锚定令牌匹配：
用于免训练自回归图像编辑的隐式结构锁定

胡泰航 [1*]; 李林轩 [1*]; 王凯 [3,4†]; 王亚星 [2,1†]; 杨健 [1], 程明明 [2,1]
[1] 媒体计算实验室, 计算机学院, 南开大学, [2] 南开国际先进研究院, 深圳福田
[3] 巴塞罗那自治大学计算机视觉中心, [4] 香港城市大学 (东莞)
{hutaihang00, linxuanli520}@gmail.com
kai.wang@cityu-dg.edu.cn, {csjyang,cmm,yaxing}@nankai.edu.cn

图 1. 我们的方法，隐式结构锁定（*ISLock*），能够实现属性/对象替换（左）、添加/删除对象（中）和样式/状态转换（右）任务，同时保持其他信息与原始图像一致。*ISLock* 也适用于不同的基于 AR 的模型，第一行基于 LlamaGen [49] 生成，而第二行基于 Lumina-mGPT [30] 生成

摘要

文本到图像生成在扩散模型方面取得了突破性的进展，通过交叉注意力机制实现了高保真合成和精确的图像编辑。最近，自回归 *(AR)* 模型再次成为强大的替代方案，利用下一个令牌生成来匹配扩散模型。然而，由于结构控制上的根本差异，现有的为扩散模型设计的编辑技术无法直接迁移到 *AR* 模型。具体而言，*AR* 模型在图像编辑过程中存在注意力图的空间贫乏和结构误差的连续累积，从而破坏了对象布局和全局一致性。在本研究中，我们引入了隐式结构锁定 *(ISLock)*，这是第一个无需训练的 *AR* 视觉模型编辑策略。*ISLock* 不依赖于显式的注意力机制或微调，而是通过锚令牌匹配 *(ATM)* 协议将自注意力模式与参考图像动态对齐，从而保留结构蓝图。通过在潜在空间中隐式地强制执行结构一致性，我们的方法 *ISLock* 能够在保持生成自主性的同时实现结构感知编辑。大量实验表明，*ISLock* 无需额外训练即可实现高质量、结构一致的编辑，并且优于或媲美传统的编辑技术。我们的研究成果为高效灵活的基于自回归 *(AR)* 的图像编辑开辟了道路，进一步弥合了扩散生成模型和自回归生成模型之间的性能差距。代码将在 *https://github.com/hutaiHang/ATM* 公开发布。

*同等贡献, † 共同通讯

1. 介绍

近年来，以扩散模型为中心的文本到图像（T2I）生成技术取得了革命性的突破 [7, 19, 20, 26, 44]。诸如稳定扩散 [11, 39] 和 Imagen [45] 之类的基础模型不仅推动了艺术创作 [3, 13–15] 的创新，而且还催化了丰富的下游任务生态系统——从文本引导的图像编辑 [18, 27, 53, 56] 和拖动操作 [33, 46] 到图像引导的 T2I 生成 [34, 35, 61, 62, 64, 65]。这些方法通常通过在去噪过程中探索潜在特征 [34, 62] 或注意力权重 [36, 53] 来实现像素级的图像处理。

最近，自回归模型（AR）在图像生成领域重新兴起。受大型语言模型（LLM）成功案例 [1, 29, 51] 的启发，近期视觉自回归模型 [30, 50, 59, 60]（包括 LlamaGen [49] 和 Emu3 [57]）将图像视为离散 token 序列 [10, 55]，通过下一 token 预测重建高保真视觉效果。这些方法在长距离连贯性上媲美扩散模型，同时具备独特优势：其序列化生成机制天然支持局部编辑，并能与多模态语

言模型无缝集成。但 AR 模型在图像编辑任务中的潜力仍未充分挖掘——其下一 token 生成范式与扩散模型中完整隐向量的并行去噪机制存在根本差异，导致现有基于扩散的编辑技术无法直接迁移。

这一挑战源于图像生成过程中结构控制机制的显著差异。在扩散模型中，文本到图像的空间对应关系是通过交叉注意力图明确建立的：粗略的结构信息在早期去噪阶段被锁定，从而通过局部注意力调整保持全局一致性 [18, 21, 53]。相比之下，自回归 (AR) 模型遵循独特的结构生成逻辑：每个令牌预测严格依赖于先前的序列，结构信息并非在任何单一阶段集中确定，而是在生成过程中逐步演变。这种机制引入了两个关键问题：*(1) 注意力图的空间贫困*：AR 模型中的文本到图像注意力图缺乏精确的结构对应*，这使得它们作为编辑锚点不可靠（如图 4 所示）。*(2) 结构误差的顺序累积*：对目标词条的简单修改（例如，将“cat”改为“dog”）会导致潜在状态的局部偏移。这些偏差会通过自回归依赖链传播，最终扭曲全局结构，例如物体姿态和场景布局（如图 2 所示）。

近期尝试 [28, 35] 通过在大规模配对图像编辑数据集 [2, 12] 上微调模型参数或引入蒸馏损失来缓解这些问题。然而，此类方法需要大量训练数据和计算资源，同时牺牲了零样本编辑的灵活性。因此，核心挑战显现：如何基于对注意力机制的深层理解，以免训练方式在文本到图像自回归模型中实现结构一致性编辑？

在本研究中，我们首先对自回归 (AR) 图像生成中固有的结构控制机制进行了系统研究。尽管先前的研究已经广泛探讨了扩散模型中注意力引导的编辑 [18, 36, 53, 56]，但注意力图与 AR 框架中结构布局之间的关系仍未得到深入探索。现有的基于扩散的方法 [18, 53] 表明，移植参考注意力图可以有效地增强结构一致性。然而，我们的实验揭示了该方法应用于 AR 模型时的一个根本局限性：注入外部注意力图会破坏 AR 模型固有的注意力动态，导致与生成图像中语义上下文的一致性丧失（如图 2-中图所示，布局结构得以保留，但内容发生了扭曲。）这种破坏表现为纹理模糊、物体比例扭曲和光照不一致，这些现象是由于并行注意力注入和 AR 模型的顺序依赖关系之间固有的不兼容性引起的。为了应对这一挑战，我们提出了隐式结构锁定 (*ISLock*)，作为首个无需零样本训练的 AR 视觉生成模型编辑策略。我们方法的核心是锚令牌匹配 (*ATM*) 策略。我们并非采用强力的注意力图移植方法，而是在自回归解码过程中选择性地匹配令牌，方法是识别那些隐藏表示与原始序列中锚令牌相似度最高的令牌。此过程引入了隐式注意力对齐，使模型能够自然地计算注意力图，在适应局部语义编辑（例如，将毛发纹理从“cat”转换为“dog”）的同时保持结构一致性。需要注意的是，注意力一致性是 *ISLock* 的副产品，而非显式约束的产物。*ISLock* 同时实现了两个关键目标：*(1) 保留结构蓝图*：编辑序列的自注意力图与参考图像的自注意力图保持结构一致性。*(2) 保持生成自主性*：允许模型动态调整注意力模式以适应编辑内容的语义要求，同时非编辑区域保持不变。

