CTRLGEN: OPTIMAL CONTROL FOR TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-based text-to-image models achieve strong generation quality and efficiency, yet conditional generation still relies heavily on classifier-free guidance. Existing guidance methods mainly adjust when or how strongly the conditional correction is applied, while its direction and magnitude are usually regulated together. This coupling limits local strength control, and unconstrained spatial and temporal magnitude variations can produce excessive responses during sampling. We identify guidance magnitude as the key variable for controlling strength while preserving the model-predicted direction and suppressing excessive local magnitude variations. Based on this observation, we propose CtrlGen, a training-free framework that decomposes the conditional correction into local direction and magnitude and treats the latter as a state-dependent control variable. At each sampling state, CtrlGen solves a regularized convex control problem that balances guidance fidelity, control energy, spatial smoothness, and state consistency. The resulting optimality condition is a screened elliptic equation with an explicit spectral solution, enabling frequency-selective magnitude control without iterative optimization. The optimized magnitude is recombined with the preserved direction to form a feedback guidance rule without modifying the pretrained model. Extensive experiments show consistent improvements over existing guidance methods in generation quality, text alignment, and preference-oriented metrics.

1 INTRODUCTION

Text-to-image generation has evolved from iterative diffusion models toward more efficient generative dynamics. Early text-conditioned diffusion models achieved high-quality synthesis and semantic alignment through guided sampling (Nichol et al., 2021; Balaji et al., 2022). Later work reduced sampling cost through discrete representations (Chang et al., 2023), consistency models (Song & Dhariwal, 2024), flexible single- and multi-step generation (Frans et al., 2025), and average-velocity modeling (Geng et al., 2026). In parallel, transport methods established learned velocity fields as a general framework for generation, including rectified flow (Liu et al., 2023), interpolant-based models (Ma et al., 2024), and large-scale flow transformers (Esser et al., 2024). These advances place increasing importance on the generative dynamics themselves and motivate more precise inference-time control of the learned velocity field across sampling trajectories and architectures.

Guidance has evolved along with these models. Classifier-Free Guidance controls conditional generation by scaling the difference between conditional and unconditional predictions. Recent methods move beyond a fixed global rule through higher-order interactions (Sun et al., 2024), restricted or adaptive guidance intervals (Kynkäänniemi et al., 2024; Castillo et al., 2025), spatial strength control (Shen et al., 2024), directional decomposition (Sadat et al., 2024), and manifold-constrained correction (Chung et al., 2025). For continuous generative dynamics, flow guidance analyzes conditional effects on flow trajectories (Feng et al., 2025), while inference-time steering (Singhal et al., 2025) and dynamic CFG (Papalampidi et al., 2026) further adapt the guidance process across different stages of the sampling trajectory without retraining the generator during inference time.

Despite these advances, several limitations remain: (1) Coupled direction and magnitude: Most methods regulate the conditional correction as a whole, although its direction determines where the flow moves and its magnitude determines how strongly it moves. This coupling makes guidance strength difficult to control without changing the model-predicted correction orientation. (2) Unregulated

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.

