FLOWFOCUS: FOCUSED TRAJECTORY REGULARIZATION FOR INVERSION-FREE IMAGE EDITING

Anonymous authors
Paper under double-blind review



Figure 1: Representative editing results produced by FlowFocus, where green and red text denote the original prompt tokens and the substituted target tokens, respectively. FlowFocus is capable of handling a wide range of image editing tasks (*e.g.,* object modification, addition and removal, attribute replacement, and style transformation.)

ABSTRACT

Inversion-free image editing offers an efficient alternative to inversion-based approaches, yet achieving substantial semantic changes while preserving source content unrelated to the desired edit remains challenging. Existing trajectory-regularized methods improve source consistency, but typically apply preservation constraints uniformly across the latent representation, overlooking that different spatial regions may require different degrees of modification. In this work, we propose **FlowFocus**, an inversion-free image editing framework with focused trajectory regularization. Our key insight is that the editing velocity can be decomposed into a counterfactual semantic response induced by changing the textual condition and a state-induced response arising from latent variation. Their relative strengths provide a model-intrinsic and time-varying signal that indicates where semantic modification is required. Based on this, FlowFocus constructs a continuous counterfactual edit field and incorporates it into a trajectory-level optimal-control objective, replacing global scalar regularization with spatially and temporally adaptive control. This formulation strengthens source preservation where semantic changes are unnecessary while retaining sufficient flexibility in regions relevant to the target edit, without relying on external masks, auxiliary perception models, or additional supervision. Experiments on the PIE benchmark demonstrate consistent improvements in source preservation and editing quality over state-of-the-art methods. The code will be released after publication.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Text-guided image editing aims to modify visual content according to a natural-language instruction while preserving information irrelevant to the requested change. Recent advances in generative modeling, particularly diffusion models (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022; Peebles & Xie, 2023) and rectified flow models (Lipman et al., 2023; Liu et al., 2023), have substantially improved the visual quality and semantic controllability of image synthesis, providing powerful generative priors. By exploiting these priors, image editing methods can perform diverse semantic modifications without training generative models from scratch, enabling applications in content creation, image retouching, and interactive media generation (Meng et al., 2022; Xie et al., 2023; Brooks et al., 2023; Dalva et al., 2025; Jiang et al., 2026; Li et al., 2026; Kulikov et al., 2026).

A major line of research performs editing through an *inversion-first* paradigm. These methods first recover a corresponding latent or noise representation and subsequently regenerate the image under the target condition. For instance, DDIB (Su et al., 2023) inversion provides a widely adopted deterministic inversion mechanism for diffusion models, while recent rectified flow approaches (Rout et al., 2025; Yang et al., 2025b) further improve inversion through trajectory optimization, numerical integration, or latent refinement. Although these methods achieve strong reconstruction and editing performance, the additional inversion stage introduces computational overhead and makes the subsequent editing process dependent on the accuracy of the recovered source trajectory.

To address this, recent *inversion-free* approaches provide an attractive alternative by directly transporting the source image toward the target distribution without explicitly recovering its initial noise representation. FlowEdit (Kulikov et al., 2025) establishes a representative framework in this direction by constructing a source-to-target editing velocity from coupled source and target conditional flows. Such a formulation considerably simplifies the editing pipeline and improves inference efficiency. However, directly constructing the source-to-target trajectory can lead to unstable evolution and insufficient preservation of source content, causing unintended modifications in regions unrelated to the editing instruction. FlowAlign (Kim et al., 2026) approaches the problem from a trajectory perspective and introduces optimal-control-based regularization to stabilize the source-to-target transformation. Specifically, it penalizes deviation from the original editing velocity while imposing terminal source consistency, thereby propagating preservation constraints throughout the editing trajectory. This formulation substantially improves the stability of inversion-free editing.

Nevertheless, the source-consistency regularization is typically applied uniformly across the latent representation, without distinguishing regions that require semantic modification from those that should remain unchanged. This spatially isotropic treatment introduces a preservation–editability conflict: stronger regularization suppresses unintended variations but may also hinder necessary semantic changes, whereas weaker regularization improves editability at the cost of source consistency. This issue is particularly pronounced for localized edits, where substantial deviation may be required only within a limited spatial region. These observations raise a natural question: *can the pretrained conditional flow itself reveal where source consistency should be enforced and where semantic deviation should be preserved?*

Our key observation is that the inversion-free editing velocity contains complementary information about both semantic intervention and latent-state variation. By decomposing it into a counterfactual semantic response, induced by changing the source condition to the target condition at a fixed latent state, and a state-induced response, caused by latent variation under the same source condition, their relative strengths provide a model-intrinsic and time-varying indication of where the requested edit should take effect. Based on this observation, we propose **FlowFocus**, an inversion-free image editing framework with focused trajectory regularization. Rather than relying on externally specified editing regions (Couairon et al., 2023; Lai et al., 2026) or predefined segmentation masks (Wang et al., 2023; Xu et al., 2026), FlowFocus constructs a continuous counterfactual edit field from the conditional flow responses and uses it to adapt the geometry of the trajectory-level optimal-control objective. This replaces global scalar regularization with spatially and temporally adaptive control, providing greater flexibility in edit-dominant regions while enforcing stronger source consistency elsewhere. Extensive experiments on the PIE benchmark (Ju et al., 2024) demonstrate that FlowFocus consistently improves the balance between source preservation and target-driven editing.

Our **main contributions** can be summarized as follows: 1) We propose FlowFocus, an inversion-free image editing framework that addresses the preservation–editability conflict by adapting source-

consistency regularization to the spatial and temporal characteristics of the requested edit. 2) We introduce a counterfactual decomposition of the editing velocity and derive a continuous edit field that enables focused trajectory regularization through a spatially adaptive optimal-control formulation, without relying on external masks or specified editing regions. 3) Extensive experiments on PIE demonstrate the effectiveness of FlowFocus, with consistent improvements in source preservation and editing quality over state-of-the-art methods.

2 Related Work

2.1 Inversion-based Editing

Inversion-based methods typically reconstruct a latent trajectory from the source image and subsequently perform target-conditioned generation along the reversed trajectory. Early diffusion-based approaches employ ordinary differential equation (ODE) inversion to deterministically map an input image into a noisy latent representation, which has been widely adopted by editing methods (Hertz et al., 2023; Mokady et al., 2023; Ju et al., 2024). However, numerical approximation during inversion can lead to reconstruction errors that accumulate along the trajectory and compromise source fidelity. More recently, inversion-based editing has been extended to rectified flow models. RF-Inv (Rout et al., 2025) formulates the inversion process from an optimal-control perspective to better balance reconstruction and editability, while RFEdit (Wang et al., 2025) and FireFlow (Deng et al., 2025) improve inversion efficiency and accuracy through higher-order numerical integration. Other approaches (Hu et al., 2025; Zhou et al., 2025; Guo et al., 2026; Yang & Ye, 2026) further inject source/prompt features or attention information during target-conditioned generation to retain image structure and appearance. Despite substantial improvements, inversion-based methods inherently require recovering and following a source-specific trajectory before editing, introducing additional computation and making editing performance dependent on inversion accuracy.

2.2 Inversion-Free Editing

Recent studies have explored inversion-free editing to avoid the explicit recovery. FlowEdit (Kulikov et al., 2025) establishes a source-to-target editing paradigm by coupling the source and target flows, enabling direct image transformation without an inversion stage. FIA (Yang et al., 2026) introduces a unified frequency-interactive attention mechanism (Yang et al., 2025a) that facilitates explicit feature-level interactions between the source and target branches. FlowAlign (Kim et al., 2026) formulates source preservation through an optimal-control perspective and regularizes the source-to-target trajectory to improve editing consistency. Despite their promising performance, existing inversion-free methods (Xu et al., 2024; Beaudouin et al., 2026; Tang et al., 2026) still face a fundamental tension between preserving the source image and introducing the semantic changes required by the target prompt. In particular, source-preservation constraints are typically imposed globally, without explicitly distinguishing regions relevant to the intended edit from those that should remain unchanged. This observation motivates a more focused regularization strategy that selectively balances source preservation and target-driven editing.

