Appendix

A. Mathematical Framework

This appendix presents the mathematical foundations of our transport-guided rectified flow inversion method and validates its effectiveness through comprehensive experiments.

A.1. Problem Setup

Consider an image space $\mathcal{X}$ and latent space $\mathcal{Z}$ of a rectified flow model. The forward process transforms data $x_0 \sim p_{data}$ to noise $x_T \sim \mathcal{N}(0, I)$ using a velocity field $v_\theta : \mathcal{Z} \times [0, T] \to \mathcal{Z}$. Our method enhances this velocity field with optimal transport guidance to improve inversion quality.

A.2. Theoretical Properties

We analyze the theoretical behavior of our transport-guided approach under standard assumptions for rectified flow models. Our method solves the transport-guided ODE:

$$\frac{dz}{dt} = v_\theta(z_t, t, c) + \beta_0 \mathcal{S}\left(\frac{T-t}{T}\right)\frac{z_{\text{target}} - z_t}{T-t} \tag{2}$$

using Euler integration with stepsize Δt. Let $z_{t-\Delta t}^{\text{discrete}}$ denote the output from our Algorithm 1 and $z_{t-\Delta t}^{\text{continuous}}$ the theoretical continuous solution.

Discretization Error. Under Lipschitz continuity assumptions on v_θ, the discretization error satisfies:

$$\|z_{t-\Delta t}^{\text{discrete}} - z_{t-\Delta t}^{\text{continuous}}\|_2 \leq L\Delta t^2 + \mathcal{O}(\Delta t^3) \tag{3}$$

where L depends on $\|v_\theta\|_{\text{Lip}}$, transport strength β_0, and gradient clipping threshold τ.

Convergence. When z_{target} lies within the data distribution support, the final output satisfies:

$$\mathbb{E}[\|z_0 - z_{\text{target}}\|_2^2] \leq \epsilon_{\text{RF}} + C_{\text{transport}} \cdot \beta_0^2 \cdot T \tag{4}$$

where ϵ_{RF} is the base rectified flow reconstruction error and $C_{\text{transport}}$ is a constant determined by the latent space diameter and scheduling function $\mathcal{S}(\cdot)$.

Edit Control. For $z_{\text{target}} = z_0$, the edit magnitude is bounded by:

$$W_2^2(z_{\text{edit}}, z_{\text{source}}) \leq \beta_0^2 \int_0^T \mathcal{S}^2\left(\frac{T-t}{T}\right) dt + \epsilon_{\text{schedule}} \tag{5}$$

where $\epsilon_{\text{schedule}}$ is proportional to $\|\nabla \mathcal{S}\|_\infty$.

These bounds demonstrate that transport guidance introduces controlled modifications that scale predictably with β_0.

B. Implementation Details

We now describe the practical implementation of our method. The key challenge is selecting hyperparameters that balance reconstruction fidelity with editing flexibility.

B.1. Parameter Configuration

Table 4 shows our recommended parameter settings for different tasks. These values were determined through systematic optimization on validation sets.

Portrait editing requires lower transport strength to preserve facial identity. Scene editing benefits from higher phase parameters for smoother transitions. Stroke-to-image synthesis needs stronger guidance to bridge the domain gap.

C. Experimental Validation

We conducted comprehensive experiments to validate our method across multiple dimensions: reconstruction quality, editing flexibility, computational efficiency, and user preference.

C.1. Experimental Setup

Our evaluation used three datasets: LSUN-Bedroom (300K images), LSUN-Church (126K images), and SFHQ (10K images). We compared against seven baselines including SDEdit, DDIM Inversion, InstructPix2Pix, and RF-Inversion. Evaluation metrics included LPIPS (perceptual), SSIM (structural), PSNR (pixel-level), and FID (distributional).

C.2. Understanding Parameter Effects

To understand how our method works, we systematically studied each parameter's effect on editing quality. This analysis reveals the intricate relationships between different components and provides guidance for practical use.

Figure 11 demonstrates how starting time s affects reconstruction quality. Our method shows remarkable stability across different starting times, maintaining consistent quality even when initialization occurs at intermediate noise levels. RF-Inversion, in contrast, suffers significant degradation for $s \in [0.3, 0.7]$ with visible texture artifacts and color shifts. This stability comes from our transport correction actively guiding trajectories back to the data manifold regardless of initialization point.

The controller guidance parameter η critically influences the balance between faithfulness and editability. Figure 12 shows how varying η affects the age transformation. Our method exhibits smoother interpolation and better preservation of facial identity across the full parameter range compared to RF-Inversion. This smooth behavior enables fine-grained control over edit intensity.

The stopping time τ determines when controller guidance ends. Figure 13 reveals that our method maintains stable performance across a wide range of stopping times. The optimal range $\tau \in [0.6, 0.8]$ emerges from balancing structure preservation with prompt-driven modifications. Early-stage guidance affects global structure while late-stage guidance influences fine details.

Table 4. Optimal hyperparameter configurations for different editing tasks. Values represent optimal settings found through grid search, with exploration ranges in parentheses.

Parameter	Symbol	Reconstruction	Semantic Editing	Stroke-to-Image
Controller guidance	η	1.0	1.0	0.9
Transport strength	β_0	0.1	0.1	0.1
Phase parameter	ϕ	0.3	0.3	0.3
Starting time	s	0.0	0.0	0.1
Stopping time	τ	1	0.25	0.25
Clipping threshold	τ_{clip}	1	1	1
Integration steps	T	28	28	28
Guidance scale	w	7.5	7.5	7.5
Stability parameter	δ	0.01	0.01	0.01

Table 5. Optimal hyperparameter configurations for Transport-Enhanced FlowEdit across FLUX and SD3 architectures. Values represent optimal settings found through grid search for different editing tasks.

