Faithful Image Editing via Degraded Representations

Anonymous authors
Paper under double-blind review

Abstract

Rectified flow and diffusion-based models currently represent the state-of-the-art in image editing, leveraging powerful pre-trained generative priors to produce visually compelling modifications. Despite their impressive capabilities, maintaining faithfulness to the source image – preserving structure and photometric characteristics while satisfying a target prompt – remains a persistent challenge in this domain. Direct traversal between source and target distributions in rectified flow frameworks offers a promising direction for improving fidelity. However, identifying trajectories that are both semantically effective and strictly structure-preserving remains an open problem. In this work, we propose an optimization- and inversion-free image editing framework that is, in principle, agnostic to the underlying generative backbone. Our central insight is to operate within a carefully designed degraded representation space that constrains editing trajectories and suppresses unintended collateral modifications to the target. We first establish the existence of such degraded representations for generative-prior-based editing and then develop a principled method to project editing trajectories onto this space. The resulting method, Editing via Degraded Representations (EDR), systematically eliminates unfaithful trajectory deviations while preserving the flexibility required to satisfy the target text prompt. Extensive quantitative and qualitative evaluations demonstrate that EDR achieves precise, high-quality edits with superior fidelity, establishing a new state-of-the-art in faithful image editing. *Code will be released upon acceptance.*

1 Introduction

Diffusion models Sohl-Dickstein et al. (2015); Ho et al. (2020) and Rectified Flow (RF) models Liu et al. (2023); Esser et al. (2024) have revolutionized image synthesis. Leveraging these powerful generative priors has consequently emerged as a dominant paradigm for image editing Kulikov et al. (2025); Brack et al. (2024); Mokady et al. (2023); Rout et al. (2025), enabling flexible and photorealistic manipulations grounded in the strong structural representations learned by diffusion and RF models Rombach et al. (2022); Esser et al. (2024).

Despite their impressive generative capabilities, existing approaches often struggle to produce faithful edits, particularly when operating on real images Garibi et al. (2024). Faithful editing requires preserving the global structure and color distribution of the source image while restricting modifications strictly to those necessary for satisfying the target prompt. A primary line of research Kawar et al. (2023); Valevski et al. (2023); Zhang et al. (2023) addresses this challenge through test-time optimization, adapting the generative model to the source image prior for editing. Although this strategy achieves high fidelity, it incurs substantial computational overhead due to per-image optimization, limiting its practicality for large-scale or real-time applications.

Addressing that, a second line of work explores optimization-free approaches Meng et al. (2021); Mokady et al. (2023); Tumanyan et al. (2023). Early methods along this line injected random noise into the source image followed by denoising for editing. This frequently resulted in significant drift and an unstable fidelity–editability trade-off Meng et al. (2021). To reduce this drift, inversion techniques were introduced to recover the latent noise that reconstructs the source image via a deterministic ordinary differential equation (ODE) trajectory. While standard DDIM inversion Song et al. (2020) avoids explicit optimization, numerical

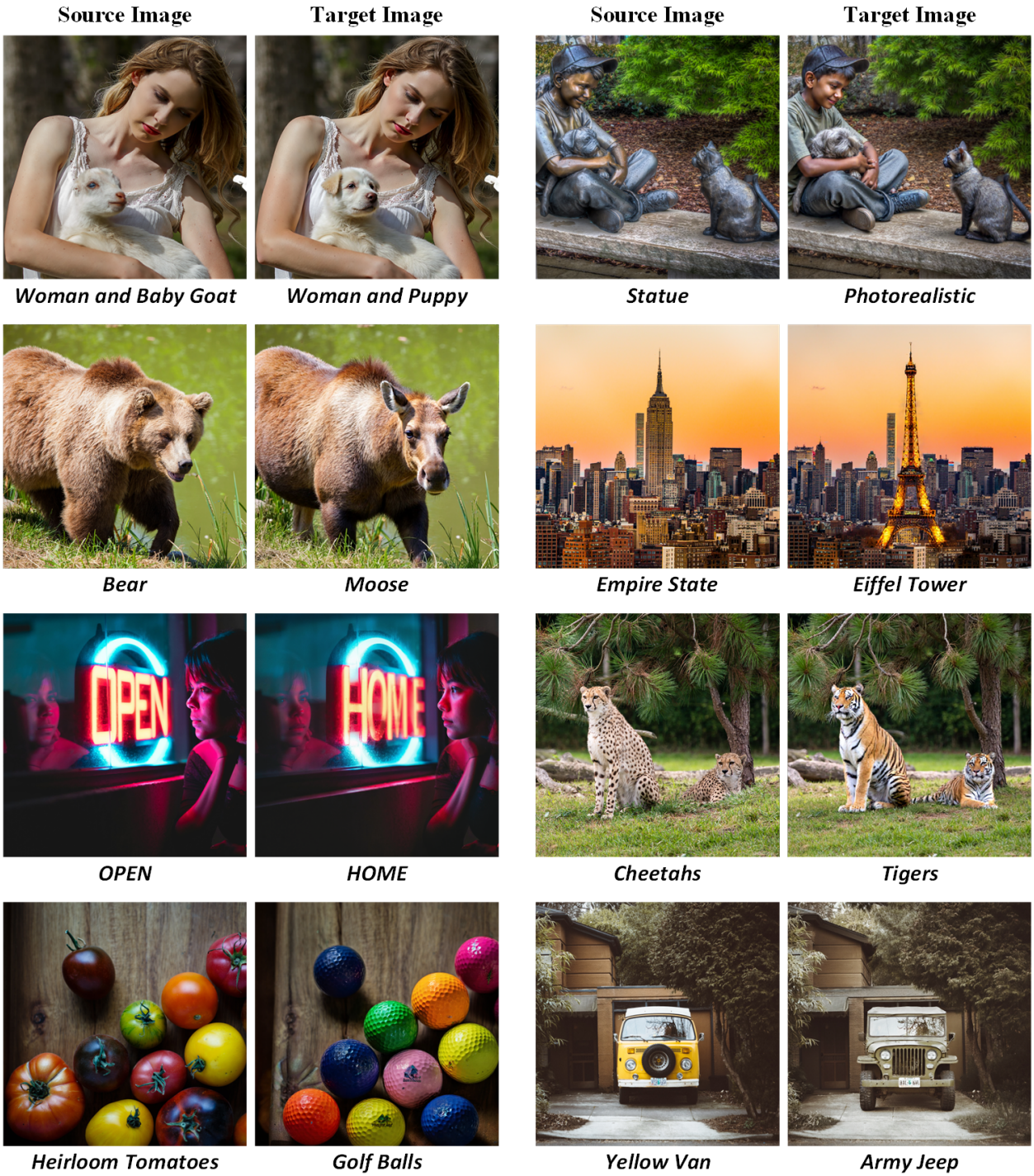

Figure 1: Editing samples using the proposed method on FLUX.1-dev (top two rows) and SD3 (bottom two rows). Our approach projects the editing trajectories onto degraded representations that are invariant to insignificant intensity and structural variations. The resulting edits effectively prevent unintended side effects. Note that no explicit or implicit masks are used in our method.

