FLOWOPT: FAST OPTIMIZATION THROUGH WHOLE FLOW PROCESSES FOR TRAINING-FREE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

The remarkable success of diffusion and flow-matching models has ignited a surge of works on adapting them at test time for controlled generation tasks. Examples range from image editing to restoration, compression and personalization. However, due to the iterative nature of the sampling process in those models, it is computationally impractical to use gradient-based optimization to directly control the image generated at the end of the process. As a result, existing methods typically resort to manipulating each timestep separately. Here we introduce FlowOpt – a zero-order (gradient-free) optimization framework that treats the entire flow process as a black box, enabling optimization through the whole sampling path without backpropagation through the model. Our method is both highly efficient and allows users to monitor the intermediate optimization results and perform early stopping if desired. We prove a sufficient condition on FlowOpt's step-size, under which convergence to the global optimum is guaranteed. We further show how to empirically estimate this upper bound so as to choose an appropriate step-size. We demonstrate the effectiveness of FlowOpt in the context of image editing, showcasing two use cases: (i) inversion (determining the initial noise that generates a given image), and (ii) directly steering the edited image to be similar to the source image while conforming to the target text prompt. In both settings, our method achieves state-of-the-art results while using roughly the same number of neural function evaluations (NFEs) as existing methods.

1 INTRODUCTION

Diffusion and flow matching models have emerged as powerful generative frameworks, achieving state-of-the-art (SotA) results on image, video, and audio generation (Ho et al., 2020; Song et al., 2021a; Rombach et al., 2022; Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023). However, as opposed to their generative adversarial network (GAN) predecessors, flow models generate samples through an iterative process that often involves dozens of neural function evaluations (NFEs). This makes it challenging to adapt them at inference time for solving controlled generation tasks. Indeed, while GANs naturally lend themselves to gradient-based optimization for directly minimizing losses on the generator's output (Menon et al., 2020), in flow models this approach is computationally impractical. As a result, methods that use pre-trained flow models for controlled generation typically intervene in each step of the sampling process separately, without employing any direct supervision on the final result. This strategy is used *e.g.*, for image restoration, image editing (using inversion techniques), and image compression (Kawar et al., 2022; Tumanyan et al., 2023; Pan et al., 2023; Qi et al., 2023; Huberman-Spiegelglas et al., 2024; Hong et al., 2024; Cohen et al., 2024; Garibi et al., 2024; Manor & Michaeli, 2024; Elata et al., 2025; Wang et al., 2025; Martin et al., 2025; Deng et al., 2025; Ohayon et al., 2025; Samuel et al., 2025).

Recently, Ben-Hamu et al. (2024) demonstrated the great potential of employing optimization through the whole flow process in the context of solving inverse problems with pre-trained flow models. Unlike other methods, this approach directly controls the generated image, and thus avoids accumulation of approximation errors that can build up throughout the flow path. However, performing gradient-based optimization is not scalable to reasonably sized models and image dimensions. In fact, even with a small flow-matching model, small images (128×128), and memory-saving techniques like gradient checkpointing, this approach takes approximately 15 minutes to run on a single input.

Figure 1: FlowOpt. We propose a zero-order (gradient-free) framework for optimization through an unrolled flow sampling process. FlowOpt can efficiently optimize losses on the target image, even when working with large models and high resolution images. We leverage our framework for text-based image editing, demonstrating state-of-the-art results on both FLUX (first and third rows) and Stable Diffusion 3 (second row). Fine details are visible upon zooming in.

In this work, we introduce FlowOpt – a zero-order (gradient-free) optimization framework for directly minimizing loss functions on the target image without backpropagating through the model. Specifically, unrolling the sampling process, a flow model can be viewed as a chain of neural networks, which we refer to as "denoisers". Our approach treats this entire chain of denoisers as a black box, and enables optimization with respect to arbitrary loss functions. Here we specifically focus on image-editing objectives. The avoidance of backpropagation enables working with large flow models and treating large images. Furthermore, it allows using a small number of flow timesteps, which is in contrast with inversion-based techniques that often require many timesteps to avoid error accumulation. Taken together, these features enable FlowOpt to achieve SotA results at a number of NFEs comparable to existing methods. Additionally, FlowOpt allows monitoring the intermediate optimization results. Thus, at the same budget of NFEs as existing methods, FlowOpt in fact provides multiple candidate edited images (one per optimization step) from which the user can choose.

Zero-order optimization has been previously used in several computer vision contexts (Tao et al., 2017; Milanfar, 2018; Chen et al., 2019; Tu et al., 2019). FlowOpt is a generalization of the method of Tao et al. (2017), with the difference that the update in each optimization step is multiplied by a step-size η (the method of Tao et al. (2017) corresponds to FlowOpt with $\eta = 1$). As we show, this modification is of dramatic importance. Specifically, we prove a sufficient condition on η under which convergence to the global minimum is guaranteed, and show that for popular flow models this bound is orders of magnitude smaller than 1. We demonstrate that FlowOpt indeed converges when η is chosen smaller than the bound, and fails to converge when it significantly exceeds the bound.

We demonstrate the effectiveness of FlowOpt for both image reconstruction (inversion) and direct image editing (Fig. 1), using the FLUX-1.dev (Black Forest Labs, 2024) and Stable Diffusion 3 (SD3) (Esser et al., 2024) text-to-image (T2I) models. We show that FlowOpt provides an efficient solution to these tasks, delivering SotA performance at running times comparable to existing methods.

