

TweezeEdit: 带有路径正则化的一致且高效的图像编辑

Jianda Mao, Kaibo Wang, Yang Xiang, and Kani Chen*

Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong SAR

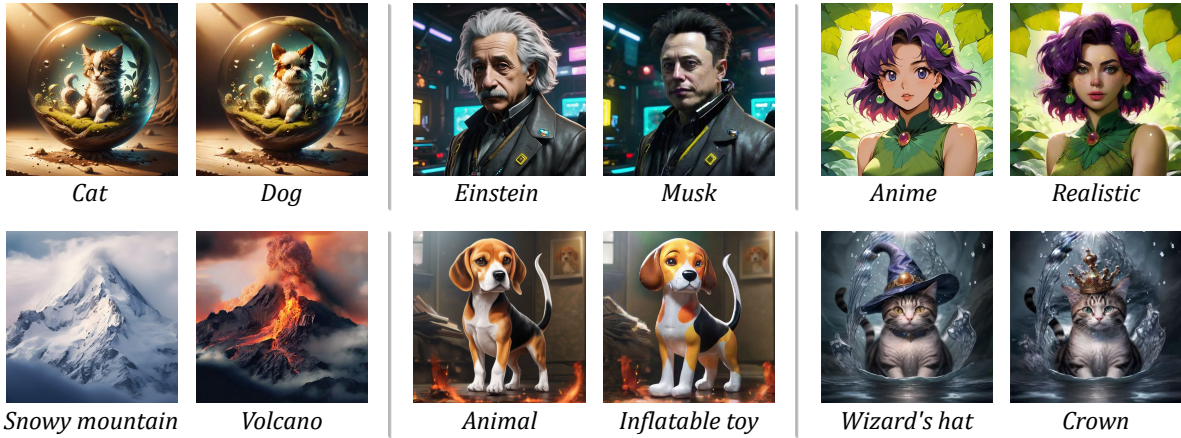


Figure 1: 使用 TweezeEdit 编辑的图像示例，这是一种使用预训练一致性模型进行文本驱动编辑的无调优和反转框架。它有效地保留了源图像语义，同时与目标提示对齐。

ABSTRACT

大规模预训练扩散模型使用户能够通过文本指导编辑图像。然而，现有方法往往过分对齐目标提示，而不能充分保留源图像语义。这些方法通过将目标图像显式或隐式地产生自源图像的逆过程噪声，称为逆锚定。我们认为这种策略在语义保留方面是不理想的，并且由于编辑路径过长而效率低下。我们提出了一种称为 TweezeEdit 的调整和无需逆转的框架，以实现一致且高效的图像编辑。我们的方法通过规范化整个去噪路径，而不是仅依赖逆锚定，解决了这些限制，保证了源语义的保留并缩短了编辑路径。在梯度驱动的正则化指导下，我们利用一致性模型沿直接路径高效地注入目标提示语义。大量实验表明，TweezeEdit 在语义保留和目标对齐方面表现优于现有方法。值得注意的是，它仅需 12 步（每次编辑 1.6 秒），彰显了其在实时应用中的潜力。

1 介绍

最近，在大规模扩散模型方面的进展 (Rombach et al., 2022; Luo et al., 2023; Labs, 2024) 已经使得文本驱动的图像编辑成为可能 (Xu et al., 2024; Rout et al., 2024; Kulikov et al., 2024; Hertz et al., 2022)，在这种方法中，输入一张源图像和目标提示，生成一个编辑后的图像。然而，现有的方法往往无法保持源图像的语义内容，过于贴合目标提示并需要大量附加控制来进行修正 (Shuai et al., 2024)。当前的方法通常采用确定性反向过程来提取反演锚点（源图像的反演噪声）。虽然从理论上讲，反演锚点可以包含源图像的信息，但它们在实践中往往无法产生预期的类似去噪路径，这在编辑过程中导致过度对齐 (Mokady et al., 2023; Miyake et al., 2023)。

这一挑战主要由两个问题引起：(1) 反演中的数值误差导致信息丢失，阻止反演锚点完全重构源图像；(2) 对扩散模型输出缺乏约束，导致从反演锚点到源图像和目标图像去噪路径的不受控偏离，引发意外变化。尽管最近的努力 (Kulikov et al., 2024) 通过使用采样的反演锚点插值源图像和目标图像之间的直接路径来缓解反演误差，但它们仍然受到估计不准确和反演模式固有过度对齐的影响。此外，(Cao et al., 2023; Hertz et al., 2022) 强加路径约束的技术通常需要入侵性修改，依赖于特定模型的设计并增加资源需求。基于微调的方

*Correspondence to: Kani Chen <makchen@ust.hk>.

法 (Mokady et al., 2023; Zhang et al., 2023) 通过调整文本嵌入或参数来将输出与源图像对齐, 然而它们需要大量计算并可能导致过拟合, 从而损害一般生成能力。

为了解决这些局限性, 我们提出了 TweezeEdit, 这是一种无须调整和反演的高效且语义一致的图像编辑框架。不同于仅依赖反演锚点的方法, 我们的方法规范了源图像和目标图像之间的整个去噪路径差异。这种策略类似于收紧镊子的手臂, 限制扩散模型的输出, 仅限于提示相关区域进行编辑。我们的方法不仅增强了源图像的保留, 还缩短了直接路径。此外, 我们采用基于梯度的正则化来引导沿该路径的更新, 消除了对结构修改的需求。使用一致性模型作为骨干, TweezeEdit 减少了采样步骤和累积错误, 实现了高效编辑, 同时通过与一致性模型的连接, 自然扩展到噪声和速度预测模型。

通过大量实验, 我们证明了 TweezeEdit 在提示对齐和语义保留方面优于当前最先进的方法。作为一种与架构无关的解决方案, TweezeEdit 无缝集成了注意力控制来进行优化。定量和定性评估突出了它在一致性关键任务和感知质量方面的优越性。利用一致性模型, TweezeEdit 在大约 12 个采样步骤中实现编辑, 在保持质量的前提下减少了延迟。

我们的贡献有三个方面:

1. 我们通过对整个去噪路径进行正则化来扩展反转锚定范式, 从而增强源语义的保留并缩短编辑路径。
2. 我们提出了 TweezeEdit, 这是一种通过梯度引导的编辑算法, 它避免了反演和架构变更, 并且通过一致性模型加速。
3. 我们通过实验证明了 TweezeEdit 在编辑任务中的有效性和效率。

在基于扩散的图像编辑中, 从源图像及其描述 (源提示) 开始, 中心挑战在于根据目标提示修改图像, 同时保持与源图像的一致性。

基于微调的方法 (Ruiz et al., 2023; Dong et al., 2023; Zhang et al., 2023; Mokady et al., 2023) 通过优化文本嵌入或模型参数来强制模型根据源提示重构源图像。虽然这些方法在保持一致性方面有效, 但计算量很大, 并且由于过拟合可能会损害模型的生成能力。

无调节方法利用预训练的扩散模型进行图像编辑, 无需微调。这些方法依赖反转锚点 (即来源图像的反转噪声) 以保持结构和语义一致性, 但由于反转误差, 它们在重构保真度方面面临挑战。例如, 在估计初始噪声时, DDIM (Song et al., 2020a) 在其反向去噪过程中存在误差累积 (Mokady et al., 2023; Miyake et al., 2023)。RF-inversion (Rout et al., 2024) 通过整合基于源图像的条件向量场提高一致性, 但仍受到反转不准确的限制。此外, 对这些向量场的直接修改可能会降低生成质量或引入语义失真。FlowEdit (Kulikov et al., 2024) 使用基于采样的反转噪声近似来避免显式反转, 但仍然在锚点精度上存在困难。Virtual Inversion (Xu et al., 2024) 为源和目标图像生成随机路径, 并利用源路径的重构误差校准目标路径。然而, 由于路径的随机性, 这种方法需要额外的约束。

基于注意力的方法, 无需调整的一类方法, 使用扩散模型的注意力机制来指导图像编辑。在基于 U-Net 的模型中 (Rombach et al., 2022), 有些方法 (Hertz et al., 2022; Cao et al., 2023) 在生成过程中通过利用源图像重建的注意力图来优化注意力。StableFlow (Avrahami et al., 2024) 将这些方法扩展到基于变压器的模型 (Peebles & Xie, 2023), 提高了源目标的一致性。然而, 这些方法需要进行架构更改并且计算量更大, 降低了可扩展性和效率。

2 预备知识

2.1 扩散模型

扩散模型 (Ho et al., 2020) 是生成模型, 学习逆转逐渐增加的噪声过程来生成图像。在实践中, 扩散模型通常在一个潜在空间 (Rombach et al., 2022) 中运行, 通过一个编码器-解码器架构对图像进行编码而获得。为了符号的简单性, 我们将扩散过程表述为直接作用于图像 z_0 上。

在前向扩散 ($t = 0 \rightarrow +\infty$, 在实际操作中离散化为 $t \in \{0, 1, \dots, T\}$) 中, 干净的图像 z_0 逐渐被破坏成高斯噪声。在时间步 t 的噪声图像 z_t 如下:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (1)$$