discretization errors accumulate, often leading to structural inconsistencies during editing. Subsequent works proposed mathematically exact inversion schemes Wallace et al. (2023); Pan et al. (2023) or straight-line RF trajectories Lu et al. (2023); Esser et al. (2024) to improve consistency. However, the exact inverted noise corresponding to the source image may conflict with the target prompt, resulting in weak or unnatural edits Huberman-Spiegelglas et al. (2024). More recently, FlowEdit Kulikov et al. (2025) introduced an inversion-free framework that infers a direct trajectory between the source and target distributions. Although this formulation reduces transition cost and improves fidelity relative to inversion-based methods, it does not guarantee that the inferred trajectory strictly preserves the desirable content. Moreover, to stabilize the editing path, it requires aggregating multiple model predictions at each time step, increasing inference complexity.

In this work, we investigate the editing via generative prior from a trajectory-centric perspective to enable effective yet faithful image editing (see Figure 1). Editing generative trajectories in latent space can traverse multiple directions while still satisfying a target prompt. While some of these directions correspond to semantically meaningful transformations, others introduce unintended variations arising from stochasticity or entangled latent representations shaped by spurious correlations. We show that such trajectories can be systematically constrained by projecting them onto a degraded representation where selected characteristics are suppressed or rendered invariant. We illustrate our key intuition in Figure 2. Consider the representations of an image (red square) and one of its edited versions (blue circle). When mapped to degraded forms—such as edge-based or blurred representations—specific attributes are removed: edge representations discard intensity information, whereas blurred representations suppress fine structural details. In this degraded space, trajectories contain none or minimal components along the directions corresponding to the suppressed characteristics. Consequently, projecting editing trajectories onto degraded trajectories enables effective removal of latent components associated with these characteristics, thereby eliminating unintended deviations.

Building upon this insight, we propose Editing via Degraded Representations (EDR), a method that approximates the projection of editing trajectories onto a tailored degraded space. Our degraded representation combines Gaussian structural smoothing with dynamic range reduction, inducing invariance to subtle structural perturbations as well as incidental intensity and color shifts that frequently arise during editing. By constraining trajectories within this space, EDR enforces high-fidelity edits that preserve the foundational structure and color statistics of the source image. At the same time, the degraded representation remains sensitive to substantial semantic modifications when required by the target prompt, ensuring editability is not unduly restricted.

Our contributions are summarized as follows:

- We introduce a principled mechanism for constraining generative editing trajectories via projection onto *degraded representations*, thereby suppressing entangled and spurious latent directions.
- We propose Editing via Degraded Representations (EDR) method, which maps editing trajectories to a custom degraded space combining *Gaussian spatial smoothing* and *dynamic range reduction*, achieving invariance to unintended structural and intensity shifts, respectively.
- We conduct extensive quantitative and qualitative evaluations demonstrating that EDR achieves highly faithful edits and outperforms state-of-the-art methods maintaining exceptional source fidelity.

2 Related Work

Our work relates to diffusion- and flow-based image editing, and is particularly relevant to controlled editing with rectified flow. Below, we discuss this relevance, contextualizing our contribution with respect to recent advances.

Rectified Flow Models. ODE-based generative formulations are increasingly replacing stochastic diffusion processes as the dominant paradigm in generative modeling. Rectified Flow (RF) Liu et al. (2023) introduced a deterministic transport perspective, learning straightened trajectories between source and data

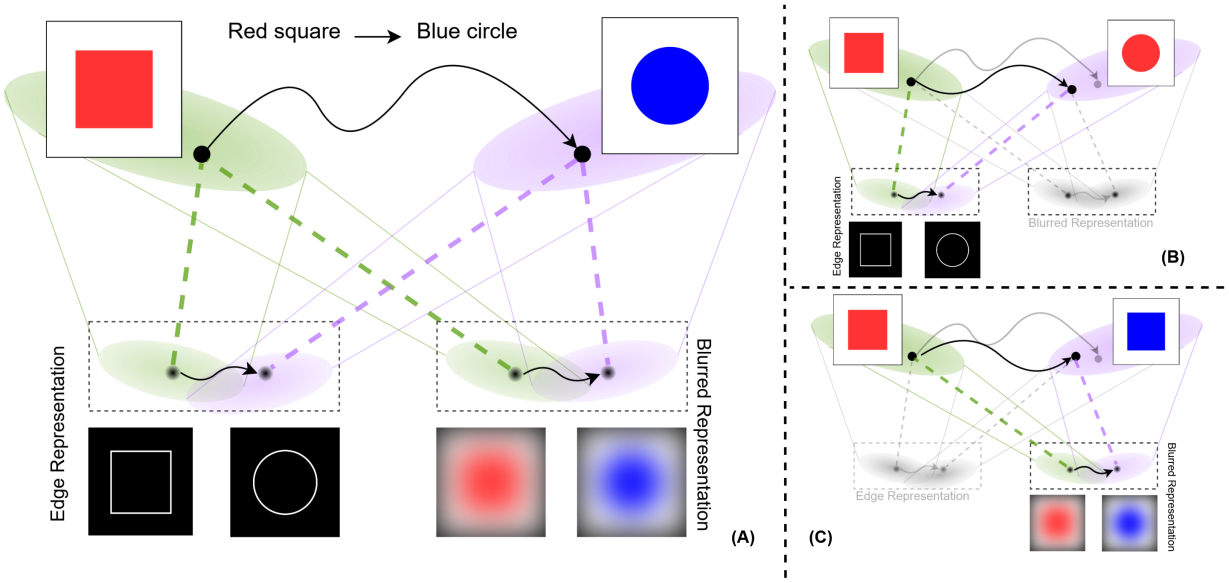

Figure 2: Illustration of our key insight for faithful image editing via degraded representation. **(A)** A source image (red square) and its target edited version (blue circle) get mapped onto their degraded (edge and blurred) representations. The edge representation is inherently invariant to color changes, while the blurred representation is largely insensitive to structural changes. **(B)** The editing trajectory gets projected onto the edge-based degraded representation, which ignores color variations. This prevents the editing process from altering the colors of the source image. **(C)** The editing trajectory is projected onto the blurred degraded representation. Since this representation is largely insensitive to structural information, the resulting trajectory does not modify the structure, thereby the resulting output retains the structure of the source image.

distributions. Related frameworks, including Flow Matching Lipman et al. (2022) and Stochastic Interpolants Albergo et al. (2025), similarly focus on directly learning vector fields that define probability transport paths. Compared to conventional diffusion models Sohl-Dickstein et al. (2015); Song & Ermon (2019); Ho et al. (2020), these approaches offer improved sampling efficiency and more stable training dynamics. However, while rectified flows provide cleaner generative trajectories, they do not inherently address the problem of trajectory control during editing; namely, how to restrict the path to preserve specific structural and photometric properties of a given image. In the context of rectified flow models, our work directly addresses this gap by introducing a mechanism for constraining editing the trajectories.

