

Prompt-to-Prompt Image Editing with Cross Attention Control

摘要

随着文本驱动的图像合成模型的发展，文本驱动的图像编辑技术成为了一个新兴的研究领域。本课程论文 [1] 旨在复现并验证“Prompt-to-Prompt Image Editing with Cross Attention Control”论文中提出的方法，该方法通过仅编辑文本提示来实现对预训练扩散模型生成的图像的直观编辑。我们首先详细阐述了原论文中的核心思想，即利用交叉注意力层来控制图像空间布局与文本提示中单词之间的关系。接着，我们描述了复现过程中的关键步骤，包括模型的搭建、训练过程、以及如何通过修改文本提示来控制图像编辑的具体操作。我们的实验结果表明，复现的方法能够成功地在多种图像和文本提示上实现高质量的合成和对编辑提示的高保真度。此外，我们还探讨了在复现过程中遇到的挑战，以及可能的改进方向。本论文的复现工作不仅验证了原论文方法的有效性，也为未来的图像编辑研究提供了参考和启示。

关键词：文本驱动图像编辑；交叉注意力控制；预训练扩散模型；Prompt-to-Prompt编辑框架

1 引言

在深度学习和人工智能技术的推动下，图像生成和编辑领域取得了显著进展。特别是在文本驱动的图像合成模型方面，如DALL-E 2 [3]、Imagen [4]和Stable Diffusion等，它们能够根据文本提示生成多样化的图像，展现出了强大的语义理解和视觉合成能力。这些模型不仅能够理解复杂的文本描述，还能够生成与之匹配的图像，为图像编辑提供了新的可能性。

然而，尽管这些模型在图像合成方面取得了巨大成功，它们在图像编辑方面的能力却相对有限。特别是在保留原始图像内容和结构的同时进行编辑时，即使是对文本提示的微小修改也常常导致完全不同的结果。为了解决这一挑战，“Prompt-to-Prompt Image Editing with Cross Attention Control”论文提出了一种创新的图像编辑框架，该框架通过控制文本提示来实现对图像的精确编辑，而无需任何额外的图像处理或掩码操作。

这项工作的意义在于，它不仅能够提高图像编辑的直观性和便捷性，还能够在不牺牲图像质量的前提下，实现对图像的局部或全局编辑。这种方法的核心在于交叉注意力层，它负责将文本提示中的单词与图像的空间布局相联系，从而实现对图像编辑的精细控制。通过这种方法，可以实现包括局部编辑（如替换一个单词）、全局编辑（如添加一个描述）以及控制单词在图像中的反映程度等多种编辑应用。

2 相关工作

在深度学习和人工智能技术的推动下，图像生成和编辑领域取得了显著进展，尤其是在文本驱动的图像合成模型方面。以下是几个关键点：引文的bib文件统一粘贴到**refs.bib**中，引用方式 [2]。

2.1 文本驱动的图像合成技术进展

最新的研究进展表明，大规模文本到图像合成模型如Imagen和Stable Diffusion等，不仅能够理解复杂的文本描述，还能够生成与之匹配的图像，如图 1 所示。这些模型的发展为图像编辑技术提供了新的可能性和挑战，尤其是在如何精确控制图像内容以匹配文本提示方面。



图 1. 文生图示例

2.2 交叉注意力层

交叉注意力层在图像编辑中扮演着至关重要的角色，Cross-Attention 层是控制图像空间布局 (spatial layout) 和 prompt 中分词 (token) 关系的关键，高维张量 cross-attention maps 可以在 pixels 和 tokens 间建立连接。以 text-conditioned diffusion 的一次图像生成过程为例，pixels 和 tokens 间的联系如图 2。这种机制使得模型能够识别文本中的关键词汇与图像中相应区域的关联，实现精确的文本驱动编辑，这对于图像编辑的精确性和灵活性至关重要。

