TweezeEdit: Consistent and Efficient Image Editing with Path Regularization

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

Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong SAR
makchen@ust.hk

Abstract

Recent progress in training-free image editing has enabled existing text-to-image diffusion models to be directly adapted into text-guided image editors without additional training. However, existing methods often over-align with target prompts while inadequately preserving source image semantics. These approaches generate target images explicitly or implicitly from the inversion noise of the source images, termed the inversion anchors. We identify this strategy as suboptimal for semantic preservation and inefficient due to elongated editing paths. We propose *TweezeEdit*, a tuning- and inversion-free framework for consistent and efficient image editing. Our method addresses these limitations by regularizing the entire denoising path rather than relying solely on the inversion anchors, ensuring source semantic retention and shortening editing paths. Guided by gradient-driven regularization, we efficiently inject target prompt semantics along a direct path using a consistency model. Extensive experiments demonstrate TweezeEdit's superior performance in semantic preservation and target alignment, outperforming existing methods. Remarkably, it requires only 12 steps (1.6 seconds per edit), underscoring its potential for real-time applications. The appendix is available in the extended version.

Code — https://github.com/hdsfade/TweezeEdit
Extended version — https://arxiv.org/abs/2508.10498

1 Introduction

Recent progress in training-free image editing (Xu et al. 2024; Rout et al. 2024; Kulikov et al. 2024; Hertz et al. 2022) has enabled existing text-to-image diffusion models (Rombach et al. 2022; Luo et al. 2023; Labs 2024) to be directly adapted into text-guided image editors without additional training. However, existing methods often fail to preserve the semantic content of the source image, over-aligning with the target prompt and necessitating extensive corrections through additional control (Shuai et al. 2024). Current approaches typically employ a deterministic reverse process to derive *inversion anchors* (the inverted noise of source images). Although inversion anchors theoretically encapsulate source image information, they often fail to generate expected similar denoising paths between source and target images in practice, resulting in over-alignment during editing (Mokady et al. 2023; Miyake et al. 2023).

Two primary issues contribute to this challenge: (1) numerical errors in inversion cause information loss, preventing the inversion anchors from fully reconstructing the source image, and (2) the lack of constraints on the diffusion model's output results in uncontrolled divergence in the denoising path of source and target images, triggering unexpected changes. Although recent efforts (Kulikov et al. 2024) mitigate inversion errors by interpolating direct paths between source and target images using sampled inversion anchors, they still suffer from estimation inaccuracies and inherent over-alignment from the inversion paradigm. Moreover, techniques (Cao et al. 2023; Hertz et al. 2022) that impose path constraints often require intrusive modifications, which rely on model-specific designs and increase computational demands. Tuning-based methods (Mokady et al. 2023; Zhang, Xiao, and Huang 2023) align output with the source image by tuning text embeddings or parameters, yet they require heavy computation cost and may overfit, harming general generative ability.

To address these limitations, we propose *TweezeEdit*, a tuning-free, inversion-free framework for efficient and semantically consistent image editing. Unlike methods that depend solely on inversion anchors, our approach regularizes the entire denoising path difference between the source and target images. This strategy constrains the output of the diffusion model, akin to tightening the arms of tweezers, restricting edits exclusively to prompt-relevant regions. Our method not only enhances source retention but also shortens the direct path. Additionally, we employ gradient-based regularization to guide updates along this path, eliminating the need for architectural modifications. Using consistency models as the backbone, TweezeEdit reduces sampling steps and cumulative errors for efficient editing, while naturally extending to noise and velocity prediction models through their connection to consistency models.

Through extensive experiments, we demonstrate that TweezeEdit outperforms state-of-the-art (SOTA) methods in prompt alignment and semantic preservation. As an architecture-agnostic solution, TweezeEdit seamlessly integrates with attention control for refinement. Quantita-

*These authors contributed equally.
†Corresponding author.

