Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(3), 832-843
Published Online May 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.153077

基于多位置轻量适配器的文生图细节优化方法

马宇航

北京建筑大学理学院，北京

收稿日期：2026年3月23日；录用日期：2026年4月13日；发布日期：2026年5月25日

摘　要

目前，人工智能技术的快速发展带来了庞大的应用需求，图像生成已成为计算机视觉领域的重要研究方向。扩散模型凭借其优异的生成性能和稳定可控的训练过程，成为当前生成高质量图像的主流框架。然而，扩散模型仍面临细节保真度不足的问题，这限制了其在定制化应用中的灵活性与可控性。因此，如何在维持模型轻量化的同时有效提升生成质量与鲁棒性，成为扩散模型高效适配的核心挑战。针对通用文本到图像生成中存在的空间细节模糊与语义对齐偏差，本文设计了基于多位置轻量适配器的细节增强架构。该架构在残差块的输出端嵌入EBlock适配器，利用深度可分离卷积增强局部空间细节；在注意力层输出后接入DAT适配器，通过低秩映射对通道特征进行调制以提升文本－图像语义对齐。两类适配器均采用零初始化与渐进式激活机制，并结合知识蒸馏实现原始模型知识保留。该架构仅引入约0.35%的额外参数，在COCO2017数据集上Laplacian方差和FID分别为0.032和17.8；在Flickr30k数据集上FID进一步降至16.3，显著提升了生成图像的细节质量与语义一致性。

关键词

扩散模型，轻量化适配器，图像生成，细节优化

Detail Enhancement Method for Text-to-Image Based on Multi-Position Lightweight Adapters

Yuhang Ma

College of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: March 23, 2026; accepted: April 13, 2026; published: May 25, 2026

Abstract

Currently, the rapid development of artificial intelligence technology has brought about substantial application demands, and image generation has become an important research direction in the field of

文章引用: 马宇航. 基于多位置轻量适配器的文生图细节优化方法[J]. 人工智能与机器人研究, 2026, 15(3): 832-843.
DOI: 10.12677/airr.2026.153077

computer vision. Diffusion models, with their excellent generation performance and stable and controllable training process, have become the mainstream framework for generating high-quality images. However, diffusion models still face the problem of insufficient detail fidelity, which limits their flexibility and controllability in customized applications. Therefore, how to effectively improve generation quality and robustness while maintaining model lightweightness has become the core challenge for efficient adaptation of diffusion models. Addressing the spatial detail blurring and semantic alignment deviation in general text-to-image generation, this paper designs a detail enhancement architecture based on multi-position lightweight adapters. This architecture embeds an EBlock adapter at the output of the residual block to enhance local spatial details using depthwise separable convolution; it also incorporates a DAT adapter after the attention layer output to modulate channel features through low-rank mapping to improve text-image semantic alignment. Both adapters adopt zero initialization and progressive activation mechanisms, and combine knowledge distillation to preserve the knowledge of the original model. This architecture introduces only about 0.35% additional parameters, achieving a Laplacian variance and FID of 0.032 and 17.8 on the COCO2017 dataset, respectively; on the Flickr30k dataset, the FID further decreases to 16.3, significantly improving the detail quality and semantic consistency of generated images.

Keywords

Diffusion Model, Lightweight Adapters, Image Generation, Detail Enhancement

1. 引言

图像生成是计算机视觉与人工智能交叉领域的重要方向，其核心目标是能够生成符合视觉规律的合理的真实内容。该领域的技术进展主要受到生成式模型架构创新的推动。早期研究主要基于变分自编码器 VAE [1]等概率生成模型，但所生成图像有着清晰度不足的问题。生成对抗网络 GAN [2]的提出标志着重要进展，通过对抗训练机制显著提升了生成结果的视觉真实感，然而其训练的不稳定与模式崩溃问题也成为研究焦点。近年来，扩散模型通过渐进式的去噪过程，在生成质量与训练稳定性之间取得了比较好的平衡，已经成为当前高质量图像生成的主流模型。随着通用图像生成技术的逐渐发展，如何实现精准可控的内容定制成为越来越重要的研究议题，而其中对于细节优化逐渐成为一个重要的研究方向。其所追求的细节重建，不仅限于对特定主体的外观与色彩的匹配，更在于对其独特身份的高频视觉特征的准确还原与一致保持。例如在人物肖像生成中，需捕捉特定的面部特征(如痣，胡须等)、发丝纹理乃至表情肌理；在对于艺术风格的个性化生成中，则需再现色彩过渡，风格特点等细节。这种对细节的极高要求，使得对生成模型提出了更改的要求。通用生成模型基于海量图像或图像文本对进行训练，倾向于学习生成视觉合理结果，往往会忽视掉具体细节。因此，该方向的核心研究价值在于探索如何在有限数据与计算资源的约束下，实现用户概念的高效的表征学习、适配与可控合成，从而缩小生成模型的在通用和专用之间的差距，并满足对视觉细节保真度的严格需求。

2. 相关工作

在通用文生图模型中实现精细化控制，是提升其生成结果准确性的关键。基于适配器的方法通过冻结预训练模型的参数，并在其外部附加轻量级的可训练控制网络，将额外指导信号注入生成过程，在此方向上贡献显著。