Diffusion- and Flow-Based Image Editing. Diffusion-based editing leverages powerful pre-trained generative priors to enable realistic and semantically aligned modifications Rombach et al. (2022). Existing methods largely fall into three categories: training-based methods, test-time optimization, and optimization-free strategies.

Training-based methods Brooks et al. (2023); Wasserman et al. (2025) learn dedicated editing networks that produce fast and faithful modifications. While effective, they rely on synthetic supervision and incur substantial training costs. Even reduced-resource alternatives Zhang et al. (2025) remain dependent on additional learning procedures. In contrast, our method requires neither retraining nor synthetic supervision, operating directly with pre-trained models. *Test-time optimization approaches* Kawar et al. (2023); Valevski et al. (2023) adapt the generative model to each source image to enhance faithfulness. Although this yields high-quality edits, it introduces significant inference latency due to per-image fine-tuning. Our method avoids per-instance optimization entirely while maintaining high structural and photometric fidelity. *Optimization-free approaches* attempt to balance efficiency and faithfulness but often lack principled trajectory control. Attention-injection methods Tumanyan et al. (2023); Hertz et al. (2022); Alaluf et al. (2024); Patashnik

et al. (2023); Hertz et al. (2024) guide generation through semantic feature manipulation, yet they do not explicitly regulate latent transport directions, leaving edits susceptible to unintended structural or color drift. Mask-based strategies Couairon et al. (2022); Avrahami et al. (2023) spatially restrict edits but do not constrain global distributional shifts. Inversion-based techniques Song et al. (2020); Huberman-Spiegelglas et al. (2024); Garibi et al. (2024); Mokady et al. (2023); Wallace et al. (2023) recover the latent noise corresponding to the source image to reduce drift. However, exact inversion often conflicts with target prompts, resulting in weak edits, while approximate inversion accumulates discretization errors that degrade structure.

Recent flow-based editors such as RF Edit Wang et al. (2024) and FlowEdit Kulikov et al. (2025) advance toward Flow Matching formulations. FlowEdit Kulikov et al. (2025) removes explicit inversion by directly coupling source and target vector fields using the ODE structure of flow models Liu et al. (2023). This significantly improves structural fidelity relative to diffusion-based inversion methods. Nevertheless, stabilizing the learned trajectory requires aggregating multiple model predictions at each time step, increasing computational cost. More fundamentally, these methods aim to find a suitable trajectory but do not explicitly suppress undesirable latent directions during transport.

In contrast, our approach introduces an orthogonal perspective: rather than refining inversion, modifying attention, or averaging vector fields, we constrain editing trajectories through projection onto a carefully designed degraded representation. This mechanism explicitly removes components associated with unintended structural and photometric variations, enabling faithful edits without retraining, per-image optimization, or multi-step aggregation.

Rectified Flows in Large-Scale Generative Modeling. Recent large-scale text-to-image systems, including Stable Diffusion 3 (SD3) Esser et al. (2024) and FLUX Labs et al. (2025), adopt rectified flow formulations, demonstrating superior sample quality, improved prompt alignment, and scalable training dynamics. Beyond image synthesis, RF-based methods have been extended to unsupervised domain translation Wang et al. (2024), video generation Davtyan et al. (2023); Ifriqi et al. (2025); Jin et al. (2024), and inverse problem solving Zhu et al. (2024). While these works establish rectified flows as a powerful generative backbone, they do not address controlled editing under strict fidelity constraints. Our work builds upon the strengths of rectified flows while introducing an explicit trajectory-constraint mechanism tailored to faithful image editing.

3 Methodology

Image editing with diffusion and RF models enables flexible transformation between source and target concepts by navigating learned generative trajectories. However, these editing trajectories frequently deviate along unintended latent directions, compromising fidelity to the source image. In Section 3.1, we provide a theoretical analysis demonstrating how such arbitrary deviations can be systematically constrained by regulating the trajectory geometry in latent space. Building upon this foundation, Section 3.2 introduces a custom degraded representation and presents our method to enforce trajectory constraints using this representation to achieve faithful, structure-preserving image edits.

3.1 Constraining Editing Trajectories

Consider RF models as the backbone of the editing process, where generative modeling is formulated as learning a transport map between two probability distributions, π_0 and π_1. In conventional generative settings, π_0 denotes the data distribution over images, while π_1 is a standard Gaussian prior $\mathcal{N}(0, I)$. The RF models parameterize a time-dependent velocity field that induces transporting trajectories between these distributions. The transport is governed by an ODE:

$$dZ_t = v_\theta(Z_t, t)\, dt, \quad Z_0 \sim \pi_0,\ Z_1 \sim \pi_1, \tag{1}$$

where Z_t denotes the state at time $t \in [0, 1]$ and $v_\theta : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ is a time-dependent velocity field driving the flow, parametrized by a neural network. RF assumes a linear interpolation trajectory (straight

line) between a data sample Z_0 and a noise sample Z_1:

$$Z_t = tZ_1 + (1-t)Z_0. \tag{2}$$

Consequently, the ideal velocity field v that generates this trajectory is constant along the path: $v(Z_t, t) = Z_1 - Z_0$. A neural network $v_\theta(Z_t, t, c)$ is trained to approximate this velocity field, based on condition c. To edit a real source image X_{src}, we define two text-conditioned velocity fields: $V^{\text{src}}(Z_t, t)v_\theta(Z_t, t, c_{\text{src}})$ corresponding to the source description, and $V^{\text{tar}}(Z_t, t)v_\theta(Z_t, t, c_{\text{tar}})$, the target field corresponding to the desired edit. The editing process typically begins by extracting the initial noise map corresponding to the source image, traversing the forward process defined by the ODE:

$$dZ_t^{\text{src}} = V^{\text{src}}(Z_t^{\text{src}}, t)\, dt, \tag{3}$$

where the integration starts at $t = 0$ from the source image $Z_0^{\text{src}} = X_{\text{src}}$ and proceeds to $t = 1$ to reach the noise map Z_1^{src}. Subsequently, the edited image is obtained by solving the reverse ODE using the target velocity field:

$$dZ_t^{\text{tar}} = V^{\text{tar}}(Z_t^{\text{tar}}, t)\, dt. \tag{4}$$

This process is solved backward in time, starting at $t = 1$ with the extracted noise ($Z_1^{\text{tar}} = Z_1^{\text{src}}$) and integrating to $t = 0$ to reach the edited image Z_0^{tar}.