*由于 AR 模型中的下一个令牌预测机制，当前令牌会大量关注前一个令牌。

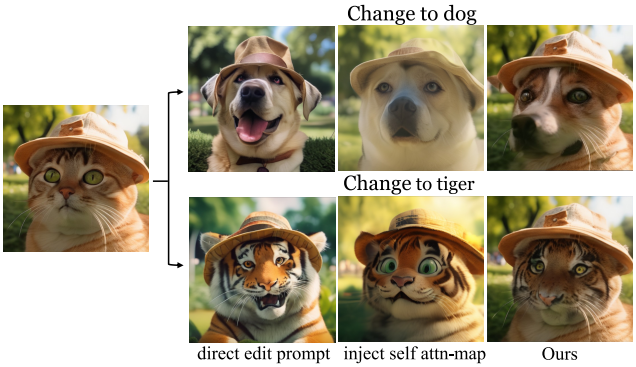

图 2. （左图）直接修改目标令牌会导致结构性错误连续累积，从而造成内容严重失真。（中图）简单的注意力注入会破坏内容连贯性。（右图）相比之下，我们提出的隐式结构锁定 (*ISLock*) 通过提出的锚令牌匹配 (*ATM*) 策略有效地缓解了这些问题。

通过在广泛使用的 PIE-Bench [23] 数据集上与现有的基于扩散和 AR 模型的图像编辑方法进行广泛比较，我们提出的方法 *ISLock* 在文本引导的图像编辑中取得了用户满意的性能。总而言之，我们的主要贡献包括：

- 我们对自回归（AR）图像生成中的注意力机制进行了深入研究，揭示了现有的基于扩散的结构控制方法应用于 AR 模型时的局限性。
- 基于这一发现，我们提出了 第一免训练编辑方法，即隐式结构锁定（*ISLock*），通过在自回归解码期间隐式匹配注意模式来近似结构布局，从而克服文本引导编辑中的空间不一致问题。
- 我们引入了一种令牌匹配机制，即锚定令牌匹配（*ATM*），它通过识别潜在空间中的锚令牌来隐式地保留关键结构元素，同时允许注意力连贯性作为副产品自然出现，确保语义一致性而不破坏生成自主性。
- 在 PIE-Bench [23] 上进行的大量实验表明，我们的方法在基于 AR 的图像编辑中显著保持了结构一致性和视觉保真度。

2. 相关工作

自回归图像生成 受到 LLM [1, 29] 中的序列预测范式的启发，自回归 (AR) 模型将图像生成重新表述为图像令牌序列预测任务。先驱 PixelCNN [54] 通过逐像素的条件概率建模实现了图像合成，但其有限的感受野限制了全局相干性。离散表示学习领域的后续研究解决了这一瓶颈：VQ-VAE [42] 和 VQGAN [10] 建立了可学习的离散码本，将图像压缩为令牌序列，为可扩展的 AR 建模奠定了基础。MaskGIT [6] 创新性地引入了掩码预测机制，通过双向上下文建模实现并行解码，同时保留了自回归特性。

随着 LLMs 的进展，研究人员探索了跨模态扩展——LlamaGen [49] 调整了 LLaMA 架构 [51] 用于视觉 token 建模，取得了与扩散模型 [7, 44, 47] 相当的视觉效果；而 Emu3 [57] 构建了统一的自回归空间，用于

多模态联合训练。值得注意的是，最近的研究 [9, 60] 进一步将视觉理解和生成集成到一个自回归框架中，展现出任务泛化潜力。然而，现有的基于自回归 (AR) 的方法在可控图像编辑方面面临根本性的挑战。它们的顺序生成过程固有地会在局部修改过程中累积误差，从而导致与精确图像编辑任务所需的严格空间一致性相冲突的差异。

文本引导的图像编辑 文本引导的图像编辑旨在根据语义提示修改图像内容，同时保留不相关的区域。传统的基于 GAN 的方法 [16, 25, 37]，通过在 CLIP 的引导下优化 GAN 潜在空间来实现局部编辑 [22, 31, 38] ，但它们的性能受到预训练 GANs 容量的限制。最近基于扩散的方法已成为主流，但它们通常需要复杂的反演和潜在优化过程来平衡保真度和可控性。例如，基于优化的反演方法 [17, 27, 32, 56] 可以细化潜在噪声以实现精确重建，并操纵交叉注意力图（例如，Prompt-to-Prompt [18]）以保留结构。InstructPix2Pix [2] 通过对成对编辑数据进行训练来绕过反转，但依赖于 Prompt-to-Prompt [18] 来生成大规模图像指令对。为了减少数据依赖性，一些研究 [5, 8] 提取连续视频帧作为编辑样本，以模拟真实世界的编辑动态。最近的研究 [35] 通过在配对数据集上进行微调，将 InstructPix2Pix 框架应用于自回归模型。

然而，大多数现有的编辑方法由于其底层架构而仍然分散 - 虽然扩散模型占据主导地位，但 AR 模型仍未得到充分探索，主要是因为缺乏零样本编辑框架。一些先前的研究 [28, 35] 尝试通过对大规模配对图像编辑数据集进行计算成本高昂的微调来解决这一限制，但通常以牺牲零样本灵活性为代价。相比之下，我们的方法统一了 AR 范式内的空间控制和语义编辑，从而无需优化或额外的模型组件。

3. 方法

本研究旨在通过系统地研究控制基于自回归 (AR) 生成图像结构的关键因素，利用文本到图像的自回归 (AR) 模型 [49] 实现零样本文本引导的图像编辑。我们首先在 3.1 节简要回顾文本到图像 AR 模型的范式，这为我们的方法奠定了基础框架。接下来，在 3.2 节，我们通过一系列分析实验确立了我们的研究动机，这些实验重点强调了基于自回归 (AR) 的编辑中结构保留的挑战。最后，我们将在 3.3 节介绍我们的方法——隐式结构锁定 (*ISLock*)，并详细说明其设计和有效性。图 3 展示了 *ISLock* 的概述。

3.1. 准备工作

AR 视觉生成模型 LlamaGen [49] 通过顺序预测图像令牌，将文本合成图像。其架构由两个协同工作的关键组件组成：一个 VQ-Autoencoder [10, 42]，用于将图像转换为离散的令牌序列；以及一个自回归变换器 f_θ，用于学习这些图像令牌的联合分布。给定输入图像 $x \in \mathcal{R}^{H\times W\times 3}$，特征编码器 $\mathcal{E}_{VQ}$ 首先将其映射到潜在表示 $z^e \in \mathcal{R}^{h\times w\times d}$，其中 d 表示特征维度。通过最近邻量化，每个空间特征向量 $z^e_{i,j}$ 被投影到码本原型 $z^q_{i,j} \in \mathcal{V}$，生成离散令牌序列 $\mathbf{Z} = \{z_1, ..., z_{h\times w}\}$，其中 $\mathcal{V}$ 是学习到的码本向量集。为了实现文本条件生成，LlamaGen [49] 集成了预先训练的 T5 [41] 文本编码器 τ_ξ，将文本提示 $\mathcal{P}$ 映射到嵌入序列 $c = \tau_\xi(\mathcal{P})$。文本嵌入通过线性层投影到 Transformer 输入空间，并添加到图像令牌序列的前面。随后，AR 模型 f_θ 自回归地预测以文本为条件的图像令牌的联合分布：

$$P(z|c) = \prod_{i=1}^{N} P(z_i|z_{<i}, c) \tag{1}$$

其中 $N = h \times w$。在生成过程中，连接序列 $[c; z_{<i}]$ 通过堆叠的 Transformer 层进行处理，并使用因果掩蔽来强制执行自回归约束。在每一层 l，自注意力机制如下：

$$\mathbf{A}_l = \text{Softmax}\left(\frac{\mathbf{Q}_l\mathbf{K}_l^\top}{\sqrt{d_k}}\right) \tag{2}$$

