Multi-turn Consistent Image Editing

Zijun Zhou[1,3] Yingying Deng[2*] Xiangyu He[3] Weiming Dong[3] Fan Tang[1†]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Science and Technology Beijing
[3]MAIS, Institute of Automation, Chinese Academy of Sciences

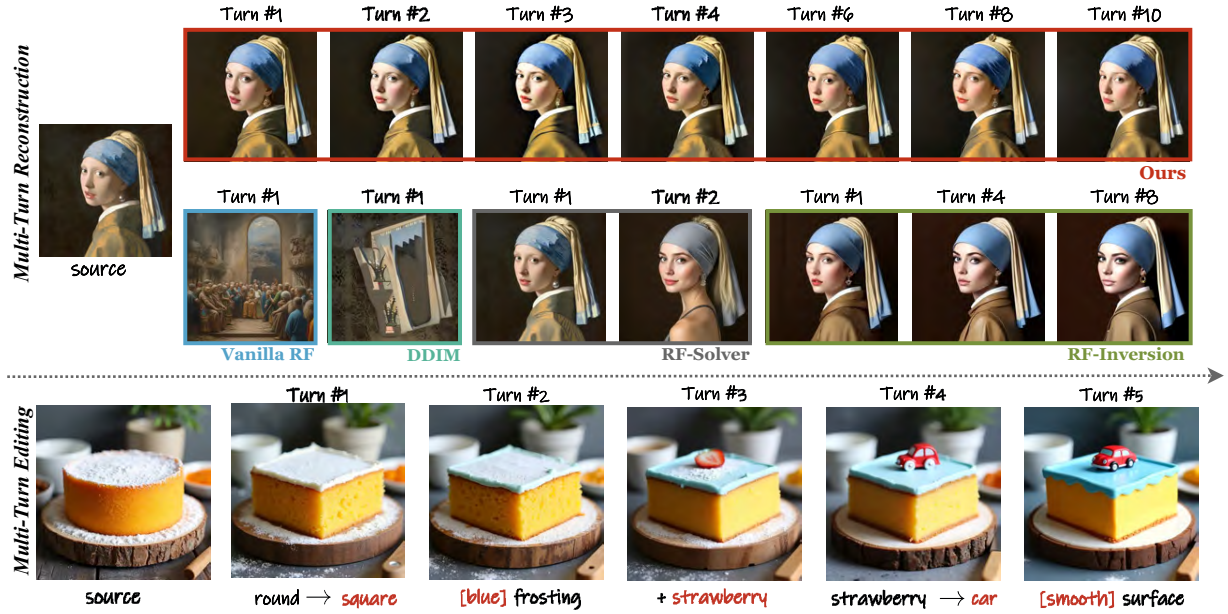


Figure 1. Our method efficiently preserves the original image's features during multi-turn image reconstruction. Additionally, it enables flexible editing capabilities in multi-turn editing tasks, providing the user with an iterative editing framework.

Abstract

Many real-world applications, such as interactive photo retouching, artistic content creation, and product design, require flexible and iterative image editing. However, existing image editing methods primarily focus on achieving the desired modifications in a single step, which often struggles with ambiguous user intent, complex transformations, or the need for progressive refinements. As a result, these methods frequently produce inconsistent outcomes or fail to meet user expectations. To address these challenges, we propose a multi-turn image editing framework that enables users to iteratively refine their edits, progressively achieving more satisfactory results. Our approach leverages flow matching for accurate image inversion and a dual-objective Linear Quadratic Regulators (LQR) for stable sampling, effectively mitigating error accumulation. Additionally, by analyzing the layer-wise roles of transformers, we introduce a adaptive attention highlighting method that enhances editability while preserving multi-turn coherence. Extensive experiments demonstrate that our framework significantly improves edit success rates and visual fidelity compared to existing methods. The code is available at: https://zhouzj-dl.github.io/Multi-turn_Consistent_Image_Editing/.

1. Introduction

Current image editing methodologies often strive for a single-step editing solution that perfectly aligns with a given textual prompt. This paradigm, however, proves inadequate for practical applications like product design, where user specifications are often inherently ambiguous and necessitate progressive refinement. A more effective framework should incorporate iterative editing capabilities, enabling users to sequentially refine outputs through multiple edit-

*Corresponding author: Yingying Deng, email: dyy15@outlook.com
†Project lead: Fan Tang, email: tfan.108@gmail.com

