

DreamBooth: 微调文本到图像扩散模型，实现主体驱动生成

摘要

本研究聚焦于 DreamBooth 方法的实验复现，旨在探索其在特定主体图像生成中的应用。通过一系列实验复现并扩展了基于 stable-diffusion-v1-4 生成模型的功能，聚焦于特定对象的图像生成、语义信息融合、艺术风格迁移以及新视角合成四个方面。研究发现该方法能使模型在少量样本下生成高保真且多样化的主体图像，但也存在对复杂场景生成不够精准、主体特征稳定性有差异和语义理解边界等局限性。实验结果表明，基于该生成模型的方法在各类图像生成任务中表现出了较好的效果，尤其是在语义控制和风格迁移上具有较强的适应性。本复现也为深入理解 DreamBooth 技术提供了实践依据，对其在图像生成领域的进一步应用具有参考价值。

关键词：文本到图像生成；模型微调；图像合成

1 引言

近些年来，大型文本到图像模型如 stable-diffusion [1]，由于其训练过程中从大量图像标题对中学习到的强语义先验，具有很强大的文生图能力。例如，这样的先验知识就会把将“狗”这个词汇与不同姿势和上下文出现在图像中的各种狗的实例绑定在一起，为文生图技术的发展为创意表达提供了新的途径。但是现有模型缺乏模拟重现特定主体外观的能力，也缺乏在不同背景下合成特定主体的表现能力。而 DreamBooth 的出现为解决这一问题带来了新的思路，它通过独特的微调策略，使模型能够学习并绑定特定主体与标识符，从而在不同场景下生成高保真的主体图像。本实验复现旨在深入探索 DreamBooth 在实际应用中的表现，通过对特定主体的实验操作，验证其在多种任务中的有效性，并分析其优势与局限。

2 相关工作

此部分对课题内容相关的工作进行简要的分类概括与描述，二级标题中的内容为示意，可按照行文内容进行增删与更改，若二级标题无法对描述内容进行概括，可自行增加三级标题，后面内容同样如此，引文的 bib 文件统一粘贴到 `refs.bib` 中并采用如下引用方式。

2.1 图像合成技术

传统图像合成技术在融合主体与新背景时面临诸多挑战，如场景融合效果不自然、对新颖姿势处理能力有限等，难以实现复杂场景下的高质量合成 [2]。

2.2 文本到图像编辑与合成

基于 GANs 与 CLIP 结合的方法在结构化场景表现尚可，但处理多样化主体时存在不足，虽有改进尝试，但仍受限制；而扩散模型虽在生成质量上有进步，但多数文本驱动方法在主体特定情境生成方面有待提升，全局编辑难以满足精细控制需求。最近的大型文本到图像模型，如 Imagen [3]、DALL-E2 [4]、Parti [5] 和 CogView2 [6] 展示了前所未有的语义生成。这些模型不提供对生成图像的细粒度控制，并且仅使用文本指导。具体来说，在合成图像中一致地保留主体的身份是极具挑战性的。

2.3 可控生成模型

当前在可控生成模型方面，已有多种方法探索。一些扩散模型技术在引导图像变化方面有进展，但在主体身份保持和新颖样本生成上存在改进空间 [7]；而有些方法基于掩码的方法在限制特定区域的同时，难以实现主体身份精确保持和多样化生成。本研究的微调方法使我们能够将主题嵌入到模型的输出域中，从而生成主题的新颖图像，并保留其关键视觉特征。

3 本文方法

3.1 总体目标

利用 DreamBooth 方法，借助少量特定主体图像，实现主体在不同场景、属性、艺术风格和视角下的多样化生成，同时确保主体关键特征得以保留。

3.2 个性化主体设置

3.2.1 微调策略

运用预训练的文本到图像扩散模型，通过对高斯噪声的逐步去噪学习数据分布。模型依据输入噪声和文本提示生成的条件向量生成图像，并使用平方误差损失函数优化模型参数，以保证生成图像与真实图像的一致性。采用特定主体（如狗、背包、茶壶）的少量图像微调模型，发现模型在上述损失函数合理的设置下可有效整合新信息，避免过度拟合，且能准确表示主体。