The transition from source to the target distributions via passing through Gaussian noise distribution can also be interpreted as a *direct path* Kulikov et al. (2025) between the source and target distributions. Under this interpretation, instead of transporting samples through the intermediate Gaussian distribution, one can construct a direct path that aligns the forward trajectories of the source and target. A point along this path is defined as:

$$Z_t^{\text{dir}} = Z_0^{\text{src}} + Z_t^{\text{tar}} - Z_t^{\text{src}}. \tag{5}$$

Accordingly, the editing process begins at $t = 1$ with $Z_1^{\text{dir}} = Z_0^{\text{src}}$, and progressively follows the inverse path until reaching $Z_0^{\text{dir}} = Z_0^{\text{tar}}$ at $t = 0$. The corresponding velocity field is then given by

$$V_t^{\text{dir}}(Z_t^{\text{dir}}, t) = V_t^{\text{tar}}(Z_t^{\text{tar}}, t) - V_t^{\text{src}}(Z_t^{\text{src}}, t), \tag{6}$$

which defines the direction of the editing trajectory (see Figure 3a). Importantly, the direct path interpretation extends beyond inverted noise and is equally applicable to cases where Z_1^{tar} is randomly sampled from Gaussian noise.

The trajectory induced by V_t^{dir} may potentially take multiple forms, leading to a variety of possible edits. Among these, we seek trajectories that are *constrained* to exclusively modify the intended features dictated by the target prompt. To achieve this, we propose projecting the editing trajectories onto arbitrarily degraded representations where certain image characteristics get suppressed. For instance, the edge representation of an image captures structural information while being blind to intensity and color characteristics. Conversely, an intensely blurred representation preserves overall intensity but is largely invariant to structural details. As illustrated in Figure 2, projecting the editing trajectory onto a degraded representation causes the trajectory to ignore changes to the characteristics that the representation is invariant to.

We can approximate the degraded representation of V_t^{dir} at each timestep t, denoted as $\tilde{V}_t$ as

$$\tilde{V}_t = \mathcal{D}\left(\frac{Z_t^{\text{dir}} - Z_{t-\Delta t}^{\text{dir}}}{\Delta t}\right), \tag{7}$$

where $\mathcal{D}$ is an arbitrary degradation function, and Δt is the temporal difference between each step. The projection of V_t^{dir} on $\tilde{V}_t$ is given by

$$V_t^{\text{proj}} = \frac{\langle V_t^{\text{dir}}, \tilde{V}_t \rangle}{\|\tilde{V}_t\|^2}\, \tilde{V}_t. \tag{8}$$

Following the projected velocity vector V_t^{proj} at each timestep t during the editing process, ignores changes in the directions that are undefined or suppressed in the corresponding degraded representation.

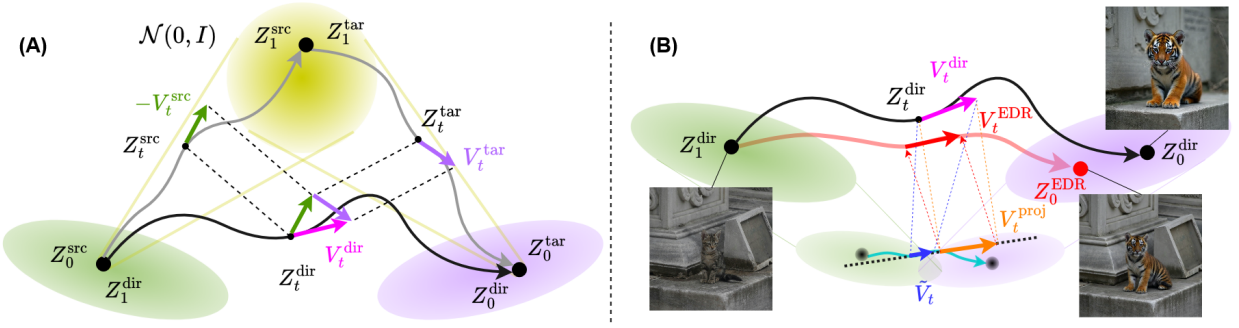

Figure 3: **Proposed EDR Method. (A)** Reinterpretation of the editing process as a direct path, where the velocity V_t^{dir} is given by the difference between the target and source velocities, $V_t^{\text{tar}} - V_t^{\text{src}}$, as if they were generated directly from the inverted or even random Gaussian noise. **(B)** Our proposed method maps the direct velocity vector V_t^{dir} to a degraded representation obtained by a combination of Gaussian spatial smoothing and dynamic range reduction. This representation is invariant to minor structural and intensity modifications, which are commonly caused by secondary editing effects. The initial update velocity vector is then projected onto this degraded counterpart to suppress these unintended modifications, resulting in more precise and focused edits.

3.2 Editing through Degraded Representations (EDR)

Here, we introduce our method to leverage the capability of restraining editing trajectories along undesired directions, ensuring that the edits remain precise and faithful to the source image. We categorize the modifications occurring during the editing process into two main types: primary modifications, which implement the intended edit and are often substantial, and secondary modifications, which constitute minor side effects. For example, consider editing an image of a blue car into a red car. In an ideal editing process, the color is successfully changed as intended (primary modification), but the car shape may inadvertently become more *sporty* due to spurious correlation between red color and sports cars, representing a secondary modification.

In principle, the primary modification represents the main focus of the model, ensuring that the target image aligns with the prompt text embedding, as opposed to secondary modifications, which constitute minor side effects that are not explicitly guided by the prompt. Hence, we need to project the editing trajectory onto a representation space that is invariant to subtle, insignificant modifications which form the secondary modifications in our context.

We characterize modifications into structural and color distribution shifts, allowing us to target each specifically through our degraded representation. To design a representation that is invariant to insignificant amount of these two modifications, we use Gaussian spatial smoothing and dynamic range reduction. Formally, for Z_t^{dir}, our proposed combined degraded representation is

$$\mathcal{D}_{\text{c}}(Z_t^{\text{dir}}) = I_{\min} + (I_{\max} - I_{\min})\left(G_\sigma * Z_t^{\text{dir}}\right), \quad 0 < I_{\min} < I_{\max} < 1, \tag{9}$$

where G_σ is the Gaussian kernel with standard deviation σ and the size equal to 6σ, while $I_{\min}$ and $I_{\max}$ define the target intensity range. In our proposed method, we use $\mathcal{D}_{\text{c}}(\cdot)$ in place of $\mathcal{D}(.)$ in Equation 7.

Projecting V_t^{dir} onto this degraded representation constrains the editing trajectory against incidental structural and intensity shifts, ensuring the transformation remains aligned with the source image's foundational characteristics. Importantly, this mechanism does not strictly confine the model to the source state; since the designed degraded representation $\mathcal{D}_c$ remains sensitive to substantial structural and color transformations, it permits significant semantic changes when necessitated by the target prompt.