现有研究主要集中于利用参考图像作为控制条件。ControlNet [3]通过复制预训练 U-Net 的架构作为可训练的副本，并采用零卷积层进行连接，构建了一个能够将边缘图、深度图、人体姿态图等多种空间条件输入的通用控制框架。IP-Adapter [4]设计了独立的图像提示编码器，通过将参考图像特征与文本嵌入在特征空间中对齐，实现了无需修改文本提示的图像内容与风格驱动。InstantID [5]专注于保持人脸身份细节，通过联合编码面部图像与文本特征，并嵌入轻量级身份适配模块，在无需微调主体模型的情况下实现了高保真人脸生成。

然而，在仅依赖纯文本描述的场景下，实现对复杂精细细节的生成面临更大挑战。这要求模型具备对自然语言更深层、更细粒度的理解能力，能够将包含多重属性与空间关系的描述准确映射为相应视觉特征。当前多数基于适配器的可控生成工作依赖于参考图像提供的视觉先验，对于仅通过文本细化的控制研究相对不足。而另一种方法是引入大语言模型 LLM，通过 LLM 强大的语言处理能力来精确理解文本提示，更好地指导图像生成。例如，SUR-adapter [6]在推理阶段虽然保持了参数高效与轻量的特性，但其训练与语义建模过程依赖于大语言模型，因此并未构成一种完全独立的轻量适配器解决方案。因此探索不依赖外部大规模模型、仅生成机制来实现细节控制的方法，是一项至关重要且富有挑战性的前沿课题。

本文针对个性化生成和通用文生图任务中空间细节模糊与语义对齐偏差，设计 EBlock 与 DAT 双路径轻量增强的适配器架构；同时通过渐进式训练与知识锚定相结合，在轻量级适配器架构下平衡新概念学习与原始知识保留。具体来说，设计 EBlock 与 DAT 双路径轻量增强架构，旨在提升空间细节与语义对齐精度。EBlock 适配器通过深度可分离卷积对残差块输出进行空间细节调制，以增强高频纹理而不破坏原有语义；DAT 适配器在交叉注意力层后引入低秩分解结构来优化文本－图像特征对齐。两者均采用零初始化和渐进式激活策略，通过两阶段优化逐步调整适配器影响强度，最终在仅引入极少的参数的情况下，在 COCO2017 与 Flickr30k 数据集上显著改善了 Laplacian 方差、边缘密度与 FID 等指标，证明了我们的方法在实现细节控制方面取得了重大进步。

3. 模型架构

3.1. Eblock 适配器和 DAT 适配器

3.1.1. Eblock 适配器

为增强扩散模型对空间细节的建模能力，我们参考了 Feijoo 与 Benito 等人[7]针对频域细节优化所提出的 EBlock 结构设计思路，并在此基础上进行调整与拓展，将其整合并适配至 Stable Diffusion 的 UNet 架构中，实现在保持语义一致性的同时有效提升生成图像的细节质量。如图 1 所示，我们在 UNet 的残差块中引入轻量级适配器，其结构基于深度可分离卷积，由深度卷积[8]与逐点卷积[9]两部分串联构成，以较低的计算代价实现对空间信息的有效增强。

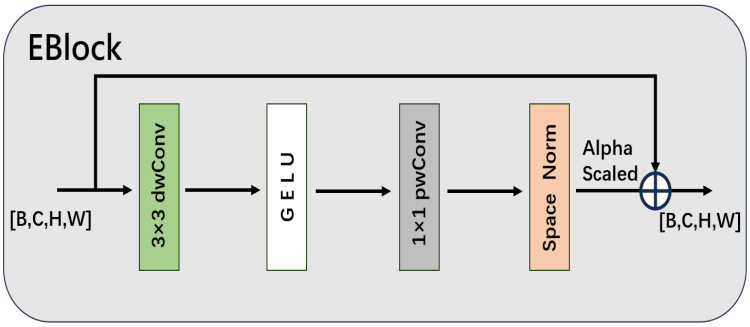

Figure 1. Schematic diagram of Eblock structure
图 1. Eblock 结构示意图

具体而言，深度卷积对输入特征图的每个通道独立进行空间卷积运算，采用 3 × 3 卷积核，仅在同一通道内提取局部特征，不进行跨通道的信息融合。逐点卷积则使用 1 × 1 卷积核，在不改变特征图空间尺寸的前提下，对不同通道的特征进行线性组合与重组，其作用等效于对每个空间位置的特征向量执行通道维度的变换。该设计先通过深度卷积捕获空间局部模式，再借助逐点卷积建立通道间的语义关联，从而在计算效率与特征表达能力之间取得良好平衡。EBlock 适配器的定义为：

$$h=\frac{\left(\mathrm{Conv}_{1\times1}\left(\mathrm{GELU}\left(\mathrm{DepthwiseConv}_{3\times3}\left(x\right)\right)\right)-\mu\right)}{\mathrm{RMS}\left(h\right)+\epsilon} \tag{1}$$

最后的输出为：

$$y=x+\alpha\cdot h \tag{2}$$

其中，x 是输入特征，μ 是通道均值，Space Norm 是空间归一化，α 是可学习的门控参数，由 $\alpha=a_{\max}\cdot\sigma\left(a_{\mathrm{raw}}\right)$ 计算得出。这种设计使得残差连接可以保证梯度的平稳传递；其次，空间归一化有效抑制了特征分布偏移；最后，门控参数 α 动态控制增强强度，从而避免对原始特征造成过度干扰。

