Oscillation Inversion: Training-Free Image and Video Enhancement Through Oscillated Latents in Large Flow Models

Yan Zheng[1], Zhenxiao Liang[1], Xiaoyan Cong[2], Yi Yang[3], Lanqing Guo[1], Yuehao Wang[1], Peihao Wang[1], Zhangyang Wang[1]

[1]University of Texas at Austin
[2]Brown University
[3]The University of Edinburgh

Abstract

We explore the oscillatory behavior observed in inversion methods applied to large-scale flow models, including text-to-image and text-to-video. By employing an augmented fixed-point-inspired iterative approach to invert real-world images, we observe that the solution does not achieve convergence, instead oscillating between distinct clusters. Through both experiments on synthetic data, text-to-image and text-to-video, we demonstrate that these oscillating clusters exhibit notable semantic coherence. We offer theoretical insights, showing that this behavior arises from oscillatory dynamics in flow models. Building on this understanding, we introduce a simple and fast distribution transfer technique that facilitates training-free image and video editing/enhancement. Furthermore, we provide quantitative results demonstrating the effectiveness of our method on tasks such as image enhancement, editing, and reconstruction. Notably, our approach enables the transformation of image-only enhancers and editors into lightweight, video-capable tools—without additional training—highlighting its practical versatility and impact.

Code —
https://github.com/VITA-Group/Oscillation-Inversion

1 Introduction

Recently, large flow models (Wang et al. 2024; Labs 2024; Lipman et al. 2022, 2024) have demonstrated exceptional performance in generating high-quality images and videos with rapid sampling. However, the underlying latent structure of rectified flow-based models differs fundamentally from the structure of DDPMs (Ho, Jain, and Abbeel 2020). This distinction makes previous inversion techniques, such as DDIM inversion (Song, Meng, and Ermon 2020), and editing methods like SDEdit (Meng et al. 2021) less viable. Therefore, adopting a new perspective for understanding and navigating the latent space of these flow-based models is essential for diverse tasks.

When attempting to invert real-world images using fixed-point iteration methods in flow-based models, we observe that the sequence of iterates does not converge to a single point, but instead oscillates between several clusters. This behavior stands in stark contrast to the fixed-point methods used in DDIM (Pan et al. 2023; Garibi et al. 2024), which ensuring smooth convergence to a single, stable solution. Further analysis reveals that these clusters exhibit coherence, indicating that the oscillations are not merely random fluctuations but may reflect underlying structural properties of the latent space. Rather than being a limitation, we believe these oscillations could introduce greater insight for image and video inversion and editing tasks.

To investigate this phenomenon, we first propose **Oscillation Inversion**, a method that uses fixed-point iteration to directly establish a one-to-one mapping between noisy latents at an intermediate timestep and the corresponding encoded image latent. The inverted latents are discovered to oscillate among several clusters. We further extend our approach in two key directions for broader downstream applications: **1) Group Inversion**: Instead of inverting a single image, we simultaneously invert a group of images, enabling blending across samples and automatically discovering on-manifold high-quality latents; and **2) Video Extension**: We generalize our method to large-scale flow models for text-to-video generation, where a training-free distribution transfer technique effectively filters out temporally inconsistent clusters, yielding coherent and stable video editing and enhancement results.

The main contributions of this work are as follows:

- We discovered an oscillatory phenomenon in the fixed-point method on large flow models and provide theoretical insights through proofs under mild assumptions and aligned experiments on synthetic data.
- We introduce Group Inversion and extend our framework to T2I and T2V editing/enhancement, demonstrating that our approach can leverage the inherent oscillatory dynamics to enhance both image and video editing tasks.
- Extensive experiments on various downstream tasks, including image and video editing and enhancement, validate our theoretical insights and demonstrate significant improvements in both perceptual quality and data fidelity.

2 Related Works

Flow Model. Diffusion models (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2022) generate data by a

stochastic differential equation (SDE)-based denoising process and probability flow ordinary differential equations (ODE) (Song et al. 2020; Lipman et al. 2022; Salimans and Ho 2022; Song et al. 2023) improves sampling efficiency by formulating the denoising process into a ODE-based process. However, probability flow ODE-based methods suffer from the computational expense of denoising via numerical integration with small step sizes. To address these issues, some simulation-free flow models have emerged, e.g. flow matching (Lipman et al. 2022) and rectified flow (Liu, Gong, and Liu 2022). Flow matching introduces a training objective for continuous normalizing flows (Chen et al. 2018; Zheng et al. 2022) to regress the vector field of a probability path. Flow learns a transport map between two distributions through constraining the ODE to follow the straight transport paths. Since the latent structure of flow models differs fundamentally from the layered manifold structure of Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020), it is valuable to explore the intrinsic characteristics of the flow models' latent space.