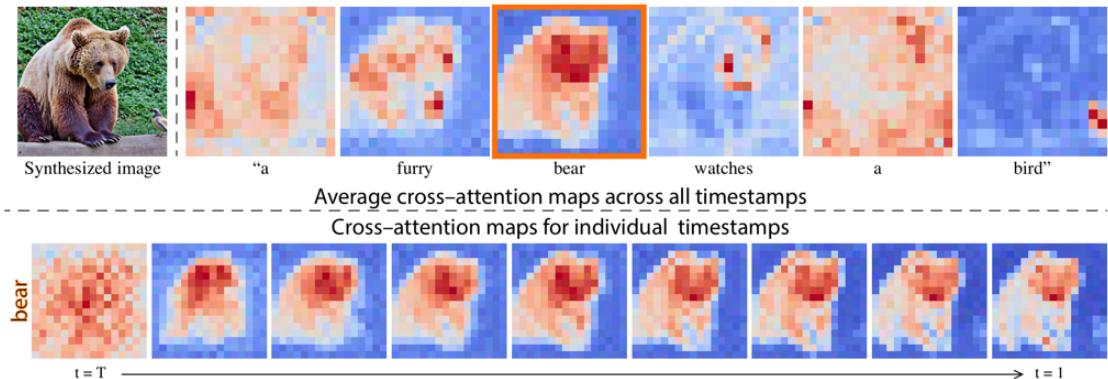


图 2. Cross-Attention层示例

2.3 Inpaint Anything技术

Inpaint Anything技术通过结合Segment Anything Model (SAM)、LaMa和Stable Diffusion等强大的视觉模型，实现了无需掩码的图像修复。用户只需点击图像中的物体，即可实现无痕移除、内容填补或场景替换，极大地扩展了图像编辑的应用范围和用户友好性。

2.4 Edit Everything技术

Edit Everything技术展示了如何通过文本提示引导图像生成和编辑。它通过结合SAM、CLIP和Stable Diffusion等组件，允许用户使用简单的文本指令来编辑图像，实现了从文本到图像的直观转换，为图像编辑提供了新的思路。

2.5 Grounded-SAM技术

Grounded-SAM技术通过检测、分割和替换图像中的对象，展示了深度学习在图像编辑领域的强大能力。这项技术不仅能够处理静态图像，还能够扩展到视频和3D场景，为图像编辑提供了更为丰富的应用场景。

3 本文方法

3.1 本文方法概述

此部分对本文将要复现的工作进行概述，图的插入如图所示：本论文提出的 Prompt-to-Prompt 方法，使用基于文本的预训练 diffusion 模型，通过调整模型生成图像过程中的 cross-attention maps，从而保证在尽可能保持原有图像的空间布局和几何外观的情况下实现图像编辑。具体操作是通过在扩散过程中向预训练模型中注入特定的 cross-attention maps，能够使得一些 pixels 去匹配对应的 tokens。而为了维持原始图像的空间布局与几何形状，可以在生成编辑图像的过程中向 cross-attention maps 中注入原始图像的特定 cross-attention maps，具体示意图如图 3 所示。这种技术主要适用于Word Swap、Adding a New Phrase和Attention Re-weighting这三种任务，也是本文复现工作的重点。