3.2.2 提示设计

将输入主体图像与 prompt：“a [标识符] [类名词]”形式绑定，其中，[标识符] 是与主体紧密相关的独特标识，通过在词汇表中查找罕见 token 并反转到文本空间的方式确定，有效降低了其先验概率，使模型能够更好地识别和区分特定主体。[类名词]（如“dog”等）则为主

体提供了一个粗略的类别描述，这样的设计使得模型能够利用其对类别先验知识的理解，与主体独特标识符的嵌入相结合，从而更好地生成多样化且符合逻辑的主体图像。

3.3 类别特定先验保持损失

原文作者发现，微调模型的所有层可以获得最大主题保真度的最佳结果，但同时也引发了两个关键问题。一是语言漂移现象，即模型在预训练时学习到的关于语言的句法和语义知识，在针对特定主体的微调过程中逐渐丧失。这表现为模型在生成与目标主体同类的其他主体时，能力下降；二是输出多样性降低的问题，文本到图像扩散模型原本具有较高的输出多样性，能够生成各种不同姿势、视角和风格的图像。然而，在对少量特定主体图像进行微调时，容易对训练图像的过拟合，生成图像失去多样性。

为了解决上述问题，提出了类别特定先验保持损失。其核心思想是通过利用模型自身生成的样本来监督模型的训练过程，从而使模型在微调开始后能够保留对类别先验的理解。具体实现方式是，首先使用冻结预训练扩散模型的采样器，从随机噪声 $z_{t1} \sim \mathcal{N}(0, I)$ 与文本条件 $c_{\text{pr}} := \Gamma(f(\text{"a [class noun]"}))$ 生成一些原始样本 $x_{\text{pr}} = \hat{x}(z_{t1}, c_{\text{pr}})$ 然后将生成的数据纳入损失函数中：

$$\mathbb{E}_{x, c, \epsilon, \epsilon', t} \left[w_t \|\hat{x}_\theta(\alpha_t x + \sigma_t \epsilon, c) - x\|_2^2 + \lambda w_{t'} \|\hat{x}_\theta(\alpha_{t'} x_{\text{pr}} + \sigma_{t'} \epsilon', c_{\text{pr}}) - x_{\text{pr}}\|_2^2 \right] \quad (1)$$