While projecting the trajectory onto our proposed degraded representation effectively enforces fidelity to the source, it potentially can compromise the quality of target image. Therefore, we utilize the projected trajectory in a decaying manner, to ensure preservation of the image quality. The update vector of the

proposed method, denoted as V_t^{EDR}, is given by

$$V_t^{\text{EDR}} = \alpha V_t^{\text{proj}} + (1-\alpha)V_t^{\text{dir}}. \tag{10}$$

Here, the coefficient α scales the contribution of the projected trajectory according to the decay rate γ as

$$\alpha = (1-t)^{\gamma}. \tag{11}$$

Furthermore, the overall extent of the modification is governed by the editing strength, denoted as t_0, which dictates the starting timestep of the generative process. A larger t_0 provides the model with a longer trajectory to satisfy complex target prompts, whereas a smaller t_0 inherently restricts the generation closer to the source image.

Thus, starting from t_0, our method prioritizes the projected update vector V_t^{proj} during the initial timesteps to guide the edit along the degraded representation. This ensures structural fidelity to the source image without deviating into irrelevant directions. As the generation progresses into the later steps, the update vector V_t^{EDR} gradually shifts its weight toward the model's native directional vector V_t^{dir}, ensuring the final output is realistic and visually coherent. Figure 3b illustrates the proposed method with $\gamma = 0$ where V_t^{EDR} is updated only with regard to V_t^{proj}.

4 Experimental Evaluation

4.1 Setup

4.1.1 Experimental Setting

We evaluate our approach on two state-of-the-art RF-based text-to-image models: Stable Diffusion 3 (SD3) Esser et al. (2024) and FLUX.1-dev Labs et al. (2025). For SD3, we set the total inference steps to T=50. The editing strength (t_0) is set to 0.76 with a decay factor $\gamma = 2$. We employ a Classifier-Free Guidance (CFG) scale of 3.5 and 13.5 for source and target images, respectively. For FLUX.1-dev, we utilize $T = 28$ inference steps. The corresponding editing parameters are configured as $t_0 = 0.9$ and $\gamma = 5$. For both models, Gaussian smoothing is applied using $\sigma = 5$ and a kernel size of 30. The reduced dynamic range is bounded between $I_{min} = 0.25$ and $I_{max} = 0.75$. We chose these hyperparameters empirically (see supplementary material).

4.1.2 Comparison Methods

We evaluate our approach against the following methods: the baseline SDEdit Meng et al. (2021), which applies editing by adding random Gaussian noise followed by denoising conditioned on target prompt, ODE Inversion as discussed in Section 3.1, and state-of-the-art flow-based methods RF inversion Rout et al. (2024), RF Edit Wang et al. (2024), iRFDS Yang et al. (2024), and FlowEdit Kulikov et al. (2025).

For the SDEdit baseline, we set the editing strength to $t_0 = 0.4$ with a target Classifier-Free Guidance (CFG) scale of 13.5 on SD3. For FLUX, we utilized $t_0 = 0.75$ with a target CFG scale of 5.5. For RF Edit, we used its official implementation which is applied only on FLUX and used default hyperparameters: 25 guidance steps, 5 injection steps, and a guidance scale of 3.5. For FlowEdit, we utilized the provided codebase and the hyperparameter settings recommended in the original paper Kulikov et al. (2025). For SD3, we set the total number of steps to $T = 50$, with the editing starting timestep at $n_{\max} = 33$, and CFG scales of 3.5 and 13.5 for the source and target conditioning, respectively. For FLUX, the parameters were set to $T = 28$ steps, $n_{\max} = 24$, and CFG scales of 1.5 and 5.5 for the source and target. Also, for ODE inversion we applied the same hyperparameters.

4.1.3 Dataset

For qualitative comparisons, we curated a diverse evaluation set comprising 80 high-resolution images. This dataset is sourced from LSDIR Li et al. (2023), DIV2K Agustsson & Timofte (2017), and the dataset introduced by Kulikov *et al.* in Kulikov et al. (2025). To ensure accurate image-to-text alignment, we utilize

Figure 4: **Qualitative results on SD3.** Our method not only achieves higher fidelity to the source image but also facilitates more effective semantic editing.

Gemma-3-12B Gemma Team & Google DeepMind (2025) to infer descriptive source prompts for each image. Subsequently, for every image-prompt pair, we manually craft two distinct target prompts to evaluate various editing scenarios, resulting in a total of 160 test cases. The quantitative evaluation of our method is executed on the dataset and prompts in Kulikov et al. (2025) to ensure a fair comparison.

4.2 Qualitative Evaluation

The most reliable evaluation of image editing quality is arguably qualitative assessment by human users. Therefore, we first evaluate our method qualitatively by performing a user study. We chose four metrics for this evaluation, namely Prompt Adherence, Scene Preservation, Pose Preservation, and overall Success. Prompt Adherence reflects the degree to which the edited image correlated with the given target prompt. The structural faithfulness is divided into two subcategories: Scene Preservation and Pose Preservation. Eventually, Overall Success gives the unified success of the editing. We performed the user study on 30 participants, each asked to rank the edited images on the four criteria given each source image and target