3.1.2. DAT 适配器

为了解决注意力层中的文本–图像语义对齐问题，我们参考了 Kai Han 和 Yunhe Wang 等人[10]的核心思路，提出了 DAT 适配器，通过在交叉注意力层输出后插入轻量级结构实现参数高效优化。如图 2 所示，首先对注意力输出进行层归一化[11]处理，随后经由低秩投影[12]与非线性激活[13]得到中间表示，我们设计特征维度的零均值通道归一化，通过中心化消除直流分量，再以 RMS 约束特征幅值，有效抑制分布偏移。通过特征维度的归一化稳定特征分布，最终通过可学习缩放系数与原始输出进行残差融合。DAT 适配器定义为：

$$h=\mathrm{LayerNorm}\left(y\right)\cdot W_1+b_1\xrightarrow{\mathrm{GELU}}\frac{\left(W_2\cdot h+b_2\right)-\mu}{\mathrm{RMS}\left(h\right)+\epsilon} \tag{3}$$

最后的输出为：

$$y'=y+\alpha\cdot h \tag{4}$$

其中 y 是原始注意力输出，$W_1\in\mathbb{R}^{d\times r}$、$W_2\in\mathbb{R}^{r\times d}$ 是低秩权重矩阵，μ 与 ϵ 分别为均值项与数值稳定常数。归一化操作沿特征维度执行，有效抑制输出分布偏移，确保增强过程与原始模型的语义一致性。门控缩放参数 α 是可学习的门控参数同样采用 sigmoid 约束。

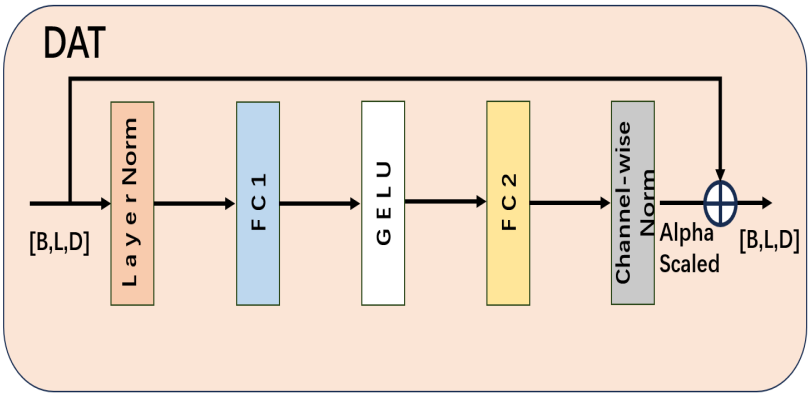

Figure 2. Schematic diagram of DAT structure
图 2. DAT 结构示意图

3.2. 基于多位置轻量适配器网络架构

本文提出的整体架构如图 3 所示，以适配器的方式无缝集成至 Stable Diffusion 1.5 的 UNet 架构中。EBlock 适配器被嵌入于残差卷积块输出端，用于增强局部空间特征的细节表达能力；DAT 适配器则主要部署在深层及中层的交叉注意力模块之后，因为这些层级所处理的特征具有更高的语义抽象度，对文本－图像对齐起到关键作用。在训练过程中，原始 UNet 的所有参数均保持冻结，仅对新增的适配器模块进行优化，从而极大降低了可训练参数量。此外，通过 identity_zero 初始化机制[14]与可学习的门控系数α协同作用，有效确保了训练初始阶段的输出稳定性，避免了因随机初始化可能导致的训练振荡或性能退化。