第二项是先验保留项，它用自己生成的图像来监督模型，图 1 阐述了这个流程是怎么使用类生成的样本和先验保留损失进行的模型微调。

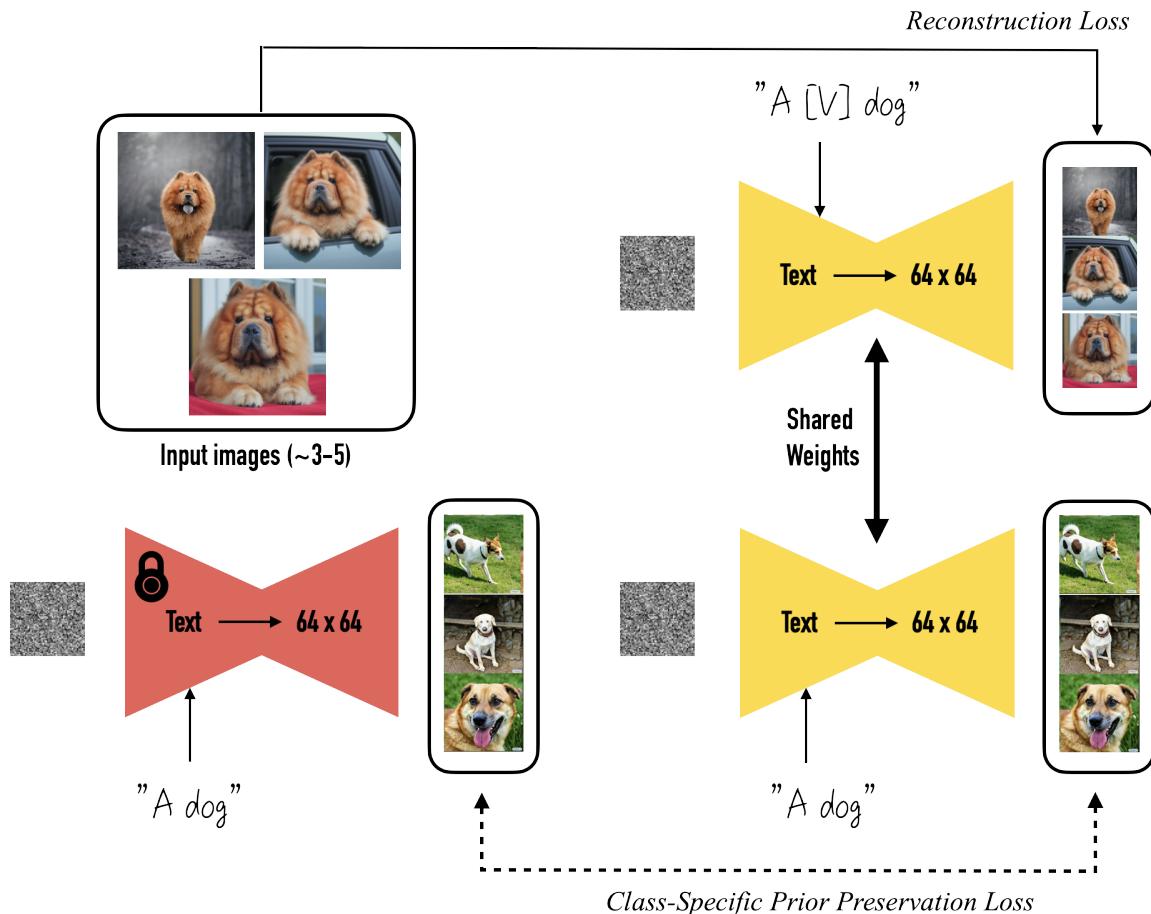


图 1. 方法示意图

图 1 中左上角的 input images 作为监督信号来绑定到这个“V”上，通过重构损失来实现；而图片下方则是先验保留损失，通过冻结前后的与训练模型的生成图片对比来确保模型不会发生语言偏移。

4 复现细节

4.1 与已有开源代码对比

此部分为必填内容。如果没有参考任何相关源代码，请在此明确申明。如果复现过程中引用参考了任何其他人发布的代码，请列出所有引用代码并详细描述使用情况。同时应在此部分突出你自己的工作，包括创新增量、显著改进或者新功能等，应该有足够差异和优势来证明你的工作量与技术贡献。在本实验复现过程中，参考了 <https://github.com/XavierXiao/Dreambooth-Stable-Diffusion> 中的部分代码实现。该开源代码为理解 DreamBooth 的基本框架和流程提供了重要基础，我主要借鉴采用了其模型架构搭建、基础训练循环以及部分数据处理模块的代码逻辑。对于开源代码主体的复现工作主要是 DreamBooth 方法可以对提供的特定主题实例进行大量文本引导的语义修改，包括重新情境化、主题属性（例如材料和物种）的修改、艺术风格迁移和视角切换。重要的是，在所有这些修改中，该方法能够保留特定主体身份和本质的独特视觉特征。像在重新情境化任务中，主体的特征不会被修改，但例如姿势等可能会改变。而如果任务是更强的语义修改，例如我们的特定主体和另一个对象之间的特征交叉融合，那么只会保留特定主体的关键特征。然而，我在实验中进行了多项创新和扩展工作，具体内容显著区别于已有开源代码的实现

4.1.1 主要任务扩展

原文中的主要实验是基于的数据集包括 30 个 subjects 和 25 个提示词 prompts，总共 750 个组合，然后去做重新语境化任务。因此，完全复现会是一个巨大的工作量，也会耗费大量计算资源，因此，在重语境这个任务上，只选择 12 个组合进行复现，故后续的性能指标会和原文结果有些许差异，但差距也在正常范围内。本实验使用了 3 个 subject (dog、backpack、teapot) 以及 4 个个性化的 prompt 进行组合。在重语境这个任务上的部分复现结果如图 2 所示，从直观的视觉效果上可以看出结果微调后的生成图片处理能够保持 input subject 的身份和细节，并且可以借助预训练 sd v1-4 强大的上下文语义来进行不同场景的主体重现，得到 input images 中没有的场景结构，以及主体在场景（例如接触、阴影、反射等）中的现实整合。