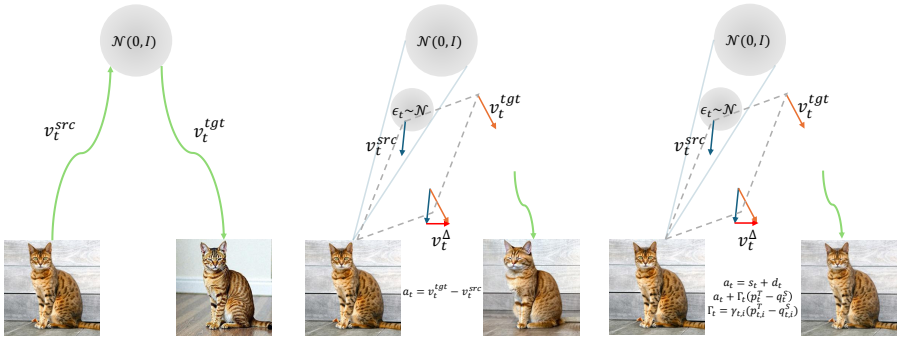
3 Method

3.1 Preliminary

Let X_0 denote a clean data sample and $X_1 \sim p_{noise} = \mathcal{N}(0, I)$ be a Gaussian sample. A continuous generative flow transports X_1 to X_0 through an ordinary differential equation (ODE)

$$dX_t = \boldsymbol{v}_t(X_t)dt, \qquad t \in [0, 1], \tag{1}$$

where $\boldsymbol{v}_t : \mathbb{R}^d \to \mathbb{R}^d$ denotes the time-dependent velocity field. In practice, $\boldsymbol{v}_t$ is parameterized by a neural network θ, yielding the sampling ODE. Once the velocity field is learned, generation amounts to sampling X_1 and integrating Eq. (1) from $t = 1$ toward $t = 0$, which is the basic flow-ODE formulation adopted by flow-matching generative models. Conditional flow matching (Lipman et al., 2023) circumvents this difficulty by defining a tractable probability path between paired endpoints and learning its conditional velocity field. For an interpolation $\boldsymbol{u}_t(X_t \mid X_0, X_1)$,



(a) Inversion-based Editing (b) Inversion-free Editing (c) Focused Trajectory Regularization

Figure 2: Overview of different image editing paradigms. (a) Inversion-based methods first recover a source-specific latent and then generate the target image. (b) Inversion-free methods directly construct a source-to-target editing trajectory. (c) FlowFocus decomposes the editing velocity into semantic and state-induced responses, derives a continuous counterfactual edit field, and uses it to spatially and temporally adapt trajectory regularization. Prompt from PIE (Ju et al., 2024): " a opened eyes cat sitting on wooden floor" → "a closed eyes cat ...".

the neural velocity field can then be trained with the following regression loss:

$$\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}_{t,X_0,X_1}\left[\left\|\boldsymbol{v}_t^{\theta}(X_t) - \boldsymbol{u}_t(X_t \mid X_0, X_1)\right\|_2^2\right], \tag{2}$$

where $\|\cdot\|_2^2$ represents mean square error. Among various flows, the linear interpolation is commonly used as rectified flow models (Liu et al., 2023)

$$X_t = (1-t)X_0 + tX_1, \tag{3}$$

Its conditional velocity for Eq. (2) has a particularly simple form:

$$\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}_{t,X_0,X_1}\left[\left\|\boldsymbol{v}_t^{\theta}(X_t) - (X_1 - X_0)\right\|_2^2\right], \tag{4}$$

The linear conditional path and its associated constant conditional velocity are also the ingredients underlying the subsequent image-editing construction. For text-conditioned generation, the velocity predictor is formulated as $\boldsymbol{v}_t^{\theta}(X_t, \boldsymbol{c})$, where $\boldsymbol{c}$ denotes the additionally received textual condition. While the pretrained model is originally designed to transport noise to images, the discrepancy between two text-conditioned velocity fields can be exploited to construct a transformation between two image semantics, without training a dedicated source-to-target flow.

3.2 Inversion-Free Editing and Trajectory Regularization

Let $\boldsymbol{x}_{src}$ be the source image with textual condition $\boldsymbol{c}_{src}$, and let $\boldsymbol{x}_{tgt}$ denote the desired edited image associated with $\boldsymbol{c}_{tgt}$. Ideally, the edit could be represented by a direct source-to-target conditional path

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_{tgt} + t\boldsymbol{x}_{src}, \qquad t \in [0,1], \tag{5}$$

with $\boldsymbol{x}_1 = \boldsymbol{x}_{src}$ and $\boldsymbol{x}_0 = \boldsymbol{x}_{tgt}$. To reuse a pretrained model, two auxiliary flows can be constructed using a shared noise variable $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$:

$$\boldsymbol{q}_t = (1-t)\boldsymbol{x}_{src} + t\boldsymbol{\epsilon}, \qquad \boldsymbol{p}_t = (1-t)\boldsymbol{x}_{tgt} + t\boldsymbol{\epsilon}, \tag{6}$$

Here $\boldsymbol{q}_t$ belongs to a source-conditioned noise-to-image path, whereas $\boldsymbol{p}_t$ analogously corresponds to the target-conditioned path. Because the two paths share the same noise endpoint, equations (Eq. (5) and Eq. (6)) satisfy

$$\boldsymbol{p}_t = \boldsymbol{q}_t + \boldsymbol{x}_t - \boldsymbol{x}_{src}, \tag{7}$$

Consequently, a source-to-target editing dynamic can be approximated by

$$\boldsymbol{a}_t = \frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{v}_t^{\theta}(\boldsymbol{p}_t, \boldsymbol{c}_{tgt}) - \boldsymbol{v}_t^{\theta}(\boldsymbol{q}_t, \boldsymbol{c}_{src}), \tag{8}$$

It captures the instantaneous direction predicted by the pretrained model for transforming the current source-associated state toward the target semantics, avoiding explicit inversion.

To stabilize this trajectory, source consistency can be introduced through a time-reversed optimal-control objective,

$$\min_{\boldsymbol{k}} \quad \int_0^1 \frac{1}{2}\|\boldsymbol{k}_t - \boldsymbol{a}_t\|_2^2 dt + \frac{\eta}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_{src}\|_2^2, \quad \dot{\boldsymbol{x}}_t = \boldsymbol{k}_t, \qquad \boldsymbol{x}_1 = \boldsymbol{x}_{src}, \tag{9}$$

where η controls source consistency. Under a first-order approximation, the regularized velocity can be expressed as

$$\boldsymbol{v}_t^r \simeq \boldsymbol{a}_t + \gamma_t(\hat{\boldsymbol{p}}_0^T - \hat{\boldsymbol{q}}_0^S), \qquad \gamma_t = \frac{\eta}{\eta t - 1} \tag{10}$$

where $\hat{\boldsymbol{p}}_0^T = \boldsymbol{p}_t - t\,\boldsymbol{v}_t^\theta(\boldsymbol{p}_t, \boldsymbol{c}_{tgt})$ and $\hat{\boldsymbol{q}}_0^S = \boldsymbol{q}_t - t\,\boldsymbol{v}_t^\theta(\boldsymbol{q}_t, \boldsymbol{c}_{src})$ are clean-state estimates. While Eq. (10) improves trajectory consistency, its regularization is spatially uniform: the scalar γ_t applies the same form of source-consistency correction to every latent location. Consequently, regions that must change to satisfy the target prompt are regularized in the same manner as regions that should remain unchanged. Our FlowFocus addresses this limitation by allowing the pretrained conditional flow itself to determine where trajectory deviation should be preserved or suppressed.

3.3 Focused Trajectory Regularization

Counterfactual decomposition. Recalling the inversion-free editing velocity $\boldsymbol{a}_t$ in Eq. (8), we obtain the exact decomposition by inserting the source condition at the same target-side state $\boldsymbol{p}_t$:

$$\boldsymbol{a}_t = \underbrace{\boldsymbol{v}_t^\theta(\boldsymbol{p}_t, \boldsymbol{c}_{tgt}) - \boldsymbol{v}_t^\theta(\boldsymbol{p}_t, \boldsymbol{c}_{src})}_{s_t} + \underbrace{\boldsymbol{v}_t^\theta(\boldsymbol{p}_t, \boldsymbol{c}_{src}) - \boldsymbol{v}_t^\theta(\boldsymbol{q}_t, \boldsymbol{c}_{src})}_{d_t}, \tag{11}$$

The first term s_t changes only the textual condition while fixing the latent state, and therefore measures how the pretrained flow responds to the counterfactual intervention: $\boldsymbol{c}_{src} \to \boldsymbol{c}_{tgt}$. We refer to it as the *counterfactual semantic response*. In contrast, d_t keeps the source condition fixed and captures variation induced by the latent-state difference between $\boldsymbol{p}_t$ and $\boldsymbol{q}_t$. We refer to it as the *state-induced response*. Importantly, we do not assume that d_t consists solely of undesirable error; legitimate contextual adaptation may also contribute to it. Instead, the relative magnitude of s_t and d_t provides a model-intrinsic measure of how strongly a local change is explained by the requested semantic intervention. FlowAlign (Kim et al., 2026) regularizes their sum with a single scalar coefficient. Differently, our goal is instead to let the relative contribution of these factors determine where source consistency should be enforced strongly and where semantic deviation should be preserved.

Counterfactual edit field. Let the latent representation contain N spatial tokens, with $s_{t,i}, d_{t,i} \in \mathbb{R}^C$ denoting their C-dimensional velocity vectors. We define a continuous edit score for each spatial location as

$$\rho_{t,i} = \frac{\|s_{t,i}\|_2^2}{\|s_{t,i}\|_2^2 + \|d_{t,i}\|_2^2 + \lambda}, \qquad \rho_{t,i} \in [0, 1], \tag{12}$$

where $\lambda > 0$ is a small numerical constant. A large $\rho_{t,i}$ indicates that the local variation is dominated by the target-conditioned semantic response, whereas a small value indicates that semantic intervention is weak relative to state-induced variation. Because the velocity fields evolve throughout generation, ρ naturally varies over both space and time. Unlike a binary mask, it requires neither thresholding nor an externally specified editing region, and can represent both localized and spatially distributed edits. Regions with high, intermediate, and low values naturally correspond to edit-dominant, transition, and preservation-dominant regimes, respectively. For spatially localized edits, the semantic response is expected to concentrate on a subset of latent locations. For spatially distributed edits, such as global appearance transformations, ρ_t can instead become broadly distributed. This allows local and global editing to be represented within the same formulation without edit-type-specific routing.

