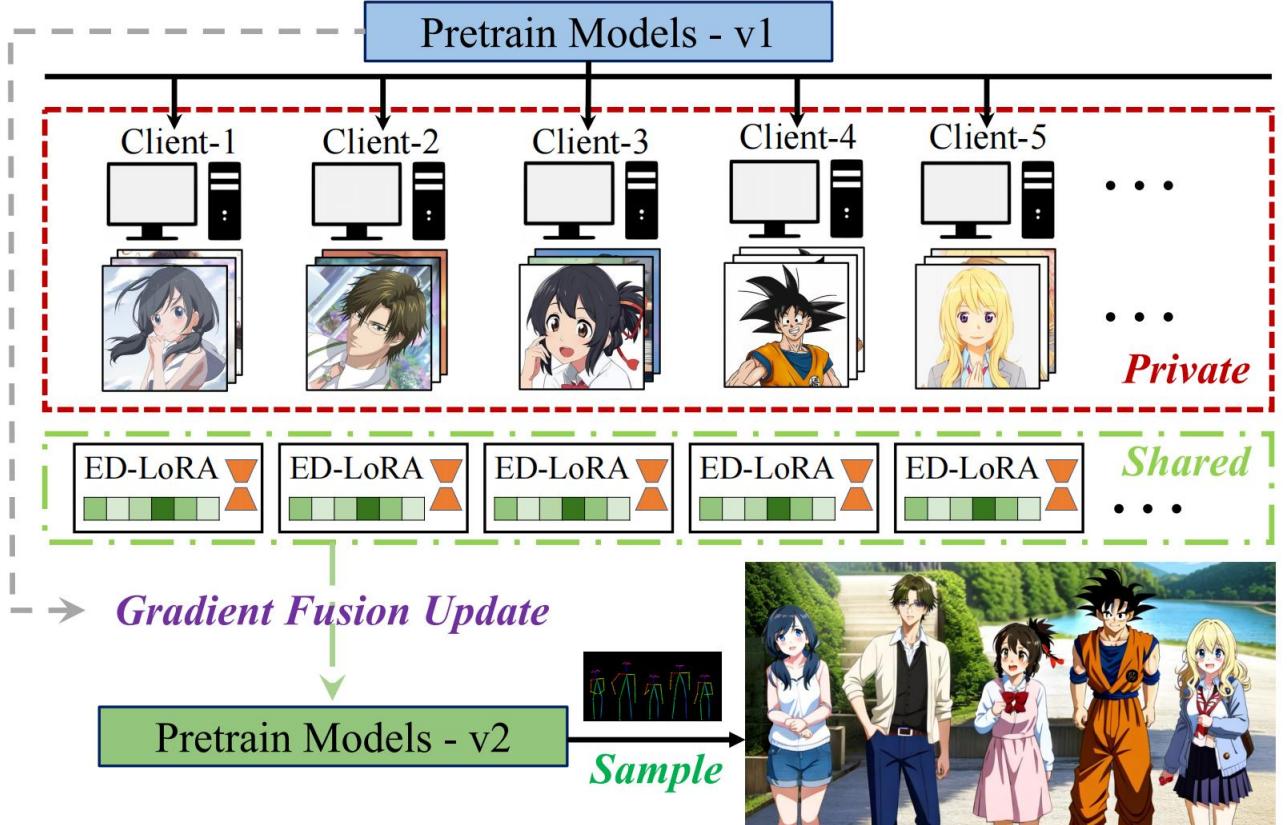


图 1. 通过mix-of-show进行分散的多概念定制的示意图

3.2 Preliminary

文本到图像扩散模型。 扩散模型属于一类生成模型，它在正向扩散过程中逐渐向图像中引入噪音，并学习反向这一过程来合成图像。当与预训练的文本嵌入相结合时，文本到图像的扩散模型能够基于文本提示生成高保真图像。在本文中，我们使用Stable Diffusion进行实验，Stable Diffusion是文本到图像扩散模型在潜在空间中操作的一种变体。给定条件 $c = \Psi(P^*)$ ，其中 P^* 为文本提示符，为预训练的CLIP文本编码器，稳定扩散的训练目标是通过最小化去噪目标

$$\mathcal{L} = \mathbb{E}_{z, c, \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2] \quad (1)$$

