Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation

摘要

这篇论文提出了一种名为 “Plug-and-Play Diffusion Features” 的新框架，用于文本驱动的图像到图像翻译。核心目标是利用预训练的文本-图像扩散模型，在不需要额外训练或微调的情况下，通过注入指导图像的特征，实现高质量的结构保留和语义翻译。文本生成图像模型的快速发展虽然可以生成高质量的图像，但在用户可控性方面存在限制，尤其是在需要控制图像结构的场景中，例如品牌设计或数字艺术。作者发现预训练扩散模型内部的中间特征和自注意力结构可以有效捕获图像的空间信息，通过修改这些特征实现对生成结构的精细控制。具体来说，他们提取指导图像的空间特征和自注意力图，将其直接注入目标图像生成过程。这种方法无需额外训练，既适用于真实图像，也适用于从文本生成的图像。论文通过定量和定性分析，证明该方法在不同任务中的优越性，包括草图、粗略绘图到真实图像的翻译、物体类别和外观的变化以及光照与颜色的调整。

关键词： 扩散模型；自注意力；无需训练；高质量生成；结构保留

1 引言

近年来，大规模文本驱动的图像生成模型（如扩散模型）的发展，极大地推动了生成式人工智能的能力。这些模型可以通过自然语言生成丰富多样的图像，表现复杂的视觉概念。然而，这些模型在真实场景内容创作中的应用仍存在一个核心挑战，即如何为用户提供对生成内容的细粒度控制能力。当前的文本驱动模型主要依赖输入的文本描述，用户无法对生成图像的结构和语义布局进行直接控制。这种局限性限制了其在品牌设计、数字艺术和营销等实际任务中的应用。

为解决上述问题，已有的研究尝试通过训练新的模型或对现有模型进行微调，以结合用户提供的引导信号（如掩模或分割图像）来控制生成结果。然而，这类方法通常需要大量计算资源，以及大规模的文本-图像训练数据，且只能针对特定的输入类型。在此基础上，论文提出了一个更加通用且无需训练的方法，旨在实现多样化的图像到图像翻译任务。这种方法充分利用了预训练的文本-图像扩散模型的内部表征，直接在生成过程中操作中间空间特征和自注意力，从而实现对图像结构和语义的控制。

2 相关工作

本文将与文本驱动的图像生成和图像到图像翻译相关的工作进行分类综述，按照当前主流方法的发展路径，主要分为传统图像到图像翻译方法、基于生成模型的文本驱动图像编辑方法，以及扩散模型在图像生成领域的应用和扩展。

2.1 传统图像到图像翻译方法

图像到图像翻译（Image-to-Image Translation, I2I）旨在实现从源域图像到目标域图像的映射，同时保留输入图像的域不变特性，例如对象的结构或场景布局。从传统方法到现代数据驱动方法，许多视觉问题被构建为 I2I 任务并得以解决。早期的深度学习方法提出了各种基于生成对抗网络（GAN）的框架，以鼓励输出图像符合目标域的分布。然而，这些方法需要源域和目标域的样本图像数据集，并且通常需要为每个翻译任务（例如，从马到斑马、从白天到夜晚、从夏天到冬天）从头开始训练。

其他方法利用预训练的 GAN，通过在其潜在空间中执行翻译来解决这一问题。一些方法还考虑了零样本 I2I 的任务，即通过单对源目标图像对训练生成器。此外，随着无条件图像扩散模型的出现，许多方法被提出以适配或扩展它们用于各种 I2I 任务。

本文关注文本驱动的图像到图像翻译任务，其中目标域不是通过图像数据集指定的，而是通过目标文本提示指定的。pnp 的方法是零样本的，不需要训练，并适用于多种 I2I 任务。

2.2 扩散模型在图像生成中的应用

扩散模型（Diffusion Models）是一类基于逐步去噪的生成式模型，通过反向模拟高斯噪声的去噪过程生成高质量图像。这些模型近年来在图像生成任务中展现出了卓越的表现，尤其是在图像到图像翻译和语义控制方面。扩散模型的核心思想包括两个过程：前向扩散过程和反向生成过程。前向过程将原始图像逐步添加高斯噪声，最终得到一个完全随机的高斯噪声图像，公式为：