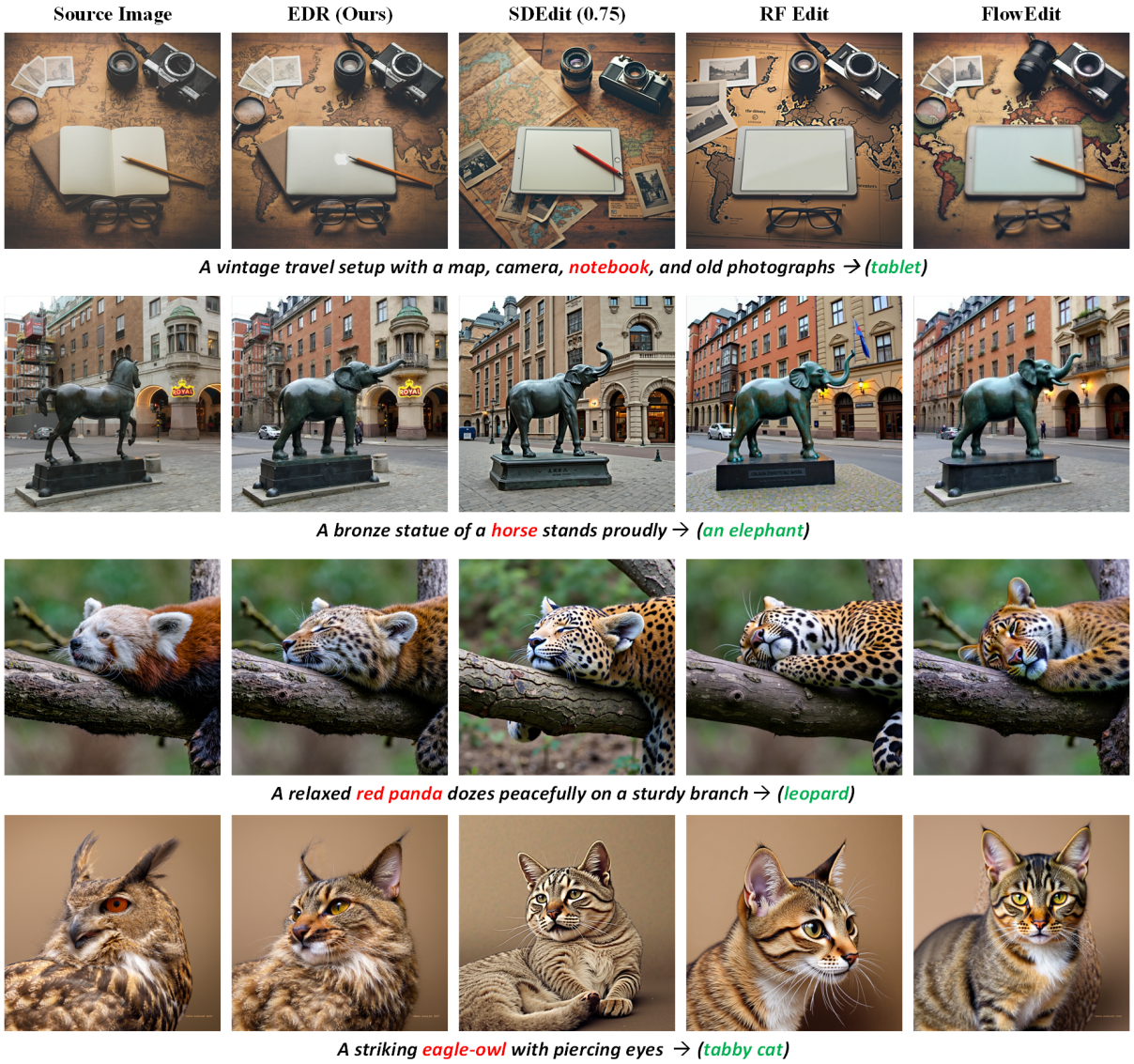

Figure 5: **Qualitative results on FLUX.** Our method produces high-quality edits while effectively preserving the background and subject pose of the source image.

Table 1: **User study on SD3.** The win rate (%) of our method against competing approaches across four criteria is presented. Our method achieves superior performance in editing effectiveness (Prompt Adherence), source image fidelity (Scene and Pose Preservation), and Overall Success compared to existing methods on SD3.

	SDEdit	ODE Inv.	FlowEdit
Prompt Adherence	79 %	96 %	72 %
Scene Preservation	83 %	74 %	67 %
Pose Preservation	77 %	75 %	78 %
Overall Success	81 %	93 %	75 %

prompt. We report the win rate of our method against compared methods on SD3 in Table 1 and on FLUX in Table 2. Representative results of our method using SD3 and FLUX are given in Figure 4 and Figure 5, respectively.

Table 2: **User study on FLUX.** We report the win rate (%) of our method against competing approaches. Our method remains significantly more faithful to the source images while maintaining prompt adherence on par with state-of-the-art baselines.

	SDEdit	RF Edit	FlowEdit
Prompt Adherence	57 %	52 %	50 %
Scene Preservation	94 %	71 %	67 %
Pose Preservation	82 %	89 %	85 %
Overall Success	89 %	83 %	78 %

Table 3: **Quantitative results on SD3.** Our proposed method achieves state-of-the-art performance, particularly in metrics evaluating structural preservation. The best values are highlighted in **bold**.

	$\text{CLIP}_{\text{txt}}\uparrow$	$\text{CLIP}_{\text{img}}\uparrow$	DINO ↑	LPIPS ↓	DreamSim ↓
SDEditMeng et al. (2021)	0.330	0.885	0.634	0.251	0.213
iRFDSYang et al. (2024)	0.335	0.822	0.534	0.376	0.327
FlowEditKulikov et al. (2025)	0.344	0.872	0.719	0.181	0.253
EDR (Ours)	**0.347**	**0.909**	**0.744**	**0.146**	**0.158**

4.3 Quantitative Evaluation

To quantitatively evaluate our proposed method, we employ the following metrics: cosine similarity of CLIP Radford et al. (2021) embeddings—specifically CLIP_{txt} to measure the alignment between the target image and the editing prompt (quantifying editing success), and CLIP_{img} to assess the similarity between the source and target images. Additionally, we report DINO Caron et al. (2021) embedding similarity to further evaluate the structural preservation between the source and target images. Also, we used LPIPS Zhang et al. (2018) to measure the perceptual distance between source and target images, and DreamSim Fu et al. (2023) that assesses object pose and holistic perceptual similarities in source and target images. In our quantitative evaluation, we adopted the dataset and the results for comparison methods from Kulikov et al. (2025). Quantitative results on SD3 and FLUX are given in Table 3 and Table 4, respectively. More results are given in the supplementary material.

5 Ablation Study

The proposed approach utilized degraded representation that is a combination of two different degradations, namely gaussian structural suppression (GSS) and dynamic range reduction (DRR). To verify that both these degradations are essential and complementary for the task, we executed experiments with dropping each of the operators. Figure 6 shows the effect of ablation of the two degradation operators. As expected, without dynamic range reduction the color palette of the edited image changed significantly. On the other hand, ablating Gaussian blur degradation, resulted in an edited image that drifted from the source image structure.

Furthermore, Table 5 compares our proposed degradation representation against baselines that omit individual components. Removing any single operator severely compromises overall fidelity. While ablating GSS and DRR yields a slight increase in editing strength CLIP_{txt}, this inflated text alignment score occurs because the generation trajectories become completely unconstrained. In this case, the model easily satisfies

Table 4: **Quantitative results on FLUX.** Our proposed method achieves state-of-the-art performance on metrics evaluating structural fidelity, while maintaining effective editing capabilities. The best values are highlighted in **bold**.

	$\text{CLIP}_{\text{txt}}\uparrow$	$\text{CLIP}_{\text{img}}\uparrow$	DINO ↑	LPIPS ↓	DreamSim ↓
SDEdit Meng et al. (2021)	0.316	0.902	0.637	0.264	0.180
RF Inversion Rout et al. (2024)	0.334	0.856	0.558	0.34	0.266
RF Edit Wang et al. (2024)	0.332	0.876	0.650	0.220	0.220
FlowEdit Kulikov et al. (2025)	**0.337**	0.875	0.682	0.223	0.252
EDR (Ours)	0.335	**0.914**	**0.739**	**0.170**	**0.136**

Figure 6: **Ablation Impact** Ablating dynamic range reduction (DRR) and gaussian structural suppression (GSS) from our proposed degraded representation. Omitting DRR leads to an editing that is not faithful to the original image intensity and color palette, and omitting GSS produces an edited image unfaithful to the source structure.

Table 5: **Ablation Study.** Dropping either of the degradation operators, GSS or DRR, results in a significant drop in metrics that measure editing fidelity.

