Wavelet-Guided Semantic Signal Compensation for Inversion-Free Image Editing

Anqi Tang*, Wenhao Sun*, and Zhaoqiang Liu(✉)

School of Computer Science and Engineering
University of Electronic Science and Technology of China
anqitang@std.uestc.edu.cn,202511081644@std.uestc.edu.cn,zqliu12@gmail.com

Abstract. Text-guided image editing aims to modify visual content according to a target prompt while preserving the background. Recent inversion-free image editing frameworks such as FlowEdit have demonstrated strong editing capability without requiring inversion. Empirically, FlowEdit can achieve substantial semantic changes under appropriate hyperparameter settings. However, we observe that under certain global attribute shifts, the editing trajectory may not effectively move away from the source distribution in the early timesteps. Our analysis suggests that in the high-noise regime, the dominant manifold-seeking flow toward the data manifold can reduce the influence of the text-conditioned direction, leading to limited global modification while background structures remain only moderately preserved. Inspired by this observation, we propose an inversion-free, frequency-aware semantic compensation strategy that strengthens the effective signal in the early stage of generation, while maintaining structural consistency in the background. The proposed method improves global editing capacity without sacrificing background fidelity.

Keywords: Inversion-Free Image Editing · Rectified Flow · Wavelet Transform

1 Introduction

Diffusion models [13, 33, 34] and rectified flows (RFs) [23, 24] have enabled high-fidelity image synthesis by progressively transforming noise into realistic images, achieving strong generation quality in practice. Building on their strong generative priors, diffusion and flow models have also become a foundation for text-guided image editing, where the goal is to change semantics under textual control while preserving the structure of input. Traditional methods [10, 11, 30, 37, 45] typically rely on explicit semantic inversion to modify content while preserving spatial layouts. However, this process is plagued by reconstruction drift and high computational overhead.

* Equal contribution
(✉) Corresponding author
Project Page: https://zoey-tang-33.github.io/wavelet-edit-project

To avoid these bottlenecks, recent inversion-free editing frameworks [2, 18, 19, 21, 27] modify images directly by dynamically estimating the velocity difference between a source trajectory and a target trajectory. By integrating this differential velocity field over time, these methods can guide the generative flow toward the target semantic domain without retracing the exact noise latent, thereby inherently retaining the structural priors of the source image. While these inversion-free methods demonstrate impressive editing capability and structural preservation, we observe a clear limitation when they are applied to global semantic modifications, such as shifting overall color attributes, as illustrated in Fig. 1. We attribute this behavior to what can be described as semantic indistinguishability in the high-noise regime. At the initialization of the generation process, the latent state is dominated by unstructured noise. In this regime, the predicted velocity field is largely governed by the fundamental generative objective of transporting the noise distribution toward the coarse natural image manifold. This dominant manifold-seeking flow tends to outweigh the comparatively subtle semantic signal introduced by the text conditioning. As a result, the target trajectory remains strongly aligned with the structural prior inherited from the source image, limiting its ability to deviate toward new global attributes. Consequently, during the early steps, when coarse structures and low-frequency components are established, the editing dynamics are primarily dictated by structural preservation rather than semantic transformation. By the time the noise level decreases and text conditioning becomes more influential, the global layout and color composition have already been largely fixed, significantly reducing the capacity for effective large-scale semantic modification.

To address this insufficient semantic deviation, it is necessary to extract a guidance signal that is disentangled from the dominant manifold-seeking flow. We propose same-point semantic probing, which evaluates both source and tar-

Fig. 1: Failure cases of inversion-free RF-based editing methods, including FlowEdit [19], FlowAlign [18], and DVRF [2], under global semantic modifications. When performing global attribute shifts, existing methods either introduce structural distortions, semantic drift, or incomplete color transformation. In contrast, our method achieves more consistent global attribute changes while preserving structural details.

get conditions at the exact same latent state. By locking the spatial coordinates, the shared generative flow creates a baseline that allows us to obtain a co-located prompt-induced direction that is less confounded by the spatial separation between source and target latents. However, directly injecting this raw signal into the editing flow can introduce high-frequency perturbations that degrade the structural integrity of the source image.

To resolve this spatial-frequency conflict, we introduce a Wavelet-based Semantic Signal Compensation framework. Leveraging the 2D discrete Haar wavelet transform (DWT) [26], we explicitly decouple the frequency bands of the semantic editing signal. We isolate the low-frequency components, while filtering out high-frequency detail coefficients. This prior is injected into the editing flow via a time-modulated update rule, ensuring maximum guidance during the early structural formation phase while quadratically decaying to preserve fidelity during the textural refinement phase.

The main contributions of this paper are summarized as follows:

- We identify and analyze a limitation of inversion-free editing under rectified flow models, where the semantic editing signal becomes weak in high-noise regimes due to the dominance of manifold-seeking flow. We show that this effect limits the accumulation of directional deviation required for global attribute changes.
- We propose a wavelet-based semantic signal compensation mechanism that extracts low-frequency directions from identical latent states while suppressing high-frequency perturbations. This design strengthens global semantic modification while preserving structural consistency.
- Extensive experiments, including quantitative evaluation, user studies, and frequency-domain analysis, demonstrate that our method achieves stronger and more consistent semantic editing across diverse tasks.

2 Related Work

Existing approaches in text-guided image editing can generally be categorized according to their generative backbones, namely diffusion models or rectified flows.

2.1 Text-Guided Image Editing with Diffusion Models

A principled family of methods first inverts the source image into the latent space of the diffusion model and then regenerates it under a modified text prompt. DDIM Inversion [12, 17] retraces the deterministic sampling trajectory to estimate the noise latent that reconstructs the input, but the approximate nature of the inverse process often leads to reconstruction drift, especially under low-step settings. To solve this issue, subsequent works have proposed more consistent inversion schemes [10, 11, 30, 37, 45], guidance-based manipulations [29, 47], and higher-order solvers [3, 9, 14].

There are also several works focusing on manipulating internal representations for structural preservation [4, 12, 36, 42]. P2P [12] controls cross-attention maps to align spatial layout, while PnP [36] and MasaCtrl [4] inject spatial features and mutual self-attention queries, respectively, to maintain geometric consistency during editing. Beyond attention-level control, FreeDiff [40] exploits frequency-domain priors by employing a training-free progressive frequency truncation strategy to maintain structural consistency along the denoising steps.