Diffusion-based Inversion. The rise of diffusion models (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2022) has unlocked significant potential of inversion methods for real image editing, which are primarily categorized into DDPM (Ho, Jain, and Abbeel 2020) based (Wu and De la Torre 2023; Huberman-Spiegelglas, Kulikov, and Michaeli 2024) and DDIM (Song, Meng, and Ermon 2020) based methods (Pan et al. 2023; Garibi et al. 2024; Li et al. 2024; Meiri et al. 2023; Zheng and Wu 2024). While DDPM-based methods yield impressive editing results, they are hindered by their inherently time-consuming and stochastic nature, due to the random noise introduced across a large number of inversion steps (Wu and De la Torre 2023; Huberman-Spiegelglas, Kulikov, and Michaeli 2024). DDIM-based methods utilize the DDIM sampling strategy to enable a more deterministic inversion process, substantially reducing computational overhead and time. However, the linear approximation behind DDIM often leads to error propagation, resulting in reconstruction inaccuracy and content loss, especially when classifier-free guidance (CFG) is applied (Mokady et al. 2023). Recent approaches, (Wallace, Gokul, and Naik 2023; Mokady et al. 2023; Pan et al. 2023; Miyake et al. 2023; Han et al. 2023; Hong et al. 2024), address these issues by aligning the diffusion and reverse diffusion trajectories through the optimization of null-text tokens (Mokady et al. 2023) or prompt embeddings (Han et al. 2023; Miyake et al. 2023). EDICT (Wallace, Gokul, and Naik 2023) and BDIA (Zhang, Lewis, and Kleijn 2023) introduce invertible neural network layers to enhance computational efficiency and inversion accuracy, though these methods suffer from notably longer inversion times. To tackle this, recent works (Meiri et al. 2023; Pan et al. 2023) (Garibi et al. 2024; Li et al. 2024) have adopted fixed-point iteration for each inversion step, mitigating numerical error accumulation and ensuring smooth convergence to a single, stable solution. Interestingly, when applied to rectified flow-based methods, the sequence of fixed-point iterates oscillates between several semantically meaningful clusters, presenting significant potential for downstream applications. RF-Inversion (Rout et al. 2024) proposes a dynamic optimal control approach that leverages the optimal mixture path from a Gaussian distribution. FlowEdit (Kulikov et al. 2024) introduces an inversion-free interpolation method, connecting source and target distributions by reusing velocities computed during inversion. However, these methods primarily focus on semantic editing—since image content and layout are mostly determined in the early diffusion steps—while our method targets low-level editing tasks such as enhancement, relighting, and recoloring for both images and videos, all while preserving the input subject and content.

3 Oscillation Inversion

3.1 Preliminary: Rectified Flow

Rectified flow (Liu, Gong, and Liu 2022) is a novel generative approach that facilitates smooth transitions between two distributions, denoted π_0 for noise and π_1 for target, by solving ordinary differential equations (ODEs). Specifically, for $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$, the transition between x_0 and x_1 is defined through an interpolation given by $X_t = (1-t)X_0 + tX_1$ for $t \in [0,1]$. (Liu, Gong, and Liu 2022) demonstrated that, starting from $Z_0 \sim \pi_0$, the following ODE can be used to obtain a trajectory that preserves the marginal distribution of Z_t at any given time t:

$$\begin{aligned} \frac{dZ_t}{dt} &= v^X(Z_t, t), \\ \text{where } v^X(x,t) &:= \mathbb{E}[X_1 - X_0 \mid X_t = x]. \end{aligned} \tag{1}$$

The solution of v^X in Eq. (1) is obtained by optimizing the following loss via stochastic coupling sampling $(X_0, X_1) \sim (\pi_0, \pi_1)$ and $t \sim \text{Uniform}([0,1])$,

$$v^X = \arg\min_v \mathbb{E}\left[\left\|(X_1 - X_0) - v(X_t, t)\right\|^2\right]. \tag{2}$$

3.2 Method

In this section, we first formulate the inversion problem for flow-based models in Section 3.2. Next, we report the oscillation phenomena observed during fixed-point iterations for the flow inversion problem in Section 3.2. We then construct an augmented fixed-point method to control these phenomena in Section 3.2. This method enables both image and video editing as well as image enhancement, which is benchmarked in Section 5.1. The concept behind our approach is illustrated in Figure 1.

Inversion Problem In practice, the large flow model in the context of generative modeling operates within the latent space of a Variational Autoencoder (VAE) (Kingma 2013), utilizing an encoder $E : \mathbb{R}^d \to \mathbb{R}^n$ and a decoder $D : \mathbb{R}^n \to \mathbb{R}^d$. The sampling process begins from Gaussian noise $z_T \sim \mathcal{N}(0, \mathbf{I})$, and the latent variable is progressively refined through a sequence of transformations. The forward generative process is defined by the following iterative formula starting from $t = T$ all the way back to $t = 0$:

$$z_{t-1} = z_t + (\sigma_{t-1} - \sigma_t)\, v_\theta(z_t, \sigma_t), \tag{3}$$

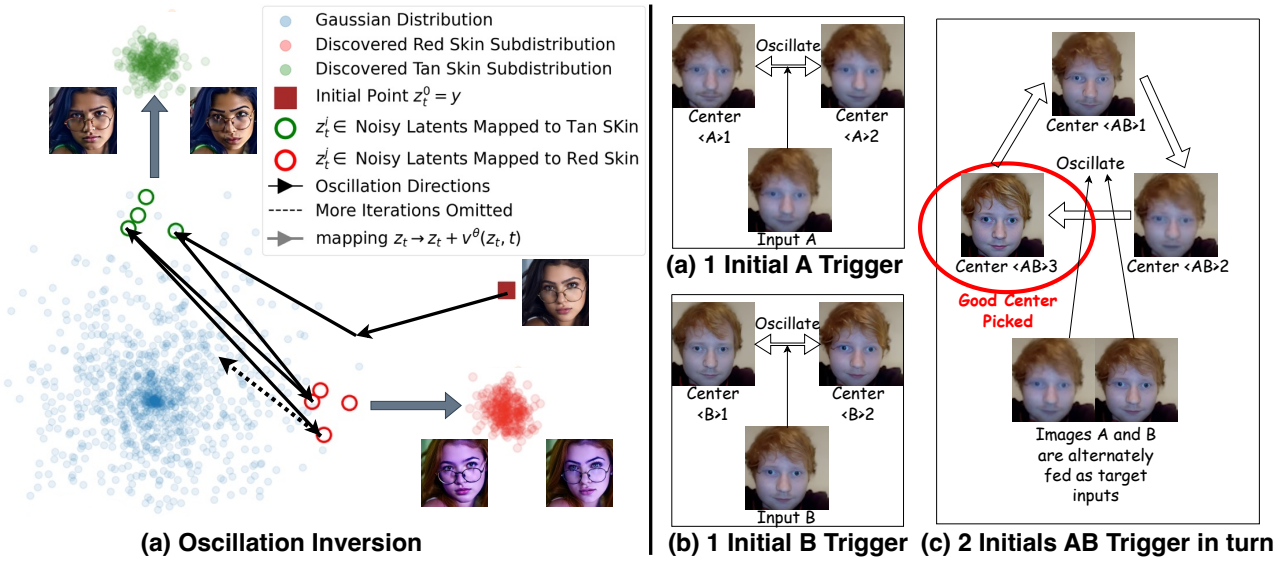

Figure 1: In the left figure (a), fixed-point iteration causes oscillation, leading to subdomains with opposite features in the case of the brown-skinned girl, resulting in more tan and red tones. In the right figure (b), we demonstrate how this oscillation can be extended by using a group input consisting of more than one image. This phenomenon can push a discovered latent (center 3) back toward a high-quality image distribution by factoring out the undesirable components (centers 1 and 2).

where $v_\theta(z_t, \sigma_t)$ represents the learned velocity field parameterized by a transformer with weights θ, and σ_t is a monotonically increasing time step scaling function depending on time t with $\sigma_0 = 0$ and $\sigma_T = 1$. Here, T denotes the total number of discredited timesteps. The final latent variable, z_0, is the output ready for decoding.

The inversion problem involves seeking for the initial noise z_T given an observed pixel image I with corresponding latent encoding $y \in \mathbb{R}^d$, such that generating from z_T using the flow model described above allows us to either reconstruct y or apply desired modifications to it.

However, the gradual process of sampling from Gaussian noise to the original y diminishes the advantage of GAN-like one-step mappings for direct latent space optimization. To address this, unlike tackling with the initial noise at $t = T$, we introduce the assumption that, at a selected intermediate timestep t_0, there exists a direct one-step mapping from the noisy latent at timestep t_0 to the clean latent at timestep 0 via a "jumping" transformation.

More specifically, assuming y is the latent code to recover, we aim to figure out the *intermediate* latent code z_{t_0} satisfying

$$z_{t_0} + (\sigma_0 - \sigma_{t_0})\, v_\theta(z_{t_0}, \sigma_{t_0}) = y. \tag{4}$$

Solving Eq. (4) is non-trivial, and in the following sections, we will describe how we find a set of approximated solutions using iterative method and analyze its oscillating properties.

Oscillations Inversion To address the inversion problem, we employ a fixed-point iteration method to approach the solution of Eq. (4). Instead of directly seeking a point z_{t_0} such that applying the one-step generative process as described in the left side of (4) yields the target latent y, we define an iterative process that refines our approximation of the inverted latent code. We define the fixed-point iteration as:

$$z_{t_0}^{(k+1)} = y - (\sigma_0 - \sigma_{t_0}) v_\theta(z_{t_0}^{(k)}, \sigma_{t_0}), \tag{5}$$

with the initial condition $z_{t_0}^{(0)} = y$.

The sequence $\{z_{t_0}^{(k)}\}_{k=0}^{\infty}$ represents successive approximations of the inverted latent code at timestep t.

As shown in Figure 2, rather than converging to a single point as suggested by Banach's Fixed-Point Theorem (Banach 1922), we empirically observed that the sequence $\{z_{t_0}^{(k)}\}_{k=0}^{\infty}$ generally oscillates among several clusters in the latent space. Each cluster corresponds to a semantically concentrated region that shares similar low-level features. This oscillatory behavior can be harnessed to explore different variations of the input image, providing a richer inversion that captures multiple aspects of the data.

Group Inversion Building upon the fixed-point method, we propose an augmented fixed-point approach—referred to as **group inversion**—which induces more stable and controllable oscillatory behavior by simultaneously inverting a group of images in a periodic fashion. Suppose we obtain their corresponding latent encodings $y_1, \ldots, y_m$ from a collection of images $I_1, \ldots, I_m$ using the VAE encoding. We could perform the iteration on the group:

$$z_{t_0}^{(k+1)} = y_{(k \bmod m)} - (\sigma_0 - \sigma_{t_0}) v_\theta(z_{t_0}^{(k)}, \sigma_{t_0}), \tag{6}$$

with initial conditions $z_{t_0}^{(0)} = y_{(1 \bmod m)}$. By inverting the images together, we enable interactions between their latent representations during the iteration process.