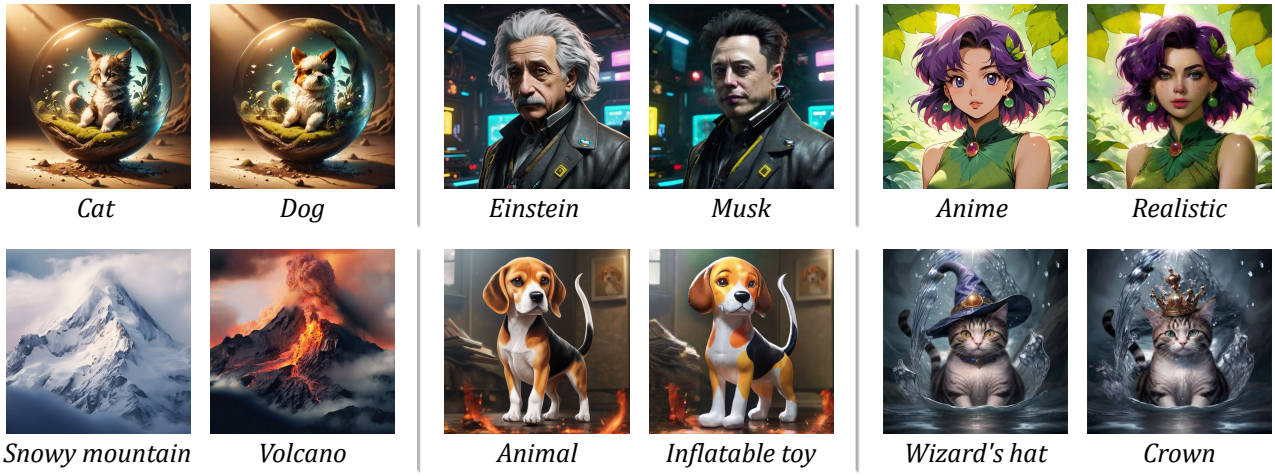


Figure 1: Examples of images edited with *TweezeEdit*, a tuning- and inversion-free framework for text-driven editing using pretrained consistency models. It effectively preserves source image semantics while aligning with target prompts.

tive and qualitative evaluations highlight its superiority in consistency-critical tasks and perceptual quality. Leveraging consistency models, TweezeEdit performs edits in about 12 steps, reducing latency while maintaining quality.

Our contributions are threefold:

1. We extend the inversion anchor paradigm by regularizing the entire denoising path, enhancing source semantic retention and shortening editing paths.
2. We propose *TweezeEdit*, a gradient-guided editing algorithm that avoids inversion and architectural changes, and is accelerated by consistency models.
3. We empirically validate TweezeEdit's effectiveness and efficiency in editing tasks.

2 Related Work

In diffusion-based image editing, given a source image and its description (source prompt), the central challenge is to modify the image according to a target prompt while preserving consistency with the source image.

Tuning-based methods (Ruiz et al. 2023; Dong et al. 2023; Zhang, Xiao, and Huang 2023; Mokady et al. 2023) enforce the models to reconstruct the source image given the source prompt by optimizing text embeddings or model parameters. While effective for consistency, these approaches are computationally intensive and may compromise the models' generative capability due to overfitting.

Tuning-free methods leverage pre-trained diffusion models for image editing without fine-tuning. These approaches rely on inversion anchors (i.e., inverted noise from source images) to preserve structural and semantic consistency, but they suffer from limited reconstruction fidelity due to inversion errors. For example, DDIM (Song, Meng, and Ermon 2020) suffers from error accumulation (Mokady et al. 2023; Miyake et al. 2023) during its reverse denoising process when estimating the initial noise. RF-Inversion (Rout et al. 2024) improves consistency by integrating conditional vector fields based on source images, but it remains limited by inversion inaccuracies. Moreover, direct modifications to these vector fields may degrade generation quality or introduce semantic distortions. FlowEdit (Kulikov et al. 2024) and Virtual Inversion (Xu et al. 2024) both employ sampling-based inversion anchors to avoid explicit deterministic inversion. However, because the source paths are constructed from sampled inversions, these methods still suffer from inaccurate anchors and the inherent stochasticity of sampling.

Attention-based methods, a subset of tuning-free methods, use diffusion models' attention mechanism to guide image editing. In U-Net-based models (Rombach et al. 2022), some approaches (Hertz et al. 2022; Cao et al. 2023) refine attention during generation by leveraging attention maps from the source image reconstruction. StableFlow (Avrahami et al. 2024) extends these to transformer-based models (Peebles and Xie 2023), improving source-target consistency. However, these methods require architectural changes and higher computational costs, reducing scalability and efficiency.

3 Preliminaries

3.1 Diffusion Models

Diffusion models (Ho, Jain, and Abbeel 2020) are generative models that learn to reverse a gradual noising process to produce images. In practice, diffusion models typically operate in a latent space (Rombach et al. 2022), obtained by encoding images through an encoder-decoder architecture. For notational simplicity, we formulate the diffusion process as acting directly on the image z_0.

During forward diffusion ($t = 0 \to +\infty$, discretized in practice as $t \in \{0, 1, \cdots, T\}$), the clean image z_0 is progressively corrupted into Gaussian noise. The noisy image z_t at timestep t follows:

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

where $\alpha_{1:T}$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ are schedule parameters. When $T \to \infty$, $\bar{\alpha}_T \to 0$ ensures $z_T \to \mathcal{N}(0, I)$.

Following DDIM (Song, Meng, and Ermon 2020), the reconstruction process reverses the forward diffusion via iterative denoising according to:

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

starting from $z_T \sim \mathcal{N}(0, I)$, where $\epsilon_\theta(z_t, t)$ is a trained noise-prediction network output. This discrete update corresponds to a numerical solver for a deterministic ODE (Song et al. 2020).

