STABLECTRL: ONE-STEP STABLE IMAGE EDITING VIA OPTIMAL CONTROL

Anonymous authors
Paper under double-blind review

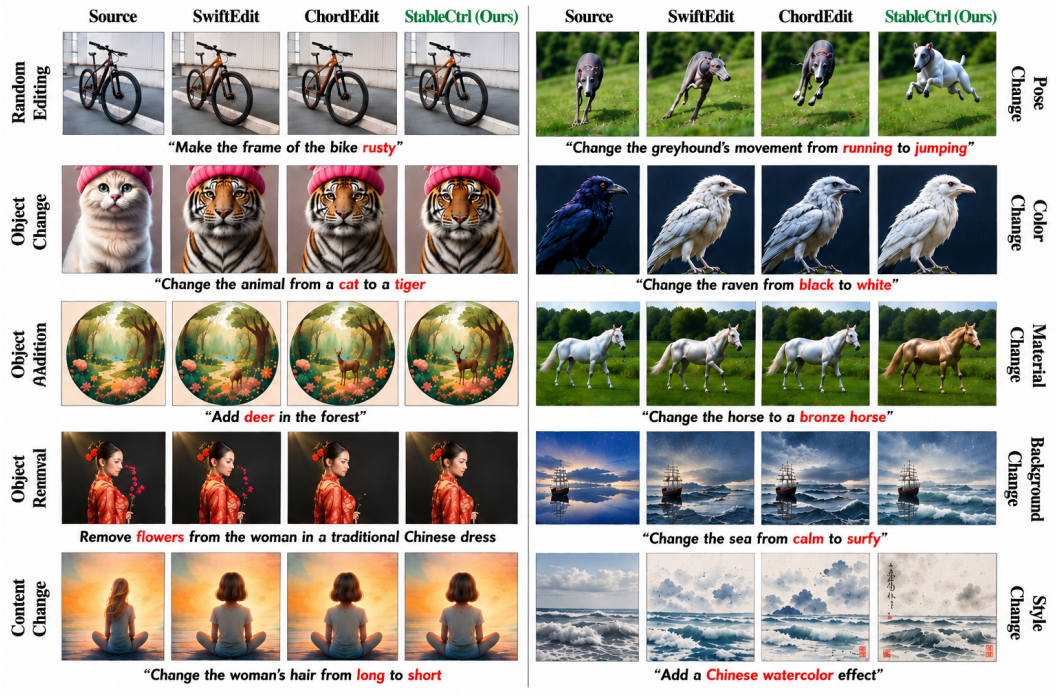

Figure 1: **Qualitative overview of StableCtrl.** Comparison with one-step editing baselines across representative object, attribute, background, and style edits. StableCtrl better preserves object geometry, scene layout, and content outside the intended editing regions across diverse scenarios.

ABSTRACT

Recent one-step text-to-image models enable real-time image editing by completing semantic transformations in a single denoising step. However, existing training-free editors often become unstable in this setting. Source-target residuals can act as high-energy control signals, leading to structural distortion and background inconsistency. To address this issue, we introduce **StableCtrl**, a model-agnostic, training-free, and inversion-free framework for stable one-step image editing. StableCtrl casts one-step editing as a regularized optimal control problem and treats source-target residuals as noisy control cues instead of directly applying them as editing directions. This formulation yields an elliptic control equation with an explicit spectral solution. The solution admits efficient one-step optimization without iterative refinement. It regularizes the control field while retaining the dominant semantic direction. The resulting operator constrains control energy, enforces spatial smoothness, and preserves temporal reference consistency. It therefore balances semantic editing with structural preservation while suppressing unstable high-frequency perturbations through spectral low-pass filtering. StableCtrl consequently produces stable control fields that preserve scene geometry and background consistency without additional model training or latent inversion. Extensive experiments on PIE-Bench show that StableCtrl consistently improves background preservation and editing robustness.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Text-guided image editing (Hertz et al., 2023; Parmar et al., 2023; Cao et al., 2023; Huberman-Spiegelglas et al., 2024; Li et al., 2024; Starodubcev et al., 2024; Xu et al., 2024; Miyake et al., 2025) aims to modify real images according to natural language instructions while preserving content unrelated to the requested edit. Early diffusion-based editing methods (Meng et al., 2021; Avrahami et al., 2022; Kim et al., 2022) perform edits through iterative diffusion sampling, instruction-guided generation (Brooks et al., 2023; Couairon et al., 2022), typically relying on multi-step denoising or latent inversion (Deutch et al., 2024; Huang et al., 2025). Recent advances in diffusion distillation (Meng et al., 2023), consistency models (Song & Dhariwal, 2024), rectified flowEsser et al. (2024), flow matchingKerrigan et al. (2023), and distilled one-step backbones (Liu et al., 2023; Sauer et al., 2024) have enabled few-step and one-step image generation, making real-time image editing possible.

Without iterative denoising, one-step editing relies on a single latent update to perform the desired semantic transformation (Zhang & Hooi, 2023; Wu et al., 2025; 2026). Raw source–target residuals can induce high-energy or spatially unstable updates, which distort image structure and alter unedited regions (Chinchuthakun et al., 2025; Ren et al., 2025; Lu et al., 2026). TurboEdit (Deutch et al., 2024) and InstantEdit (Gong et al., 2025) accelerate editing through efficient few-step diffusion or rectified flow pipelines. SwiftEdit (Nguyen et al., 2025) enables one-step editing via dedicated inversion, whereas FIA-Edit (Yang et al., 2026) improves inversion-free editing through frequency-interactive attention. ChordEdit (Lu et al., 2026) formulates one-step editing as a transport problem with temporally smoothed residual fields. Existing methods still use source–target residuals as heuristic editing directions instead of explicitly solving for a stable control field (Yan et al., 2025; Yi et al., 2026; Li et al., 2026; Cui et al., 2026). As a result, harmful high-frequency perturbations remain insufficiently controlled in one-step editing, without iterative correction during inference.

Contribution. We introduce StableCtrl, a training-free and inversion-free framework for stable one-step text-guided image editing, as illustrated in Fig. 2. Rather than directly applying source–target residuals as editing directions, StableCtrl treats them as noisy control cues and explicitly regularizes their energy, spatial smoothness, and adjacent-timestep consistency within a unified optimal control formulation. The resulting control operator admits a frequency-domain interpretation as a parameterized low-pass filter that selectively suppresses high-frequency components empirically associated with editing artifacts. Experiments on PIE-Bench (Ju et al., 2023) show improved preservation of unedited regions with competitive target-prompt alignment.

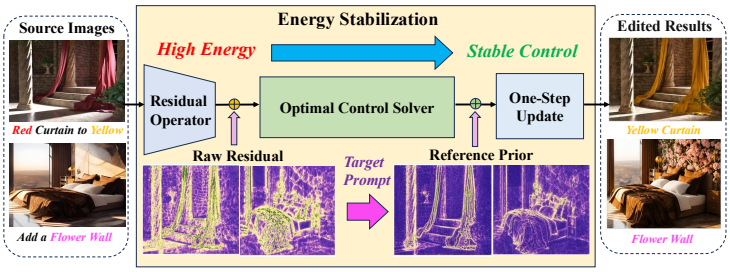

Figure 2: **One-step optimal control with StableCtrl.** StableCtrl regularizes high-energy residuals into stable control fields, enabling one-step editing with improved structural consistency and controlled semantic alignment.

We instantiate these components as follows. **(i) Regularized optimal control.** We formulate one-step image editing as a regularized optimal control problem that treats source–target residuals as noisy control cues, enabling explicit regulation of semantic editing and latent-state stability during one-step inference. **(ii) Elliptic and spectral solution.** We derive a unique elliptic control solution with energy, spatial smoothness, and temporal reference regularization, whose spectral representation characterizes frequency-selective suppression without iterative optimization. **(iii) Training-free one-step integration.** We integrate the resulting control operator into one-step editing without training or inversion, demonstrating improved structural preservation and editing robustness on PIE-Bench.

2 RELATED WORK

Text-Guided Image Editing. Diffusion-based editing has been studied through noise injection and mask estimation (Meng et al., 2021; Couairon et al., 2022), attention manipulation (Hertz et al., 2023), latent inversion (Mokady et al., 2023), and feature reuse (Tumanyan et al., 2023). Training-free

methods further employ attention control (Parmar et al., 2023; Cao et al., 2023), semantic guidance decomposition (Jeong et al., 2026), spatial regularization (Chinchuthakun et al., 2025; Soni et al., 2025), and trajectory modification (Ju et al., 2023; Xu et al., 2023; Kulikov et al., 2025). Other approaches improve reconstruction through inversion or iterative refinement (Wallace et al., 2023; Han et al., 2024; Jiao et al., 2025; Xu et al., 2026; Zhan et al., 2026), typically at additional inference cost. Among spatially regularized methods, LOCATEdit (Soni et al., 2025) optimizes cross-attention maps with a graph Laplacian, whereas our formulation applies elliptic regularization directly to the latent residual control field within a unified regularized optimal control formulation.

One-Step Image Editing. Consistency learning, rectified flow, flow matching, and model distillation enable efficient few-step or one-step generation (Song et al., 2020; Luo et al., 2023; Liu et al., 2022; Lipman et al., 2022; Liu et al., 2023; Sauer et al., 2024). Recent editors build on these advances through few-step diffusion, rectified-flow pipelines, frequency interaction, or inversion (Deutch et al., 2024; Gong et al., 2025; Alimohammadi et al., 2025; Yang et al., 2026; Nguyen et al., 2025). FeedEdit (Fu et al., 2025) regulates editing intensity through frequency-aware feedback across denoising steps, while ChordEdit (Lu et al., 2026) constructs a low-energy, variance-reduced field through dynamic optimal transport. StableCtrl instead regularizes the one-step latent residual using energy, spatial smoothness, and temporal reference terms, yielding an elliptic equation with an explicit spectral solution. Recent methods may also involve inversion, iterative refinement, or multi-stage pipelines (Zhou et al., 2025; Yang & Ye, 2026; Zhu et al., 2026; Elezabi et al., 2026). In contrast, our formulation performs explicit control regularization within a single editing step without iterative correction, enabling direct and efficient residual stabilization.

3 Preliminaries

3.1 One-Step Image Editing

Modern one-step text-to-image models generate an output through a single forward transition, enabling efficient latent editing. Let $x_0 \in \mathbb{R}^{C\times H\times W}$ denote the source latent, and let c_{src} and c_{tar} denote the source and target prompts. A one-step editor applies

$$x_1 = x_0 + u, \tag{1}$$

where $u \in \mathbb{R}^{C\times H\times W}$ is the latent editing field. For analysis, we embed this update into the controlled dynamics

$$\dot{x}(t) = v_\theta\big(x(t), t, c\big) + u(t), \qquad t \in [0, 1], \tag{2}$$

The velocity field v_θ is fixed after pre-training, while $u(t)$ provides the control input. The discrete update u is viewed as the one-step counterpart of $u(t)$. At the same latent state and timestep, the prompt-conditioned velocity difference

$$g(x, t) = v_\theta(x, t, c_{\mathrm{tar}}) - v_\theta(x, t, c_{\mathrm{src}}) \tag{3}$$

defines the source-to-target semantic direction. We regard $g(x, t)$ as the ideal semantic control. During training-free inference, it is estimated from model responses evaluated on perturbed latent samples.