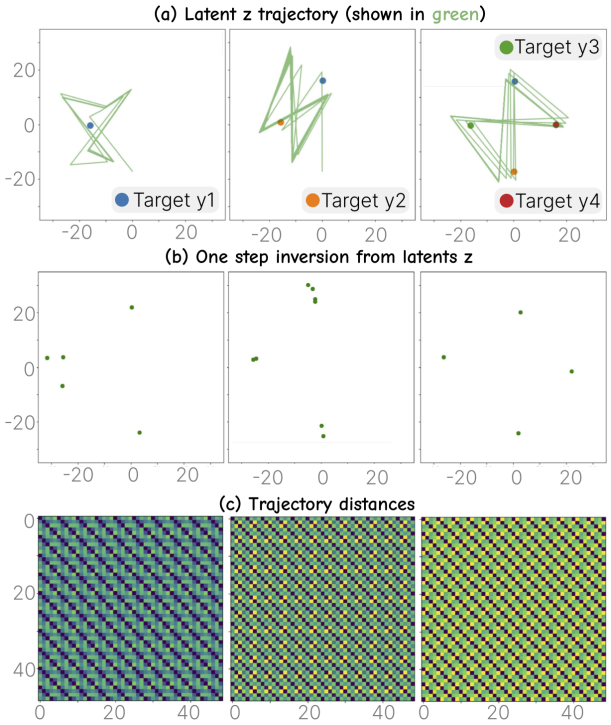

Figure 2: By training flow matching on the distribution shown in Fig. 4(a), we demonstrate that the oscillatory inversion phenomenon in large flow models closely aligns with that observed in toy data. Each column represents experiments with 1, 2, and 4 initial ys, respectively. Row (a) shows $z_{t_0}^{(k+1)}$. Row (b) shows one-step predictions. Row (c) shows trajectory distances.

Our experimental findings indicate that this collective inversion not only induces additional periodic oscillatory clusters in the latent space, but the number of clusters also tends to match the periodic number. Rather than straying off the manifold and degrading latent quality, this oscillatory process expels undesirable components while pushing left ones onto the real data distribution manifold. This behavior greatly facilitates downstream tasks such as enhancement and editing.

For example, Figure 2 illustrates this phenomenon. We first trained a rectified flow model on a toy distribution that models transitions from a large central Gaussian to a mixture of four smaller Gaussians (Figure 4(a)). As the number of periodic groups increases—from a single input y_1 in the first column to multiple inputs in the second and third columns—the trajectory becomes more regular, and the resulting clusters more concentrated.

Generalization: From Large Text-to-Image Model to Text-to-Video Model We now explain how the same phenomenon manifests in large flow models for both text-to-image and text-to-video generation, demonstrating the strong generalization ability of our discovery and proposed technique.

T2I Flow Model: In Fig. 1(c), two low-quality images, labeled A and B, are input as a group for periodic inversion in the order A, A, B. This process triggers three clusters in the latent space. Clusters 1 and 2 correspond to low-quality centers that are expelled from the real picture manifold, while cluster 3 represents the high-quality center that is pushed back into the manifold. As a result, the output exhibits quality that is even higher than that of the original images A and B. We also validate this technique on a larger-scale experiment in Section 5.1. Similarly, when A represents the original image and B represents the stroke-edited image, the beneficial cluster yields a harmonized editing result that appears natural with effects such as make-up or relighting. We refer readers to the Appendix for this experiment.

T2V Flow Model: In Fig. 3, we demonstrate that using this training-free technique, we can transform a per-frame image editor/enhancer into a video editor/enhancer. Specifically, we showcase a per-frame face identity switch in (a), a stroke edit for eye shadow in (b), and a text-based edit for mouth lipstick in (c). We treat the original video input as A and the per-frame edited, inconsistent video as B. The group input consists of A, A, and B. Surprisingly, the temporally inconsistent cluster is factored out, leaving one cluster that exhibits significant temporal consistency, which serves as the training-free smoothing result.

4 Analysis

In this section, consider a simplified mixed-Gaussian setup trained with an ideal model within the rectified flow framework to explain the oscillatory behavior observed in the empirical experiments of fixed-point inversion. We aim to show that, under certain reasonable assumptions: There does not exist any *stable* fixed point solution z_{t_0} for Eq. 4. The iterative formula Eq. 5 would *randomly* oscillate between the clusters determined by the target Gaussian mixture distribution. The group inversion introduced in Section 3.2 generalizes the concept of a non-converging fixed point, resulting in a periodic dynamic system with multiple solutions within this framework. To align with the notation of the "Flux" model's time schedule and the notation introduced in Section 3.2, we use a flipped notation here: we denote the source domain as π_1 and the target domain, which the generation is heading towards, as π_0. The pure noise distribution π_1 is assumed to be the d-dimensional standard normal distribution, and the target distribution π_0 is a Gaussian mixture distribution given by components $\mathcal{N}(\mu_c, \Sigma_c)$ and corresponding coefficient ϕ_c for $c = 1, \ldots, c_0$. Equivalently, the PDF of π_0 is determined by

$$\mathbf{P}_{\pi_0} = \sum_{c=1}^{c_0} \phi_c \cdot \mathbf{P}_{\mathcal{N}(\mu_c, \Sigma_c)}.$$

Moreover, we consider the ideal case where v^X is the precise solution to Eq. 2, which can be written as

$$v^X(x,t) = \mathbb{E}[X_0 - X_1 \mid (1-t)X_1 + tX_0 = x] \quad (7)$$

as discussed in (Liu, Gong, and Liu 2022). Here $X_t \sim \pi_t$ for $0 \le t \le 1$, while the derived random variable $X_t = tX_1 + (1-t)X_0$ is subject to another Gaussian mixture distribution.