3.2 Consistent Models

Consistency models (Song et al. 2023; Luo et al. 2023), a class of diffusion models, enhance sampling efficiency by enforcing self-consistency across timesteps. These models learn a mapping $f(z_t, t)$ that directly predicts the clean image z_0 from a noisy input z_t. In practice, they employ multistep consistency sampling to iteratively refine z_0:

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

where $\epsilon \sim \mathcal{N}(0, I)$, and σ_t denotes the noise scale at timestep t. By allowing larger step sizes k, consistency models generate images with fewer sampling steps.

3.3 Tuning-Free Image Editing

Tuning-free image editing methods based on diffusion models involve a two-step process: inversion followed by denoising. During inversion, the source image z_0^{src} is inverted into noise (inversion anchors) either explicitly via DDIM inversion or implicitly via sampling (Kulikov et al. 2024) (i.e., $z_0^{src} \to z_T$). In the subsequent denoising phase, guided by a target prompt, z_T is denoised to generate the target image z_0^{tar}, aiming to preserve consistency with the source image while aligning the output with the target prompt.

Some methods (Kulikov et al. 2024) construct an update target $z_t^{mix} = z_0^{src} + z_t^{tar} - z_t^{src}$, integrating semantics from z_t^{src} into the update to enhance consistency, rather than relying solely on z_T. This mathematically corresponds to a direct path between z_0^{src} and z_0^{tar}. Other approaches (Tumanyan et al. 2023; Cao et al. 2023; Hertz et al. 2022) enforce consistency through attention control in diffusion models.

4 TweezeEdit

Current methods depend on inversion anchors to maintain consistency with source images. However, inaccurate estimation of these anchors often results in poor consistency preservation. To address the dual challenges of preserving source semantics while achieving alignment with target prompts, we propose *TweezeEdit*, a consistent and efficient image editing framework. Our approach builds on a key intuition illustrated in Figure 2:

❖ Desired edited image: The desired edited image within the target prompt's distribution should lie close to the source image, enabling both semantic consistency and prompt alignment.

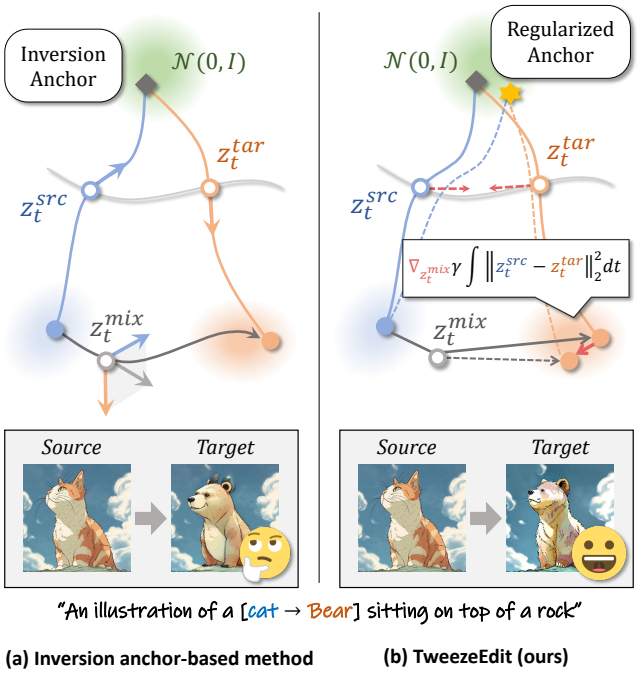


Figure 2: Comparison of inversion anchor-based method (a) and TweezeEdit (b). The inversion anchor-based method follows either the inversion-denoising path or direct path, frequently over-aligning with the target prompt due to inadequate retention of source semantics. TweezeEdit navigates the direct 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.

❖ Desired anchor: The source and desired edited image should originate from the same anchor and ensure the generated sample in the target domain remains close to the source. This provides a desired anchor for obtaining the desired edited image, rather than relying on DDIM-inverted noise (inversion anchor).

Based on this principle, TweezeEdit consists of two core components. First, we leverage a consistency model to progressively integrate target semantics along the direct path (Section 4.1). Its self-consistency, reduced sampling steps, and an additional calibration trick collectively mitigate cumulative errors and improve alignment. Second, we introduce denoising path regularization (Section 4.2), which enhances semantic preservation by constraining discrepancies between the source and target denoising trajectories throughout the entire denoising process. This effectively adjusts the underlying anchor (regularized anchor) toward one capable of jointly generating both the source image and desired edited image rather than depending solely on inversion anchors. The regularization shortens the direct path—akin to tweezers tightening their arms—allowing the editing pro-

cess to sharpen its focus while maintaining strong source fidelity.