3.2 Vanilla Residual Editing

Let $z \sim K_t(\cdot \mid x_0)$ be sampled from the forward perturbation kernel at timestep t. The observable residual is

$$R(x_0, t) = \mathbb{E}_{z\sim K_t(\cdot|x_0)}\Big[B_t\big(Q(z, t, c_{\mathrm{tar}}) - Q(z, t, c_{\mathrm{src}})\big)\Big], \tag{4}$$

where $Q(z, t, c)$ denotes the prompt-conditioned model response, and B_t projects it into the latent control space. The expectation is approximated with a finite number of noisy samples. We decompose the residual as

$$R(x_0, t) = g(x, t) + \epsilon_t, \tag{5}$$

where the residual term $\epsilon_t = R - g$ collects sampling, discretization, and model approximation errors. Since $\nabla R = \nabla g + \nabla \epsilon_t$, spatial fluctuations in the estimation error can be transferred to the edited latent. Vanilla residual editing sets $u = R$ in Eq. (1), replacing the ideal semantic control with its observable estimate without explicitly controlling its magnitude, spatial variation, or reference consistency. These limitations motivate the regularized control formulation developed in the next section.

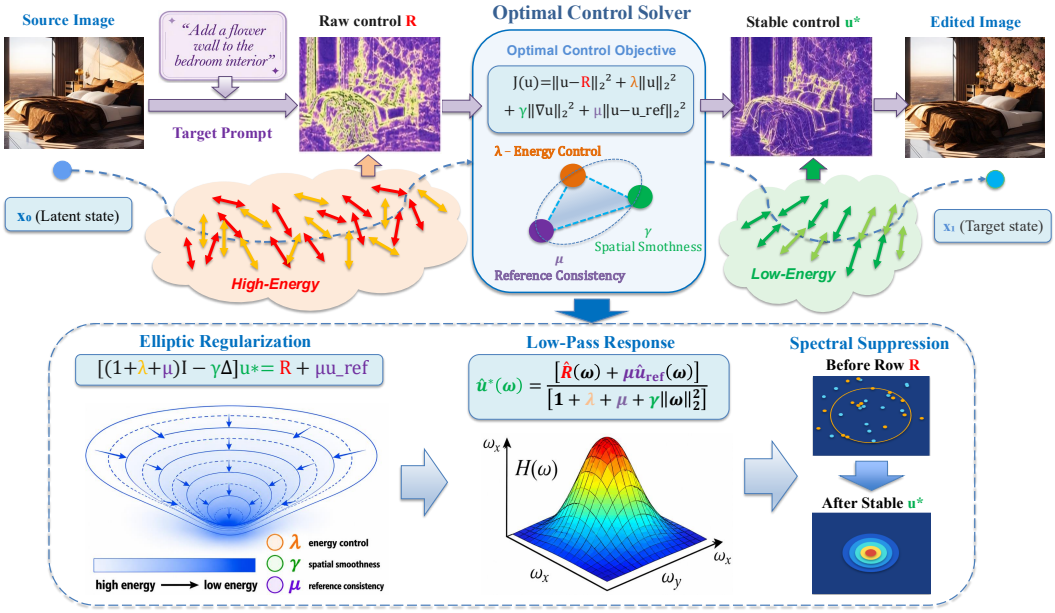

Figure 3: **Overview and theoretical interpretation of StableCtrl. Top:** StableCtrl converts the raw control R into a low-energy and spatially smooth control field u^*. The optimal control solver minimizes the regularized objective and applies the resulting field through a single latent update. **Bottom:** The optimality condition gives the elliptic equation $[(1+\lambda+\mu)I-\gamma\Delta]u^* = R+\mu u_{\text{ref}}$. Its Fourier-domain transfer function is $H(\omega) = (1+\lambda+\mu+\gamma\|\omega\|_2^2)^{-1}$. This response suppresses unstable high-frequency components while retaining the dominant editing structure.

4 StableCtrl

4.1 Optimal Control Formulation

As illustrated in Figure 3, StableCtrl transforms the raw semantic residual into an energy-regularized and spatially coherent control field. Let $x(t)$ denote the controlled latent trajectory governed by Eq. (2), and let $x_0(t)$ denote the corresponding trajectory without the additional control. We consider admissible controls in $\mathcal{U} = L^2\big(0,1;H^1(\Omega;\mathbb{R}^C)\big)$ and formulate the continuous-time control problem through the Bolza functional

$$J_{\text{cont}}(u) = \Phi\big(x(1), x_0(1), R\big) + \int_0^1 \mathcal{R}\big(u(t)\big)\,dt, \tag{6}$$

where the terminal cost measures the discrepancy between the control-induced terminal displacement and the semantic displacement specified by R. The running cost regularizes the control along the latent trajectory and is defined as

$$\mathcal{R}\big(u(t)\big) = \lambda\|u(t)\|^2_{L^2(\Omega)} + \gamma\|\nabla u(t)\|^2_{L^2(\Omega)} + \mu\|u(t)-u_{\text{ref}}(t)\|^2_{L^2(\Omega)}, \tag{7}$$

where $\lambda,\gamma,\mu \geq 0$. The three terms respectively penalize control energy, spatial variation, and deviation from the reference field u_{ref}, with $\|\nabla u\|^2_{L^2} = \int_\Omega \|\nabla u(s)\|_F^2\,ds$. This formulation separates semantic agreement at the terminal state from stability regularization along the control trajectory.

For one-step editing, we restrict the admissible set to time-independent controls, $u(t) \equiv u$ and $u_{\text{ref}}(t) \equiv u_{\text{ref}}$ for $t \in [0,1]$. Since the interval has unit length, the running cost reduces to $\int_0^1 \mathcal{R}(u)\,dt = \mathcal{R}(u)$. Under a local one-step approximation, the displacement induced by the additional control satisfies $x(1)-x_0(1) \approx u$, with the update scale absorbed into the control variable. Because R represents the desired source-to-target semantic displacement relative to the uncontrolled trajectory, we adopt the quadratic terminal surrogate

$$\Phi\big(x(1),x_0(1),R\big) \approx \|(x(1)-x_0(1))-R\|^2_{L^2} \approx \|u-R\|^2_{L^2}.$$

Under this local approximation, substituting the surrogate cost into Eq. (6) yields the one-step objective

$$J(u) = \|u - R\|^2_{L^2(\Omega)} + \lambda\|u\|^2_{L^2(\Omega)} + \gamma\|\nabla u\|^2_{L^2(\Omega)} + \mu\|u - u_{\text{ref}}\|^2_{L^2(\Omega)}, \tag{8}$$

where the data term preserves agreement with the semantic residual R, while the remaining terms regulate control energy, spatial smoothness, and reference consistency. Setting $\lambda = \gamma = \mu = 0$ recovers the direct residual solution $u^* = R$, showing that direct residual injection is a special case of the proposed formulation. The analysis characterizes this one-step surrogate, whose unique elliptic optimality condition is derived in the following subsection.

4.2 Elliptic Optimality Condition

For $\lambda, \gamma, \mu \geq 0$, the objective J in Eq. (8) is strictly convex. For any nonzero $v \in H^1(\Omega; \mathbb{R}^C)$, its second variation satisfies $2(1+\lambda+\mu)\|v\|^2_{L^2} + 2\gamma\|\nabla v\|^2_{L^2} > 0$, which ensures uniqueness whenever a minimizer exists. Differentiating $J(u^* + \varepsilon v)$ with respect to ε at $\varepsilon = 0$ yields the weak optimality condition

$$0 = \langle u^* - R, v\rangle + \lambda\langle u^*, v\rangle + \gamma\langle \nabla u^*, \nabla v\rangle + \mu\langle u^* - u_{\text{ref}}, v\rangle, \qquad \forall v \in H^1(\Omega; \mathbb{R}^C), \tag{9}$$

where all inner products are taken over $L^2(\Omega)$.

Under homogeneous Neumann boundary conditions, Green's identity gives $\langle \nabla u^*, \nabla v\rangle = -\langle \Delta u^*, v\rangle$; the same adjoint relation holds under periodic boundary conditions. Substituting this relation into Eq. (9) and using the arbitrariness of v yields the elliptic optimality equation

$$\big[(1 + \lambda + \mu)I - \gamma\Delta\big]u^* = R + \mu u_{\text{ref}}, \tag{10}$$

where the positive zeroth-order coefficient eliminates the constant-function null space associated with the Neumann Laplacian.

Let $\mathcal{L} = (1 + \lambda + \mu)I - \gamma\Delta$. For $\gamma > 0$, the associated bilinear form $a(u, v) = (1 + \lambda + \mu)\langle u, v\rangle + \gamma\langle \nabla u, \nabla v\rangle$ is continuous and coercive on $H^1(\Omega; \mathbb{R}^C)$, with $a(u, u) \geq \min\{1 + \lambda + \mu, \gamma\}\|u\|^2_{H^1}$. The Lax–Milgram theorem therefore guarantees the existence and uniqueness of a weak solution to Eq. (10). For $\gamma = 0$, the operator reduces to multiplication by the positive scalar $1 + \lambda + \mu$, and the solution remains unique. Thus, in both cases, the optimal control can be represented as

$$u^* = \mathcal{L}^{-1}\big(R + \mu u_{\text{ref}}\big), \tag{11}$$

where $\mathcal{L}^{-1}$ denotes the corresponding solution operator rather than an explicitly formed matrix inverse. This operator maps the semantic residual and reference prior to a unique, energy-regularized, and spatially coherent control field.

4.3 Spectral Filtering

Figure 3 illustrates the spectral interpretation of Eq. (11). Applying the Fourier transform to Eq. (10) and using $\mathcal{F}[\Delta u](\omega) = -\|\omega\|_2^2\hat{u}(\omega)$ converts the elliptic optimality equation into an algebraic relation at each frequency. Since $1 + \lambda + \mu + \gamma\|\omega\|_2^2 > 0$ for $\lambda, \gamma, \mu \geq 0$, the solution is uniquely determined as

$$\hat{u}^*(\omega) = \frac{\hat{R}(\omega) + \mu\hat{u}_{\text{ref}}(\omega)}{1 + \lambda + \mu + \gamma\|\omega\|_2^2}, \tag{12}$$

where the corresponding spectral gain is $H(\omega) = (1 + \lambda + \mu + \gamma\|\omega\|_2^2)^{-1}$. Because H depends only on $r = \|\omega\|_2$, the induced filtering is isotropic. For $\gamma > 0$, $\partial H/\partial r^2 = -\gamma H^2 < 0$, showing that the gain decreases monotonically with spatial frequency. StableCtrl therefore progressively attenuates rapidly varying residual components while retaining dominant low-frequency structures.

The spectral representation further provides an L^2-control stability bound. Using Parseval's identity together with $H(\omega) \leq (1 + \lambda + \mu)^{-1}$, we obtain

$$\|u^*\|_{L^2} \leq \frac{\|R\|_{L^2} + \mu\|u_{\text{ref}}\|_{L^2}}{1 + \lambda + \mu}, \tag{13}$$

which shows that the optimized control field cannot arbitrarily amplify the input residual energy. The parameter λ uniformly reduces the spectral gain, while γ controls its frequency-dependent decay. The parameter μ incorporates the reference prior while simultaneously increasing attenuation through the denominator. Together, these complementary effects yield a bounded and spatially coherent control field with frequency-selective suppression of unstable components.

Table 1: **Quantitative comparison on PIE-Bench under the no-cleanup protocol.** We compare multi-step, few-step, and one-step editing methods in terms of structural consistency, semantic alignment, and inference efficiency. T/I-free denotes training-free/inversion-free. Best results are shown in **bold**, and second-best results are underlined.

