

# 论文 Continuous, Subject-Specific Attribute Control in T2I Models by Identifying Semantic Directions 复现

## 摘要

近年来, 文本到图像 (Text to Image, T2I) 扩散模型的发展显著提升了生成图像的质量。然而, 由于自然语言的局限性 (例如“人”和“老人”之间缺少连续的中间描述集), 实现对特定对象的属性的细粒度控制仍然是一个挑战。尽管引入了许多方法来增强模型或图像生成过程以实现这种控制, 但无需为模型提供参考图像的方法仅限于实现全局细粒度属性控制或针对特定对象的粗略属性控制, 而不是同时实现。这篇文章指出, 在常用的令牌级 CLIP 文本嵌入中存在一些变化方向, 可以实现对文本到图像模型中的特定对象的语义上的属性进行细粒度的控制。基于这一观察, 提出了高效的无优化和基于鲁棒优化的两种方法, 用于从对抗文本中识别特定对象的属性的这些变化方向。相关实验结果表明, 这些方向可用于以组合方式 (即控制单个对象的多个属性) 对特定对象的属性进行细粒度控制, 从而增强文本输入的表达能力, 且无需调整扩散模型。

**关键词:** 图像生成; 扩散模型; 属性控制; 模型微调

## 1 引言

图像生成 (Text to Image, T2I) 模型的性能和图像质量在近年来取得了显著进展, 但在实现属性控制的细粒度控制方面, 仍然面临着持续的挑战。这主要源于自然语言表达的局限性, 尤其是缺乏可以供模型理解的连续描述。与图像编辑任务不同 [10], 图像生成任务无法借助参考图像或通过对待修改的图像添加掩码来保证修改对象的针对性, 因此在这种任务中, 对象实例只能通过文本提示 (prompt) 来识别。

扩散模型 (Diffusion Model) 是一类基于概率过程的生成模型, 近年来在生成任务中表现出色。其核心思想源于反向扩散过程: 首先通过逐步添加噪声将数据转换为纯噪声, 然后通过学习到的去噪过程, 逐步将噪声恢复为原始数据。扩散模型的优势在于其能够生成高质量的样本, 尤其在图像生成领域展现出超越生成对抗网络 (GAN) 和变分自编码器 (VAE) 的表现。与传统的生成方法不同, 扩散模型通过对数据进行多次扰动和恢复, 学习到数据的深层结构与分布。特别地, 随着去噪过程的迭代, 模型能够精确捕捉数据的细节, 生成质量更高的图像。近年来, 扩散模型被广泛应用于图像生成、语音合成、视频生成等多种任务, 并在多模态生成和数据恢复等方面展现了巨大的潜力。随着研究的深入, 扩散模型的应用前景愈加广阔, 成为生

成模型领域的重要研究方向. 本文专注于扩散模型在图像生成任务中的潜力, 探究更高质量和灵活的图像生成方案.

过去的方法通常专注于全局细粒度控制 [7, 11–14] 或局部粗粒度控制 [9, 21], 而在实际应用中, 已有的方法难以同时实现这两者的兼顾. 本文旨在通过细粒度属性控制的方法, 在生成过程中的对象针对性和属性幅度方面进行优化, 从而实现对特定对象的单独细粒度属性控制.

传统的直接将属性描述加入文本提示的方法, 往往只能保证粗粒度的控制, 且在组合多个描述时, 扩散模型 (Diffusion Model) 可能会忽略部分信息 [5, 14]. 本文发现, CLIP 嵌入中的某些语义方向 [18], 能够满足细粒度且针对特定对象的属性控制, 这与反演方法不同, 是通过学习具体对象实例的信息来实现 [6, 15]. 基于这一观察, 本文介绍了识别特定属性方向的方法, 并展示了如何使用这些方法来增强文本提示输入, 以组合的方式对特定主题的属性进行细粒度连续控制, 而无需调整扩散模型或在生成过程中产生额外成本.

## 2 相关工作

相比生成对抗网络 (Generative Adversarial Networks, GAN) [4, 17], 扩散模型不提供类似有序和可解释的隐空间, 因此如何实现扩散模型更强大可控的视觉控制能力成为了一个研究热点. 一种常见的方法是直接使用新的或调整现有的神经特征来引导生成过程朝向语义变化, 以此实现细粒度属性控制.

在利用扩散模型进行图像编辑 (Image to Image, I2I) 任务时, 待修改图像会作为输入传递给模型, 期望对此输入图像进行修改, 在图像编辑任务中常见的一个方法是利用掩码控制图像中希望修改的部分 (即 In-Painting 方法), 但在图像生成任务中, 通常不会向扩散模型输入图像, 而只有文本输入. 受此启发, 本文考虑在反向扩散过程使用编辑掩码标记对象来强制将修改定位到希望修改的对象实现局部编辑.