概念定制的嵌入微调。 Textual Inversion使用唯一的标记 V 表示输入概念。当提供目标概念的少量图像时，使用公式1对 V 的嵌入进行调整。微调后， V 的嵌入像预训练模型中的任何其他文本一样编码目标概念和功能的本质。为了实现更好的解纠缠和控制， P^+ 引入了概念令牌的分层嵌入，在本文中表示为 V^+ 。

低秩适应。 低秩自适应(Low-rank adaptation, LoRA)最初是为了使大型语言模型适应下游任务而提出的。它假设自适应过程中的权重变化具有较低的“内在秩”，并对权重变化进行低秩分解，得到更新后的权重 W ，即 $W = W_0 + \Delta W = W_0 + BA$ 。其中， $W_0 \in \mathbb{R}^{d \times k}$ 代表预训练模型中的原始权重， $B \in \mathbb{R}^{d \times r}$ 和 $A \in \mathbb{R}^{r \times k}$ 代表低秩因子， $r \ll \min(d, k)$ 。最近，社区采用LoRA对扩散模型进行

微调，取得了很好的结果。LoRA通常用作预训练模型中的即插即用插件，但社区也使用权重融合技术来组合多个LoRA：

$$W = W_0 + \sum_{i=1}^n w_i \Delta W_i, \quad \text{s. t. } \sum_{i=1}^n w_i = 1, \quad (2)$$

3.3 任务制定：分散的多概念定制

虽然自定义扩散试图将两个调整过的概念模型合并为一个预训练模型，但他们的研究结果表明，与多个概念共同训练可以产生更好的结果。然而，考虑到可伸缩性和可重用性，专注于合并单概念模型以支持多概念定制。我们将这种设置称为分散的多概念定制。形式上，分散的多概念定制涉及两步过程：单客户机概念微调和中心节点概念融合。如图1所示，n个客户端都拥有自己的私有概念数据，并对概念模型 ΔW_i 进行微调。这里， ΔW_i 表示网络权值的变化，特指LoRA权值。省略了讨论文本嵌入的合并，因为调整后的嵌入可以无缝地集成到预训练的模型中，而不会产生冲突。

微调后，中心节点对所有LoRA进行聚类，得到更新后的预训练权值W：

$$W = f(W_0, \Delta W_1, \Delta W_2, \dots, \Delta W_n) \quad (3)$$

其中表示对原始预训练模型权重 W_0 和n个概念LoRAs $\{\Delta W_i, i = 1 \dots n\}$ 进行操作的更新规则。一个简单的更新规则是权重融合，如公式3所示。一旦更新，新模型W应该能够生成n个LoRA中引入的所有概念。