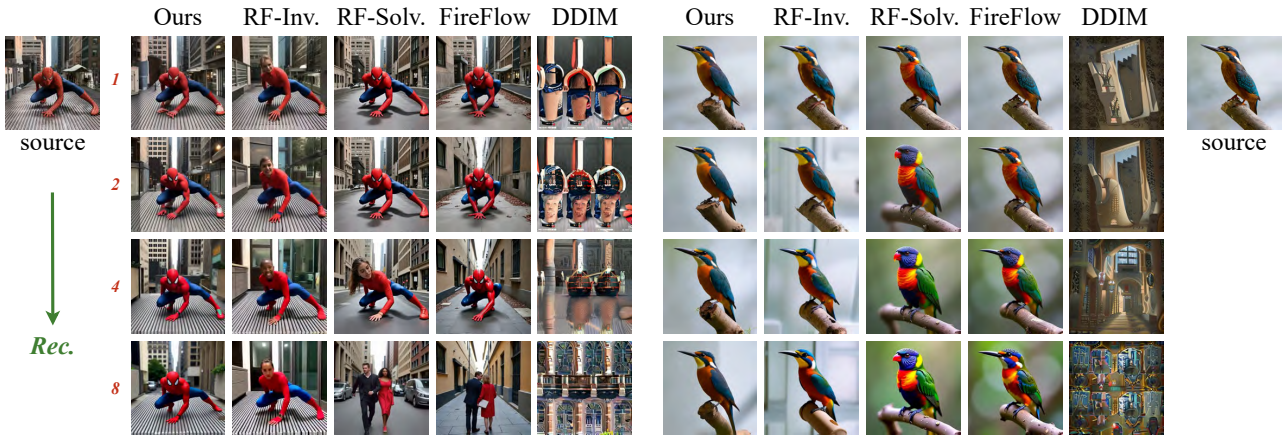


Figure 2. **Multi-turn Reconstruction Results.** This figure compares image reconstructions using our method and baseline methods across 1, 2, 4, and 8 reconstruction iterations. Our method effectively preserves color, background, structure, and semantic consistency across multiple reconstruction rounds, outperforming the baseline methods.

ing cycles. Such an approach would provide enhanced control over the final result by allowing continuous adjustments based on intermediate outcomes, as illustrated in Fig. 1. Consequently, further exploration of multi-turn image editing frameworks is essential to unlock their potential for iterative image refinement.

An intuitive approach to multi-turn image editing involves directly integrating existing single-step methods, leveraging the significant advancements in diffusion-based inversion [13, 16, 23, 24, 26, 32] and related editing techniques. These single-step methods often employ techniques such as attention map replacement [3, 4, 7, 11, 12, 19, 25, 33], mask application [3, 5, 14], and domain-specific pre-trained models [15, 17, 35] to mitigate inversion inaccuracies and preserve image structure. However, this strategy often lacks the robustness required for reliable multi-turn editing, as these techniques are insufficient to prevent the accumulation of errors across multiple iterations. Consequently, edited results in multi-turn frameworks tend to exhibit increasing artifacts and semantic biases, deviating significantly from natural image characteristics.

Flow matching [10, 21, 22] has emerged as a powerful technique for image generation and editing. By directly estimating the transformation from noisy to clean images, rather than predicting noise as in diffusion-based methods, flow matching offers a more efficient and direct framework. This results in simplified distribution transfer, fewer inference steps, and ultimately, more precise editing and reconstruction. This has led to its adoption in state-of-the-art models like SD3 [10] and FLUX.1-dev [20]. Existing image editing research has explored flow matching [1, 2, 10, 21, 22] as a method for accurate image inversion in single-turn editing. Beyond single-turn editing, ReFlow-based models have significant potential for multi-turn editing due to their efficiency in inference steps and accurate inversion, which are crucial prerequisites for this task. However, as shown in Fig. 2, challenges such as accumulated errors in multi-turn editing still need to be addressed. Additionally, the trade-off between preserving content and ensuring sufficient editing flexibility in a multi-turn framework remains unexplored.

In this paper, we present a novel framework that leverages FLUX models to facilitate robust and controllable multi-turn image editing. To ensure long-term coherence and restrict the distribution of edited images in multi-turn tasks, we integrate a dual-objective Linear Quadratic Regulator (LQR) control mechanism into our framework. This LQR mechanism considers both the outputs of preceding turns and the initial input image, establishing a long-term dependency in the editing process. Although dual-objective LQR's stabilization capability is essential for reliable multi-turn editing, the method's stringent regularization constraints may inadvertently reduce editing flexibility. To achieve a balance between stability and flexibility during the editing process, we propose an adaptive attention guidance method aimed at directing the editing focus toward salient regions. This adaptive attention mechanism utilizes medium-to-low activated regions as spatial guidance signals to generate a probabilistic editing mask. By employing attention reweighting, this approach selectively concentrates on target areas while preserving non-target regions.

The key contributions are summarized as follows:

- A dual-objective Linear Quadratic Regulator (LQR) approach that builds upon the flow matching inversion process to ensure stable image distribution across multiple editing turns.
- An adaptive attention mechanism, guided by analysis of intermediate attention layers within the DiT architecture, to enhance the precision and localization of edits.
- A multi-turn interactive image editing framework that

empowers users to iteratively refine images with consistent and predictable results.

2. Preliminary

Rectified Flow: Liu et al. [22] proposed an ordinary differential equation (ODE) model to describe the distribution transfer from $x_0 \sim \pi_0$ to $x_1 \sim \pi_1$. They defined this transfer as a straight-line path, given by $x_t = tx_1 + (1-t)x_0$, where $t \in [0,1]$. This can be expressed as the following differential equation:

$$\frac{dx_t}{dt} = x_1 - x_0. \tag{1}$$

To model this continuous process, they sought a velocity field v that minimizes the objective:

$$\min_v \int_0^1 \mathbb{E}\left[\|(x_1 - x_0) - v\left(x_t, t\right)\|^2\right] \mathrm{d}t. \tag{2}$$

In practice, this continuous ODE is approximated using a discrete process, where the velocity field $v(x_t, t)$ is parameterized by a neural network. Typically, $x_1 \sim \pi_1$ is assumed to be Gaussian noise, and $x_0 \sim \pi_0$ represents the target image. The discrete inversion process is then formulated as:

$$X_{t+\Delta t} = X_t + v(\theta, t)\Delta t, \tag{3}$$

where $v(\theta)$ denotes the neural network with parameters θ.