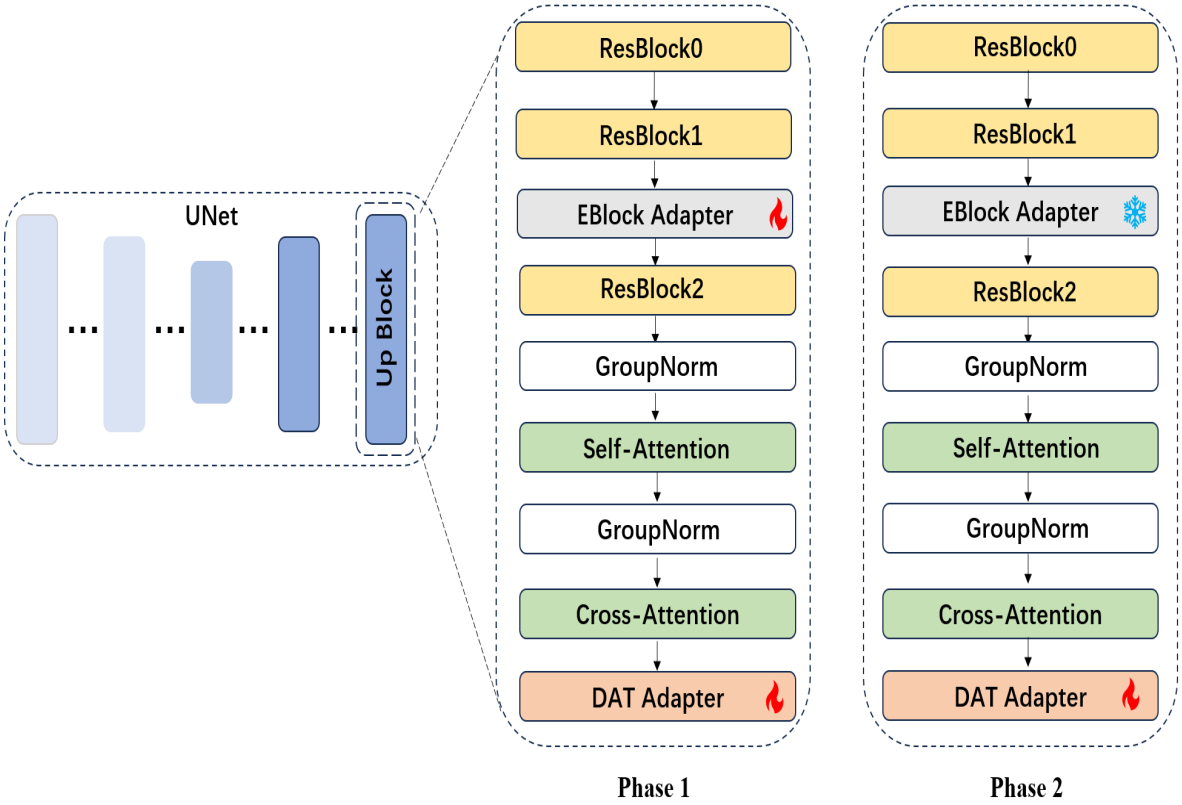

Figure 3. Network architecture diagram based on multi-location lightweight adapters
图 3. 基于多位置轻量适配器网络架构图

3.3. 渐进式训练与优化策略

3.3.1. 渐进式激活机制

如果在训练过程中同时优化所有适配器的参数，可能引发训练不稳定，例如出现收敛缓慢或损失振荡等现象。为此，我们设计了一种渐进式激活机制，在训练初期逐渐增加 DAT 的训练强度。该机制通过一个随时间变化的动态缩放因子 s，控制 DAT 门控参数 α 的实际生效强度，从而实现从无干扰到逐步增强的平滑过渡。缩放因子 s 具体的定义为：

$$s(t) = s_{\text{start}} + (s_{\text{end}} - s_{\text{start}}) \cdot \frac{t}{T_{\text{phaseA}}} \tag{5}$$

其中t是当前训练步数，T_{phaseA}是到达 A 阶段的训练步数(不和第一阶段有关)，$s_{\text{start}}=0.7$，$s_{\text{end}}=1.05$，这样的设置保证了训练前的一定步数内，适配器贡献被抑制至原始设计的 70%，使模型充分依赖预训练模型的生成能力，在后期提供 5%的增强补偿，抵消初期保守策略导致的收敛滞后。先验当$t<T_{\text{phaseA}}$时，$\alpha_{\text{eff}}=\alpha\cdot s(t)$；否则$\alpha_{\text{eff}}=\alpha$。这种设计使模型在初期依赖原始的生成先验，逐步过渡到增强模式，显著提升训练稳定性。

3.3.2. 两阶段优化策略

本文设计了一种兼顾训练效率和最终生成质量的两阶段优化策略：第一阶段同时训练 EBlock 和 DAT 适配器的所有参数，快速建立整体特征增强能力。进入第二阶段后，对 EBlock 的可训练参数进行冻结，只保留其门控参数进行更新，将训练重点放在 DAT 适配器上，进一步提高语义对齐和细节恢复的质量。具体地说，冻结操作包括：停止 EBlock 权重的梯度更新；在优化器中将相应参数组的学习率设置为零；并继续分别优化 EBlock 中的门控参数和 DAT 适配器的所有参数。该分阶段策略不仅加快了训练收敛速度，而且也提高了模型的表达能力。

3.3.3. 门控参数优化

门控参数α作为适配器中的核心调节变量，其作用是控制特征增强的强度。我们发现，若采用与主体网络相同的标准学习率(如 $1e^{-4}$)对α进行优化，其在训练初期的更新幅度往往过大，易引发优化过程的不稳定。因此，我们为α参数设计了专用的学习率缩放系数$k_{\alpha}=6.0$，以降低其相对更新速度，从而确保训练过程的平稳进行，具体定义为：

$$\text{lr}_{\alpha}=k_{\alpha}\times\text{base\_lr} \tag{6}$$

同时，为 EBlock 设置了专门的学习率缩放系数$k_{\text{eblock}}=0.25$，从而有效减缓其更新幅度。这种针对不同功能参数设计的差异化学习率调度策略，能够使各类型参数以各自适宜的速度进行优化，促进了整体模型的协调收敛，并最终显著提升了方法的综合性能。

3.4. 正则化机制与总损失函数

3.4.1. 正则化机制

为抑制过拟合并缓解概念漂移问题，我们整合了多种正则化策略：首先，引入分类器自由引导随机丢弃机制[15]，以固定概率将文本条件替换为空标记，从而增强模型在无条件生成时的鲁棒性；其次，在扩散过程中添加高斯噪声偏移，以提升低信噪比区域的特征学习质量；第三，采用梯度裁剪技术[16]，将梯度范数上限固定，以避免优化过程中的剧烈波动；最后，使用指数移动平均对模型参数进行平滑更新，衰减系数设置为 0.99，从而提高推理阶段的输出稳定性。

特别地，我们设计 alpha 参数的 L2 正则化：

$$\mathcal{L}_{\alpha\text{-reg}}=\lambda_{\alpha}\sum\alpha^{2},\lambda=10^{-4} \tag{7}$$

这防止门控参数过度增长，维持适配器的适度增强作用，避免对原始模型的过度干扰。

3.4.2. 知识锚定机制与总损失函数

为缓解适配器训练过程中可能引发的灾难性遗忘问题，本文引入知识蒸馏机制[17]。具体而言，我们在训练期间维护一个固定的教师模型，该模型由预训练模型 Stable Diffusion 1.5 初始化。通过计算学生模型与教师模型在相同输入下的输出差异，构建一项额外的约束损失$\mathcal{L}_{\text{anchor}}$，从而引导学生模型在习得新能力的同时，尽可能保留其原有的生成分布与语义理解能力。该机制有效避免了由于训练的原因造成模型