其中 $\alpha_{1:T}$ 和 $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ 是调度参数。当 $T \rightarrow \infty$ 时, $\bar{\alpha}_T \rightarrow 0$ 确保 $z_T \rightarrow \mathcal{N}(0, I)$ 。

根据 DDIM (Song et al., 2020a), 重建过程通过迭代去噪来逆转前向扩散, 具体如下: 从 $z_T \sim \mathcal{N}(0, I)$ 开始

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(z_t, t), \quad (2)$$

, 其中 $\epsilon_\theta(z_t, t)$ 是经过训练的噪声预测网络输出。该离散更新对应于确定性常微分方程的数值求解器 (Song et al., 2020b)。

一致性模型 (Song et al., 2023; Luo et al., 2023), 这是一类扩散模型, 通过在时间步长中强制自洽来提高采样效率。这些模型学习一个映射 $f(z_t, t)$, 从噪声输入 z_t 直接预测出清晰图像 z_0 。实际上, 它们采用多步一致性采样间接优化 z_0 :

$$\hat{z}_{t-k} = \sqrt{\bar{\alpha}_t} f_\theta(\hat{z}_t, t) + \sigma_t \epsilon, \quad (3)$$

, 其中 $\epsilon \sim \mathcal{N}(0, I)$, 并且 σ_t 表示时间步长 t 的噪声尺度。通过允许更大的步长 k , 一致性模型能够在较少的采样步骤中生成图像。

基于扩散模型的无调谐图像编辑方法涉及两个步骤的过程: 反演, 然后去噪。在反演过程中, 源图像 z_0^{src} 被反演为噪声 (反演锚点), 可以通过 DDIM 反演显式地实现, 也可以通过采样 (Kulikov et al., 2024) (即, $z_0^{src} \rightarrow z_T$) 隐式地实现。在随后的去噪阶段, 在目标提示的指导下, 对 z_T 进行去噪, 以生成目标图像 z_0^{tar} , 旨在与源图像保持一致的同时使输出与目标提示对齐。

一些方法 (Kulikov et al., 2024) 构建一个更新目标 $z_t^{mix} = z_0^{src} + z_t^{tar} - z_t^{src}$, 在更新过程中整合 z_t^{src} 的语义以增强一致性, 而不是仅仅依赖于 z_T 。这在数学上对应于 z_0^{src} 和 z_0^{tar} 之间的一条直接路径。其他方法 (Tumanyan et al., 2023; Cao et al., 2023; Hertz et al., 2022) 通过扩散模型的注意力控制来强制一致性。

3 TweezeEdit

当前的方法依赖于反演锚点来保持与源图像的一致性。然而, 这些锚点的不准确估计往往导致糟糕的一致性保持。为了解决在保持源图像语义的同时实现与目标提示对齐的双重挑战, 我们提出了 TweezeEdit, 一个一致且高效的图像编辑框架。我们的方法基于在 Figure 2 中展示的一个关键直觉 (相应元素在括号中标记):

Key Intuition :

- ❖ Desired edited image : The desired edited image (●) in the target prompt's distribution stays close to the source image (●), ensuring both semantic consistency and prompt alignment.
- ❖ Desired anchor : Source and desired edited images derive from an identical noise (regularized anchor, ★) with a tightened denoising path (→★→) instead of source image's DDIM-inverted noise (inversion anchor, ◆).

基于这一原则, TweezeEdit 集成了两个核心组件。首先, 我们利用一致性模型沿直接路径 (Section 3.1) 逐步整合目标语义。通过减少采样步骤和结合校准技巧, 该模型减轻了累积误差并增强了对齐。其次, 我们引入去噪路径正则化 (Section 3.2), 通过约束整个去噪路径而不是仅依赖反演锚定来增强语义保留。这一方法还缩短了直接路径——类似于镊子收紧机械臂, 使得正则化在保持原始信息完整性的同时增强编辑聚焦。

3.1 具有一致性模型的直接路径

直接路径插值。为了将目标提示注入源图像, 我们使用一致性模型在它们之间构建一个插值路径。我们首先将源图像和目标图像之间的理想直接路径定义为:

$$z_t^{mix} = z_0^{src} + \sqrt{\bar{\alpha}_t}(z_0^{tar} - z_0^{src}), \quad (4)$$

其中 $\bar{\alpha}_t$ 是扩散噪声调度器 (具有 $\bar{\alpha}_0 = 1$ 和 $\bar{\alpha}_T = 0$), $z_t^{mix} (t = 1, \dots, T)$ 表示满足 $z_T^{mix} = z_0^{src}$ 和 $z_0^{mix} = z_0^{tar}$ 的插值。

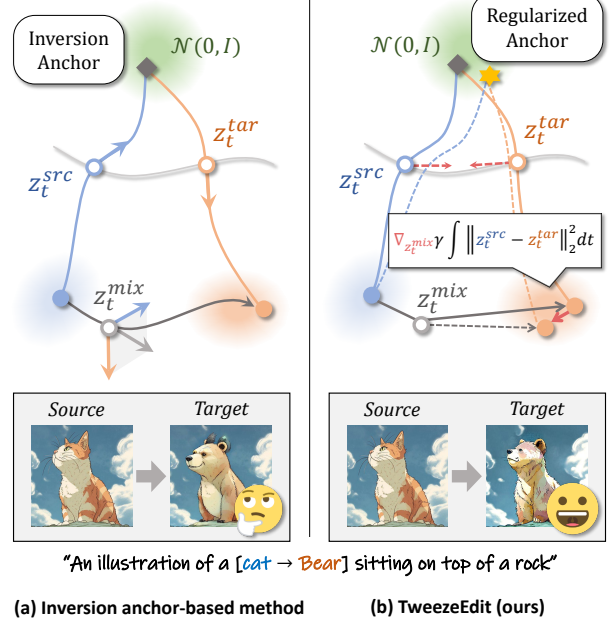


Figure 2: Comparison of inversion anchor-based method (a) and TweezeEdit (b). The inversion anchor-based method follows either the inversion-denoising path (→◆→) or equivalent direct path (→), frequently over-aligning with the target prompt due to inadequate retention of source semantics. TweezeEdit navigates the direct path via a consistency model, implicitly calibrating the anchor through the gradient (→) of denoising path regularization. Our method tightens subsequent denoising paths (shown as dashed lines), producing target images that better preserve source content.

由于 z_0^{tar} 不可用, 我们使用以下方法近似 z_t^{mix} :

值得注意的是, z_t^{src} 和 z_t^{tar} 都是从具有共享噪声 ϵ 的扩散过程中得出的, 即 $z_t^{src/tar} = \sqrt{\alpha_t} z_0^{src/tar} + \sqrt{1 - \alpha_t} \epsilon$ 。理论上, 共享噪声最小化了 z_t^{src} 和 z_t^{tar} 之间距离的上界, 而在经验上它可以防止噪声伪影 (详情请参见附录 A.1)。为了符号简洁, src/tar 表示该方程适用于两种情况。这种共享噪声公式产生了关键关系 $\sqrt{\alpha_t}(z_0^{tar} - z_0^{src}) = z_t^{tar} - z_t^{src}$, 因此 z_t^{mix} 可以等效地表示为:

$$z_t^{mix} = z_0^{src} + z_t^{tar} - z_t^{src}. \quad (5)$$

z_t^{src} 和 z_t^{tar} 的演化使一致性模型能够迭代地完善 $f(z_t^{tar}, t)$, 不断产生对 z_0^{tar} 直接路径模拟的更好近似。

基于一致性模型的编辑。一个直接的方法是从给定的 z_0^{src} 中采样 z_t^{src} , 并使用 Eq 5 将 z_t^{tar} 表示为 $z_t^{mix} - z_0^{src} + z_t^{src}$ 。然后从一致性模型中获得去噪的 $f(z_t^{src}, t)$ 和 $f(z_t^{tar}, t)$, 并根据 Eq ?? 更新 z_{t-1}^{mix} , 迭代直到 $z_0^{mix} = z_0^{tar}$ 。

然而, 这种方法面临两个挑战: (1) 采样的 z_t^{src} 可能不符合预期的锚点。我们通过噪声路径正则化进行即时噪声校正来解决这个问题, 详见 Section 3.2。 (2) $f(z_t^{tar}, t)$ 中的更新错误可能会降低与目标提示的对齐度。TweezeEdit 通过采用一致性模型作为去噪算法来解决此问题, 有效减少累积错误, 并启用校准技巧以改善目标对齐度。

当已知 z_0 时, 可以准确推导出 $f(z_t^{src}, t)$ 的预测误差。在此基础上, 我们通过应用来自 (Ju et al., 2023) 的校准技巧, 提高 $f(z_t^{tar}, t)$ 的精确度, 调整预测为:

$$\hat{f}(z_t^{tar}, t) = f(z_t^{tar}, t) + z_0^{src} - f(z_t^{src}, t), \quad (6)$$