High-order Solver: To enhance the precision of this discretization, RF-Solver [34] and FireFlow [8] employ second-order ODE solvers, which reduce the approximation error from $\mathcal{O}(\Delta t^2)$ to $\mathcal{O}(\Delta t^3)$ for the same step size Δt as used in standard ODE methods. This improvement enables comparable results with fewer sampling steps. In practice, these methods implement the standard midpoint method, increasing accuracy by evaluating the velocity field at an intermediate point. In the discrete setting, for time $t \in [0,1]$ and a positive time increment $\Delta t > 0$, the inversion process updates the state forward in time according to:

$$X_{t+\Delta t} = X_t + v(\theta, t + \frac{\Delta t}{2})\Delta t. \tag{4}$$

Additionally, FireFlow [8] introduces an acceleration technique by caching intermediate velocity field results, reducing the required sampling steps to eight with the same truncation error as midpoint method.

Linear Quadratic Regulator (LQR) Control: RF-Inversion [31] introduces the Linear Quadratic Regulator (LQR) method to effectively guide image generation. When dealing with images or noise originating from atypical distributions, an explicit guidance term is incorporated. This ensures that images from atypical distributions can be inverted into typical noise, and likewise, atypical noise can be transformed back into typical images. Assuming $x_1 \sim \pi_1$ represents the Gaussian noise space and $x_0 \sim \pi_0$ represents the image space, the discrete inversion process over time $t \in [0,1]$ is described by:

$$X_{t+\Delta t} = X_t + [v_t(X_t) + \eta(v_t(X_t \mid X_1) - v_t(X_t))]\Delta t. \tag{5}$$

This process guides the inversion toward typical noise. In this equation, $v_t(x_t \mid x_1)$ is derived by solving an LQR problem, resulting in $v_t(x_t \mid x_1) = \frac{x_1 - x_t}{1-t}$.

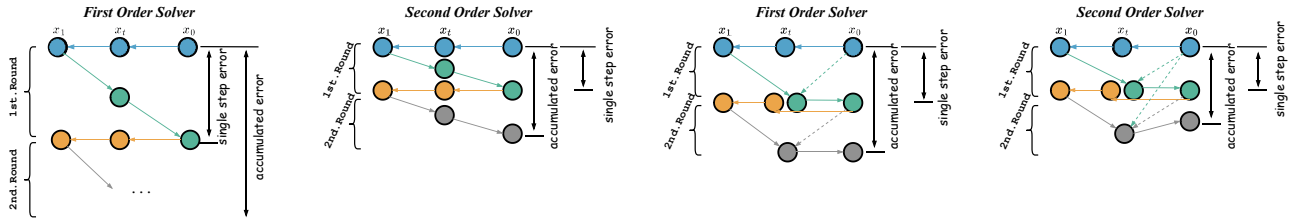
3. Motivation

Single step error v.s. multi-round error. In image editing, flow matching acts as a discrete approximation of a continuous ordinary differential equation (ODE). While employing high-order solvers [8] or increasing the number of timesteps [34] reduces single-step errors—potentially enhancing editing performance in theory—practical implementations encounter notable challenges under multi-round constraints. When the forward and reverse processes are performed multiple times, especially in iterative editing scenarios, multi-round truncation errors become a significant concern as shown in Fig. 2.

Multi-round truncation error arises not just from individual steps but from the accumulation of these errors over a sequence of operations. High-order methods do minimize local truncation errors, but when these methods are applied iteratively, the cumulative error can become substantial, overshadowing initial gains in precision from reducing single-step errors. The reversibility of the process also introduces an additional layer of complexity. Numerical methods are typically not perfectly reversible; the pathway through which errors propagate in the forward direction may differ from that in the reverse direction. This asymmetry can further exacerbate the accumulation of errors, especially over multiple editing cycles.

In practical applications of the ReFlow model, these considerations highlight the limitations of reducing single-step error alone, as shown in Fig. 3b. Instead, comprehensive strategies are needed to address the cumulative nature of global truncation errors and stability challenges in multi-round editing processes.

Single step guidance v.s. multi-turn guidance. Another group of methods [31] relies on the source image as a reference, performing precise single-step edits. However, these methods falter in multi-round editing contexts where cumulative error becomes a critical issue. The crux of the problem lies in the way the original LQR-based approach references only the last edited image Y_i, gradually diverging from the source image Y_0 over multiple iterations. While



(a) Vanilla ReFlow with 1^{st} order ODE solver

(b) FireFlow/RF-solver with 2^{nd} order ODE solver

(c) RF-inversion with 1^{st} order ODE solver

(d) Our approach with 2^{nd} order ODE Solver

Figure 3. We visualize the differences in single-step and multi-round accumulative errors during inversion (←) and editing (↘) across different ReFlow-based editing methods. (a) Vanilla ReFlow struggles with structure preservation during inversion due to the truncation error of the Euler method. (b) While a second-order ODE solver reduces truncation error in a single step, the accumulated error over multiple editing rounds remains significant. (c) Incorporating the source image as guidance (dotted ↙) via LQR improves performance in a single step but becomes less effective as accumulated errors increase with more steps. (d) Our approach addresses this issue by integrating both techniques, leveraging a dual-objective LQR coupled with a high-order solver to enhance stability and accuracy.