的先验生成能力被破坏的问题，具体定义为：

$$\mathcal{L}_{\text{anchor}} = \left\| y_{\text{student}} - y_{\text{teacher}} \right\|^2 \tag{8}$$

最终的总损失函数任务损失与知识锚定损失共同构成：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{anchor}} \cdot \mathcal{L}_{\text{anchor}} \tag{9}$$

其中 $\mathcal{L}_{\text{task}}$ 为 MSE 损失函数，$\lambda_{\text{anchor}} = 0.3$。这种设计有效平衡了新能力学习与原始知识保留，在保持细节增强的同时，确保基础生成能力不退化。

4. 实验结果

4.1. 实验配置

本实验基于 Stable Diffusion 1.5 进行。核心创新在于对 U-Net 架构的轻量化结构化增强，提出了两种轻量级的适配器分别为 EBlock 适配器，对残差块输出进行空间细节化增强；以及 DAT 适配器，在交叉注意力层后嵌入低秩分解结构，通过带门控的残差路径对注意力输出进行精细化校准。两种适配器均采用恒等初始化(identity-zero)确保训练起始阶段模型行为与原始 SD1.5 一致。

训练采用渐进式调度策略：针对通用文生图任务场景，在前 1200 步(占总训练步数的 30%)逐步提升 DAT 适配器的有效缩放因子(从 0.70 线性增长至 1.05)，实现平滑的过渡激活；3000 步后冻结 EBlock 的训练参数(仅保留 α 可训练)，并在最后 1000 步专注于优化训练 DAT 适配器参数。而对于个性化微调任务场景，前 600 步逐步提升 DAT 适配器的有效缩放因子，1500 步后冻结 Eblock 的训练参数，后 500 步训练 DAT。

优化器采用 AdamW，基础学习率 1×10^{-4}，并配合 800 步线性 warmup 与余弦退火衰减。

所有实验均在单张 NVIDIA A100-PCIE-40GB GPU (40 GB 显存)上进行。在通用任务中批大小(batch size)设为 8，个性化微调任务中批大小设为 1。通过梯度检查点、VAE slicing 以及 xFormers 高效注意力实现等技术进行显存优化，将峰值显存控制在 38 GB 以内。图像预处理仅包括 512 × 512 的随机裁剪与归一化，未引入复杂的几何或光度变换，以聚焦于评估适配器结构本身的泛化能力。

4.2. 定性实验

本小节系统评估所提出方法在通用文本到图像生成与个性化生成两大任务场景下的性能表现，通过定性对比实验验证多位置轻量适配架构的有效性。

如图 4 所示，针对通用文生图任务场景中，模型在面对多样化文本提示时展现出优秀的细节生成与语义对齐能力。例如提示“a ginger cat sunbathing on a windowsill”，生成图像不仅准确还原主体布局更在微观结构上呈现丰富细节，猫科动物的绒毛层次清晰可辨，符合生物特征；在“a butterfly resting on a blooming flower branch”蝴蝶的触角得到精细化表达；对于“Sunlight breaking through storm clouds over a wheat field”，模型能够渲染出光线透过云层时的体积感与明暗过渡。这种细节增强能力源于 EBlock 适配器对残差特征的空间调制：其深度可分离卷积结构有效捕获局部梯度变化，不破坏原始语义的前提下注入高频细节信号。

为进一步衡量本方法的相对优势，我们在图 5 中将本文与其他纯文生图领域的模型(无参考图像输入)进行横向对比，分别针对 SD1.4，SD1.5 和 SDXL 的基于 UNet 架构的扩散模型进行比较，可以看出，本文的模型在生成质量上，全面优于 SD1.4，在细节方面比 SD1.5 处理的更好，如对于文本提示“Waterfall cascading down cliffs”里瀑布遇到岩石的折射和在“Elephant walking through savanna”对于象牙的生成。而对于 SDXL 来说，我们的模型细节更加多样，空间布局更为细腻。