While FreeDiff also leverages frequency decomposition, our method differs in operating domain and control objective. First, FreeDiff operates directly in the latent space via fast Fourier transform (FFT) [6], aiming to constrain the generation by imposing the structural prior of source. This is a restrictive operation designed to prevent structure collapse in diffusion models. In contrast, our method operates in the velocity space of rectified flow models. Our goal is injection rather than constraint. We apply a wavelet transform to the prompt-induced editing signal and inject its low-frequency component into the velocity field to drive the necessary global deviations.

2.2 Text-Guided Image Editing in Rectified Flows

Within rectified flow models, inversion-based editing has been advanced by RF-Edit [38], which introduces a training-free high-order solver via Taylor expansion of the nonlinear ordinary differential equation term. RF-Inv [32] formulates semantic image inversion under rectified stochastic differential equations, enabling text-guided editing within the RF framework. FireFlow [7] achieves a 3× speed-up with a second-order solver. ABM-Flow [25] adopts a multi-step predictor–corrector scheme with mask-guided latent injection. DNA-Edit [41] aligns noise trajectories directly to facilitate semantic manipulation in rectified flow models.

A class of inversion-free methods bypasses explicit reconstruction altogether. FlowEdit [19] formulates editing by aligning the velocity fields of the source and target trajectories. Building upon FlowEdit, recent works have focused on stabilizing the editing trajectory. FlowAlign [18] and TweezeEdit [27] assume shared noise conditions to simplify velocity-difference estimation under optimal control and path-regularization objectives. DVRF [2] generalizes these formulations with an additional correction term.

While highly efficient, we observe that the editing signal of FlowEdit often diminishes in the high-noise regime due to the dominant manifold-seeking transport, leading to limited global semantic modifications. To address this semantic indistinguishability, our method introduces a wavelet-guided compensation mechanism to enforce stronger global semantic guidance during the early sampling steps.

In the broader context of image editing, frequency-domain constraints have been introduced to disentangle semantic changes from structural preservation. Recent inversion-free editing methods incorporate frequency-aware mechanisms, including FFT-based frequency stochasticity injection and frequency interactive attention, as well as wavelet-based multi-scale decomposition frameworks [35,43,

44]. Different from them, our method retains the original pre-trained flow model without attention modifications.

3 Preliminaries

In this section, we summarize the formulation of rectified flow, and then describe the translation-coupled editing mechanism used in FlowEdit. For brevity, we provide a concise overview, and defer the full derivations and detailed discussions to the supplementary material.

3.1 Rectified Flow

Generative flow models aim to construct a transport between two distributions, a simple source distribution p_1 (typically standard Gaussian noise) to the complex data distribution p_0 [23].Throughout this work, we follow the formulation in FlowEdit [19], where $t = 1$ corresponds to pure noise and $t = 0$ corresponds to clean data. The generative process is defined by an ordinary differential equation (ODE) flowing from $t = 1$ to $t = 0$, governed by a velocity field $\boldsymbol{v}_t$ with $\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}_t)$.

Rectified flow [24] adopts a linear interpolation path between noise and data: $\boldsymbol{x}_t = t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0$. We discretize the time interval $[0, 1]$ into N subintervals with timesteps $1 = t_0 > t_1 > \cdots > t_N = 0$. Starting from initial noise $\boldsymbol{x}_{t_0} \sim p_1$, we apply a first-order Euler discretization to numerically integrate the ODE:

$$\boldsymbol{x}_{t_{i+1}} = \boldsymbol{x}_{t_i} + (t_{i+1} - t_i)\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_i}, t_i), \quad i \in \{0, \ldots, N-1\}, \tag{1}$$

where $\boldsymbol{v}_{\boldsymbol{\theta}}$ denotes the pre-trained rectified flow model.

3.2 FlowEdit

FlowEdit [19] formulates image editing by constructing a direct path from the source distribution to the target distribution via rectified flow. Given a source image $\boldsymbol{x}^{\mathrm{src}}$, for any time step $t \in [0, 1]$, we independently sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and define $\boldsymbol{x}_t^{\mathrm{src}} = (1-t)\boldsymbol{x}^{\mathrm{src}} + t\boldsymbol{\epsilon}$. The idea of FlowEdit is to define the target trajectory $\boldsymbol{x}_t^{\mathrm{tar}}$ as a deviation from the source trajectory, driven by the semantic difference. Let $\boldsymbol{e}_{\mathrm{src}}$ and $\boldsymbol{e}_{\mathrm{tar}}$ denote the text embeddings of the source and target prompts, respectively. We define the delta velocity $\Delta\boldsymbol{v}_t$ between the target and source vector fields as:

$$\Delta\boldsymbol{v}_t \triangleq \boldsymbol{v}_\theta(\boldsymbol{x}_t^{\mathrm{tar}}, t; \boldsymbol{e}_{\mathrm{tar}}) - \boldsymbol{v}_\theta(\boldsymbol{x}_t^{\mathrm{src}}, t; \boldsymbol{e}_{\mathrm{src}}). \tag{2}$$

FlowEdit constructs a target flow that preserves the structural dynamics of the source flow while integrating this semantic velocity. This corresponds to the coupled differential equation:

$$\mathrm{d}\boldsymbol{x}_t^{\mathrm{tar}} = \mathrm{d}\boldsymbol{x}_t^{\mathrm{src}} + \Delta\boldsymbol{v}_t \mathrm{d}t. \tag{3}$$

By discretizing Eq. (3), we obtain the practical Euler update rule that combines the exact source transport with the semantic edit:

$$\boldsymbol{x}^{\text{tar}}_{t_{i+1}} = \boldsymbol{x}^{\text{tar}}_{t_i} + \underbrace{(\boldsymbol{x}^{\text{src}}_{t_{i+1}} - \boldsymbol{x}^{\text{src}}_{t_i})}_{\text{source transport}} + \underbrace{(t_{i+1} - t_i)\Delta\boldsymbol{v}_{t_i}}_{\text{semantic edit}}. \tag{4}$$

This formulation ensures that the target generation faithfully follows the structural layout of the source image through the transport term, while $\Delta\boldsymbol{v}_{t_i}$ injects the necessary semantic alterations.