2 Related Work

T2I diffusion and flow-based models (Saharia et al., 2022; Ramesh et al., 2022) generate images by steering a diffusion or flow process according to a text prompt provided by the user. Latent diffusion and flow-based variants (Rombach et al., 2022; Vahdat et al., 2021; Dao et al., 2023) follow the same principle but operate in a lower-dimensional latent space, improving computational efficiency while

preserving visual fidelity. Many methods utilize these T2I foundation models for downstream tasks like image editing in a zero-shot manner.

A common approach for performing image editing with pre-trained diffusion/flow models is to start with an inversion stage (Song et al., 2021a) (often referred to as DDIM or ODE inversion), whose goal is to extract the initial noise that would generate the input image if used in a regular sampling process. Once this initial noise is obtained, it is used for sampling a new image, by using a text prompt that describes the desired edit. However, inversion methods introduce approximation errors that accumulate across the flow timesteps, and lead to significant reconstruction inaccuracies (Mokady et al., 2023; Huberman-Spiegelglas et al., 2024).

One line of work focuses on improving the precision of ODE-inversion. Wang et al. (2025) employ a high-order Taylor expansion to more accurately approximate the nonlinear components of the flow. Deng et al. (2025) propose a solver that reuses intermediate velocity vector approximations. Yet, despite improving numerical accuracy, such methods still operate on each timestep separately and do not promote direct alignment with the given image during the inversion. Therefore, they still suffer from accumulation of errors that can degrade overall performance.

A different approach is to optimize each denoising timestep independently (Mokady et al., 2023; Pan et al., 2023; Hong et al., 2024; Garibi et al., 2024; Miyake et al., 2025; Samuel et al., 2025). For instance, Mokady et al. (2023) optimize the unconditional null prompt embedding used in classifier-free guidance (CFG) (Ho & Salimans, 2021) during the reverse process, aligning latent variables obtained through DDIM inversion. While effective, this approach requires storing all latent variables and optimized embeddings in memory, which becomes prohibitive for a large number of timesteps. Furthermore, repeated backward passes through each timestep render such methods impractical for interactive editing with large-scale models. Hong et al. (2024) propose a gradient-based inversion scheme applied independently at each timestep, however their method is computationally expensive and time-intensive, particularly for modern large-scale T2I models. Pan et al. (2023) and Garibi et al. (2024) mitigate this by introducing fixed-point iteration strategies that iteratively refine approximations of predicted states along the diffusion trajectory. However, all these methods rely on optimizing each timestep independently, ignoring the input image in each optimization step. This leads to accumulation of local approximation errors that degrade overall performance.

There exist several optimization-based methods that may superficially seem similar to FlowOpt, as they neglect the Jacobian of the denoiser and thus avoid backpropagation through the model. These include Score Distillation Sampling (SDS) (Poole et al., 2023), Delta Denoising Score (DDS) (Hertz et al., 2023), Posterior Distillation Sampling (PDS) (Koo et al., 2024), and inverse Rectified Flow Distillation Sampling (iRFDS) (Yang et al., 2025). However, these methods still optimize each timestep separately by randomly sampling a timestep in each optimization step and performing an update based on that timestep alone. This is in contrast with FlowOpt, which performs optimization through the whole chain of denoisers simultaneously.

Recently, Patel et al. (2025) proposed FlowChef, a method that initializes the sampling process from white Gaussian noise, and then performs zero-order optimization at each denoising timestep separately. Unlike FlowOpt, this method does not treat the entire flow process as a black box. A detailed comparison between the two methods is provided in App. K.

Finally, Ben-Hamu et al. (2024) proposed D-Flow, a method that like FlowOpt, optimizes across the entire generative process. However, their framework relies on gradient-based optimization and requires repeated backpropagation through the entire chain of denoisers. This makes the method computationally intensive and impractical for high-resolution, real-world applications – precisely the setting we aim to address with FlowOpt.

3 Preliminaries and Notation

Probability flow ODE (Song et al., 2021b) and flow-matching models (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023) generate images by numerically solving an ODE over a time parameter t. Focusing for simplicity on the flow-matching formalism, the ODE takes the form

$$d\boldsymbol{z}_t = \boldsymbol{v}_t(\boldsymbol{z}_t, c)\, dt, \quad t \in [0, 1]. \tag{1}$$

This ODE is designed such that when initialized at $t = 1$ with a sample from some source distribution (usually taken to be an isotropic Gaussian), $\boldsymbol{z}_1 \sim \pi_1$, and run backwards in time until $t = 0$, it yields