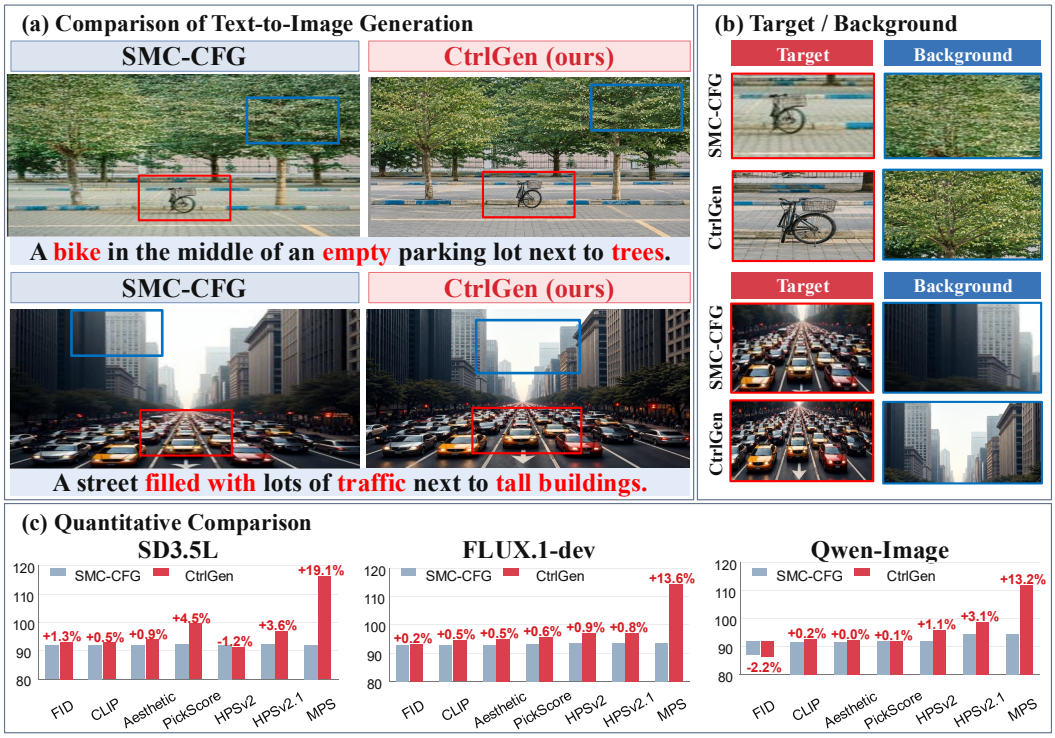

Figure 1: **Overview of CtrlGen.** (a) CtrlGen improves prompt-relevant details over SMC-CFG while preserving coherent scene structure. (b) Target and background crops reveal improved local structure and contextual consistency. (c) Consistent quantitative gains across SD3.5L, FLUX.1-dev, and Qwen-Image further demonstrate the effectiveness of magnitude-controlled guidance.

local variation: Global or coarse guidance adjustment provides limited control over spatially varying responses, so excessive local magnitudes can introduce unstable corrections.

To address these limitations, we introduce CtrlGen, a training-free control framework for flow-based text-to-image generation. The core module, Magnitude Flow Control (MagFlow), decouples guidance direction and magnitude and treats the latter as the control variable. At each spatial location, MagFlow preserves the model-predicted direction and optimizes its magnitude through a regularized control objective that balances guidance fidelity, control energy, spatial smoothness, and state consistency. The resulting optimality condition with an explicit spectral solution, which enables frequency-selective magnitude regulation without iterative optimization. At each sampling step, the optimized magnitude is recombined with the preserved direction to form a state-dependent guidance rule. Extensive experiments demonstrate consistent improvements in generation quality, text alignment, and preference-oriented metrics (see Figure 1). Our contributions are as follows:

- We introduce a direction–magnitude decomposition of the guidance field and isolate the magnitude as an explicit control variable while preserving the model-predicted correction direction.
- We formulate each state-wise magnitude update as a regularized convex control problem and derive its unique elliptic optimum together with an explicit spectral solution, providing principled energy, spatial, and temporal regularization without iterative optimization.
- We integrate the resulting magnitude controller into flow-based sampling as a training-free state-dependent feedback rule and demonstrate consistent gains across multiple flow-based generators.