4.1 Direct Path with Consistency Model

Direct path interpolation. To inject target prompt semantics into the source image, we construct an interpolation path using consistency models. The self-consistency of consistency models constrains $\|f_\theta(z_t, t) - f_\theta(z_{t-1}, t-1)\|$, providing implicit temporal regularization and mitigating the discontinuities that arise from timestep-independent sampling in the inversion-free methods. We first define the ideal direct path between the source and target images as:

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

where $\bar{\alpha}_t$ is the noise scheduler, and $z_t^{mix}(t = 1, \cdots, T)$ represents the interpolation that satisfies $z_T^{mix} = z_0^{src}$ and $z_0^{mix} = z_0^{tar}$.

Since z_0^{tar} is unavailable, we approximate z_t^{mix} by:

$$z_t^{mix} = z_0^{src} + \sqrt{\bar{\alpha}_t}(f(z_{t+1}^{tar}, t+1) - z_0^{src}). \tag{5}$$

When z_0^{src} is known, the prediction error of $f(z_t^{src}, t)$ can be computed accurately. Building on this, we refine the estimate of $f(z_t^{tar}, t)$ using the **calibration trick** from (Ju et al. 2023):

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

which calibrates $f(z_t^{tar}, t)$ by leveraging the similarity of source and target denoising paths. Substituting this into Eq 8 yields:

$$z_{t-1}^{mix} = z_0^{src} + \sqrt{\bar{\alpha}_{t-1}}(f(z_t^{tar}, t) - f(z_t^{src}, t)). \tag{7}$$

Consistency model-based editing. For Eq 4, we obtain an equivalent form:

$$\begin{aligned} z_t^{mix} &= z_0^{src} + (\sqrt{\bar{\alpha}_t} z_0^{tar} + \sqrt{1-\bar{\alpha}_t}\epsilon) \\ &\quad - (\sqrt{\bar{\alpha}_t} z_0^{src} + \sqrt{1-\bar{\alpha}_t}\epsilon) \\ &= z_0^{src} + z_t^{tar} - z_t^{src} \end{aligned} \tag{8}$$

The evolution of z_t^{src} and z_t^{tar} enables the consistency model to iteratively refine $f(z_t^{tar}, t)$, yielding progressively better approximations of z_0^{tar} for simulating the direct path.

A straightforward updating approach is to sample z_t^{src} from a given z_0^{src} and compute z_t^{tar} via $z_t^{tar} = z_t^{mix} - z_0^{src} + z_t^{src}$ using Eq 8. The model predicts $f(z_t^{src}, t)$ and $f(z_t^{tar}, t)$, and z_{t-1}^{mix} is updated via Eq 7, iterating until $z_0^{mix} = z_0^{tar}$.

However, this approach faces two challenges: (1) The sampled z_t^{src} may not match the desired anchor. We address this using on-the-fly noise rectification via denoising path regularization (detailed in Section 4.2). (2) Update errors in $f(z_t^{tar}, t)$ can degrade alignment with the target prompt. TweezeEdit addresses this issue by employing a consistency model as the denoising algorithm, reducing cumulative and discontinuous errors.

Using Eq 7, we progressively edit z_t^{mix} from z_0^{src} to z_0^{tar}. Compared with other models, the consistency model offers two key advantages: (1) Fewer sampling steps mitigate potential errors and improve editing efficiency. (2) Consistency models are inherently robust to noise (Song et al. 2023) and enable yielding more temporally coherent denoising trajectories. This property naturally aligns with our on-the-fly noise regularization framework.

Algorithm 1: 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 (Eq 7)
  $v_t = z_0^{src} + \sqrt{\bar{\alpha}_{t-1}}(\hat{z}_0^{tar} - \hat{z}_0^{src})$;
  // Gradient of path regularization (Eq 10)
  $\nabla_{z_t^{mix}} R_t = \hat{\gamma}_t \left[z_t^{src} - z_t^{tar} - \frac{\dot{\bar{\alpha}}_t}{4\sqrt{\bar{\alpha}_t}}(\hat{z}_0^{src} - \hat{z}_0^{tar}) \right]$;
  // Update step (Eq 11)
  $z_{t-1}^{mix} = v_t - \nabla_{z_t^{mix}} R_t$;
end
return Edited image $z_0^{tar} = z_0^{mix}$

4.2 Denoising Path Regularization

The Role of Inversion Anchors. Inversion anchors serve to maintain consistency between the edited image and the source image. However, their effectiveness in preserving source semantics relies on two assumptions that often fail in practice:

1. *Inversion anchors completely preserve information from the source image.* While theoretically valid, in practice, the estimation of the inversion anchors suffers from discretization and approximation errors, leading to information loss.
2. *Diffusion model updates ($t \rightarrow t-1$) modify only regions corresponding to prompt differences for semantic consistency between z_t^{src} and z_t^{tar}.* In reality, updates uncontrollably modify existing elements, e.g., introducing undesired changes in demeanor or body shape in Figure 2 (a).