Focused control geometry. A direct use of ρ_t as a multiplicative mask on $\boldsymbol{a}_t$ would alter the editing velocity heuristically and break the optimal-control interpretation established in Sec. 3.2. Instead, we use ρ_t to modify the geometry in which trajectory deviation is penalized. Specifically, we define

$$r_{t,i} = \delta + (1 - \delta)\rho_{t,i}, \qquad 0 < \delta \leq 1, \tag{13}$$

and replace Eq. (9) with

$$\min_{\boldsymbol{k}} \quad \int_0^1 \frac{1}{2}\sum_{i=1}^N r_{t,i}\|\boldsymbol{k}_{t,i} - \boldsymbol{a}_{t,i}\|_2^2 dt + \frac{\eta}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_{src}\|_2^2, \quad \dot{\boldsymbol{x}}_t = \boldsymbol{k}_t, \qquad \boldsymbol{x}_1 = \boldsymbol{x}_{src}, \tag{14}$$

Here $r_{t,i}$ determines how strongly the controlled velocity should remain close to the original editing velocity. At edit-dominant locations, $\rho_{t,i}$ is large and $r_{t,i}$ approaches 1, preserving the target-driven editing direction. At preservation-dominant locations, a smaller $r_{t,i}$ allows the terminal source-consistency objective to exert greater influence. Thus, the counterfactual field does not overwrite the editing dynamics; instead, it determines where the trajectory should remain faithful to the editing velocity and where stronger source correction is permitted.

Proposition 1. *Under the linear conditional flow and a first-order frozen-field approximation, the locally optimal editing dynamics satisfy*

$$\frac{d\boldsymbol{x}_t}{dt} \simeq \boldsymbol{a}_t + \Gamma_t(\hat{\boldsymbol{p}}_0^T - \hat{\boldsymbol{q}}_0^S), \tag{15}$$

where Γ_t is a spatially adaptive linear operator that independently rescales the source-consistency correction at each latent location. Specifically, for the i-th spatial token,

$$[\Gamma_t(\hat{\boldsymbol{p}}_0^T - \hat{\boldsymbol{q}}_0^S)]_i = \gamma_{t,i}(\hat{\boldsymbol{p}}_{0,i}^T - \hat{\boldsymbol{q}}_{0,i}^S), \qquad \gamma_{t,i} = \frac{\eta}{\eta t - r_{t,i}}, \tag{16}$$

for $\eta t \neq r_{t,i}$. Here, $\gamma_{t,i}$ is shared across the feature channels at location i.

Proposition 1 replaces the global scalar correction with a spatially and temporally adaptive operator determined by the counterfactual edit field. When $\delta = 1$, we have $r_{t,i} = 1$ for all i, and Eq. (16) reduces to $\gamma_t = \eta/(\eta t - 1)$, recovering the global trajectory regularization in Eq. (10). Hence, the global trajectory regularization (*e.g.,* FlowAlign) is a special case of our formulation. FlowFocus therefore generalizes global trajectory control while preserving its optimal-control interpretation. The proof and more details are deferred to Appendix A.1.

4 Experiments

4.1 Experimental Setup

Benchmark. The experiments were conducted on PIE (Ju et al., 2024), a comprehensive benchmark for text-guided image editing. The benchmark consists of 700 samples, each comprising a source image, a source prompt, and a target prompt. It covers a broad range of image editing tasks, such as object addition and removal, object replacement, color and material modification, style transfer, background change, action editing, texture modification, and spatial repositioning.

Evaluation metrics. Following prior works, we adopted six commonly used metrics to evaluate editing performance, background preservation and text-image alignment: Structure Distance (Ju et al., 2024) measures the structural consistency between the edited and source images, while PSNR, LPIPS (Zhang et al., 2018), MSE, and SSIM (Wang et al., 2004) assess content preservation in the unedited regions. CLIP (Radford et al., 2021) similarity scores evaluate the semantic consistency between target prompts and edited results across both the entire image and the edited region.

Baselines. We compared our method with a set of representative image editing approaches, covering both diffusion-based and rectified flow-based methods. The diffusion-based baselines include P2P (Hertz et al., 2023), PnP (Tumanyan et al., 2023), Null-text Inv (Mokady et al., 2023), and PnP-Inv (Ju et al., 2024). For rectified flow-based editing, we consider RF-Inv (Rout et al., 2025), FireFlow (Deng et al., 2025), DCEdit (Hu et al., 2025), FlowEdit (Kulikov et al., 2025), iRFDS (Yang et al., 2025b), SplitFlow (Yoon et al., 2025), DNAEdit (Xie et al., 2025), FIA (Yang et al., 2026), FlowAlign (Kim et al., 2026), and DRFS (Beaudouin et al., 2026). Unless otherwise specified, all baselines were evaluated using their publicly available implementations and configurations.

Implementation details. Our method was implemented on Stable Diffusion (Esser et al., 2024) series (SD3 and SD3.5), using 50 sampling steps for image generation. We set classifier-free guidance (CFG) scales to 5.5 and 13.5 for the source and target prompts, respectively. All experiments were conducted on a single NVIDIA L40S GPU. More details are provided in Appendix A.2.

4.2 Comparisons with State-Of-The-Arts

Qualitative evaluation. As shown in Fig. 3, FlowFocus produced high-quality editing results across diverse editing tasks, demonstrating strong adaptation to different target instructions. In particular,

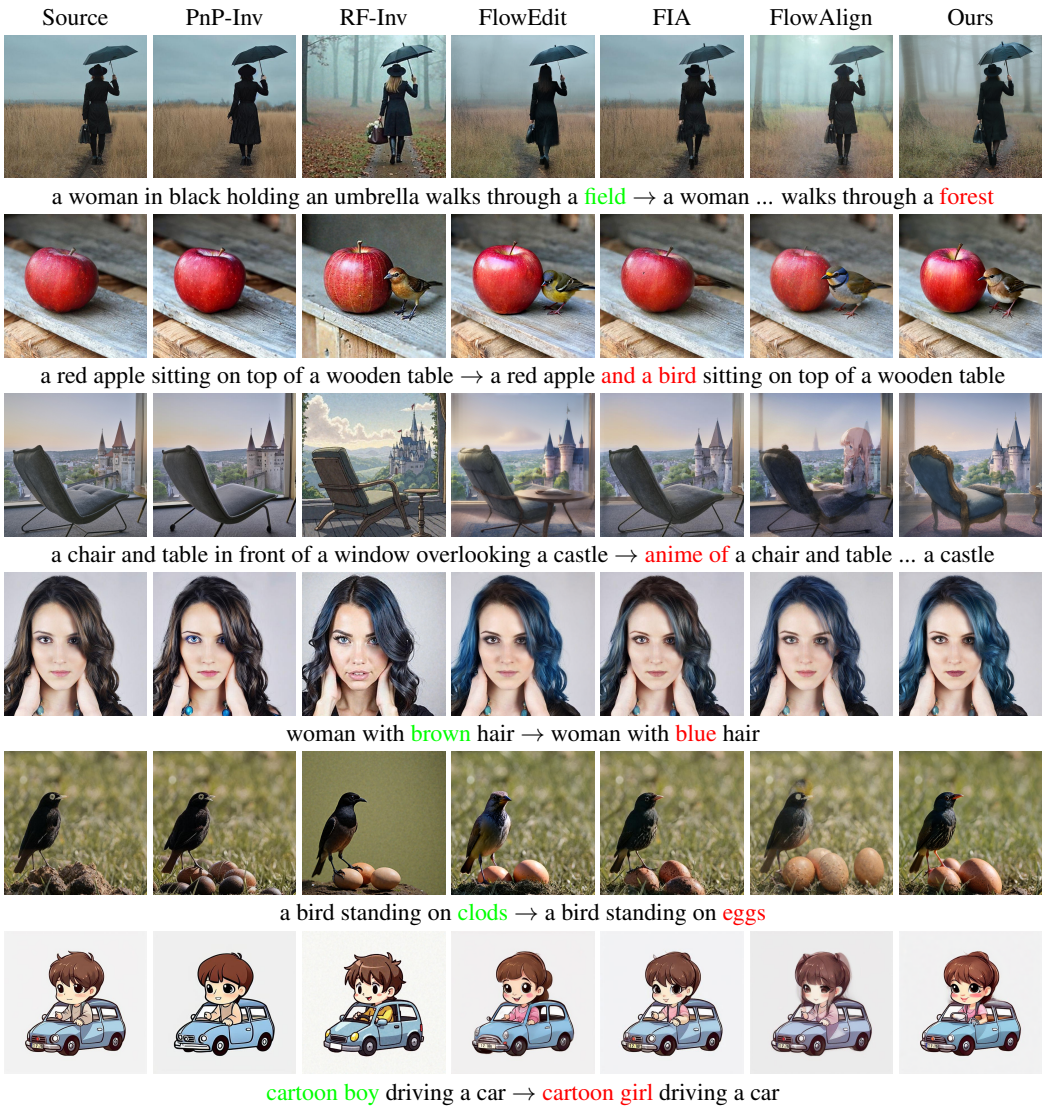


Figure 3: Qualitative comparison of existing text-guided image editing methods across six representative editing types: "change background", "add object", "change style", "change color", "change content" and "change object". More qualitative examples are provided in Appendix A.4.

our method achieved more faithful semantic changes while effectively preserving content unrelated to the requested edit. For example, in the first row of Fig. 3, FlowFocus successfully transformed the surrounding scene into a forest environment with clear trees and green grass, whereas FIA and FlowAlign exhibited insufficient changes toward the target description. Meanwhile, FlowFocus preserved the woman's appearance and accessories, including the shape of the handbag, hair color, and hat, which were unnecessarily altered by methods such as RF-Inv and FlowEdit. Similarly, in the last example, although several competing methods successfully changed the boy into a girl, they introduced unintended modifications to the car color or license plate, while FlowFocus performed the desired semantic edit with substantially better preservation of the remaining scene. These comparisons demonstrate that FlowFocus effectively concentrates modifications on content relevant to the target prompt while suppressing unnecessary changes elsewhere.