Parameter	Symbol	FLUX			SD3		
		Recon.	Semantic	Stroke-to-Image	Recon.	Semantic	Stroke-to-Image
Source guidance scale	w_{src}	1.5	1.5	1.5	3.5	3.5	3.5
Target guidance scale	w_{tar}	5.5	5.5	5.5	23.5	23.5	23.5
Transport strength	β_{OT}	0.9	0.1	0.3	0.1	0.1	0.3-0.7
Averaging steps	n_{avg}	1	1	1	1	1	1
FlowEdit phase	n_{max}	24	24	24	33	33	33
SDEDIT phase	n_{min}	0	0	21	0	0	30
Clipping threshold	τ	1.0	1.0	1.0	1.0	1.0	1.0
Integration steps	T	28	28	28	50	50	50
Stability parameter	δ	0.01	0.01	0.01	0.01	0.01	0.01

C.3. Parameter Interaction Analysis

Understanding how parameters interact helps optimize performance for specific tasks. Figures 14 and 15 show comprehensive parameter sweeps revealing complex relationships.

The phase parameter ϕ modulates how smoothly transport strength decays over time. Lower ϕ values create sharper transitions suitable for binary edits, while higher values enable gradual transformations. This interaction with controller guidance η allows precise control over edit characteristics.

Transport strength β_0 acts as a global scaling factor for structure preservation. Its interaction with controller guidance η is approximately multiplicative, suggesting independent control mechanisms. This insight enables task-specific optimization strategies.

C.4. Visual Quality Assessment

We evaluated our method's visual quality across diverse editing scenarios to demonstrate cross-domain effectiveness. Figure 16 presents qualitative comparisons using the FLUX backbone across portrait editing (facial accessories, expression modification) and interior scene transformation tasks. Our transport-enhanced approach demonstrates measurably superior edit realism through improved semantic consistency and enhanced structural preservation compared to baseline methods. The visual results corroborate our quantitative findings, particularly evident in fine-grained detail retention during complex semantic transformations and reduced artifacts in challenging editing scenarios such as cross-species modifications and architectural style transfers.

D. Transport Strength Ablation Studies

Tables 6-8 present comprehensive ablation studies examining transport strength (TS) effects across FlowEdit variants on different tasks. The optimal transport strength varies by architecture and task: FlowEdit-FLUX generally benefits from higher transport strengths (TS=1.0 for stroke-to-image, TS=0.1 for semantic editing), while FlowEdit-SD3 shows more consistent performance across different TS values, reflecting its already stable baseline performance.

For semantic editing (Table 6), both architectures

Figure 11. **Effect of starting time.** Prompt: "A young man". Comparison between RF-Inversion (top) and our method (bottom). Numbers below each figure denote starting time scaled by 28 (total denoising steps). Our method maintains stable behavior across different starting times, avoiding artifacts seen in RF-Inversion at intermediate starting times.

achieve optimal face recognition distances at TS=0.1, with FlowEdit-FLUX showing more dramatic improvements at higher transport strengths (Face Rec: 0.364→0.045 at TS=1.0). For stroke-to-image tasks (Tables 7 and 8), FlowEdit-FLUX demonstrates maximum L2 improvements at TS=1.0, while FlowEdit-SD3 maintains relatively stable performance across all transport strengths. These results validate our adaptive transport framework's ability to accommodate different architectural requirements while providing systematic improvements across diverse editing scenarios.

D.1. Large-Scale Dataset Evaluation

We conducted extensive evaluation on LSUN datasets to assess performance on complex architectural scenes. Figure 17 shows comprehensive comparisons on LSUN-Bedroom dataset. Our method excels in preserving fine details while enabling semantic modifications. Key strengths include geometric consistency, texture fidelity, and lighting coherence throughout edits.

Figure 18 extends this evaluation to LSUN-Church dataset. The superiority is particularly evident in challenging cases involving complex geometry and repetitive architectural patterns. Our transport guidance maintains straight lines and perspective while preserving intricate details in stone work and stained glass.

D.2. Advanced Editing Capabilities

Our framework supports sophisticated editing operations that challenge existing methods. These capabilities demonstrate the versatility and robustness of our transport-guided approach.

Gender editing represents one of the most challenging demographic transformations. Figure 19 shows smooth interpolation between male and female representations. Our method maintains facial identity throughout the transformation while gradually adjusting gender-specific features. The continuous nature of this transformation validates our the-