In summary, inversion anchors expect semantic consistency in paired $(z_T^{src}, z_T^{tar}) \cdots (z_t^{src}, z_t^{tar}) \cdots (z_0^{src}, z_0^{tar})$, which is essentially the similarity between the denoising paths $z_T^{src} \cdots z_0^{src}$ and $z_T^{tar} \cdots z_0^{tar}$. Thus, rather than relying solely on inaccurate inversion anchors, regularizing the similarity across the entire denoising path is more effective.

Denoising path regularization. We incorporate the distance between continuous denoising paths as a regulariza-

Method	Structure Distance$_{10^3}$ ↓	Unedited Region Preservation PSNR ↑	LPIPS$_{10^3}$ ↓	MSE$_{10^4}$ ↓	SSIM$_{10^2}$ ↑	Editing Alignment Whole ↑	Edited ↑	Efficiency Inv-Free	Steps↓
DDIM (SD1.5)	79.54	17.36	220.13	243.13	70.52	27.08	23.90	✘	50
VI (LCM: SD1.5)	113.83	13.93	292.99	518.98	59.37	**27.68**	**24.37**	✔	15
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
StableFlow (Flux)	16.44	24.24	76.10	64.41	**89.43**	23.98	20.96	✘	50
DDIM (SD1.5) + P2P	69.99	17.87	208.90	219.56	71.63	25.28	22.57	✘	50
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: Quantitative results on PIE-Bench. Inv-Free indicates whether explicit inversion is avoided. Whole and Edited refer to CLIPScore for the full image and edited region. ↑: higher is better, ↓: lower is better. Bold: best results. Our approach TweezeEdit achieves competitive edits with fewer steps and high consistency in unedited areas.

tion term during the update of z_t^{mix} in Eq 7, defined as

$$R_t = \gamma_t \int_{t-1}^{t} \|z_\tau^{src} - z_\tau^{tar}\|_2^2 \, \mathrm{d}\tau, \tag{9}$$

where γ_t denotes the predefined regularization strength at step t.

To approximate the continuous-time integral, we apply the integral mean value theorem with a Taylor expansion and derive the gradient regularization term with respect to z_t^{mix} (please see Appendix A.2 for details):

$$\nabla_{z_t^{mix}} R_t \approx \hat{\gamma}_t \left[z_t^{src} - z_t^{tar} - \frac{\dot{\bar{\alpha}}_t}{4\sqrt{\bar{\alpha}_t}} (f(z_t^{src}, t) - f(z_t^{tar}, t)) \right], \tag{10}$$

where we set $\Delta_t \approx \frac{1}{2}$ and define $\hat{\gamma}_t := 2\gamma_t\big(-1 + \frac{\dot{\bar{\alpha}}_t}{4\bar{\alpha}_t}\big)$ for simplicity. Further details regarding the choice of $\hat{\gamma}_t$ can be found in Appendix A.3.

Finally, the update of z_t^{mix} in Eq 11 can be divided into two parts: the target-prompt editing direction and the source-image preservation direction. Our algorithm is summarized in Algorithm 1. We set the update interval to 1 only for clarity of exposition. Our method naturally scales to larger intervals. In practice, we select only 12-15 timesteps in $\{1, \cdots, T\}$, which significantly reduces computational steps while maintaining performance.

$$\hat{z}_{t-1}^{mix} = \underbrace{z_0^{src} + \sqrt{\bar{\alpha}_{t-1}}(f(z_t^{tar}, t) - f(z_t^{src}, t))}_{\text{target editing}} \underbrace{-\nabla_{z_t^{mix}} R_t}_{\text{source preserving}} . \tag{11}$$

By incorporating the regularization term R_t at each update, we enforce consistency across the full denoising path as $\sum_{t=1}^{T} R_t = \int_{t=0}^{T} \gamma_\tau \|z_\tau^{src} - z_\tau^{tar}\|_2^2 \mathrm{d}\tau$. This requires z_t^{tar} to retain source image semantics and introduces only target prompt-related changes to avoid gradient penalties. Unlike inversion anchors, our approach extends regularization across the entire denoising path and dynamically calibrates the regularized anchors, enhancing semantic consistency.

Besides, our approach does not explicitly compute anchors. Instead, gradient-driven updates guide z_t^{mix} along the direct path, which circumvents two limitations: (1) **Reduced editing path length.** Traditional implicit or explicit inversion-anchor methods often incur inconsistent denoising paths, which in turn lead to the longer direct path. By regularizing the denoising paths, we shorten the direct path. (2) **Architecture-agnostic updates.** Prior methods often compensate for semantic loss through architecture-specific interventions (e.g., attention control). Our gradient-driven regularization operates without model intrusion, ensuring compatibility across diverse architectures and reducing computational overhead.