其中 $\mathbf{H}_l$ 表示层 l 的隐藏状态，W_l^Q, W_l^K, W_l^V 是投影矩阵，并且 $\mathbf{Q}_l = \mathbf{H}_l W_l^Q, \mathbf{K}_l = \mathbf{H}_l W_l^K, \mathbf{V}_l = \mathbf{H}_l W_l^V$。最后一层应用 MLP，然后进行 softmax 运算，以预测 $\mathcal{V}$ 上的下一个令牌分布。生成的令牌序列 $\mathbf{s}$ 最终通过解码器 $\mathcal{D}_{\text{VQ}}$ 解码为 RGB 图像空间，从而完成文本到图像的合成流程。

3.2. 结构信息分析

为了揭示 AR 视觉生成模型的内在结构控制机制，我们开展了一系列系统的实验分析，重点关注其 注意力动态和 序列敏感性。我们的观察揭示了一个关键的洞见：尽管连接文本和图像令牌的交叉注意力图缺乏重要的空间信息，[†]，图像令牌之间的自注意力图展现出丰富的结构信息。

如图 4所示，我们将主成分分析 (PCA) 分解应用于自注意力矩阵 $A \in \mathcal{R}^{(h\times w)\times(h\times w)}$ 到三维空间。可视化结果表明，语义相似的令牌倾向于表现出连贯的注意力模式，这表明空间结构是通过图像令牌自组织形成的，而不是仅仅依赖于明确的文本提示指导。然而，在将参考图像的注意力图注入目标生成过程作为保持结构一致性的简单方法时，我们观察到了明显的伪影和全局扭曲（如图 2所示）。我们将其归因于参考注意力图和目标序列的潜在动态之间的上下文不匹配。

如图 5所示，对 序列敏感度的进一步研究表明，对生成序列中前 20% 的 token 进行扰动会导致结构相似性指数 (SSIM) 显著下降，变化量为 0.56 ± 0.02。这种影响明显高于对序列后 20% 进行扰动时的影响，后 20% 的 SSIM 变化量为 $\Delta\text{SSIM} = 0.08 \pm 0.05$。此外，后期扰动引起的扭曲主要集中在高频细节区域，例如精细纹理和边缘。这种观察到的渐进式结构固化现象与理论预期相符：由于 Transformer 中的因果注意力机制，早期令牌与所有后续位置相互作用，从而在塑造全局图像结构方面发挥关键作用。相比之下，后期令牌更

[†]在 AR 模型中，交叉注意力与扩散模型的不同之处在于，它在文本和图像令牌之间共享 QKV 投影，而扩散模型则使用单独的映射。更多详情请参阅补充材料

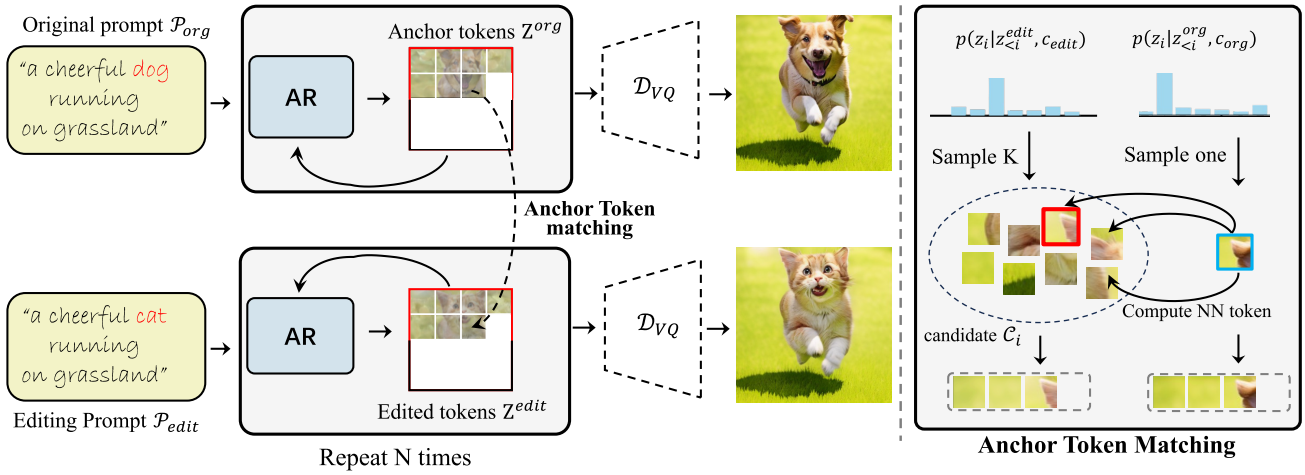

图 3. 我们的方法 *ISLock* 通过锚定令牌匹配 (*ATM*) 实现隐式结构锁定。我们的方法并非依赖于直接的注意力注入（这通常会在基于 AR 的视觉生成中引入失真），而是通过从 K 个候选令牌中识别与参考令牌距离最小的候选令牌来选择最佳编辑令牌。

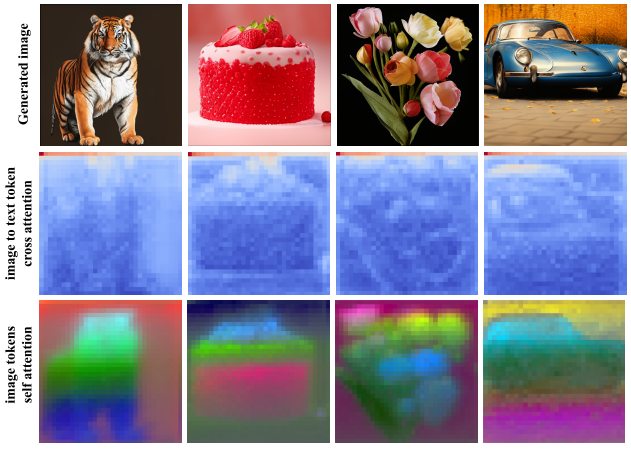

图 4. 使用 LlamaGen 模型生成最后一个 token 后，交叉注意力（第二行）和自注意力（第三行）的可视化效果。我们观察到，从图像到文本 token 的交叉注意力包含极少的结构信息，而自注意力图则表现出与结构布局更强的语义对齐性。

多地受到先前生成的令牌的局部上下文约束，导致对整体结构的影响较小。这凸显了早期令牌在自回归图像生成过程中定义全局图像结构的重要性。

3.3. 隐式结构锁定 (*ISLock*)

基于上一节的观察，我们发现直接注入注意力图会导致严重的伪影和失真，从而破坏生成图像的连贯性。为了解决这个问题并确保图像编辑过程中的结构一致性，我们提出了一个利用锚令牌匹配 (*ATM*) 的自适应解码框架。我们的 *ISLock* 并非显式地进行注意力移植，而是隐式地将结构锁定在潜在空间中。