3.3 Two-Dimensional Haar Wavelet Transform

Given a spatial tensor $X \in \mathbb{R}^{H\times W}$, the single-level 2D Haar transform [26] decomposes each non-overlapping 2×2 block into four sub-bands (cA, cH, cV, cD) representing local averages and directional differences. The forward transform is defined as

$$cA = \tfrac{1}{2}(a + b + c + d), \tag{5}$$
$$cH = \tfrac{1}{2}(a - b + c - d), \tag{6}$$
$$cV = \tfrac{1}{2}(a + b - c - d), \tag{7}$$
$$cD = \tfrac{1}{2}(a - b - c + d), \tag{8}$$

and the inverse transform exactly reconstructs the original samples. The transform is applied independently to each channel. Using the 2D Haar wavelet transform, we write the forward and inverse operators as $\text{DWT}(\cdot)$ and $\text{IDWT}(\cdot)$. The cA encodes the local spatial average, while cH, cV, and cD encode horizontal, vertical, and diagonal local differences, respectively. The single-level inverse transform reconstructs the original samples from these coefficients.

4 Method

We present a training-free augmentation to FlowEdit [19] that addresses its limited effectiveness in global attribute modification, such as large color shifts, as illustrated in Fig. 1. The overall mechanism is illustrated in Fig. 2.

4.1 Motivation: Semantic Indistinguishability in the High-Noise Regime

Although FlowEdit produces visually compelling edits in many settings, we identify a failure mode that limits its capability for global semantic modifications. Specifically, while the geometric editing signal $\Delta\boldsymbol{v}_{t_i}$ is non-zero during the high-noise phase in practice, it provides only weak and unreliable semantic directionality, which is often insufficient to initiate significant global attribute changes.

Because the early steps determine the global layout and dominant attributes of the generated image, the lack of informative guidance in this regime constrains

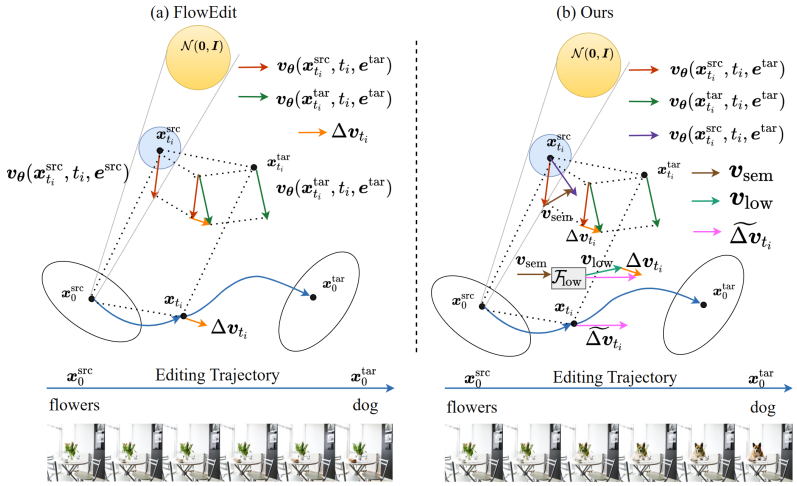

Fig. 2: Illustration of trajectory editing with FlowEdit and our method. FlowEdit directly applies a single residual direction $\Delta \boldsymbol{v}_{t_i}$ derived from the source and target velocity fields. Our method first estimates a co-located prompt-induced editing direction ($\boldsymbol{v}_{\text{sem}}$), extracts its spatially coherent low-frequency component ($\boldsymbol{v}_{\text{low}}$) as an auxiliary global compensation, and combines it with the original geometric update to obtain $\widetilde{\Delta \boldsymbol{v}}_{t_i}$.The bottom row visualizes editing trajectory, where the scene gradually replaces the flowers with a dog while preserving the background layout.

global semantic modification. Consequently, the target latent $\boldsymbol{x}_t^{\text{tar}}$ accumulates insufficient directional momentum along the probability flow trajectory, yielding only weak global changes (e.g., uniform color shifts). Instead, it remains closely aligned with the source trajectory and exhibits only minor attribute deviations. When the noise level finally becomes low enough for the prompt to exert stronger semantic control, the global structure and color composition have already been fixed, making the edit ineffective.

Analogous to the Tweedie formula in diffusion models, we recover the instantaneous expectation of the clean data $\hat{\boldsymbol{x}}_0$ by projecting the flow forward. Under the linear-trajectory assumption of rectified flow, the one-step estimator is [28]:

$$\hat{\boldsymbol{x}}_0(\boldsymbol{x}_{t_i}) = \boldsymbol{x}_{t_i} - t_i \cdot \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_i}, t_i). \tag{9}$$

Eq. (9) provides a deterministic preview of the final structure from any intermediate state t_i. As shown in Fig. 3, we examine $\hat{\boldsymbol{x}}_0$ computed along the target trajectory of FlowEdit. Despite the target prompt specifying a distinct attribute (e.g., a brown lizard), the semantic shift is largely suppressed by the source bias: only faint target components appear, and the trajectory ultimately returns to the source appearance, causing the edit to fail.

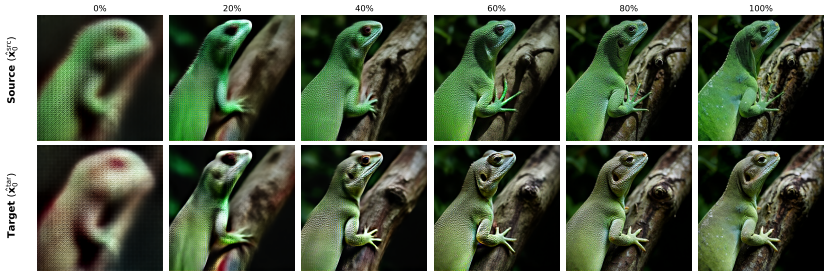

Fig. 3: Visualizing Semantic Signal Indistinguishability. Top Row: Source trajectory. **Bottom Row:** Target trajectory conditioned on a prompt requesting a global attribute change (green → brown).

4.2 Problem Formulation

Building on the analysis above, we formalize the main quantities involved and describe the principles guiding the proposed method.