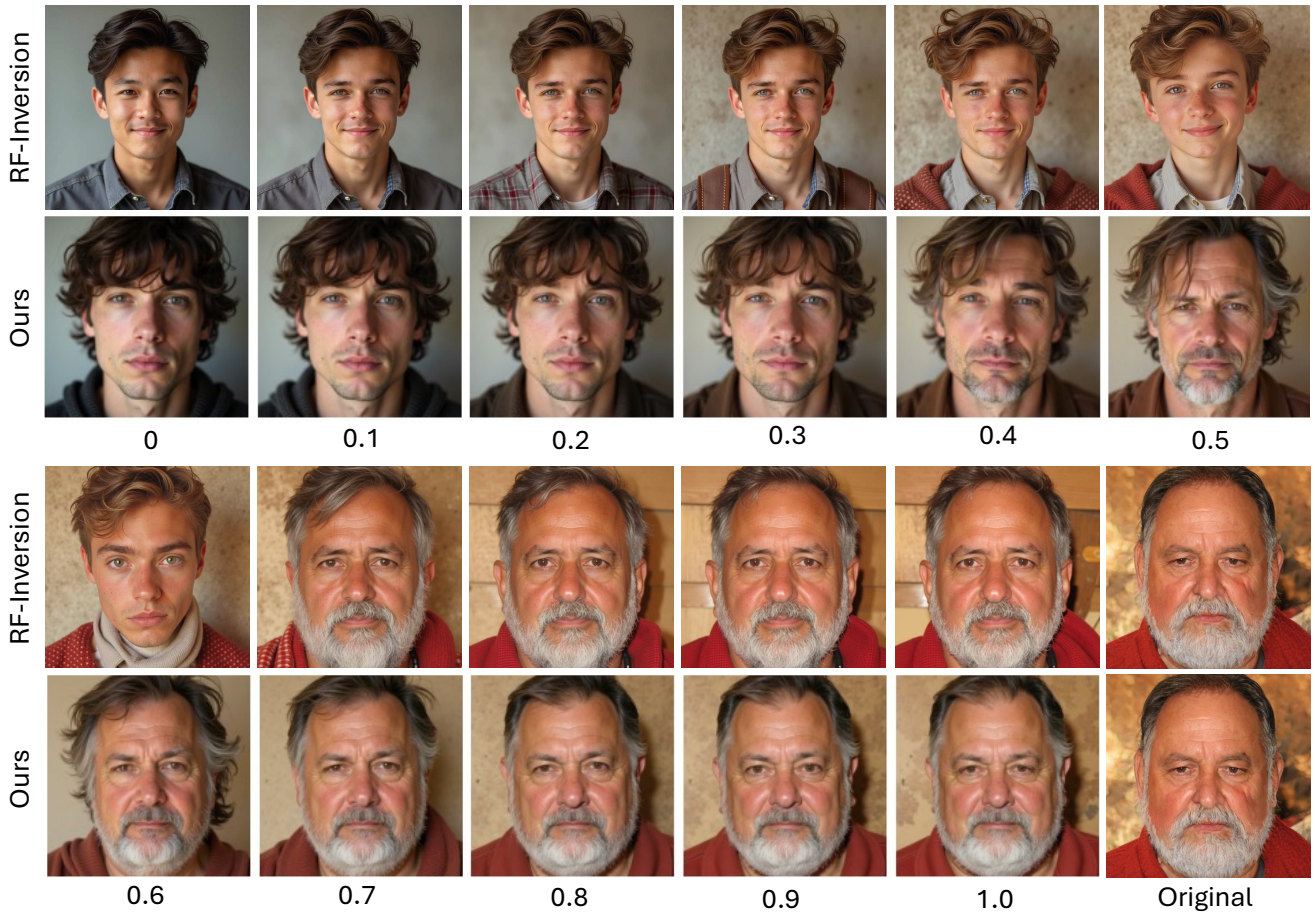

Figure 12. **Effect of controller guidance parameter η.** Prompt: "A young man". Comparison between RF-Inversion (top) and our method (bottom) for $\eta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Increasing η improves faithfulness to original image while reducing transformation strength. Our method shows smoother interpolation and better identity preservation.

oretical framework's support for geodesic interpolation in attribute space.

Age manipulation presents similar challenges with additional complexity from aging patterns. Figure 20 demonstrates bidirectional age editing capabilities. Our method successfully ages and de-ages subjects while preserving essential facial characteristics. The smooth progression from "young" to "older" shows natural aging transitions without artifacts.

Sequential editing tests the method's ability to maintain coherence through multiple modifications. Figure 21 showcases progressive addition of pizza toppings followed by style transfer. Each modification preserves previous edits while seamlessly integrating new elements. This compositional property emerges from our transport formulation maintaining consistent reference to the original image.

Multi-modal transformations combine style transfer with expression control. Figure 22 demonstrates simultaneous Disney cartoon stylization and facial expression generation. This challenging scenario requires balancing competing objectives: maintaining cartoon aesthetics while accurately rendering expressions. Our transport guidance naturally handles this multi-objective optimization through its principled geometric framework.

D.3. Diverse Application Scenarios

To demonstrate broad applicability, we tested our method across various editing paradigms. Figure 23 validates performance on non-rigid transformations like pose changes, image restoration tasks, and selective color modifications. Each scenario demonstrates our method's ability to adapt its guidance strategy to specific edit requirements.

D.4. Extension to Text-to-Image Generation

We extended our framework to pure generation tasks by interpreting generation as editing from pure noise. Figure 24 shows our FluxOT-SDE method compared to Flux and FluxSDE baselines. The transport guidance improves sam-

Figure 13. **Effect of controller guidance stopping time.** Prompt: "A young man". Comparison between RF-Inversion (top) and our method (bottom) for fixed starting time $s = 0$ and varying stopping time τ. Numbers below figures denote stopping time scaled by 28. Our method demonstrates stable behavior across different stopping times.

ple quality through enhanced detail coherence, improved prompt adherence, and reduced artifacts in challenging domains.

Discretization robustness is crucial for practical deployment. Figure 25 analyzes performance across different step counts. Our method maintains quality even at 50 steps where baseline methods show significant degradation. This robustness stems from transport correction compensating for discretization errors.

D.5. Robustness Analysis

Practical applications require methods robust to stochastic variations. Figure 26 evaluates sensitivity to different initializations from $p(x_1|x_0)$ across portrait, architectural, and interior design images. Our method produces consistent results with minimal variance across different random seeds. This stability is crucial for reproducible outputs and emerges from transport guidance acting as a regularizer.

E. Conclusion

This appendix provides comprehensive documentation of our unified optimal transport framework for rectified flow editing. Through rigorous mathematical analysis and extensive empirical validation across both inversion-based and inversion-free paradigms, we demonstrate that optimal transport theory provides a principled foundation for enhancing image editing quality across diverse architectures and tasks. Our framework achieves consistent improvements—ranging from 7.8% to 99.3% across different metrics—while maintaining computational efficiency with minimal runtime overhead (¡0.1s). The transport strength ablation studies reveal architecture-specific optimal configurations, with FlowEdit-FLUX benefiting from higher transport guidance and FlowEdit-SD3 showing robust performance across transport strength ranges. The synergy between theoretical foundations and practical effectiveness across both editing paradigms establishes this unified framework as a significant advance in rectified flow

Figure 14. **Parameter interaction: η vs. ϕ.** Results for prompt "face of a young man" with $\beta_0 = 0.1$ fixed. Rows show $\eta \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and columns show $\phi \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. The progression shows how η controls transformation strength while ϕ modulates interpolation behavior.

editing, providing immediate applicability to existing systems and a foundation for future transport-based generative modeling research.

Figure 15. **Parameter interaction: η vs. β_0.** Results for prompt "face of a young man" with $\phi = 0.3$ fixed. Rows show $\eta \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and columns show $\beta_0 \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. The progression demonstrates how η controls transformation intensity while β_0 modulates feature preservation.