使用动态窗口进行锚令牌匹配 给定原始提示 $\mathcal{P}_{org}$ 和编辑提示 $\mathcal{P}_{edit}$，AR 模型会根据预测分布 $p(z_i|z_{<i}, c)$ 采样下一个令牌。对于 $\mathcal{P}_{org}$，它会从分布 $p(z_i|z_{<i}, c_{org})$ 中采样一个令牌 z_i^{org}，并将其作为生成编辑图像时相应位置的锚令牌。我们的动态锚令牌匹配 (*ATM*) 框架会隐式对齐结构，同时保持编辑的灵活性。在生成编辑序

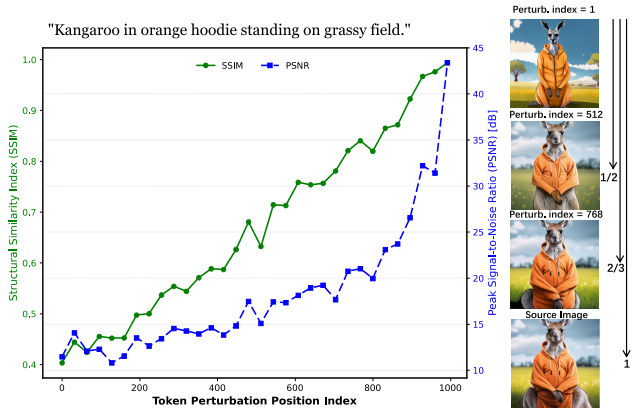

图 5. 在 AR 生成的不同阶段对图像令牌进行扰动会导致图像质量指标（SSIM 和 PSNR）发生不同的变化，如左侧曲线所示。早期扰动主要影响整体结构几何，而后期扰动仅影响高频细节，如右侧生成结果所示。

列 $\mathbf{Z}^{\text{edit}}$ 的步骤 i 中，我们从条件分布 $p(z_i|z^{\text{edit}}_{<i}, c_{edit})$ 中采样 K 个候选令牌 $\mathcal{C}_i = \{z_i^{(1)}, ..., z_i^{(K)}\} \subset \mathbb{R}^d$。我们计算每个候选 $z_i^{(k)}$ 与参考锚点 z_i^{ref} 之间的潜在空间欧几里得距离，如下所示：

$$s^{(k)} = \|z_i^{(k)} - z_i^{\text{org}}\|_2^2 \tag{3}$$

然后，我们选择距离最小的匹配候选作为输出，这个过程类似于最近邻 (NN) 计算。这种 *ATM* 策略确保了编辑序列和参考序列的隐藏状态轨迹之间的局部对齐，从而引导自注意力机制生成结构一致的注意力图。通过利用隐式结构约束而非直接注入注意力图，我们的方法 *ISLock* 减轻了语义冲突并保持了上下文连贯性，从而实现了更稳定、更结构感知的图像编辑。

为了适应不同生成阶段中不同的结构约束，我们进一步引入了 *动态窗口机制*。自适应滤波窗口 $\mathcal{W}_i \subseteq \mathcal{C}_i$ 会根据解码进度按比例调整其大小，如下所示：

$$|\mathcal{W}_i| = \lfloor K \cdot (1 - \alpha \cdot \frac{i}{N}) \rfloor, \quad \alpha \in [0, 1] \tag{4}$$

其中我们默认设置 $\alpha = 0.6$。初始化 ($i = 0$) 时，窗口保留 100% 的候选集 ($|\mathcal{W}_i| = K$)，以强制对基础结构进行严格的结构对齐。随着生成的进行，窗口线性缩小，在 $i = 0.5N$ 时达到 70% 的容量 ($|\mathcal{W}_i| \approx 0.7K$)，在完成 ($i = N$) 时达到 40% ($|\mathcal{W}_i| \approx 0.4K$)。这种设计保证了持续的自适应性，因为早期阶段通过广泛的候选集优先考虑结构保真度，而后期阶段则通过更严格的约束逐步强调上下文连贯性。最终的编辑令牌如下：

$$z_i^{\text{edit}} = \arg\min_{k \in \mathcal{W}_i} s^{(k)} \tag{5}$$

通过这种方式，我们的方法 *ISLock* 实现了渐进式潜在空间对齐，通过隐式结构指导有效地缓解了模式不匹配。

自适应约束松弛 (*AdaCR*). 为了平衡 结构蓝图和 生成自主性，我们提出了一种基于自适应阈值的自主性保护方案。相似度阈值 τ 作为约束调节器：

$$z_i^{\text{edit}} = \begin{cases} \arg\min s^{(k)} & \text{if } \min s^{(k)} \leq \tau \\ \arg\max p(z_i | z_{<i}^{\text{edit}}, c_{edit}) & \text{otherwise} \end{cases} \tag{6}$$

该机制包含两项保障措施：(1) 候选窗口预过滤，以确保 AR 生成质量；(2) 动态阈值，以防止过度约束错误。此外，用户可以根据具体需求调整 τ，在文本到图像的生成多样性和与原始输入图像的相似性之间取得平衡。较大的 τ 可以保留与原始图像的更大相似性，而较小的 τ 则可以提高生成多样性。

最后，我们的 *ISLock* 框架主要构建于锚令牌匹配 (*ATM*) 策略之上，并辅以额外的 动态窗口和 *AdaCR* 技术。这些模块协同工作，在自回归 (AR) 生成模型中实现无需训练的文本引导图像编辑。

4. 实验

4.1. 实验设置

基准 我们在 LlamaGen [49] 上构建了我们的方法 *ISLock* ，以 512×512 的分辨率生成图像。默认情况下，我们将 $K = 150, \tau = 1.0$ 设置为超参数。为了建立一个严谨的无需训练的 AR 图像编辑评估框架，我们设计了一个与我们方法的操作范式相符的“生成到编辑”流程。鉴于我们的方法处理 AR 生成图像的内在需求，我们在 PIE-Bench 数据集 ([24]，图像编辑的标准基准) 的一个精选子集上评估了我们的方法。虽然现有方法通常支持 PIE-Bench 中的所有 10 种编辑类型，但我们专注于与我们的无需训练的 AR 范式兼容的 *5* 个基本类别：对象替换、对象添加、对象移除、风格迁移和属性修改。PIE-Bench [23] 精选子集中的每个案例都包含成对的原始提示和编辑提示，我们首先使用 LlamaGen 根据原始提示生成图像，然后应用 *ISLock* 根据编辑提示编辑这些图像。

指标 为了全面评估我们的方法，我们采用了三个核心指标：(1) 通过结构距离 [52] 衡量原始图像和编辑图像之间的结构一致性；(2) 通过 PSNR、LPIPS [63]、MSE 和 SSIM [58] 量化背景区域之间的背景保留（使用通过 Grounded-SAM [43] 生成的前景蒙版）；以及 (3) 通过 CLIP Score [40] 衡量整幅图像和编辑蒙版中区域的语义对齐。这种多维度评估框架确保在无需训练的 AR 图像编辑中对结构完整性和语义保真度进行严格验证。

比较方法 遵循 [35]，我们比较了当前基于扩散的文本驱动图像编辑方法，这些方法大致可分为两种范式：基于反转的方法和无反转方法。基于反转的技术，包括 Prompt-to-Prompt [18]、Null-text Inversion [32]、PnPInversion [24]、Pix2Pix-Zero 和 MasaCtrl [4]，通常依赖于优化输入图像的反转潜在表示，以在编辑过程中保持结构一致性。相比之下，无反转方法（例如 InstructPix2Pix [2] 和 MGIE [12]）通过利用替代策略来绕过显式潜在反转。