Quantitative evaluation. Table 1 presents the quantitative comparison on the PIE benchmark across diffusion- and rectified flow-based editing methods. Overall, FlowFocus achieved a favorable balance between source preservation and target-driven editing across both backbones. In particular, FlowFocus-SD3.5 achieved the best Structure Distance, PSNR, LPIPS, and MSE, while obtaining a competitive SSIM. Compared with recent SD3.5-based methods (*e.g.,* FIA and DRFS), these results demonstrate that our focused regularization effectively suppresses unintended modifications in

Table 1: Comparison with existing methods on the PIE (Ju et al., 2024) benchmark for each group (diffusion-based and rectified flow-based), reporting editing performance, background preservation, and text-image alignment. The best results are **bold** while the second best are underlined.

Method	Structure	Background Preservation				CLIP Similarity	
	Distance$_{\times 10^3}$↓	PSNR↑	LPIPS$_{\times 10^3}$↓	MSE$_{\times 10^4}$↓	SSIM$_{\times 10^2}$↑	Whole↑	Edited↑
P2P (Hertz et al., 2023)	68.11	17.59	211.70	219.72	70.41	24.29	21.96
PnP (Tumanyan et al., 2023)	29.91	21.87	115.96	88.46	79.40	25.24	22.06
Null-text-Inv (Mokady et al., 2023)	15.77	26.03	62.33	36.73	84.21	24.80	21.88
PnP-Inv (Ju et al., 2024)	24.42	22.62	109.88	81.66	79.69	25.02	22.70
RF-Inv-FLUX (Rout et al., 2025)	41.63	20.48	182.08	132.74	70.72	24.26	22.88
FireFlow-FLUX (Deng et al., 2025)	28.72	23.77	124.50	70.98	82.29	25.78	22.74
DCEdit-FLUX (Hu et al., 2025)	22.36	25.41	94.17	48.09	85.60	25.47	22.71
FlowEdit-FLUX (Kulikov et al., 2025)	27.66	21.58	112.62	94.49	82.98	25.41	22.62
iRFDS-SD3 (Yang et al., 2025b)	63.37	19.62	188.64	182.94	74.79	25.17	21.67
FlowEdit-SD3 (Kulikov et al., 2025)	27.16	22.37	104.57	86.19	88.50	26.30	23.21
SplitFlow-SD3 (Yoon et al., 2025)	25.18	22.81	103.16	81.84	83.36	26.37	23.35
FlowAlign-SD3 (Kim et al., 2026)	26.74	25.39	73.25	54.03	87.82	25.28	22.24
DNAEdit-SD3.5 (Xie et al., 2025)	14.19	26.66	74.57	32.76	88.63	25.63	22.71
FIA-SD3.5 (Yang et al., 2026)	13.19	27.10	57.88	29.56	88.75	25.77	22.89
FlowAlign-SD3.5 (Kim et al., 2026)	15.14	26.54	55.49	30.71	88.62	25.40	22.63
DRFS-SD3.5 (Beaudouin et al., 2026)	12.30	26.99	56.34	30.31	**89.20**	26.22	23.11
FlowFocus-SD3 (Ours)	22.79	26.12	88.66	52.61	88.12	**27.13**	**23.88**
FlowFocus-SD3.5 (Ours)	**11.71**	**27.55**	**55.45**	**28.80**	89.17	26.85	23.55

Table 2: Ablation study on the key components of FlowFocus using SD3.5.

Method	Structure	Background Preservation				CLIP Similarity	
	Distance$_{\times 10^3}$↓	PSNR↑	LPIPS$_{\times 10^3}$↓	MSE$_{\times 10^4}$↓	SSIM$_{\times 10^2}$↑	Whole↑	Edited↑
Baseline (Kulikov et al., 2025)	23.05	24.31	89.78	48.49	86.18	26.46	23.42
+ Global regularization	15.14	26.54	55.49	30.71	88.62	25.40	22.63
+ Focused regularization (fixed)	12.91	27.50	**54.70**	29.83	88.15	26.08	23.05
+ Focused regularization (adaptive)	**11.71**	**27.55**	55.45	**28.80**	**89.17**	**26.85**	**23.55**

regions irrelevant to the desired edit. Importantly, the improved source preservation does not come at the expense of editability. FlowFocus-SD3 achieved the highest Whole and Edited CLIP scores of 27.13 and 23.88, respectively, outperforming existing SD3-based approaches (*e.g.,* FlowEdit and FlowAlign) by a clear margin. When equipped with the stronger SD3.5 backbone, FlowFocus maintained competitive text-image alignment while substantially improving background fidelity. These results demonstrate that our focused regularization effectively preserves regions irrelevant to the desired edit while maintaining sufficient flexibility to achieve the target semantic changes. More quantitative results across different editing categories are provided in Appendix A.3.

4.3 Human Evaluation

Since automatic metrics cannot fully capture perceptual editing quality, we additionally conducted a human preference study following the protocol of FlowAlign. Differently, we separately evaluated the two key aspects considered in our method, namely *editing accuracy* and *source preservation*. We directly used the random subset (140 samples) from PIE and compared FlowFocus against four representative baselines: PnP-Inv, RF-Inv, FlowEdit, and FlowAlign. For each sample, one baseline was randomly assigned, with the comparisons balanced across the four methods. Twenty participants took part in the evaluation, and each comparison was independently assessed by at least three participants. Given the source image and editing instruction, participants separately judged 1) *which result better satisfies the requested semantic edit* and 2) *which result better preserves source content unrelated to the edit*. For each criterion, participants chose between the two results. The human preference results are summarized in Fig. 4 and demonstrate that FlowFocus achieves more faithful target edits while better preserving source content, further validating the effectiveness of focused trajectory regularization. More details are provided in Appendix A.5.

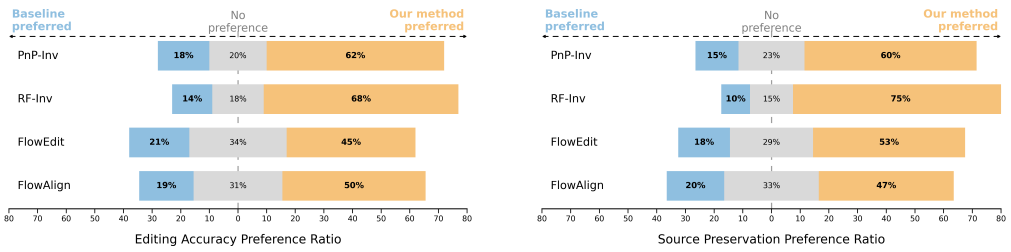


Figure 4: Human preference results for editing accuracy and source preservation. Each bar reports the proportion of responses favoring the baseline (blue), indicating a tie (gray), or favoring FlowFocus (orange). Bars are centered at zero to emphasize directional preference.

4.4 Ablation Study

Component analysis. Table 2 studies the contribution of each component under the SD3.5 backbone. Starting from the inversion-free baseline, introducing global trajectory regularization substantially improved source preservation (*e.g.,* reducing Distance from 23.05 to 15.14). However, this improvement was accompanied by a noticeable decrease in text-image alignment (*e.g,* Edited CLIP score dropping from 23.42 to 22.63), reflecting the preservation–editability conflict induced by uniform regularization. In contrast, focused control effectively mitigated this trade-off. The fixed variant, which uses the counterfactual edit field ρ computed at the middle timestep throughout the trajectory, further improved structural and background preservation while recovering semantic alignment. More importantly, the adaptive formulation, where ρ evolves along the editing trajectory, achieved the strongest overall performance. These results demonstrate the benefit of focused regularization and further confirm the importance of adapting the focus over time.

Effect of δ and ω. Fig. 5 analyzes the influence of the target CFG scale (ω) and the focused-regularization parameter (δ). As expected, CFG controls the trade-off between semantic alignment and source consistency: increasing ω generally improved CLIP similarity by strengthening the target guidance, but also led to larger Structural Distance due to increased deviation from the source image. Considering this trade-off, we set $\omega = 13.5$ as the default value. Beyond adjusting the global editing strength through CFG, FlowFocus adaptively modulates source-consistency regularization across spatial locations using the counterfactual edit field. The parameter δ controls the degree of this spatial adaptation through Eq. (13), and we therefore examined its sensitivity. Increasing δ from 0.5 to 0.75 consistently improved CLIP similarity while introducing only a minor increase in Structural Distance. However, further increasing it to $\delta = 1$ removed the spatial adaptation, reducing FlowFocus to global trajectory regularization and degrading the overall preservation–editability balance. We therefore used $\delta = 0.75$ in all experiments. Additional hyperparameter analyses and quantitative results are provided in Appendix A.3.

