方法不同，作者希望避免依赖于用户定义的掩码来指示编辑的位置。实际上，生成的图像的结构和外观不仅取决于随机种子，还取决于像素通过扩散过程与文本嵌入之间的交互。通过修改交叉注意力层中发生的像素到文本的交互，能够提供由文本到文本的图像编辑能力。更具体地说，注入输入图像 I 的交叉注意力图使我们能够保留原始构图和结构。

具体而言，在通过整个扩散过程生成图片时，文本提示会在扩散的每一步和带噪声图像一起输入到模型中生成预测的噪声来指导模型生成文本特征对应的图像。

文本对图像的影响发生在交叉注意力层[7]。通过每个文本特征的空间注意力图，图像特征和文本特征融合在一起。具体而言，带噪声图像的深度空间特征通过线性变换 lq 投影到查询矩阵 Q，文本特征通过线性变换 lk 和 lv 被投影到键矩阵 K 和值矩阵 V。然后通过 Q 乘 K 的转置除以缩放因子再进行 softmax 运算得到注意力图 M。其中，矩阵 M 的元素 Mij 定义了第 j 个文本特征对像素 i 的权重。最后，M 与 V 相乘得到新的空间特征。

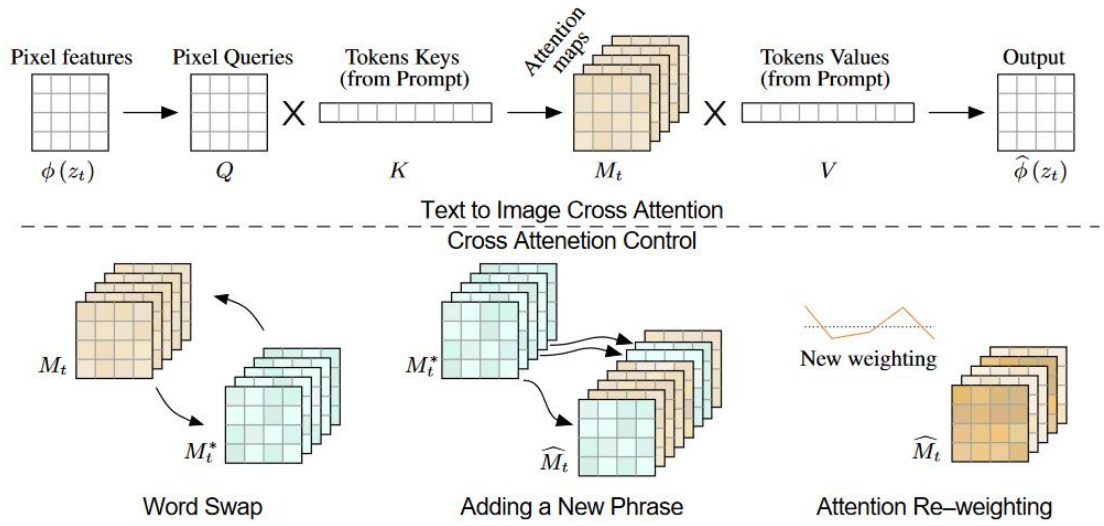


图 3 方法总览

作者提出三种基于修改注意力图的编辑方法，分别是词语替换、添加新元素以及改变词语权重。

Word Swap，用于在图像生成中替换原始提示中的词汇。例如，将“一辆红色自行车”（a big red bicycle）替换为“一辆红色汽车”（a big red car）。作者提出一种更柔和的注意力约束：Edit(Mt, M∗t, t) := {M∗t if t < τ Mt otherwise}。其中τ是一个时间戳参数，用于确定注入应用的步骤。构图是在扩散过程的早期步骤中确定的。因此，通过限制注入步骤的数量，可以引导新生成图像的构图，同时留出必要的自由度以适应新词语。

Adding a New Phrase，即 向提示中添加新的词汇。为了保留两个提示中共同的细节，只对两个提示中共同的词汇应用注意力注入。具体来说，使用了一个对齐函数 A，它接收目标提示 P∗中的一个词汇索引，并输出在 P 中对应的词汇索引，如果没有匹配则输出 None。编辑函数定义为：(Edit(Mt, M∗t, t))i,j := {(M∗t)i,j if A(j) = None (Mt)i,A(j) otherwise}。索引 i 对应于像素值，而 j 对应于文本词汇。同样可以设置一个时间戳τ来控制应用注入的扩散步骤的数量。这种编辑使得从提示到提示的转换能力更加多样化，比如风格化、指定对象属性或全局操作。

Attention Re-weighting，用于加强或减弱每个词汇对生成图像的影响程度。具体而言，将指定词汇 j*的注意力图以参数 c ∈ [−2, 2]进行缩放，从而产生更强/更弱的效果。其他注意力图保持不变。也就是说：(Edit(Mt, M∗t, t))i,j := {c · (Mt)i,j if j = j∗ (Mt)i,j otherwise}