$$x_t = \sqrt{\alpha_t} \cdot x_0 + \sqrt{1 - \alpha_t} \cdot z$$

其中，x_0 是原始图像，$z \sim N(0, I)$ 是标准高斯噪声，α_t 是时间步 t 对应的噪声控制参数。反向过程通过训练一个神经网络 $\epsilon_\theta(x_t, t)$ 来预测每个时间步中的噪声，从而逐步还原无噪图像。具体来说，每一步的去噪过程为：

$$x_{t-1} = x_t - \epsilon_\theta(x_t, t)$$

这一过程最终生成一个与原始图像分布一致的图像。本研究利用了一个预训练的文本条件潜变量扩散模型（Latent Diffusion Model，简称 LDM），也被称为 Stable Diffusion。该扩散过程应用于预训练图像自动编码器的潜在空间中。模型基于一个 U-Net 架构，并受指导提示 P 的条件约束。U-Net 的各层由残差块、自注意力块和交叉注意力块组成，如图 2(b) 所示。残差块将上一层 $l-1$ 的图像特征 ϕ_t^{l-1} 进行卷积，以生成中间特征 f_t^l。

在自注意力块中，特征被投影为查询 q_t^l、键 k_t^l 和值 v_t^l，其输出由以下公式给出：

$$f_t^l = A_t^l v_t^l, \quad \text{其中} \quad A_t^l = \text{Softmax}\left(q_t^l k_t^{lT}\right)$$

这一步操作允许图像特征之间的长距离交互。最后，交叉注意力在空间图像特征与文本提示 P 的词嵌入之间计算完成。

2.3 基于语言模型的图像编辑方法

随着语言-视觉模型的快速发展，大量方法被提出用于执行各种类型的文本驱动图像编辑。这些方法尝试结合 CLIP [4] 模型，它提供了一个丰富且强大的图像-文本联合嵌入空间，并与预训练的无条件图像生成器（如 GAN 或扩散模型）相结合。例如，DiffusionCLIP 通过微调扩散模型来实现文本驱动的图像操作。与之类似，其他方法利用 CLIP 和语义损失来指导扩散过程以实现图像到图像翻译任务。Text2LIVE 专注于编辑真实世界图像中的对象外观，针对单一图像-文本对训练生成器，而无需额外的训练数据，从而避免了传统生成器在保持原始内容的高保真度和满足目标编辑之间的权衡。

尽管这些方法在文本驱动的语义编辑方面表现出色，但在仅通过视觉数据学习的生成式先验（通常是特定领域或 ImageNet 数据）与 CLIP 学习的更丰富文本-图像指导信号之间仍存在差距。最近，文本生成图像模型通过在训练期间直接将图像生成过程与文本条件绑定，成功缩小了这一差距。这些模型在从文本生成高质量和多样化图像方面表现出了前所未有的能力，可以捕获复杂的视觉概念（如对象交互、几何形状或构图）。然而，这些模型对生成内容的控制能力仍然有限。这种限制引发了对开发能够应用这些非约束性文本生成图像模型来实现内容控制方法的强烈兴趣。

3 本文方法

3.1 本文方法概述

本文的目标是给定一个指导图像 I^G 和一个目标文本提示 P，生成一张新的图像 I^*。生成的图像需要符合目标文本的语义 P，同时保留指导图像 I^G 的结构和语义布局。本方法基于 Stable Diffusion 预训练的固定文本-图像潜变量扩散模型（Latent Diffusion Model, LDM），模型核心是一个 U-Net 架构，用于在生成过程中结合指导图像和目标文本的特征。在生成过程中，通过操控模型的空间特征可以实现对生成结构的细粒度控制，图的插入如图 1所示。