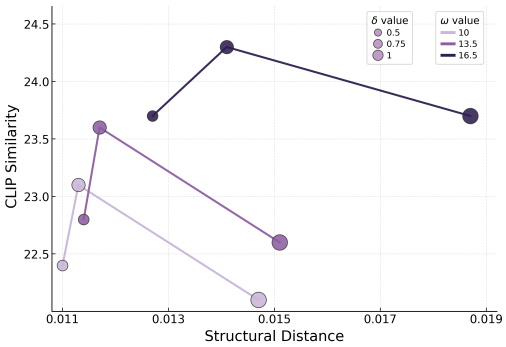


Figure 5: Sensitivity analysis of the CFG scale ω and focused-regularization parameter δ on the preservation–editability trade-off.

5 Conclusion

In this paper, we present **FlowFocus**, an inversion-free image editing framework that addresses the preservation–editability conflict in trajectory-regularized flow editing. We decompose the editing velocity into counterfactual semantic and state-induced responses, and use their relative strengths to construct a time-varying edit field. This field further enables focused trajectory regularization by adapting the optimal-control geometry across spatial locations, without relying on external hard masks. Experiments on the PIE benchmark demonstrate consistent improvements in both source preservation and target-driven editing over state-of-the-art image editing methods.

AI USE STATEMENT

In this work, Large Language Models (LLMs) were used to aid in the polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. In this study, no animal experimentation was involved. We provided a human study to assess the predicted results for editing performance evaluation without any specific identifiable information. All datasets used (the PIE benchmark) were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. All code will be made publicly available to facilitate replication and verification after the paper gets accepted. The experimental setup, including model configurations, resources, and hardware details, is described in detail in the paper. We also provide a complete proof of the core claims in the appendix. Additionally, the datasets used in this paper are publicly available, ensuring consistent and reproducible evaluation results. We believe these measures will enable other researchers to reproduce our work and further advance the field.

REFERENCES

Gaspard Beaudouin, Minghan Li, Jaeyeon Kim, Sung-Hoon Yoon, and Mengyu Wang. Delta rectified flow sampling for text-to-image editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18662–18672, 2026.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023.

Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. Fluxspace: Disentangled semantic editing in rectified flow models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13083–13092, 2025.

Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. Fireflow: Fast inversion of rectified flow for image semantic editing. In *International Conference on Machine Learning*, 2025.

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 *International Conference on Machine Learning*, 2024.

Zehui Guo, Zhen Wang, Junwei Shu, Yang Li, Changbo Wang, and Long Chen. h-flow: Flexible flow-based image editing via doob'sh-transform. In *European Conference on Computer Vision*, pp. 542–558, 2026.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

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

Yihan Hu, Jianing Peng, Yiheng Lin, Ting Liu, Xiaochao Qu, Luoqi Liu, Yao Zhao, and Yunchao Wei. Dcedit: Dual-level controlled image editing via precisely localized semantics. *arXiv preprint arXiv:2503.16795*, 2025.

Yilei Jiang, Zhen Wang, Yanghao Wang, Jun Yu, Yueting Zhuang, Jun Xiao, and Long Chen. Flowdc: flow-based decoupling-decay for complex image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25757–25766, 2026.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *International Conference on Learning Representations*, pp. 23395–23422, 2024.

Jeongsol Kim, Yeobin Hong, Jonghyun Park, and Jong Chul Ye. Flowalign: Trajectory-regularized, inversion-free flow-based image editing. In *International Conference on Learning Representations*, pp. 121814–121846, 2026.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025.

Vladimir Kulikov, Roni Paiss, Andrey Voynov, Inbar Mosseri, Tali Dekel, and Tomer Michaeli. Versatile editing of video content, actions, and dynamics without training. In *European Conference on Computer Vision*, pp. 448–466, 2026.

Yongwen Lai, Chaoqun Wang, and Shaobo Min. Fusionedit: Semantic fusion and attention modulation for training-free image editing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 11907–11911, 2026.

Kun Li, Michael Ying Yang, and Sami Sebastian Brandt. Query-guided spatial–temporal–frequency interaction for music audio–visual question answering. In *International Conference on Learning Representations*, pp. 81734–81752, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182, 2023.

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*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF conference on computer vision and pattern recognition*, pp. 10674–10685, 2022.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *International Conference on Learning Representations*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *International Conference on Learning Representations*, 2023.

Anqi Tang, Wenhao Sun, and Zhaoqiang Liu. Wavelet-guided semantic signal compensation for inversion-free image editing. In *European Conference on Computer Vision*, pp. 671–688, 2026.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. In *International Conference on Machine Learning*, volume 267, pp. 64044–64058, 2025.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18359–18369, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

Chenxi Xie, Minghan Li, Shuai Li, Yuhui Wu, Qiaosi Yi, and Lei Zhang. Dnaedit: Direct noise alignment for text-guided rectified flow editing. *Advances in Neural Information Processing Systems*, 38:124456–124474, 2025.

Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22428–22437, 2023.

Rui Xu, Yang Yong, Shunzi Yang, Ruihao Gong, and Chengtao Lv. Maskflow: Precise, consistent and seamless regional image editing. *arXiv preprint arXiv:2608.06929*, 2026.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9454–9461, 2024.

Desong Yang and Mang Ye. Directedit: Step-level accurate inversion for flow-based image editing. In *International Conference on Machine Learning*, 2026.

Kaixiang Yang, Xin Li, Yuxi Li, Qiang Li, and Zhiwei Wang. Fsi-edit: Frequency and stochasticity injection for flexible diffusion-based image editing. *Advances in Neural Information Processing Systems*, 38:133100–133132, 2025a.

Kaixiang Yang, Boyang Shen, Xin Li, Yuchen Dai, Yuxuan Luo, Yueran Ma, Wei Fang, Qiang Li, and Zhiwei Wang. Fia-edit: Frequency-interactive attention for efficient and high-fidelity inversion-free text-guided image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11613–11621, 2026.

Xiaofeng Yang, Cheng Chen, Fayao Liu, and Guosheng Lin. Text-to-image rectified flow as plug-and-play priors. In *International Conference on Learning Representations*, pp. 13896–13920, 2025b.

Sung-Hoon Yoon, Minghan Li, Gaspard Beaudouin, Congcong Wen, Muhammad Rafay Azhar, and Mengyu Wang. Splitflow: Flow decomposition for inversion-free text-to-image editing. *Advances in Neural Information Processing Systems*, 38:153207–153225, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Zijun Zhou, Yingying Deng, Xiangyu He, Weiming Dong, and Fan Tang. Multi-turn consistent image editing. In *IEEE/CVF International Conference on Computer Vision*, pp. 15792–15801, 2025.

A Appendix

This appendix is organized as follows::

- Sec. A.1 provides the theoretical derivation and additional justification of the proposed focused trajectory regularization.
- Sec. A.2 presents further implementation details for FlowFocus and the compared baselines.
- Sec. A.3 reports category-wise comparisons with representative inversion-free methods, together with additional ablation studies on the key hyperparameters of FlowFocus.
- Sec. A.4 provides additional qualitative results of FlowFocus on the PIE benchmark.
- Sec. A.5 describes the protocol and detailed settings of the human preference study.
- Sec. A.6 compares the inference runtime of FlowFocus with representative baselines.
- Sec. A.7 discusses the limitations of FlowFocus, presents representative failure cases, and outlines potential directions for future work.
- Sec. A.8 discusses the broader impacts and potential implications of the proposed generative editing framework.

A.1 Proof of Proposition 1

We provide the derivation of Proposition 1. Our proof first establishes the optimal solution for terminal-regularized control under a spatially varying quadratic metric, and then specializes this result to the focused trajectory regularization used in FlowFocus.

Let

$$R_t = \mathrm{Diag}\,(r_{t,1}, \ldots, r_{t,N}) \otimes I_C, \qquad M_t := R_t^{-1}, \tag{17}$$

where

$$r_{t,i} = \delta + (1-\delta)\rho_{t,i}, \tag{13}$$

Since $0 < \delta \leq 1$ and $\rho_{t,i} \in [0,1]$, we have $R_t \succ 0$. The running cost can therefore be written as

$$\frac{1}{2}(\boldsymbol{k}_t - \boldsymbol{a}_t)^\top R_t (\boldsymbol{k}_t - \boldsymbol{a}_t), \tag{18}$$

Throughout the derivation, $\boldsymbol{a}_t$ and R_t are treated as locally fixed fields evaluated from the current pretrained-flow predictions, consistent with the first-order frozen-field approximation used in the main text.