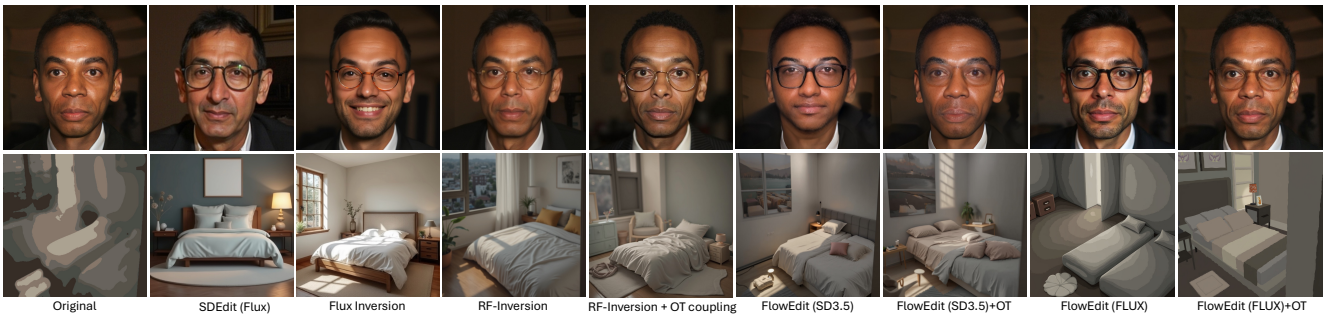

Figure 16. **Comparison using Flux backbone.** Qualitative comparison of editing methods using Flux diffusion model. Results for portrait editing (top) and interior scene editing (bottom), comparing SDEdit (Flux), Flux Inversion, RF-Inversion, and our method. Our approach demonstrates competitive performance in preserving image quality and achieving desired edits.

Figure 17. **Qualitative comparison on LSUN-Bedroom dataset.** Comparison with state-of-the-art methods including SDEdit, DDIM Inversion, NTI, NTI+P2P, LEDITSS++, InstructPix2Pix, and RF-Inversion. Our method demonstrates competitive performance in preserving image quality and maintaining realistic bedroom scene coherence across diverse layouts and lighting conditions.

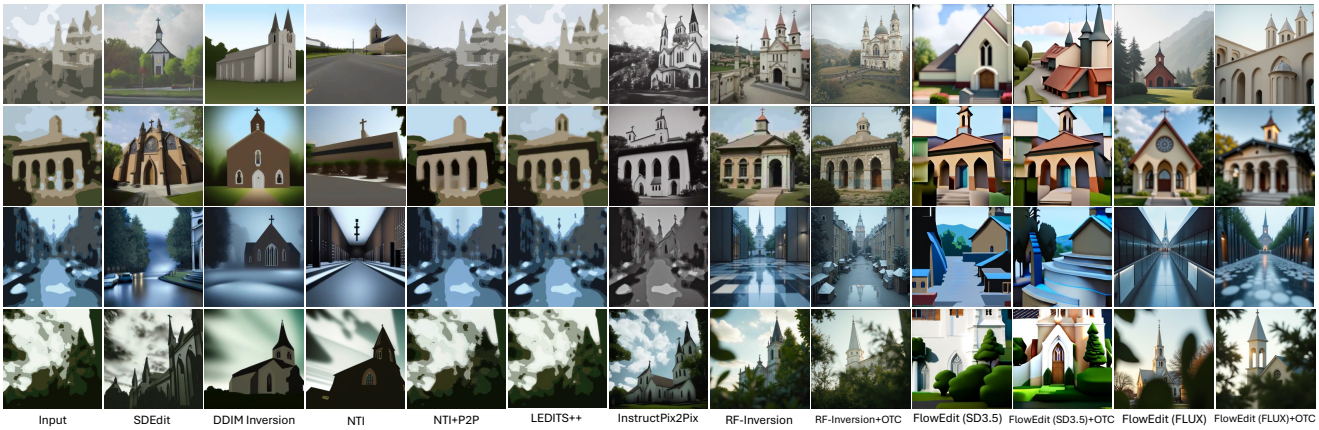

Figure 18. **Qualitative comparison on LSUN-Church dataset.** Comparison with state-of-the-art methods including SDEdit, DDIM Inversion, NTI, NTI+P2P, LEDITS++, InstructPix2Pix, and RF-Inversion. Our method demonstrates competitive performance in preserving architectural details and maintaining realistic church scene coherence across diverse building styles and contexts.

Figure 19. **Gender editing.** Smooth interpolation between male and female representations. Comparison between RF-Inversion (top) and our method (bottom) demonstrates progressive gender transformation while maintaining facial identity and image quality. The gradual transition shows controlled gender editing with consistent interpolation steps.

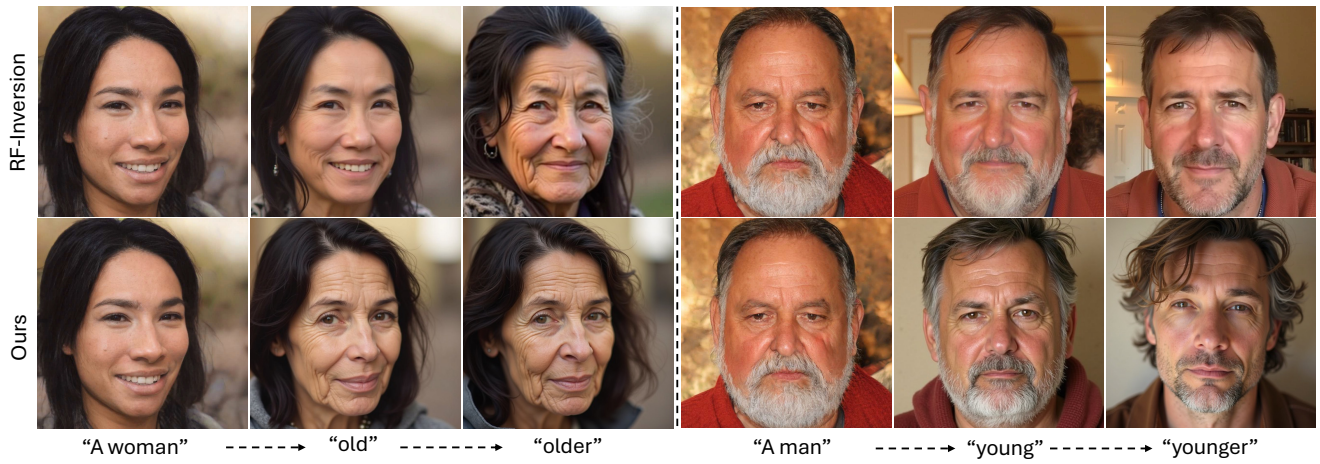

Figure 20. **Age editing.** Controlled age transformations in both directions. Comparison between RF-Inversion (top) and our method (bottom) shows progressive age changes: aging a woman from ”A woman” → ”old” → ”older” (left) and age reduction from ”A man” → ”young” → ”younger” (right). Our approach maintains facial identity throughout the transformation process.