2 Related Work

Diffusion and Flow-Based Generative Models. Diffusion models generate samples by reversing a noise process (Ho et al., 2020; Song et al., 2020a;b) and have become standard tools for image

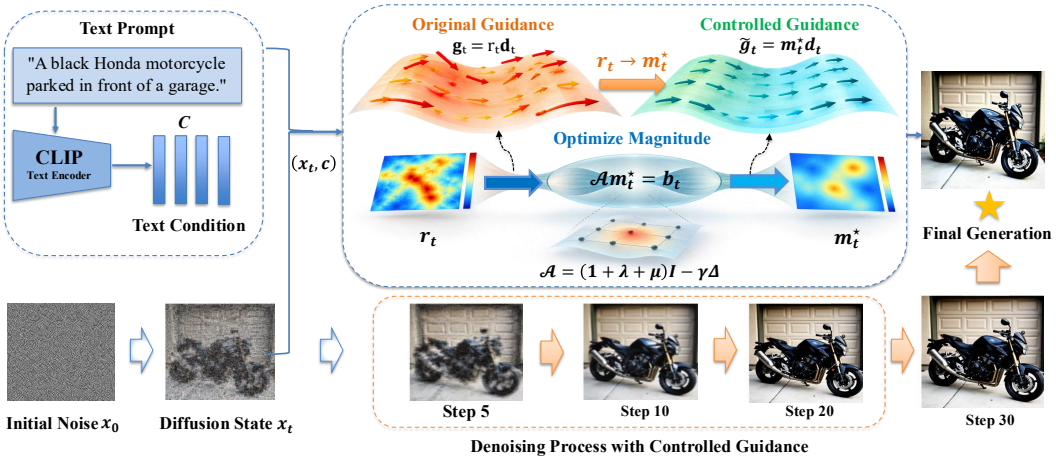

Figure 2: **CtrlGen architecture and inference pipeline.** Starting from the current latent state $\mathbf{x}_t$ and text condition c, the guidance field $\mathbf{g}_t$ is decomposed into its local magnitude r_t and direction $\mathbf{d}_t$. The model-predicted direction is preserved, while the magnitude r_t is regulated to obtain the optimized field $m_t^\star$. The optimized magnitude is then recombined with the preserved direction to construct the controlled guidance $\tilde{\mathbf{g}}_t = m_t^\star \mathbf{d}_t$. This state-dependent control is applied throughout sampling and transports the initial noise toward the final generated image.

synthesis (Rombach et al., 2022; Peebles & Xie, 2023). Flow matching and rectified flow instead learn continuous velocity fields for distribution transport (Lipman et al., 2023; Liu et al., 2022). These models extend continuous-time generative modeling (Grathwohl et al., 2018) and support recent systems such as Stable Diffusion 3.5 (Esser et al., 2024), FLUX (Black Forest Labs, 2024), and Qwen-Image (Wu et al., 2025). Their pretrained velocity fields provide the basis for flexible inference-time guidance and control across different sampling trajectories and model architectures.

Guidance for Generative Models. Classifier Guidance uses an external classifier (Dhariwal & Nichol, 2021), while Classifier-Free Guidance (CFG) combines conditional and unconditional predictions (Ho & Salimans, 2022). CFG is widely used for conditional generation (Saharia et al., 2022; Ruiz et al., 2023). Later methods improve guidance through manifold correction (Chung et al., 2025), restricted intervals (Kynkäänniemi et al., 2024), sampling schedules (Lin et al., 2024), adaptive scaling (Wang et al., 2024), nonlinear correction (Zheng & Lan, 2023), direction refinement (Sadat et al., 2024), and rectified diffusion guidance (Xia et al., 2025). Recent work extends CFG to flow models through velocity scaling (Fan et al., 2025) and predictor-corrector guidance (Saini et al., 2026). CtrlGen instead separates local guidance into direction and magnitude. The controller regulates only the magnitude and preserves the predicted direction across spatial locations and sampling states.