4.2. 实验结果

定性比较 图 6 定性比较了我们的方法与其他编辑方法在各种编辑任务中的效果。基于指令的方法，例如 InstructPix2Pix [2] 和 MGIE [12]，在涉及对象添加和全局风格迁移的任务中表现良好——例如，在第二行（猫 → 带项圈的猫）和最后一行（照片 → 水彩画）中。然而，这些方法在对象替换任务中表现不佳，经常会引入意想不到的全局背景变化。例如，在第三行（马 → 斑马）中，变换会同时改变草地的纹理和背景颜色。同样，在第一行（玫瑰 → 黄玫瑰）中，背景颜色也受到了明显的影响。此外，当这些方法失效时，可能会引入严重的视觉伪影，如在第四行（兔子 → 松鼠）中所示，编辑会导致图像饱和度不自然，内容完全扭曲。相比之下，基于反转的方法，例如空文本反转 [32] 和 PnP 反转 [23]，在背景保留和编辑对齐之间取得了更好的平衡。然而，它们对交叉注意力图的依赖带来了一个根本性的限制：它们在物体移除任务中表现不佳。这在第五行（戴墨镜的男人 → 男人）中很明显，两种方法都失败了。

ISLock 在所有五种编辑类型中都展现出强大的潜力：对象添加、对象移除、对象替换、属性修改和样式迁移。它有效地保留了结构一致性，同时确保了局部和精确的修改，使其成为文本驱动图像编辑的更通用的解决方案。

定量比较. 表格 1 全面比较了 *ISLock* 与基于扩散的编辑方法在多个关键指标上的差异。与之前主要依赖稳定扩散 [44] 的研究不同，*ISLock* 首次探索了基于自回归 (AR) 模型的无需训练的结构化图像编辑方法。虽然基础模型能力的根本差异必然会带来新的挑战，但我们的结果在所有评估维度上都展现了极具竞争力的性能。

我们还实现了两个简单的基线：朴素提示修改 (*NPM*)，它直接修改提示中的目标词；以及 PnP-AR [18]，它替换编辑过程中从原始图像生成的注意力图。如表格 1 所示，这些基线难以在结构一致性和背景保留之间取得平衡，其结构距离得分明显较高（分别为 113.95 和 103.94），而背景保真度较差。相比之下，*ISLock* 获得了令人满意的结构距离（31.79↓，仅次于基于反转的方法），表明在编辑操作过程中具有出色的结构保留效果。

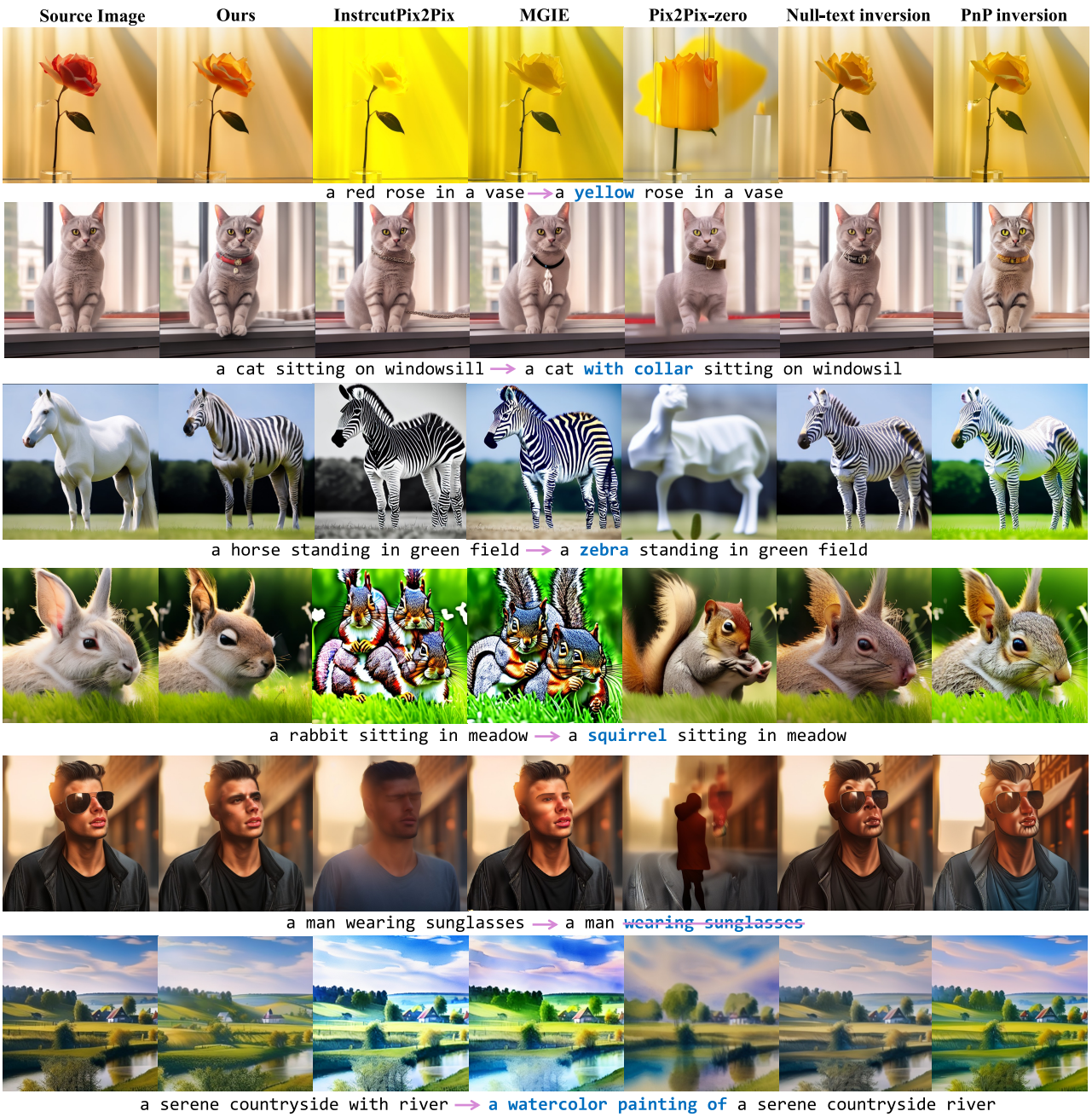

图 6. 与各种文本引导的图像编辑方法进行定性比较

此外，我们的方法保持了良好的背景保真度，在峰值信噪比 (PSNR) 和 SSIM 指标上的表现与领先的基于指令的无反转编辑方法 (InstructPix2Pix [2]、MGIE [12]) 相当。值得注意的是，我们的方法在整幅图像和编辑区域均获得了较高的 CLIP 相似度得分 (24.19&21.33，仅次于 PnPInversion 和 P2P)，突显了其在保持语义一致性的同时将编辑内容与目标文本提示对齐的能力。

虽然某些基于扩散的方法优于 *ISLock*，但这些方法专门针对扩散框架进行了优化。我们的方法在自回归 (AR) 范式中开创了无需训练的结构控制，为 AR 图像处理开辟了一条新的技术途径。