(a) Ablation Study on SD3

	$\text{CLIP}_{\text{txt}}\uparrow$	$\text{CLIP}_{\text{img}}\uparrow$	DINO ↑	LPIPS ↓	DreamSim ↓
w/o GSS	0.347	0.873	0.612	0.287	0.278
w/o DRR	0.349	0.865	0.604	0.233	0.291
EDR	0.347	0.909	0.744	0.146	0.158

(b) Ablation Study on FLUX

	$\text{CLIP}_{\text{txt}}\uparrow$	$\text{CLIP}_{\text{img}}\uparrow$	DINO ↑	LPIPS ↓	DreamSim ↓
w/o GSS	0.336	0.847	0.582	0.291	0.225
w/o DRR	0.340	0.822	0.556	0.258	0.239
EDR	0.335	0.914	0.739	0.170	0.136

the target prompt only by ignoring structural preservation, which is reflected in the severe degradation across all fidelity metrics CLIP_{img}, DINO, LPIPS, and DreamSim.

6 Conclusion

In this work, we addressed the challenge of achieving faithful and effective image editing utilizing flow-based models. We demonstrated that the generation trajectories between two distinct distributions can be systematically restricted from moving in arbitrary directions. This is accomplished by projecting the trajectories onto degraded representations that are inherently invariant to such changes. Consequently, we proposed a method of constraining editing trajectories to prevent unintended visual deviations by leveraging these degraded representations, wherein specific image properties, namely pixel intensities and overall structure, are suppressed.

References

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. URL `http://www.vision.ee.ethz.ch/~timofter/publications/Agustsson-CVPRW-2017.pdf`.

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–12, 2024.

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023.

Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8861–8870, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Aram Davtyan, Sepehr Sameni, and Paolo Favaro. Efficient video prediction via sparsely conditioned flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23263–23274, 2023.

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

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. In *European Conference on Computer Vision*, pp. 395–413. Springer, 2024.

Gemma Team and Google DeepMind. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL `https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf`.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4775–4785, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12469–12478, 2024.

Tariq Berrada Ifriqi, John Nguyen, Karteek Alahari, Jakob Verbeek, and Ricky TQ Chen. Flowception: Temporally expansive flow matching for video generation. *arXiv preprint arXiv:2512.11438*, 2025.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6007–6017, 2023.

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

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1775–1787, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Yawen Lu, Qifan Wang, Siqi Ma, Tong Geng, Yingjie Victor Chen, Huaijin Chen, and Dongfang Liu. Transflow: Transformer as flow learner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18063–18073, 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. *arXiv preprint arXiv:2108.01073*, 2021.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Zhihong Pan, Riccardo Gherardi, Xiufeng Xie, and Stephen Huang. Effective real image editing with accelerated iterative diffusion inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15912–15921, 2023.

Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23051–23061, 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. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 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. *arXiv preprint arXiv:2410.10792*, 2024.

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 *The Thirteenth International Conference on Learning Representations*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1921–1930, 2023.

Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22532–22541, 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. *arXiv preprint arXiv:2411.04746*, 2024.

Navve Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. Paint by inpaint: Learning to add image objects by removing them first. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18313–18324, 2025.

Xiaofeng Yang, Cheng Chen, Xulei Yang, Fayao Liu, and Guosheng Lin. Text-to-image rectified flow as plug-and-play priors. *arXiv preprint arXiv:2406.03293*, 2024.

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

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6027–6037, 2023.

Yixuan Zhu, Wenliang Zhao, Ao Li, Yansong Tang, Jie Zhou, and Jiwen Lu. Flowie: Efficient image enhancement via rectified flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13–22, 2024.

Appendix

A Additional Results

Additional results on SD3 and FLUX using the proposed method EDR are presented in Figure A.

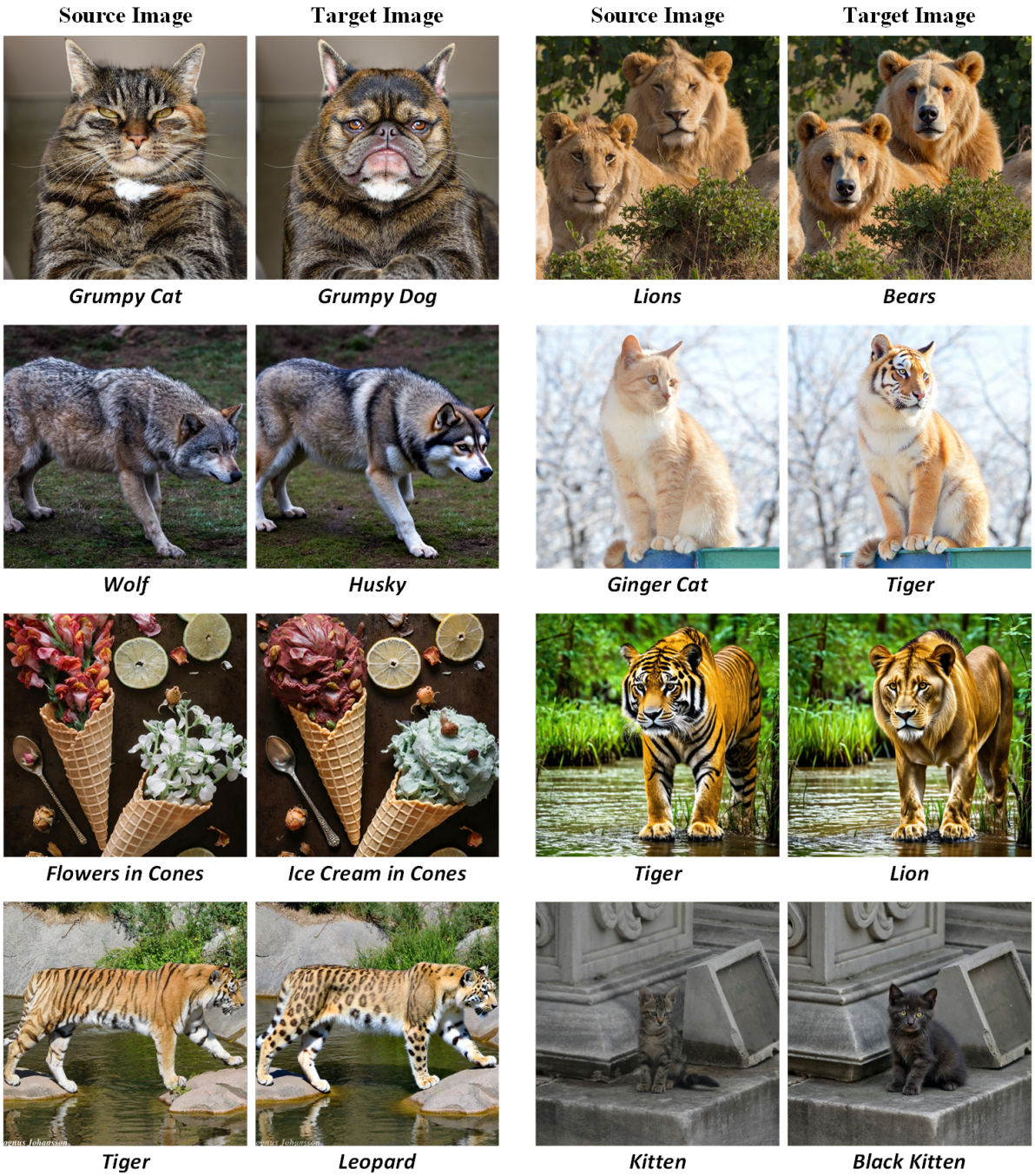

Figure A: Additional results of EDR and FLUX (top 2 rows) and SD3 (bottom 2 rows)