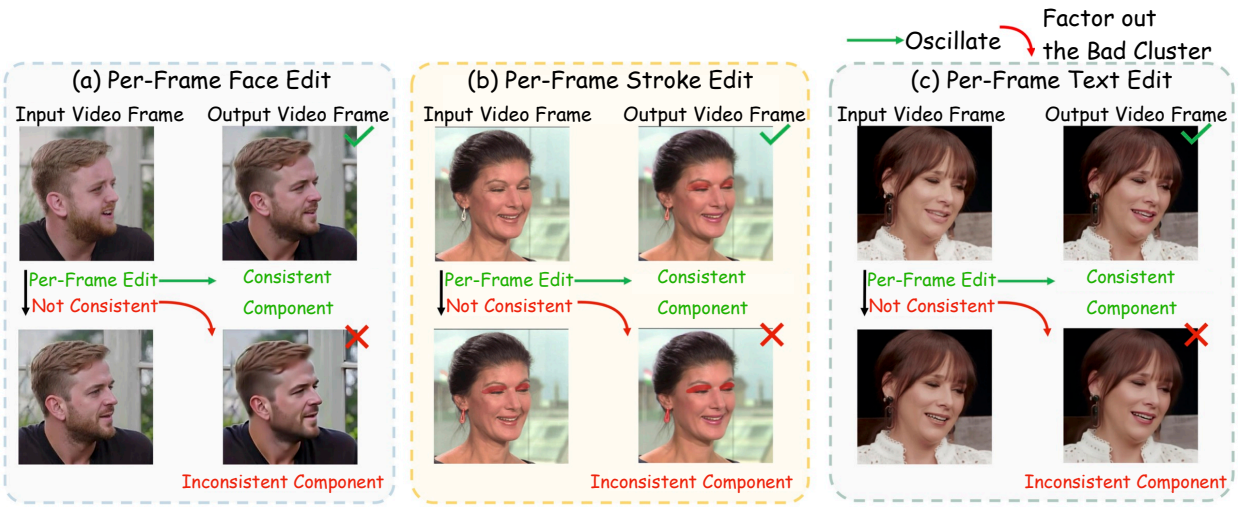

Figure 3: Training-free video editing via oscillation inversion. Each group illustrates a different per-frame editing task: (a) face identity modification, (b) stroke-based makeup editing, and (c) text-guided makeup editing. Inside each group, the two left images are input. The top-right image shows the optimized output, where unstable clusters (inconsistent components) are removed. The bottom-right image highlights these inconsistencies, which arise from per-frame edits and lack temporal coherence. Our oscillation-based optimization enhances correct information, ensuring stable and temporally consistent video editing.(Please check videos in the supplementary material.)

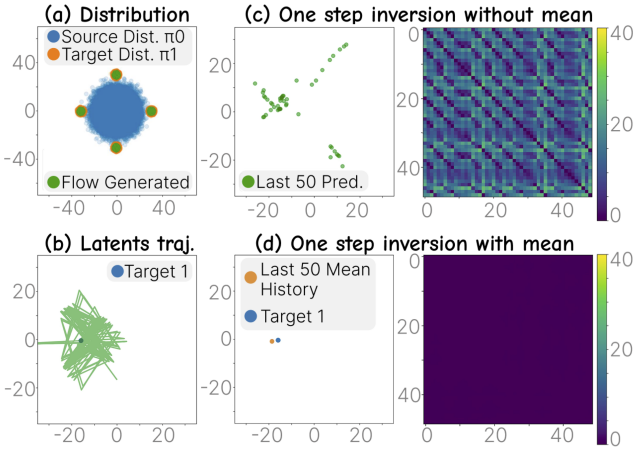

Figure 4: a) Toy flow matching setting. b), c), d) Results on our toy experiment show that taking the average among clusters achieves near-exact one-step inversion, as demonstrated in Theorem 2.

Let $\gamma = t_0$ be the intermediate timestamp that we are interested in. The inversion problem Eq. 4 is equivalent to figuring out the fixed points z of function $f(z; y, \gamma)$, which is defined by

$$\begin{aligned} f(z; y, \gamma) &:= y - (\sigma_0 - \sigma_\gamma) v_\theta(z, \sigma_\gamma) \\ &= y + \gamma v_\theta(z, \sigma_\gamma), \end{aligned} \tag{8}$$

in which we assume $\sigma(t)$ is set to trivial t without losing generality. In the rest of this section we may write $f(z)$ with ignoring y and γ whenever the context is clear.

By fixed point method, we could seek fixed points of Eq. 8 with iterative formula

$$z^{(k+1)} := f(z^{(k)}) \tag{9}$$

with some initial $z^{(0)}$.

4.1 Analysis of Fixed Points

In this section, we present a series of theorems that collectively demonstrate the instability of the function $f(z)$ near its roots under certain mild conditions. Specifically, we examine the magnitude of the spectral norm of the Jacobian $\|J_f(z)\|$ at the root z.

First, we establish a foundational result that relates $f(\cdot)$ to the conditional expectation involving the source and target distributions.