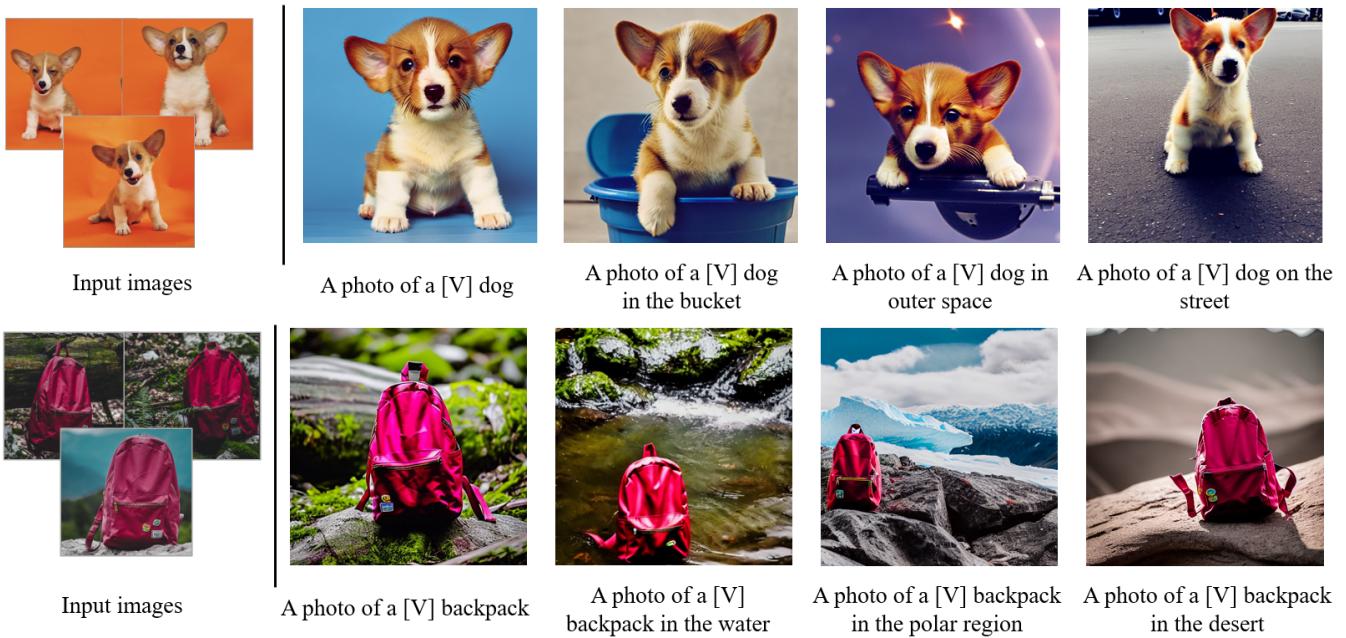


图 2. 重新语境化

对标原文实验，我在这个任务上也做了量化指标的计算，原文中主体保真性的定量分析使用三个指标：DINO、CLIP-I、CLIP-T，其中，DINO 和 CLIP-I 指标属于 subject fidelity 主体保真度的范围，也即是用于评估特定物体的生成是否相似，而 CLIP-T 指标属于 prompt fidelity 保真度的范围，也即用于评估生成图片和输入的文本之间特征的相似度。原文中与 Textual Inversion 的对比实验中，还比较了基于 Imagen 和 SD 的 DreamBooth，通过以上三个指标全面衡量 DreamBooth 的性能，由于 Imagen 并未开源，所以本实验的复现是基于 Stable Diffusion，相关数据展示在表 1 中，其中，method 中的 Real Images、DreamBooth (Imagen) 和 Textual Inversion (Stable Diffusion) 是原论文中数据，而 DreamBooth (Stable Diffusion) 是本人基于 12 个组合，每一个组合 8 张图片，分别在以上三个指标上重新计算的结果，而原文在这一个 method 中的实验结果分别是：0.668、0.803、0.305。

表 1. 主体保真度 (DINO、CLIP-I) 和提示保真度 (CLIP-T、CLIP-T-L) 定量指标比较

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Real Images	0.774	0.885	N/A
DreamBooth (Imagen)	0.696	0.812	0.306
DreamBooth (Stable Diffusion)	0.787	0.859	0.304
Textual Inversion(Stable Diffusion)	0.569	0.780	0.255