为了控制目标属性  $A_i$  在图像中的表达, 这种生成方法的性质使图像有限的部分成为影响生成过程的目标. 为了控制图像属性表达, 一种选择是通过在全局重新定义属性表达式来修改扩散模型. 另一种选择是使用动态预测的或固定的辅助方向直接修改起始噪声隐空间  $x_T$  或中间噪声隐空间  $x_t$ . 此外, 也可以通过影响扩散模型如何解释文本提示嵌入  $e$  实现. 最后, 可以直接修改文本提示嵌入  $e$ , 影响文本编码器  $\epsilon_{\text{CLIP}}$ . 类似 [6, 8, 15, 20] 的方法以前研究了这种插入实例外观信息的方法, 但没有研究细粒度属性控制.

为了保证本文的方法可以稳定地生成包括特定对象实例的图像, 对扩散模型的个性化进行了研究. 基于微调的方法包括 DreamBooth [19], 它利用先验保留损失来调整视觉骨干, 文本反转 [6] 等方法则将模型的调整限制在要优化的添加嵌入向量上, 以绑定特定的对象实例. 但在复杂组成结构的场景中获得增强的控制仍然是文本到图像模型的一个重大挑战.

大量的工作研究了在生成对抗网络中 CLIP 嵌入空间对生成任务的指导作用 [1–3, 16], 这些研究指出 CLIP 嵌入空间存在与全局语义变化相关联的编辑方向, 并可用于指导图像生成过程.

本文提出了学习这些语义方向, 并阐述如何将其用于精细控制目标对象的属性的方法. 该方法不需要修改扩散模型本身, 也不会在生成过程中引入额外的计算开销.

### 3 本文方法

#### 3.1 本文方法概述

本文从文本-图像对 (Text-Prompt pair) 中学习语义变化，并提出了两种不同的实现思路。直接从对抗文本提示中提取属性语义变化的方法不需要优化模型，实现简单，可以满足实现对属性的控制。本次复现主要关注的是学习鲁棒的细粒度属性变化方向的方法，这种方法通过识别文本提示与图像两者变化的关系，学习一个鲁棒的修改增量对扩散模型进行微调。

#### 3.2 从文本-图像对中学习语义变化

此前的研究表明，预训练的 T2I 扩散模型的重建损失可以实现将图像中对象属性的表示反向传播到文本嵌入中 [6, 8, 15, 20]。这表明，扩散模型能够解释那些不完全位于 CLIP 模型所提供的文本嵌入空间中的点。

尽管对象属性信息是在像素级重建的，但我们发现这种通用方法也能够直接学习语义信息。在使用单个图像/提示对  $(x_0, \text{prompt})$  时，我们向图像添加随机噪声，并通过扩散模型  $x_0(\cdot)$  反向传播正则化重建损失。我们更新了一个可学习的增量  $\Delta e$ ，并将其添加到文本嵌入  $e$  中，在最小化正则化重建损失的过程中学习属性变化方向增量  $\Delta e$ 。其中，噪声  $\epsilon$  在每一步都会被随机重新绘制。

$$\mathcal{L}(x_0, e + \Delta e) = \mathbf{E}_{\epsilon \sim N(0, I), t \sim \mathcal{U}(0, T)} [\omega(t) \|x_0 - \hat{x}_0(\alpha_t x_0 + \sigma_t \epsilon | e + \Delta e, t)\|_2^2] \quad (1)$$

本文的方法确实能够学习到文本嵌入增量  $\Delta e$ ，该增量捕获了生成图像集与目标图像之间的语义差异（在仅给出文本提示的条件下）。这种方法大大缩小了原始文本提示生成图像与目标图像之间的语义差距。此外，原始文本嵌入和应用增量后的嵌入之间的线性插值展示了从原始生成图像到目标图像的清晰语义过渡。这表明，CLIP 嵌入空间在语义上至少是局部平滑的。然而，本文的实验中也观察到在插值过程的某些子集中，图像发生了显著变化，这表明嵌入空间并非全局平滑的。