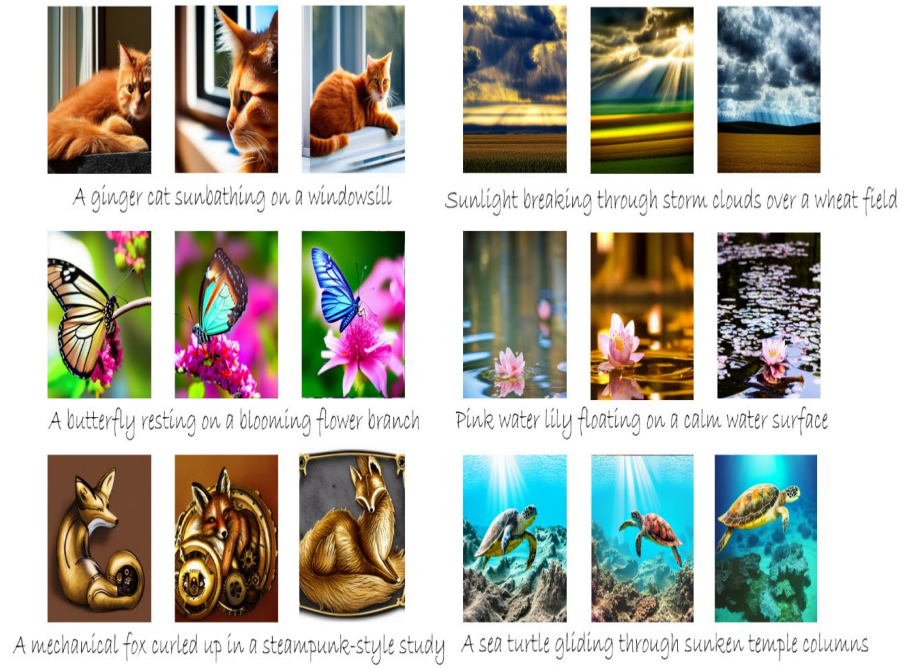

Figure 4. Generation results
图 4. 生成结果图

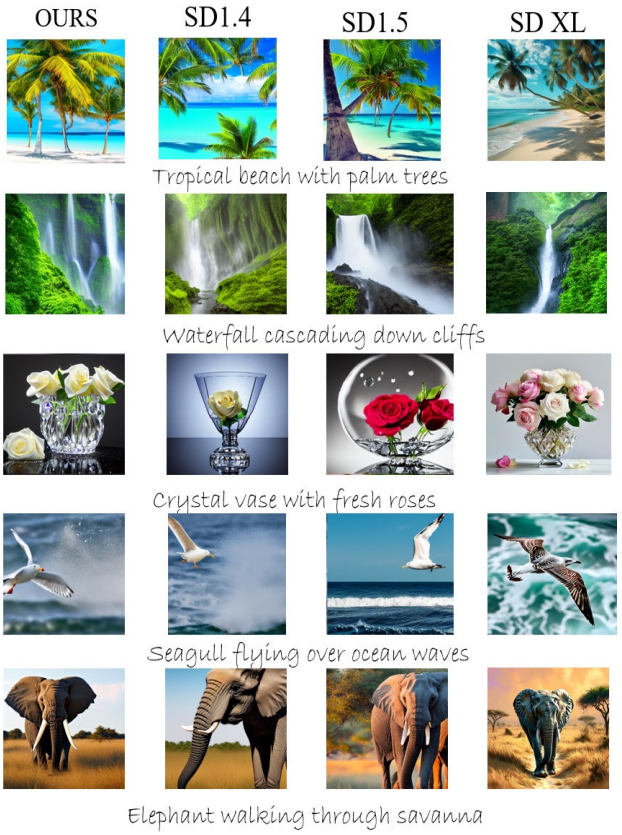

Figure 5. Comparison chart of generation results with different text-to-image models
图 5. 与不同文生图模型的生成结果对比图

4.3. 定量实验

为全面评估模型在通用文文生图任务上的性能，我们在 COCO2017 数据集上对所提方法与其他的基于 UNet 架构的扩散模型进行了定量比较。为确保公平对比，将 SDXL 生成的高分辨率图像统一降采样至 512 × 512 后计算各项指标。

如表 1 所示，本方法在 Laplacian Variance、Entropy 及 FID 三项关键指标上均优于对比的模型，说明我们的方法展现出卓越的细节丰富度、信息复杂度与生成质量。特别值得注意的是，在 FID 指标上，本章节提出的方法不仅显著优于基准模型 SD1.5，更以微弱优势超越了参数规模更大的 SDXL，验证了结构化适配策略在参数效率与生成质量平衡方面的有效性。在 Edge Density 指标上，本方法以 0.0679 略低于 SDXL 的 0.075，这反映了 SDXL 这种更大规模参数的模型在刻画复杂轮廓结构方面的优势。然而，本方法在 Laplacian Variance 指标上以 0.032 显著优于 SDXL 的指标 0.030，表明在纹理与微观细节的生成上更具优势。综合四项指标分析，本方法在保持与 SDXL 相当的边缘结构复杂度的同时，通过更丰富的高频细节表达(Laplacian Variance)和更优的全局感知质量(FID)，实现了整体生成质量的提升。

值得注意的是，相较于基准模型 SD1.5，本方法在所有评估指标上均取得全面领先：Laplacian Variance 提升 28.0%，Edge Density 提升 3.0%，Entropy 提升 0.4%，FID 降低 15.6%。这一结果充分验证了多位置的轻量化适配器架构在细节增强与语义保真之间的有效平衡，为扩散模型的性能提升提供了新的技术路径。

Table 1. Comparison of relevant indicators with different text-to-image models
表 1. 与不同文生图模型的相关指标对比

Model	Laplacian Variance	Edge Density	Entropy	FID
SD1.4	0.022	0.050	7.50	25.0
SD1.5	0.025	0.066	7.58	21.1
SDXL	0.030	0.075	7.60	18.0
OURS	0.032	0.068	7.61	17.8

为全面验证模型在不同视觉内容分布下的泛化性能，我们在两个具有显著差异的基准数据集上进行了系统性评估：COCO2017 (以自然场景和日常物体为主)与 FLICKR30k (侧重人物与场景交互)。如表 2 所示，本方法在两类数据集上均表现出优异的细节表达能力与生成质量。

Table 2. Comparison of model metrics across different datasets
表 2. 不同数据集下的模型指标对比

Dataset	Laplacian Variance	Edge Density	Entropy	FID
COCO2017	0.032	0.068	7.61	17.8
FLICKR30k	0.031	0.068	7.55	16.3

在 COCO2017 数据集上，模型取得了在 Laplacian Variance、Edge Density 与 Entropy 这几个指标上分别取得了 0.032，0.068 和 7.61 的优异表现，表明其在复杂自然场景中能够生成丰富细腻的高频细节。证实了本方法在保持语义一致性的同时，有效提升了生成图像的视觉质量。