此外，我对原文中的艺术风格迁移、新视角的合成以及属性的修改做了些扩展任务。首先是艺术风格迁移，旨在将输入图像的内容与特定的艺术风格进行融合，从而生成具有独特艺术魅力的新图像。不同于文章中的艺术风格迁移，原文中做了梵高、米开朗基罗以及维米尔的画作风格迁移，我保留了梵高的艺术风格作为比较，再对塞尚的风格进行迁移尝试，也引入更多世界各地的地域特色艺术风格，比如中国的山水墨画风格、日本的浮世绘风格进行迁移尝试，取得不错的效果，如图 3所示。



图 3. 艺术风格迁移

DreamBooth 的微调方法，使生成的图像不仅在风格上独具特色，还能精准地保持主体的本质特征，实现内容与风格的完美和谐统一。

其次是新的视角合成，即从已知的图像视角中推断并生成主体在全新视角下的图像表现。原文中尝试了从 input 的几张只有正面视角去尝试推断俯视、仰视视角的图像表现，在本实验中，还扩展了从侧身角度、背身角度，这包括对主体在不同视角下的几何形状变化、光照效果的动态调整、遮挡关系的合理处理以及背景环境的自然融合等多方面的研究，实验结果如图 4 所示，实验发现模型能够生成更加逼真、自然且逻辑连贯的新视角图像，就如同真实拍摄主体从不同角度观察所得到的效果一样，从而为虚拟现实、增强现实、产品展示等多个领域提供更强大的图像生成支持。

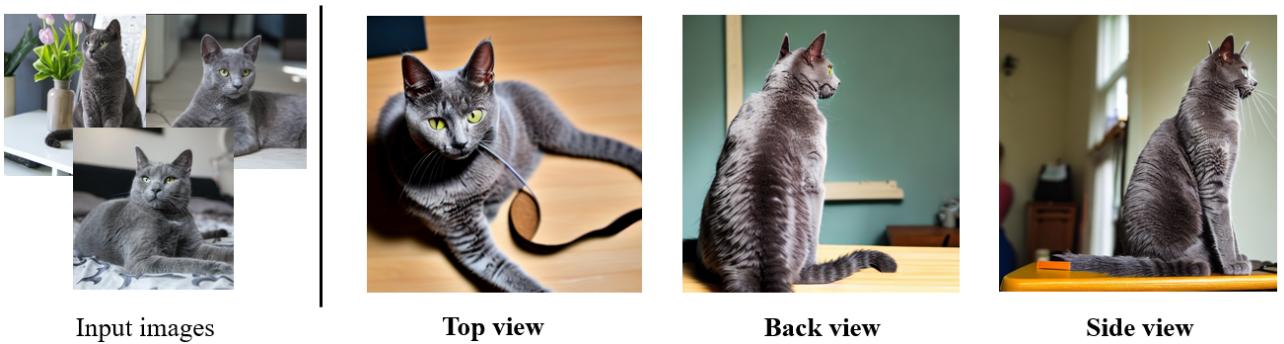


图 4. 新视角合成

最后是关于属性的修改，原文实验做了茶壶的材质修改以及微调的主体与其他物种进行融合的实验，本实验在这方面，显示扩展了物种融合的品种，包括狼、考拉，然后还基于微调的图像进行情感融合，将不同的情感赋予同一个特定的主体，研究生成图片的情感表现是否与所给的 prompt 一致以及是否能能表现出该情感，实验结果如图 5 所示。



图 5. 属性修改

4.1.2 多主体绑定策略对比

深入研究了在同一模型中使用同一个标识符 (V) 绑定多个主体 (subject) 和一个模型中不同 V 绑定不同 subject 的效果差异。结果发现，在使用同一个 V 绑定多个 subject 时，发现模型在生成过程中出现了生成崩溃的现象。例如，当试图用同一个 V 同时表示一只狗和一个背包时，生成的图像在特征融合和区分上出现混乱，如图 6 所示，使用的都是同一个 prompt：“A photo of a sks backpack”，但是由于模型第一次训练时这个 [V] (sks) 与我们 input 的几张小狗图像绑定，进行二次训练重新绑定后，不仅无法准确呈现任何一个主体的特征，导致生成的图像质量严重下降，还方向背景保留了在第一次训练时 input 的几张小狗图片的背景，因为正常情况下像图 5 中的 Normal，重新生成后不加特定的限制，有大概率是会出现与微调图片背景相似的生成结果的。随后再做了对比实验，采用一个模型中不同 V 绑定不同 subject 的策略时，模型能够稳定地生成不同主体在各种场景下的高质量图像，主体特征保持清晰且准确，不同主体之间没有出现混淆或特征丢失的情况。