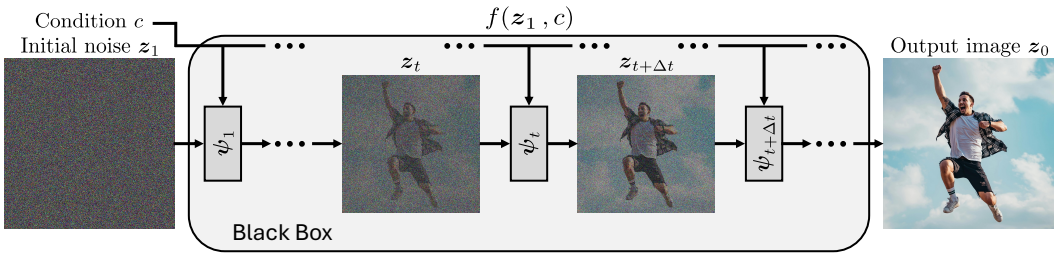

Figure 2: A whole flow process as a black box. We encapsulate the flow process as a black box function f, which receives an initial noise $\boldsymbol{z}_1$ and text conditioning c, and outputs a clean sample $\boldsymbol{z}_0$. Each internal step within the black box is given by $\psi_t(\boldsymbol{z}_t, c) = \boldsymbol{z}_t + \boldsymbol{v}_t(\boldsymbol{z}_t, c)\Delta t$, where $\boldsymbol{v}_t$ is the text-conditioned velocity predicting network.

a sample from a desired target distribution (*e.g.* the distribution of natural images), $\boldsymbol{z}_0 \sim \pi_0$. The function $\boldsymbol{v}_t(\cdot, \cdot)$ is a time dependent vector field that optionally accepts a condition c (*e.g.*, a text prompt) in its second argument. In practice, this velocity field is implemented by a neural network, which we refer to as "denoiser", and the ODE is discretized and solved numerically as

$$\boldsymbol{z}_{t+\Delta t} = \boldsymbol{z}_t + \boldsymbol{v}_t(\boldsymbol{z}_t, c)\,\Delta t, \tag{2}$$

where Δt is the (negative) discretization step.

Unrolling Eq. (2), the sample $\boldsymbol{z}_0$ generated at the end of the flow process can be written as a function of the initial noise $\boldsymbol{z}_1$, namely $\boldsymbol{z}_0 = f(\boldsymbol{z}_1, c)$. This function is given by

$$f(\boldsymbol{z}_1, c) = \boldsymbol{z}_1 + \sum_i \boldsymbol{v}_{t_i}(\boldsymbol{z}_{t_i}, c)\,\Delta t, \tag{3}$$

where $t_i = 1 + i\,\Delta t$ (see Fig. 2). For notational simplicity, we henceforth omit the condition c whenever it is clear from the context. Furthermore, we sometimes use $f(\cdot)$ to denote the mapping from some intermediate timestep $t < 1$ to timestep $t = 0$. Our method treats the function $f(\cdot)$ as a black box in the sense that it can be evaluated but its Jacobian cannot be computed.

Commonly, the flow process is defined in the latent space of an encoder $\mathcal{E}(\cdot)$, so that the final image is obtained by passing the generated sample $\boldsymbol{z}_0$ through the corresponding decoder $\mathcal{D}(\cdot)$.

4 Method

Given a source image $\boldsymbol{y}$, a text prompt c_{src} describing it, and a target text prompt c_{tar} describing a desired edit, our goal is to generate an edited image $\boldsymbol{y}_{\text{edit}}$ that conforms to c_{tar} while being as similar as possible to $\boldsymbol{y}$. Like previous approaches, we rely on a pre-trained flow model. However, in contrast to existing methods we propose to achieve this by directly optimizing over the vector $\boldsymbol{z}_t$ at some timestep t (usually taken to be 1), such that the image $\boldsymbol{z}_0$ at the end of the flow process is close to $\boldsymbol{y}$.

Formalizing this mathematically, we are interested in $\boldsymbol{z}_t^* = \arg\min_{\boldsymbol{z}_t} \mathcal{L}(f(\boldsymbol{z}_t, c), \boldsymbol{y})$, where $\mathcal{L}$ is some dissimilarity measure. Let us focus on the L^2 loss (see App. E for other losses). In this case,

$$\boldsymbol{z}_t^* = \arg\min_{\boldsymbol{z}_t} \frac{1}{2}\|f(\boldsymbol{z}_t, c) - \boldsymbol{y}\|^2. \tag{4}$$

This optimization problem can be used in two distinct ways. (i) **Inversion:** setting $c = c_{\text{src}}$ in Eq. (4) leads to a $\boldsymbol{z}_t^*$ that reconstructs the input image with the source prompt. (ii) **Direct editing:** setting $c = c_{\text{tar}}$ in Eq. (4) leads to a $\boldsymbol{z}_t^*$ that directly approximates the input image with the target prompt. In both cases, once $\boldsymbol{z}_t^*$ is obtained, it can be used to generate the edited image by performing sampling with the target prompt, $\boldsymbol{y}_{\text{edit}} = f(\boldsymbol{z}_t^*, c_{\text{tar}})$.

Using gradient descent to solve Eq. (4) would lead to the iterations

$$\boldsymbol{z}_t^{(i+1)} \leftarrow \boldsymbol{z}_t^{(i)} - \eta\, \boldsymbol{J}(\boldsymbol{z}_t^{(i)})^\top \left(f(\boldsymbol{z}_t^{(i)}) - \boldsymbol{y}\right), \tag{5}$$

where η is the step size and $\boldsymbol{J}(\boldsymbol{z}_t^{(i)})$ is the Jacobian of $f(\cdot)$ with respect to $\boldsymbol{z}_t^{(i)}$. However, as mentioned above, backpropagation through whole flow processes is computationally impractical.