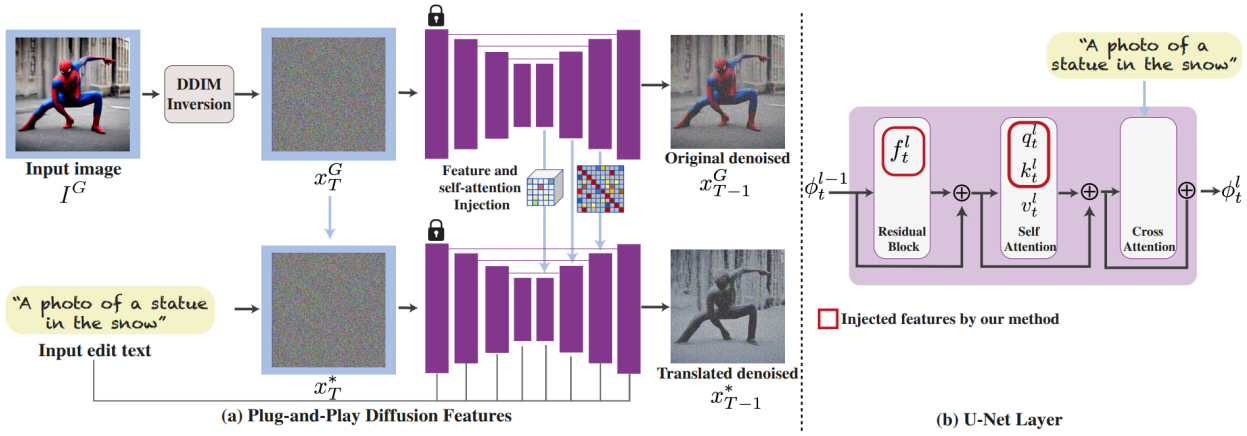

图 1. Plug-and-Play Diffusion Feature 方法流程图

3.2 特征提取模块

给定输入图像 I^G，通过 DDIM [5] 反向扩散过程将其映射到扩散模型的潜变量空间，得到初始潜变量表示 x_T^G。这一步骤将真实图像嵌入到模型的潜在空间中。在扩散模型生成的中间阶段，提取指导图像 x_T^G 的中间空间特征 f_t^l。这些特征在扩散模型的 U-Net 架构中由残差块（Residual Block）生成，编码了输入图像的局部语义信息和结构信息。图的插入如图 2所示。

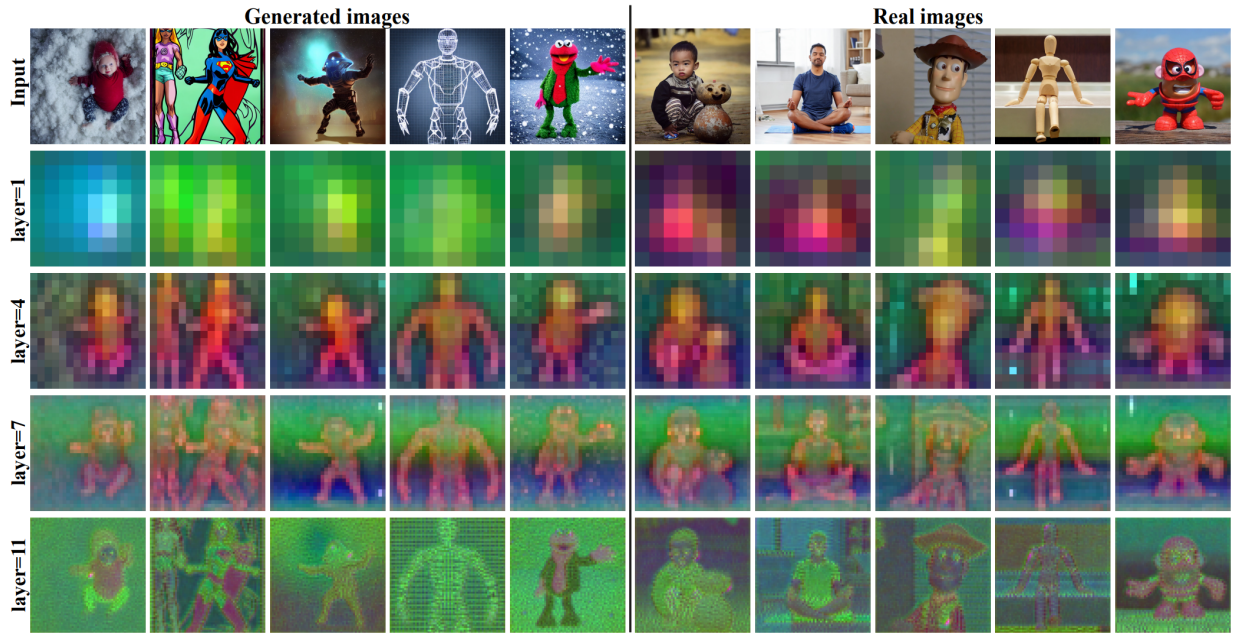

图 2. 对所有图像提取的特征应用主成分分析（PCA），并可视化了前三个主要成分。

从指导图像中获取自注意力矩阵 A_t^l，它反映了图像中不同位置之间的关联性（例如形状和布局的结构信息）。这一矩阵由 U-Net 的自注意力模块生成。提取的空间特征和自注意力矩阵被设计为多层次的表征，能够细粒度地捕捉图像的语义和结构信息。该模块的目标是从输入指导图像中提取关键特征（空间特征和自注意力矩阵），为后续的特征注入过程提供指导。这些提取的特征将被直接注入到目标图像的生成过程中，以确保生成的图像既符合目标文本提示的语义要求，又能保留输入图像的结构和布