Type	Method	Structure			CLIP		Property	Efficiency			
		PSNR↑	MSE $\times 10^3$ ↓	LPIPS $\times 10^3$ ↓	Whole↑	Edit.↑	T/I-free	Step	NFE	Time (s)↓	VRAM (MiB)↓
Multi-step ≥ 30	DDIM + MasaCtrl (Cao et al., 2023)	21.25	8.58	106.59	24.13	21.13	✓/✗	50	100	55.20	12272
	Direct Inv. + MasaCtrl (Cao et al., 2023)	21.78	7.99	87.38	24.42	21.38	✓/✗	50	100	79.10	12272
	DDIM + PnP (Tumanyan et al., 2023)	21.26	8.42	113.58	25.45	22.54	✓/✗	50	100	28.01	9262
	Direct Inv. + PnP (Tumanyan et al., 2023)	21.43	8.10	106.26	25.48	22.63	✓/✗	50	100	28.03	9262
	SDEdit (Meng et al., 2021)	20.91	10.82	121.35	23.88	20.94	✓/✗	50	100	28.50	10240
	Prompt-to-Prompt (Hertz et al., 2023)	21.67	9.34	110.28	24.92	21.71	✓/✗	50	100	26.10	11050
	FlowEdit (Kulikov et al., 2025)	22.17	7.69	104.81	**26.64**	**23.69**	✓/✓	33	33	7.22	17140
Few-step 4	TurboEdit (Deutch et al., 2024)	21.44	9.49	108.60	24.66	21.79	✓/✓	4	4	2.69	13826
	InfEdit (Xu et al., 2023)	24.14	6.82	55.69	24.89	21.88	✓/✓	4	4	1.41	6502
	InstantEdit (Gong et al., 2025)	23.80	4.21	60.92	24.97	21.82	✓/✗	4	8	1.30	16270
One-step	SwiftEdit (Nguyen et al., 2025)	21.71	8.22	91.22	24.93	21.85	✗/✗	1	2	0.54	15060
	ChordEdit (Lu et al., 2026)	22.04	7.13	111.22	25.12	22.58	✓/✓	1	2	0.38	6988
	ChordEdit (w/o prox.) (Lu et al., 2026)	23.80	6.56	89.20	24.11	21.88	✓/✓	1	1	0.43	**5914**
	StableCtrl	**27.46**	**2.86**	**55.55**	23.34	21.26	✓/✓	1	1	**0.26**	**5914**

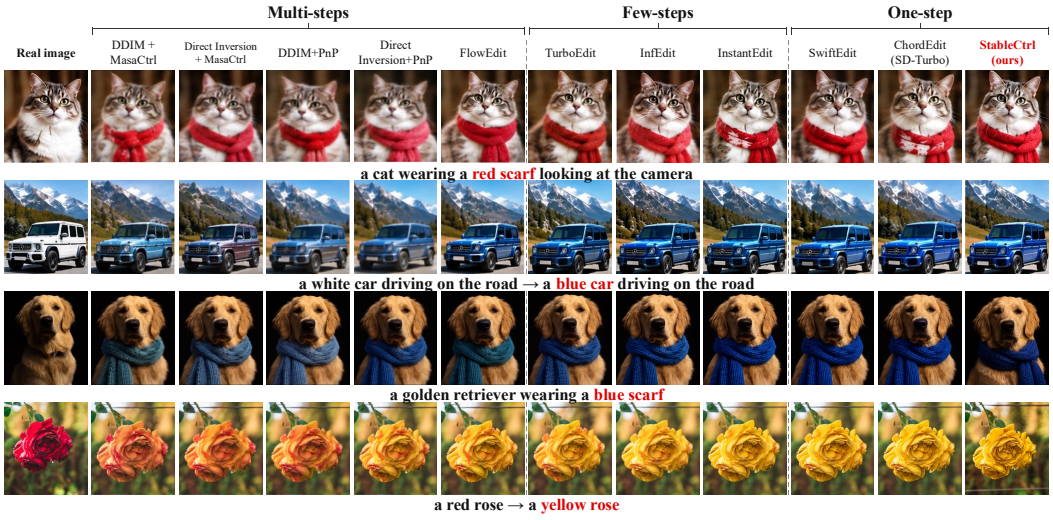

Figure 4: **Qualitative comparison across inference regimes.** We compare representative multi-step, few-step, and one-step methods for real-image editing. StableCtrl achieves the intended edits while better preserving object structure and scene content in a single step.

5 Experiments

5.1 Experimental Setup

Dataset and Evaluation Metrics. Experiments are conducted on PIE-Bench (Ju et al., 2023), a standard benchmark for text-guided image editing. The benchmark contains 700 editing pairs spanning ten edit categories, including object, attribute, color, style, and background editing. Each sample provides a source image, source and target prompts, and a ground-truth editing mask. We follow the official evaluation protocol. Editing fidelity is evaluated using PSNR, LPIPS (Zhang et al., 2018), and structure distance. Semantic alignment with the target prompt is measured by CLIP-Edited (Radford et al., 2021). We also report the average inference time and GPU memory.

Implementation Details. All experiments use pretrained SD-Turbo (Sauer et al., 2024) without additional training or fine-tuning. Following ChordEdit (Lu et al., 2026), we use shared-noise Monte

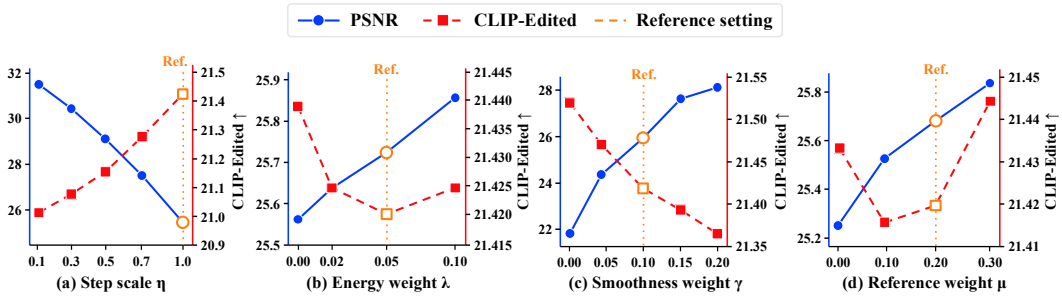

Figure 5: **Controlled-variable analysis of StableCtrl.** Each panel varies one hyperparameter while fixing the others. Blue solid and red dashed curves show PSNR and CLIP-Edited, respectively. Orange markers denote the selected setting within this controlled-variable study.

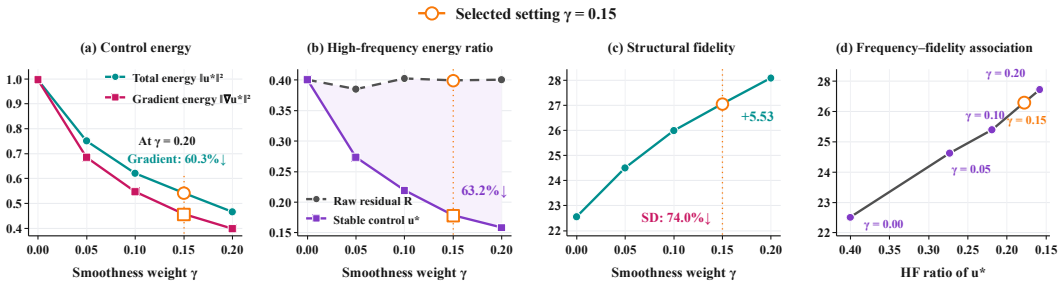

Figure 6: **Frequency and energy analysis of StableCtrl.** Panels (a)–(d) analyze the effects of γ on control energy, frequency characteristics, and structural fidelity. Orange markers denote the adopted setting ($\gamma = 0.15$). This trend is consistent with the spectral filtering interpretation.

Carlo estimation with $n = 1$ by default. Unless otherwise specified, the default configuration used for evaluation is $\lambda = 0.08$, $\gamma = 0.15$, $\mu = 0.25$, $\eta = 0.9$, and $\delta = 0.15$, with $u_{\text{ref}} = R_{t-\delta}$. For the controlled-variable study in Fig. 5, we independently evaluate one-at-a-time sweeps around $(\eta, \lambda, \gamma, \mu) = (1.0, 0.05, 0.10, 0.20)$, allowing minor run-to-run variation across independently evaluated sweeps while isolating individual effects. NFE counts sequential model evaluations rather than individual conditional branches, which are computed jointly in a shared batch; StableCtrl thus requires exactly one NFE. All methods use identical metrics, random seed, and PIE-Bench masks under the no-cleanup protocol. Experiments use PyTorch on a single NVIDIA RTX 4090 GPU.

5.2 Results and Analysis

Comparison with State-of-the-Art Methods. Table 1 compares multi-step, few-step, and one-step editors on PIE-Bench under the no-cleanup protocol. Among one-step methods, our method achieves the highest structural fidelity, with a lower CLIP-E score reflecting the trade-off between structural preservation and semantic alignment. Using the same SD-Turbo backbone, it improves over ChordEdit (w/o prox.) from 23.80 to 27.46 in PSNR and reduces LPIPS from 89.20×10^{-3} to 55.55×10^{-3}. Runtime decreases from 0.43 s to 0.26 s with the same 5914 MiB VRAM. The full ChordEdit configuration is also included for comparison. Inference requires one elliptic solve and one latent update, without inversion or iterative optimization. Figure 1 compares the one-step results with SwiftEdit (Nguyen et al., 2025) and ChordEdit (Lu et al., 2026). Across different edit types, the results show better preservation of object boundaries, scene layout, and unedited regions while retaining the intended semantic changes across a range of editing scenarios.

Hyperparameter Sensitivity. Figure 5 independently varies each hyperparameter around the reference configuration $\eta = 1.0$, $\lambda = 0.05$, $\gamma = 0.10$, and $\mu = 0.20$. Increasing η strengthens semantic editing but gradually reduces structural fidelity, whereas increasing λ slightly improves PSNR with little change in CLIP-E. The smoothness weight γ produces the largest fidelity gain, highlighting the importance of spatial regularization. Increasing μ improves structural preservation with only minor changes in semantic alignment. Overall, the selected configuration, marked in orange, provides a favorable balance between structural fidelity and semantic alignment across the evaluated settings.

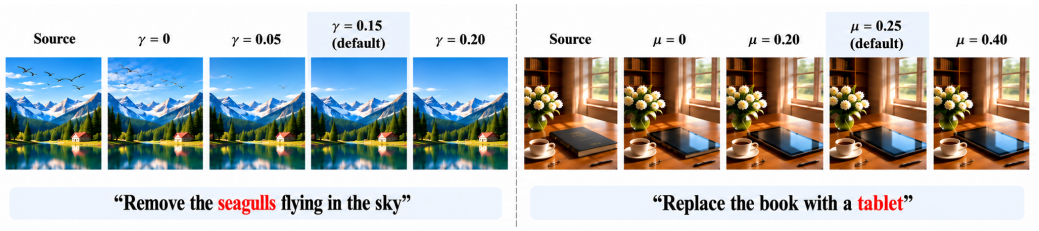

Figure 7: **Qualitative analysis of key regularization weights.** We visualize the effects of the smoothness weight γ and reference weight μ across representative editing tasks. The adopted settings provide a favorable balance between effective semantic editing and structural preservation.

Table 2: **Component ablation of StableCtrl.** Bold values indicate the best result for each evaluation metric across all settings.

Setting	Structural Preservation			Semantic
	PSNR↑	LPIPS$\times 10^3$ ↓	SD$\times 10^3$ ↓	CLIP-E↑
w/o γ-term	21.50	122.00	16.80	**21.94**
w/o λ-term	25.54	76.00	7.00	21.44
w/o μ-term	25.23	81.00	7.90	21.44
w/o ref. signal	25.70	74.00	**6.60**	21.42
StableCtrl	**27.46**	**56.00**	7.40	21.26

Table 3: **Ablation of reference-prior design.** Alternative reference constructions are compared under the same controlled setting, separately from the objective-component ablation in Table 2.