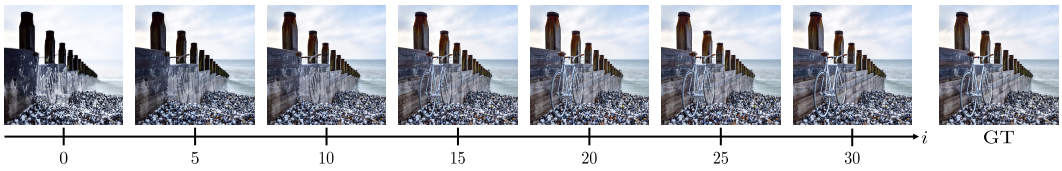

Figure 3: Image inversion with FlowOpt. Intermediate samples $\boldsymbol{z}_0^{(i)} = f(\boldsymbol{z}_t^{(i)}, c)$ attained during our zero-order optimization through a chain of 10 denoising steps (FLUX) for the task of reconstruction (inversion), *i.e.*, with $c = c_{\text{src}}$. Notice the missing details in the early steps, such as the bicycle and the horizon. As the iterations progress, the reconstruction converges to the ground truth image.

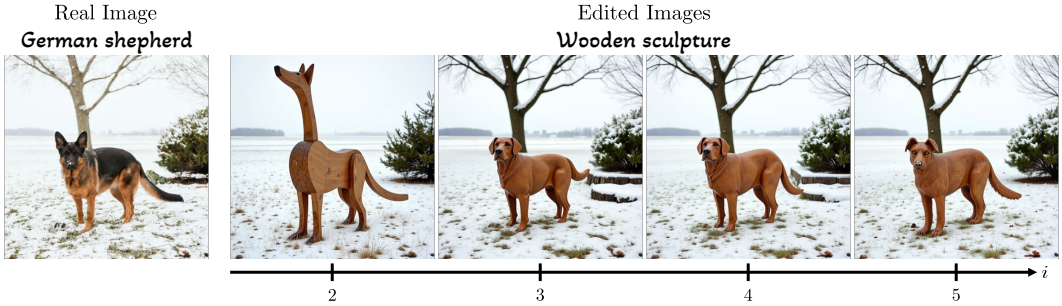

Figure 4: Direct image editing with FlowOpt. Intermediate samples $\boldsymbol{z}_0^{(i)} = f(\boldsymbol{z}_t^{(i)}, c)$ attained during our zero-order optimization through a chain of 15 denoising steps (FLUX) for direct image editing, *i.e.*, with $c = c_{\text{tar}}$. Notice the misalignment in the dog's body structure in the first iterations.

Therefore, as an alternative, here we propose to simply ignore the Jacobian. This leads to the zero-order (gradient-free) iterations

$$\boldsymbol{z}_t^{(i+1)} \leftarrow \boldsymbol{z}_t^{(i)} - \eta \left(f(\boldsymbol{z}_t^{(i)}) - \boldsymbol{y} \right). \tag{6}$$

Figure 3 demonstrates the progression of those iterates when used for inversion (with the source prompt). Figure 4 demonstrates the progression of the iterates when used for direct editing (with the target prompt). Algorithm 1 summarizes the proposed method.

Before providing a theoretical convergence guarantee, two comments are in place. First, when $\eta = 1$, Eq. (6) degenerates to the method of Tao et al. (2017). However, as we show below, η is of crucial importance, as the maximal step size allowing convergence is much smaller than 1 for modern flow-matching models. Second, it is insightful to note that for flow-matching models, Eq. (6) is equivalent to using gradient descent with step-size η while applying the `stop-grad` operator on the output of the velocity prediction network. Similarly, for probability flow ODE models (Song et al., 2021b), (a.k.a. DDIM (Song et al., 2021a)), Eq. (6) is equivalent to using gradient descent with step size $\sqrt{\alpha_T}\eta$ while applying `stop-grad` on the noise prediction network (following the notation of Song et al. (2021a)). The derivations of those observations are provided in App. G.

The iterations of Eq. (6) can be written as $\boldsymbol{z}_t^{(i+1)} = g(\boldsymbol{z}_t^{(i)})$, where $g(\boldsymbol{u}) \triangleq \boldsymbol{u} - \eta(f(\boldsymbol{u}) - \boldsymbol{y})$. By the Banach fixed-point theorem, if $g(\cdot)$ is a contractive mapping[1] then there exists a unique point satisfying $\boldsymbol{z}_t^* = g(\boldsymbol{z}_t^*)$, and thus $f(\boldsymbol{z}_t^*) = \boldsymbol{y}$. Furthermore, in this case the iterations converge to this unique solution. This fact can be used to obtain a sufficient condition on the step size η under which the iterations are guaranteed to converge to the global minimum (see proof in App. F).