, 通过利用源和目标之间的去噪路径相似性来校准 $f(z_t^{tar}, t)$ 。然后, z_{t-1}^{mix} 的预测 $\hat{z}_{t-1}^{mix}$ 变为:

$$\begin{aligned} \hat{z}_{t-1}^{mix} &= z_0^{src} + \sqrt{\alpha_{t-1}}(\hat{f}(z_t^{tar}, t) - z_0^{src}) \\ &= z_0^{src} + \sqrt{\alpha_{t-1}}(f(z_t^{tar}, t) - f(z_t^{src}, t)). \end{aligned} \quad (7)$$

利用 Eq 7, 我们可以通过一致性模型逐步将 z_t^{mix} 从 z_0^{src} 编辑为 z_0^{tar} 。与其他扩散模型相比, 一致性模型提供了两个优势: (1) 更少的采样步骤和自洽约束有助于减轻潜在的累积误差, 同时不会受到在 DDIM 反演中引入的累积误差的影响 (Mokady et al., 2023; Miyake et al., 2023)。一致性模型的稳定性允许使用更大的步长 ($t \rightarrow t - k$) 和更少的迭代次数, 从而提高编辑效率。 (2) 不同于预测噪声或速度的扩散模型, 一致性模型本质上对噪声具有鲁棒性, 允许其在各种噪声水平下进行去噪 (Song et al., 2023)。此特性自然符合我们即兴噪声正则化框架。

3.2 去噪路径正则化

反转锚点的作用。反转锚点用于保持编辑图像与源图像之间的一致性。然而, 它们在保持源语义方面的有效性依赖于两个在实践中常常不成立的假设:

1. 反演锚定点完全保留了来自源图像的信息。虽然在理论上有效, 但在实际中, 反演锚定点的估计会受到离散化和逼近误差的影响, 从而导致信息丢失。
2. 扩散模型更新 ($t \rightarrow t - 1$) 仅修改与提示差异对应的区域, 以确保 z_t^{src} 和 z_t^{tar} 之间的语义一致性。实际上, 更新会不受控制地修改现有元素, 例如在 Figure 2 (a) 中引入不需要的举止或体型变化。

总之, 反转锚点在配对的 $(z_T^{src}, z_T^{tar}) \cdots (z_t^{src}, z_t^{tar}) \cdots (z_0^{src}, z_0^{tar})$ 中期待语义一致性, 这本质上去噪路径 $z_T^{src} \cdots z_0^{src}$ 与 $z_T^{tar} \cdots z_0^{tar}$ 之间的相似性。因此, 与其仅依赖不准确的反转锚点, 规范整个去噪路径的相似性更为有效。

去噪路径正则化。在 Eq 7 中更新 z_t^{mix} 的过程中, 我们将连续去噪路径的距离作为正则化项加入, 定义为

$$R_t = \gamma_t \int_{t-1}^t \|z_\tau^{src} - z_\tau^{tar}\|_2^2 d\tau, \quad (8)$$

, 其中 γ_t 表示在步骤 t 时预定义的正则化强度。

为了逼近连续时间积分, 我们使用积分中值定理结合泰勒展开, 推导出 z_t^{mix} 的梯度正则化项 (详细信息请参见附录 A.2):

$$\nabla_{z_t^{mix}} R_t \approx \hat{\gamma}_t \left[z_t^{src} - z_t^{tar} - \frac{\dot{\alpha}_t}{4\sqrt{\alpha_t}} (f(z_t^{src}, t) - f(z_t^{tar}, t)) \right], \quad (9)$$

Algorithm 1: 夹取编辑

Input : Source image z_0^{src} , source prompt P^{src} , target prompt P^{tar} , regularization scheduler $\hat{\gamma}_t$ and consistency model f

Output: Edited image z_0^{tar}

$z_T^{mix} = z_0^{src}$ // 初始化;

for $t \leftarrow T$ **to** 1 **do**

 在去噪路径中获取样本

 Sample $\epsilon \sim \mathcal{N}(0, I)$;

$z_t^{src} = \sqrt{\bar{\alpha}_t} z_0^{src} + \sqrt{1 - \bar{\alpha}_t} \epsilon$;

$z_t^{tar} = z_t^{mix} - z_0^{src} + z_t^{src}$;

 一致性模型的预测

$\hat{z}_0^{src} = f(z_t^{src}, t, P^{src})$;

$\hat{z}_0^{tar} = f(z_t^{tar}, t, P^{tar})$;

 直接路径中的编辑方向

$v_t = z_0^{src} + \sqrt{\bar{\alpha}_{t-1}}(\hat{z}_0^{tar} - \hat{z}_0^{src})$ // Eq 7;

 路径正则化的梯度

$\nabla_{z_t^{mix}} R_t = \hat{\gamma}_t \left[z_t^{src} - z_t^{tar} - \frac{\bar{\alpha}_t}{4\sqrt{\bar{\alpha}_t}}(\hat{z}_0^{src} - \hat{z}_0^{tar}) \right]$ // Eq 9;

 更新步骤

$z_t^{mix} = v_t - \nabla_{z_t^{mix}} R_t$ // Eq 10;

end

return Edited image $z_0^{tar} = z_0^{mix}$

这里我们设置 $\Delta_t \approx \frac{1}{2}$ 并定义 $\hat{\gamma}_t := 2\gamma_t(-1 + \frac{\bar{\alpha}_t}{4\bar{\alpha}_t})$ 以简化处理。关于选择 $\hat{\gamma}_t$ 的更多细节可以在附录 A.3 中找到。

最后, Eq 10 中显示的 z_t^{mix} 更新可以分为两个部分: 朝向目标提示的编辑方向和保持源图像语义的方向。我们的算法在 Algorithm 1 中进行了总结。出于阐述清晰的目的, 我们将更新间隔设置为 1。我们的方法自然可以扩展到更大的间隔。在实践中, 我们只在 $\{1, \dots, T\}$ 中选择 12-15 个时间步长, 这大大减少了计算步骤, 同时保持了性能。

$$z_{t-1}^{mix} = \underbrace{z_0^{src} + \sqrt{\bar{\alpha}_{t-1}}(f(z_t^{tar}, t) - f(z_t^{src}, t))}_{\text{target editing}} - \underbrace{\nabla_{z_t^{mix}} R_t}_{\text{source preserving}}. \quad (10)$$

通过在每次更新步骤中加入正则化项 R_t , 我们在整个去噪路径上执行一致性, 如 $\sum_{t=1}^T R_t = \int_{t=0}^T \gamma_\tau \|z_\tau^{src} - z_\tau^{tar}\|_2^2 d\tau$ 。这要求 z_t^{tar} 保留源图像的语义, 仅引入与目标提示相关的变化以避免梯度惩罚。与反转锚点不同, 我们的方法在整个去噪路径上扩展正则化, 并动态校准正则化的锚点, 增强语义一致性。

此外, 我们的方法不需显式地计算锚点。相反, 梯度驱动的更新引导 z_t^{mix} 沿直接路径, 从而规避了两个限制: (1) 减少编辑路径长度。传统的反演锚点需要计算强度大且不精确的 DDIM 反演, 限制了效率。即使通过采样 (Kulikov et al., 2024) 来估计锚点, 不一致的去噪路径也会导致更长的直接路径。通过正则化去噪路径, 我们缩短了直接路径。(2) 架构无关的更新。先前的方法通常通过特定架构的干预 (例如, 注意力注入) 来弥补语义损失。我们的梯度驱动正则化在不干涉模型的情况下进行, 确保了跨各种架构的兼容性并减少了计算开销。

尽管我们的方法基于一致性模型, 具有自一致性和更少降噪步骤等优势, 但由于干净图像预测与其输出之间的固有关系, 它仍可以应用于其他噪声或速度预测模型 (详细信息见 A.4)。同时, 连续一致性模型 (Lu & Song, 2024) 允许将这些模型转化为一一致性模型。

4 实验

我们在 PIE-Bench (Ju et al., 2023) 上评估我们的方法, 该基准包含来自 4 个类别 (动物、人物、室内场景、室外场景) 的 700 张图像, 并涵盖 10 种类型的编辑任务, 如对象添加、修改和风格转移。每个实例提供源提示、目标提示以及指示编辑区域的掩码。详细的实验设置在附录 B.1 中描述。

我们通过多项指标评估编辑的保真度: 用结构距离评估全局一致性 (Tumanyan et al., 2023), MSE、PSNR、SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018) 用于评估未编辑区域的保留情况, 以及用 CLIPScore (Hessel et al., 2021) 衡量目标提示与编辑结果之间的对齐程度。