3.4 Mix-of-show

在本节中，我们将介绍Mix-of>Show，其中包含用于单客户端概念微调的ED-LoRA和用于中心节点概念融合的梯度融合

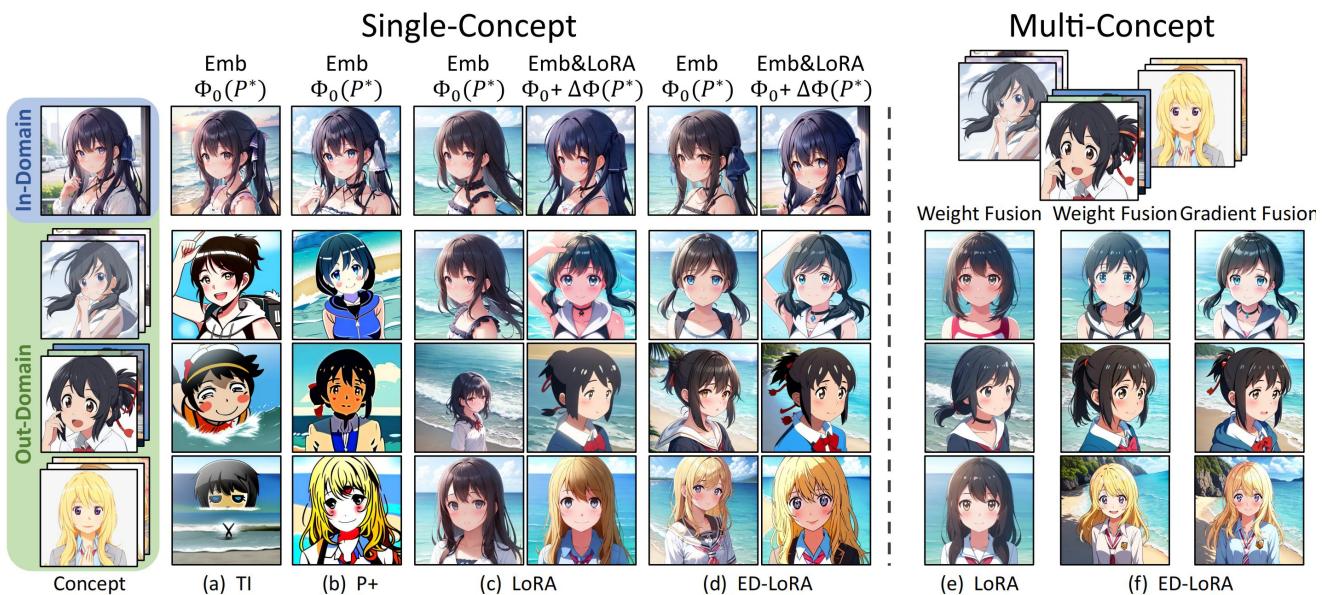


图 2. 嵌入微调(即文本反演(TI)和P+)和联合嵌入权微调(即LoRA和我们的ED-LoRA)之间的单概念和多概念定制。 P^* = “一个V的照片，在海滩附近”。 Φ_0 和 $\Delta\Phi$ 为预训练模型和LoRA权值。

单客户机概念微调:ED-LoRA: Vanilla LoRA由于概念冲突的问题，不适合去中心化的多概念定制。为了更好地理解这种限制，我们首先研究嵌入和LoRA权重在概念微调中的不同作用。

单概念调谐设置。我们研究嵌入微调(即 Textual Inversion)和P+以及基于单个概念定制的

联合嵌入权微调(即LoRA)。我们对域内概念(即直接从预训练模型中采样)和域外概念进行了实验。预训练模型的权重，包括unet θ 和文本编码器 ψ ，表示为 $\Phi_0 = \{\theta_0, \psi_0\}$ 。给定一个包含概念V的文本提示P*，我们使用预训练的权重 $\Phi_0(P^*)$ 可视化概念V的调谐嵌入，并使用 $(\Phi_0 + \Delta\Phi)(P^*)$ 可视化调谐嵌入以及LoRA权重。

根据图2的实验结果，作者对现有的嵌入调谐和联合嵌入权调谐方法得出以下两点观察结果。观察1：嵌入能够捕获预训练模型域中的概念，而LoRA有助于捕获域外信息。在图2(a, b)中，我们观察到嵌入微调方法(如 Textual Inversion和P+)难以捕获域外概念。这是因为它们试图在嵌入中编码所有域外细节(例如，动画风格或未由预训练模型 Φ_0 建模的细节)，导致语义崩溃。然而，对于从模型中采样的域内概念，嵌入微调准确地编码了嵌入中的概念身份，受益于通过预训练的模型权重 Φ_0 对概念细节的准确建模。此外，当与LoRA联合调整嵌入时，嵌入不再产生过饱和输出。这是因为外域信息是通过LoRA权移(即 $(\Phi_0 + \Delta\Phi)$)的预训练模型捕获的。

观察2：现有的LoRA权重对大多数概念标识进行编码，并将语义相似的嵌入投影到视觉上不同的概念上，导致概念融合过程中的冲突。在图2(c)所示的联合嵌入-LoRA微调结果中，我们观察到使用预训练模型 $\Phi_0(P^*)$ 直接可视化嵌入会产生语义上相似的结果。但是，当加载LoRA权重($\Phi_0 + \Delta\Phi(P^*)$)时，可以准确捕获目标概念。这表明大多数概念标识是在LoRA权重中编码的，而不是嵌入本身。然而，当试图在单个模型中支持多个语义相似的概念时，根据相似的嵌入确定要对哪个概念进行采样就会出现问题，从而导致概念冲突。如图2(e)所示，当融合到一个模型中时，每个单独概念的身份就丢失了。