**复现过程**

原论文只在谷歌的文生图模型 Imagen[3]上进行了实现，为了使其达到一个更好的效果，我在 Stable Diffusion XL[4]上重新进行了实现，并对代码进行了优化，大幅降低了对显存的占用，提高了生成速度。SDXL 于 2023 年 7 月公布，是当前开源的性能最好的文生图模型。SDXL 主要由三部分组成：基于 CLIP[5]的文本编码器，UNet[6]以及 VAE[8]。扩散的主要过程依赖于 UNet，因此只需关注 UNet 的结构即可。

UNet 的结构较为复杂，由多种模块堆叠而成。主要包括 Resnet 模块，Transformer 模块和 downsample，upsample 模块。由于文本对图像的影响只会发生在 Transformer 模块，因此我们只需要关注 Transformer 模块的结构。而 Transformer 模块又是由多个 basic transformer 模块串联。basic transformer 模块则是由 self attention 模块和 cross attention 模块连接得到。因此，对 self attention 和 cross attention 中的注意力图进行操作即可实现基于注意力的图像生成控制。

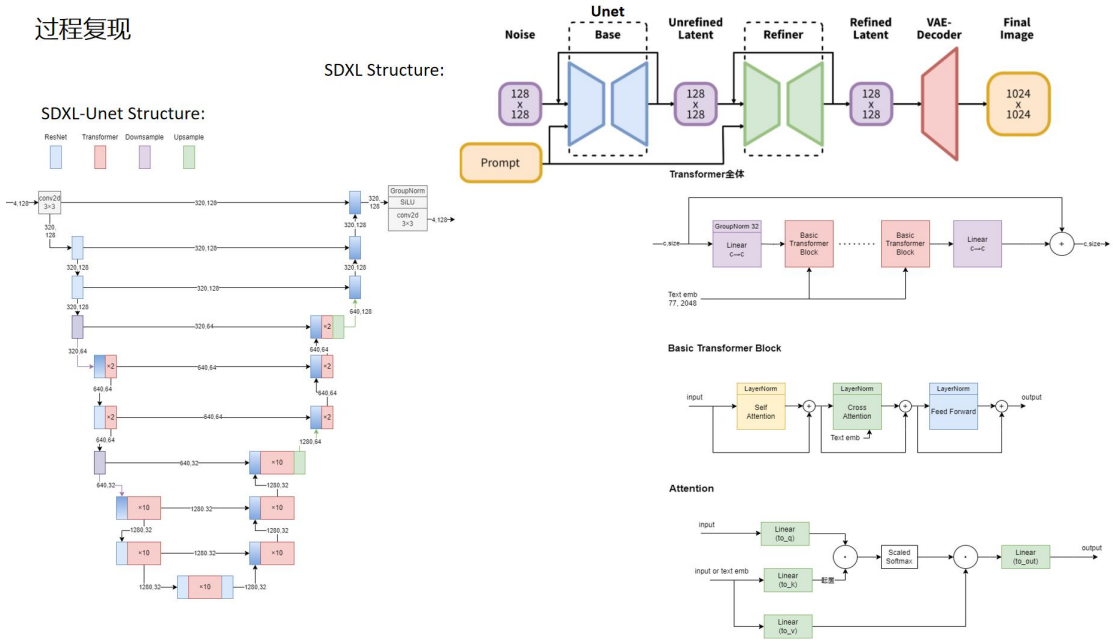


图 4 SDXL 结构分析

图 5 展示了在 sdxl 中的注意力图的可视化。每一个文本特征都有其对应的注意力图，将对应文本特征的各个时间以及 UNet 位置的注意力图求平均即得到以下结果。可以发现每个文本特征都较大程度作用于图像的特定位置而对其他位置没什么影响。比如 bear 这个词引导熊的大概形状，而 riding 引导动作，bicycle 引导自行车的形状。

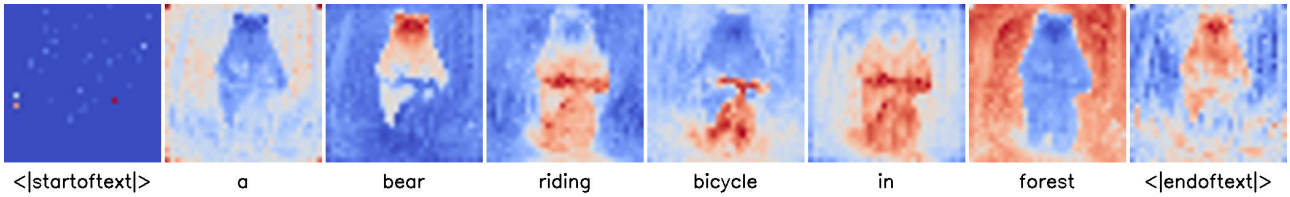


图 5 注意力图可视化

图 6 展示的是 Word Swap 即图像元素替换的一个应用。Mt*是目标图片的注意力图，Mt 是原图片的注意力图。具体实现方法是，在生成目标图片的扩散过程的早期，将没有改变的文本特征的注意力图替换为原图片的注意力图。比如原文本是“A bear is riding a bicycle in the forest”，可以将 bicycle 替换成 motorcycle，可以在保持熊的动作以及周围环境不变的情况下将自行车换成摩托车。将 in the forest 替换成 on the beach 可以改变环境。或者将熊替换成狗来骑车。还有一个对照组就是在没有施加注意力控制的情况下整个构图包括熊的形象和动作都发生了改变。