Lemma 1. *Consider the reverse-time optimal-control problem*

$$\dot{\boldsymbol{x}}_t = \boldsymbol{k}_t, \qquad \boldsymbol{x}_{t_0} = \boldsymbol{x}_{start}, \tag{19}$$

with objective

$$\int_0^{t_0} \frac{1}{2}(\boldsymbol{k}_t - \boldsymbol{a}_t)^\top R_t (\boldsymbol{k}_t - \boldsymbol{a}_t)\, dt + \frac{\eta}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_{src}\|_2^2, \tag{20}$$

Define

$$\Sigma_{t_0} := \int_0^{t_0} M_t \, dt, \qquad \boldsymbol{b}_{t_0} := \int_{t_0}^{0} \boldsymbol{a}_t \, dt, \tag{21}$$

Provided $I - \eta\Sigma_{t_0}$ is nonsingular, the optimal control is

$$\boldsymbol{k}_t^\star = \boldsymbol{a}_t - M_t\xi, \tag{22}$$

where

$$\xi = \eta(I - \eta\Sigma_{t_0})^{-1}\left(\boldsymbol{x}_{start} + \boldsymbol{b}_{t_0} - \boldsymbol{x}_{src}\right), \tag{23}$$

Proof. The corresponding Hamiltonian is

$$H(\boldsymbol{x}_t, \boldsymbol{k}_t, \xi_t) = \frac{1}{2}(\boldsymbol{k}_t - \boldsymbol{a}_t)^\top R_t(\boldsymbol{k}_t - \boldsymbol{a}_t) + \xi_t^\top \boldsymbol{k}_t, \tag{24}$$

The it gives

$$\boldsymbol{k}_t^\star = \boldsymbol{a}_t - M_t\xi_t, \tag{25}$$

Under the frozen-field approximation, the Hamiltonian has no explicit dependence on $\boldsymbol{x}_t$. Therefore, $\xi_t = \xi$ is constant over the local trajectory.

Integrating Eq. (25) backward from t_0 to 0 gives

$$\begin{aligned}\boldsymbol{x}_0 &= \boldsymbol{x}_{start} + \int_{t_0}^{0} (\boldsymbol{a}_t - M_t\xi)\, dt \\ &= \boldsymbol{x}_{start} + \boldsymbol{b}_{t_0} + \Sigma_{t_0}\xi,\end{aligned} \tag{26}$$

The positive sign of the last term in Eq. (26) follows from the reversed integration interval. The terminal condition induced by the source-similarity loss is

$$\xi = \eta(\boldsymbol{x}_0 - \boldsymbol{x}_{src}), \tag{27}$$

Substituting Eq. (26) into Eq. (27) yields

$$(I - \eta\Sigma_{t_0})\,\xi = \eta\left(\boldsymbol{x}_{start} + \boldsymbol{b}_{t_0} - \boldsymbol{x}_{src}\right). \tag{28}$$

Provided $I - \eta\Sigma_{t_0}$ is nonsingular,

$$\xi = \eta\left(I - \eta\Sigma_{t_0}\right)^{-1}\left(\boldsymbol{x}_{start} + \boldsymbol{b}_{t_0} - \boldsymbol{x}_{src}\right). \tag{29}$$

which proves Lemma 1. □

Proof of Proposition 1. We now apply Lemma 1 to the inversion-free editing velocity

$$\boldsymbol{a}_t = \boldsymbol{v}_t^\theta(\boldsymbol{p}_t, \boldsymbol{c}_{tgt}) - \boldsymbol{v}_t^\theta(\boldsymbol{q}_t, \boldsymbol{c}_{src}), \tag{8}$$

At the current timestep t, the first-order frozen-field approximation gives

$$R_s \simeq R_t, \qquad \boldsymbol{a}_s \simeq \boldsymbol{a}_t, \qquad s \in [0, t], \tag{30}$$

Consequently,

$$\Sigma_t = \int_0^t M_s \, ds \simeq tM_t, \qquad \boldsymbol{b}_t = \int_t^0 \boldsymbol{a}_s \, ds \simeq -t\boldsymbol{a}_t, \tag{31}$$

Using Eq. (25) and Eq. (29), the locally optimal control becomes

$$\boldsymbol{k}_t^\star \simeq \boldsymbol{a}_t - \eta M_t\left(I - \eta t M_t\right)^{-1}\left(\boldsymbol{x}_t - t\boldsymbol{a}_t - \boldsymbol{x}_{src}\right). \tag{32}$$

From the auxiliary-flow construction,

$$\boldsymbol{x}_t - \boldsymbol{x}_{\mathrm{src}} = \boldsymbol{p}_t - \boldsymbol{q}_t. \tag{33}$$

Combining this identity with Eq. (8), we obtain

$$\begin{aligned}\boldsymbol{x}_t - t\boldsymbol{a}_t - \boldsymbol{x}_{src} &= \boldsymbol{p}_t - \boldsymbol{q}_t - t\left[\boldsymbol{v}_t^\theta(\boldsymbol{p}_t, \boldsymbol{c}_{tgt}) - \boldsymbol{v}_t^\theta(\boldsymbol{q}_t, \boldsymbol{c}_{src})\right] \\ &= \hat{\boldsymbol{p}}_0^T - \hat{\boldsymbol{q}}_0^S,\end{aligned} \tag{34}$$

where

$$\hat{\boldsymbol{p}}_0^T = \boldsymbol{p}_t - t\boldsymbol{v}_t^\theta(\boldsymbol{p}_t, \boldsymbol{c}_{tgt}), \qquad \hat{\boldsymbol{q}}_0^S = \boldsymbol{q}_t - t\boldsymbol{v}_t^\theta(\boldsymbol{q}_t, \boldsymbol{c}_{src}), \tag{35}$$

Therefore,

$$\frac{d\boldsymbol{x}_t}{dt} \simeq \boldsymbol{a}_t + \Gamma_t \left(\hat{\boldsymbol{p}}_0^T - \hat{\boldsymbol{q}}_0^S\right), \tag{36}$$

where

$$\Gamma_t := -\eta M_t \left(I - \eta t M_t\right)^{-1}, \tag{15}$$

Since R_t is block diagonal over spatial locations, so are M_t and Γ_t. At the i-th spatial location,

$$[M_t]_i = \frac{1}{r_{t,i}} I_C, \tag{37}$$

Hence, for $\hat{\boldsymbol{p}}_{0,i}^T := [\hat{\boldsymbol{p}}_0^T]_i$ and $\hat{\boldsymbol{q}}_{0,i}^S := [\hat{\boldsymbol{q}}_0^S]_i$,

$$\left[\Gamma_t \left(\hat{\boldsymbol{p}}_0^T - \hat{\boldsymbol{q}}_0^S\right)\right]_i = \gamma_{t,i} \left(\hat{\boldsymbol{p}}_{0,i}^T - \hat{\boldsymbol{q}}_{0,i}^S\right), \tag{38}$$

with

$$\begin{aligned} \gamma_{t,i} &= -\frac{\eta}{r_{t,i}} \left(1 - \frac{\eta t}{r_{t,i}}\right)^{-1} \\ &= \frac{\eta}{\eta t - r_{t,i}}, \end{aligned} \tag{39}$$

Substituting Eq. (13) gives

$$\gamma_{t,i} = \frac{\eta}{\eta t - \delta - (1-\delta)\rho_{t,i}}, \tag{40}$$

Finally, when $\delta = 1$, we have $r_{t,i} = 1$ for every spatial location, and Eq. (40) reduces to

$$\gamma_{t,i} = \frac{\eta}{\eta t - 1}, \tag{41}$$

which recovers the global trajectory-regularization coefficient. Therefore, global trajectory regularization is a special case of the proposed focused formulation. This completes the proof of Proposition 1. □

A.2 More Implementation details

We provide additional implementation details for FlowFocus and the compared baselines. FlowFocus was implemented on latent rectified-flow models, using Stable Diffusion 3 (SD3[1]) and Stable Diffusion 3.5 (SD3.5[2]) (Esser et al., 2024) from the Hugging Face Diffusers library. For comparisons under the same model family, we used identical pretrained backbones whenever supported to minimize performance differences caused by the underlying generative model. Unless otherwise specified, baseline methods were evaluated using their publicly released implementations and recommended configurations. All quantitative metrics were computed using the official PIE evaluation code.[3] Experiments were conducted on a single NVIDIA L40S GPU unless stated otherwise. The implementation of FlowFocus will be **made publicly available** upon acceptance.

A.3 Additional Experimental Analysis

Category-wise editing performance. To further examine the robustness of FlowFocus across different editing scenarios, we report category-wise results on PIE in Table 3. The benchmark contains nine editing categories: "change object", "add object", "delete object", "change attribute content", "change attribute pose", "change attribute color", "change attribute material", "change background", and "change style". We additionally report results on the provided 140-sample random subset. The random subset is drawn from the same 700 benchmark samples and is not counted as an additional category.

[1] https://huggingface.co/stabilityai/stable-diffusion-3-medium

[2] https://huggingface.co/stabilityai/stable-diffusion-3.5-medium

[3] https://github.com/cure-lab/PnPInversion/tree/main/evaluation

Table 3: Category-wise quantitative comparison on PIE.