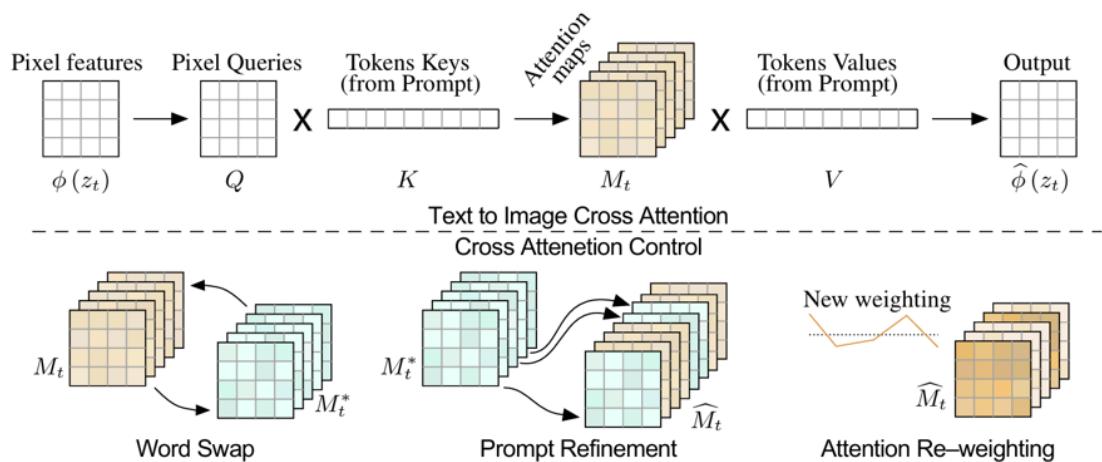


图 3. Cross-Attention 控制

3.2 编辑图像生成

记原始图像为 \mathcal{I} ,原始 prompt 文本为 \mathcal{P} ,编辑后图像为 \mathcal{I}^* ,编辑后 prompt 文本为 \mathcal{P}^* ,随机种子为 s 。 $DM(z_t, \mathcal{P}, t, s)$ 表示 t 时刻的逆扩散过程,输出隐空间的噪声图像 z_{t-1} 和 cross-attention map M_t 。 $DM(z_t, \mathcal{P}, t, s)\{M \leftarrow \widehat{M}\}$ 表示替换该步骤中的 M 为 \widehat{M} , $Edit(M_t, M_t^*, t)$ 表示 t 时刻的 cross-attention map 是 M_t 或 M_t^* 。

生成编辑图像时,同时使用 diffusion 模型分别对 \mathcal{P} 和 \mathcal{P}^* 进行图像生成,然后将 \mathcal{P} 生成图像过程中的 M_t 注入 M_t^* 得到新的 cross attention map \widehat{M} ,最后再用 \widehat{M} 生成新的 z_{t-1}^* 用于下一轮迭代,详细过程如图 4:

Algorithm 1: Prompt-to-Prompt image editing

```
1 Input: A source prompt  $\mathcal{P}$ , a target prompt  $\mathcal{P}^*$ , and a random seed  $s$ .  
2 Output: A source image  $x_{src}$  and an edited image  $x_{dst}$ .  
3  $z_T \sim N(0, I)$  a unit Gaussian random variable with random seed  $s$ ;  
4  $z_T^* \leftarrow z_T$ ;  
5 for  $t = T, T - 1, \dots, 1$  do  
6    $z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t, s)$ ;  
7    $M_t^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s)$ ;  
8    $\widehat{M}_t \leftarrow Edit(M_t, M_t^*, t)$ ;  
9    $z_{t-1}^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s_t)\{M \leftarrow \widehat{M}_t\}$ ;  
10 end  
11 Return  $(z_0, z_0^*)$ 
```

图 4. 编辑图像生成过程

3.3 Word Swap

当想要替换文本中的某个单词时,只需用原始图像的 cross-attention maps M_t 替换目标图像的 cross-attention maps M_t^* ,这样就可以在维持原始图像空间布局的情况下表示新的语义。然而,当涉及到大的结构修改时,如将“Photo of a cat riding on a bicycle”中的“bicycle”替换为“car”,这种注意力注入可能过于约束对象的保留,从而导致 prompt 中新加的 token 语义无法显现。

如图 5 所示,右上角是原图像和原提示。在每一行中,通过替换文本中的某个单词并注入原图像的cross-attention map来修改图像的内容,注入范围从扩散过程的 0% (左侧) 到 10% (右侧)。一方面,如果不本文提出的方法,则无法保证保留任何原图像内容。另一方面,在所有扩散步骤中注入cross-attention map可能会过度约束,从而导致文本提示的保真度低,例如,第3行, car完全变成了bicycle。