Method	Structure		Unedited Region Preservation			Editing Alignment		Efficiency	
	Distance $_{10^3}$ ↓	PSNR ↑	LPIPS $_{10^3}$ ↓	MSE $_{10^4}$ ↓	SSIM $_{10^2}$ ↑	Whole ↑	Edited ↑	Inv-Free	Steps ↓
DDIM (SD1.5)	79.54	17.36	220.13	243.13	70.52	27.08	23.90	✗	50
DDIM (SD1.5) + P2P	69.99	17.87	208.90	219.56	71.63	25.28	22.57	✗	50
FlowEdit (Flux)	22.77	23.08	100.96	74.31	86.29	25.19	22.29	✓	28
RF-inversion (Flux)	55.08	19.27	227.59	164.21	66.78	25.22	22.49	✗	28
Stable-flow (Flux)	16.44	24.24	76.10	64.41	89.43	23.98	20.96	✗	50
VI (LCM: SD1.5)	113.83	13.93	292.99	518.98	59.37	27.68	24.37	✓	15
VI (LCM: SD1.5) + P2P	27.86	21.82	86.62	124.45	80.89	24.76	21.71	✓	15
TweezeEdit (SD1.5)	23.96	22.30	82.62	83.61	82.11	25.87	22.45	✓	25
TweezeEdit (Flux)	20.92	23.49	82.72	72.87	87.70	25.23	22.30	✓	28
TweezeEdit (LCM: SD1.5)	17.36	24.62	81.90	54.42	80.40	25.54	22.30	✓	12
TweezeEdit (LCM: SD1.5)+P2P	13.63	25.59	67.36	43.71	82.65	24.75	21.61	✓	12
TweezeEdit (LCM: SDXL1.0)	22.42	24.13	98.45	57.33	83.43	26.01	22.79	✓	15

Table 1: PIE-Bench 的定量结果。Inv-Free 表示是否避免了显式求逆。Whole 和 Edited 分别指整个图像和编辑区域的 CLIPScore。↑：值越高越好，↓：值越低越好。加粗：最佳结果。我们的方法 TweezeEdit 以更少的步骤实现了具有竞争力的编辑，并在未编辑区域保持高度的一致性。



Figure 3: TweezeEdit 在路径正则化强度上的表现。从左到右：正则化强度的增加提升了源的一致性，其中适当的强度在一致性和目标对齐之间达到平衡。

我们将 TweezeEdit 与代表性的无调优方法进行比较：DDIM (Song et al., 2020a), VI (Xu et al., 2024), FlowEdit (Kulikov et al., 2024) 和 RF-inversion (Rout et al., 2024)。DDIM 使用 SD1.5 (Rombach et al., 2022) 实现，而 VI 采用 SD1.5 的潜在一致性模型 (LCM) (Luo et al., 2023) 变体。FlowEdit 和 RF-inversion 使用 Flux (Labs, 2024)。TweezeEdit 利用 SD1.5、Flux 以及来自 SD1.5 和 SDXL1.0 (Podell et al., 2023) 的 LCM。此外，我们评估了 DDIM 和 VI 的 P2P 增强 (Hertz et al., 2022) 版本，其中 P2P 是一种基于 Unet (Rombach et al., 2022) 的注意力控制方法。我们还包括使用 Flux 的 StableFlow，这是一种基于 DiT (Peebles & Xie, 2022) 的注意力控制方法。实现细节在附录 B.2 中提供。

我们评估了 TweezeEdit 在保持一致性和生成高质量编辑方面的有效性和效率。定量结果。Table 1 表明，TweezeEdit 在多种范式下表现出强劲的鲁棒性，包括噪声 (SD1.5)、速度 (Flux) 和纯净图像预测 (LCM)，在保持一致性方面表现出色，同时保持了编辑性能。指标比较基于均值，p 值使用 Wilcoxon 符号秩检验计算。在 SD1.5 上，TweezeEdit 在 LPIPS 上比 DDIM 表现得更好，超过了 -137.51 ($p < 0.01$)。在 Flux 上，TweezeEdit 在所有一致性指标上均优于 RF-inversion 和 FlowEdit，同时获得更高的整体 CLIPScore。Stable-flow 为了保持一致性牺牲了可编辑性，在整体和编辑的 CLIPScore 中远低于 TweezeEdit (分别为 -1.25 、 $p < 0.01$ 和 -1.34 、 $p < 0.01$)。在 LCM 上，即使 VI 使用 P2P 进行了增强以提高一致性，TweezeEdit (不含 P2P) 在一致性和对齐上都超过了它 (PSNR: $+2.8$, $p < 0.01$ ；编辑的 CLIPScore: $+0.59$, $p < 0.01$)。我们的方法在仅有 12 步推理和自一致性情况下，在 LCM 上表现显著优于在 Flux 上的表现。与 SD1.5 相比，它在保持一致性方面取得了显著的改进 (MSE: -29.19 , $p < 0.01$)，编辑成本极小 (编辑的 CLIPScore: -0.15 , n.s.)。整合 P2P 进一步提高了我们方法的一致性，代价是对齐略有下降。除此之外，TweezeEdit 与架构无关的设计使得其能够轻松适配 SDXL1.0 以实现更好的对齐。

如 Figure ?? 所示，我们的方法在编辑图像时有效地保留了源保真度。例如，(1) 它将鹿的姿势从向后改为向前，同时保留其身份特征 (Figure ?? (a-1))；(2) 它在不改变背景的情况下将水果替换为比萨 (Figure ?? (a-2))；(3) 它给猫添加了帽子，同时保持其原始姿势和视觉属性 (Figure ?? (a-3))。相比之下，DDIM 在错误累积和过度对齐目标提示方面存在困难。VI+P2P 在一致性保留方面效果突出，但在编辑任务中失败，例如在 Figure ?? 中遗漏了帽子。FlowEdit 和 RF-Inversion 会引入不必要的伪影，修改背景 (Figure ?? (a-2, a-3)) 或改变角色属性 (Figure ?? (a-1, a-3))。更多视觉比较可在附录 ?? 中找到。

Metrics	Path Regularization Steps (#)				
	0	2	4	6	8
SD $\uparrow$	82.1	57.96	34.25	17.36	9.71
PSNR $\uparrow$	15.84	17.82	21.03	24.62	27.19
LPIPS $_{10^3} \downarrow$	231.04	183.33	125.30	81.90	60.23
MSE $_{10^4} \downarrow$	345.00	222.09	113.37	54.42	32.42
SSIM $_{10^2} \uparrow$	64.96	69.99	75.69	80.40	82.89
Whole $\uparrow$	27.08	26.78	26.43	25.54	24.22
Edited $\uparrow$	23.76	23.77	23.27	22.30	21.15

Table 2: 在路径正则化中不同早期步骤的表现（总步骤数：12；粗体值表示最佳结果）。增加步骤可以提高一致性，但会降低目标对齐度，最佳折中点在 6 步。

4.1 路径正则化分析

路径正则化对于 TweezeEdit 至关重要，它在保持图像一致性的同时实现灵活编辑。我们通过三个方面分析其影响：早期步骤正则化、与基于注意力的方法的集成，以及在一致性与目标对齐之间的平衡。另外，附录 C.3 的实验确认路径正则化对随机起点和轻微梯度强度扰动保持了鲁棒性。

早期步骤正则化。在扩散模型中，早期生成步骤在塑造图像结构方面起着至关重要的作用。因此，我们将路径正则化限制为 12 步过程中的前 m 步。正如 Table 2 所示，增加 m 可以增强未编辑区域的一致性，但逐渐降低编辑能力。为了找到最佳平衡，将路径正则化应用于一半的步骤（12 步中的 6 步）可以在保留源图像的真实性的同时，避免对后期步骤的过度编辑限制。

与注意力方法的协同作用。路径正则化与注意力控制框架兼容。Figure ?? (b) 表明，路径正则化与 P2P 的结合在显著提高输出一致性的同时减少了独立 P2P 中的伪影。

平衡一致性与对齐。路径正则化使用户能够校准一致性和文本对齐之间的权衡。如 Figure 3 所示，增加正则化强度会使输出从严格的文本对齐转向结构保留（例如，在将猫转换为狗时保留背景）。这种可调性支持多样化的定制需求而无需架构修改，使得 TweezeEdit 能够适应以一致性为重点和以创造性为驱动的编辑场景。

我们使用 IR (Xu et al., 2023)、HPSv2 (Wu et al., 2023)、PickScore (Kirstain et al., 2023) 和 AES (Schuhmann et al., 2022) 评估编辑图像的视觉质量。我们的比较包括 TweezeEdit、FlowEdit、RF-Inversion 和基于 Flux 的 Stable Flow。我们的方法 TweezeEdit 在所有指标上都取得了最佳表现。例如，它的 IR 得分比第二好的方法高出 3.81。详细结果可以在附录 C.1 中找到。

4.2 基于文本的翻译编辑

基于文本的翻译编辑任务需要极强的一致性维护，尤其是在进行大量视觉修改时。正如在 Figure 4 中所展示的，TweezeEdit 能够有效地在翻译任务中保持语义一致性。

在这项工作中，我们解决了扩散模型中文本驱动图像编辑中语义忠实度损失的关键挑战，其中现有方法过度对齐目标提示，并由于反演不准确和无约束的去噪而无法保留源内容。我们的框架 TweezeEdit 引入了一种无反演范式，它对源图像和目标图像之间的去噪轨迹进行正则化，从而通过一致性模型实现高效编辑。通过限制与提示相关区域的偏差，并使用梯度引导更新，TweezeEdit 在不改变架构的情况下实现了与源一致、与目标对齐的编辑。广泛的评估表明其在各种任务中的卓越效果和效率，凸显了路径正则化在弥合创作意图与生成模型限制之间差距的潜力。