3.3 特征注入模块

令 x_T^G 为通过 DDIM 反向扩散从 I^G 映射得到的初始噪声。对于目标文本提示 P，翻译图像 I^* 的生成使用相同的初始噪声，即 $x_T^* = x_T^G$。在反向生成过程的每一步 t，pnp 从去噪步骤 $z_{t-1}^G = \epsilon_\theta(x_t^G, \emptyset, t)$ 中提取指导特征 $\{f_t^l\}$。这些特征随后被注入到 I^* 的生成过程中，即在 x_t^* 的去噪步骤中，pnp 将生成的特征 $\{f_t^{*l}\}$ 用 $\{f_t^l\}$ 替代。该操作表示为：

$$z_{t-1}^* = \hat{\epsilon}_\theta(x_t^*, P, t; \{f_t^l\})$$

其中，$\hat{\epsilon}_\theta(\cdot; \{f_t^l\})$ 表示注入特征 $\{f_t^l\}$ 后修改的去噪步骤。如果未注入特征，则 $\hat{\epsilon}_\theta(x_t^*, P, t; \emptyset) = \epsilon_\theta(x_t^*, P, t)$。当在更深层注入特征时，结构得以保留，但输入外观的信息（例如红色 T 恤或蓝

色牛仔裤的阴影）泄露到生成图像中（如第 4-11 层）。为了在保留 I^G 的结构和减少其外观影响之间取得更好的平衡，该过程不修改深层特征，而是利用自注意力层。

3.4 自注意力模块

自注意力模块在将空间特征线性投影为查询和键后，计算它们之间的相似性 A_t^l。这些相似性与自相似性（self-similarity）的概念紧密相关，该概念已被广泛应用于经典和现代工作中设计结构描述符。论文中通过调整注意力矩阵 A_t^l 来实现对生成内容的精细控制。

图 3 展示了给定图像的注意力矩阵 A_t^l 的主要成分。可以看到，在浅层，注意力与图像的语义布局对齐，按照语义部分对区域进行分组。随着层数的增加，注意力捕获了更高频率的信息。 3所示。

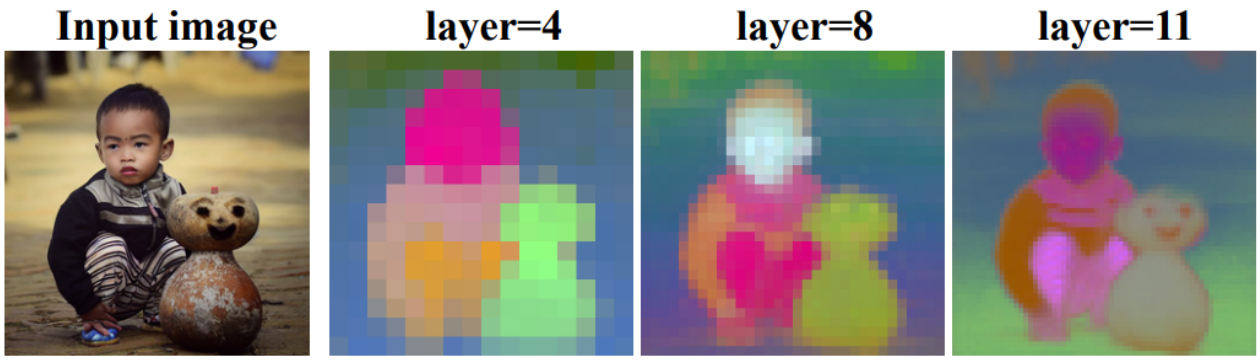

图 3. Self-attention visualization。

4 复现细节

4.1 与已有开源代码对比

代码参考了两份 pnp 的代码版本，一份是原论文中的 pnp 版本，github 链接为 https://github.com/Mi and-play，还有一份是基于 diffusers 库实现的版本，github 链接为 https://github.com/MichalGeyer/pnp-diffusers，这篇工作的复现参考了这两份代码版本，基于 diffusers 库的版本不需要手动下载模型，直接参考 diffusers 库的使用方法可以直接从 huggingface 上下载模型。