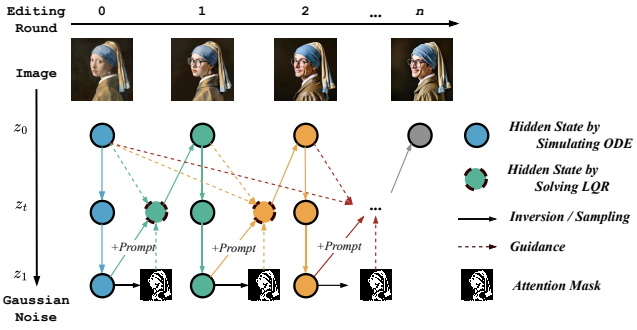


Figure 4. **Multi-turn editing pipeline.** In each editing iteration, a high-accuracy rectified flow inversion maps the image back to the Gaussian noise space, followed by sampling to generate the edited images. To better constrain the distribution of edits across multiple turns, the original image and previous editing results serve as guidance during subsequent sampling. Additionally, a highlighted region in the attention mask further preserves the content structure of the edited outputs.

well-suited for single-step optimization as Y_i equals to Y_0, this technique accumulates discrepancies across successive rounds of edits due to its inability to realign with the original image's core characteristics, as shown in Fig. 3c. Multiple condition generation addresses this shortcoming by incorporating both Y_0 and Y_n as simultaneous conditions for transformation. This dual-reference approach ensures that each round of editing remains anchored to the source image's foundational elements, thereby minimizing drift over time, shown in Fig. 3d.

4. Method

4.1. Dual-objective LQR Guidance

We develop an optimal control strategy to efficiently transform any image X_0 (whether corrupted or not) into a state that reflects multiple random noise conditions, represented by samples $X_1 \sim p_1, X_2 \sim p_2, \ldots, X_n \sim p_n$.

$$V(c) \coloneqq \int_0^1 \frac{1}{2} \left\| c\left(Z_t, t\right) \right\|_2^2 \, \mathrm{d}t + \sum_i^n \frac{\lambda_i}{2} \left\| Z_1 - X_i \right\|_2^2, \\ \mathrm{d}Z_t = c\left(Z_t, t\right) \, \mathrm{d}t, \quad Z_0 = \mathbf{X}_0. \tag{6}$$

This formulation is equivalent to leveraging a weighted average approach in a d-dimensional vector space $\mathbb{R}^d$ to achieve a balanced transformation:

$$V(c) \coloneqq \int_0^1 \frac{1}{2} \left\| c\left(Z_t, t\right) \right\|_2^2 \, \mathrm{d}t + \frac{\lambda}{2} \left\| Z_1 - \hat{X} \right\|_2^2, \\ \mathrm{d}Z_t = c\left(Z_t, t\right) \, \mathrm{d}t, \quad Z_0 = \mathbf{X}_0, \tag{7}$$

where $\hat{X} = \frac{\sum_{i=1}^n \lambda_i X_i}{\sum_{i=1}^n \lambda_i}$ represents the weighted synthesis of the noise samples. The function $V(c)$ quantifies the total energy of the control $c : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$. By optimizing $V(c)$ over the set of permissible controls, denoted by $\mathcal{C}$, we address the multi-condition generation challenge through a Linear Quadratic Regulator (LQR) framework.

Proposition 1 *Given $Z_0 = \mathbf{X}_0$ and the composite target $\hat{X} = \frac{\sum_{i=1}^n \lambda_i X_i}{\sum_{i=1}^n \lambda_i}$, the optimal control solution for the LQR problem (7), denoted by $c^*(\cdot, t)$, aligns with the conditional vector field $u_t(\cdot | X_1, ..., X_n)$, guiding the transformation along the interpolated path $X_t = t\hat{X} + (1-t)X_0$. Specifically, this results in $c^*(\mathbf{z}_t, t) = u_t\left(\mathbf{z}_t | \hat{X}\right) = \frac{\hat{X} - \mathbf{z}_t}{1-t}$.*

Based on Proposition 1 of multi-objective LQR guidance, we establish a framework for iterative image inversion and sampling, constraining the distribution of edited images per round to enable accurate editing. Additionally, we solve the second-order ODE (Eq. (4)) using the FireFlow acceleration algorithm [8], enhancing the speed of single-step simulations within the framework.

In practice, we employ a single-objective LQR for the inversion process. Let the clean image space be denoted by $x_0 \sim \pi_0$ and the Gaussian noise space by $x_1 \sim \pi_1$. For the k-th editing step, the inversion process employs a single-objective LQR to map an image—whether corrupted or uncorrupted—back to the Gaussian noise space π_1, using a second-order ODE solver:

$$\begin{aligned} X^k_{t+\Delta t} = X^k_t + \big[& v_{t+\frac{\Delta t}{2}}(X^k_t) \\ & + \eta \left(v_{t+\frac{\Delta t}{2}}(X^k_t \mid X^k_1) - v_{t+\frac{\Delta t}{2}}(X^k_t) \right) \big] \Delta t. \end{aligned} \tag{8}$$

For sampling, we apply dual-objective LQR control within an invertible flow model, using both the initial image and the previous edit result as guidance. Specifically, at the k-th editing turn, the initial image is denoted as X^k_0 and the result from the $(k-1)$-th editing turn as X^{k-1}_0. Given a time interval $\Delta t > 0$, the dual-objective LQR sampling process is defined as follows:

$$\begin{cases} X^k_{t-\Delta t} = X^k_t - \big[v_{t-\frac{\Delta t}{2}}(X^k_t) + \\ \quad \eta \left(v_{t-\frac{\Delta t}{2}}(X^k_t \mid X^{\text{dual}}_0) - v_{t-\frac{\Delta t}{2}}(X^k_t) \right) \big] \Delta t, \\ X^{\text{dual}}_0 = X^0_0 + \lambda (X^{k-1}_0 - X^0_0), \end{cases} \tag{9}$$

where η and λ are parameters controlling the influence of the guidance terms, $v_t\left(X^k_t \mid X^{\text{dual}}_0\right)$ is intended to encapsulate the dual-objective influence.