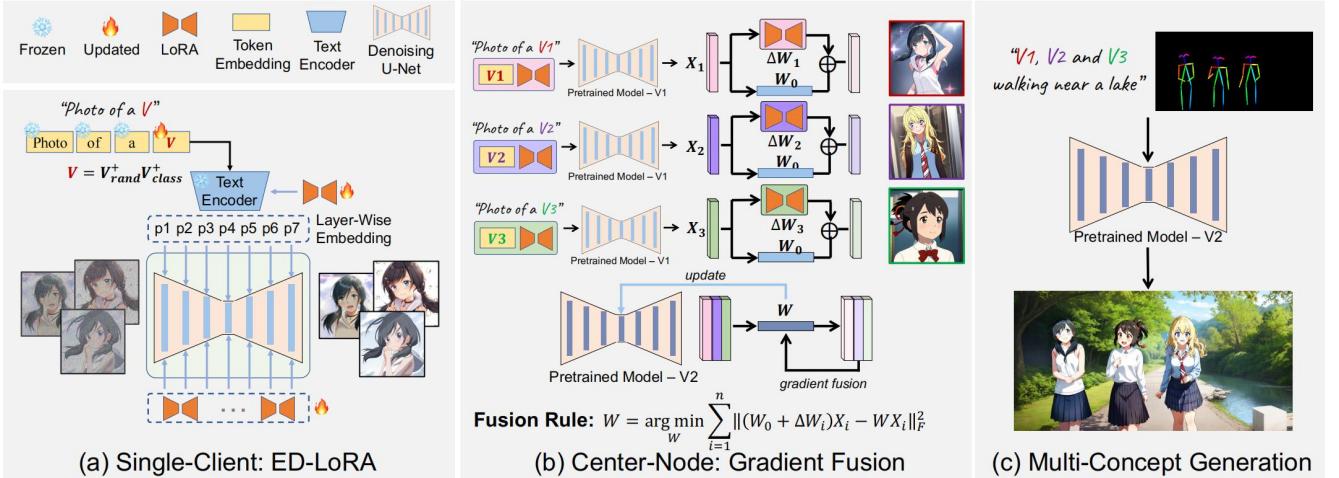


图 3. 混合显示的pipeline。在单客户端概念微调中，ED-LoRA采用分层嵌入和多词表示。在中心节点，采用梯度融合将多个概念LoRA融合在一起，支持自定义概念的组合。

Mix-of-show的解决方案:ED-LoRA。基于上述观察，我们的ED-LoRA被设计成在使用LoRA权重捕获剩余细节的同时，在嵌入中保留更多的域内本质。为此，我们通过分解嵌入来增强嵌入的表达性。如图3所示，我们采用分层嵌入。并为概念令牌($V = V_{\text{rand}}^+ V_{\text{class}}^+$)创建一个多世界表示 V_{rand}^+ 。在这里， V_{rand}^+ 随机初始化以捕获不同概念的方差，而 V_{class}^+ 初始化基于语义类保持语义意义。这两个符号在概念微调期间都是可学习的。如图2(d)所示，ED-LoRA的学习嵌入有效地保留了预训练模型域内给定概念的本质，而LoRA则有助于捕获其他细节。

区域可控采样：直接多概念采样经常会遇到对象缺失和属性绑定的挑战。虽然空间可控采样方法(如ControlNet和T2IAdapter)可以解决多概念生成中缺失对象的问题，但它们不能准确地将概念绑定到特定的键位或草图。仅仅表示所期望的概念和属性通过文本提示可能导致属性绑定问题，如图4(a)所示，其中三个人的身份混合，并且“红裙子”被错误地分配给其他概念。