Theorem 1. *Assume that* $\inf_{\boldsymbol{u}_1 \neq \boldsymbol{u}_2} \frac{\langle \boldsymbol{u}_1 - \boldsymbol{u}_2, f(\boldsymbol{u}_1) - f(\boldsymbol{u}_2) \rangle}{\|\boldsymbol{u}_1 - \boldsymbol{u}_2\| \|f(\boldsymbol{u}_1) - f(\boldsymbol{u}_2)\|} > 0$ *and* $\sup_{\boldsymbol{u}_1, \boldsymbol{u}_2} \frac{\langle \boldsymbol{u}_1 - \boldsymbol{u}_2, f(\boldsymbol{u}_1) - f(\boldsymbol{u}_2) \rangle}{\|f(\boldsymbol{u}_1) - f(\boldsymbol{u}_2)\|^2} < \infty$. *If the step size η satisfies*

$$0 < \eta < 2 \inf_{\boldsymbol{u}_1, \boldsymbol{u}_2} \frac{\langle \boldsymbol{u}_1 - \boldsymbol{u}_2, f(\boldsymbol{u}_1) - f(\boldsymbol{u}_2) \rangle}{\|f(\boldsymbol{u}_1) - f(\boldsymbol{u}_2)\|^2} \tag{7}$$

then there is a unique $\boldsymbol{z}_t^*$ satisfying $f(\boldsymbol{z}_t^*) = \boldsymbol{y}$ and the iterations of Eq. (6) converge to this $\boldsymbol{z}_t^*$.

[1] $g(\cdot)$ is a contractive mapping if it satisfies $\|g(\boldsymbol{u}_1) - g(\boldsymbol{u}_2)\| \leq \gamma \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|$ for some $\gamma < 1$ and all $\boldsymbol{u}_1, \boldsymbol{u}_2$.

Algorithm 1: Flow Zero-Order Optimization (FlowOpt)

Require: step size η, number of iterations N, condition c, input image $\boldsymbol{y}$

Initialization: $\boldsymbol{z}_t^{(0)} \in \mathbb{R}^d$

for $i \leftarrow 0, \ldots, N-1$ **do**

$\quad \boldsymbol{z}_0^{(i)} = f(\boldsymbol{z}_t^{(i)}, c)$

$\quad \boldsymbol{z}_t^{(i+1)} \leftarrow \boldsymbol{z}_t^{(i)} - \eta(\boldsymbol{z}_0^{(i)} - \boldsymbol{y})$

$\boldsymbol{z}_0^{(N)} = f(\boldsymbol{z}_t^{(N)}, c)$

Return $\{\boldsymbol{z}_0^{(i)}\}_{i=0}^{N}$

Table 1: Step sizes guaranteeing convergence. Column 2 shows the estimated sufficient condition of Eq. (7) and column 3 reports the step size we chose for each model (see App. F for details).

Model	Sufficient condition (Eq. (7))	Our chosen step size
FLUX	$\eta < 2.70 \cdot 10^{-3}$	$\eta = 2.5 \cdot 10^{-3}$
SD3	$\eta < 1.67 \cdot 10^{-2}$	$\eta = 1.0 \cdot 10^{-2}$

The bound in Eq. (7) depends only on the flow model $f(\cdot)$. It can thus be computed once for each model in order to choose the step size. In App. F we approximate this upper bound for the FLUX and SD3 models by drawing many pairs of samples $\boldsymbol{u}_1, \boldsymbol{u}_2$. As we show, the right-hand side of Eq. (7) is smallest when $\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|$ is small. Tab. 1 shows the bounds estimated for the two models, and the step sizes we chose for our experiments.

As can be seen, the bounds in Tab. 1 are significantly smaller than 1, suggesting that the method of Tao et al. (2017) is inapplicable in our setting. Indeed, Fig. 5 shows the reconstruction error along the iterations for several choices of η when used for inversion with SD3 (results for FLUX are presented in App. F). When setting $\eta = 10^{-2}$, which is below the bound of $1.67 \cdot 10^{-2}$, the iterations converge. However, when using larger step sizes, like $4 \cdot 10^{-2}$ or $5 \cdot 10^{-2}$, the iterations fail to converge. The setting of this experiment is as in Sec. 5.1. For additional convergence results with other image dimensions, please see App. J.

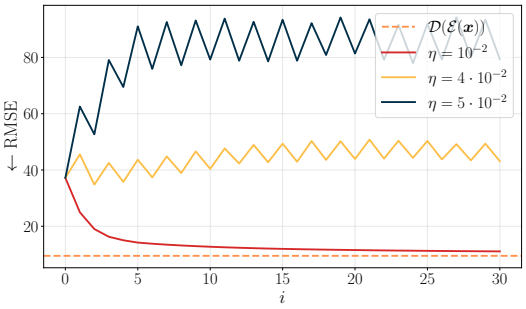

Figure 5: Convergence analysis. The plot shows RMSE in pixel space vs. number of iterations for the task of inversion, averaged over a dataset. The step size we use (red) satisfies the sufficient condition of Eq. (7) and thus leads to convergence. Step sizes that are $4\times$ and $5\times$ larger (yellow and black) do not satisfy the condition and do not lead to convergence. The dashed orange line is the minimal RMSE achievable in this setting. It corresponds to passing images through the encoder and decoder.

5 Experiments

We compare FlowOpt against competing methods on two tasks: image reconstruction (inversion) and text-based image editing. We show results with FLUX-1.dev in the main text and with SD3 in App. D. We use the step sizes reported in Tab. 1 and initialize our algorithm with the UniInv (Jiao et al., 2025) inversion method (see App. C for details). All images are of dimensions 1024×1024.