在 FLICKR30k 数据集上，模型进一步展现出对人物与场景交互场景的适应性：Laplacian Variance (0.031)与 Edge Density (0.068)指标与 COCO2017 保持相近，表明细节建模能力具有跨数据集的稳定性；Entropy 指标略降至 7.55，这与人物场景构图相对规整的数据特性相符合；而 FID 指标显著降低至 16.3，

表明模型在人物生成任务中实现了更优的分布对齐能力。这一结果验证了我们的方法在不同内容域下的泛化适应性即 EBlock 适配器有效捕获细节特征，DAT 适配器则维持跨域语义一致性。

综合两项数据集的评估结果，本方法在 Laplacian Variance 与 Edge Density 指标上保持稳定(均值分别为 0.0315 与 0.068)，证明其细节建模能力具有数据集无关性即不受数据集分布差异的显著影响；而在 FID 指标上展现出数据集特异性(COCO2017:17.8, FLICKR30k:16.3)，表明模型对人物场景的表征更为精准。这种差异化可能与 DAT 适配器对交叉注意力流的精细化校准有关：在人物场景中，模型能够更准确地定位主体与环境的语义关联，从而生成更具视觉一致性的结果。

4.4. 消融实验

为系统验证 DAT 与 EBlock 模块对生成质量的独立贡献，我们在控制其他训练条件一致的前提下进行了消融实验。如表 3 所示，DAT 模块使 Laplacian Variance 提升 3.8% (0.026 → 0.027)，Edge Density 提升 6.7% (0.060 → 0.064)，FID 降低 4.8% (21.0 → 19.9)，表明其在优化整体感知质量方面的有效性。EBlock 模块使 Laplacian Variance 提升 4.8% (0.021 → 0.022)，Edge Density 提升 3.4% (0.058 → 0.060)，但 FID 略升 1.7% (24.5 → 24.9)，表明其在局部细节增强方面具有优势，但在全局语义一致性上需与 DAT 模块协同优化。

Table 3. Comparison of indicators between Eblock and DAT
表 3. Eblock 和 DAT 的指标对比

	Laplacian Variance	Edge Density	Entropy	FID
No DAT	0.026	0.060	7.54	21.0
NO Eblcok	0.021	0.058	7.31	24.5

对比“w/o DAT”与“with DAT”生成结果，DAT 模块在保持整体构图合理性的同时，显著改善了场景中的各个主体的语义关系与布局合理性。在“w/o DAT”中可以看到空间布局并不符合客观规律，比如灯塔在房子的下面，而在 with DAT 之中场景语义一致性得到体现，空间布局符合客观规律。

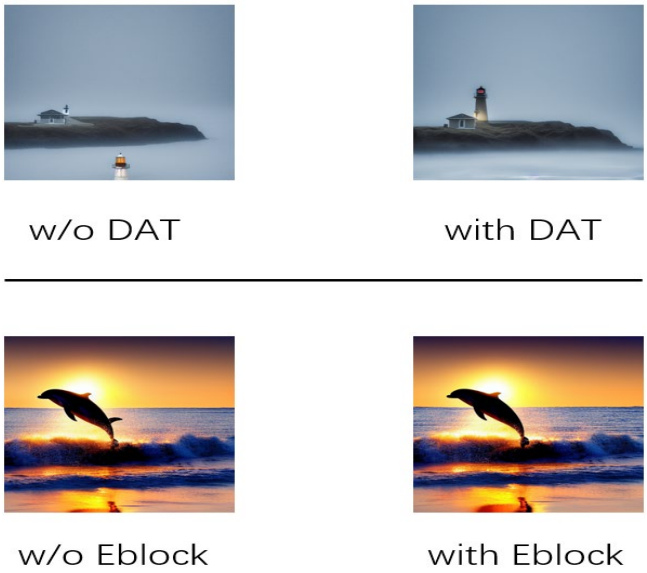

Figure 6. The influence of Eblock and DAT on the generated results
图 6. Eblock 与 DAT 对生成结果的影响

对比“w/oEBlock”与“with EBlock”生成结果，EBlock 模块在改善边缘结构与纹理细节方面展现出独特优势。在“w/o EBlock”中，生成的海豚多了一个鱼鳍，而在“with EBlock”没有生成，符合一般海豚的样子，对比可以看出 Eblock 在高频细节方面的优越性。详细对比如图 6 所示。

为探究 EBlock 模块在 UNet 架构中的最优部署位置，我们系统性地评估了不同残差块间隔策略对生成质量的影响。如图 7 所示，我们分别测试了三种部署方案：每个残差块后均插入 EBlock 模块(EBlock every 1)；每隔两个残差块插入(EBlock every 2)；每隔三个残差块插入(EBlock every 3)。

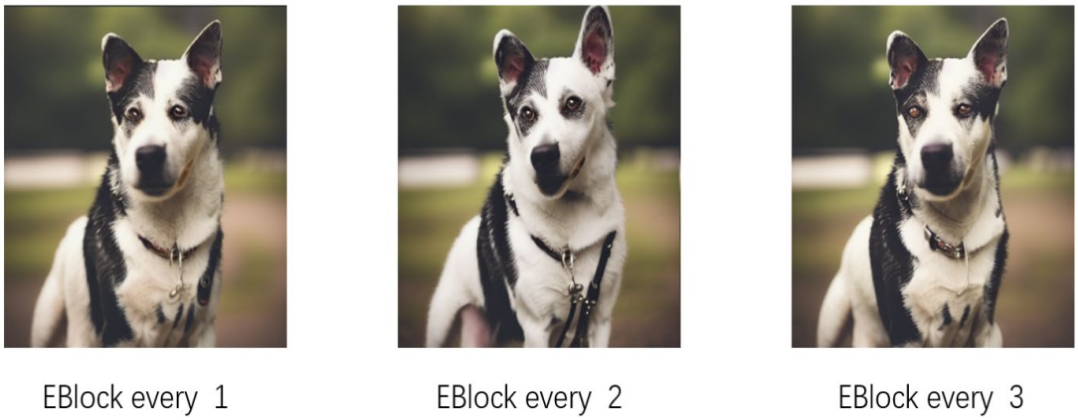

Figure 7. Comparison chart of generation results with different text-to-image models
图 7. EBlock every 对生成结果的影响