为了解决这些挑战，mix-of-show提出了一种称为区域可控采样的方法。该方法利用全局提示

和多个区域提示来描述基于空间条件的图像。全局提示提供整体上下文，而区域提示指定特定区域内的主题，包括它们的属性和来自全局提示的上下文信息(例如，“靠近湖泊”)。为了实现这一点，我们引入了区域感知交叉注意。给定一个全局提示 P 和 n 个区域提示 P_g^* ，我们首先通过交叉注意合并全局提示 $h = \text{softmax}\left(\frac{Q(z)K(P_g^*)}{\sqrt{d}}\right) \cdot V(P_g^*)$ 。然后，我们通过 $z_i = z \odot M_i$ 提取区域潜在特征，其中 M_i 表示与 $P_{r_i}^*$ 指定的区域相关联的二进制掩码。我们使用 $h_i = \text{softmax}\left(\frac{Q(z_i)K(P_{r_i}^*)}{\sqrt{d}}\right) \cdot V(P_{r_i}^*)$ 获得区域特征。最后，我们将全局输出中的特征替换为区域特征： $h[M_i] = h_i$ 。如图4(b)所示，区域可控采样允许对主题和属性进行精确分配，同时保持和谐的全局环境。

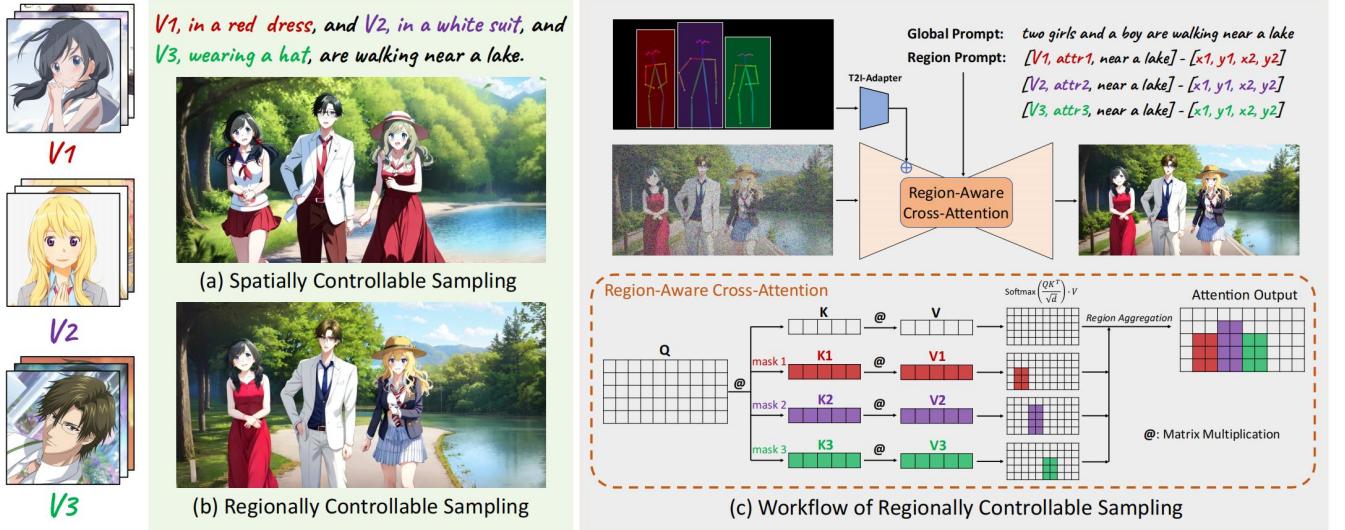


图4:多概念生成的区域可控采样。

4 复现细节

4.1 与已有开源代码对比

我使用了Mix-of-show的开源代码，修改了其中gradient_fusion.py中的代码。具体来说我修改了其中zip-LoRA融合部分的代码，新设定了两个可学习的参数，分别与原zipLoRA矩阵相乘，实现两个矩阵之间正交，之后再按照原方法进行融合。

4.2 实验环境搭建

编程语言: python

显卡: P100

操作系统: Linux

Cuda: 11. 3

Python: 3. 10

Pytorch: 1. 12. 1

预训练模型: 现实风格使用ChilloutMix，动漫风格使用Anything-v4

数据集: 使用Mix-of-show提供的包括真人的3个数据集，动漫的2个人物数据集和风格的两个数据集，其中动漫的是maao和maidane，真人的是Harry_Potter、Hermione_Granger、Thanos，风