关于这些学习到的语义编辑增量  $\Delta e$  与初始文本提示之间的关系，在语义上我们的训练方法对完整文本提示嵌入的适应性良好，基本满足缩小了原始生成的图像和目标图像之间的差距，即这种方式学习到的方向增量  $\Delta e$  可以满足按预期调整图像属性。我们发现，仅应用学习到的编辑增量  $\Delta e$  的特定对象的编辑方向增量  $\Delta e_{S_j}$  就足以获得属性  $S_j$  的一定程度上的解耦编辑结果。这种局部编辑在语义上接近完整编辑，同时对图像其余部分的影响很小。我们只有在训练后才添加增量  $\Delta e$  的掩码，这意味着我们的训练过程将语义信息与对应于每个属性  $S_j$  相关联，并且这些方向增量  $\Delta e_{S_j}$  可直接在语义层面上影响对象  $S_j$  属性  $A_i$  的表达。这提供了一种简单的可解释的方法来定位修改。

在学习到方向增量  $\Delta e$  后，在图像采样生成阶段，我们将输入的文本提示经过 CLIP 编码器后得到的文本嵌入  $e$  与通过超参数  $\alpha_i$  控制的方向增量  $\Delta e$  相加得到最终的文本嵌入提示  $e'$  以替代原本的文本嵌入  $e$ 。其中  $\alpha_i$  控制属性变化程度。

$$e'(e, \alpha_i \Delta e_{A_i})_{[S_j]} = e_{[S_j]} + \alpha_i \Delta e_{A_i} \quad (2)$$

### 3.3 从对抗文本提示中提取属性语义变化

首先, 本文提出了一种不需要优化模型的方法, 从对抗文本提示中识别影响特定属性  $A_i$  的文本嵌入空间中的语义方向 (例如 “年龄” 属性的 “老人” 与 “老人”, 两个提示都使用相同的名词 “人” 表示对象  $S_j$ , 而属性形容词的语义方向相反) . 由于 CLIP 文本编码器已经将语义属性聚合到对应的对象中, 我们首先分别获得对抗文本中正提示和负提示的 CLIP 文本嵌入结果  $\epsilon_{\text{CLIP}}(\text{prompt}_{A_i,+})_{[S_j]}$  和  $\epsilon_{\text{CLIP}}(\text{prompt}_{A_i,-})_{[S_j]}$ . 然后, 我们通过 (3) 式计算两个文本嵌入中对象的属性表达的差异

$$\Delta e_{A_i} = (\epsilon_{\text{CLIP}}(\text{prompt}_{A_i,+}))_{[S_j]} - (\epsilon_{\text{CLIP}}(\text{prompt}_{A_i,-}))_{[S_j]} \quad (3)$$

这直接得到了一个可以指导属性控制的方向增量  $\Delta e_{A_i}$ , 这个方向增量与特定对象的特定属性绑定 (例如上例中的名词 “人”的属性 “年龄”), 这种方法可以简单实现对属性的控制, 且不需要进行优化模型. 为了获得更为可靠的结果, 本文利用大量对抗文本提示识别的结果平均以增加方向增量  $\Delta e_{A_i}$  修改属性的可靠性.

### 3.4 学习鲁棒的细粒度属性变化方向

在上一节的基础上, 本文提出了一种基于优化扩散模型的方法, 在训练过程中, 我们利用扩散模型生成的图像  $x_0$  中的变化与模型得到的文本提示的关系, 识别文本提示与属性控制结果的关系, 并在模型预测的隐空间找到属性控制的表现, 并通过扩散模型反向传播这个结果, 学习方向增量  $\Delta e$ , 这种方法只是对扩散模型进行微调, 不需要重新训练模型.

在每个优化步中, 向 CLIP 文本编码器  $\epsilon_{\text{CLIP}}$  输入一组对抗文本提示, 通常包括一个基准文本和一对语义方向相反的对抗文本, 在扩散模型的隐空间, 首先根据文本提示通过扩散模型随机采样一个基准图像  $\hat{x}_{0,a}$ , 并根据此基准图像在随机时间步  $t$  处预测中间结果  $\hat{x}_{t,a}$ , 再经过采样过程分别得到具有正提示和负提示的预测结果的隐空间表示  $\hat{x}_{0,+}$ 、 $\hat{x}_{0,-}$ , 结合控制属性变化程度的超参数  $\alpha_i$ , 得到供扩散模型学习的目标隐空间表示  $\hat{x}_{0,target}$ .

$$\hat{x}_{0,target}(\alpha_i) = \hat{x}_{0,a} + \alpha_i \cdot (\hat{x}_{0,+} - \hat{x}_{0,-}) \quad (4)$$