References

Avrahami, O., Patashnik, O., Fried, O., Nemchinov, E., Aberman, K., Lischinski, D., and Cohen-Or, D. Stable

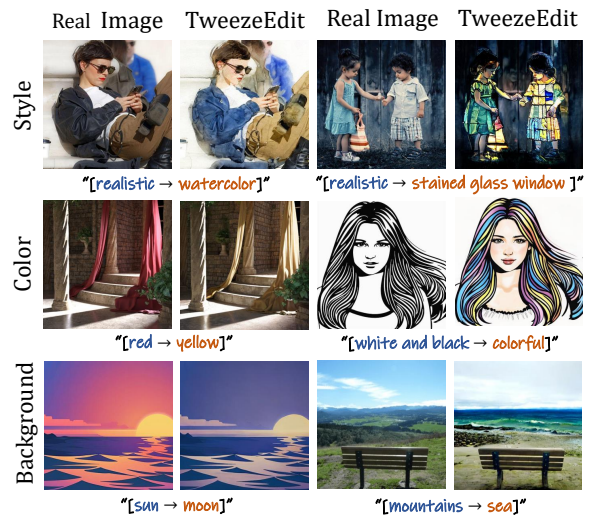


Figure 4: TweezeEdit 在一致性关键翻译任务中的表现。

- flow: Vital layers for training-free image editing. *arXiv preprint arXiv:2411.14430*, 2024.
- Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., and Zheng, Y. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22560–22570, 2023.
- Dong, W., Xue, S., Duan, X., and Han, S. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7430–7440, 2023.
- Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- Hessel, J., Holtzman, A., Forbes, M., Min, R. L. B., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Ju, X., Zeng, A., Bian, Y., Liu, S., and Xu, Q. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023.
- Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.
- Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., and Michaeli, T. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024.
- Kwon, G. and Ye, J. C. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022.
- Labs, B. F. Flux. <https://github.com/black-forest-labs/flux>, 2024. URL <https://github.com/black-forest-labs/flux>.
- Lu, C. and Song, Y. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024.
- Luo, S., Tan, Y., Huang, L., Li, J., and Zhao, H. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.
- Miyake, D., Iohara, A., Saito, Y., and Tanaka, T. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023.

- Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.
- Peebles, W. and Xie, S. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.
- Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.
- Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Rout, L., Chen, Y., Ruiz, N., Caramanis, C., Shakkottai, S., and Chu, W.-S. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.
- Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.
- Shuai, X., Ding, H., Ma, X., Tu, R., Jiang, Y.-G., and Tao, D. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024.
- Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020a. URL <https://arxiv.org/abs/2010.02502>.
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.
- Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

- Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. 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*, pp. 1921–1930, 2023.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., and Li, H. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.
- Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935, 2023.
- Xu, S., Huang, Y., Pan, J., Ma, Z., and Chai, J. Inversion-free image editing with language-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9452–9461, 2024.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.
- Zhang, S., Xiao, S., and Huang, W. Forgedit: Text guided image editing via learning and forgetting. *arXiv preprint arXiv:2309.10556*, 2023.
- Zhou, Z., Chen, D., Wang, C., and Chen, C. Fast ode-based sampling for diffusion models in around 5 steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7777–7786, 2024.

5

附录

A 方法

A.1 共享噪声假设

在主文的 Section 3.1 中, 我们假设 z_t^{src} 和 z_t^{tar} 具有相同的噪声。从理论上讲, 这一假设通过消除 Eq 5 中的随机项, 使 z_t^{src} 和 z_t^{tar} 之间的距离的上界最小化。从经验上看, 它减少了由随机性引起的差异, 否则这些差异会产生如 Figure 5 中所示的噪声结果。

A.2 路径正则化梯度

为了近似连续时间上的积分, 我们应用积分中值定理, 并在 t 处展开泰勒级数: 对于某些 $\Delta_t \in [0, 1]$,

$$\begin{aligned} R_t &= \gamma_t \|z_{t-\Delta_t}^{src} - z_{t-\Delta_t}^{tar}\|_2^2 \\ &\approx \gamma_t \|z_t^{src} - z_t^{tar} - \Delta_t(\dot{z}_t^{src} - \dot{z}_t^{tar})\|_2^2. \end{aligned} \quad (11)$$

虽然定理在高维中并不严格成立, 但对于几乎是直线的扩散模型轨迹 (Zhou et al., 2024), 它仍然近似有效, 而一致性模型的自相容性进一步降低了这种近似的影响。

从关系 $z_t^{src} - z_t^{tar} = \sqrt{\alpha_t}(z_0^{src} - z_0^{tar})$, 我们推导出 $\dot{z}_t^{src} - \dot{z}_t^{tar} = \frac{\dot{\alpha}_t}{2\sqrt{\alpha_t}}(z_0^{src} - z_0^{tar})$, 其中 $z_0^{src/tar}$ 可以通过 $f(z_t^{src/tar}, t)$ 来近似。代入此结果得

$$R_t \approx \gamma_t \|z_t^{src} - z_t^{tar} - \frac{\Delta_t \dot{\alpha}_t}{2\sqrt{\alpha_t}}(f(z_t^{src}, t) - f(z_t^{tar}, t))\|_2^2. \quad (12)$$

随着 z_t^{mix} 沿直接路径更新, 我们使用相对于 z_t^{mix} 的 R_t 的梯度对其更新进行正则化。请注意, z_t^{src} (来自 z_0^{src} 的噪声) 与 z_t^{mix} 独立, 从而得到 $\frac{\partial z_t^{src}}{\partial z_t^{mix}} = 0$ 。从 $z_t^{tar} = z_t^{mix} - z_0^{src} + z_t^{src}$ 和近似 $z_t^{tar} \approx \sqrt{\alpha_t}f(z_t^{tar}, t) + \sqrt{1-\alpha_t}\epsilon$, 我们得到 $\frac{\partial z_t^{tar}}{\partial z_t^{mix}} = I$ 和 $\frac{\partial f(z_t^{tar}, t)}{\partial z_t^{mix}} = \frac{I}{\sqrt{\alpha_t}}$ 。因此, R_t 的梯度表示为:

$$\nabla_{z_t^{mix}} R_t \approx \hat{\gamma}_t \left[z_t^{src} - z_t^{tar} - \frac{\dot{\alpha}_t}{4\sqrt{\alpha_t}}(f(z_t^{src}, t) - f(z_t^{tar}, t)) \right], \quad (13)$$

A.3 γ_t 的选择

。在实际操作中, 我们提前简化了主文本中的 Eq 13。从 $z_t^{src} - z_t^{tar} = \sqrt{\alpha_t}(z_0^{src} - z_0^{tar})$ 出发, 并再次使用 Eq 6 近似 z_0^{tar} 。我们有:

$$\nabla_{z_t^{mix}} R_t \approx \hat{\gamma}_t [(f(z_t^{src}, t) - f(z_t^{tar}, t))], \quad (14)$$

, 其中我们设置 $\Delta_t \approx \frac{1}{2}$ 并定义 $\hat{\gamma}_t := -2\gamma_t(-1 + \frac{\dot{\alpha}_t}{4\sqrt{\alpha_t}})(-1 + \frac{\dot{\alpha}_t}{4\sqrt{\alpha_t}})$ 以简化说明。请注意, 这个 $\hat{\gamma}_t$ 的定义与主文本中的定义不同, 但是为了在这个推导中保持清晰, 我们重复使用了这个符号。为了确保语义的注入, 我们限制 $|z_t^{tar} - z_t^{src}| \geq |z_{t-1}^{tar} - z_{t-1}^{src}|$, 因此我们在 $[-(\sqrt{\alpha_t} - \sqrt{\alpha_{t-1}}), 0]$ 范围内选择 $\hat{\gamma}_t$ 。当 $\hat{\gamma}_t = 1$ 时, 我们有

$$|z_t^{tar} - z_t^{src}| = |z_{t-1}^{tar} - z_{t-1}^{src}|,$$

, 在这种情况下实现了最大程度的图像一致性保持。或者, 我们可以直接使用 $z_t^{tar} - z_t^{src} = z_{t-1}^{tar} - z_{t-1}^{src}$ 来绕过扩散模型的预测, 从而节省计算开销。相反, $\hat{\gamma}_t = 0$ 实现了最大的编辑效果。更新后的算法在 Algorithm 2 中总结。

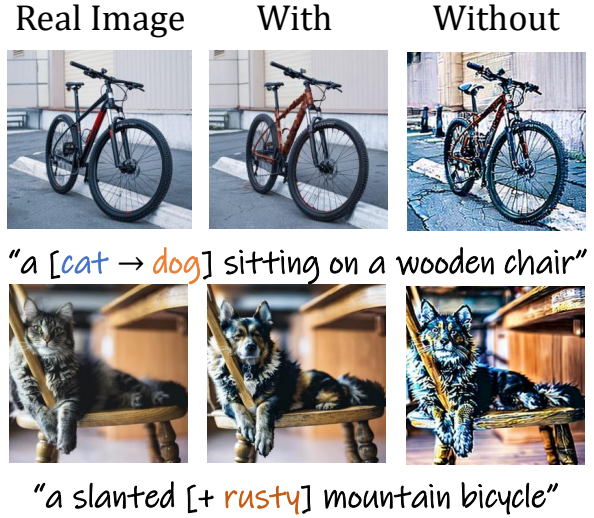


Figure 5: 编辑结果在有共享噪声和无共享噪声情况下的比较

Algorithm 2: TweezeEdit

Input : Source image z_0^{src} , source prompt P^{src} , target prompt P^{tar} , regularization scheduler $\hat{\gamma}_t$ and consistency model f