Reference	PSNR↑	LPIPS$\times 10^3$ ↓	CLIP-E↑	Time (s)↓
Reference-free	$\mathbf{27.50 \pm 0.15}$	$\mathbf{56.04 \pm 2.15}$	21.22 ± 0.02	0.233 ± 0.002
Current R_t	26.24 ± 0.19	70.28 ± 2.94	$\mathbf{21.32 \pm 0.02}$	$\mathbf{0.231 \pm 0.000}$
Future $R_{t+\delta}$	26.85 ± 0.18	61.85 ± 2.50	21.29 ± 0.02	0.288 ± 0.001
Bidirectional	27.13 ± 0.17	59.17 ± 2.31	21.27 ± 0.02	0.289 ± 0.000
Random	26.88 ± 0.12	61.47 ± 0.09	21.24 ± 0.01	0.232 ± 0.001
Previous $R_{t-\delta}$	27.29 ± 0.15	57.91 ± 2.16	21.27 ± 0.02	0.232 ± 0.001

Frequency and Energy Analysis. Figure 6 studies the effect of the smoothness weight γ, with $\lambda = 0.08$, $\mu = 0.25$, and $\eta = 0.9$. As γ increases, both the total energy $\|u^*\|_2^2$ and gradient energy $\|\nabla u^*\|_2^2$ decrease, together with the high-frequency ratio of the optimized control field. In contrast, the high-frequency ratio of the raw residual R remains nearly unchanged because regularization is applied only when solving for u^*. Panels (c) and (d) further show that reduced high-frequency energy is associated with higher PSNR and lower structural deviation. These results support the spectral filtering interpretation in Section 4.3. We select $\gamma = 0.15$, marked in orange, based on the structure–semantic trade-off in the controlled-variable analysis.

Ablation of Objective Terms. Table 2 evaluates the contribution of individual components in the StableCtrl objective. Specifically, w/o γ sets $\gamma = 0$, w/o λ sets $\lambda = 0$, and w/o μ-term sets $\mu = 0$, thereby removing the corresponding regularization term. For the reference-signal ablation, we remove the temporal reference information while retaining the associated regularization strength. All remaining components and inference settings are kept unchanged from the full StableCtrl configuration. Removing the smoothness term causes the largest degradation in structural fidelity, highlighting the importance of spatial regularization. The energy and reference components provide complementary constraints on the control field, with individual removals yielding different trade-offs across structural metrics. The complete objective achieves the best overall PSNR and LPIPS while maintaining competitive structural fidelity and semantic alignment across diverse editing scenarios.

Reference-Prior Design. Table 3 separately studies how different reference constructions affect the structure–semantic trade-off. Unlike the objective-term ablation in Table 2, this experiment evaluates alternative reference constructions under a separate controlled setting and therefore does not correspond to removing the reference component from the full objective. We consider a reference-free baseline together with current, future, bidirectional, random, and previous-residual references. The reported $\pm$ values indicate the observed run-to-run variation across repeated evaluations rather than standard deviation. Among these alternatives, the previous residual $R_{t-\delta}$ provides a favorable balance between structural preservation and semantic alignment, and is therefore adopted in StableCtrl.

Temporal Reference Interval. We vary the interval δ used to construct $R_{t-\delta}$, under a separate controlled parameter study. As shown in Table 4, increasing δ consistently improves structural fidelity, with diminishing gains beyond $\delta = 0.15$, while CLIP-E decreases slightly. This trend further indicates that the reference interval directly controls the practical trade-off between structural preservation and semantic alignment. We therefore use $\delta = 0.15$ by default, which captures most of the structural improvement without further reducing semantic alignment.

Table 4: **Effect of the reference interval δ.** † denotes the default setting used throughout the main experiments; bold indicates the best result.

δ	PSNR↑	LPIPS$\times 10^3$ ↓	SD$\times 10^3$ ↓	CLIP-E↑
0.05	26.87 ± 0.17	62.04 ± 2.49	8.45 ± 0.56	$\mathbf{21.32 \pm 0.02}$
0.10	27.15 ± 0.16	59.22 ± 2.24	8.30 ± 0.41	21.29 ± 0.02
$0.15^{\dagger}$	27.29 ± 0.15	57.91 ± 2.16	7.89 ± 0.46	21.27 ± 0.02
0.20	27.36 ± 0.15	57.29 ± 2.13	7.68 ± 0.43	21.24 ± 0.03
0.30	$\mathbf{27.45 \pm 0.15}$	$\mathbf{56.55 \pm 2.14}$	$\mathbf{7.46 \pm 0.44}$	21.24 ± 0.02

Table 5: **Effect of the Monte Carlo sample count.** †/‡ denote the default/recommended settings; bold indicates the best result.

n	PSNR↑	LPIPS$\times 10^3$ ↓	CLIP-E↑	Time (s)↓
$1^{\dagger}$	27.29 ± 0.15	57.91 ± 2.16	$\mathbf{21.27 \pm 0.02}$	$\mathbf{0.233 \pm 0.003}$
2	28.62 ± 0.04	43.98 ± 0.29	21.15 ± 0.01	0.325 ± 0.001
$4^{\ddagger}$	29.52 ± 0.01	36.50 ± 0.14	21.10 ± 0.02	0.522 ± 0.001
8	$\mathbf{30.14 \pm 0.03}$	$\mathbf{33.00 \pm 0.03}$	21.07 ± 0.01	0.912 ± 0.001



Figure 8: **Visualization of the StableCtrl inference process.** From left to right, we show the source image, raw control field R, reference prior u_{ref}, optimized control u^*, and edited result. The reference prior retains the dominant spatial structure, while the elliptic solver suppresses scattered high-frequency responses to produce a smoother, spatially coherent control field.

Effect of the Monte Carlo Sample Count. We vary the number of shared-noise Monte Carlo samples $n \in \{1, 2, 4, 8\}$ in a controlled study. As shown in Table 5, increasing n improves structural fidelity at the cost of inference time, with diminishing gains beyond $n = 4$ and only small changes in CLIP-E. We use $n = 1$ by default for efficient inference and $n = 4$ when additional structural fidelity is preferred, providing a flexible quality–efficiency trade-off in practice.

Control-Field Visualization. Figure 8 illustrates how the reference prior and elliptic regularization reshape the control field. The raw residual R contains scattered high-frequency responses, while $u_{\mathrm{ref}} = R_{t-\delta}$ retains the dominant spatial structure from a temporally adjacent reference. The optimized field u^* attenuates irregular high-frequency components while preserving spatially coherent editing signals, resulting in the intended semantic change with improved source structure preservation.

6 Conclusion

StableCtrl formulates one-step image editing as a regularized control problem that converts semantic residuals into spatially smooth control fields. The resulting elliptic optimality equation admits an explicit spectral solution, providing a training-free and inversion-free mechanism for regulating one-step updates. The control operator attenuates high-frequency residual components, while the energy, smoothness, and reference terms jointly constrain updates that cannot be progressively refined through multiple denoising steps. Experiments on PIE-Bench demonstrate improved structural fidelity while retaining semantic editing capability and efficient one-step inference. Overall, StableCtrl provides a regularized control framework for stabilizing residual-based one-step image editing.

Limitations. StableCtrl nevertheless relies on a local control approximation and an isotropic smoothness prior. As a result, performance may degrade particularly for large geometric changes, thin structures, or inaccurate residual estimates. Moreover, the current formulation uses fixed regularization parameters across spatial locations and highly complex editing scenarios, which may limit its adaptability to spatially heterogeneous edits. Spatially adaptive regularization, anisotropic control operators, and trajectory-level control therefore offer promising directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and assistance with code development. All AI-assisted content, mathematical derivations, experimental implementations, and generated text were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies inference-time control for pretrained text-to-image generative models and does not introduce new training data or collect human-subject data. The proposed method inherits potential risks and biases from the underlying pretrained generative models.

REPRODUCIBILITY STATEMENT

We provide the complete mathematical formulation, optimization procedure, and implementation details required to reproduce the proposed method. Hyperparameter settings, evaluation protocols, and additional experimental results are provided in the main paper and appendix. Code will be released upon publication.

REFERENCES

Amirhossein Alimohammadi, Aryan Mikaeili, Sauradip Nag, Negar Hassanpour, Andrea Tagliasacchi, and Ali Mahdavi-Amiri. Cora: Correspondence-aware image editing using few step diffusion. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18208–18218, 2022.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

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*, pp. 22560–22570, 2023.

Worameth Chinchuthakun, Tossaporn Saengja, Nontawat Tritrong, Pitchaporn Rewatbowornwong, Pramook Khungurn, and Supasorn Suwajanakorn. LUSD: Localized update score distillation for text-guided image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

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

Haowang Cui, Rui Chen, Tao Luo, Tao Guo, Zheng Qin, and Jiaze Wang. Sam-flow: Source-anchored masked flow for training-free image editing. *arXiv preprint arXiv:2606.06228*, 2026.

Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models. In *SIGGRAPH Asia Conference*, pp. 1–12, 2024.

Omar Elezabi, Eduard Zamfir, Zongwei Wu, and Radu Timofte. Language-free generative editing from one visual example. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1078–1088, 2026.

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

Fengyi Fu, Lei Zhang, Mengqi Huang, and Zhendong Mao. Feededit: Text-based image editing with dynamic feedback regulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2670, 2025.

Yiming Gong, Zhen Zhu, and Minjia Zhang. Instantedit: Text-guided few-step image editing with piecewise rectified flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16808–16817, 2025.

Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4291–4301, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

Wootaek Jeong, Junghyo Sohn, Jee Seok Yoon, and Heung-Il Suk. Deconstructing guidance: A semantic hierarchy for precise diffusion model editing. In *International Conference on Learning Representations*, 2026.

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

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.

Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Functional flow matching. *arXiv preprint arXiv:2305.17209*, 2023.

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*, pp. 2426–2435, 2022.

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

Minghan Li, Jeremy Moebel, and Mengyu Wang. Rethinking one-step image editing through chordedit: Reproduction, simplification, and new insights. *arXiv preprint arXiv:2606.14042*, 2026.

Ruibin Li, Ruihuang Li, Song Guo, and Lei Zhang. Source prompt disentangled inversion for boosting image editability with diffusion models. In *European Conference on Computer Vision*, pp. 404–421. Springer, 2024.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2023.

Liangsi Lu, Xuhang Chen, Minzhe Guo, Shichu Li, Jingchao Wang, and Yang Shi. Chordedit: One-step low-energy transport for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14407, 2026.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023.

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

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

Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, and Cuong Pham. Swiftedit: Lightning fast text-guided image editing via one-step diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21492–21501, 2025.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH Conference*, pp. 1–11, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yufan Ren, Zicong Jiang, Tong Zhang, Søren Forchhammer, and Sabine Süsstrunk. Fds: Frequency-aware denoising score for text-guided latent diffusion image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2651–2660, 2025.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

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

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024.

Achint Soni, Meet Soni, and Sirisha Rambhatla. LOCATEdit: Graph laplacian optimized cross attention for localized text-guided image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. doi: 10.1109/ICCV51701.2025.01747.

Nikita Starodubcev, Mikhail Khoroshikh, Artem Babenko, and Dmitry Baranchuk. Invertible consistency distillation for text-guided image editing in around 7 steps. *Advances in Neural Information Processing Systems*, 37:12496–12527, 2024.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

Unsplash. Unsplash lite dataset. https://huggingface.co/datasets/sentence-transformers/unsplash-lite, 2022. Accessed July 2026.

Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22532–22541, 2023.

Junyi Wu, Zhiteng Li, Haotong Qin, Xiaohong Liu, Linghe Kong, Yulun Zhang, and Xiaokang Yang. Flashedit: Decoupling speed, structure, and semantics for precise image editing. *arXiv preprint arXiv:2509.22244*, 2025.

Tao Wu, Senmao Li, Yaxing Wang, Shiqi Yang, Kai Wang, and Joost van de Weijer. Training-free image inversion for one-step diffusion models. *Pattern Recognition*, pp. 114063, 2026.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. *arXiv preprint arXiv:2312.04965*, 2023.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9452–9461, 2024.