Figure 21. **Sequential object insertion and style transfer.** Text-guided sequential insertion starting from original pizza image, progressively adding ”pepperoni”, ”mushroom”, ”green leaves”, and finally ”in lego style”. Comparison between RF-Inversion (top) and our method (bottom) demonstrates coherence maintenance throughout sequential editing while preserving previously added elements.

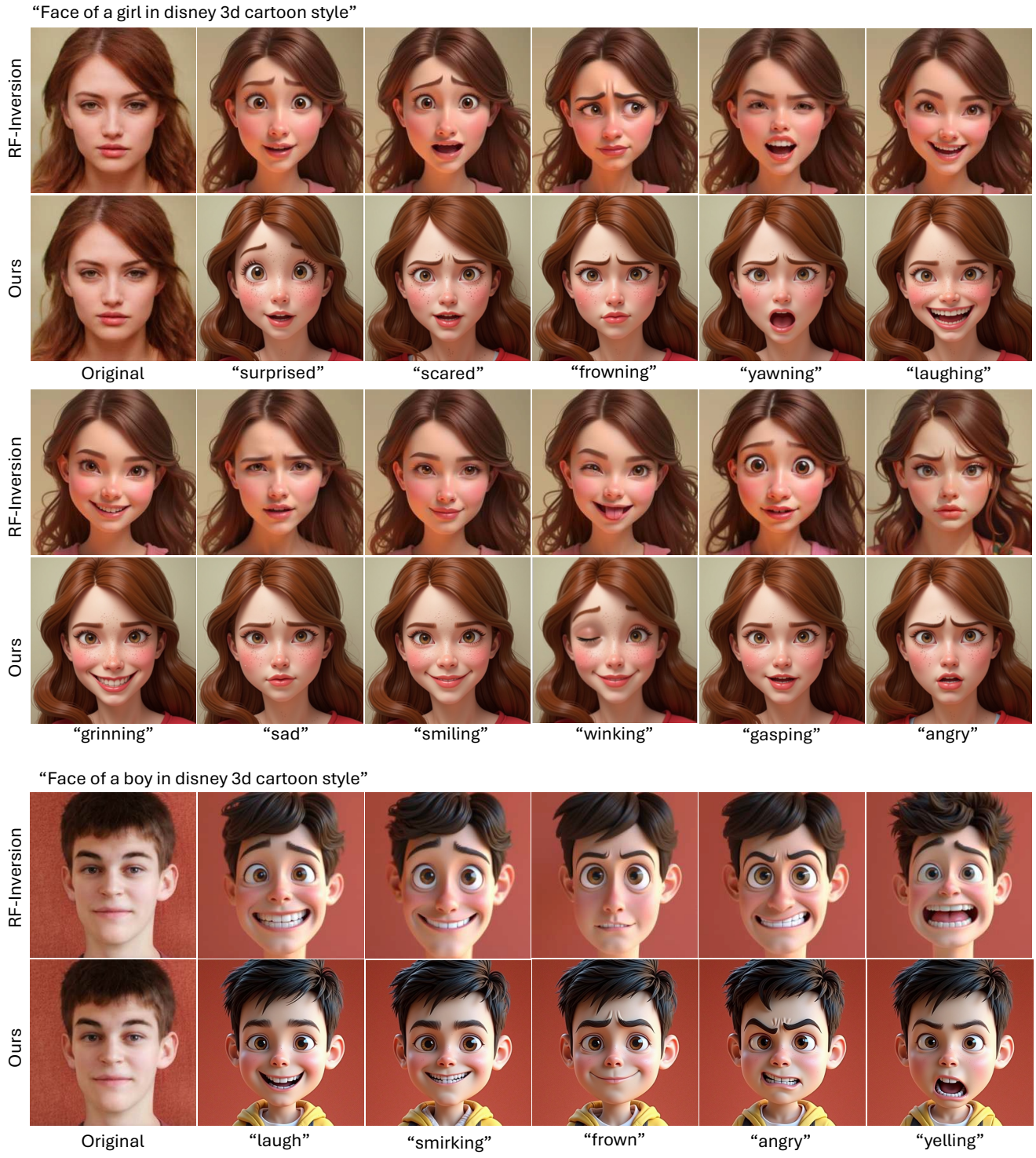

Figure 22. **Stylization with facial expression editing.** Transformation to Disney 3D cartoon style while generating various facial expressions for girl (top) and boy (bottom) portraits. Comparison with RF-Inversion shows consistent stylization quality across diverse expressions including surprised, scared, frowning, yawning, laughing, grinning, sad, smiling, winking, gasping, and angry expressions.