Lemma 1. *Let π_1 be the source distribution and π_0 be the target distribution that the rectified flow transports, following the notations from Section 4. As in Eq. 8, the function f is given by $f(z) = y + \gamma v^X(z, \gamma)$ where $v^X(x, t) = \mathbb{E}[X_0 - X_1 \mid X_t = x]$. Then we provide the explicit form of $v^X(x, t)$:*

$$v^X(x,t) = \frac{1}{t^d \cdot \pi_t(x)} \int_z \pi_0(z) \cdot \pi_1\left(\frac{x-(1-t)z}{t}\right) \cdot z \, dz, \tag{10}$$

where $\pi_t(x)$ is the probability density function of X_t.

Next, we explore the local convergence behavior of $f(z)$ when the target distribution is a mixture of Gaussians and the source distribution is a standard Gaussian as displayed in Fig 4(a). This setting allows us to analyze the properties of the Jacobian of $f(z)$.

Theorem 1 (Informal)**.** *Suppose PDF of source noise distribution $\pi_0(z)$ is d-dimensional standard Gaussian density*

and the target distribution $\pi_1(x)$ is a mixture of Gaussian densities centered around multiple centers.

Let z_0 be one of the fixed points for $f(z)$, we show that under mild conditions the Jacobian $J_f(z_0) = \nabla_z v^X(z_0, t)$ has at least one singular value greater than 1, or equivalently, $\|J_f(z_0)\| > 1$ where $\|\cdot\|$ represents the matrix operator norm, implying that the iterative process will never converge to z_0 in a stable manner.

The model assumption in Theorem 1 is reasonable when modeling the process from a pure random noise distribution to one of several potential clusters with each cluster corresponding to a distinct image class. This result implies that $f(z)$ exhibits instability near its roots due to the large singular values of its Jacobian. The instability leads to multiple roots forming compact clusters, each associated with an attraction field. Points within this field are drawn toward the corresponding cluster but never converge to its center.

To illustrate the impact of this instability on iterative methods, we present the following theorem.

Theorem 2 (Informal)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable function, and let $y \in \mathbb{R}^n$. Suppose that the equation $f(z) = z$ has a unique solution $z^* \in \mathbb{R}^n$.*

Consider the fixed-point iteration defined by

$$z_{i+1} = f(z_i),$$

with an initial point z_0 close to z^*.

Then under mild conditions, the sequence $\{z_i\}$ as defined by Eq. (9) oscillates between two compact clusters with means z' and z'', which are significantly different. Furthermore, the average of these cluster means approximates the solution z^*: $(z' + z'')/2 \approx z^*$

This theorem demonstrates that due to the instability near the root, the iterative process does not converge directly to z^* but instead oscillates between clusters. The average of these clusters approximates the exact solution of the inversion problem in Equation 4.

In summary, these theorems collectively establish that under mild conditions, $f(z)$ not only exhibits instability near its fixed points but also possesses a set of clusters derived from the unstable fixed point. Each cluster is compact, with its mean approximating the underlying solution, and is associated with an attraction field that draws points toward it without stable convergence to its center.

5 Applications

Experiment Settings All of our image experiments are based on the 'black-forest-labs/FLUX.1-schnell' checkpoint. We run the experiments on a single A6000 GPU with 48GB of memory. All images are cropped and resized to 512×512 pixels. The oscillation inversion consistently runs for 20 iterations, taking 8.74 seconds per image. All of our video experiments are based on the 'HunyuanVideo' checkpoint.

5.1 Enhancement and Editing

We use real-world degraded (low-quality) images for the qualitative assessment and apply simulated noise, blur, and low-resolution degradation to the CelebA validation dataset (Liu et al. 2015) for the quantitative analysis. For the latter, we follow the protocols of previous studies, using metrics like PSNR and LPIPS to measure performance.

We compare our approach against existing image restoration and enhancement methods, including BlindDPS, DIP, GDP, and BIRD (Chung et al. 2023; Ulyanov, Vedaldi, and Lempitsky 2018; Fei et al. 2023; Chihaoui, Lemkhenter, and Favaro 2024). Positioned as a post-processing image enhancement technique, our method is compared to Piscart, chosen for its strong identity preservation and efficient batch processing. Figure 5 showcases visual examples of our approach restoring richer details in real-world degraded images, such as noise and blur, outperforming existing methods. Notably, we observed consistent distribution oscillation during recovery tasks, indicating our inversion method effectively introduces a distribution transfer mechanism. In large-scale experiments, we identified the best-performing distribution center for each task through manual inspection, applying it consistently across the dataset. We direct readers to Section 3.2 for a detailed explanation of how group inversion is applied to image and video editing tasks, and to the appendix for quantitative metrics and Supplementary materials for additional results including makeup transfer.

5.2 Transfer Light-weight Image Enhancer into Video Enhancer for free

We conducted two types of degradation experiments in input videos, namely Gaussian blur and down-sample of different levels on VFHQ dataset(Xie et al. 2022). As shown in Figure 6, each degraded video was first enhanced frame by frame using Topaz Image Enhancer and reassembled as a video. Due to the lack of temporal consistency, these results exhibit visible artifacts (e.g., phantom eyeglasses and iris-color distortions). Our method fuses the Topaz outputs and the degraded input following 3.2, preserving the enhanced sharpness from Topaz while maintaining the coherent motion and appearance of the original frames. We report the different metrics under Gaussian Blur with $\sigma = 4$ in Table 1 due to space limit and refer extensive quanitative and visual results in appendix and supplementary materials.