Editing task	Number	FlowEdit		FlowAlign		Ours	
		PSNR↑	CLIP↑	PSNR↑	CLIP↑	PSNR↑	CLIP↑
0-random	140	23.88	22.34	25.73	22.46	28.06	23.55
1-change object	80	22.61	22.59	25.35	21.54	25.80	23.12
2-add object	80	26.24	25.32	26.27	24.52	27.72	25.16
3-delete object	80	24.49	21.78	28.87	19.25	28.07	20.87
4-change attribute content	40	24.99	20.16	27.52	21.79	27.11	21.46
5-change attribute pose	40	25.60	22.87	28.79	21.91	29.15	22.42
6-change attribute color	40	24.91	20.85	27.42	20.85	27.65	21.71
7-change attribute material	40	26.58	24.36	25.73	23.68	27.48	29.01
8-change background	80	24.63	24.16	26.46	22.79	27.68	23.10
9-change style	80	21.91	27.86	25.52	26.49	26.99	27.68
Overall	700	24.31	23.42	26.54	22.63	27.55	23.55

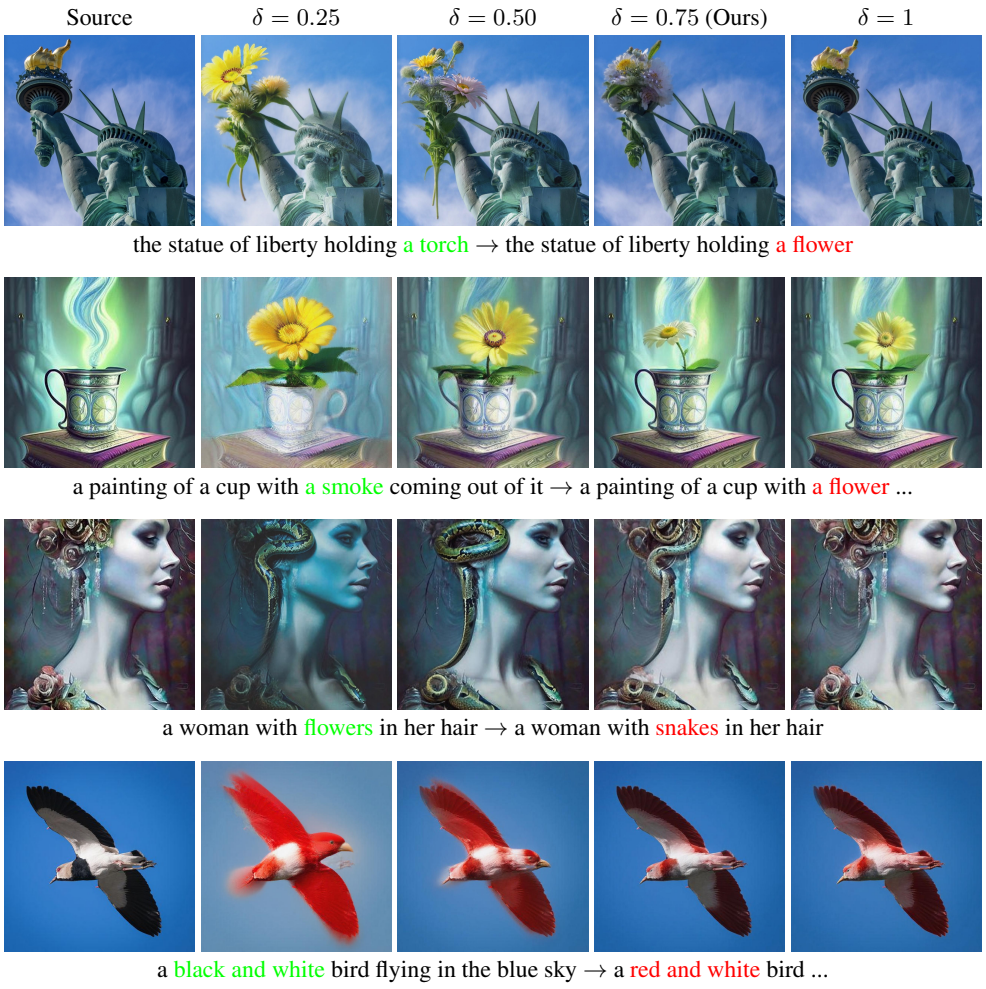


Figure 6: Qualitative ablation results for δ.

We compared FlowFocus with two representative inversion-free methods, FlowEdit and FlowAlign, using PSNR to measure background preservation and CLIP similarity to assess target alignment. FlowFocus showed consistently strong performance across diverse editing categories and achieved the best overall results on both metrics. In particular, the improvements were observed across both localized edits, such as object and attribute manipulation, and more spatially distributed changes,

such as background and style editing. These results further indicate that our proposed focused regularization can adapt to different spatial extents of semantic modification while maintaining a favorable balance between source preservation and target alignment.

Ablation on δ. In the main text, we analyze the joint effect of the focused-regularization parameter δ and target CFG scale ω on the preservation–editability trade-off in Sec. 4.4 and Fig. 5. To complement the analysis, Fig. 6 provides qualitative comparisons under a fixed CFG scale of $\omega = 13.5$, highlighting how different values of δ affect the editing behavior.

As δ controls the degree of spatial adaptation in FlowFocus, smaller values allow stronger source-consistency correction in preservation-dominant regions, while larger values make the regularization progressively closer to the global formulation. Although the optimal operating point can vary slightly across individual samples, $\delta = 0.75$ consistently provided a favorable balance between target-driven modification and source structure preservation across diverse editing cases. In contrast, overly small or large values tend to bias the result toward excessive preservation or insufficiently focused regularization, respectively. For example, in the first row of Fig. 6, the focused edit region accurately covered the torch and replaced it with flowers, while preserving the spatial extent of the original object; notably, the length of the flower stem remained consistent with the length of the torch handle in the source image. These qualitative observations are consistent with the quantitative trends reported in the main text.

Effect of CFG scale ω. As discussed in Sec. 4.4, the target CFG scale ω plays an important role in balancing semantic alignment and source preservation. Following prior work (Kim et al., 2026; Beaudouin et al., 2026), we further evaluated FlowFocus under $\omega \in \{10, 13.5, 16.5\}$ and compared it with two representative inversion-free baselines, FlowEdit and FlowAlign. We report CLIP similarity for target alignment and PSNR for source preservation. For FlowAlign, its regularization coefficient ζ was fixed to the default setting recommended in the original implementation, such that the comparison mainly reflects the influence of CFG. As shown in Fig. 7, all methods exhibited a similar trend: increasing ω strengthened target guidance and improved CLIP similarity, but gradually reduced PSNR due to larger deviations from the source image. Despite this inherent trade-off, FlowFocus consistently maintained substantially higher background preservation across all CFG settings while also achieving stronger semantic alignment than FlowEdit and FlowAlign at the corresponding guidance scales. In particular, the operating curve of FlowFocus remained clearly separated from those of the competing methods, indicating that focused trajectory regularization provides a more favorable balance between target-driven editing and source consistency over a broad range of guidance strengths.

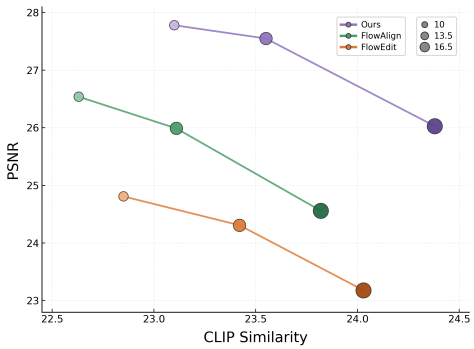


Figure 7: Effect of the target CFG scale ω on different inversion-free editing methods.

A.4 More Qualitative Results

To further demonstrate the effectiveness of our method, we conducted additional qualitative comparisons on text-guided image editing tasks. Specifically, Fig. 8 highlights visual evaluations against competing baselines across various editing categories and object types. These additional results illustrate that FlowFocus delivers superior editing performance while better preserving the underlying source structure.

A.5 Human Preference Test Protocol

We conducted the human evaluation on a subset of 140 samples selected from PIE. Each pairwise comparison was independently evaluated by at least three participants. With 140 samples, this resulted in 420 pairwise evaluations and 840 criterion-level judgments across the two evaluation dimensions. For each evaluation, participants were shown the source image, the corresponding editing instruction, and two edited results. Although the editing methods operated on source–target text