格化的是sketch_style和felt_style。

4.3 界面分析与使用说明

单客户机概念微调：

步骤一：修改配置

在微调之前，必须在相应的配置文件中指定数据路径并调整某些超参数。以下是一些需要修改的基本配置设置。

```
datasets:  
  train:  
    # Concept data config  
    concept_list: datasets/data_cfgs/edlora/single-concept/characters/anime/hina_amano.json  
    replace_mapping:  
      <TOK>: <hina1> <hina2> # concept new token  
  val_vis:  
    # Validation prompt for visualization during tuning  
    prompts: datasets/validation_prompts/single-concept/characters/test_girl.txt  
    replace_mapping:  
      <TOK>: <hina1> <hina2> # Concept new token  
  
models:  
  enable_edlora: true # true means ED-LoRA, false means vallina LoRA  
  new_concept_token: <hina1>+<hina2> # Concept new token, use "+" to connect  
  initializer_token: <rand-0.013>+girl  
  # Init token, only need to revise the later one based on the semantic category of given concept  
  
val:  
  val_during_save: true # When saving checkpoint, visualize sample results.  
  compose_visualize: true # Compose all samples into a large grid figure for visualization
```

步骤二：开始微调

我使用2个P100的GPU调整每个概念。与LoRA类似，社区用户可以在一个GPU上启用梯度积累，xformer，梯度检查点。

步骤三：采样

中心节点概念融合

步骤一：收集概念模型

收集所有需要扩展预训练模型的概念模型，并相应地修改
datasets/data_cfgs/MixofShow/multi-concept/real/*中的配置。

```

[
  {
    "lora_path": "experiments/EDLoRA_Models/Base_Chilloutmix/characters/edlora_potter.pth", # ED-Lo
    "unet_alpha": 1.0, # usually use full identity = 1.0
    "text_encoder_alpha": 1.0, # usually use full identity = 1.0
    "concept_name": "<potter1> <potter2>" # new concept token
  },
  {
    "lora_path": "experiments/EDLoRA_Models/Base_Chilloutmix/characters/edlora_hermione.pth",
    "unet_alpha": 1.0,
    "text_encoder_alpha": 1.0,
    "concept_name": "<hermione1> <hermione2>"
  },
  ...
  # keep adding new concepts for extending the pretrained models
]

```

步骤二：概念融合

步骤三：采样

4.4 创新点

LoRA更新矩阵是稀疏的，不同LoRA层的更新矩阵 ΔW 是稀疏的，即， ΔW 中的大多数元素的幅度非常接近于零，因此对微调模型的输出影响很小。对于每一层，可以根据它们的大小对所有元素进行排序，并将最低的元素归零，直到某个百分位数。并且，高度对齐的LoRA权重合并较差。两个独立训练的LoRA的权值矩阵的列可能包含未解纠缠的信息，即它们之间的余弦相似度可以不为零。LoRA权重列之间的对齐程度在决定最终合并的质量方面起着重要作用：如果直接将彼此具有非零余弦相似性的列相加，则会导致它们关于多个概念的信息的叠加，导致合并模型失去准确综合输入概念的能力。而当列彼此正交且余弦相似度等于零时，可以避免这种信息损失。

为了防止合并过程中的信号干扰，我用一个可学习的系数乘以每列，这样列之间的正交性就可以实现。LoRA更新是稀疏的这一事实允许忽略每个LoRA中的某些列，从而促进最小化干扰的任务。该方法有两个目标：(1)最小化内容两个LoRA之间的干扰，由两个LoRA之间的余弦相似度定义；(2)通过最小化混合LoRA生成的图像与原始图像之间的差异，保留合并LoRA独立生成各自主题的能力。

5 实验结果分析

单概念微调：ED-LoRA复现结果



图5:提示文本为“a potter in front of eiffel tower”，potter概念ED-LoRA的复现结果



图6: 原文的实验结果

从图5和图6的实验结果中可以看到，当脸部很小时，面部生成结果会扭曲。推测为由于面部特征细节很多，如果生成区域分辨率过小，则会导致无法区分面部细节，以至于崩溃。



a hermione sit in front of mount fuj



a hermione near the sea



a hermione sit near the sea



图7:将Harry_Potter、Hermione_Granger、Thanos三个概念融合后，从中提取出单概念复现的实验结果

从图7复现结果中可以看到，即使将三个概念进行了融合，从中提出出单个概念的特征保留能力是非常强的，没有出现明显的特征丢失或特征融合的现象。

多概念区域控制复现结果



图8:主文本提示为“three people near the sea”，三个区域的文本提示分别为“a potter, in Hogwarts uniform, holding hands, near the sea”、“a hermione, girl, in Hogwarts uniform, near the sea”、“a thanos, purple armor, near the sea”，复现结果