Method	flow_L1	flicker	T-LPIPS	CLIP_TSC
Baseline	5.0897	0.1317	0.0215	0.9910
Ours	5.1495	0.1376	0.0179	0.9922

Table 1: Average metrics across all test clips. Lower is better for flow_L1, flicker, and T-LPIPS; higher is better for CLIP_TSC.

5.3 Reconstruction

We performed an evaluation on image reconstruction task on the COCO Validation set, utilizing the default captions as text prompts. We followed the same settings in (Pan et al. 2023). We selected the existing inversion methods, including DDIM (Song, Meng, and Ermon 2020), NULL Text (Mokady et al. 2023) , AIDI (Pan et al. 2023),

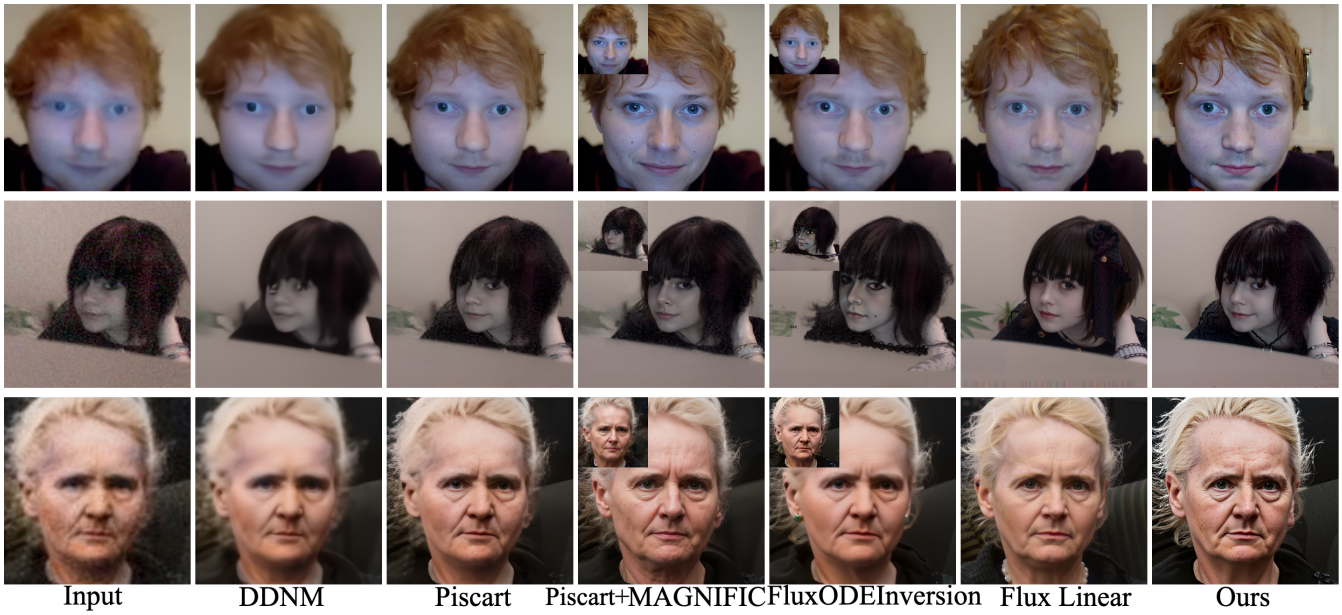

Figure 5: Image Enhancer by Fusing Two Lower-quality Distributions. Top-left insets for MAGNIFIC and FluxODEInversion show direct-input ablations, which perform worse than using Piscart as an intermediate and are shown as smaller crops.

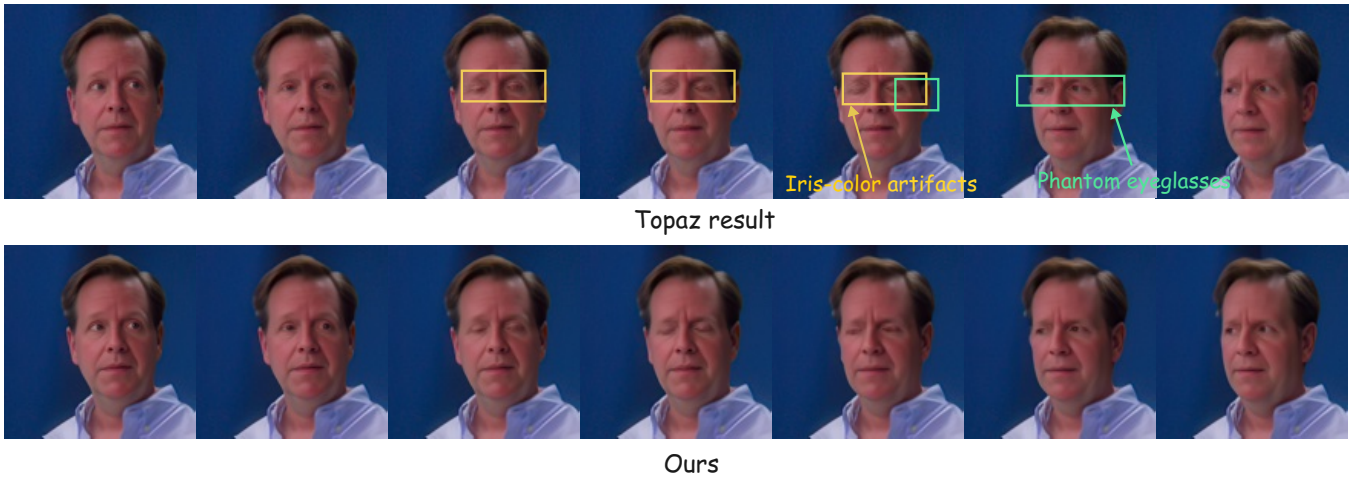

Figure 6: Visual comparison over time. The top row shows per-frame results from the Topaz image enhancer, while the bottom row shows our method. Across successive frames, our method suppresses phantom eyeglasses and iris-color distortions while preserving temporal coherence and detail.