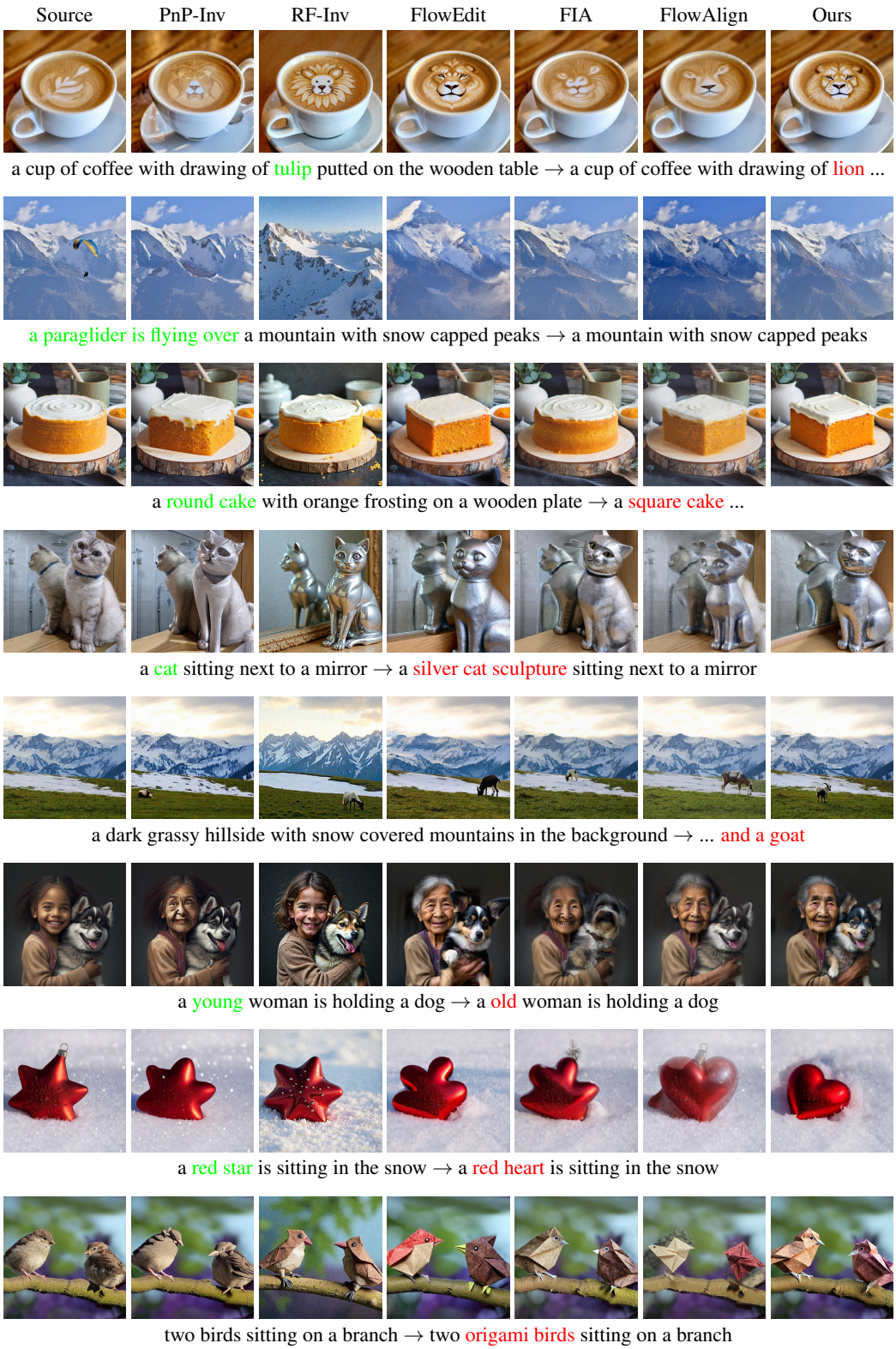


Figure 8: Additional qualitative comparison results from PIE.

Method	P2P	PnP	RF-inv	FireFlow	FlowEdit	FIA	FlowAlign	Ours
Time	36.62	19.82	78.59	35.41	3.60	7.18	3.23	3.93

Table 4: Runtime comparison of different image editing methods. Runtime was measured in seconds and averaged over 20 editing samples.

pairs, we presented the editing instructions provided by PIE to participants instead, which makes the intended modification more explicit and reduces unnecessary cognitive load. The identities of the compared methods were not disclosed, and the display order of the two edited results was randomly shuffled for each comparison such that FlowFocus is not associated with a fixed position.

Participants independently assessed the results according to two criteria: *editing accuracy*, which measures how faithfully the edited result satisfies the requested modification, and *source preservation*, which evaluates how well content unrelated to the requested edit remains consistent with the source image. For each criterion, participants selected the first result, the second result, or "Tie" when neither result exhibits a clear advantage. The responses were aggregated separately for the two criteria and reported as win/tie/loss ratios against each baseline (as shown in Fig. 4).

A.6 Runtime Comparison

We report the inference runtime of representative open-source methods in Table 4. For a fair comparison, most methods were evaluated on a single NVIDIA L40S GPU, while RF-Inv and FireFlow were run on an NVIDIA A100 (80GB) due to their higher memory requirements. Runtime is averaged over 20 editing samples at a resolution of 512×512 and includes the complete pipeline from model input to saving the edited image. All inversion-free methods were evaluated with the same SD3 (Esser et al., 2024) backbone.

As shown in Table 4, inversion-free approaches were substantially more efficient than inversion-based counterparts. FlowFocus required 3.93 seconds per image, remaining close to FlowEdit and FlowAlign, while being considerably faster than others. The modest overhead over FlowEdit mainly arised from computing the counterfactual edit field and spatially adaptive regularization at each timestep. Nevertheless, FlowFocus preserved the efficiency advantage of inversion-free editing while providing improved editing quality and source preservation.

A.7 Limitations and Failure Cases

Despite its strong balance between editability and source preservation, FlowFocus still exhibits several limitations, particularly in visually complex scenes. When the source image contains dense fine-grained details involving diverse colors, shapes, scales, and poses, preserving all edit-irrelevant content remains challenging even if the requested modification is spatially localized. As shown in the first row of Fig. 9, although FlowFocus successfully added the deer according to the target prompt, the highly complex flower regions were not perfectly preserved by any method, including ours. This suggests that the current focused regularization may still be insufficient for maintaining very fine-grained local structures under substantial semantic modification. A second limitation arises when the desired edit requires smooth spatial completion rather than simple semantic replacement. For example, in object-removal cases, trajectory regularization can overly preserve source-related structure around the removed region, leaving residual traces of the original object. In the third row of Fig. 9, most inversion-free methods, including FlowFocus, successfully removed the bus, yet faint ghosting and part of its original contour remained visible. This indicates that stronger source consistency can occasionally hinder the synthesis of seamless content across regions that require complete structural reconstruction.

Future work could further improve the flexibility of focused regularization by incorporating boundary-aware or context-adaptive control, allowing the model to relax source constraints when broader structural changes are necessary while retaining strong preservation elsewhere. Extending the counterfactual edit field to better capture fine-grained appearance changes and long-range contextual dependencies may also improve robustness in highly detailed scenes and challenging tasks.

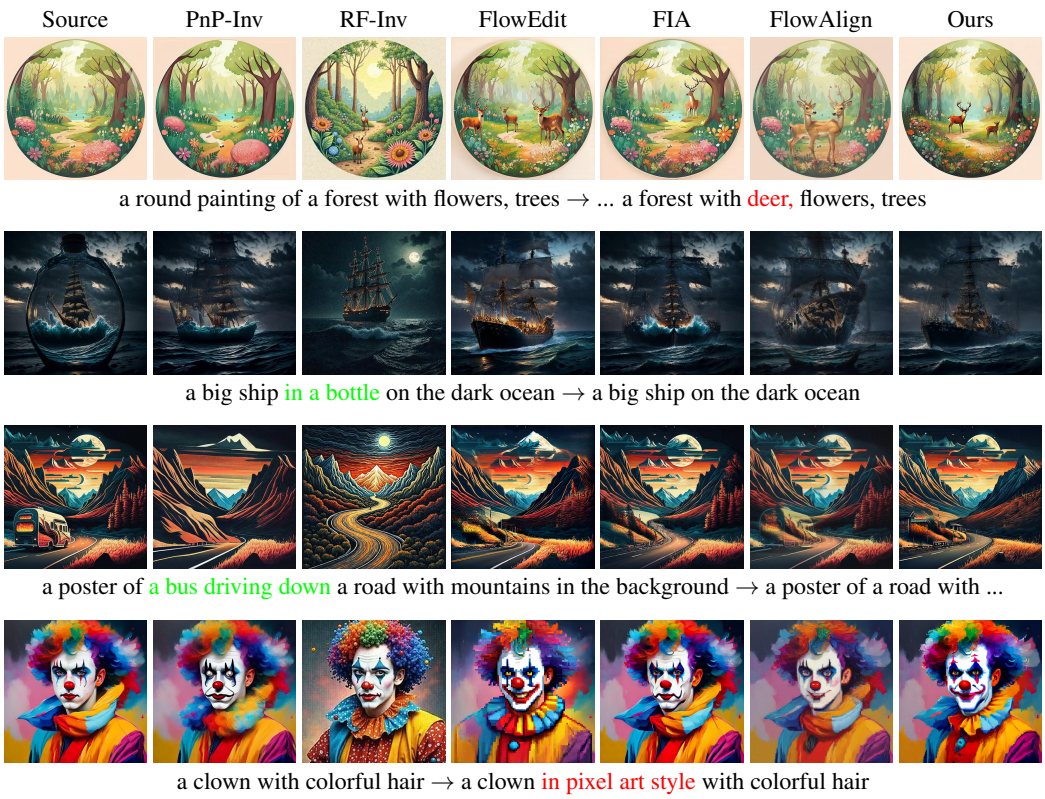


Figure 9: Failure cases from PIE.

A.8 Broader Impacts

While our generative framework offers substantial benefits for creative workflows, empowering artists and designers with precise tools for visual content creation, it also introduces potential ethical considerations. Specifically, like many advanced image synthesis systems, the technology carries a risk of misuse for generating misleading or harmful media, raising concerns regarding authenticity and privacy. To address these challenges, we advocate for responsible deployment strategies, including controlled or gated access models, alongside the parallel development of robust attribution and detection mechanisms to counteract malicious exploitation.