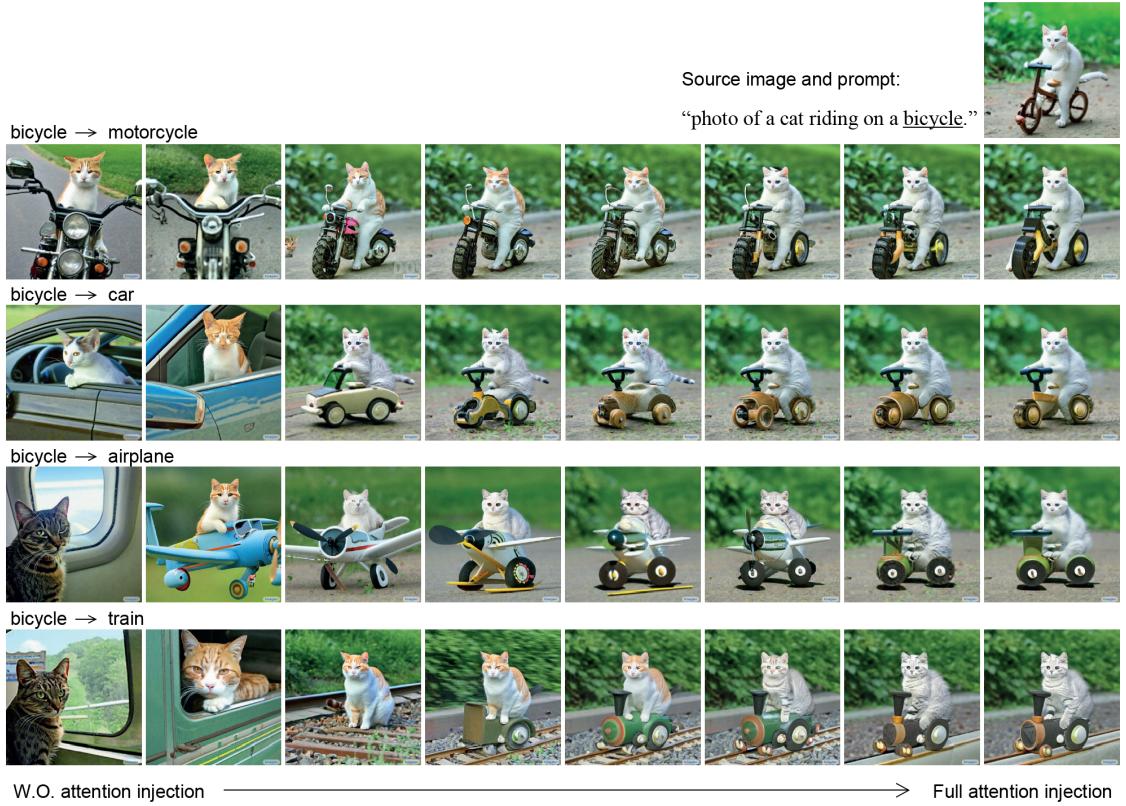


图 5. 不同扩散范围示例图

故本论文提出一种松弛策略来解决这个问题，通过控制注入时间，从而在保持构图的同时，允许必要的自由度以适应新的提示语。在扩散过程中设置时间节点 τ ， τ 之前按编辑后的生成编辑图像，使其更贴近新的提示文本的描述，从而引导图像生成向目标方向调整， τ 之后再进行注入替换，此时模型生成的细节逐渐确定，减少修改，保留原图像的空间布局。

$$Edit(M_t, M_t^*, t) = \begin{cases} M_t^* & \text{if } t < \tau \\ M_t & \text{otherwise} \end{cases}$$

3.4 Adding a New Phrase

Adding a New Phrase 是指用户向提示语中添加新的词，例如， $P = \text{"castle next to a river"}$ 到 $P' = \text{"children drawing of a castle next to a river"}$ 。为了保留共同的细节，只在两个提示语中共同的token上应用cross-attention注入。定义一个对齐函数 A ，该函数接收目标提示语 P 中的 token 索引，并输出在原提示语 P 中对应的 token 索引，如果没有匹配，则输出 None。编辑函数由以下式子给出：

$$(Edit(M_t, M_t^*, t))_{i,j} := \begin{cases} (M_t^*)_{i,j} & \text{if } A(j) = \text{None} \\ (M_t)_{i,A(j)} & \text{otherwise} \end{cases}$$