Yangyang Xu, Wenqi Shao, Yong Du, Haiming Zhu, Yang Zhou, Jiayuan Xie, Ping Luo, and Shengfeng He. Invert your prompt: Editing-aware diffusion inversion. *International Journal of Computer Vision*, 134(4):163, 2026.

Zexuan Yan, Yue Ma, Chang Zou, Wenteng Chen, Qifeng Chen, and Linfeng Zhang. Eedit: Rethinking the spatial and temporal redundancy for efficient image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17474–17484, 2025.

Desong Yang and Mang Ye. Directedit: Step-level accurate inversion for flow-based image editing. *arXiv preprint arXiv:2605.02417*, 2026.

Kaixiang Yang, Boyang Shen, Xin Li, Yuchen Dai, Yuxuan Luo, Yueran Ma, Wei Fang, Qiang Li, and Zhiwei Wang. Fia-edit: Frequency-interactive attention for efficient and high-fidelity inversion-free text-guided image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11613–11621, 2026.

Hongzhu Yi, Zhongtian Luo, Tong Li, Yiyan Fan, and Jungang Xu. Bridging the manifold gap: Riemannian residual line search for one-step image editing. *arXiv preprint arXiv:2606.24844*, 2026.

Zheyuan Zhan, Hongchen Li, Can Wang, Yinfei Ma, Mingzhen Huang, Ruoshi Bai, Jiawei Chen, Siwei Lyu, and Defang Chen. Conditioning matters: Stabilizing inversion and attention in diffusion image editing. *arXiv preprint arXiv:2606.14125*, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yifan Zhang and Bryan Hooi. Hipa: Enabling one-step text-to-image diffusion models via high-frequency-promoting adaptation. *arXiv preprint arXiv:2311.18158*, 2023.

Zijun Zhou, Yingying Deng, Xiangyu He, Weiming Dong, and Fan Tang. Multi-turn consistent image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15792–15801, 2025.

Zifeng Zhu, Jiaming Han, Jiaxiang Zhao, Minnan Luo, and Xiangyu Yue. Gide: Unlocking diffusion llms for precise training-free image editing. *arXiv preprint arXiv:2603.21176*, 2026.

StableCtrl: One-Step Stable Image Editing via Optimal Control Supplementary Material

SUPPLEMENTARY CONTENTS

A StableCtrl: Variational Control for Stable One-Step Editing **14**

A.1 Continuous Optimal Control 14

A.2 Running Cost Design 15

A.3 One-Step Approximation 16

B Variational Derivation **16**

B.1 First-Order Optimality Condition 16

C Elliptic Analysis **17**

C.1 Well-Posedness of the Elliptic Operator 17

C.2 Characterization of the Solution 18

C.3 Fourier-Domain Solution 18

C.4 Spectral Properties 19

D Additional Ablation Results **20**

D.1 Comparison with Simple Regularization Baselines 20

D.2 Cross-Backbone Evaluation 20

D.3 Category-wise Analysis 21

D.4 Qualitative Analysis of Key Hyperparameters 22

E Additional Qualitative Results **23**

A STABLECTRL: VARIATIONAL CONTROL FOR STABLE ONE-STEP EDITING

A.1 CONTINUOUS OPTIMAL CONTROL

We interpret StableCtrl as an external control applied to the latent dynamics of a pretrained generator. A continuous optimal-control formulation is first introduced, after which the one-step approximation and residual-based surrogate lead to the discrete variational objective used in the main paper.

Controlled Latent Dynamics. Let $x(t) \in \mathcal{X} \subseteq \mathbb{R}^d$, $t \in [0,1]$, denote the vectorized latent state of a pretrained text-conditioned generator. Without editing control, its trajectory follows $\dot{x}(t) = v_\theta(x(t), t, c)$ from $x(0) = x_{\text{src}}$, where v_θ is the pretrained velocity field and c is the text condition. To introduce spatial regularization, we associate the vectorized latent state with a field over Ω and take the admissible controls from $\mathcal{U} = L^2([0,1]; H^1(\Omega; \mathbb{R}^C))$, where C denotes the number of latent channels. The editing control $u(t)$ then acts additively on the pretrained dynamics:

$$\underbrace{\dot{x}(t)}_{\text{controlled transport}} = \underbrace{v_\theta\big(x(t), t, c\big)}_{\text{pretrained dynamics}} + \underbrace{u(t)}_{\text{editing control}}, \qquad x(0) = x_{\text{src}}. \tag{A.1}$$

After spatial discretization, $u(t)$ has the same representation as the latent velocity field. Integrating Eq. (A.1) from 0 to t expresses the controlled state as the accumulated effect of the pretrained

transport and external control. Let x_u and x_0 denote the controlled and uncontrolled trajectories, respectively. Subtracting their integral forms yields

$$x_u(t) - x_0(t) = \underbrace{\int_0^t u(\tau)\, d\tau}_{\text{direct control effect}} + \underbrace{\int_0^t \Delta v_\theta(\tau)\, d\tau}_{\text{induced dynamics}}, \tag{A.2}$$

where $\Delta v_\theta(\tau) = v_\theta(x_u(\tau), \tau, c) - v_\theta(x_0(\tau), \tau, c)$. Equation (A.2) therefore separates the direct displacement induced by the control from the indirect change caused by subsequent evaluations of the state-dependent velocity field.

Continuous Optimal-Control Formulation. The controlled dynamics above provide the state equation, while the editing objective determines which admissible control should be selected. We formulate text-guided editing as a constrained optimal-control problem that balances terminal target agreement with regularization of the control trajectory:

$$\begin{aligned} \min_{u \in \mathcal{U}} \quad & \underbrace{\Phi\big(x(1)\big)}_{\text{terminal agreement}} + \underbrace{\int_0^1 \mathcal{R}\big(u(t), t\big)\, dt}_{\text{control regularization}} \\ \text{s.t.} \quad & \dot{x}(t) = v_\theta\big(x(t), t, c\big) + u(t), \qquad x(0) = x_{\text{src}}. \end{aligned} \tag{A.3}$$

Here, Φ measures agreement of the terminal state with the desired editing target, whereas $\mathcal{R}$ regulates the admissible control trajectory. Equation (A.3) gives the continuous problem underlying StableCtrl, but solving it directly would require repeated evaluations of the controlled generator trajectory. StableCtrl therefore replaces the terminal objective by a local source–target residual surrogate and adopts the constant-control approximation $u(t) \equiv u$, leading to the discrete variational objective developed in the main paper.

A.2 Running Cost Design

The running cost constrains control magnitude, spatial variation, and reference consistency. These terms provide the regularization component of StableCtrl, while residual fidelity is introduced under the one-step surrogate formulation.

Regularized running cost. Let $V := H^1(\Omega; \mathbb{R}^C)$, where $\Omega \subset \mathbb{R}^2$, and let $u(t), u_{\text{ref}}(t) \in V$ for a.e. $t \in [0, 1]$. We regularize three complementary properties of the control field: its overall magnitude, spatial variation, and consistency with a reference control. The resulting running cost is

$$\mathcal{R}(u, t) = \underbrace{\lambda \|u(t)\|_{L^2}^2}_{\text{control energy}} + \underbrace{\gamma \|\nabla u(t)\|_{L^2}^2}_{\text{spatial smoothness}} + \underbrace{\mu \|u(t) - u_{\text{ref}}(t)\|_{L^2}^2}_{\text{reference consistency}}, \tag{A.4}$$

where all norms are taken over Ω. Equivalently, writing $E_{\text{eng}} := \|u\|_{L^2}^2$, $E_{\text{spa}} := \|\nabla u\|_{L^2}^2$, and $E_{\text{ref}} := \|u - u_{\text{ref}}\|_{L^2}^2$, the cost is $\mathcal{R} = \lambda E_{\text{eng}} + \gamma E_{\text{spa}} + \mu E_{\text{ref}}$. Accordingly, Eq. (A.4) uses λ to limit the overall control magnitude, γ to penalize rapid spatial variation, and μ to encourage consistency with the reference control field. Together, these terms discourage excessively strong or spatially irregular latent updates.

Effective reference regularization. The magnitude and reference penalties are not independent quadratic effects: together, they can be interpreted as a single penalty centered at an effective reference field. For $s := \lambda + \mu > 0$, define $\alpha := \mu/s$ and $u_{\text{eff}} := \alpha u_{\text{ref}}$. Completing the square yields

$$\underbrace{\lambda \|u\|_{L^2}^2 + \mu \|u - u_{\text{ref}}\|_{L^2}^2}_{\text{energy and reference penalties}} = \underbrace{s \|u - u_{\text{eff}}\|_{L^2}^2}_{\text{effective reference penalty}} + \underbrace{\frac{\lambda\mu}{s} \|u_{\text{ref}}\|_{L^2}^2}_{\text{independent of } u}. \tag{A.5}$$

Because the last term in Eq. (A.5) is independent of the optimized control, the first two penalties are equivalent, up to an additive constant, to $s\|u - u_{\text{eff}}\|_{L^2}^2$. Hence, $s = \lambda + \mu$ determines the total zeroth-order regularization strength, $\alpha = \mu/s$ determines its effective reference center, and γ independently controls spatial smoothness.

A.3 One-Step Approximation

We next impose a time-independent control approximation and interpret the source–target residual as a noisy semantic control cue. Combining the resulting residual surrogate with the running regularizer gives the one-step variational objective used by StableCtrl.

Time-independent control. One-step editing requires a single spatial control rather than a time-varying control trajectory. We therefore restrict the admissible control and reference fields to $u(\xi,t)=\bar{u}(\xi)$ and $u_{\mathrm{ref}}(\xi,t)=\bar{u}_{\mathrm{ref}}(\xi)$ for $t\in[0,1]$. Since the time horizon has unit length, any time-independent spatial functional satisfies $\int_0^1 E(u(t))\,dt = E(\bar{u})$. Applying this reduction to Eq. (A.4) gives

$$\mathcal{R}_{1\mathrm{step}}(\bar{u}) = \underbrace{\lambda\|\bar{u}\|_{L^2}^2}_{\text{control energy}} + \underbrace{\gamma\|\nabla\bar{u}\|_{L^2}^2}_{\text{spatial smoothness}} + \underbrace{\mu\|\bar{u}-\bar{u}_{\mathrm{ref}}\|_{L^2}^2}_{\text{reference consistency}}. \tag{A.6}$$

Equation (A.6) therefore removes temporal dependence while retaining all three regularization mechanisms of the continuous running cost.

Residual-based semantic surrogate. At the one-step evaluation time t_0, the observable semantic cue is the difference between target- and source-conditioned velocity fields,

$$R := \underbrace{v_\theta\big(x_{t_0},t_0,c_{\mathrm{tar}}\big)}_{\text{target-conditioned velocity}} - \underbrace{v_\theta\big(x_{t_0},t_0,c_{\mathrm{src}}\big)}_{\text{source-conditioned velocity}}. \tag{A.7}$$

Rather than using R directly as the editing direction, StableCtrl treats it as a noisy observation of the desired control field. For vectorized fields $\mathbf{R},\mathbf{u}\in\mathbb{R}^{NC}$, where N is the number of spatial locations and C the number of latent channels, we assume $\mathbf{R}=\mathbf{u}+\boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2 I_{NC})$. The corresponding negative log-likelihood, up to a constant independent of $\mathbf{u}$, is proportional to $\|\mathbf{u}-\mathbf{R}\|_2^2$. Absorbing its scale into the objective weights therefore gives the discrete fidelity $D_h(\mathbf{u};\mathbf{R}) := \|\mathbf{u}-\mathbf{R}\|_2^2$.