Output: Edited image z_0^{tar}

$z_T^{mix} = z_0^{src}$ // initialization ;

for $t \leftarrow T$ **to** 1 **do**

//Obtain samples in denoising path

Sample $\epsilon \sim \mathcal{N}(0, I)$;

$z_t^{src} = \sqrt{\bar{\alpha}_t} z_0^{src} + \sqrt{1 - \bar{\alpha}_t} \epsilon$;

$z_t^{tar} = z_t^{mix} - z_0^{src} + z_t^{src}$;

//Consistency model's prediction

$\hat{z}_0^{src} = f(z_t^{src}, t, P^{src})$;

$\hat{z}_0^{tar} = f(z_t^{tar}, t, P^{tar})$;

//Editing direction in direct path

$v_t = z_0^{src} + \sqrt{\bar{\alpha}_{t-1}}(\hat{z}_0^{tar} - \hat{z}_0^{src})$;

//Gradient of denoising path regularization

$\nabla_{z_t^{mix}} R_t = \hat{\gamma}_t [(\hat{z}_0^{src} - \hat{z}_0^{tar})]$;

//Update step

$z_t^{mix} = v_t - \nabla_{z_t^{mix}} R_t$;

end

return Edited image $z_0^{tar} = z_0^{mix}$

A.4 将 TweezeEdit 应用于噪声和速度预测器

噪声/速度（线性路径）预测器与清晰图像预测器之间的关系可以表示为：

$$f_\theta(z_t, t) \approx \frac{z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}}$$

$$f_\theta(z_t, t) \approx z_t - t \cdot v_\theta(z_t, t)$$

基于这种关系，我们可以利用上述公式预测 $z_0^{\hat{src}}$ 和 $z_0^{\hat{tar}}$ 来更新 Eq 10。

对于噪声预测器，

$$z_0^{\hat{src}} = \frac{z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, t, P^{src})}{\sqrt{\bar{\alpha}_t}}$$

$$z_0^{\hat{tar}} = \frac{z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, t, P^{tar})}{\sqrt{\bar{\alpha}_t}}$$

对于速度预测器，

$$\hat{z}_0^{src} = z_t - t \cdot v_\theta(z_t, t, P^{src})$$

$$\hat{z}_0^{tar} = z_t - t \cdot v_\theta(z_t, t, P^{tar})$$

B 实验细节**B.1 度量标准**

在本节中，我们具体说明所用的指标。使用了五个指标来量化一致性。

- 结构距离 (Tumanyan et al., 2023) 是一种相似度量，通过 DINO-ViT 模型 (Kwon & Ye, 2022) 的自相似距离来量化两幅图像之间的结构相似性。DINO-ViT 模型通过自监督学习提取出稳健且语义上有意义的特征。
- 峰值信噪比 (PSNR) 量化最大可能信号功率与失真噪声功率的比率，以测量重建图像与原始图像的质量。

Method	IR	AES	HPSv2	PS
FLowEdit (Flux)	73.11	28.02	22.16	6.75
RF-inversion (Flux)	69.88	27.74	21.99	6.43
Stable Flow (Flux)	30.43	27.26	21.37	6.51
TweezeEdit (Flux)	76.92	28.05	22.22	6.76

Table 3: 感知质量评估结果

- 学习感知图像块相似性 (LPIPS) (Zhang et al., 2018) 是一种感知质量度量, 通过比较从深度神经网络中提取的特征激活来衡量两幅图像之间的相似性。
- 均方误差 (MSE) 通过测量两幅图像之间平均平方差来评估其相似性。
- 结构相似性指数 (SSIM) (Wang et al., 2004) 通过比较亮度、对比度和结构信息来评估两幅图像之间的相似性。

一个指标用于量化对齐。

- CLIPScore (Hessel et al., 2021) 通过计算图像和文本之间的 CLIP 嵌入的余弦相似性来评估图像和字幕的对齐程度。整体 CLIPScore 评估整个图像与文本的匹配程度, 而编辑 CLIPScore 则侧重于图像的编辑区域与文本进行比较。

四个指标用于量化图像感知质量。

- 美学评分 (AES) (Schuhmann et al., 2022) 基于美学预测器评估图像的美学质量, 该预测器在 LAION-5B 数据集上的 CLIP 图像嵌入上进行训练, 预测从 1 到 10 的美学分数, 反映人类的主观偏好。
- ImageReward (IR) (Xu et al., 2023) 基于一个通用的文本到图像人类偏好奖励模型 (RM) 来量化人类偏好, 该模型在 137k 对专家比较数据上训练得到。
- 人类偏好评分 v2 (HPSv2) (Wu et al., 2023) 使用一个评分模型预测人类对生成图像的偏好, 该模型是一个在 HPD v2 上微调的 CLIP 模型, HPD v2 是一个由 433,760 对图像中 798,090 个有人类偏好选择的数据集。
- PickScore (PS) (Kirstain et al., 2023) 使用基于 CLIP 的评分函数预测人类对生成图像的偏好, 该评分函数是在 Pick-a-Pic 上训练的, 这是一个大型开放的数据集, 其中包含通过网络应用收集的文本到图像提示与真实用户偏好配对的信息。

B.2 实现细节

对于所有其他方法, 我们维持了它们的官方超参数设置。实验配置如下: DDIM (SD1.5) 使用 50 个生成步骤, 源和目标提示的分类器自由引导 (CFG) 尺度为 7.5。DDIM(SD1.5)+P2P 遵循与 DDIM 相同的基础参数, 同时加入 P2P 特定的设置, 自我替换步骤为 0.6, 交叉替换步骤为 0.4。Flowedit (Flux) 设置为 28 个步骤和 n-max 值为 24, 使用源提示的 CFG 尺度为 1.5, 目标提示的为 5.5。虚拟反向 (VI) (LCM: SD1.5) 使用 12 个步骤 (与 P2P 结合时扩展到 15 个), 源提示的 CFG 尺度设置为 1.0, 目标提示的为 2.3。VI(LCM: SD1.5)+P2P 配置继承了 VI 的基础参数, 同时加入自我替换步骤 1.0 和交叉替换步骤 0.7 的设置。

对于 TweezeEdit, 我们实现了以下配置。TweezeEdit (SD1.5) 使用 25 步, CFG 比例为 3.5 (源) 和 7.5 (目标), 在第一个步骤中用 $\gamma = -(\sqrt{\alpha_t} - \sqrt{\alpha_{t-1}})$ 正则化。Tweezedit (LCM: SD1.5) 使用 12 步, 对于两个提示均使用 1.5 的 CFG 比例, 在前 6 个步骤中用 $\gamma = -(\sqrt{\alpha_t} - \sqrt{\alpha_{t-1}})$ 进行路径正则化。Tweezedit (LCM: SD1.5+P2P) 维持与 Tweezedit (LCM: SD1.5) 相同的 CFG 和总步数, 并在前三个步骤中应用 $\gamma = -(\sqrt{\alpha_t} - \sqrt{\alpha_{t-1}})$ 进行路径正则化, 同时使用自替换步骤 0.6 和交替替换步骤 0.4 的 P2P 配置。TweezeEdit (Flux) 运行 28 步, 源提示 CFG 为 1.5, 目标提示 CFG 为 5.5, 在前 4 个步骤中用 $\gamma = -0.8(\sqrt{\alpha_t} - \sqrt{\alpha_{t-1}})$ 正则化。最后, TweezeEdit (LCM: SDXL1.0) 运行 15 步, 对两个提示使用相同的 CFG 值, 并在前三个步骤中用 $\gamma = -(\sqrt{\alpha_t} - \sqrt{\alpha_{t-1}})$ 正则化。

所有实验均在一台配置 Ubuntu 22.04.5 LTS 操作系统的单个 NVIDIA RTX A6000 GPU (48GB) 上进行。每个编辑任务都通过算法的一次运行来执行, 并记录结果以供评估。

Method	Steps	path regularization	time (s)
TweezeEdit(SD1.5)	12	6	1.63 $\pm$ 0.01
TweezeEdit(SD1.5)+P2P	12	3	2.49 $\pm$ 0.01
TweezeEdit(SDXL)	15	3	2.27 $\pm$ 0.01