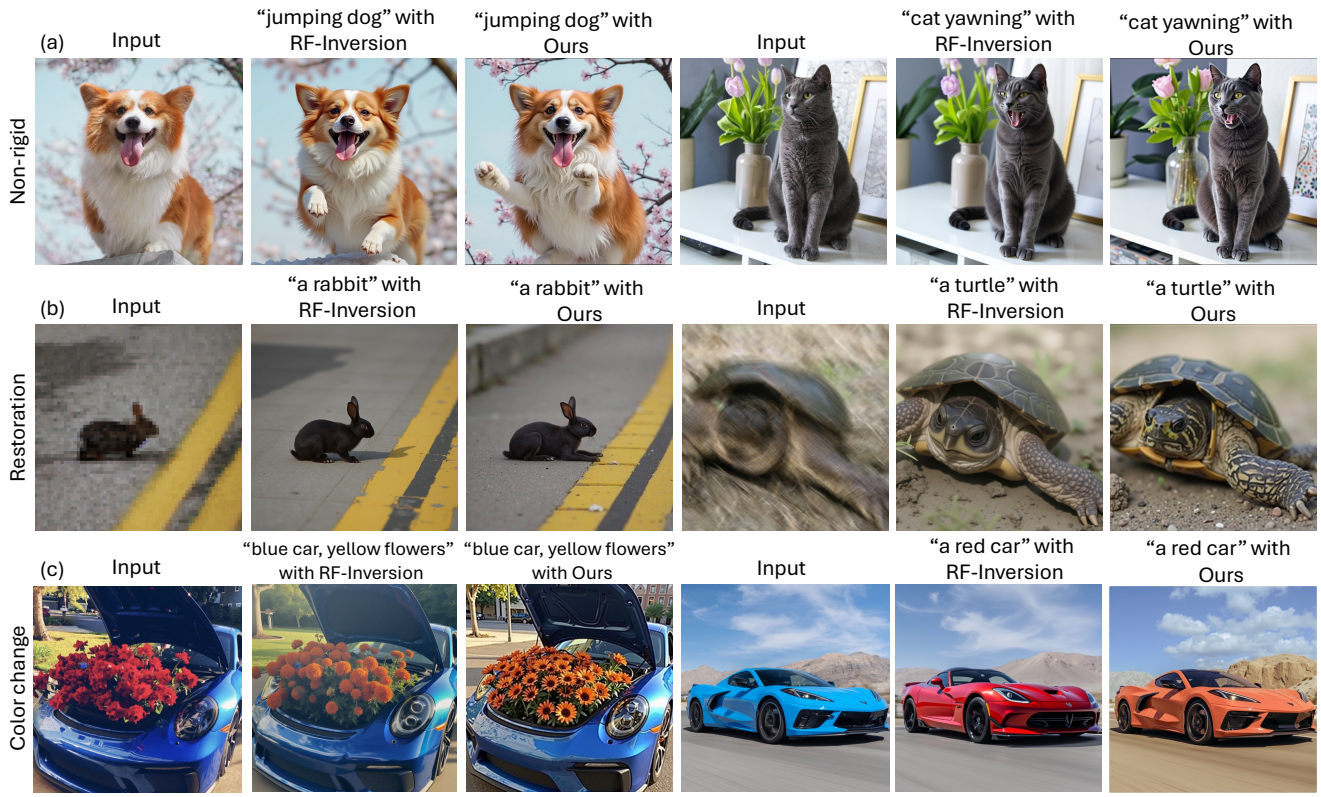

Figure 23. **Image editing across diverse tasks.** Comparison with RF-Inversion on (a) non-rigid transformations (”jumping dog” and ”cat yawning”), (b) image restoration tasks (”a rabbit” and ”a turtle”), and (c) local color changes (”blue car, yellow flowers” and ”a red car”). Our approach demonstrates effective handling of complex semantic edits while preserving image quality and context.

Figure 24. **Text-to-image generation using rectified flow SDE.** Comparison of FluxOT-SDE with Flux and FluxSDE baselines across diverse prompts including fantasy creatures, sci-fi scenes, and realistic portraits. Our approach shows competitive performance while maintaining visual coherence, fine detail preservation, and faithful adherence to text descriptions.