For a uniform two-dimensional grid with spacing h, the scaled discrete fidelity converges to its continuous counterpart:

$$\underbrace{h^2 D_h(\mathbf{u};\mathbf{R})}_{\text{discrete residual fidelity}} \longrightarrow \underbrace{\|\bar{u}-R\|_{L^2(\Omega)}^2}_{\text{continuous residual fidelity}}, \qquad h\to 0. \tag{A.8}$$

Defining $D(\bar{u};R) := \|\bar{u}-R\|_{L^2(\Omega)}^2$, Eq. (A.8) motivates the local surrogate replacement $\Phi(x(1)) \rightsquigarrow D(\bar{u};R)$, where $\rightsquigarrow$ denotes a surrogate replacement rather than exact equality or direct temporal discretization.

One-step variational objective. The residual surrogate provides semantic fidelity, whereas Eq. (A.6) provides the three stability regularizers. Combining them produces the central one-step objective:

$$J_{1\mathrm{step}}(\bar{u}) = \underbrace{\|\bar{u}-R\|_{L^2}^2}_{\text{residual fidelity}} + \underbrace{\lambda\|\bar{u}\|_{L^2}^2}_{\text{control energy}} + \underbrace{\gamma\|\nabla\bar{u}\|_{L^2}^2}_{\text{spatial smoothness}} + \underbrace{\mu\|\bar{u}-\bar{u}_{\mathrm{ref}}\|_{L^2}^2}_{\text{reference consistency}}. \tag{A.9}$$

The StableCtrl control is the minimizer $\bar{u}^* = \arg\min_{\bar{u}\in V} J_{1\mathrm{step}}(\bar{u})$. Thus, Eq. (A.9) preserves agreement with the semantic residual while explicitly constraining control magnitude, spatial variation, and reference consistency. In the subsequent analysis, we write $\bar{u}\equiv u$ and $\bar{u}_{\mathrm{ref}}\equiv u_{\mathrm{ref}}$ for simplicity.

B Variational Derivation

B.1 First-Order Optimality Condition

We derive the first-order optimality condition for the one-step surrogate objective introduced in Eq. (A.9). The derivation concerns the regularized control field under this local surrogate, rather than the global optimality of the full nonlinear generator trajectory. We first compute the first variation and then obtain the weak optimality condition, from which the elliptic control equation follows in the next subsection.

Admissible perturbation. Let $V := H^1(\Omega;\mathbb{R}^C)$ and assume $\lambda,\gamma,\mu\ge 0$. For the objective $J(u) = \|u-R\|_{L^2}^2 + \lambda\|u\|_{L^2}^2 + \gamma\|\nabla u\|_{L^2}^2 + \mu\|u-u_{\mathrm{ref}}\|_{L^2}^2$, consider an arbitrary perturbation

$v \in V$ and define $u_\varepsilon := u + \varepsilon v$ for $\varepsilon \in \mathbb{R}$. Since J is quadratic, differentiating its exact expansion at $\varepsilon = 0$ gives

$$\frac{1}{2}\delta J(u; v) = \underbrace{\langle u - R, v\rangle}_{\text{residual fidelity}} + \underbrace{\lambda\langle u, v\rangle}_{\text{control energy}} + \underbrace{\gamma\langle \nabla u, \nabla v\rangle}_{\text{spatial regularization}} + \underbrace{\mu\langle u - u_{\text{ref}}, v\rangle}_{\text{reference consistency}}. \tag{B.1}$$

Here and below, all inner products are taken over Ω, with Euclidean contraction over latent channels. The factor $1/2$ removes the common factor generated by differentiating the squared norms and does not alter the stationary point. Introducing $a_0 := 1 + \lambda + \mu$ and $f := R + \mu u_{\text{ref}}$, Eq. (B.1) becomes $\frac{1}{2}\delta J(u; v) = a_0\langle u, v\rangle + \gamma\langle \nabla u, \nabla v\rangle - \langle f, v\rangle$. This representation separates the zeroth-order control term, spatial regularization, and residual-reference forcing.

Weak first-order optimality condition. Because the surrogate objective is Fréchet differentiable, its minimizer $u^* \in V$ must have zero first variation in every admissible direction. Substituting the collected form above gives the weak optimality condition

$$\underbrace{a_0\langle u^*, v\rangle}_{\text{zeroth-order control}} + \underbrace{\gamma\langle \nabla u^*, \nabla v\rangle}_{\text{spatial regularization}} = \underbrace{\langle f, v\rangle}_{\text{residual-reference forcing}}, \qquad \forall v \in V. \tag{B.2}$$

Since $a_0 = 1 + \lambda + \mu \geq 1$ and $\gamma \geq 0$, the quadratic objective is strictly convex in u; hence the stationary point in Eq. (B.2) is the unique global minimizer of the one-step surrogate objective. In particular, this uniqueness does not require strictly positive regularization weights because the residual fidelity term already contributes the positive L^2 quadratic component.

Equivalently, Eq. (B.2) states that $\int_\Omega (a_0 u^* - f) \cdot v \, d\xi + \gamma \int_\Omega \nabla u^* : \nabla v \, d\xi = 0$ for every $v \in V$. Under sufficient regularity, applying Green's identity to this weak form yields the corresponding Euler–Lagrange equation. The associated natural boundary condition is derived explicitly in the next subsection rather than imposed at this stage.

C Elliptic Analysis

C.1 Well-Posedness of the Elliptic Operator

We establish well-posedness of the elliptic control equation through its weak formulation. Continuity and coercivity of the associated bilinear form yield existence, uniqueness, and continuous dependence of the optimal control on the residual and reference forcing. These results characterize the one-step surrogate control problem rather than the full nonlinear generator trajectory.

Weak formulation. Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain, $V := H^1(\Omega; \mathbb{R}^C)$, and $V^* := H^{-1}(\Omega; \mathbb{R}^C)$. Define $a_0 := 1 + \lambda + \mu \geq 1$ and $f := R + \mu u_{\text{ref}}$, with $\lambda, \mu \geq 0$ and $\gamma > 0$. Applying Green's identity to the weak condition derived in Eq. (B.2) gives the elliptic system

$$\underbrace{(a_0 I_C - \gamma\Delta)u^*}_{\mathcal{L}u^*} = \underbrace{R + \mu u_{\text{ref}}}_{f} \quad \text{in } \Omega, \qquad \partial_n u^* = 0 \quad \text{on } \partial\Omega. \tag{C.1}$$

Here, Δ acts channel-wise. The homogeneous Neumann condition in Eq. (C.1) is the natural boundary condition associated with the H^1 spatial regularizer; it is not an additional data-dependent constraint.

The corresponding weak problem is to find $u^* \in V$ such that $a(u^*, v) = \ell_f(v)$ for every $v \in V$, where $a(u, v) := a_0\langle u, v\rangle_{L^2} + \gamma\langle \nabla u, \nabla v\rangle_{L^2}$ and $\ell_f(v) := \langle f, v\rangle_{V^*, V}$. All inner products include Euclidean contraction over the C latent channels.

Continuity and coercivity. Equip V with $\|u\|_V^2 := \|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2$. Cauchy–Schwarz and the positivity of a_0 and γ give the two estimates

$$\underbrace{|a(u, v)|}_{\text{continuity}} \leq C_{\mathcal{L}}\|u\|_V\|v\|_V, \qquad \underbrace{a(u, u)}_{\text{coercivity}} \geq c_{\mathcal{L}}\|u\|_V^2, \tag{C.2}$$

where $C_{\mathcal{L}} := \max\{a_0, \gamma\}$ and $c_{\mathcal{L}} := \min\{a_0, \gamma\} > 0$. Moreover, $|\ell_f(v)| \leq \|f\|_{V^*}\|v\|_V$, so ℓ_f is bounded on V. Equation (C.2) therefore verifies the continuity and coercivity assumptions required by the Lax–Milgram theorem.

Existence, uniqueness, and stability. Lax–Milgram consequently gives a unique weak solution $u^* \in V$. Testing the weak equation with $v = u^*$ and using Eq. (C.2) yields the stability estimate

$$\boxed{\|u^*\|_V \le c_{\mathcal{L}}^{-1} \|R + \mu u_{\mathrm{ref}}\|_{V^*}}. \tag{C.3}$$

Equation (C.3) bounds the magnitude and spatial variation of the one-step control by the combined residual and reference forcing. It is a stability statement for the control solution and does not by itself imply stability of the final generated image under arbitrary perturbations.

The same argument applied to two forcing fields $f_1, f_2 \in V^*$ and their solutions u_1, u_2 gives

$$\|u_1 - u_2\|_V \le c_{\mathcal{L}}^{-1} \|f_1 - f_2\|_{V^*}. \tag{C.4}$$

Thus, the solution map $f \mapsto u^*$ is Lipschitz continuous from V^* to V: bounded changes in the combined forcing induce bounded changes in the StableCtrl control field. When $\gamma = 0$, the spatial operator reduces separately to the pointwise relation $a_0 u = f$; the H^1-coercivity argument above is specifically for the elliptic case $\gamma > 0$.

C.2 Characterization of the Solution

We next connect the weak elliptic solution to the variational minimizer. This equivalence shows that the elliptic equation is not an independent post-processing rule, but the first-order optimality system of the one-step StableCtrl objective.

Operator representation. Define $A : V \to V^*$ through $\langle Au, v\rangle_{V^*,V} := a(u, v)$. The weak problem in Eq. (C.1) is therefore $Au^* = f$ in V^*. By the well-posedness result above, A is boundedly invertible, and the solution admits the operator representation

$$\boxed{u^* = A^{-1}\big(R + \mu u_{\mathrm{ref}}\big) = \mathcal{L}^{-1}\big(R + \mu u_{\mathrm{ref}}\big)}. \tag{C.5}$$

Here, $\mathcal{L}^{-1}$ denotes the elliptic solution operator associated with the boundary-value problem in Eq. (C.1), rather than an explicitly formed dense matrix inverse. This distinction separates the analytical operator notation from its numerical realization.

Variational characterization. The one-step objective can be written as $J(u) = a(u, u) - 2\ell_f(u) + C_0$, where $C_0 := \|R\|_{L^2}^2 + \mu\|u_{\mathrm{ref}}\|_{L^2}^2$ is independent of u. For any $w \in V$, set $e := w - u^*$. Using the weak identity $a(u^*, e) = \ell_f(e)$, the cross term cancels exactly, giving

$$\underbrace{J(w) - J(u^*)}_{\text{objective gap}} = \underbrace{a(w - u^*, w - u^*)}_{\text{elliptic energy}} \ge \underbrace{c_{\mathcal{L}}\|w - u^*\|_V^2}_{\text{coercive lower bound}} \ge 0. \tag{C.6}$$

Equality in Eq. (C.6) holds only when $w = u^*$. Hence, the weak elliptic solution in Eq. (C.5) is exactly the unique global minimizer of the one-step surrogate objective. The qualifier "one-step surrogate" is essential here: the result does not assert global optimality of the underlying nonlinear generator trajectory.

C.3 Fourier-Domain Solution

We next derive a mode-wise spectral representation of the same elliptic operator. For analytical convenience, this subsection uses a periodic domain so that the Laplacian is diagonal in the Fourier basis. This changes the spectral basis used for the analysis, not the structure of the elliptic multiplier: under homogeneous Neumann conditions on a rectangular domain, the corresponding representation uses cosine eigenmodes of the Neumann Laplacian.