图 6. 微调流程

而采用一个模型中不同 V 绑定不同 subject 的策略时，模型能够稳定地生成不同主体在各种场景下的高质量图像，主体特征保持清晰且准确，不同主体之间没有出现混淆或特征丢

失的情况。

4.2 实验环境搭建

首先是配置环境，根据所给的 environment.yaml 进行 ldm 环境的配置。然后关于实验的主体扩散模型的选择，要微调稳定扩散模型，需要提前下载好预训练的稳定扩散模型。可以在 HuggingFace 上下载。自行决定使用哪个版本的 checkpoint，但本次实验使用的是 sd-v1-4-full-ema.ckpt。

4.3 界面分析与使用说明

首先是正则化图片的生成，在每一次训练前，还需要创建一组用于正则化的图像，因为 Dreambooth 的微调算法需要这样做。该算法的细节已在第三节解释过。生成正则化图像的文本提示可以是 `<class>` 的照片，其中 `<class>` 是描述对象类的单词，例如 dog。执行指令使用 stable-diffusion-v1-4 生成正则化图像后，保存在特定路径下。

然后是 training 阶段，可以仿照图 7 中的指令进行模型的微调训练，详细配置见项目中 configs/stable-diffusion/v1-finetune_unfrozen.yaml。值得注意的是，默认的学习率是 1.0×10^{-6} ，因为实验发现 Dreambooth 论文中的 1.0×10^{-5} 导致可编辑性差；参数 reg_weight 对应于 Dreambooth 论文中正则化的权重，默认设置为 1.0。对于 Dreambooth 中的占位符单词 [V]，原论文通过在 T5-XXL 分词器中使用一个罕见的单词来实现这一点。为了简单起见，在这里使用一个随机的单词 “sks” 并且将其硬编码。

```
python main.py --base configs/stable-diffusion/v1-finetune_unfrozen.yaml
    -t
    --actual_resume /path/to/original/stable-diffusion/sd-v1-4-full-ema.ckpt
    -n <job name>
    --gpus 0,
    --data_root /root/to/training/images
    --reg_data_root /root/to/regularization/images
    --class_word <xxx>
```

图 7. 实验结果示意

最后的 Generation，经过上述训练后，使用训练保存的 checkpoint 并通过图 8 中的命令，通过运行命令来获得个性化样本。

```
python scripts/stable_txt2img.py --ddim_eta 0.0
                                --n_samples 8
                                --n_iter 1
                                --scale 10.0
                                --ddim_steps 100
                                --ckpt /path/to/saved/checkpoint/from/training
                                --prompt "photo of a sks <class>"
```

图 8. 实验结果示意

4.4 创新点

4.4.1 强化正则化策略

本实验对正则化图像的数量进行了探索，发现增加正则化图像的数量能够显著增强模型的正则化效果，进而提升编辑能力。最初使用 8 张图像用于正则化，但进一步尝试了更多数量像 100 张图像的情况，在实验对比后发现，更多的正则化图像有助于模型更好地学习到稳定的特征表示，从而在对主体进行各种编辑操作（如属性修改、风格迁移等）时，能够生成更合理、更符合预期的图像。例如，在进行复杂的艺术风格迁移时，更多的正则化图像使得模型能够更准确地把握主体结构，避免因风格转换而导致的主体变形或特征丢失。

4.4.2 优化学习率设置

经过对比几个学习率，发现 DreamBooth 模型的学习率进行了优化调整。在参考原论文中学习率 1.0×10^{-5} 的基础上，发现该学习率在应用到 sd 预训练模型后，出现编辑能力较差的情况。通过对比 3 个学习率，最终确定默认学习率为 1.0×10^{-6} 时，模型在训练过程中能够更稳定地收敛，并且在生成图像时能够更好地平衡主体特征保持与编辑灵活性。