Definition 1 (Semantic Editing Signal). *Given a noisy latent $\boldsymbol{x}$ at step i, source prompt embedding $\boldsymbol{e}_{\text{src}}$, and target prompt embedding $\boldsymbol{e}_{\text{tar}}$, the semantic editing signal is defined as*

$$\boldsymbol{v}_{\text{sem}}(\boldsymbol{x},\, t_i) \;\triangleq\; \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x},\, t_i;\, \boldsymbol{e}_{\text{tar}}) - \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x},\, t_i;\, \boldsymbol{e}_{\text{src}}). \tag{10}$$

Remark 1. The geometric editing signal $\Delta\boldsymbol{v}_{t_i}$ evaluates the velocity network at two distinct latent points $\boldsymbol{x}_t^{\text{tar}}$ and $\boldsymbol{x}_t^{\text{src}}$ under different prompts, so its magnitude can reflect both the prompt-conditioning difference and the spatial separation between the two inputs. By contrast, $\boldsymbol{v}_{\text{sem}}$ evaluates both prompts at the same latent point, thereby removing the geometric confound caused by distinct input locations. We use this co-located prompt-induced direction as an auxiliary editing signal, rather than as an assumption that all semantic changes are represented by a particular frequency band.

Definition 2 (Low-Frequency Extraction Operator). *Let L denote the number of Haar wavelet decomposition levels. The* low-frequency extraction operator $\mathcal{F}_{\text{low}} : \mathbb{R}^{B\times C\times H\times W} \to \mathbb{R}^{B\times C\times H\times W}$ *is defined as the composition of an L-level forward two-dimensional Haar decomposition, zeroing of all detail sub-bands at every level, and L-level inverse reconstruction:*

$$\mathcal{F}_{\text{low}}(X) = \text{IDWT}^{(1:L)}\Big(cA^{(L)},\; \{\mathbf{0},\mathbf{0},\mathbf{0}\}_{l=1}^{L}\Big), \tag{11}$$

where $cA^{(L)}$ is the L-th level approximation coefficient obtained by iterating the single-level forward transform on successive approximation sub-bands.

4.3 Wavelet-Based Semantic Signal Compensation

Multi-Level Decomposition and Low-Frequency Extraction Building upon the single-level Haar transform introduced in Sec. 3.3, we construct a multi-level decomposition to isolate progressively coarser structural components of the trajectory. The L-level forward decomposition iterates the single-level transform on the approximation sub-band only. Setting $cA^{(0)} = X$, the recursion for $l = 1, \ldots, L$ is

$$\big(cA^{(l)},\ cH^{(l)},\ cV^{(l)},\ cD^{(l)}\big) = \mathrm{DWT}\big(cA^{(l-1)}\big). \tag{12}$$

The operator $\mathcal{F}_{\mathrm{low}}$ defined in Definition 2 can be equivalently interpreted as a cascaded inverse reconstruction starting from $cA^{(L)}$, where all detail sub-bands are set to zero at each level.

Semantic Editing Signal After computing the noise-averaged geometric signal $\Delta \boldsymbol{v}_{t_i}$, one additional network evaluation is performed at the source noisy latent $\boldsymbol{x}_{t_i}^{\mathrm{src}}$ under the target prompt. The semantic editing signal is defined as

$$\boldsymbol{v}_{\mathrm{sem}}(\boldsymbol{x}_{t_i}^{\mathrm{src}},\, t_i) = \boldsymbol{v}_{\boldsymbol{\theta}}\big(\boldsymbol{x}_{t_i}^{\mathrm{src}},\, t_i;\, \boldsymbol{e}_{\mathrm{tar}}\big) - \boldsymbol{v}_{\boldsymbol{\theta}}\big(\boldsymbol{x}_{t_i}^{\mathrm{src}},\, t_i;\, \boldsymbol{e}_{\mathrm{src}}\big)\,. \tag{13}$$

Because both evaluations share the same input latent $\boldsymbol{x}_{t_i}^{\mathrm{src}}$, the difference $\boldsymbol{v}_{\mathrm{sem}}$ captures the co-located directional response induced by the change in text conditioning, while avoiding the confound introduced by evaluating the two prompts at distinct latent points. Since the source-prompt velocity $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_i}^{\mathrm{src}}, t_i; \boldsymbol{e}_{\mathrm{src}})$ is already available, this additional evaluation requires only one extra forward pass per active timestep under the target prompt at $\boldsymbol{x}_{t_i}^{\mathrm{src}}$. This design choice distinguishes our approach from FlowAlign [18] and CVC [21]. While they also utilize an additional network inference, they apply this correction to the target latent $\boldsymbol{x}_{t_i}^{\mathrm{tar}}$. In contrast, our method anchors the additional evaluation at the source latent $\boldsymbol{x}_{t_i}^{\mathrm{src}}$, so that the resulting signal serves as a prompt-conditioned auxiliary direction rather than a direct rectification of the edited trajectory.

Frequency-Selective Injection and the Modified Update Rule While $\boldsymbol{v}_{\mathrm{sem}}$ provides a useful prompt-induced auxiliary direction, its raw form may also contain spatially incoherent perturbations caused by the random forward noise embedded in $\boldsymbol{x}_{t_i}^{\mathrm{src}}$. Directly injecting the full signal can therefore disturb fine structures in the output. Accordingly, the low-frequency extraction operator is applied to obtain

$$\boldsymbol{v}_{\mathrm{low}} = \mathcal{F}_{\mathrm{low}}\big(\boldsymbol{v}_{\mathrm{sem}}(\boldsymbol{x}_{t_i}^{\mathrm{src}},\, t_i)\big)\,, \tag{14}$$

which retains the spatially slowly varying component of $\boldsymbol{v}_{\mathrm{sem}}$. This component is used only as a complementary compensation for globally coherent changes, such as overall color distribution, material appearance, and coarse spatial organization, which are often difficult to initiate during the high-noise phase. Importantly, we do not assume that all semantic attributes are low-frequency, nor do we low-pass the entire editing update. The original geometric signal $\Delta \boldsymbol{v}_{t_i}$ remains

active in Eq. (15), allowing high-frequency structures, textures, and fine-grained attributes to still be handled by the standard editing dynamics.