Spectral representation. For the periodic analysis, let $\Omega = \mathbb{T}^2 := (\mathbb{R}/2\pi\mathbb{Z})^2$ and $f := R + \mu u_{\mathrm{ref}}$. For $k = (k_1, k_2) \in \mathbb{Z}^2$, define $\varphi_k(x) := e^{\,\mathrm{i}k\cdot x}$. Expanding u^* and f in this basis gives coefficients $\widehat{u}_k^*, \widehat{f}_k \in \mathbb{C}^C$, with $\widehat{f}_k = \widehat{R}_k + \mu\widehat{u}_{\mathrm{ref},k}$. Since $\Delta\varphi_k = -\|k\|_2^2\varphi_k$, Eq. (C.1) decouples mode by mode as

$$\underbrace{\big(1 + \lambda + \mu + \gamma\|k\|_2^2\big)}_{\text{mode-dependent elliptic denominator}}\ \widehat{u}_k^* = \underbrace{\widehat{R}_k + \mu\widehat{u}_{\mathrm{ref},k}}_{\text{residual-reference forcing}}, \qquad k \in \mathbb{Z}^2. \tag{C.7}$$

Therefore, each spatial mode has the explicit representation

$$\boxed{\widehat{u}_k^* = \frac{\widehat{R}_k + \mu \widehat{u}_{\mathrm{ref},k}}{1 + \lambda + \mu + \gamma \|k\|_2^2}}. \tag{C.8}$$

The denominator in Eq. (C.8) is strictly positive for $\lambda, \mu, \gamma \geq 0$, so each mode is uniquely determined and the same scalar response is applied independently across latent channels.

Spectral multiplier. Define $H(k) := (1+\lambda+\mu+\gamma\|k\|_2^2)^{-1}$. Then Eq. (C.8) becomes $\widehat{u}_k^* = H(k)\widehat{f}_k$, showing directly that the periodic realization of $\mathcal{L}^{-1}$ is diagonal in the Fourier basis. The reconstructed control is obtained by applying this scalar multiplier to each spectral coefficient and transforming back to the spatial domain. No dense spatial matrix needs to be explicitly formed or inverted.

C.4 Spectral Properties

We now analyze what the multiplier in Eq. (C.8) implies for different spatial modes and regularization parameters. The conclusions below concern the spectral response of the control operator; empirical associations between these modes and editing artifacts are evaluated separately in the experiments.

Radial frequency response. Let $a_0 := 1 + \lambda + \mu$ and define $h(r) := (a_0 + \gamma r^2)^{-1}$ for $r \geq 0$. For $\gamma > 0$, its derivative is negative away from the zero mode, and the response satisfies

$$\underbrace{0 < H(k) \leq a_0^{-1}}_{\text{bounded modal gain}}, \qquad \underbrace{H(k) \sim \left(\gamma \|k\|_2^2\right)^{-1}}_{\text{high-frequency decay}}, \qquad \|k\|_2 \to \infty. \tag{C.9}$$

Thus, the elliptic solution operator attenuates increasingly high-frequency control components, with asymptotic decay $\mathcal{O}(\|k\|_2^{-2})$. This is the precise sense in which StableCtrl has a low-pass spectral response; it does not assume that every high-frequency component is necessarily an editing artifact.

Parameter-wise spectral sensitivity. To separate the roles of the three regularization weights, define $B_k := 1 + \lambda + \gamma\|k\|_2^2$ and $D_k := B_k + \mu$. Equation (C.8) can then be written as $\widehat{u}_k^* = D_k^{-1}(\widehat{R}_k + \mu \widehat{u}_{\mathrm{ref},k})$. Differentiating with respect to λ and γ gives

$$\underbrace{\partial_\lambda \widehat{u}_k^* = -D_k^{-1}\widehat{u}_k^*}_{\text{frequency-uniform attenuation}}, \qquad \underbrace{\partial_\gamma \widehat{u}_k^* = -\|k\|_2^2 D_k^{-1}\widehat{u}_k^*}_{\text{frequency-selective attenuation}}. \tag{C.10}$$

Equation (C.10) shows that λ modifies all modal denominators uniformly, whereas the effect of γ increases with the Laplacian eigenvalue and therefore acts more strongly on high-frequency spatial variation.

The role of μ is different because it enters both the denominator and the reference contribution in the numerator. Writing

$$\widehat{u}_k^* = \underbrace{D_k^{-1}\widehat{R}_k}_{\text{residual contribution}} + \underbrace{\mu D_k^{-1}\widehat{u}_{\mathrm{ref},k}}_{\text{reference contribution}}, \tag{C.11}$$

makes this distinction explicit. As μ increases, D_k^{-1} decreases while μD_k^{-1} increases, and the limiting behavior is

$$\widehat{u}_k^*|_{\mu=0} = B_k^{-1}\widehat{R}_k, \qquad \lim_{\mu\to\infty} \widehat{u}_k^* = \widehat{u}_{\mathrm{ref},k}. \tag{C.12}$$

Hence, λ controls frequency-uniform shrinkage, γ controls frequency-selective spatial smoothing, and μ shifts the solution from residual-driven control toward the reference field. Together, these effects explain the distinct roles of the three terms in Eq. (A.9) without interpreting any single parameter as an independent measure of editing quality.

From Theory to Empirical Validation. The analysis above establishes the mathematical behavior of StableCtrl under the one-step surrogate: the regularized objective admits a unique control solution, the associated elliptic operator depends continuously on the residual and reference forcing, and its spectral representation provides frequency-selective attenuation controlled by λ, γ, and μ. These properties characterize the control operator itself, but do not alone determine how the resulting regularization affects perceptual editing quality, semantic alignment, or robustness across practical editing scenarios. We therefore complement the theoretical analysis with additional experiments that examine these effects empirically, including component ablations, parameter sensitivity, reference design, frequency behavior, and computational trade-offs.

D Additional Ablation Results

The following experiments complement the main-paper evaluation by examining questions that are not fully resolved by aggregate metrics and component-wise ablations alone. In particular, the one-step objective in Eq. (A.9) jointly regulates control magnitude, spatial variation, and reference consistency, while the spectral solution in Eq. (C.8) predicts frequency-selective attenuation through the elliptic operator. These analytical properties characterize the control field but do not by themselves establish whether the observed preservation gains arise from the proposed joint regularization rather than weaker editing, generic smoothing, or simple combinations of individual regularizers. They also do not establish whether the same control formulation exhibits consistent behavior across different one-step generation backbones or whether the gains persist across different editing categories. This evaluation connects the theoretical behavior of the control operator with its practical effects on editing quality and robustness. We therefore supplement the main results with matched-semantic regularization baselines, cross-backbone evaluation, and category-wise analysis, examining these empirical questions from complementary perspectives.

D.1 Comparison with Simple Regularization Baselines

The spectral analysis in Eq. (C.9) shows that StableCtrl attenuates increasingly high-frequency control components, while Eq. (A.9) additionally regulates control magnitude and reference consistency. To determine whether the preservation gains observed in the main paper can instead be explained by reduced editing strength, generic spatial smoothing, or simple combinations of these effects, we compare StableCtrl with several matched-semantic alternatives.

Scalar Shrinkage applies a uniform scaling $u = \alpha R$, whereas Gaussian Smoothing applies a spatial Gaussian filter $u = G_\sigma * R$. Scalar + Gaussian combines the two operations by applying both magnitude shrinkage and spatial smoothing to the residual. We further include Reference Blending, which directly combines the semantic residual with the temporal reference without solving the joint regularized control problem. These alternatives progressively isolate the effects of control magnitude, spatial smoothing, and temporal reference information. The regularization strengths are adjusted to approximately match the CLIP-E score of StableCtrl, controlling for semantic editing strength.

As shown in Table S1(a), both Scalar Shrinkage and Gaussian Smoothing improve structural preservation over the direct residual, confirming that reducing residual magnitude or high-frequency variation is beneficial. Combining shrinkage with Gaussian smoothing further improves the preservation metrics, while direct reference blending provides an additional gain by incorporating temporal information. Nevertheless, under comparable semantic editing strength, StableCtrl achieves the best PSNR, LPIPS, and structural distance among these alternative regularization strategies. This comparison indicates that the preservation gains cannot be explained solely by weaker editing, generic smoothing, their simple combination, or direct reference blending. Instead, jointly regularizing control magnitude, spatial smoothness, and reference consistency provides additional structural benefits at a comparable semantic editing strength across these controlled comparisons.

D.2 Cross-Backbone Evaluation

The preceding experiments use SD-Turbo as the default one-step generator. To examine whether the proposed control formulation depends strongly on this particular backbone, we additionally integrate StableCtrl with InstaFlow, SwiftBrush-v2, and SDXL-Lightning without backbone-specific training or fine-tuning. The same control formulation is retained across these models, while the corresponding direct residual solution serves as the backbone-wise reference.

As shown in Table S1(b), StableCtrl exhibits a consistent structural-preservation trend across the evaluated backbones, improving PSNR, LPIPS, and structural distance relative to direct residual editing. The magnitude of the improvement varies across generators, while semantic alignment remains comparable with a moderate structure–semantic trade-off. These results provide empirical evidence that the proposed regularized control operator can transfer across different one-step generation backbones without backbone-specific training.

As shown in Table S1(a), both Scalar Shrinkage and Gaussian Smoothing improve structural preservation over the direct residual, confirming that reducing residual magnitude or high-frequency variation

Table S1: **Additional controlled evaluations of StableCtrl. (a)** Matched-semantic comparisons examine whether the preservation gain can be explained by residual shrinkage, generic smoothing, their combination, or simple reference blending. **(b)** Cross-backbone evaluation examines whether the same control formulation transfers across different one-step generation backbones without backbone-specific training.

Method	PSNR↑	LPIPS$\times 10^3$ ↓	SD$\times 10^3$ ↓	CLIP-E↑
Direct Residual	23.87	89.21	13.56	**21.88**
Scalar Shrinkage	25.60	73.52	9.42	21.29
Gaussian Smoothing	26.44	65.85	8.81	21.27
Scalar + Gaussian	26.71	62.84	8.36	21.27
Reference Blending	26.93	60.91	8.02	21.28
StableCtrl	**27.46**	**55.55**	**7.45**	21.26

Backbone / Method	PSNR↑	LPIPS↓	SD↓	CLIP-E↑
SD-Turbo / Direct	23.87	89.21	13.56	**21.88**
SD-Turbo / **Ours**	**27.46**	**55.55**	**7.45**	21.26
InstaFlow / Direct	22.94	96.38	14.21	**21.73**
InstaFlow / **Ours**	**26.31**	**64.72**	**8.36**	21.19
SwiftBrush-v2 / Direct	23.42	91.67	13.08	**21.81**
SwiftBrush-v2 / **Ours**	**26.78**	**61.34**	**7.92**	21.24
SDXL-Lightning / Direct	24.16	85.73	12.47	**22.06**
SDXL-Lightning / **Ours**	**27.12**	**58.91**	**7.61**	21.52

Table S2: **Category-wise preservation performance on PIE-Bench.** Results are evaluated on unedited regions across different editing groups. StableCtrl uses the final configuration with $\delta = 0.15$. Bold values indicate the better result for each metric.

		ChordEdit (w/o prox.)		StableCtrl	
Edit Group	**Editing Type**	**PSNR↑**	**LPIPS$\times 10^3$ ↓**	**PSNR↑**	**LPIPS$\times 10^3$ ↓**
Random	Random editing	24.90	79.0	**28.59**	**52.0**
Object	Change object	21.73	120.0	**26.39**	**67.0**
	Add object	24.97	70.0	**28.22**	**46.0**
	Delete object	21.81	120.0	**25.14**	**72.0**
Attribute	Attribute content	24.67	86.0	**28.18**	**54.0**
	Attribute pose	25.59	74.0	**28.09**	**56.0**
	Attribute color	23.65	86.0	**27.06**	**56.0**
	Attribute material	25.40	77.0	**29.78**	**48.0**
Scene	Change background	23.48	69.0	**27.01**	**41.0**
	Change style	23.21	132.0	**27.81**	**84.0**
Overall		23.80	89.2	**27.46**	**55.5**

is beneficial. Combining shrinkage with Gaussian smoothing further improves preservation, while Reference Blending provides an additional gain by incorporating temporal reference information. All matched-semantic alternatives use their standard residual-processing operations, with the regularization strength adjusted only to approximately match the semantic editing strength of StableCtrl rather than to optimize preservation metrics. Under this matched-semantic protocol, StableCtrl further improves PSNR, LPIPS, and structural distance over all simple alternatives. This comparison indicates that the preservation gains cannot be explained solely by weaker editing, generic spatial smoothing, their simple combination, or direct reference blending. Instead, jointly regularizing control magnitude, spatial smoothness, and temporal reference consistency provides additional structural benefits at a comparable semantic editing strength.