Figure B: EDR for style editing

Furthermore, our method is highly effective and faithful for complex style editing tasks (Figure B), successfully adopting target aesthetics while strictly preserving the source image's structural integrity.

B Hyperparameter Selection

The hyperparameters used in our method were selected based on a comprehensive set of experiments. Specifically, we executed a grid search over the standard deviation σ (used for Gaussian structural smoothing) and the reduced dynamic range $[I_{\min}, I_{\max}]$ on SD3. We evaluated these combinations by plotting CLIP_{txt} to measure editing success against LPIPS to measure source image fidelity. As shown in Figure C, the tested combinations form a distinct trade-off curve. We select the configuration $\sigma = 5$ and $[I_{\min}, I_{\max}] = [0.25, 0.75]$ (green triangle), as it sits on the Pareto front and achieves the best empirical balance without compromising either metric.

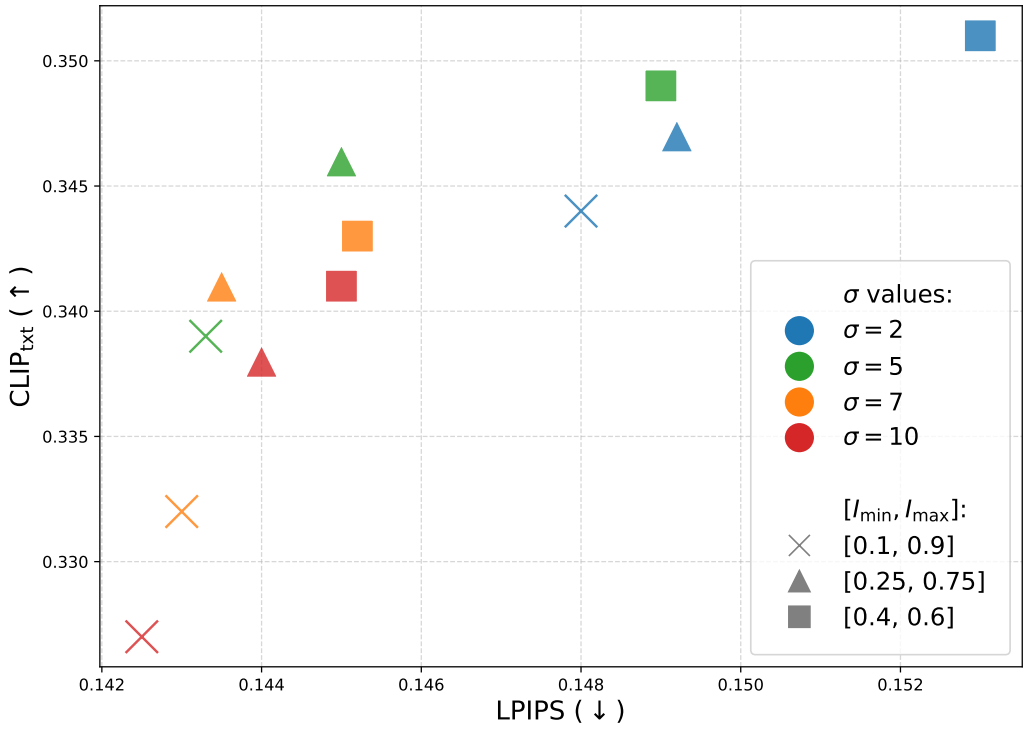

Figure C: **Hyperparameter Grid Search.** Quantitative evaluation of different configurations for the structural smoothing factor σ and the intervention bounds $[I_{\min}, I_{\max}]$. The scatter plot illustrates the trade-off between editability (measured by CLIP_{txt}, where higher is better) and source image fidelity (measured by LPIPS, where lower is better). The green triangle denotes our selected optimal configuration ($\sigma = 5$ and $[I_{\min}, I_{\max}] = [0.25, 0.75]$) situated on the Pareto front, achieving the best balance between successful text-driven editing and structural preservation.

C Pseudocode

The complete pseudocode of the proposed method (EDR) is presented in Algorithm A.

Algorithm A: Editing via Degraded Representations (EDR)

Input: Source image X_{src}, source and target prompts c_{src} and c_{tar}, degradation map $\mathcal{D}_c(.)$ (standard deviation σ, reduced dynamic range $[I_{\min}, I_{\max}]$) based on Eq. 9 in the main paper, total timesteps T, editing strength t_0, and decay rate γ

Output: Edited image X_{tar}

$n \leftarrow \lfloor t_0 \times T \rfloor$
$t \leftarrow t_0$
$Z_t^{\text{dir}} \leftarrow X_{\text{src}}$
$\Delta t \leftarrow \frac{t_0}{n}$
for $i = n$ **to** 1 **do**
  $N \sim (0, I)$
  $Z_t^{\text{src}} \leftarrow (1-t)Z_t + tN$
  $Z_t^{\text{tar}} \leftarrow Z_t^{\text{dir}} + Z_t^{\text{src}} - Z_0^{\text{src}}$
  $V_t^{\text{dir}}(Z_t^{\text{dir}}, t) \leftarrow V_t^{\text{tar}}(Z_t^{\text{tar}}, t) - V_t^{\text{src}}(Z_t^{\text{src}}, t)$
  $\tilde{V}_t = \mathcal{D}_{\text{c}}(\frac{\Delta Z_t^{\text{dir}}}{\Delta t})$
  $V_t^{\text{proj}} = \frac{\langle V_t^{\text{dir}}, \tilde{V}_t \rangle}{\|\tilde{V}_t\|^2} \tilde{V}_t$
  $\alpha = (1-t)^{\gamma}$
  $V_t^{\text{EDR}} = \alpha V_t^{\text{proj}} + (1-\alpha) V_t^{\text{dir}}$
  $Z_{t-\Delta t}^{\text{dir}} \leftarrow Z_t^{\text{dir}} + V_t^{\text{EDR}}$
  $t \leftarrow t - \Delta t$
return $X_{\text{tar}} = Z_t^{\text{dir}}$