Although our method is based on consistency models, which offer advantages such as self-consistency and fewer denoising steps, it can still be applied to other noise or velocity prediction models due to the inherent relationship between clean-image prediction and their outputs (details are provided in A.4). Meanwhile, continuous consistency models (Lu and Song 2024) allow these models to be distilled into consistency models.

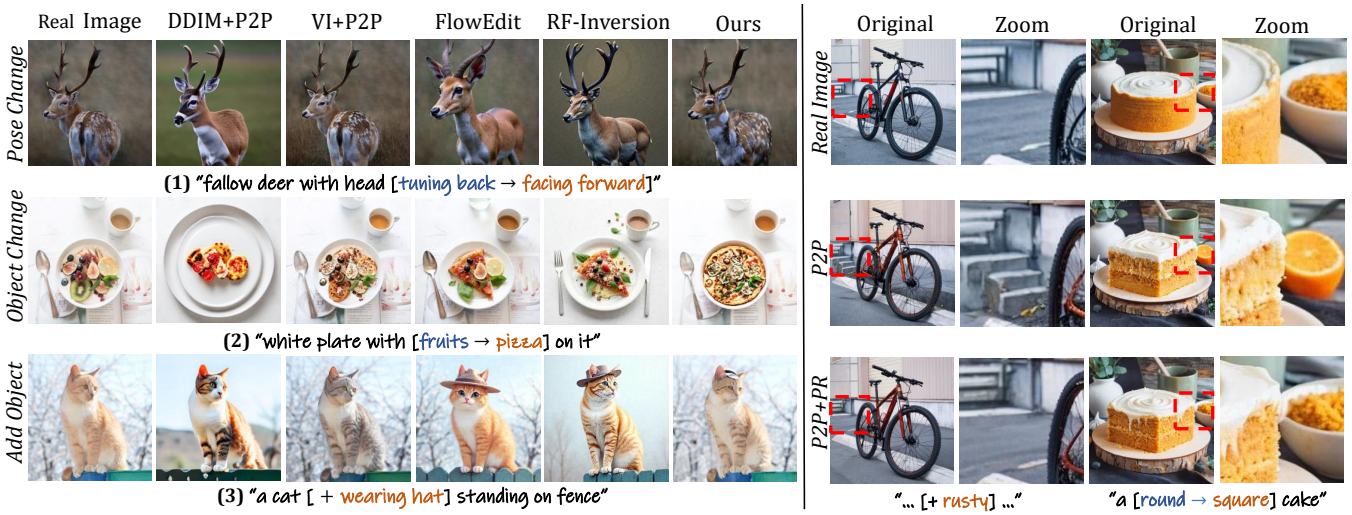
5 Experiments

5.1 Experimental Setup

Dataset. We evaluate our method on PIE-Bench (Ju et al. 2023), which comprises 700 images across 4 categories (animals, people, indoor scenes, outdoor scenes) and covers 10 types of editing tasks. Each instance provides a source prompt, a target prompt and a mask indicating the edited regions. The detailed experimental setup is described in Appendix B.1.

Evaluation Metrics. We assess editing fidelity through multiple measures: Structure Distance for global consistency (Tumanyan et al. 2023), MSE, PSNR, SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018) for evaluating the preservation of unedited regions, and CLIPScore (Hessel et al. 2021) for measuring alignment between target prompts and edited results.

Baselines. We compare TweezeEdit with representative tuning-free methods: DDIM (Song, Meng, and Ermon



(a) Qualitative comparisons on PIE-Bench

(b) Performance of PR with P2P

Figure 3: (a) Our approach achieves a superior balance between target prompt alignment and unedited region preservation. (b) The zoom column shows an enlarged view of the red dashed box region in the original column. PR and P2P work synergistically, with PR enhancing P2P's ability to preserve unedited regions.

2020), VI (Xu et al. 2024), FlowEdit (Kulikov et al. 2024) and RF-Inversion (Rout et al. 2024). DDIM is implemented using SD1.5 (Rombach et al. 2022), while VI adopts the latent consistency model (LCM) (Luo et al. 2023) variant of SD1.5. FlowEdit and RF-Inversion use Flux (Labs 2024). TweezeEdit leverages SD1.5, Flux and LCMs from SD1.5 and SDXL1.0 (Podell et al. 2023). Additionally, we evaluate P2P-enhanced (Hertz et al. 2022) versions of DDIM and VI, where P2P is a Unet-based (Rombach et al. 2022) attention control method. We also include StableFlow, a DiT-based (Peebles and Xie 2022) attention control method with Flux. Implementation details are provided in Appendix B.2.

5.2 Quantitative and Qualitative Analysis

We evaluate the effectiveness and efficiency of TweezeEdit in preserving consistency and producing high-quality edits.