实验结果表明，EBlock every 2 配置在细节增强与语义保真度之间实现了最佳平衡：在保持主体结构完整性的前提下，模型能够有效增强边缘细节(如犬类毛发的层次感、皮质项圈的纹理表现)。相比之下，EBlock every 1 配置因过度增强高频细节而导致边缘结构过度锐化，产生非自然的纹理噪声；而 EBlock every 3 配置则因增强密度不足，无法充分提升关键区域的细节表现力。

5. 结论

本文提出了基于多位置的轻量化适配器的文生图细节优化方法，针对文生图扩散模型在通用生成存在的细节问题，重新设计并提出了两种全新的轻量化的适配器模块分别为 Eblock 和 DAT。EBlock 嵌入于残差块输出端，负责增强空间局部细节；DAT 则部署在交叉注意力层之后，用于提升文本－视觉特征的语义对齐精度。该方法在完全冻结原始模型主体参数的条件下，仅训练两种适配器，实现了生成质量的整体提升。

实验表明，该架构在通用文生图任务中显著增强了纹理层次、边缘锐度与光影交互的物理合理性，消融实验则进一步验证了两种适配器对于提升细节质量的重要性。Eblock 模块在局部细节建模方面具有独特优势，DAT 模块在语义结构保持方面发挥关键作用，二者协调实现了质量最优，而知识蒸馏机制的引入，进一步提升了训练稳定性和生成保真度。

本工作为扩散模型的轻量化质量增强提供了新的思路，其核心思想通过轻量化的适配器在关键位置注入针对性的增强信号，而不是依靠增加模型的参数。该方法兼具参数高效与性能提升的优点，具备较强的可迁移性与应用潜力。

参考文献

[1] Kingma, D.P. and Welling, M. (2013) Auto-Encoding Variational Bayes. https://doi.org/10.48550/arXiv.1312.6114

[2] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., *et al*. (2014) Generative Adversarial Networks. 2014 *Advances in Neural Information Processing Systems*, Montreal, 8-13 December 2014. https://doi.org/10.48550/arXiv.1406.2661

[3] Zhang, L., Rao, A. and Agrawala, M. (2023) Adding Conditional Control to Text-to-Image Diffusion Models. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 3813-3824. https://doi.org/10.1109/iccv51070.2023.00355

[4] Ye, H., Zhang, J., Liu, S., *et al*. (2023) IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. https://doi.org/10.48550/arXiv.2308.06721

[5] Wang, Q., Bai, X., Wang, H., *et al*. (2024) InstantID: Zero-Shot Identity-Preserving Generation in Seconds. https://doi.org/10.48550/arXiv.2401.07519

[6] Li, M., Yang, T., Kuang, H., *et al*. (2024) Controlnet++: Improving Conditional Controls with Efficient Consistency Feedback. In: *Lecture Notes in Computer Science*, Springer, 129-147. https://doi.org/10.1007/978-3-031-72667-5_8

[7] Feijoo, D., Benito, J.C., Garcia, A. and Conde, M.V. (2025) DarkIR: Robust Low-Light Image Restoration. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 10879-10889. https://doi.org/10.1109/cvpr52734.2025.01016

[8] Chollet, F. (2017) Xception: Deep Learning with Depthwise Separable Convolutions. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1800-1807. https://doi.org/10.1109/cvpr.2017.195

[9] Lin, M., Chen, Q. and Yan, S. (2013) Network in Network. https://doi.org/10.48550/arXiv.1312.4400

[10] Han, K., Wang, Y., Guo, J. and Wu, E. (2024) ParameterNet: Parameters Are All You Need for Large-Scale Visual Pre-training of Mobile Networks. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 15751-15761. https://doi.org/10.1109/cvpr52733.2024.01491

[11] Ba, J.L., Kiros, J.R. and Hinton, G.E. (2016) Layer Normalization. https://doi.org/10.48550/arXiv.1607.06450

[12] Aghajanyan, A., Gupta, S. and Zettlemoyer, L. (2021) Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing* (*Volume* 1: *Long Papers*), Online, August 2021, 7319-7328. https://doi.org/10.18653/v1/2021.acl-long.568

[13] Hendrycks, D. and Gimpel, K. (2016) Gaussian Error Linear Units (GELUs). https://doi.org/10.48550/arXiv.1606.08415

[14] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Identity Mappings in Deep Residual Networks. In: *Lecture Notes in Computer Science*, Springer, 630-645. https://doi.org/10.1007/978-3-319-46493-0_38

[15] Ho, J. and Salimans, T. (2022) Classifier-Free Diffusion Guidance. https://doi.org/10.48550/arXiv.2207.12598

[16] Pascanu, R., Mikolov, T. and Bengio, Y. (2013) On the Difficulty of Training Recurrent Neural Networks. 2013 *International Conference on Machine Learning*, Atlanta, 16-21 June 2013, 1310-1318.

[17] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. https://doi.org/10.48550/arXiv.1503.02531