其中，索引 i 对应于图像像素，而 j 对应于文本 token。

这里给出一些论文中的示例，如图 6 所示。比如说，给小熊带的墨镜加个修饰词，变成方形的，五颜六色的等等；给定小熊手中的饮料种类，啤酒、咖啡等等。



图 6

还有下面这组图 7，也是同样的操作，被洪水淹没的街道，或者给定环境，在秋天、在冬天、在晚上等等，但是总体的布局都是没有改变的，都保留了原始图像的空间布局和几何外观。



图 7

3.5 Attention Re-weighting

Attention Re-weighting 是指用户可能希望增强或减弱某个词对生成图像的影响程度。例如，考虑提示语 $P = "A photo of a blossom tree."$ ，假設想让树变得不茂盛。为了实现这种操作，使用参数 $c \in [-2, 2]$ 来缩放分配给 token j^* 的 attention map，从而使效果更强或更弱。其余的 attention map 保持不变，即：

$$(\text{Edit}(M_t, M_t^*, t))_{i,j} := \begin{cases} c \cdot (M_t)_{i,j} & \text{if } j = j^* \\ (M_t)_{i,j} & \text{otherwise} \end{cases}$$

这里提到的参数 c 可以对编辑效果进行精细和直观的调节。

这里同样给两组示例，如图 8 和图 9

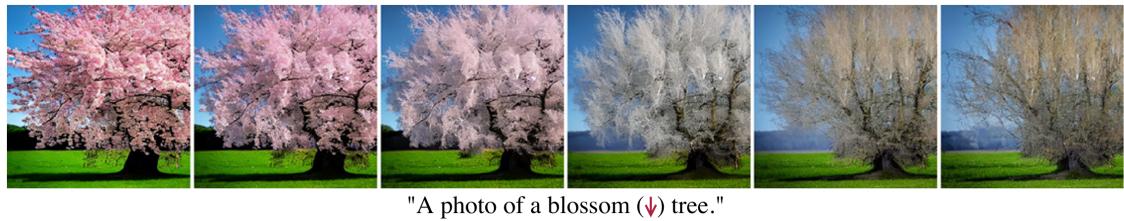


图 8



图 9

4 复现细节

4.1 与已有开源代码对比

参考代码为<https://github.com/bloc97/CrossAttentionControl>, 这是一份非官方基于Stable Diffusion实现的代码, 原论文中的prompt2prompt是基于Imagen实现的, 这篇工作的复现是基于这份非官方的代码进行复现的。在这份非官方的代码中, 作者使用的是diffusers库中的Stable Diffusion, 从huggingface上自动下载模型, 不需要手动下载, 这份代码通过修改Stable Diffusion中的指定的层, 来实现特定层的特征注入。

4.2 实验环境搭建

这篇工作使用了Stable Diffusion模型, 实验环境首先应该安装Diffusers库中的Stable Diffusion模型需要的配置文件, 可以直接在Stable Diffusion的项目文件下找到environment.yaml, 安装完官方的Stable Diffusion需要的依赖之后, 在安装该目录文件下的requirement.txt文件中的指定的依赖项。

4.3 界面分析与使用说明

项目的运行不提供可视化界面, 但是使用.ipynb文件, 便于分布调试, 使用时直接运行.ipynb文件的每个单元格, prompt2prompt中的不同的功能在不同的单元格中, 使用时可以直接选择相应功能, 同时, 可以直接在单元格中修改提示词内容, 可以实现不同的文本相对的图像的编辑操作。

4.4 创新点

p2p原论文主要是针对提示词的修改, 通过修改后图像对原始图像cross attention的注入, 来保存原始图像的结构, 创新点参考了pnp中图像编辑的方法 [5], pnp中使用源图像解码器的