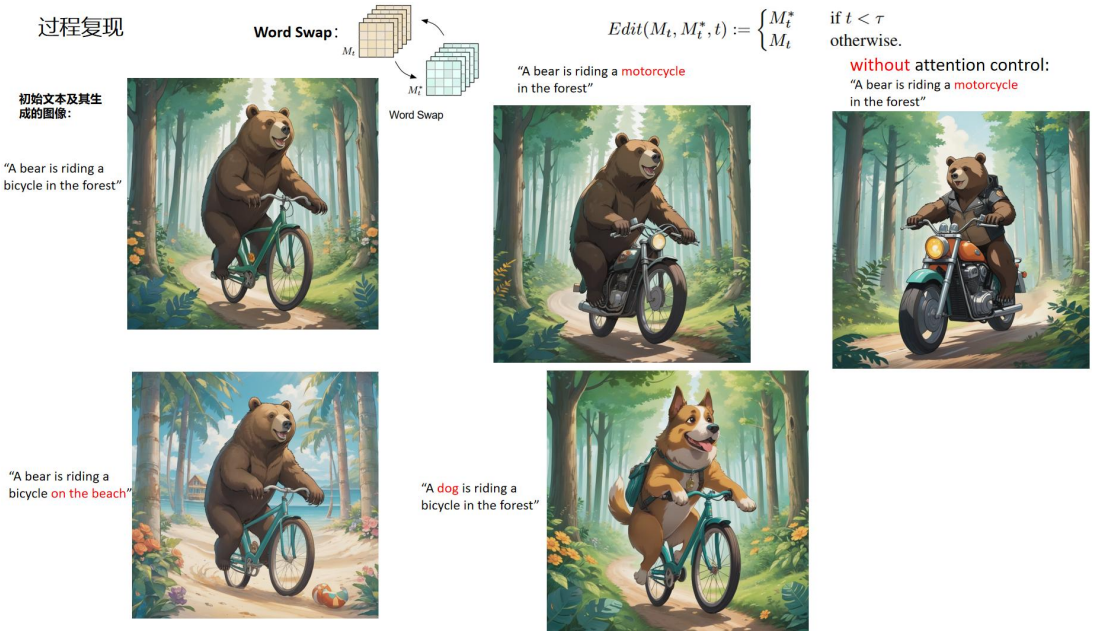


图 6 Word Swap 应用

图 7 展示的是给图片添加新元素的应用。具体原理是判断哪些文本特征是原文本中存在，将将那些文本特征对应的注意力图插入到新文本的注意力图中。一个具体的例子是，原文本是“A cute cat”。我想给小猫添加一件衣服，就将文本修改为“A cute cat，with clothes”，可以发现在构图不变情况下小猫穿上了衣服。也可以给小猫添加一个画风，比如说印象主义。也可以给小猫添加一个环境背景，比如说在街道上。

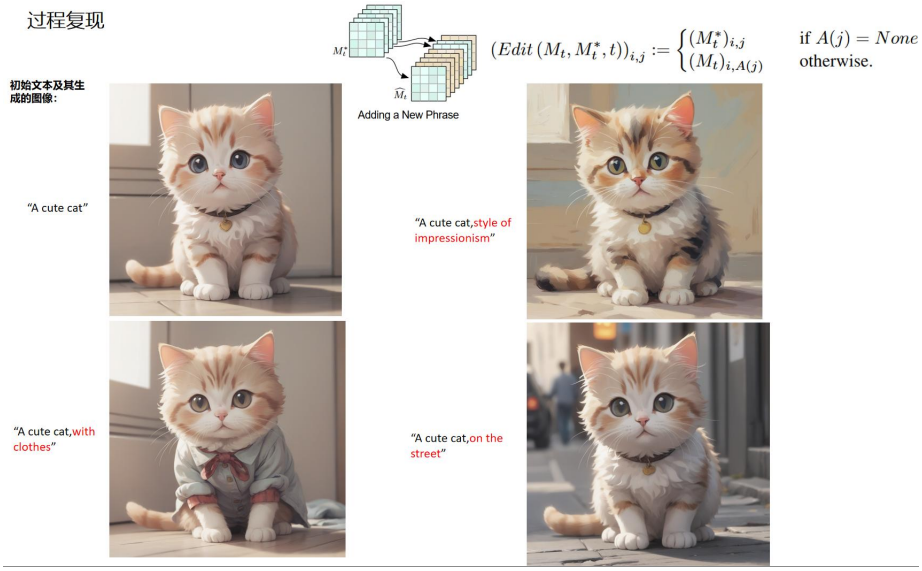


图 7 给图片添加新元素

参考文献

[1] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626.

[2] Avrahami, O., Lischinski, D., & Fried, O. (2022). Blended diffusion for text-driven editing of natural images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18208-18218).

[3] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 35, 36479-36494.

[4] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., ... & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.

[5] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[6] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[8] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.