The injection is modulated by the time-dependent weight $w(t_i) = \lambda\, t_i^2$, where $\lambda \geq 0$ controls the overall strength of the injected signal across timesteps. This quadratic schedule satisfies four desirable properties: at $t_i \approx 1$ the weight approximates λ, maximally activating the compensation precisely where $\Delta \boldsymbol{v}_t$ is least informative; at $t_i \approx 0$ the weight vanishes, reducing the ODE to the original FlowEdit update and preserving fine-detail fidelity; the schedule is smooth and monotone, requiring no explicit threshold or switching criterion; and setting $\lambda = 0$ recovers the original FlowEdit ODE exactly for all t_i. The complete modified update is

$$\boldsymbol{x}_{t_{i+1}} = \boldsymbol{x}_{t_i} + (t_{i+1} - t_i)\left(\underbrace{\Delta \boldsymbol{v}_{t_i}}_{\text{geometric signal}} + \underbrace{\lambda\, t_i^2 \cdot \mathcal{F}_{\text{low}}\big(\boldsymbol{v}_{\text{sem}}(\boldsymbol{x}_{t_i}^{\text{src}},\, t_i)\big)}_{\text{global prompt compensation}} \right). \quad (15)$$

The geometric signal $\Delta \boldsymbol{v}_{t_i}$ tracks the evolving discrepancy between the coupled noisy latents as the edited trajectory diverges from $\boldsymbol{x}^{\text{src}}$, while the low-frequency auxiliary term provides a complementary global prompt-conditioning bias, especially in the high-noise regime where global attributes are difficult to establish.

For editing tasks confined to a specific spatial region, an optional binary mask M may be supplied. The masked combined editing signal $\widetilde{\Delta \boldsymbol{v}}_{t_i}$ is

$$\widetilde{\Delta \boldsymbol{v}}_{t_i} = M \odot \Big(\Delta \boldsymbol{v}_{t_i} + \lambda\, t_i^2\, \boldsymbol{v}_{\text{low}} \Big), \quad (16)$$

where $\odot$ denotes element-wise multiplication in the spatial domain.

4.4 Frequency-Domain Analysis

To examine the frequency-selective behavior of the proposed compensation, we analyze the spectral characteristics of the pixel-domain editing signal. Given a source image $\boldsymbol{x}^{\text{src}}$, we compute the edit delta $\delta = \boldsymbol{x}_{t_N} - \boldsymbol{x}^{\text{src}}$ for both FlowEdit and our method. Each delta map is converted to grayscale, transformed using a 2D discrete Fourier transform, and radially averaged to obtain the power spectral density (PSD) as a function of spatial frequency k. This analysis is intended to characterize how the compensation redistributes edit energy across frequency bands, rather than to serve as a standalone proof of semantic correctness. As shown in Fig. 4, our method consistently produces higher spectral energy than the baseline in the low-frequency range, while remaining close to the baseline in the high-frequency regime. The frequency-wise gain curve further confirms this trend: the average gain is larger for $k < 25$, whereas only moderate amplification is observed at higher frequencies. These results support the intended behavior of the wavelet-guided compensation: it strengthens spatially coherent global changes without suppressing the high-frequency pathways. Therefore, the frequency-domain analysis should be interpreted as evidence for selective global compensation, complementing the visual and quantitative editing results.

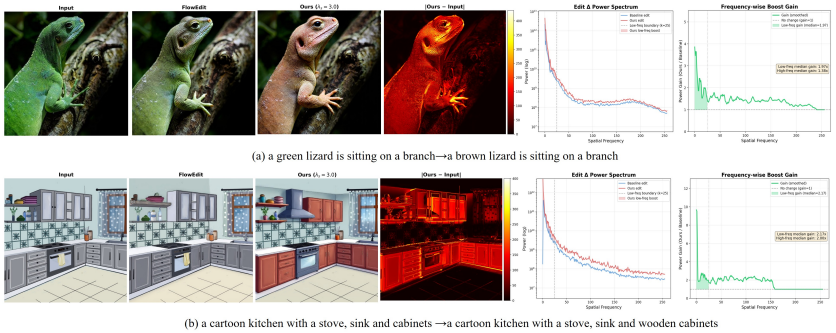

Fig. 4: Frequency-domain comparison of the editing delta. Left: qualitative editing examples. Middle: radially averaged power spectral density (PSD) of the pixel-domain edit delta for FlowEdit (blue) and ours (red). The gray dashed line indicates the low-frequency boundary ($k = 25$). Right: frequency-wise power gain (Ours/Baseline).

4.5 Pipeline of the Algorithm

Algorithm 1 presents the complete procedure.

Algorithm 1 Wavelet-Guided Semantic Signal Compensation Framework

Require: Pre-trained rectified flow model $\boldsymbol{v}_{\boldsymbol{\theta}}$, source latent $\boldsymbol{x}^{\mathrm{src}}$, source prompt embedding $\boldsymbol{e}_{\mathrm{src}}$, target prompt embedding $\boldsymbol{e}_{\mathrm{tar}}$, timestep schedule $\{t_i\}_{i=0}^{N}$, mask M, editing window $[n_{\min}, n_{\max}]$, semantic strength $\lambda \geq 0$, wavelet levels L, the operator $\mathcal{F}_{\mathrm{low}}$