4.2 实验环境搭建

pnp 是基于 Stable Diffusion 实现的，首先需要下载 Stable Diffusion 库所需要的依赖，在 Stable Diffuison 的 github 官方项目地址下可以找到 environment.yaml 文件，从这个文件进行依赖项的下载，然后 diffusers 库版本的 pnp 中的 requirement.txt 文件中还提供额外的依赖项，从这个 txt 文件中下来其他的依赖项。

4.3 界面分析与使用说明

首先用执行 process.py 文件，需要提供目标文件的文件路径，以及对目标文件进行 DDIM 反演的提示词，然后图像反演的 latents 被保存到 output 文件夹，之后执行 pnp.py 文件，提

供目标文件 latens 特征的文件夹路径，进行 pnp 操作，提示词可以在命令行中进行调整。

4.4 创新点

pnp 论文中主要实现了对原始图像结构特征的保留，生成的图像在结构上可以更好地跟原始图像的结构进行对齐，pnp 主要是对解码器的残差块的特征，以及自注意力模块的 q,k 注入到目标图像的生成过程当中，为了更好地保存目标图像相对于源图像的结构，创新点参考 masactrl，将源图像与目标图像的 cross attention 层计算出一个 attention mask，其中该 mask 中存有丰富的目标图像的结构信息，将该信息注入到目标图像的生成过程，可以更好地保存源图像的结构。

5 实验结果分析

5.1 跟现有方法进行比较

与当前技术水平的基准方法进行比较，这些方法可用于多样化的文本驱动的图像到图像 (I2I) 任务，包括：SDEdit [3]：在三个不同的噪声水平下运行；P2P (Prompt-to-Prompt) [2]；DiffuseIT；VQGAN-CLIP [1] 。Text2LIVE；FlexIT；DiffusionCLIP。需要注意的是，P2P 需要一个与目标文本对齐的源文本提示。因此，pnp 在 ImageNet-R-TI2I 基准上对 P2P 进行了定性和定量比较，该基准通过提供的标签自动创建了源和目标提示的对齐。对于真实的引导图像，应用了 DDIM [5] 反演来与源文本对齐。

图 4 显示了 pnp 与基准方法的样本结果比较。如图所示，pnp 能够成功地将多样化的输入进行翻译，适用于真实和生成的引导图像。在所有情况下，pnp 的结果都表现出对引导布局的高度保留和对目标提示的高保真度。这与 SDEdit 形成对比，后者在以下两方面存在内在权衡：当噪声水平较低时，引导结构得到了很好的保留，但几乎没有改变外观；当噪声水平较高时，可以实现更大的外观变化，但引导结构会受到破坏。VQGAN-CLIP 展现了类似的行为，但整体图像质量较低。同样，DiffuseIT 对引导形状具有较高的保真度，但对外观的变化有限。与 P2P 方法相比，可以看出其在生成的引导图像上的结果（前三行）显示了对目标文本的高保真度，但对布局的保留较为粗略。例如，第一行结果中的鸭子数量不同，或者第二行中鼠标形状的偏离。此外，当应用于真实图像时，他们的方法在偏离引导图像外观并满足目标编辑方面存在困难（第 4-8 行）。这是因为他们在采样时使用源文本对 DDIM 反演进行了约束，需要使用较低的引导尺度。而相比之下，pnp 的方法使用空文本提示进行了 DDIM 反演，这使能够在生成时使用任意的引导尺度或提示。 4所示。