残差块特征的注入，来更好的保存图像的结构，残差块中含有图像丰富的结构特征，注入残差块可以更好地保存原始图像的结构，及在对p2p中注入原始图像的cross attention的同时，注入原始图像去噪过程中的解码器的残差块的特征，来更好的保持原始图像的结构。

5 实验结果分析

本部分对实验所得结果进行展示和分析，主要包括Word Swap、Adding a New Phrase和Attention Re-weighting这三种任务，由于原论文使用的随机种子未知，故本次复现使用的文本与原论文有些许不同，但并不影响实验效果的展示。

5.1 Word Swap

如图 10所示，左图Source image，是基于文本“*A young boy playing in a field, on a hill overlooking a green valley*”生成的，中间这张将文本中的“boy”换成“girl”并保留空间布局、几何图形和语义，最右边这张将文本中的“green”换成“dry”。



图 10

5.2 Adding a New Phrase

如图 11所示，左边这张灰暗的图像由文本“*castle next to a river*”经过diffusion模型生成，右侧这张可爱稚嫩的图片在原提示文本的基础上增加了“*children drawing of a*”，使之较原始图片的风格焕然一新，但并没有改变图片的空间布局和几何外观等。



`prompt = "castle next to a river"`

`prompt = " children drawing of a castle next to a river"`

图 11

再来看下面这组图 12，也是同样的操作，左侧图片由文本“`a car on the side of the street`”生成；中间这张在原文本的基础上增加了修饰词“`flooded`”，将街道修改成被淹没的街道；右边这张在原文本的基础上增加“`at night`”，为图片增添在晚上的环境氛围。编辑后的图像都体现了编辑后文本的含义，保真度较高，并且都很好保持了原照片的布局结构，效果非常优异。



图 12

5.3 Attention Re-weighting

最后一组是增强或减弱某个单词在图像中的作用效果，这里同样给一组示例，如图 13。左侧是由提示文本“`a fluffy red ball`”生成的初始图像，右侧的第一行逐渐增强文本中“`fluffy`”的程度，第二行逐渐减弱“`fluffy`”的程度，对比效果非常明显。第一行红球的毛茸程度逐渐增强，第二行逐渐减弱，变得非常光滑。



图 13

6 总结与展望

在本论文中，成功复现了一种创新的图像编辑方法——Prompt-to-Prompt，它通过编辑文本提示来实现对图像的直观编辑，无需用户指定具体的编辑区域。该方法利用了cross-attention控制机制，这一机制在text-conditioned diffusion模型中发挥着关键作用，它控制着图像空间布局与文本提示中每个token之间的交互。通过精确注入特定的cross-attention maps，Prompt-to-Prompt能够在保持图像原有空间布局和几何外观的同时进行编辑。这项工作标志着利用文本语义功能为用户提供简单直观的图像编辑方法的第一步，使用户能够在语义、文本空间中导航，实现图像的增量变化，而不是在每次文本操作后从头开始生成图像。

尽管本文的复现工作取得了积极的结果，但仍存在一些挑战和限制。首先，当前的反转过程可能会在某些测试图像中引入可见的失真，这是一个需要在未来工作中解决的问题。其次，生成高质量的图像需要用户提出合适的文本提示，这对于复杂的构图来说可能是具有挑战性的。此外，目前的注意力图分辨率较低，限制了执行更精确的局部化编辑的能力。为了缓解这种情况，未来的工作可以考虑在更高分辨率的图层中加入Cross-Attention层。最后，当前的方法还不能用于在图像上空间上移动现有对象，这需要未来的研究来探索。

参考文献

- [1] 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.
- [2] Yilin Liu, Jiale Chen, Shanshan Pan, Daniel Cohen-Or, Hao Zhang, and Hui Huang. Split-and-Fit: Learning B-Reps via Structure-aware Voronoi Partitioning. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 43(4):108:1–108:13, 2024.

- [3] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [4] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.
- [5] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.