5.1 Image Reconstruction (Inversion)

For inversion, we use $c = c_{\text{src}}$ in Eq. (4), setting it to a text prompt describing the source image. We set the number of flow steps in FLUX (number of denoisers) to $T = 10$ and evaluate the reconstruction error for various numbers of NFEs by varying the number of FlowOpt iterations N. Specifically, we

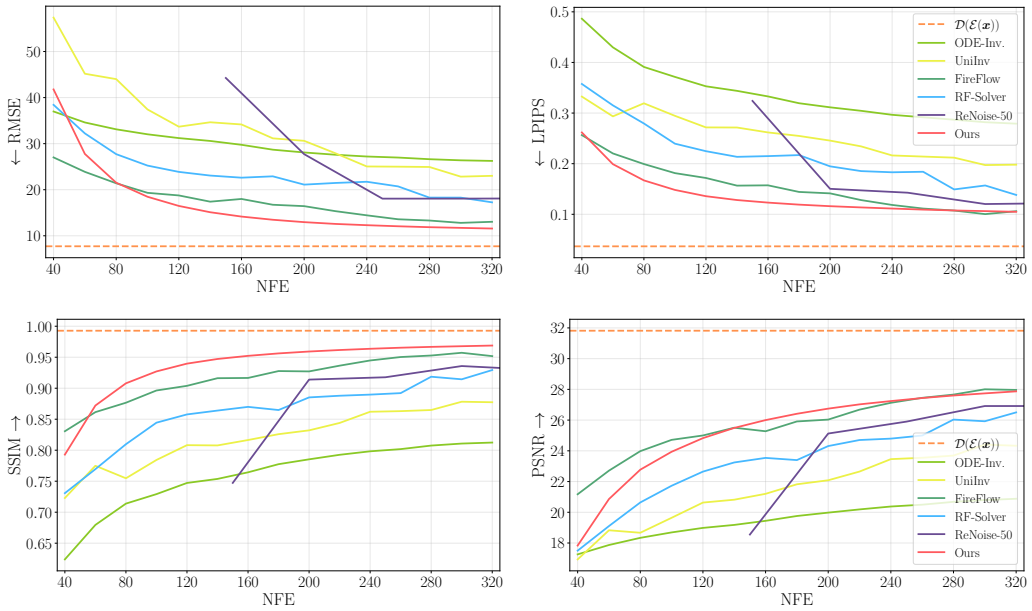

Figure 6: Reconstruction accuracy vs. NFEs for inversion. The plots depict pixel-space RMSE, LPIPS, SSIM, and PSNR as a function of the number of NFEs for several inversion methods. The dashed bound corresponds to passing the images through the encoder and decoder. FlowOpt achieves favorable reconstruction quality under 240 NFEs, which is the regime of practical interest.

have NFE $= T(N+2)$, as T NFEs are used for the initialization, NT NFEs for the optimization process, and T NFEs for the final sampling process.

We randomly choose 100 real images from the DIV2K dataset (Agustsson & Timofte, 2017), and resize and center-crop them to dimension 1024×1024. For the source prompts, we caption each image with BLIP (Li et al., 2022) and then manually refine the prompt.

We compare FlowOpt to several inversion methods: naive ODE Inversion, RF-Solver (Wang et al., 2025), FireFlow (Deng et al., 2025), UniInv (Jiao et al., 2025), and ReNoise (Garibi et al., 2024). We use the official implementations of all methods except for ODE Inversion and ReNoise (that lacks an implementation for flow models), which we implemented by ourselves. To ensure a fair comparison, we set the number of timesteps for each method such that the total NFE count is the same for all methods. Specifically, for FireFlow and UniInv, which use a single forward pass per timestep, we set $T = \frac{\text{NFE}}{2}$. For RF-Solver, which uses two forward passes per timestep for inversion and two for sampling, we set $T = \frac{\text{NFE}}{4}$. For ReNoise, we used $T = 50$ and set the number of ReNoise steps so as to achieve the desired NFE count. We note that we evaluated ReNoise with various hyperparameter settings and chose the one that achieved the best results.

Figure 6 shows the reconstruction accuracy achieved by all methods as a function of the NFEs. The figure reports pixel-space RMSE, PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). As can be seen, FlowOpt achieves the best reconstruction results over a wide range of NFE counts. In App. B we show that the same trend is obtained with empty text prompts, both with the CFG parameter of FLUX set to 0 and with it set to 1 (these options differ as FLUX is a distilled model).

5.2 Image Editing

Accurate inversion does not necessarily lead to good editing results. Indeed, even for synthetic images, for which the initial noise map is known, plain editing-by-inversion leads to unsatisfactory results (Kulikov et al., 2025; Huberman-Spiegelglas et al., 2024) (see App. I for further discussion). Accordingly, for the task of editing we employ our direct optimization approach, where the target text prompt $c = c_{\text{tar}}$ is used in Eq. (4). In this case, we do not necessarily want a large number of iterations, to avoid getting too close to the original image. We therefore use $N \in \{2, 3, 4, 5\}$.