4.4.3 实验扩展

本实验对原文的四大实验任务均进行了复现，但是在复现过程中又做了扩展，以此来进一步探究挖掘 DreamBooth 的微调能力。比如：在重新情景化能力中，除了常见的自然场景，尝试将主体置于极端环境中，如外太空。观察模型在生成这些罕见场景下主体图像时的表现，包括主体与极端环境元素的融合效果；又如在视角合成中，探索比原文实验更为刁钻的侧身和背后视角，由于微调的几张图像中只有该主体的正面视角，所以极大考验了微调后的模型对主体外部特征的理解能力，推测并生成合理的新视角的想象力，对物体空间结构和逻辑的把握能力。

5 实验结果分析

在复现 DreamBooth 的实验过程中，遵循了原文的核心方法和技术路线。通过对特定主体（如狗、背包、茶壶等）的图像生成实验，在多个任务上取得了与原文相近的定量与定性分析结果。

在保真度的评估上，我们采用了与原文一样的 CLIP-I、CLIP-T 和 DINO 指标。通过计算生成图像与真实图像在特征空间中的相似度，我们发现，尽管我们复现的 subject 与 prompt 组合数量不及原文丰富，但在已有的实验组合中，生成图像能够较好地保留主体的关键特征，以及生成的图像在遵循提示信息方面表现良好。例如，在生成狗的图像时，其品种特征（如毛色、体型、面部结构等）在不同场景和视角下都能得到较为准确的呈现，与原文中高保真度的生成效果相符。这表明我们的复现模型在学习主体特征表示方面具有较高的准确性，并且能够理解并将提示中的语义信息转化为相应的图像内容。

然后在其他几个任务上，像在艺术风格迁移中，模型能够有效地将主体的特征与艺术风格相结合，即使在设置的扩展实验中，生成的图像在色彩运用、笔触表现等方面体现了相应

艺术风格的特点，同时保持了主体的可辨识度，与原文在风格迁移效果上具有一致性；而在属性修改方面，我们不仅实现了原文中的物种融合，还针对特定主体赋予不同的情感，结果也是可以生成不错的基于微调图像的不同情感表现的新图像，整体上能体现出模型在属性修改任务中的能力和潜力。

当然，虽然复现实验取得了一定的成果，但也存在一些局限性。由于复现的 subject 与 prompt 组合相对较少，可能无法完全涵盖模型在各种复杂情况下的表现。在一些罕见的主体或极端提示条件下，模型生成的图像可能出现一些不符合预期的情况，如主体特征的轻微扭曲或场景元素的不合理组合。这些局限性为我们进一步优化模型和拓展实验提供了方向，未来我们将通过增加实验数据的多样性和复杂性，进一步提升模型的性能和泛化能力。

6 总结与展望

本实验对 DreamBooth 进行了复现，并在多个方面开展了深入研究。通过采用 DINO、CLIP-I 和 CLIP-T 指标，对主体保真度和提示保真度进行评估，结果表明，尽管复现的 subject 与 prompt 组合有限，但在已有的实验中，模型在关键指标上与原文表现出相似的趋势，有效验证了复现的准确性和有效性。在其他任务上的实验，像重新情境化方面、艺术风格迁移上、新视角合成，都展示了微调后模型在不同场景下主体适应性、逻辑理解和融合方面的卓越能力。

然而，实验还存在可以改进的地方。复现的组合数量相对较少，且在处理极端或模糊提示时模型能力有待提升。未来研究将着重增加实验数据的多样性和复杂性，进一步优化模型参数和训练策略，如继续探索正则化图像数量、等对模型性能的影响，改进标识符 [V] 的选择方法，以提升模型在各种任务中的性能和泛化能力，为 DreamBooth 技术的发展提供更有力的支持，推动其在图像生成领域更广泛的应用。

参考文献

- [1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 10684–10695, June 2022.
- [2] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2020.
- [3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Proc. Conf. on Neural Information Processing Systems*, 35:36479–36494, 2022.
- [4] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

- [5] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.
- [6] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Proc. Conf. on Neural Information Processing Systems*, 35:16890–16902, 2022.
- [7] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2023.