普遍性 我们的方法适用于各种 AR 基础模型。我们在补充材料中纳入了基于 Lumina-mgpt [30] 的定性实验。

4.3. 消融研究

窗口大小影响 $|\mathcal{W}|$. 如图 7和表 2所示，对窗口大小 $|\mathcal{W}|$ 的消融研究揭示了其在平衡结构保存和模型生成灵活性方面的关键作用。当 $|\mathcal{W}| = 1$ 时，我们的方法退化为简单的即时修改，导致编辑控制力下降。随着 $|\mathcal{W}|$ 的增加，编辑图像与原始图像之间的结构距离减小，表明图像构图保存得更好。然而，这是以生成灵活性的降低为代价的。如图 7所示，当 $|\mathcal{W}|$ 过大时，伸出舌头的

方法	文生图模型	结构	背景保存				CLIP 相似性	
		Distance ↓	PSNR ↑	LPIPS ↓	MSE ↓	SSIM ↑	Whole ↑	Edited ↑
Prompt-to-Prompt [18]	SD1.4	88.46	16.80	270.38	241.89	69.93	26.70	21.43
Null-text Inversion [32]	SD1.4	18.42	25.68	77.70	42.92	85.71	24.55	20.73
Pix2pix-zero [36]	SD1.4	59.43	19.71	193.44	147.19	76.48	23.56	19.76
MasaCtrl [4]	SD1.4	34.20	21.59	124.35	83.60	81.31	22.90	18.52
PnPInversion [23]	SD1.5	24.81	22.16	114.15	74.07	81.81	25.56	21.50
InstructPix2Pix [2]	SD1.5	67.49	19.69	164.27	235.62	76.98	23.37	20.48
MGIE [12]	SD1.5	53.46	20.62	131.13	205.09	79.55	22.67	19.58
NPM	LlamaGen	113.95	12.14	377.84	725.98	53.67	**24.71**	21.28
PnP-AR	LlamaGen	103.94	13.20	328.49	600.20	58.25	23.56	20.65
ISLock (Ours)	LlamaGen	**31.79**	**19.75**	**136.21**	**161.17**	**76.71**	<u>24.19</u>	**21.33**

Table 1. 我们主要与基于自回归 (AR) 的方法进行比较。我们应该以某种方式突出我们的方法。成绩最好和次好的基于自回归 (AR) 的方法分别用 **粗体**和 <u>下划线</u> 标记。

图 7. 窗口大小 $|\mathcal{W}|$ 的影响。随着窗口大小的增加，结构保持性得到改善，但灵活性降低。动态窗口策略可以实现更好的平衡。

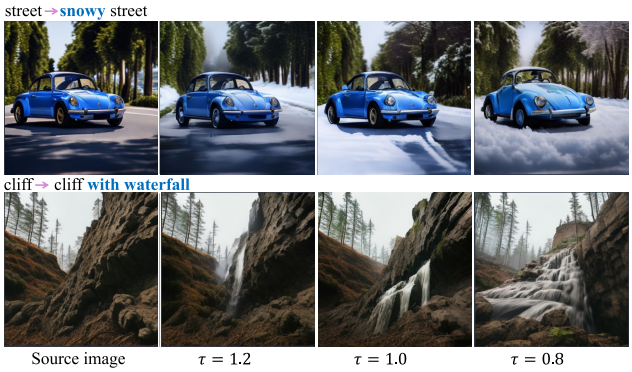

图 8. 阈值 τ 影响图像编辑的强度。随着阈值的降低，编辑效果更加明显，例如雪覆盖率增加、瀑布流量增大。

Win. Size	Struc. Dist.↓	Clip Sim.↑	S/C↓
$\|\mathcal{W}\| = 50$	60.83	**24.79**	2.45
$\|\mathcal{W}\| = 100$	38.03	24.33	1.56
$\|\mathcal{W}\| = 150$	**30.39**	22.18	1.37
Dynamic (Ours)	<u>31.79</u>	<u>24.19</u>	**1.31**

Table 2. 窗口大小的消融研究 $|\mathcal{W}|$.

“快乐狗” 会被修改为闭嘴，蛋糕上的巧克力涂层会被替换为一层普通的面包。这种影响进一步反映在 CLIP 相似度的下降中（表格 2），表明与目标提示的一致性较弱。相比之下，我们的动态窗口策略（如图 7最后一列所示）在结构一致性和语义对齐之间取得了更好的平衡，从而获得了最佳的结构距离/CLIP 相似度比率。

阈值效应 τ. 如图 8所示，阈值 τ 显著影响修改的程度，尤其是在对象添加等高方差编辑任务中。在图像编辑过程中，有时会出现没有候选令牌与锚令牌紧密匹配的情况。在这些情况下，调整 τ 可以控制编辑强度。较低的 τ 允许保留更多来自原始采样的令牌，从而导致更显著的修改。这可以在图 8 中观察到，其中降低 τ 会导致积雪更厚（第一行）和瀑布流量增加（第二行）。相反，较高的 τ 可以更好地保持与原始图像的一致性，从而保持更克制的编辑。

5. 结论

在这项工作中，我们解决了自回归（AR）模型中文本引导图像编辑的基本挑战，而无需修改模型参数或依赖明确的注意力操作。通过引入隐式结构锁定（*ISLock*），我们通过一种名为锚令牌匹配（*ATM*）的新型候选匹配协议实现无训练图像编辑，该协议在保持生成自主性的同时对齐原始图像的结构。与直接注意力图移植不同，我们的方法 *ISLock* 确保结构保存是潜在空间结构的自然结果，从而使 AR 模型能够在各种编辑任务中保持空间连贯性。大量实验表明，*ISLock* 在结构感知自回归编辑中取得了有竞争力的表现，弥合了 AR 模型和扩散模型之间的差距。

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 11

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 3, 5, 6, 7, 11

[3] Muhammad Atif Butt, Kai Wang, Javier Vazquez-Corral, and Joost van de Weijer. Colorpeel: Color prompt learning with diffusion models via color and shape disentanglement. In *European Conference on Computer Vision*, pages 456–472. Springer, 2024. 1

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *ICCV*, 2023. 5, 7, 11

[5] Mingdeng Cao, Xuaner Zhang, Yinqiang Zheng, and Zhihao Xia. Instruction-based image manipulation

by watching how things move. *arXiv preprint arXiv:2412.12087*, 2024. 3

[6] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022. 2

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 1, 2

[8] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. *arXiv preprint arXiv:2412.07774*, 2024. 3

[9] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 3

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1, 2, 3

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1

[12] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 2, 5, 6, 7, 11

[13] Daniel Geng, Inbum Park, and Andrew Owens. Factorized diffusion: Perceptual illusions by noise decomposition. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. 1

[14] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24154–24163, 2024.