D.3 Category-wise Analysis

The main-paper evaluation reports aggregate preservation performance over PIE-Bench, whereas the following analysis examines whether these gains are distributed across editing categories rather than dominated by a small subset of examples. Table S2 reports PSNR and LPIPS on unedited regions across the ten PIE-Bench editing categories. StableCtrl improves both metrics across all categories, increasing the overall PSNR from 23.80 to 27.46 dB and reducing LPIPS from 0.0892 to 0.0555. The improvements span object, attribute, background, and style edits, indicating that the aggregate preservation gain is not driven by a small subset of editing types.



Figure S1: **Qualitative analysis of key StableCtrl hyperparameters.** Each panel varies one parameter while fixing the remaining parameters to the final configuration. (a) The smoothness weight γ controls spatial-frequency attenuation. (b) The reference weight μ controls the contribution of the temporally adjacent reference field. (c) The step scale η controls the magnitude of the one-step latent update. (d) The Monte Carlo sample count n controls the number of samples used to estimate the control. The adopted settings are marked as default.

D.4 Qualitative Analysis of Key Hyperparameters

The main paper quantitatively analyzes the sensitivity of StableCtrl to its key hyperparameters. Figure S1 complements this analysis by visualizing their effects on representative edits. Unlike the controlled-variable study in the main paper, all non-varied parameters are fixed to the final configuration: $\lambda = 0.08$, $\gamma = 0.15$, $\mu = 0.25$, $\eta = 0.9$, $n = 1$, and $\delta = 0.15$. The examples cover diverse object, attribute, background, style, and illumination edits.

Effect of the smoothness weight γ. Figure S1(a) varies $\gamma \in \{0, 0.05, 0.15, 0.20\}$. According to Eqs. (C.8) and (C.9), increasing γ strengthens the attenuation of high-frequency control components, producing spatially smoother control effects. This trend is visible in the lamp-insertion and seagull-removal examples, where larger γ improves local consistency and reduces residual variation. The adopted $\gamma = 0.15$ is consistent with the quantitative frequency and energy analysis in the main paper.

Effect of the reference weight μ. Figure S1(b) varies $\mu \in \{0, 0.20, 0.25, 0.40\}$. According to the residual–reference decomposition in Eq. (C.11), increasing μ shifts the control toward the temporally adjacent reference field, while $\mu = 0$ removes reference consistency. The book-to-tablet and pose-change examples illustrate the balance between structural preservation and semantic modification. We use $\mu = 0.25$, consistent with the reference-prior analysis in the main paper.

Effect of the step scale η. Figure S1(c) varies $\eta \in \{0.5, 0.7, 0.9, 1.3\}$. Unlike the regularization parameters, η directly scales the applied one-step control, with smaller values producing more conservative updates and larger values strengthening the semantic modification. This progression is visible in the Van Gogh style-transfer and sunset-background examples, where excessive scaling can also affect global appearance and local details. We use $\eta = 0.9$ in the final configuration.

Effect of the Monte Carlo sample count n. Figure S1(d) evaluates $n \in \{1, 2, 4, 8\}$. The sample count affects the Monte Carlo estimate used to construct the control cue rather than the elliptic regu-

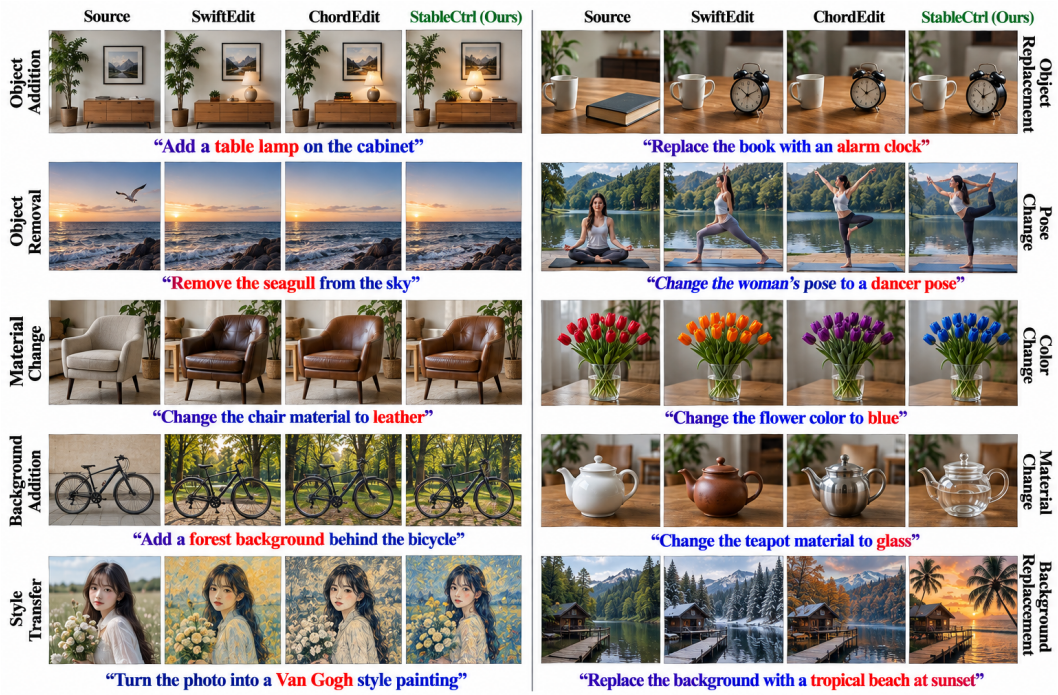

Figure S2: **Extended qualitative comparison across editing types.** All source images are real photographs. We compare StableCtrl with SwiftEdit (Nguyen et al., 2025) and ChordEdit (Lu et al., 2026) across object, attribute, material, background, and style edits. The examples complement the main-paper comparison by covering a broader range of editing instructions and visual contexts.

larization operator. Increasing n reduces sampling variation at additional inference cost, consistent with the quantitative results showing diminishing structural gains beyond $n = 4$. We retain $n = 1$ for efficient inference and use $n = 4$ as an optional quality–efficiency setting.

Overall Observation. The four parameters affect distinct components of StableCtrl: γ controls spatial regularization, μ balances residual and reference contributions, η scales the applied control, and n determines the Monte Carlo estimation budget. Figure S1 visually complements the quantitative analyses and illustrates how these factors jointly affect structural preservation, semantic modification, and inference efficiency across diverse editing conditions and parameter settings.

E Additional Qualitative Results

Comparison across Editing Types. We provide qualitative results to complement the PIE-Bench evaluation across diverse scenarios. Figure S2 extends the comparison with SwiftEdit (Nguyen et al., 2025) and ChordEdit (Lu et al., 2026) to object, attribute, material, background, and style edits using real photographs under the same inference configuration. While all methods produce semantically relevant modifications, StableCtrl better preserves local object structure and surrounding scene content across these editing types. The differences are particularly visible in fine-grained attributes and regions outside the intended edit, where unnecessary structural changes are reduced.

Results on Additional Real Images. We further evaluate StableCtrl on real photographs from publicly available image collections, including Unsplash Lite (Unsplash, 2022), without image-specific parameter adjustment. As shown in Fig. S3, StableCtrl supports both localized edits and broader appearance changes while largely preserving spatial layout and content unrelated to the edit. This behavior remains visually consistent across diverse scenarios and different edit scales, from local attribute changes and object insertion to global appearance and environmental modifications. Together, Figs. S2 and S3 provide complementary qualitative evidence across editing types and real-world visual conditions, while the quantitative evaluation remains grounded in PIE-Bench.

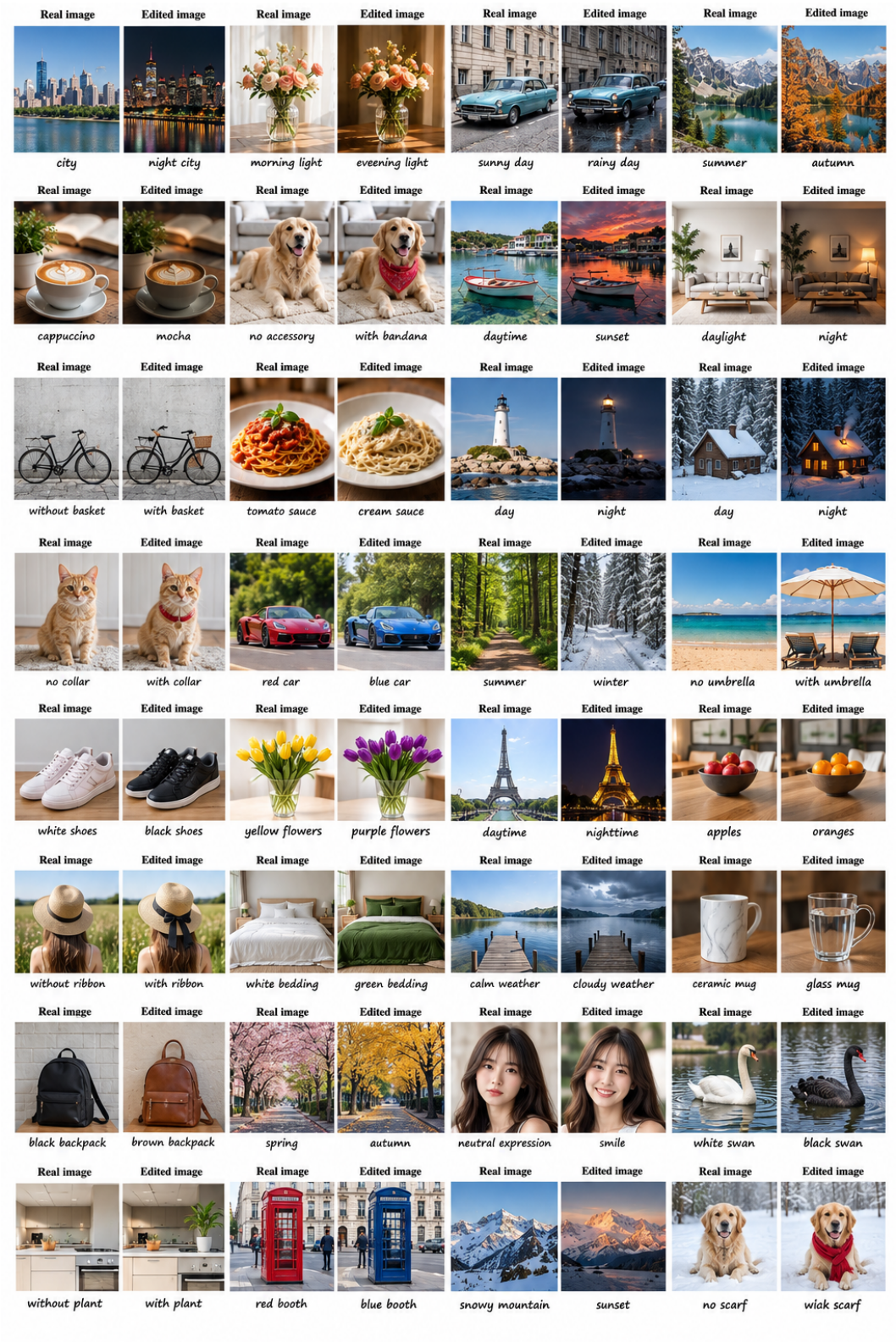

Figure S3: **Additional qualitative results.** Each pair shows a real source photograph and the corresponding StableCtrl result. The examples cover illumination, weather, season, appearance, attribute modification, and object insertion under diverse real-world visual conditions. StableCtrl realizes the requested modifications while largely preserving scene structure and object geometry.