4.2. Adaptive Attention Guidance

Our framework leverages flow reversal and LQR-based optimal control for distributional consistency across iterative edits. While LQR ensures stability, its strong regularization can limit editability. To balance stability and editability, we introduce adaptive attention modulation, guiding edits towards salient regions for precise, localized modifications while preserving unaffected areas.

Unlike Stable Diffusion [28, 30], which processes image and text information through cross-attention [3, 12, 19], FLUX utilizes double blocks to jointly process text and image embeddings. Xu et al. [36] found that FLUX's lower-left self-attention quadrant encodes text-to-image spatial influence. With each column representing a token's modulation, we exploit this column-wise interaction for fine-grained analysis of token activation dynamics.

As shown in Fig. 5, which illustrates a token mapping column reshaped into a visualization attention map, different FLUX double blocks exhibit distinct editing behaviors. As shown in the top row, the first and third double blocks primarily influence the entire image, while the second and twelfth focus on the main object. Notably, the sixteenth and eighteenth blocks precisely activate the region corresponding to "monkey," aligning with the desired editing area. This analysis reveals a discernible trend: highly activated maps tend to perform global editing, while lower activated maps focus on finer details.

Given that maintaining coherence across multiple editing turns is essential for effective multi-turn image editing, we emphasize the importance of performing finer and more localized edits in each turn. To achieve this, we propose adaptively identifying and using medium-to-low activated maps as guidance in our framework. This process generates a mask that highlights the focus area for editing, reducing the impact on unaffected regions.

We employ the attention map at time-step t and block l, defined as:

$$s_{t,l} = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}} \right). \tag{10}$$

Following prior work [6, 9], we rescale the attention values to the interval $[0, 1]$ via:

$$s'_{t,l} = \sigma\left(10 * \left(\text{normalize}\left(s_{t,l}\right) - 0.5\right)\right), \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid function and $\text{normalize}(\cdot)$ applies min-max normalization. Let $S'_t = \{s'_{t,1}, s'_{t,2}, \ldots, s'_{t,19}\}$ denote the set of 19 rescaled self-attention maps at step t. To adaptively select **medium-low activated maps** for editing guidance, we define an activation magnitude function $activation(s_{t,l})$, where $a_{t,l} = activation(s_{t,l}) = \sum s_{t,l}$ represents the sum of all elements in the attention map $s_{t,l}$. The maps in S'_t are then sorted in ascending order by activation level, resulting in the sequence:

$$\begin{aligned} A_t = \mathit{Sort}\left\{a_{t,1}, a_{t,2}, \ldots, a_{t,19}\right\} = \left\{a'_{t,1}, a'_{t,2}, \ldots, a'_{t,19}\right\}, \\ \text{where } a'_{t,1} \leq \ldots \leq a'_{t,19}. \end{aligned} \tag{12}$$

Let $A_{i:j} = \{a'_{t,l} \mid l \in \mathbb{Z}, i \leq l \leq j\}$ denote the subset of maps indexed from i to j $(1 \leq i < j \leq 19)$, corresponding to medium-low activation levels. The mask M_t is generated by averaging these selected maps:

$$\bar{v}_{i:j} = \frac{1}{j - i + 1} \sum_{l=i}^{j} a'_{t,l}, \tag{13}$$

and thresholding the result to amplify focused regions while suppressing others:

$$M_t = \begin{cases} h_{\text{factor}} & \text{if } \bar{v}_{i:j} \geq \tau \\ r_{\text{factor}} & \text{otherwise} \end{cases} \tag{14}$$

where h_{factor} and r_{factor} control amplification/reduction, and τ is a predefined threshold. Finally, M_t modulates the attention computation at step $t+1$:

$$s_{t+1,l} = \text{softmax}\left(\frac{Q \cdot K^\top}{\sqrt{d}} \right) \odot M_t, \tag{15}$$

where $\odot$ denotes element-wise multiplication. Since early steps show weak correlations between noisy image and text tokens, the attention guidance begins at $t = 5$.

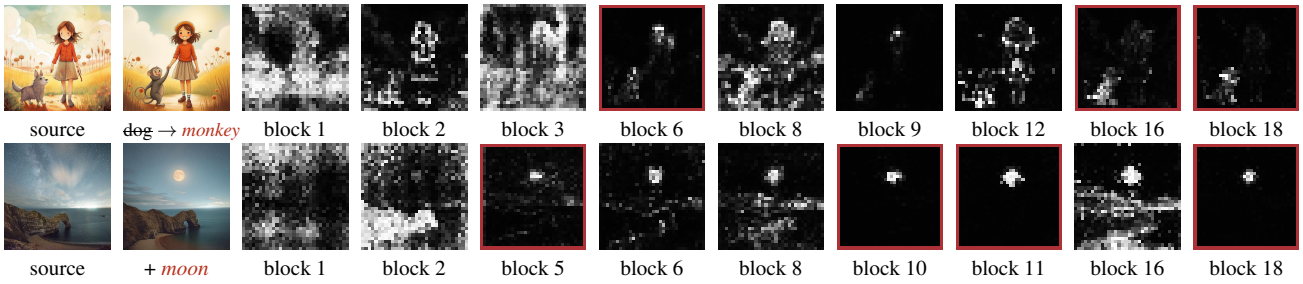


Figure 5. **Self-attention map visualizations** from selected FLUX double blocks (19 total) illustrate layer-specific roles in the editing process (e.g., global, local, details). Top row: attention maps corresponding to the "monkey" text token. Bottom row: maps for the "moon" token. The attention map highlighted by a red box denotes correctly activated maps.