Quantitative Results. Table 1 shows that TweezeEdit performs robustly across various paradigms, including noise- (SD1.5), velocity- (Flux), and clean-image predictors (LCM), achieving superior consistency preservation while maintaining editing performance. Metric comparisons are based on mean values, with p-values computed using the Wilcoxon signed-rank test. On SD1.5, TweezeEdit outperforms DDIM in LPIPS by -137.51 ($p < 0.01$). On Flux, TweezeEdit exceeds RF-inversion and FlowEdit across all consistency metrics while achieving higher Whole CLIP-Score. StableFlow sacrifices editability for consistency, significantly underperforming TweezeEdit in both Whole and Edited CLIPScore (-1.25, $p<0.01$ and -1.34, $p<0.01$, respectively). On LCM, even when VI is augmented with P2P for consistency enhancement, TweezeEdit (without P2P) surpasses it in both consistency and alignment (PSNR: $+2.8$, $p<0.01$; Edited CLIPScore: $+0.59$, $p<0.01$). Our method performs notably better on LCM with reduced inference steps (only 12 steps) and self-consistency, surpassing its performance on Flux. Compared to SD1.5, it achieves significant consistency improvement (MSE: -29.19, $p < 0.01$) with minimal editing cost (Edited CLIPScore: -0.15, n.s.). Integrating P2P further improves consistency with a minor trade-off in alignment. Beyond these models, TweezeEdit's architecture-agnostic design enables seamless adaptation to SDXL1.0 to achieve superior alignment.

Qualitative evaluation. As shown in Figure 3, our method achieves effective image edits while preserving source fidelity. For example: (1) it changes a deer's pose from backward to forward while preserving its identity (Figure 3 (a-1)); (2) it replaces fruit with pizza without altering the background (Figure 3 (a-2)); and (3) it adds a hat to a cat while maintaining its original pose and appearance (Figure 3 (a-3)). In contrast, DDIM struggles with error accumulation and over-alignment to the target prompt. VI+P2P preserves consistency but fails to apply the intended edits, such as missing the hat in Figure 3. FlowEdit and RF-Inversion introduce artifacts, either modifying backgrounds (Figure 3 (a-2, a-3)) or altering subjects' attributes (Figure 3 (a-1, a-3)). More visual comparisons are provided in Appendix D.1.

5.3 Path Regularization Analysis

Path regularization is pivotal to TweezeEdit, balancing consistency preservation with flexible editing. We analyze its impact through three dimensions: early-step regularization, integration with attention-based methods and balance between consistency and target alignment. Experiments in Appendix C.3 further confirm path regularization's robustness to random starts and slight gradient strength perturbations.

Early-step regularization. In diffusion models, early generation steps play a crucial role in shaping image structure. We thus restrict path regularization to the first m steps

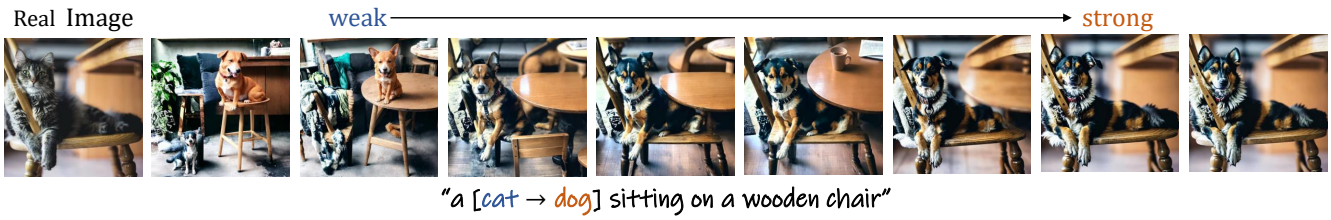


Figure 4: TweezeEdit performance across path regularization strengths. Left to right: increasing strength enhances source consistency, where an appropriate intensity strikes a balance between consistency and target alignment.

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

Table 2: Performance across different early steps with path regularization (total steps: 12; bold values denote best results). Increasing steps boost consistency but reduce goal alignment, with an optimal tradeoff at 6 steps.

of the 12-step process. As shown in Table 2, increasing m enhances consistency in unedited regions but gradually reduces editing capability. To strike an optimal balance, applying path regularization to half the steps (6 of 12) preserves source image fidelity while avoiding excessive editing constraints on later steps.

Synergy with attention-based methods. Path regularization is compatible with attention-control frameworks. Figure 3 (b) demonstrates that the combination of path regularization and P2P substantially improves output consistency while reducing artifacts in standalone P2P.

Balancing consistency and alignment. Path regularization empowers users to calibrate the trade-off between consistency and text alignment. As illustrated in Figure 4, increasing regularization strength shifts outputs from strict text alignment to structural preservation (e.g., retaining background when converting a cat to a dog). This tunability supports diverse customization needs without architectural modifications, making TweezeEdit adaptable to both consistency-focused and creativity-driven editing scenarios.