Ensure: Edited latent $\boldsymbol{x}_{t_N}$

Initialize $\boldsymbol{x}_{t_0} \leftarrow \boldsymbol{x}^{\mathrm{src}}$
for $i = N - n_{\max}, \ldots, N-1$ **do**
  if $n_{\min} \leq N - i \leq n_{\max}$ **then** ▷ Editing window
    $\boldsymbol{x}^{\mathrm{src}}_{t_i} \leftarrow (1-t_i)\,\boldsymbol{x}^{\mathrm{src}} + t_i\,\boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
    $\boldsymbol{x}^{\mathrm{tar}}_{t_i} \leftarrow \boldsymbol{x}_{t_i} + \boldsymbol{x}^{\mathrm{src}}_{t_i} - \boldsymbol{x}^{\mathrm{src}}$
    $\Delta\boldsymbol{v}_{t_i} \leftarrow \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}^{\mathrm{tar}}_{t_i}, t_i; \boldsymbol{e}_{\mathrm{tar}}) - \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}^{\mathrm{src}}_{t_i}, t_i; \boldsymbol{e}_{\mathrm{src}})$
    $\boldsymbol{v}_{\mathrm{sem}} \leftarrow \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}^{\mathrm{src}}_{t_i}, t_i; \boldsymbol{e}_{\mathrm{tar}}) - \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}^{\mathrm{src}}_{t_i}, t_i; \boldsymbol{e}_{\mathrm{src}})$
    $\boldsymbol{v}_{\mathrm{low}} \leftarrow \mathcal{F}_{\mathrm{low}}(\boldsymbol{v}_{\mathrm{sem}})$ ▷ L-level Haar extraction, as in Eq. (11)
    $\widetilde{\Delta\boldsymbol{v}}_{t_i} \leftarrow \Delta\boldsymbol{v}_{t_i} + \lambda \cdot t_i^2 \cdot \boldsymbol{v}_{\mathrm{low}}$
    $\widetilde{\Delta\boldsymbol{v}}_{t_i} \leftarrow M \odot \widetilde{\Delta\boldsymbol{v}}_{t_i}$
    $\boldsymbol{x}_{t_{i+1}} \leftarrow \boldsymbol{x}_{t_i} + (t_{i+1} - t_i)\,\widetilde{\Delta\boldsymbol{v}}_{t_i}$
  else if $N - i < n_{\min}$ **then** ▷ Standard denoising
    $\boldsymbol{x}_{t_{i+1}} \leftarrow \boldsymbol{x}_{t_i} + (t_{i+1} - t_i)\,\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_i}, t_i; \boldsymbol{e}_{\mathrm{tar}})$
  end if
end for
return $\boldsymbol{x}_{t_N}$

5 Experiments

In this section, we evaluate the performance of our proposed method against inversion-based and inversion-free baselines. We conduct comprehensive experiments to demonstrate the effectiveness of our approach in terms of editing quality, structure preservation, and background consistency.

5.1 Setup

Our method is implemented on FLUX [20], Stable Diffusion 3 (SD3) [8], and SD3.5. We report the main quantitative and qualitative results on FLUX, with additional SD3 and SD3.5 results in the supplementary material.

Dataset and Baselines. For evaluation, we utilize PIE-Bench [16], consisting of 700 source images and target prompts organized into ten diverse editing scenarios. Additional high-resolution evaluation on EditBench [22] is provided in Supplementary Sec. D. We compare our method against a comprehensive set of baselines. For diffusion-based methods, we compare with PnP [36] and MasaCtrl [4]. Additionally, we evaluate FreeDiff [40], which employs progressive frequency truncation to preserve structural consistency by utilizing low-frequency components from the source image during the editing process. For the inversion-based method, we compare with RF-Inv [32], StableFlow [1], RF-Edit [38], FireFlow [7] and DNA-Edit [41]. For inversion-free methods, we compare with FlowEdit [19] in the main FLUX evaluation, and provide additional comparisons with FlowAlign [18] and DVRF [2] in Supplementary Sec. B. Comparisons with recent frequency-aware methods, including FSI-Edit [43] and FIA-Edit [44], are reported in Supplementary Sec. C. Runtime analysis is provided in Supplementary Sec. F.

Metrics. We report three categories of metrics: structural consistency, background preservation, and text–image consistency. Structure distance abbreviated as Struct. in Tab. 1, measures structural similarity via the cosine similarity between DINO [5] feature self-similarity matrices. Background preservation evaluates non-edited regions using peak signal-to-noise ratio (PSNR) [15], structural similarity index measure (SSIM) [39], mean squared error (MSE), and learned perceptual image patch similarity (LPIPS) [46], computed outside the annotated editing masks. For text–image consistency, we compute CLIP [31] similarity on both the full image and the edited region to assess global alignment and local editing accuracy.

Implementation Details. We follow the FlowEdit [19] setup for the sampling configuration. Specifically, we adopt $N = 28$ sampling steps with the editing window defined as $n_{\min} = 0$ and $n_{\max} = 24$ for FLUX. We set the semantic strength parameter $\lambda = 3.0$ and the wavelet level to $L = 3$ for FLUX.

Table 1: Quantitative comparison on the PIE-Bench. Comparison across all methods without grouping by backbone. The best and second-best results are shown in **bold** and underline, respectively.

Method	Model	Struct.$\times 10^3$ ↓	Background Preservation				CLIP Similarity	
			PSNR ↑	LPIPS$\times 10^3$ ↓	MSE$\times 10^4$ ↓	SSIM$\times 10^2$ ↑	Whole ↑	Edited ↑
PnP [36]	Diffusion	23.47	22.48	105.70	79.83	80.22	<u>25.44</u>	22.60
MasaCtrl [4]	Diffusion	23.68	22.66	87.54	80.66	81.89	24.37	21.36
FreeDiff [40]	Diffusion	<u>17.98</u>	24.78	89.02	54.81	81.92	25.16	22.15
RF-Inv [32]	FLUX	64.61	17.97	242.00	221.24	64.92	25.29	22.91
StableFlow [1]	FLUX	19.25	23.04	<u>77.09</u>	84.78	87.22	24.06	21.18
RF-Edit [38]	FLUX	24.45	24.41	113.44	56.46	83.84	25.03	22.28
FireFlow [7]	FLUX	27.27	23.08	128.65	71.11	81.25	25.34	<u>22.92</u>
FlowEdit [19]	FLUX	27.61	22.23	111.32	90.98	83.64	25.39	22.62
DNA-Edit [41]	FLUX	**16.83**	<u>25.21</u>	86.48	<u>48.11</u>	<u>87.23</u>	24.82	22.12
Ours	FLUX	30.95	**26.17**	**54.02**	**39.24**	**90.60**	**25.52**	**23.05**

5.2 Quantitative Results

As shown in Tab. 1, our method achieves the best CLIP scores while also leading across background preservation metrics. Although DNA-Edit obtains a strong structure distance, its lower CLIP scores indicate weaker semantic editing. In contrast, our approach provides stronger target alignment while maintaining higher PSNR, lower LPIPS/MSE, and higher SSIM, showing that wavelet-guided compensation better balances edit strength and background fidelity in inversion-free editing.

5.3 Qualitative Results

Fig. 5 presents visual comparisons across a range of editing tasks, including both local attribute manipulation and more complex structural modifications. As illustrated in the figure, our method achieves precise semantic modification while preserving the identity and overall structure of the original image. For example, in the open-mouth panda case, the mouth region is successfully modified to reflect the target instruction, while the identity, pose, and surrounding environment of panda remain largely unchanged. Similar behavior can be observed in other examples, where the target attributes are effectively realized while preserving the main content of the original image. These results demonstrate that our approach effectively balances semantic editing strength and structural consistency across diverse editing scenarios.