[15] Alex Gomez-Villa, Kai Wang, Alejandro C Parraga, Bartlomiej Twardowski, Jesus Malo, Javier Vazquez-Corral, and Joost van de Weijer. The art of deception: Color visual illusions and diffusion models. *CVPR*, 2025. 1

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[17] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Yuxiao Chen, Di Liu, Qilong Zhangli, et al. Improving negative-prompt inversion via proximal guidance. *arXiv preprint arXiv:2306.05414*, 2023. 3

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *ICLR*, 2023. 1, 2, 3, 5, 7, 11

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[20] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[21] Taihang Hu, Linxuan Li, Joost van de Weijer, Hongcheng Gao, Fahad Shahbaz Khan, Jian Yang, Ming-Ming Cheng, Kai Wang, and Yaxing Wang. Token merging for training-free semantic binding in text-to-image synthesis. *Advances in Neural Information Processing Systems*, 37:137646–137672, 2025. 2

[22] Yue Jiang, Yueming Lyu, Bo Peng, Wei Wang, and Jing Dong. Cmsl: Cross-modal style learning for few-shot image generation. *Machine Intelligence Research*, pages 1–17, 2025. 3

[23] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *ICLR*, 2023. 2, 5, 7, 11

[24] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 5

[25] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. *CVPR*, 2022. 3

[26] Daiqing Li, Aleks Kamko, Ali Sabet, Ehsan Akhgari, Linmiao Xu, and Suhail Doshi. Playground v2, 2023. 1

[27] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing, 2023. 1, 3

[28] Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. *arXiv preprint arXiv:2410.02705*, 2024. 2, 3

[29] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 1, 2

[30] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 1, 6, 11, 13

[31] Ming Liu, Yuxiang Wei, Xiaohe Wu, Wangmeng Zuo, and Lei Zhang. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration. *Science China Information Sciences*, 66 (5):151101, 2023. 3

[32] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *CVPR*, 2023. 3, 5, 7, 11

[33] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. DragonDiffusion: Enabling Drag-style Manipulation on Diffusion Models. *ICLR*, 2024. 1

[34] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 1

[35] Jiteng Mu, Nuno Vasconcelos, and Xiaolong Wang. Editar: Unified conditional generation with autoregressive models. *arXiv preprint arXiv:2501.04699*, 2025. 1, 2, 3, 5

[36] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, 2023. 1, 2, 7, 11

[37] 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. 3

[38] Zhexi Peng, He Wang, Yanlin Weng, Yin Yang, and Tianjia Shao. Unsupervised image translation with distributional semantics awareness. *Computational Visual Media*, 9(3):619–631, 2023. 3

[39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[40] 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

[41] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3

[42] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2, 3

[43] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 5

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2, 5

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[46] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing. *CVPR*, 2024. 1

[47] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd-if. https://github.com/deep-floyd/IF, 2023. 2

[48] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. LLama-Gen:Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. 2024. 11, 13

[49] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 1, 2, 3, 5, 11

[50] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2025. 1

[51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[52] 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. 5

[53] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *CVPR*, 2023. 1, 2

[54] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 2

[55] Junke Wang, Yi Jiang, Zehuan Yuan, Bingyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems*, 37: 28281–28295, 2025. 1

[56] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, and Joost van de Weijer. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *NeurIPS*, 2023. 1, 2, 3

[57] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 1, 2

[58] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402. Ieee, 2003. 5

[59] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 1

[60] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *ICLR*, 2025. 1, 3

[61] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[64] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: conditional control for one-shot text-driven video editing and beyond. *Science China Information Sciences*, 68(3):132107, 2025. 1

[65] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models. *NeurIPS*, 2023. 1

A. 实现细节

A.1. 方法配置

我们的实现基于 LlamaGen [49] 和 Lumina-mGPT [30] 的官方代码库。对于基于 LlamaGen 的编辑，我们采用候选窗口大小 $K = 150$ 和相似度阈值 $\tau = 1.0$，并在 VQ-AutoEncoder 码本的潜在空间中使用欧氏距离度量。Lumina-mGPT 实现采用 $K = 100$ 和 $\tau = 0.4$，通过第一层 Transformer 嵌入的余弦相似度来测量 token 距离。所有实验均在 NVIDIA 3090 GPU 上进行。

A.2. 基线实施

对于我们比较的基于扩散的方法，包括 P2P [18]、Null-text inversion [32]、PnPInversion [23]、Pix2Pix-zero [36]、MasaCtrl [4]、InstructPix2Pix [2] 和 MGIE [12]，我们利用了它们的官方实现。对于我们实现的两个简单的基于自回归 (AR) 模型的基线方法，我们进行了如下实现：

- **简单修改提示 (*NPM*)**: 该基线将原始提示修改为编辑后的提示，同时保持所有其他变量（例如，未编辑的单词和随机种子）不变。
- **PnP-AR**: 在这个基线中，我们保存了在原始提示生成过程中计算出的逐令牌和逐层注意力图。之后，在从编辑后的提示生成图像时，这些注意力图会被直接替换到相应的令牌位置和层上。

A.3. 评估基准

我们的方法侧重于五种基本的编辑类型：对象替换、对象添加、对象删除、风格迁移和属性修改。对于每种编辑类型，我们从 PIE-Bench [23] 数据集的相应类别中随机选择 10 个示例。每个示例包含一个原始提示和一个编辑后的提示。

我们首先使用 LlamaGen 根据原始提示生成图像，然后应用我们的方法根据编辑后的提示编辑这些图像。由于 LlamaGen 在从短提示生成高质量结果方面存在固有局限性 [49]，我们使用 GPT-4o mini [1] 作为提示增强器，在生成之前对提示进行细化和改进。

B. 注意力图分析

B.1. AR 中的注意力机制

如图 9所示，对于像 LlamaGen 这样的自回归图像生成模型，文本提示首先由文本编码器编码以获得文本令牌，这些文本令牌作为整个生成序列的前缀令牌。在生成每个后续图像令牌的过程中，会同时使用前面的图像令牌和整个文本令牌集来计算注意力。我们论文中提出的自注意力图源自图像自注意力机制，如图 9右下角所示；而交叉注意力图则源自图像到文本的交叉注意力机制，如图 9左下角所示。

B.2. 注意力可视化

我们的方法并非显式地注入注意力图，而是通过锚令牌匹配隐式地实现结构保留，这自然而然地带来了注意力图的一致性。如图 10 所示，与使用 NPM 方法获得的注意力图相比，我们的方法生成的编辑图像的注意力图与原始图像的注意力图自然对齐。