图9:原文实验结果

从图8和图9实验结果中可以看到使用文中提出的区域控制的方法，可以对多个概念进行较好的区分，但是在一些细节部分还是出现了特征融合、特征丢失的情况。

使用修改后的代码进行风格化的实验结果

原图



单个lora



融合lora





图10:使用正交LoRA进行风格化的生成结果

从图10实验结果中可以看到，使用了正交LoRA进行融合后生成特定风格的图片效果并不是非常理想，部分风格化的特征在融合过程中还是消失了。



图11:使用正交LoRA进行多人物概念融合的生成结果

从图11实验结果中可以看到，使用了正交LoRA进行人物概念融合后，生成其中单一概念人物时会出现糟糕的结果，无法保持原来人物的特征。推测是由于人物概念之间有许多相似的特征，如果使用参数与LoRA矩阵相乘强行来使两个人物概念LoRA正交，会改变其中一些关键特征，从而无法保持原来人物概念的完整性。

6 总结与展望

在Mix-of-show中，作者探索了分散的多概念定制，并强调了现有方法（如LoRA调优和权重融合）的局限性，这些方法在这种情况下会受到概念冲突和身份丢失的影响。为了克服这些挑战，作者提出了Mix-of-Show框架，该框架结合了用于单客户端概念调优的ED-LoRA和用于中心节点概念融合的梯度融合。ED-LoRA在嵌入中保留了个体概念的本质，避免了冲突，而梯度融合则最大限度地减少了概念融合过程中的身份损失。在多概念生成中引入区域可控采样来处理属性绑定。复现结果也证明Mix-of-Show可以成功地生成人物的多个定制概念的复杂组合。在此基础上，我对Mix-of-show的概念融合部分进行了改动，使用一种无缝合并独立训练LoRA的新方法，通过利用关于预训练LoRA权重的理解，我尝试了使LoRA间正交来解决特征融合的问题。但目前的尝试虽然在风格化上有不错的结果，但是在人物概念融合上效果并不是很理想。

在未来的工作中，一方面，我将进一步挖掘LoRA特征融合的底层原理，尝试对人物概念进行细分。另一方面我将尝试在空间层面对生成图像进行控制，进一步将不同LoRA概念控制在目标生成区域，在保证生成图像整体一致的前提下，让各自的区域间特征不互相干扰。

参考文献

- [1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
- [2] Kingma D P, Welling M. An introduction to variational autoencoders[J]. Foundations and Trends® in Machine Learning, 2019, 12(4): 307-392.
- [3] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.
- [4] Gal R, Alaluf Y, Atzmon Y, et al. An image is worth one word: Personalizing text-to-image generation using textual inversion[J]. arXiv preprint arXiv:2208.01618, 2022.
- [5] Voynov A, Chu Q, Cohen-Or D, et al. P+: Extended Textual Conditioning in Text-to-Image Generation[J]. arXiv preprint arXiv:2303.09522, 2023.
- [6] Ruiz N, Li Y, Jampani V, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 22500-22510.
- [7] Kumari N, Zhang B, Zhang R, et al. Multi-concept customization of text-to-image diffusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 1931-1941.
- [8] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.
- [9] Han L, Li Y, Zhang H, et al. Svdiff: Compact parameter space for diffusion fine-tuning[J]. arXiv preprint arXiv:2303.11305, 2023.
- [10] Liu Z, Feng R, Zhu K, et al. Cones: Concept neurons in diffusion models for customized generation[J]. arXiv preprint arXiv:2303.05125, 2023.

- [11] Shi J, Xiong W, Lin Z, et al. Instantbooth: Personalized text-to-image generation without test-time finetuning[J]. arXiv preprint arXiv:2304.03411, 2023.
- [12] Wei Y, Zhang Y, Ji Z, et al. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation[J]. arXiv preprint arXiv:2302.13848, 2023.
- [13] Jia X, Zhao Y, Chan K C K, et al. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models[J]. arXiv preprint arXiv:2304.02642, 2023.
- [14] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Artificial intelligence and statistics. PMLR, 2017: 1273-1282.