在优化过程中, 参数  $\alpha_i$  会随机取值, 学习不同编辑程度下图像的变化, 使模型经过优化后可以实现属性编辑的平滑控制. 同时, 我们根据方向增量  $\Delta e$  通过扩散模型采样一个依据  $e'$  得到的隐空间表示  $\hat{x}_0$ , 通过最小化扩散模型的隐空间表示和目标隐空间表示之间的差异, 利用损失函数  $\mathcal{L}_{\text{delta}}$  训练模型对方向增量  $\Delta e$  解释能力. 通过损失函数  $\mathcal{L}_{\text{delta}}$  将  $\Delta e$  反向传播, 通过最小化此损失函数学习  $\Delta e$ .

$$\mathcal{L}_{\text{delta}} = \mathbf{E}_{\alpha_i} \left[ \omega(t) \| \hat{x}_{0,target}(\alpha_i) - \hat{x}_0(x_{t,a}|e'(e, \alpha_i \Delta e_{A_i}), t) \|_2^2 \right] \quad (5)$$

## 4 复现细节

### 4.1 与已有开源代码对比

本文作者公开了相应的源代码, Github 链接为 <https://github.com/CompVis/attribute-control>. 在本次的复现过程中, 参考了作者的源代码, 在充分理解其工作原理和实现细节的

基础上，基于实验环境调整了一些优化的超参数。在作者的源代码中包括一种不需要优化模型，仅通过对抗文本得到修改增量  $\Delta e$  的方法 *learn\_delta\_naive\_clip* (3.3 节) 和一种通过损失函数优化模型的方法 *learn\_delta* (3.4 节)，总的来说，本文的方法属于轻量级的微调，期望开销较小，本次复现主要关注后一种方法。在作者的学习方向增量  $\Delta e$  的算法中使用了固定的学习率，本次复现增加了学习率余弦败火机制，以尝试减少模型优化的开销的同时增强学习优化后生成图像的质量。修改的部分代表性代码如图 1 所示。

```
lr = min_lr + (max_lr - min_lr) * (1+ np.cos((global_step + 1) / max_steps * np.pi)) / 2
optimizer = torch.optim.AdamW(delta.parameters(), lr=lr)
print_lr = optimizer.param_groups[0]['lr']
logger.info(f'Globalstep {global_step} Optimizer learn rate:{print_lr}')
```

图 1. 增加的学习率余弦败火机制代码

## 4.2 实验环境搭建

本次实验环境为 Python3.9.20，运行在 Linux 系统上，部分核心代码库及其版本为 torch2.5.1，numpy2.0.2，diffusers0.31.0，einops0.8.0。本文作者在 requirements.txt 文件中有给出所有需要的代码库的列表。本文采用的模型为 SDXL (stable-diffusion-xl-base-1.0)。

## 4.3 创新点

我们对本文提出的方法进行了改进尝试。1) 为了加速优化学习的过程，本文在计算损失时采用了累计机制，为此我们首先尝试了修改累计步数和学习轮次等超参数以期望在可接受的开销范围内得到更好的结果。2) 本文在优化学习的过程中设置了固定的学习率，我们尝试了增加学习率余弦败火机制以在学习过程中平滑地降低学习率，以更精细地方式学习一个更好的目标方向增量  $\Delta e$ 。

## 5 实验结果分析

本次复现对不同修改下得到的方向增量  $\Delta e$  在实际生成图像的效果进行了对比。这里列出有代表性的几项结果：包括本文作者提供的预先学习好的方向增量  $\Delta e$  (作为结果参考)，本次复现中修改了部分学习优化超参数的复现版本方向增量  $\Delta e$ ，本次复现中增加了学习率余弦败火机制后的复现版本方向增量  $\Delta e$ 。在设计输入的文本提示时，采用了两条复杂度不同的输入进行测试，包括一条简单的单对象文本提示“a photo of a beautiful man”和一条包括复杂环境和多个对象的文本提示“a photo of a write man sitting in a chair and a write woman standing next to him in a beautiful garden”，在这两条文本提示中，约定“man”作为被识别的进行细粒度局部编辑的对象。生成的结果图像如下。