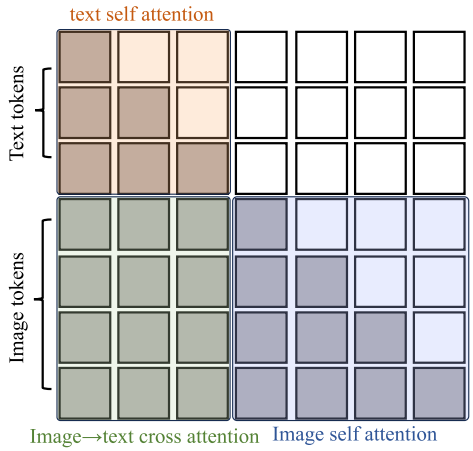

图 9. LlamaGen 中的注意力机制图解 [48]。

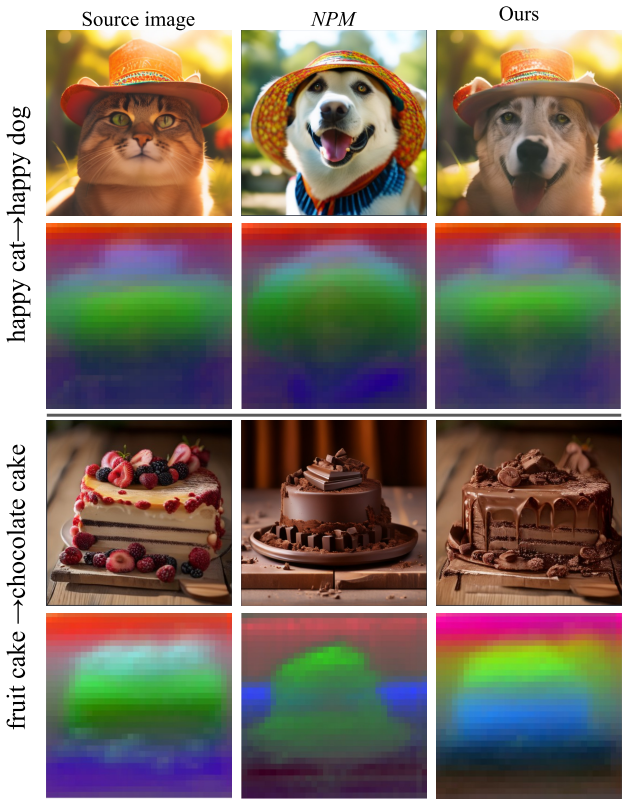

图 10. 注意力图的消融研究。与 *NPM* 相比，我们的方法 *ISLock* 在注意力图生成过程中自然地实现了原始图像与编辑图像生成过程的更好对齐。

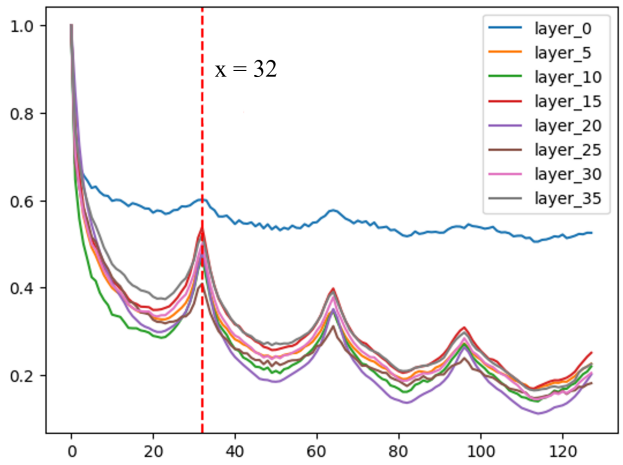

图 11. 自回归图像生成模型往往倾向于为相邻位置的令牌分配更大的注意力分数。图中的值使用最小-最大归一化进行归一化。

B.3. 注意力局部性

我们观察到，在 LlamaGen 等自回归模型中，在注意力计算过程中，token 倾向于将更高的注意力权重分配给与其位置相邻的 token。如图 11 所示，当前 token 分配的注意力分数会随着与当前 token 距离的增加而降低。然而，注意力分数会以 32 个 token 为间隔周期性地增加。出现这种现象的原因是，在生成 512×512 图像时，LlamaGen 使用 VQ-Autoencoder 将图像编码到 32×32 的潜在空间中。位于距离当前 token 32 倍数位置的 token 位于潜在空间的同一列中，从而在这些间隔处具有更高的注意力分数。这也解释了为什么在主要论文的图 4 中从图像 token 到文本 token 的交叉注意力显示最早的图像 token 具有最高的注意力分数。

C. 更多可视化比较

在图 12中，我们展示了额外的编辑结果，证明我们的方法在不同的编辑类型和基于 AR 的模型中具有很好的泛化能力。

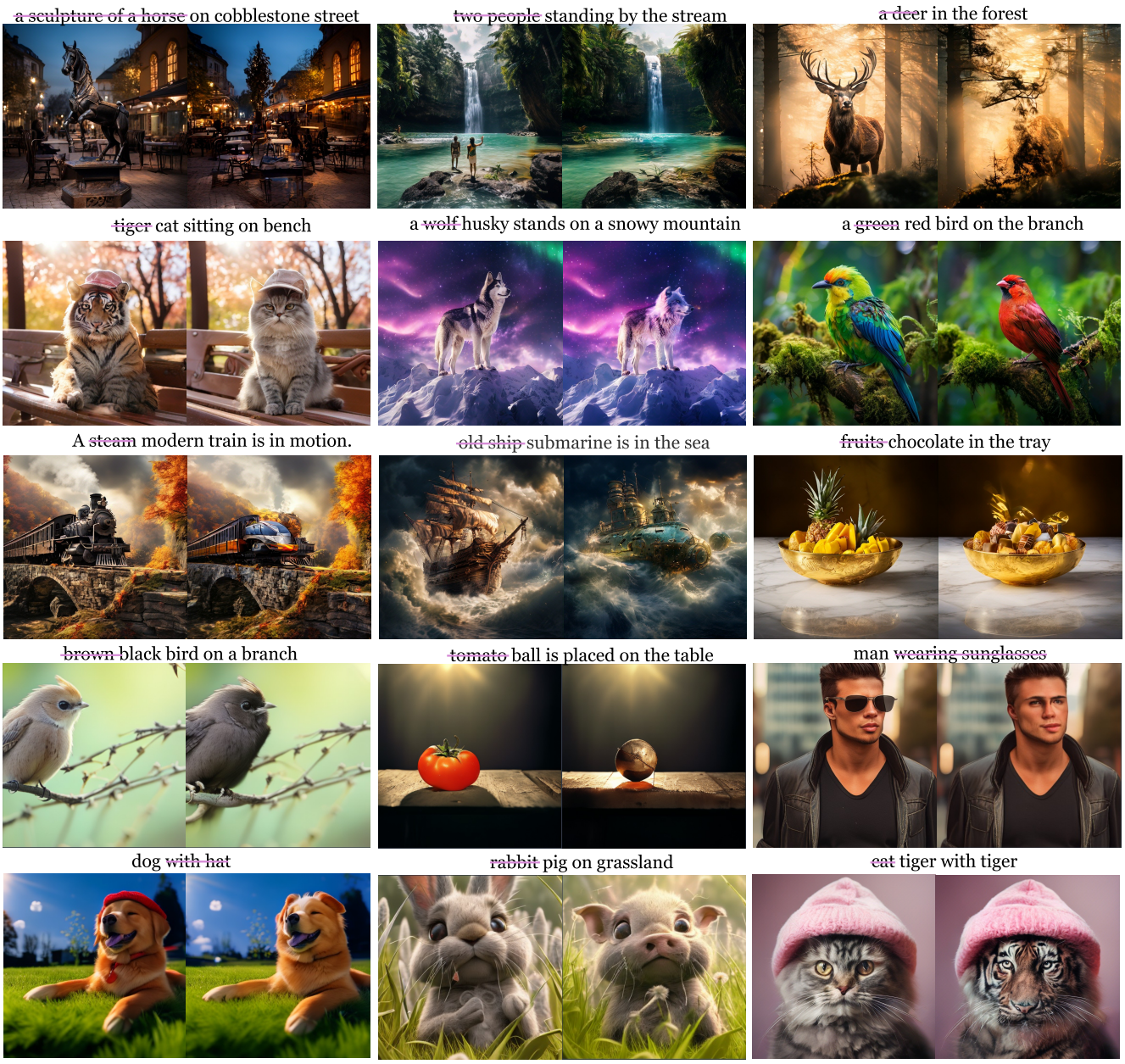

图 12. 我们的方法 *ISLock* 的更多可视化结果，其中前三行我们的方法与 lumina-mgpt [30] 集成，后两行我们的方法与 LlamaGen [48] 协同工作。