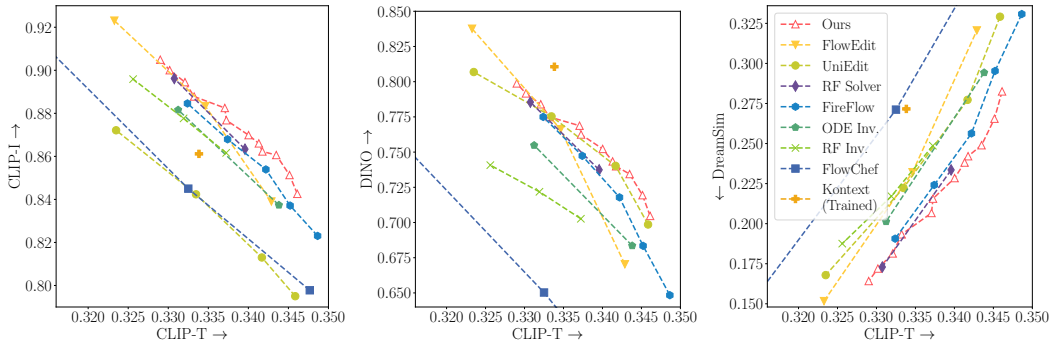

Figure 7: Editing quantitative comparisons. Semantic preservation of different editing methods evaluated using CLIP-Image, DINOv3 and DreamSim as functions of text adherence, measured by CLIP-Text. Connected markers represent different set of hyperparameters (see App. B). Our method achieves the most favorable balance between semantic preservation and text adherence.

Figure 8: FlowOpt editing results. Our method successfully preserves the object's semantics and structure, as well as the background details, all the while loyally adhering to the target text prompt. Fine details are visible upon zooming in.

We set the number of flow steps to $T = 15$ and perform the optimization on the latent vector at timestep $n_{\max} \in \{14, 13, 12\}$ (corresponding to t in Eq. (4)). The total number of NFEs is given by $\text{NFE} = n_{\max}(N + 2)$. We use the default CFG of 3.5. All visual results in the paper were obtained with $n_{\max} = 13$, except for Fig. 1, whose hyperparameters are provided in App. H.

We evaluate all methods on the dataset of Kulikov et al. (2025), which we enriched with additional images and editing prompts. In total, our dataset consists of 90 real images of dimensions 1024×1024 from the DIV2K dataset and from royalty free online sources (Pexels, 2025; PxHere, 2025). Each image was captioned by LLaVA-1.5 (Liu et al., 2024) and manually refined. For each image, we manually created target editing prompts. Overall, this led to about 400 text-image pairs.

We compare our method against all aforementioned methods, in addition to FlowEdit (Kulikov et al., 2025), FlowChef (Patel et al., 2025) and RF-Inversion (Rout et al., 2025). These three methods were excluded from the inversion experiments of Sec. 5.1 as they do not use inversion in the regular sense (FlowEdit is inversion-free and RF-Inversion and FlowChef explicitly incorporate the source image into the denoising process). For ODE Inversion, we apply the same number of NFEs as our

method. For other methods, we use the hyperparameters reported in the papers or in the official implementations. We performed a hyperparameter search for all methods that provided more than a single set of hyperparameters. Additional details and the final hyperparameters chosen for each method are provided in App. B. In addition to this set of zero-shot methods, we further compare to FLUX Kontext (Black Forest Labs et al., 2025), a trained text-based editing model.

Figures 1, 8 and S1 showcase the diverse editing capabilities of our method, including object replacement, style changes, and texture editing. FlowOpt achieves high quality, text adherent edits that also remain loyal to the source image semantics. Figure 10 presents qualitative comparisons between FlowOpt and other methods. As can be observed, our edits maintain superior alignment with the source image's structure while simultaneously adhering to the target text. For example, when turning the horse into a zebra (first row), FlowOpt successfully preserves the leg positions. Note that FLUX Kontext is a trained model; therefore, its capacity for changing the color palette of the source image is larger. For additional comparisons, see App. B.

Figure 7 presents a numerical evaluation of the results obtained for various hyperparameters. We use cosine similarity on CLIP image and text embeddings (Radford et al., 2021) to measure adherence to the original image and to the target text prompt, respectively. For image adherence, we also use cosine similarity between DINOv3 embeddings (Caron et al., 2021; Siméoni et al., 2025), as well as DreamSim (Fu et al., 2023). As can be seen, our method achieves the best tradeoff between text adherence and structure preservation.

Additionally, we evaluate our method via a user study, in which each participant was shown the reference image, an edit instruction, and two editing results – one from our method and another from a competing method. The order of the two editing results was random. We compared our method to FireFlow, FlowEdit and RF Solver, which achieve the most comparable results to FlowOpt in terms of CLIP-Text and CLIP-Image measures. Users were asked 3 two-alternative forced questions to select their preferred editing result: (i) visual fidelity between the reference image and the edited result, (ii) text alignment between the edited instruction and the edited result, and (iii) overall. We collect 60 user responses, covering a sample size of 600 for each question asked for each method. The results are reported in Fig. 9, where the error bars correspond to 95% confidence intervals, computed using the Wilson method (Wilson, 1927). These results support the quantitative results in Fig. 7. Additional details on the user study are provided in App. B.2.3.

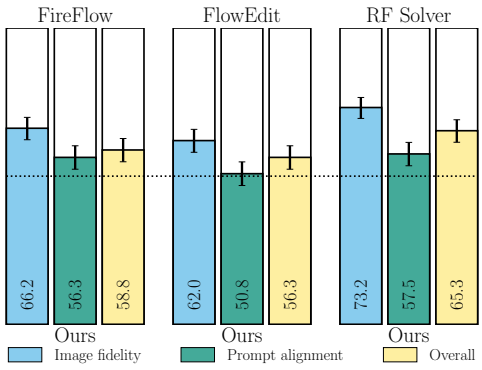

Figure 9: Human perceptual study. The bar plots report the percentages of users that preferred our method over competing methods in (i) image fidelity, (ii) text alignment, and (iii) overall. Error bars show 95% confidence intervals.