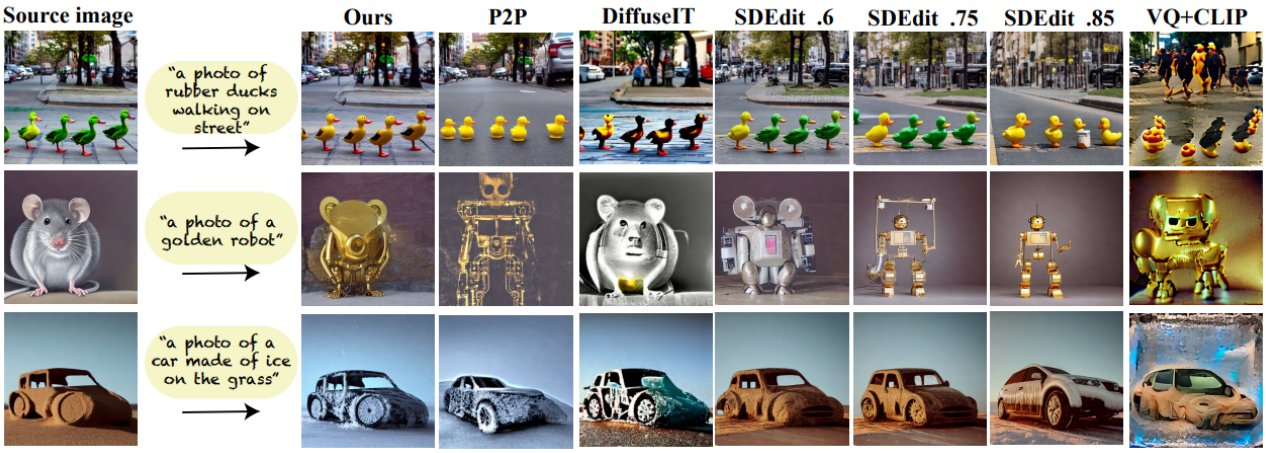

图 4. 与现有方法进行对比

5.2 baseline 效果展示

通过测量 CLIP 余弦相似度（值越高越好）和 DINO-ViT 自相似距离（值越低越好）[6]，分别量化文本保真度和结构保留情况。我们在三个基准上报告这些指标：(a) Wild-TI2I，其中包含我们方法的消融实验；(b) ImageNet-R-TI2I；以及 (c) Generated-ImageNet-R-TI2I。需要注意的是，由于提示词限制，只能在 (b) 和 (c) 中与 P2P 进行对比。所有基线方法在同时实现低结构距离和高 CLIP 分数方面都存在困难。pnp 方法在所有基准上展现了这两个方面的更好平衡。图 5所示。

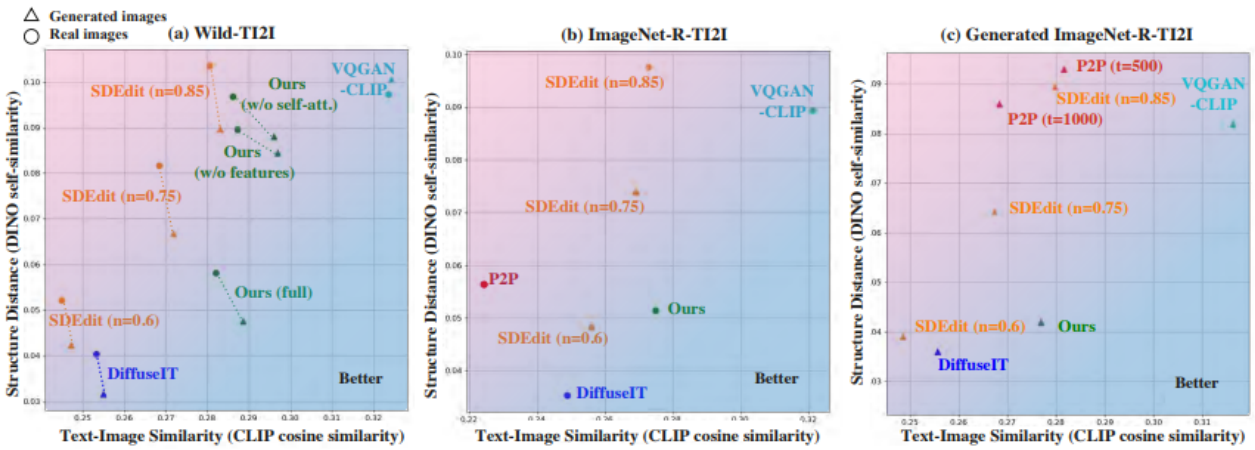

图 5. 与现有方法进行对比

6 总结与展望

论文深入研究了扩散模型中的中间空间特性，发现这些特性可以通过简单操作实现对生成结构的细粒度控制。通过从指导图像中提取空间特征及其自注意力信息，并在生成过程中直接注入这些特征，框架能够在保持图像语义结构的同时，根据目标文本生成新图像。方法适用于多种任务，如从草图生成现实图像、更改对象类别与外观、以及全局属性（如光照与颜色）的调整。方法依赖输入图像与目标文本之间的语义关联，因此对于没有明显语义关联

的输入（如随机颜色分割图）表现较差。DDIM 反演过程中，对于纹理较少的图像可能会出现低频信息泄漏，影响生成效果。这项工作展示了预训练文本到图像扩散模型的潜力，并为未来研究提供了重要启示。

参考文献

[1] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022.

[2] 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.

[3] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[4] 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.

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[6] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022.