Table 4: 微操编辑的效率

Exp	Distance (10^3)	PSNR	LPIPS (10^3)	MSE (10^4)	SSIM (10^2)	Whole	Edited
SV	17.62 $\pm$ 0.24	24.59 $\pm$ 0.08	82.50 $\pm$ 0.59	54.42 $\pm$ 0.69	80.34 $\pm$ 0.10	25.43 $\pm$ 0.10	22.30 $\pm$ 0.05
GP	17.81 $\pm$ 0.05	24.77 $\pm$ 0.40	83.33 $\pm$ 0.14	55.36 $\pm$ 0.15	80.24 $\pm$ 0.01	25.48 $\pm$ 0.01	22.36 $\pm$ 0.01

Table 5: 敏感性实验结果

C 定量结果

C.1 感知质量评估结果

我们评估了基于 Flux 的 TweezeEdit、FlowEdit、RF-Inversion 和 Stable Flow 的感知质量。如 Table 3 所示，我们的方法在多个指标上实现了最佳视觉质量，包括 IR (Xu et al., 2023)、AES (Schuhmann et al., 2022)、HPSv2 (Wu et al., 2023) 和 PS (Kirstain et al., 2023)。这表明，我们的编辑方法通过有效平衡一致性和对齐性来确保高质量的输出。

C.2 TweezeEdit 的运行时效率

如 Table 4 所示，TweezeEdit 高效地处理图像，展示了其适合实时应用的能力。这里，我们使用附录 A.2 中的绕过技巧来节省计算开销。

C.3 路径正则化的鲁棒性分析

我们通过两个关键实验 (Table 5) 实验证实了路径正则化的稳健性：(1) 种子变化 (SV)：当改变随机初始化时，结果几乎没有波动，这表明我们的方法 (公式 11) 在不同的起始条件下保持了稳定的锚点。(2) 梯度扰动 (GP)：在梯度中引入高斯噪声导致的性能变化很小，确认了对小优化扰动的不敏感性。总体而言，这些结果强调了我们的方法在随机初始化和梯度噪声下保持稳定。

在本节中，我们展示了编辑结果的视觉比较，以评估 TweezeEdit (LCM: SD1.5 无 P2P) 相对于 FlowEdit (Flux)、RF-Inversion (Flux) 和 StableFlow (Flux)。在此实验中，TweezeEdit 在其 15 步生成过程的前 10 步中应用路径正则化。定性结果显示，与使用先进基础模型 Flux 的基线方法相比，TweezeEdit 在更好地保留原始内容的同时，保持了文本对齐的编辑。



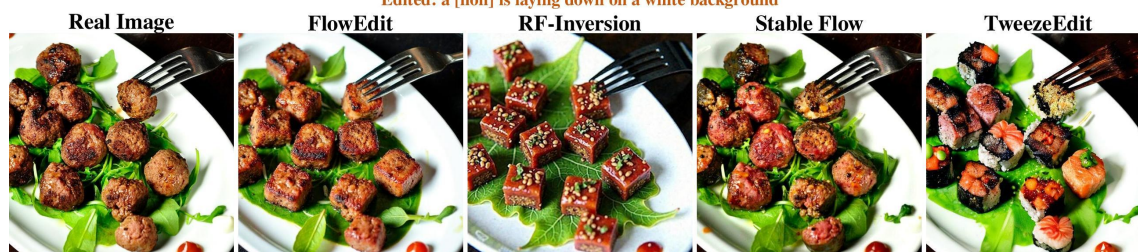
Original: a slanted mountain bicycle on the road in front of a building
 Edited: a slanted [rusty] mountain bicycle on the road in front of a building



Original: a [cat] sitting on a wooden chair
 Edited: a [dog] sitting on a wooden chair



Original: a [dog] is laying down on a white background
 Edited: a [lion] is laying down on a white background



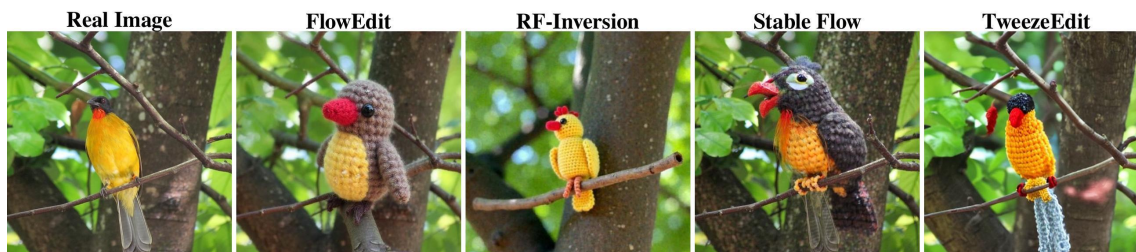
Original: meat [balls] on white plate
 Edited: meat [sushi] on white plate



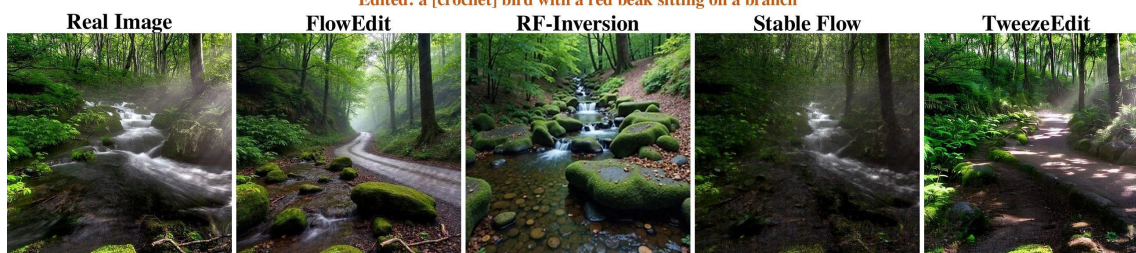
Original: photo of a statue [in front view]
 Edited: photo of a statue [in side view]



Original: white [tiger] on brown ground
 Edited: white [cat] on brown ground



Original: a yellow bird with a red beak sitting on a branch
Edited: a [crochet] bird with a red beak sitting on a branch



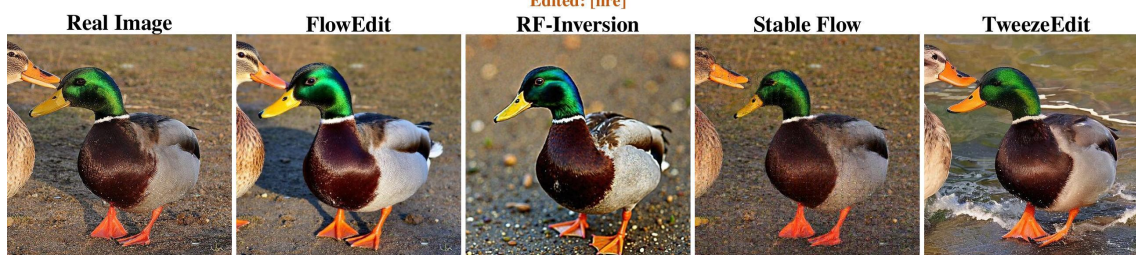
Original: a [stream] in a lush green forest with rocks
Edited: a [road] in a lush green forest with rocks



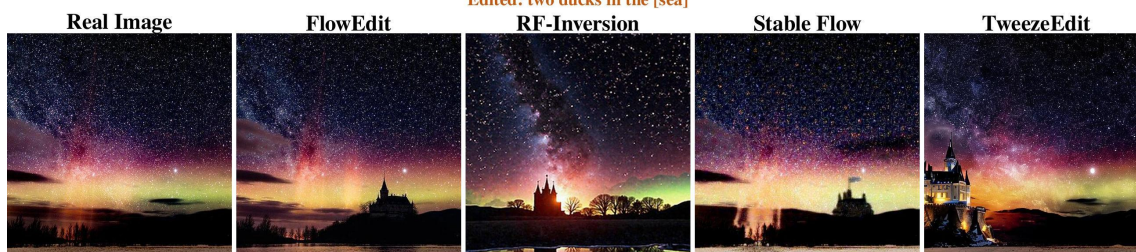
Original: a cartoon of a [boy] walking his dog in autumn
Edited: a cartoon of a [girl] walking his dog in autumn



Original: [smoke]
Edited: [fire]



Original: two ducks in the [dirt]
Edited: two ducks in the [sea]



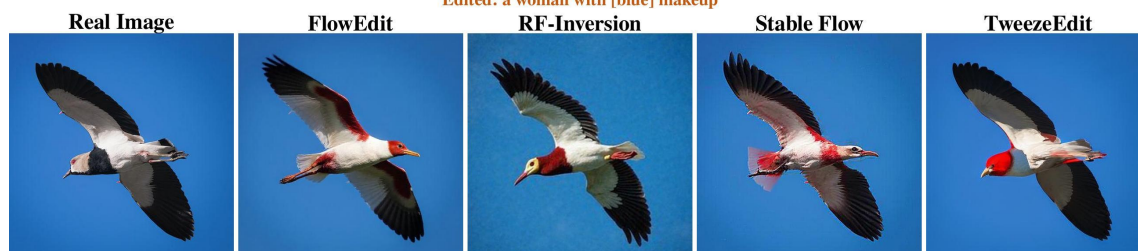
Original: the arctic sky
Edited: [a castle in] the arctic sky