5. Experiment

5.1. Implementation Details

Baselines: We compare our method against flow-based inversion methods including RF-Inverison [31], StableFlow [2], RF-Solver [34], FireFlow [8] and FlowEdit [18]. We also consider diffusion-based inversion methods including MasaCtrl [3], and PnPInversion [33].

Datasets: Existing benchmarks lack support for evaluating multi-turn image editing performance. To address this, we extended PIE-Bench [16], a benchmark originally designed for single-turn image editing, which provides images paired with editing instructions. Using GPT-4 Turbo [27], we generated four additional rounds of editing instructions, conditioned on the original prompt and prior editing prompts. This augmented dataset facilitates robust benchmarking of both single-turn and multi-turn image editing tasks.

Metrics: To demonstrate balance of our method between content preservation and editability, we employ the following evaluation metrics: CLIP-T[29] measures prompt-image consistency; CLIP-I measures the similarity between the original and edited images; and FID [37] assesses the overall generation quality.

Settings: Our method used 15 steps for both inversion and sampling, with parameters $\eta = 0.9$ and $\lambda = 0.7$ in Eq. (9) for the initial 4 sampling steps, $i = 10$ and $j = 14$ in Eq. (13), $h_{factor} = 2.0$ and $r_{factor} = 0.8$ in Eq. (14). Baseline methods were implemented using official code and settings: StableFlow[2] (50 steps), FlowEdit[18] (28 steps). FireFlow [8] (8 steps, both in its original form and without the attention's V replacement variant). RF-Solver[34] was implemented with 25 steps, accounting for its second-order ODE solver (50 effective steps totally) with V replacement. MasaCtrl[3] and PnPInversion[16] used Stable Diffusion's standard 50-step inversion and sampling.

5.2. Multi-turn Reconstruction

Figure 6 presents the MSE reconstruction results for long-term performance (ten-turn reconstruction) of our method compared to the baselines. The results demonstrate that our method remains stable and has fewer drift issues. The qualitative results of multi-turn reconstruction can be seen in Fig. 2. Flow matching shows great potential for multi-turn reconstruction or editing. FireFlow [8] and RF-Solver [34] perform exceptionally well in single-step reconstruction, indicating that solving second-order ODEs in flow matching improves inversion accuracy. However, these two methods still suffer from accumulated errors, causing the distribution drift in the reconstructed images. RF-Inversion[31] maintains semantic consistency and distribution well but tends to enforce certain patterns in the images. Our method preserves the distribution and produces natural-looking results even across multiple editing rounds.

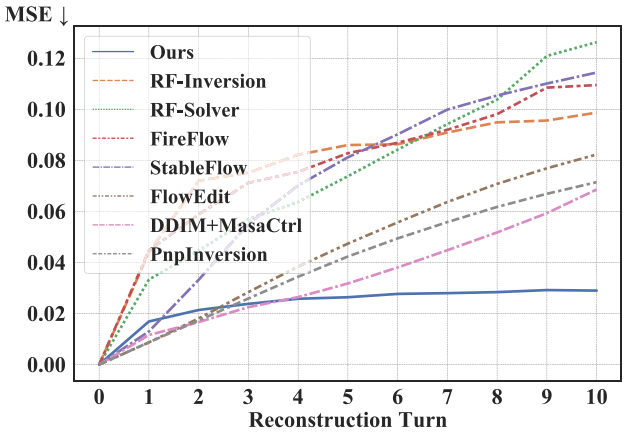


Figure 6. **MSE error across ten reconstruction turns.**

5.3. Multi-turn Editing

Figure 7 provides a qualitative comparison of multi-turn editing results, illustrating the performance of our method and several baselines. In our experiments, diffusion-based methods, including MasaCtrl [3] and PnPInversion, performed poorly in multi-turn editing, failing to preserve the original image structure and generate accurate edits. While RF-Inversion [31], RF-Solver[34], and StableFlow [2] demonstrate accurate inversion by maintaining the orig-