ReNoise (Garibi et al. 2024) as competing methods. We refer readers to Appendix for reconstruction quantitative results.

6 Ablations

Although all our results are presented using "Flux-schnell," a distilled version, we observe the same phenomena in "Flux-dev" as well. Additionally, the video experiment is conducted on "Hunyuan-Video," which is not distilled and is trained using flow matching—a generalized framework. We do not observe the same phenomenon in SD3. Although SD3 belongs to the flow family, they employ different training strategies: SD3 adjusts flow weights at intermediate timesteps. Nonetheless, validation on state-of-the-art video and image generation models based on large flow architectures strongly supports our discovery.

7 Conclusions

We introduce Oscillation Inversion, a training-free editing framework for flow-based diffusion models, validated on image and video checkpoints. Our augmented fixed-point method leverages oscillatory behavior for on-manifold latent discovery, yielding high-quality, natural results. Theoretical insights and experiments confirm its effectiveness.

References

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

Chen, R. T.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. K. 2018. Neural ordinary differential equations. *NIPS*.

Chihaoui, H.; Lemkhenter, A.; and Favaro, P. 2024. Blind Image Restoration via Fast Diffusion Inversion. *arXiv preprint arXiv:2405.19572*.

Chung, H.; Kim, J.; Kim, S.; and Ye, J. C. 2023. Parallel diffusion models of operator and image for blind inverse problems. In *CVPR*.

Fei, B.; Lyu, Z.; Pan, L.; Zhang, J.; Yang, W.; Luo, T.; Zhang, B.; and Dai, B. 2023. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*.

Garibi, D.; Patashnik, O.; Voynov, A.; Averbuch-Elor, H.; and Cohen-Or, D. 2024. ReNoise: Real Image Inversion Through Iterative Noising. *arXiv preprint arXiv:2403.14602*.

Han, L.; Wen, S.; Chen, Q.; Zhang, Z.; Song, K.; Ren, M.; Gao, R.; Stathopoulos, A.; He, X.; Chen, Y.; et al. 2023. Improving Tuning-Free Real Image Editing with Proximal Guidance. *arXiv preprint arXiv:2306.05414*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NIPS*.

Hong, S.; Lee, K.; Jeon, S. Y.; Bae, H.; and Chun, S. Y. 2024. On Exact Inversion of DPM-Solvers. In *CVPR*.

Huberman-Spiegelglas, I.; Kulikov, V.; and Michaeli, T. 2024. An Edit Friendly DDPM Noise Space: Inversion and Manipulations. arXiv:2304.06140.

Kingma, D. P. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Labs, B. F. 2024. Flux. https://github.com/black-forest-labs/flux. Accessed: 2024-09-24.

Li, R.; Li, R.; Guo, S.; and Zhang, L. 2024. Source Prompt Disentangled Inversion for Boosting Image Editability with Diffusion Models. *arXiv preprint arXiv:2403.11105*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Lipman, Y.; Havasi, M.; Holderrieth, P.; Shaul, N.; Le, M.; Karrer, B.; Chen, R. T.; Lopez-Paz, D.; Ben-Hamu, H.; and Gat, I. 2024. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*.

Meiri, B.; Samuel, D.; Darshan, N.; Chechik, G.; Avidan, S.; and Ben-Ari, R. 2023. Fixed-point Inversion for Text-to-image diffusion models. *arXiv preprint arXiv:2312.12540*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

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

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *CVPR*.

Pan, Z.; Gherardi, R.; Xie, X.; and Huang, S. 2023. Effective real image editing with accelerated iterative diffusion inversion. In *ICCV*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

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

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *NIPS*.

Salimans, T.; and Ho, J. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

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

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

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *CVPR*.

Wallace, B.; Gokul, A.; and Naik, N. 2023. Edict: Exact diffusion inversion via coupled transformations. In *CVPR*.

Wang, C.; Guo, Z.; Duan, Y.; Li, H.; Chen, N.; Tang, X.; and Hu, Y. 2024. Target-Driven Distillation: Consistency Distillation with Target Timestep Selection and Decoupled Guidance. *arXiv preprint arXiv:2409.01347*.

Wu, C. H.; and De la Torre, F. 2023. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*.

Xie, L.; Wang, X.; Zhang, H.; Dong, C.; and Shan, Y. 2022. Vfhq: A high-quality dataset and benchmark for video face

super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 657–666.

Zhang, G.; Lewis, J. P.; and Kleijn, W. B. 2023. Exact diffusion inversion via bi-directional integration approximation. *arXiv preprint arXiv:2307.10829*.

Zheng, Y.; and Wu, L. 2024. InverseMeetInsert: Robust Real Image Editing via Geometric Accumulation Inversion in Guided Diffusion Models. *arXiv preprint arXiv:2409.11734*.

Zheng, Y.; Wu, L.; Liu, X.; Chen, Z.; Liu, Q.; and Huang, Q. 2022. Neural volumetric mesh generator. *arXiv preprint arXiv:2210.03158*.