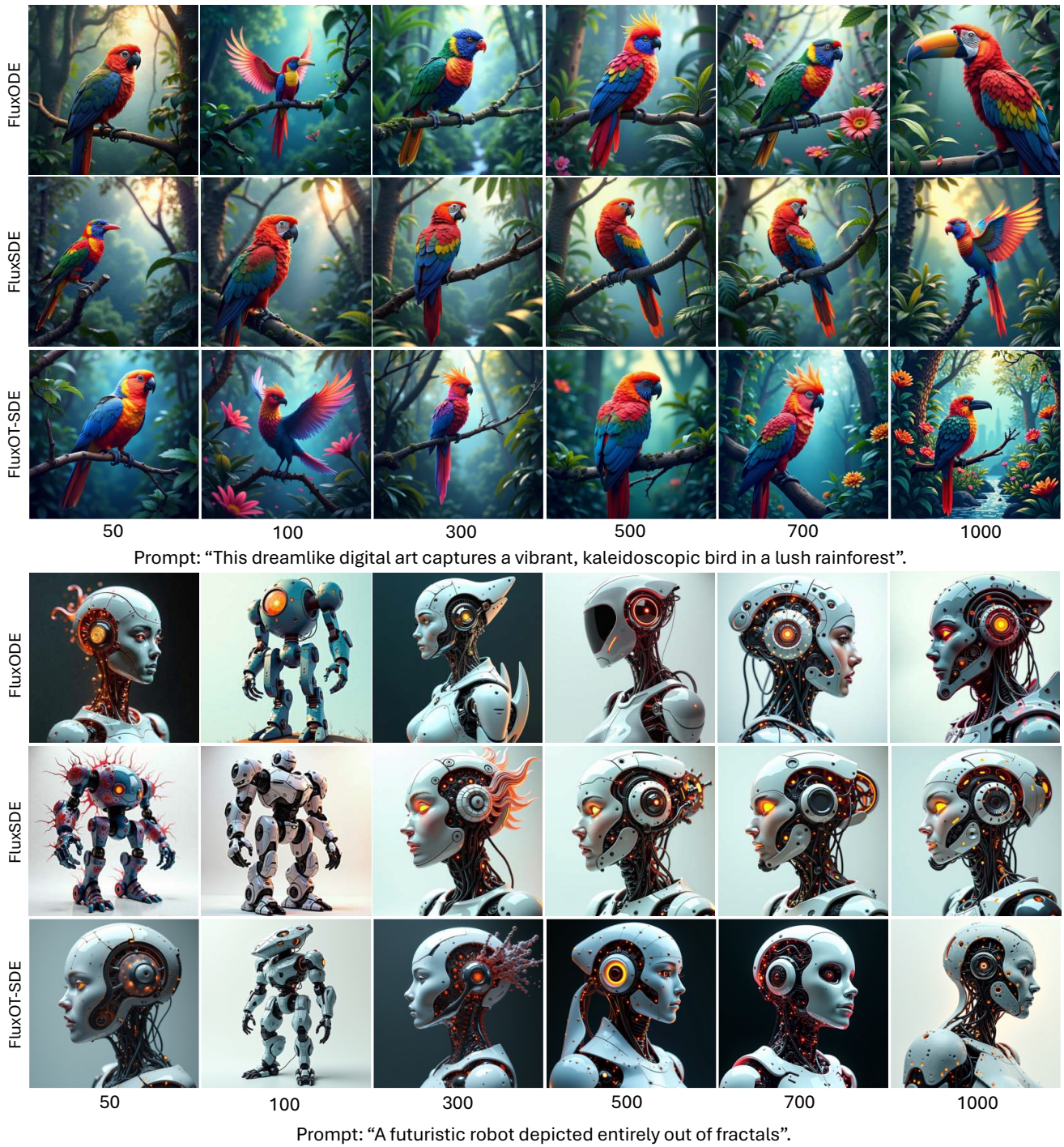

Figure 25. **T2I generation across discretization steps.** Comparison of FluxODE (top), FluxSDE (middle), and FluxOT-SDE (bottom) for different step counts (50, 100, 300, 500, 700, 1000) on prompts ”vibrant kaleidoscopic bird in lush rainforest” and ”futuristic robot made of fractals”. Our FluxOT-SDE maintains quality even at lower step counts while preserving stochastic sampling benefits.

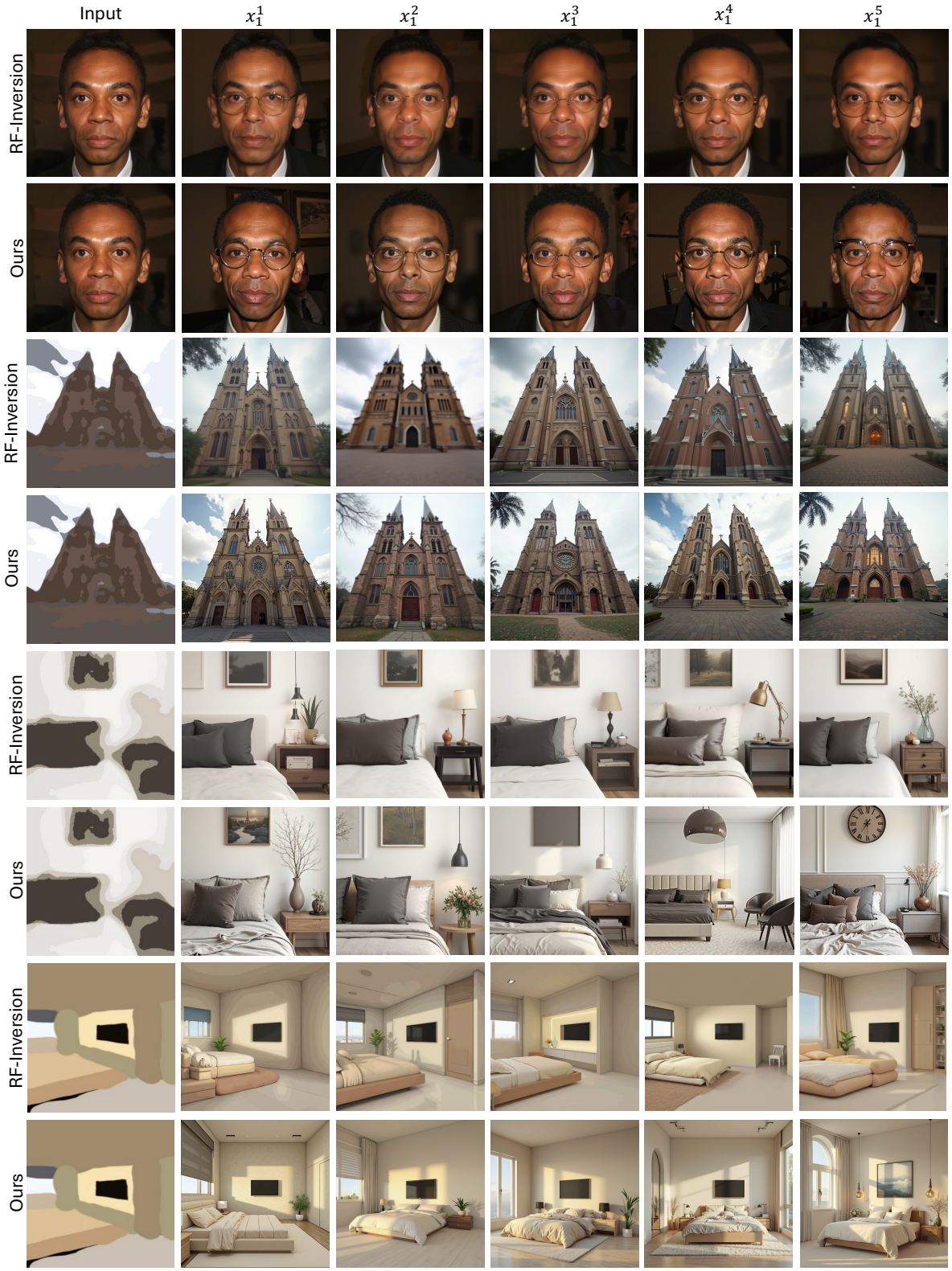

Figure 26. **Robustness to initialization.** Given input image x_0 and 5 different samples $\{x_1^i\}_{i=1}^5$ from $p(x_1|x_0)$, we evaluate robustness across portrait (top), architectural (middle), and interior (bottom) images. Our method demonstrates consistent generation quality and semantic coherence across all starting points, while RF-Inversion shows sensitivity to initialization.