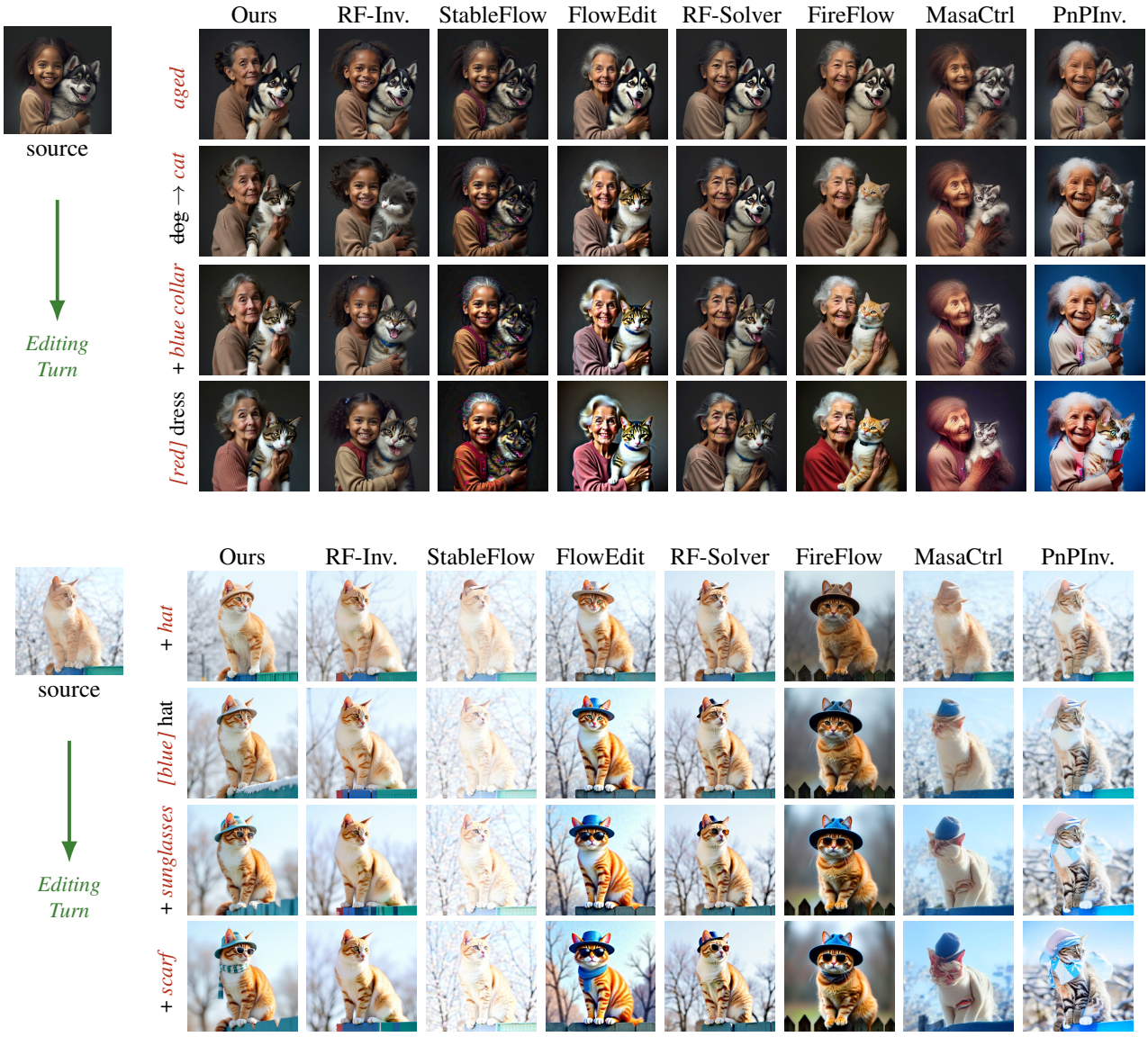


Figure 7. **Qualitative comparison of multi-turn editing results against baseline methods.** Note that our method effectively preserves the original image structure while achieving high-quality edits.

inal image structure, they often fail to produce the desired edits. For example, RF-Solver and StableFlow are unable to transform a "dog" into a "cat" (top subfigure) or add a "scarf" (bottom subfigure). FireFlow [8] and FlowEdit [18] successfully perform the edits specified by the text prompts, but they compromise the original image structure to varying degrees, with FlowEdit exhibiting a tendency to generate images with increasing artifacts over multiple editing rounds. Our method overcomes these limitations by achieving a more adaptable balance between structure preservation and successful editing, allowing for both accurate and meaningful image manipulations.

Table 1 presents quantitative results for the fourth editing turn, highlighting our approach's advantages in multi-turn scenarios. Our method achieves a relatively high CLIP-T score, demonstrating successful alignment with the editing prompt, while simultaneously maintaining high CLIP-I scores, indicating effective content preservation. Notably, our method also achieves the best FID score, suggesting that the generated images retain the characteristics of natural images and exhibit minimal distribution bias after multiple editing iterations.

5.4. Ablation Study

To evaluate the contribution of key components, we conducted ablation studies on: (1) single-objective versus dual-objective LQR (Sec. 4.1) in multi-turn editing; (2) attention map activation levels (Sec. 4.2) for editing performance;

Methods	FID ↓	CLIP-T ↑	CLIP-I ↑	Steps
RF-Inv.	5.740	24.094	<u>0.904</u>	28
StableFlow	20.624	24.234	0.899	50
FlowEdit	14.547	26.703	0.894	28
RF-Solver	11.581	25.516	**0.906**	25
FireFlow	7.970	26.500	0.897	8
FireFlow$-v$	12.375	**28.281**	0.873	8
MasaCtrl	10.811	23.797	0.886	50
PnPInv.	10.262	25.765	0.872	50
Ours	<u>5.553</u>	<u>26.831</u>	0.894	15
Ours	**5.396**	25.828	0.902	8

Table 1. **Quantitative results of fourth-turn editing.** Best results are highlighted in bold, and second-best are underlined. Our method achieves the best FID score while balancing CLIP-I and CLIP-T metrics effectively at the fourth editing step.

Settings	FID ↓	CLIP-T ↑	CLIP-I ↑	Steps
Single-LQR	9.886	26.484	0.892	15
High-attn	6.316	<u>26.878</u>	0.891	15
w/o attn	6.678	26.760	0.889	15
$\lambda = 0.5$	5.161	26.641	0.899	15
$\lambda = 0.9$	6.651	26.873	0.892	15
$\eta = 0.8$	6.677	26.831	0.890	15
$\eta = 1.0$	5.982	25.531	0.912	15
Ours$_{\lambda=0.7,\eta=0.9}$	5.553	26.831	0.894	15

Table 2. **Ablation study on fourth-turn editing reults.**

and (3) parameters λ and η (Eq. (9)).