5.4 Perceptual Quality Assessments

We evaluate the visual quality of the edited images using IR (Xu et al. 2023), HPSv2 (Wu et al. 2023), PickScore (Kirstain et al. 2023), and AES (Schuhmann et al. 2022). Our comparison includes TweezeEdit, FlowEdit, RF-Inversion, and StableFlow. Our method, TweezeEdit, achieves the best performance across all metrics. For example, its IR score is 3.81 higher than that of the second-best method. Detailed results can be found in Appendix C.1.

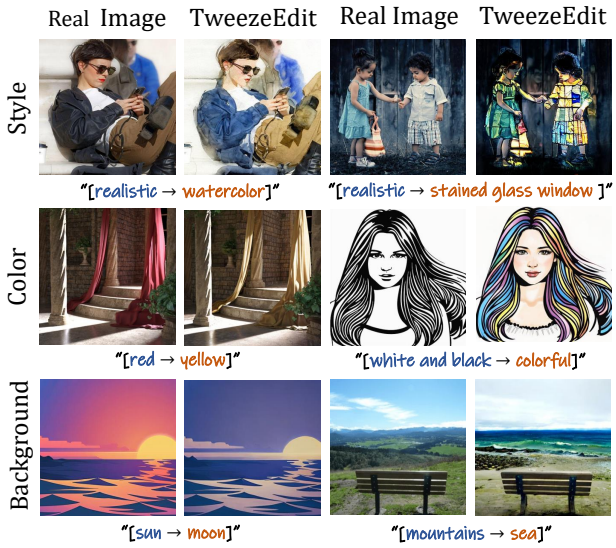


Figure 5: TweezeEdit 's performance in consistency-critical translation tasks.

5.5 Text-based Translation Editing

Text-based translation editing requires strong consistency preservation, especially when making substantial visual modifications. As demonstrated in Figure 5, TweezeEdit effectively maintains semantic consistency in these tasks.

6 Conclusion

In this work, we address the critical challenge of semantic fidelity loss in text-driven image editing with diffusion models, where existing methods over-align to target prompts and fail to preserve source content due to inversion inaccuracies and unconstrained denoising. Our framework, *TweezeEdit*, introduces an inversion-free paradigm that regularizes the denoising trajectory between source and target images, enabling efficient editing via consistency models. By constraining divergence to prompt-relevant regions and using gradient-guided updates, TweezeEdit achieves source-consistent, target-aligned edits without architectural modifications. Extensive evaluations demonstrate its superior effectiveness and efficiency across diverse tasks, highlighting the potential of path regularization to bridge the gap between creative intent and generative model limitations.

Acknowledgments

The authors are grateful to the anonymous reviewers for their thoughtful and constructive comments. This work was supported by Grant T32-615-24-R.

References

Avrahami, O.; Patashnik, O.; Fried, O.; Nemchinov, E.; Aberman, K.; Lischinski, D.; and Cohen-Or, D. 2024. Stable Flow: Vital Layers for Training-Free Image Editing. *arXiv preprint arXiv:2411.14430*.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22560–22570.

Dong, W.; Xue, S.; Duan, X.; and Han, S. 2023. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7430–7440.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hessel, J.; Holtzman, A.; Forbes, M.; Min, R. L. B.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ju, X.; Zeng, A.; Bian, Y.; Liu, S.; and Xu, Q. 2023. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36: 36652–36663.

Kulikov, V.; Kleiner, M.; Huberman-Spiegelglas, I.; and Michaeli, T. 2024. FlowEdit: Inversion-Free Text-Based Editing Using Pre-Trained Flow Models. *arXiv preprint arXiv:2412.08629*.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Lu, C.; and Song, Y. 2024. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*.

Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

Miyake, D.; Iohara, A.; Saito, Y.; and Tanaka, T. 2023. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6038–6047.

Peebles, W.; and Xie, S. 2022. Scalable Diffusion Models with Transformers. *arXiv preprint arXiv:2212.09748*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rout, L.; Chen, Y.; Ruiz, N.; Caramanis, C.; Shakkottai, S.; and Chu, W.-S. 2024. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

Shuai, X.; Ding, H.; Ma, X.; Tu, R.; Jiang, Y.-G.; and Tao, D. 2024. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *arXiv:2010.02502*.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency models. *arXiv preprint arXiv:2303.01469*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. 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*, 1921–1930.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935.

Xu, S.; Huang, Y.; Pan, J.; Ma, Z.; and Chai, J. 2024. Inversion-free image editing with language-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9452–9461.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, S.; Xiao, S.; and Huang, W. 2023. Forgedit: Text guided image editing via learning and forgetting. *arXiv preprint arXiv:2309.10556*.