6 Conclusions

We presented a zero-order (gradient-free) framework that allows efficient optimization over the initial noise in a flow process while minimizing a loss over the sample generated at the end of the process. We demonstrated the effectiveness of our approach for performing image editing using pre-trained flow models. In particular, extensive comparisons showed that our FlowOpt method achieves SotA performance on both image reconstruction and editing. We note that, similarly to other training-free editing methods, our approach still encounters difficulties in certain settings, like modifying large regions of the image (see App. L). However, taking a broader perspective, we believe that our zero-order framework opens the door for exploiting pre-trained flow-models in diverse applications (*e.g.*, restoration, compression, and personalization) and for diverse modalities (*e.g.*, image, video, and audio). We leave those extensions for future work.

Figure 10: Qualitative comparisons. FlowOpt is the only method to consistently adhere both to target text prompt, and to the original image. Fine details are visible upon zooming in. For instance, the back legs of the zebra in the first row, the posture of the bear in the second row, and the structure of the scene in the last row.

ETHICS STATEMENT

This work builds upon pre-trained generative models, and thus inherits the broader ethical considerations associated with their use. Such models may reflect or amplify societal biases present in

the training data, and their outputs could be misinterpreted or misused in sensitive applications. In addition, our approach involves large-scale flow matching models, which carry the potential risk of being repurposed for harmful or malicious purposes. We emphasize that our contributions are intended solely for advancing research in generative modeling.

Reproducibility statement

We refer to our code repository at https://anonymous.4open.science/r/FlowOpt/. The repository includes the required scripts for running the proposed approach both for image inversion and image editing, for FLUX and SD3.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, 2017. 7

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta mathematicae*, 3(1):133–181, 1922.

Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-flow: differentiating through flows for controlled generation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 3462–3483, 2024. 1, 3

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

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. 9

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. 9

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *Advances in neural information processing systems*, 32, 2019. 2

Nathaniel Cohen, Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Slicedit: zero-shot video editing with text-to-image diffusion models using spatio-temporal slices. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 9109–9137, 2024. 1

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023. 2

Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. Fireflow: Fast inversion of rectified flow for image semantic editing. In *Forty-second International Conference on Machine Learning*, 2025. 1, 3, 7

Noam Elata, Tomer Michaeli, and Michael Elad. PSC: Posterior sampling-based compression. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. 1

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. 2

Stephanie Fu, Netanel Y Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: learning new dimensions of human visual similarity using synthetic data. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 50742–50768, 2023. 9

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. 1, 3, 7

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2328–2337, 2023. 3

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

Seongmin Hong, Kyeonghyun Lee, Suh Yoon Jeon, Hyewon Bae, and Se Young Chun. On exact inversion of dpm-solvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7069–7078, 2024. 1, 3

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. 1, 3, 7

Guanlong Jiao, Biqing Huang, Kuan-Chieh Wang, and Renjie Liao. Uniedit-flow: Unleashing inversion and editing in the era of flow models. *arXiv preprint arXiv:2504.13109*, 2025. 6, 7

Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional gans. In *European Conference on Computer Vision*, pp. 428–447. Springer, 2024.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022. 1

Juil Koo, Chanho Park, and Minhyuk Sung. Posterior distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13352–13361, 2024. 3

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. 7, 8

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022. 7

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024. 8

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*, 2023. 1, 3

Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using ddpm inversion. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 34603–34629, 2024. 1

Ségolène Tiffany Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. Pnp-flow: Plug-and-play image restoration with flow matching. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 768–783, 2018.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020. 1

Peyman Milanfar. Rendition:: Reclaiming what a black box takes away. *SIAM Journal on Imaging Sciences*, 11(4):2722–2756, 2018. 2

Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2063–2072. IEEE, 2025. 3

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. 3

Guy Ohayon, Hila Manor, Tomer Michaeli, and Michael Elad. Compressed image generation with denoising diffusion codebook models. In *Forty-second International Conference on Machine Learning*, 2025. 1

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. 1, 3

Maitreya Patel, Song Wen, Dimitris N Metaxas, and Yezhou Yang. Flowchef: Steering of rectified flow models for controlled generations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15308–15318, 2025. 3, 8

Pexels. Pexels - free stock photos & videos you can use everywhere. https://www.pexels.com/, 2025. 8

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 3

PxHere. PxHere - free images & free stock photos. https://pxhere.com/, 2025. 8

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023. 1

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. 9

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

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. 1, 2

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. 8

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

Dvir Samuel, Barak Meiri, Haggai Maron, Yoad Tewel, Nir Darshan, Shai Avidan, Gal Chechik, and Rami Ben-Ari. Lightning-fast image inversion and editing for text-to-image diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 3

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 9

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. 1, 3, 5

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. 3, 5

Xin Tao, Chao Zhou, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Zero-order reverse filtering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 222–230, 2017. 2, 5, 6

Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 742–749, 2019. 2

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. 1

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021. 2

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. In *Forty-second International Conference on Machine Learning*, 2025. 1, 3, 7

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. 9

Xiaofeng Yang, Chen Cheng, Xulei Yang, Fayao Liu, and Guosheng Lin. Text-to-image rectified flow as plug-and-play priors. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

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. 7