Quantitative results for the fourth-turn editing are presented in Tab. 2. Relying solely on previous editing turn's output as single-objective LQR guidance introduces distribution bias, resulting in a significantly faster increase in FID compared to dual-objective LQR guidance, which incorporates the original image. Additionally, both highly activated attention guidance and the absence of attention mask guidance hinder content preservation. However, using highly activated attention as guidance improves editability. From Eq. (9), a smaller λ places more emphasis on the original image X_0^0 and less on the previously edited image X_0^{k-1}. A larger η gives more weight to the historical information X_0^{dual}. The quantitative ablation results align with the mathematical intuition from Eq. (9). Specifically, a smaller λ and larger η leads to reduced editability (CLIP-T) but improved content structure preservation (CLIP-I). Due to our method's high-accuracy inversion and sampling, along with its handling of historical information, it maintains strong FID performance with λ and η vary.

Figure 8 shows that single-objective LQR guidance: LQR guidance based solely on the original image restricts editability, while relying only on previous steps leads to accumulated artifacts. For the attention map ablation, we defined "low", "medium," and "high" activation levels based

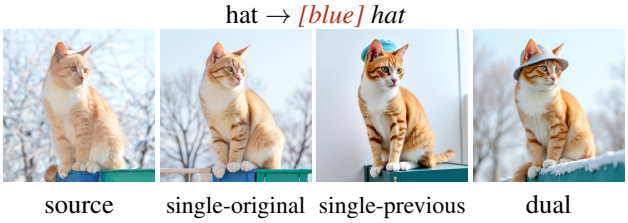


Figure 8. **Ablation study on single-objective LQR guidance.** Guidance based only on the original image limits editability, while relying only on the previous step leads to accumulated artifacts.

Figure 9. **Ablation study of adaptive attention guidance.** Results demonstrate that editing without attention guidance struggles to affect salient areas, while increasing attention map activation leads to structural damage overly aggressive edits.

on the 19 double blocks in FLUX.1-dev (Sec. 4.2), corresponding to the $12{\sim}17th$, $6{\sim}10th$, and top 5 most highly activated attention maps, respectively (Fig. 9). Results show that attention guidance is essential for effective editing, as its absence restricts edits to salient areas due to strong LQR constraints. Higher activation levels often damage the original image's structure and background, while lower levels enable precise edits, e.g., transforming a "man" into a "superhero" by targeting glasses and cloak.

6. Conclusion

This paper investigated the workflow and necessities of multi-turn image editing, explaining the limitations of existing approaches when adapted to this task. To overcome these issues, we proposed a novel framework that integrates accurate flow matching inversion with a dual-objective LQR guidance method. Furthermore, we analyzed the roles of different transformer blocks within the DiT architecture and introduced an adaptive attention map selection mechanism to improve editability while preserving unaffected areas. Our experiments demonstrate the superior performance and adaptability of our method in multi-turn editing scenarios.
Future Work: (1) We will expand our dataset with longer editing rounds to enable a more robust evaluation of multi-turn performance. (2) We plan to adapt temporal consistency techniques from video editing to improve coherence across multiple image editing turns. (3) We will explore inversion-free methods and geometric shape matching to achieve higher precision in multi-turn image editing tasks.

References

[1] Michael S. Albergo and Eric Vanden-Eijnden. Building Normalizing Flows with Stochastic Interpolants, 2023. 2

[2] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. *arXiv preprint arXiv:2411.14430*, 2024. 2, 6

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 2, 5, 6

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2

[5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[6] Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. Fluxspace: Disentangled semantic editing in rectified flow transformers. *arXiv preprint arXiv:2412.09611*, 2024. 5

[7] Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. Z$\hat{*}$: Zero-shot style transfer via attention rearrangement. *arXiv preprint arXiv:2311.16491*, 2023. 2

[8] Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. FireFlow: Fast Inversion of Rectified Flow for Image Semantic Editing, 2024. 3, 4, 6, 7

[9] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 5

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, 2024. 2

[11] Jing Gu, Nanxuan Zhao, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, Yilin Wang, and Xin Eric Wang. Swapanything: Enabling arbitrary object swapping in personalized image editing. In *European Conference on Computer Vision*, pages 402–418. Springer, 2024. 2

[12] 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. 2, 5

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[14] Nisha Huang, Fan Tang, Weiming Dong, Tong-Yee Lee, and Changsheng Xu. Region-aware diffusion for zero-shot text-driven image editing. *arXiv preprint arXiv:2302.11797*, 2023. 2

[15] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Changsheng Xu. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2

[16] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 2, 6

[17] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 2

[18] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024. 6, 7

[19] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 2, 5

[20] Black Forest Labs. Flux.1 [dev] is an open-weight, guidance-distilled model for non-commercial applications, 2024. 2

[21] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, 2023. 2

[22] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow, 2022. 2, 3

[23] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 2

[24] 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*, pages 6038–6047, 2023. 2

[25] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8110, 2024. 2

[26] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[27] OpenAI. GPT-4 Turbo. https://platform.openai.com/docs/models/gpt-4-turbo, 2023. 6

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5

[29] 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*, pages 8748–8763. PmLR, 2021. 6

[30] 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*, pages 10684–10695, 2022. 5

[31] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations, 2024. 3, 6

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[33] 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*, pages 1921–1930, 2023. 2, 6

[34] 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, 2024. 3, 6

[35] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 2

[36] Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Xiaoyu Kong, Jintao Li, Oliver Deussen, and Tong-Yee Lee. Headrouter: A training-free image editing framework for mm-dits by adaptively routing attention heads. *arXiv preprint arXiv:2411.15034*, 2024. 5

[37] 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*, pages 586–595, 2018. 6

Acknowledgements

This work was supported in part by the Beijing Natural Science Foundation under No. Z231100005923033.