5.4 Sensitivity and Design Choices

Our method introduces two design parameters: the wavelet level L and semantic strength λ. As summarized in Tab. 2, increasing L generally favors structural fidelity and background preservation, while λ controls the preservation–editability trade-off. We use $L = 3$ and $\lambda = 3.0$ as the default FLUX setting. Tab. 3 shows

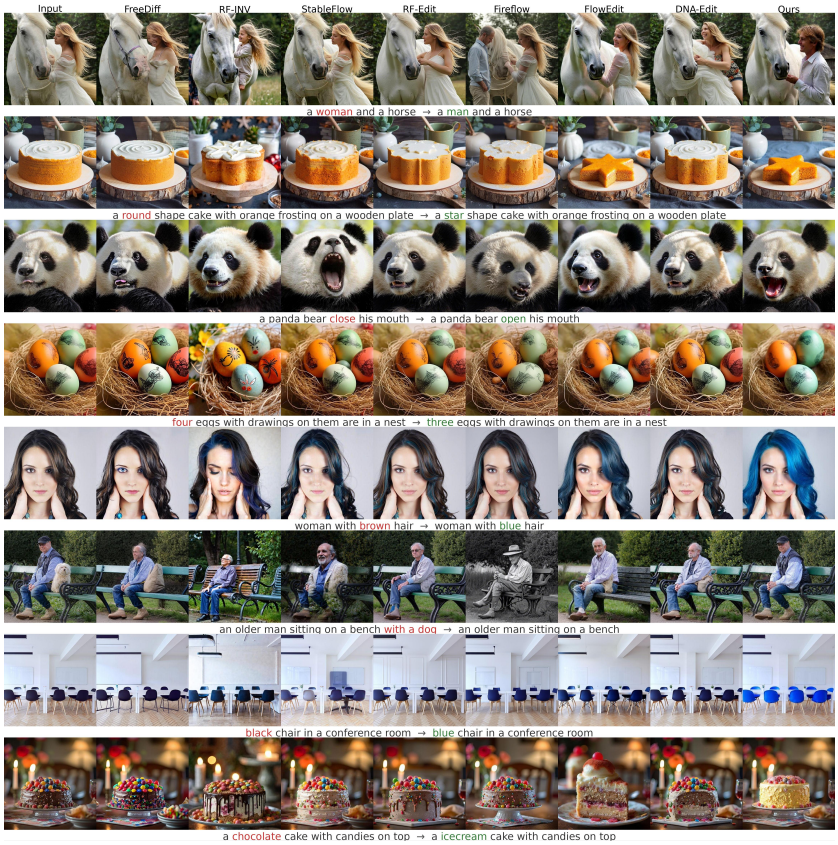

Fig. 5: Qualitative comparison on the PIE-Bench.

that low-frequency compensation performs better than full-spectrum or high-frequency-only injection. Complete grids for λ and L are provided in Supplementary Secs. B.3 and E.1, with low-pass operator and schedule ablations in Supplementary Secs. E.2 and E.3.

5.5 User Study

We further assess perceptual quality through a pairwise user study. For each baseline comparison, we randomly select 100 edited image pairs generated by our method and one of seven baselines using the same source image and edit prompt. Ten participants choose the result that better preserves unedited regions while reflecting the target edit. As shown in Tab. 4, our method is preferred over all baselines, complementing the quantitative results.

Table 2: Ablation study of the wavelet decomposition level L and the semantic strength parameter λ on PIE-Bench under FLUX. We report one-factor sensitivity slices for L and λ. Our final choice ($L = 3$, $\lambda = 3.0$) is highlighted in gray. The complete hyperparameter grid is provided in the supplementary material.

Setting	L	λ	Struct.$\times 10^3$ ↓	Background Preservation				CLIP Similarity	
				PSNR ↑	LPIPS$\times 10^3$ ↓	MSE$\times 10^4$ ↓	SSIM$\times 10^2$ ↑	Whole ↑	Edited ↑
L sweep ($\lambda = 3.0$)	1	3.0	35.31	25.26	58.48	46.12	89.95	25.56	23.01
	2	3.0	34.21	25.51	57.36	44.14	90.16	25.59	23.01
	3	3.0	30.95	26.17	54.02	39.24	90.60	25.52	23.05
	4	3.0	26.06	27.05	49.12	33.36	91.17	25.39	22.90
λ sweep ($L = 3$)	3	2.0	25.72	26.88	49.64	34.09	91.16	25.48	22.98
	3	2.5	28.30	26.52	51.83	36.59	90.88	25.54	23.05
	3	3.0	30.95	26.17	54.02	39.24	90.60	25.52	23.05
	3	3.5	33.67	25.80	56.24	41.99	90.33	25.55	23.09
	3	4.0	36.65	25.44	58.68	45.00	90.04	25.44	22.94

Table 3: Frequency-band selection on PIE-Bench under FLUX.

Method	Struct.$\times 10^3$ ↓	Background Preservation				CLIP Similarity	
		PSNR ↑	LPIPS$\times 10^3$ ↓	MSE$\times 10^4$ ↓	SSIM$\times 10^2$ ↑	Whole ↑	Edited ↑
Zero (no comp.)	27.61	22.23	111.32	90.98	83.64	25.39	22.62
Identity (full)	41.53	23.61	69.77	63.31	88.05	25.03	22.24
Highpass only	33.20	24.81	59.69	50.37	89.25	24.97	22.34
Haar wavelet	30.95	26.17	54.02	39.24	90.60	25.52	23.05

Table 4: User preference rate (%).

Method	Ours	Base	Method	Ours	Base
Fireflow	71	29	FreeDiff	58	42
StableFlow	68	32	DNA-Edit	58	42
FlowEdit	66	34	RF-Edit	58	42
RF-Inv	64	36			

6 Conclusion

In this paper, we investigate inversion-free image editing from the perspective of trajectory dynamics in rectified flow models. Through empirical and analytical observations, we identified that under certain global attribute shifts, the semantic editing signal may become weak in the early timesteps due to strong geometric coupling along the source trajectory. Motivated by this analysis, we introduced a time-dependent compensation mechanism that enhances semantic deviation when the trajectory is insufficiently responsive to the target prompt, while remaining compatible with the original formulation. The proposed strategy improves editing strength while maintaining background fidelity. Extensive experiments demonstrate that our method achieves more faithful semantic modifications and stronger perceptual preference across diverse editing scenarios.