图 2. 本文预训练的方向增量  $\Delta e$  在简单文本提示下的编辑结果

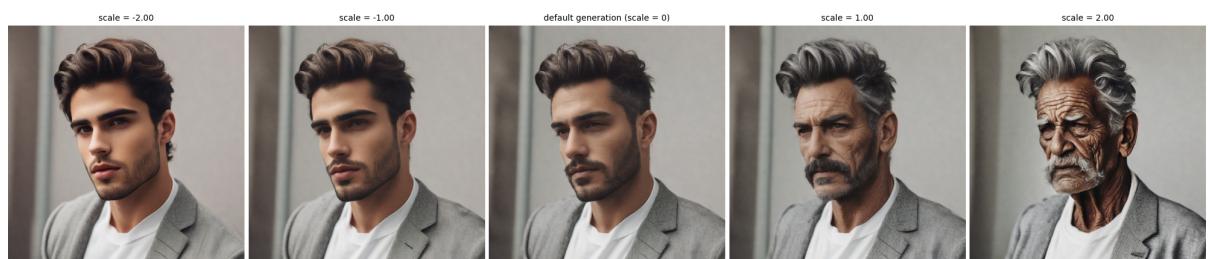


图 3. 调整训练参数后的方向增量  $\Delta e$  在简单文本提示下的编辑结果

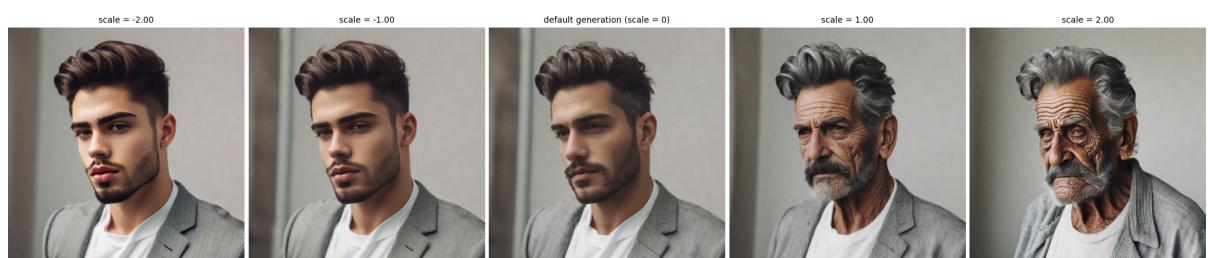


图 4. 增加学习率余弦败火机制后的方向增量  $\Delta e$  在简单文本提示下的编辑结果



图 5. 本文预训练的方向增量  $\Delta e$  在复杂文本提示下的编辑结果



图 6. 调整训练参数后的方向增量  $\Delta e$  在复杂文本提示下的编辑结果



图 7. 增加学习率余弦败火机制后的方向增量  $\Delta e$  在复杂文本提示下的编辑结果

可以看到，在简单文本提示输入下，优化后的结果均产生了更平滑的编辑效果，但在复杂文本提示输入下，对于干扰对象的属性保持仍没有较好的变化，但对于环境的属性均保持的较好。图中参数 scale 代表了修改的方向及程度。

## 6 总结与展望

在本次复现实验中，我完成了对本文作者的结果的复现，并考虑了一些基本的改进方法并进行了尝试，得到了相应改进后的结果并进行了简单的对比。可以发现达到了本文提到的结果，同时控制了优化微调模型的开销。同时，本此复现工作还存在许多不足，还需要更多的实验进行验证和进一步的优化，同时对于生成图像的结果也没有使用更具说服力的指标进行评价模型优化的结果。

此外，在未来的工作中，我考虑将模型更换为 diffusion 3 等新模型尝试进一步稳定这种针对特定对象的细粒度修改方法的效果，解决一些在目前已有的复现工作中还存在的问题，并提高编辑生成图像结果的质量。同时，考虑增加特定对象多属性组合控制等实验丰富本文方法的应用场景，验证本文提到的相关实验的结果。

## 参考文献

- [1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022.
- [2] Tsuyoshi Baba, Kosuke Nishida, and Kyosuke Nishida. Robust text-driven image editing method that adaptively explores directions in latent spaces of stylegan and clip. *arXiv preprint arXiv:2304.00964*, 2023.
- [3] Ahmet Canberk Baykal, Abdul Basit Anees, Duygu Ceylan, Erkut Erdem, Aykut Erdem, and Deniz Yuret. Clip-guided stylegan inversion for text-driven real image editing. *ACM Transactions on Graphics*, 42(5):1–18, 2023.
- [4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

- [5] Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pages 8489–8510. PMLR, 2023.
- [6] 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.
- [7] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer, 2025.
- [8] Inhwa Han, Serin Yang, Taesung Kwon, and Jong Chul Ye. Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*, 2023.
- [9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- [10] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.
- [11] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022.
- [12] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.
- [13] Nan Liu, Shuang Li, Yilun Du, Josh Tenenbaum, and Antonio Torralba. Learning to compose visual relations. *Advances in Neural Information Processing Systems*, 34:23166–23178, 2021.
- [14] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.
- [15] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

- [16] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021.
- [17] Alec Radford. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [18] 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.
- [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] 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.
- [21] Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 36:26291–26303, 2023.