Original: woman with [brown] hair
Edited: woman with [blue] hair



Original: a woman with [gold] makeup
Edited: a woman with [blue] makeup



Original: a [black and white] bird flying in the blue sky
Edited: a [red and white] bird flying in the blue sky



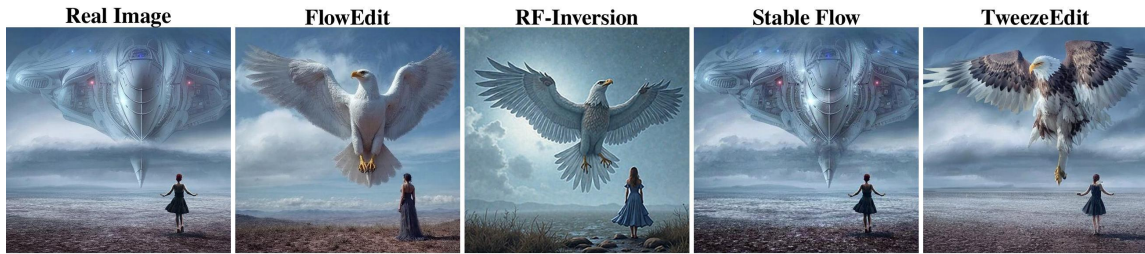
Original: a girl and her [dog] in a field
Edited: a girl and her [monkey] in a field



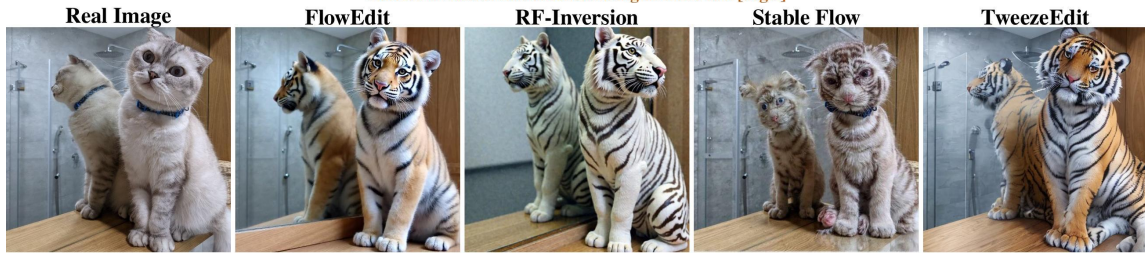
Original: a beautiful woman with [garland] on head
Edited: a beautiful woman with [hat] on head



Original: a woman's face is framed in a large [photo] on a wall
Edited: a woman's face is framed in a large [painting] on a wall



Original: a woman in a dress standing in front of a [spaceship]
Edited: a woman in a dress standing in front of a [eagle]



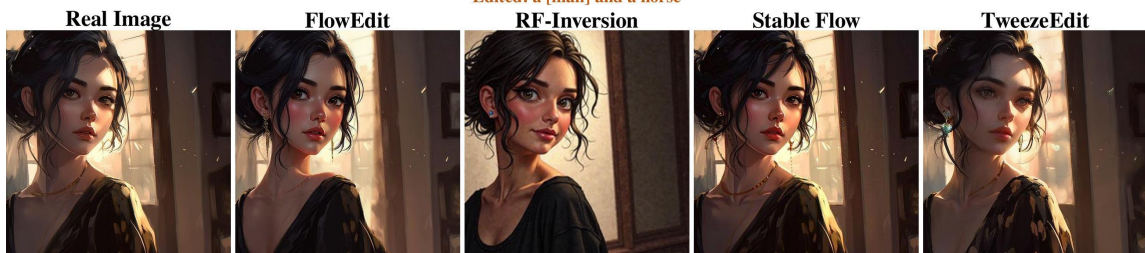
Original: a [cat] sitting next to a mirror
Edited: a [tiger] sitting next to a mirror



Original: a [horse] in a grass field
Edited: a [zebra] in a grass field



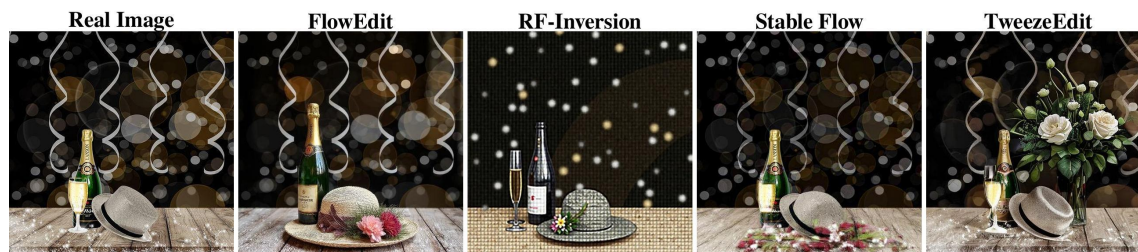
Original: a [woman] and a horse
Edited: a [man] and a horse



Original: a woman with dark hair
Edited: a woman with dark hair [wearing earrings]



Original: a black skin man with a tree head and branches on his face
Edited: a black skin man with a tree head and branches on his face [and eyes closed]



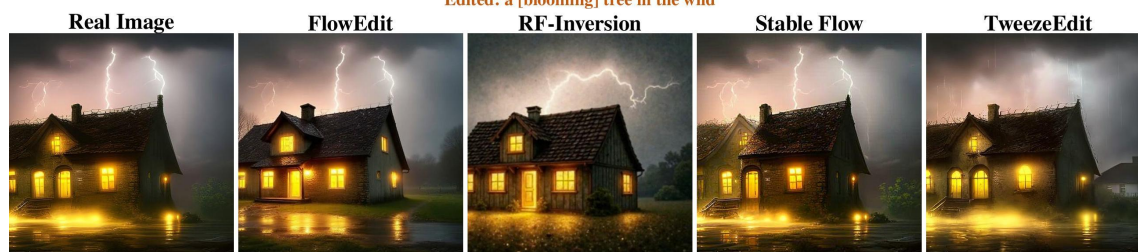
Original: a bottle of champagne and a hat on a wooden table
 Edited: a bottle of champagne and a hat [with flowers] on a wooden table



Original: a table with flowers on it
 Edited: a table with flowers [and hat] on it



Original: a [dry] tree in the wild
 Edited: a [blooming] tree in the wild



Original: a house with [lightning and] rain on it
 Edited: a house with rain on it



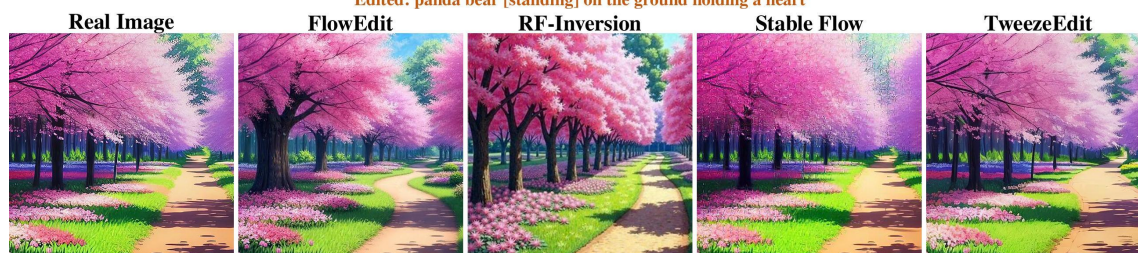
Original: a [white] dog lying on grass
 Edited: a [leopard] dog lying on grass



Original: a [long haired cat] looking up at something
 Edited: a [short haired cat] looking up at something



Original: panda bear [sitting] on the ground holding a heart
 Edited: panda bear [standing] on the ground holding a heart



Original: anime, the garden, pink, flowers, trees, path
 Edited: anime, the garden, pink, flowers, trees, [curving] path



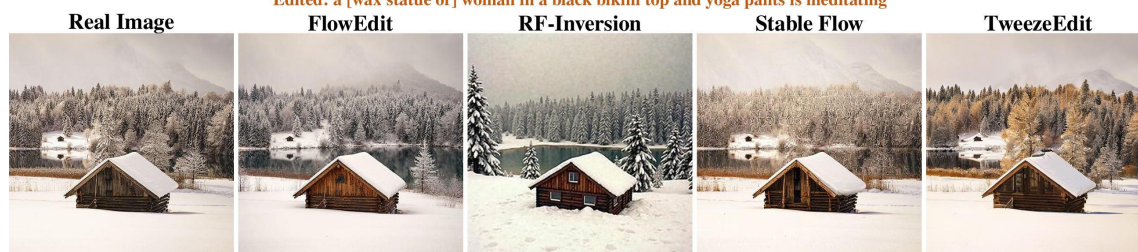
Original: a [white] kitten sitting on a leopard print blanket
 Edited: a [yellow] kitten sitting on a leopard print blanket



Original: a woman in a kimono standing in a river
 Edited: a [golden] woman [sculpture] in a kimono standing in a river



Original: a woman in a black bikini top and yoga pants is meditating
 Edited: a [wax statue of] woman in a black bikini top and yoga pants is meditating



Original: a [brwon] cabin in the snow near a lake
 Edited: a [golden] cabin in the snow near a lake