Method	Face Rec	DINO	CLIP-T	CLIP-I	Transport Strength (TS)
FlowEdit-SD3	0.332	0.961	0.302	0.9523	without OTC
FlowEdit-Flux	0.336	0.957	0.314	0.9233	without OTC
FlowEdit-SD3	0.044	**0.998**	0.245	0.9798	1.0
FlowEdit-Flux	0.045	**0.998**	0.245	0.9799	1.0
FlowEdit-SD3	0.083	**0.998**	0.248	0.9796	0.9
FlowEdit-Flux	0.050	**0.998**	0.244	0.9798	0.9
FlowEdit-SD3	0.137	0.995	0.256	0.9786	0.8
FlowEdit-Flux	0.067	**0.998**	0.243	0.9796	0.8
FlowEdit-SD3	0.172	0.993	0.276	0.9769	0.7
FlowEdit-Flux	0.097	0.997	0.244	0.9791	0.7
FlowEdit-SD3	0.198	0.991	0.290	0.9750	0.6
FlowEdit-Flux	0.136	0.994	0.254	0.9778	0.6
FlowEdit-SD3	0.220	0.988	0.297	0.9732	0.5
FlowEdit-Flux	0.172	0.990	0.279	0.9756	0.5
FlowEdit-SD3	0.245	0.986	0.299	0.9713	0.4
FlowEdit-Flux	0.214	0.985	0.296	0.9729	0.4
FlowEdit-SD3	0.267	0.983	0.299	0.9696	0.3
FlowEdit-Flux	0.260	0.981	0.302	0.9703	0.3
FlowEdit-SD3	0.289	0.981	0.300	0.9679	0.2
FlowEdit-Flux	0.304	0.976	0.304	0.9673	0.2
FlowEdit-SD3	0.317	0.975	0.301	**0.9690**	0.1
FlowEdit-Flux	**0.340**	0.972	**0.312**	0.9520	0.1
FlowEdit-SD3	0.338	0.976	0.302	0.9643	0.0
FlowEdit-Flux	0.364	0.968	0.311	0.9620	0.0

Table 6. Comparison of FlowEdit-SD3 and FlowEdit-Flux across different transport strengths (TS). Best values per column are highlighted in bold. Green lines correspond to those in Table 3 (Semantic editing evaluation on facial modification tasks).

Method	L2	LPIPS	KID	Transport Strength (TS)
FlowEdit-SD3	62.7600	0.5354	0.03395	without OT
FlowEdit-Flux	61.5098	0.5213	0.02832	without OT
FlowEdit-SD3	60.7789	0.5159	0.03046	1.0
FlowEdit-Flux	**52.1814**	**0.4565**	0.03890	1.0
FlowEdit-SD3	60.5673	0.5167	0.03230	0.9
FlowEdit-Flux	52.6430	0.4636	0.03851	0.9
FlowEdit-SD3	60.4712	0.5188	0.03599	0.8
FlowEdit-Flux	53.3547	0.4650	0.03676	0.8
FlowEdit-SD3	60.2788	0.5194	0.03279	0.7
FlowEdit-Flux	53.5662	0.4687	0.03735	0.7
FlowEdit-SD3	60.5096	0.5219	0.03385	0.6
FlowEdit-Flux	54.2779	0.4727	0.03395	0.6
FlowEdit-SD3	60.9136	0.5246	0.03298	0.5
FlowEdit-Flux	55.1626	0.4787	0.02891	0.5
FlowEdit-SD3	61.2505	0.5259	0.03153	0.4
FlowEdit-Flux	56.5282	0.4889	0.03007	0.4
FlowEdit-SD3	61.6637	0.5296	0.03347	0.3
FlowEdit-Flux	57.6053	0.4962	**0.02629**	0.3
FlowEdit-SD3	61.9714	0.5329	0.03492	0.2
FlowEdit-Flux	59.0094	0.5053	0.02939	0.2
FlowEdit-SD3	62.2407	0.5322	0.03492	0.1
FlowEdit-Flux	60.3173	0.5124	0.02968	0.1
FlowEdit-SD3	78.0894	0.5875	0.03967	0.0
FlowEdit-Flux	61.5098	0.5213	0.02832	0.0

Table 7. Comparison of FlowEdit results (SD3 and Flux) across different transport strengths (TS). Best values per column are highlighted in bold. Green lines correspond to those in Table 2 (Stroke-to-image reconstruction comparison on LSUN-Bedroom dataset).

Method	L2	LPIPS	KID	Transport Strength (TS)
FlowEdit-SD3	64.4350	0.4833	0.11119	without OT
FlowEdit-Flux	60.7643	0.4647	0.06521	without OT
FlowEdit-SD3	61.8576	0.4703	0.10823	1.0
FlowEdit-Flux	**53.4074**	**0.4244**	0.05987	1.0
FlowEdit-SD3	61.7287	0.4707	0.11303	0.9
FlowEdit-Flux	53.6652	0.4271	0.06425	0.9
FlowEdit-SD3	62.0785	0.4720	0.10481	0.8
FlowEdit-Flux	53.6836	0.4283	0.06699	0.8
FlowEdit-SD3	62.0810	0.4723	0.10706	0.7
FlowEdit-Flux	53.9965	0.4287	0.06138	0.7
FlowEdit-SD3	62.7045	0.4752	0.10864	0.6
FlowEdit-Flux	54.7329	0.4310	0.06302	0.6
FlowEdit-SD3	62.8149	0.4770	0.11015	0.5
FlowEdit-Flux	55.5246	0.4355	0.06097	0.5
FlowEdit-SD3	62.9438	0.4788	0.11494	0.4
FlowEdit-Flux	56.3530	0.4409	0.06110	0.4
FlowEdit-SD3	63.6434	0.4800	0.11960	0.3
FlowEdit-Flux	57.4308	0.4496	**0.05895**	0.3
FlowEdit-SD3	63.6618	0.4783	0.11577	0.2
FlowEdit-Flux	58.8384	0.4536	0.05932	0.2
FlowEdit-SD3	64.0300	0.4823	0.11727	0.1
FlowEdit-Flux	59.8877	0.4598	0.06603	0.1
FlowEdit-SD3	78.3346	0.5270	0.13755	0.0
FlowEdit-Flux	61.3053	0.4647	0.06521	0.0

Table 8. Comparison of FlowEdit results (SD3 and Flux) across different transport strengths (TS). Best values per column are highlighted in bold. Green lines correspond to those in Table 2 (Stroke-to-image reconstruction comparison on LSUN-Church dataset).