References

1. Avrahami, O., Patashnik, O., Fried, O., Nemchinov, E., Aberman, K., Lischinski, D., Cohen-Or, D.: Stable Flow: Vital layers for training-free image editing. In: CVPR (2025)
2. Beaudouin, G., Li, M., Kim, J., Yoon, S., Wang, M.: Delta velocity rectified flow for text-to-image editing. arXiv preprint arXiv:2509.05342 (2025)
3. Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., Passos, A.: LEDITS++: Limitless image editing using text-to-image models. In: CVPR (2024)
4. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: ICCV (2023)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
6. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex fourier series. Mathematics of Computation (1965)
7. Deng, Y., He, X., Mei, C., Wang, P., Tang, F.: FireFlow: Fast inversion of rectified flow for image semantic editing. In: ICML (2025)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
9. Feng, K., Ma, Y., Wang, B., Qi, C., Chen, H., Chen, Q., Wang, Z.: DiT4Edit: Diffusion transformer for image editing. In: AAAI (2025)
10. Garibi, D., Patashnik, O., Voynov, A., Averbuch-Elor, H., Cohen-Or, D.: ReNoise: Real image inversion through iterative noising. In: ECCV (2024)
11. Han, L., Wen, S., Chen, Q., Zhang, Z., Song, K., Ren, M., Gao, R., Stathopoulos, A., He, X., Chen, Y., et al.: ProxEdit: Improving tuning-free real image editing with proximal guidance. In: WACV (2024)
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-Prompt image editing with cross attention control. In: ICLR (2023)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
14. Hong, S., Lee, K., Jeon, S.Y., Bae, H., Chun, S.Y.: On exact inversion of DPM-Solvers. In: CVPR (2024)
15. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of PSNR in image/video quality assessment. Electronics Letters (2008)
16. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: PnP Inversion: Boosting diffusion-based editing with 3 lines of code. In: ICLR (2023)
17. Kim, G., Kwon, T., Ye, J.C.: DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In: CVPR (2022)
18. Kim, J., Hong, Y., Park, J., Ye, J.C.: FlowAlign: Trajectory-regularized, inversion-free flow-based image editing. arXiv preprint arXiv:2505.23145 (2025)
19. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: FlowEdit: Inversion-free text-based editing using pre-trained flow models. In: ICCV (2025)
20. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
21. Li, Z., Song, Y., Peng, J., Liu, T., Huang, J., Qu, X., Liu, L., Wang, W., Zhao, Y., Wei, Y.: On exact editing of flow-based diffusion models. arXiv preprint arXiv:2512.24015 (2025)

22. Lin, H., Chen, Y., Wang, J., An, W., Wang, M., Tian, F., Liu, Y., Dai, G., Wang, J., Wang, Q.: Schedule your edit: A simple yet effective diffusion noise schedule for image editing. In: NeurIPS (2024)
23. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
24. Liu, Q.: Rectified flow: A marginal preserving approach to optimal transport. arXiv preprint arXiv:2209.14577 (2022)
25. Ma, Y., Di, D., Liu, X., Chen, X., Fan, L., Su, T., Gao, Y.: Adams Bashforth Moulton solver for inversion and editing in rectified flow. arXiv preprint arXiv:2503.16522 (2025)
26. Mallat, S.G.: A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2002)
27. Mao, J., Wang, K., Xiang, Y., Chen, K.: TweezeEdit: Consistent and efficient image editing with path regularization. arXiv preprint arXiv:2508.10498 (2025)
28. Martin, S.T., Gagneux, A., Hagemann, P., Steidl, G.: PnP-Flow: Plug-and-play image restoration with flow matching. In: ICLR (2025)
29. Miyake, D., Iohara, A., Saito, Y., Tanaka, T.: Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. In: WACV (2025)
30. Pan, Z., Gherardi, R., Xie, X., Huang, S.: Effective real image editing with accelerated iterative diffusion inversion. In: ICCV (2023)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
32. Rout, L., Chen, Y., Ruiz, N., Caramanis, C., Shakkottai, S., Chu, W.: Semantic image inversion and editing using rectified stochastic differential equations. In: ICLR (2025)
33. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
34. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. ICLR (2021)
35. Sun, J., Wang, W., Sun, M., Wang, P., Zhu, X., Liu, J.: W-EDIT: A wavelet-based frequency-aware framework for text-driven image editing. In: ICLR (2026)
36. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR (2023)
37. Wallace, B., Gokul, A., Naik, N.: EDICT: Exact diffusion inversion via coupled transformations. In: CVPR (2023)
38. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming rectified flow for inversion and editing. In: ICML (2025)
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (2004)
40. Wu, W., Fan, Q., Qin, S., Gu, H., Zhao, R., Chan, A.B.: FreeDiff: Progressive frequency truncation for image editing with diffusion models. In: ECCV (2024)
41. Xie, C., Li, M., Li, S., Wu, Y., Yi, Q., Zhang, L.: DNAEdit: Direct noise alignment for text-guided rectified flow editing. In: NeurIPS (2025)
42. Xu, S., Huang, Y., Pan, J., Ma, Z., Chai, J.: Inversion-free image editing with language-guided diffusion models. In: CVPR (2024)
43. Yang, K., Li, X., Li, Y., Li, Q., Wang, Z.: FSI-Edit: Frequency and stochasticity injection for flexible diffusion-based image editing. In: NeurIPS (2025)
44. Yang, K., Shen, B., Li, X., Dai, Y., Luo, Y., Ma, Y., Fang, W., Li, Q., Wang, Z.: FIA-Edit: Frequency-interactive attention for efficient and high-fidelity inversion-free text-guided image editing. In: AAAI (2026)

45. Zhang, G., Lewis, J.P., Kleijn, W.B.: Exact diffusion inversion via bidirectional integration approximation. In: ECCV (2024)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
47. Zhao, J., Zheng, H., Wang, C., Lan, L., Huang, W., Yang, W.: Null-text guidance in diffusion models is secretly a cartoon-style creator. In: ACM MM (2023)