Control-Theoretic Generative Modeling. Optimal control provides a framework for dynamical systems (Bryson, 1975) and has been connected to generative modeling through Schrödinger bridges (De Bortoli et al., 2021) and stochastic control (Berner et al., 2022; Shi et al., 2022). Recent work studies reward-based diffusion and flow fine-tuning (Domingo i Enrich et al., 2025). FOCUS (Bill et al., 2026) uses stochastic optimal control for multi-entity fidelity, while FlowAlign (Kim et al., 2026) regularizes trajectories for inversion-free flow-based image editing. In contrast, CtrlGen preserves the predicted direction and regulates only the spatial magnitude. State-wise control maintains consistency across sampling states without model training across different model architectures.

3 Method

3.1 Preliminaries

Flow matching models learn a time-dependent velocity field that transports samples from a simple prior distribution to the data distribution (Lipman et al., 2023). Let $\boldsymbol{x}_t \in \mathbb{R}^{C\times H\times W}$ denote the latent state at time t, and let $v_\theta(\boldsymbol{x}_t, t, c)$ and $v_\theta(\boldsymbol{x}_t, t, \emptyset)$ denote the conditional and unconditional velocity predictions, respectively. Classifier-free guidance (CFG) (Ho & Salimans, 2022) constructs

the guided velocity as

$$\widehat{v}_\theta(\boldsymbol{x}_t, t, c) = v_\theta(\boldsymbol{x}_t, t, \emptyset) + w\mathbf{g}_t, \qquad \mathbf{g}_t = v_\theta(\boldsymbol{x}_t, t, c) - v_\theta(\boldsymbol{x}_t, t, \emptyset), \tag{1}$$

where w is the guidance scale and $\mathbf{g}_t$ is the conditional-unconditional velocity discrepancy. From a control perspective, the sampling dynamics can be written as $\dot{\boldsymbol{x}}_t = v_\theta(\boldsymbol{x}_t, t, \emptyset) + \mathbf{u}_t$, with $\mathbf{u}_t = w\mathbf{g}_t$, where the unconditional field defines the underlying generative transport and $\mathbf{u}_t$ provides the conditional control input. Standard CFG, however, does not distinguish the internal structure of $\mathbf{g}_t$. At each spatial location with nonzero guidance, $\mathbf{g}_t(p) \in \mathbb{R}^C$ admits a direction and a magnitude $\|\mathbf{g}_t(p)\|_2$, which determine the orientation and strength of the local conditional correction, respectively. Since $\mathbf{u}_t(p) = w\mathbf{g}_t(p)$, a global w rescales all local magnitudes uniformly and cannot regulate their spatial variation. As illustrated in Figure 2, we therefore preserve the model-predicted direction and expose the spatial magnitude as the control variable, yielding a direction-preserving state-wise magnitude parameterization developed in the following sections.

3.2 Direction–Magnitude Decoupling

At every spatial position p, the local guidance vector $\mathbf{g}_t(p) \in \mathbb{R}^C$ admits the factorization

$$r_t(p) = \|\mathbf{g}_t(p)\|_2, \qquad \mathbf{d}_t(p) = \begin{cases} \mathbf{g}_t(p)/r_t(p), & r_t(p) > 0, \\ \mathbf{0}, & r_t(p) = 0, \end{cases} \qquad \mathbf{g}_t(p) = r_t(p)\mathbf{d}_t(p), \tag{2}$$

where $r_t \in \mathbb{R}^{H\times W}$ is the spatial magnitude field and $\mathbf{d}_t \in \mathbb{R}^{C\times H\times W}$ is the corresponding directional field. Explicitly, $r_t(p) = \big(\sum_{c=1}^{C} g_{t,c}(p)^2\big)^{1/2}$, so the norm is evaluated across channels independently at each spatial location. Define the active guidance support as $S_t = \{p : r_t(p) > 0\}$. For every $p \in S_t$, $\|\mathbf{d}_t(p)\|_2 = 1$, so $r_t(p)$ and $\mathbf{d}_t(p)$ specify, respectively, the radial and directional components of the local guidance. A global norm such as $\|\mathbf{g}_t\|_F$ collapses these spatially varying amplitudes into a single scalar and therefore cannot expose $r_t(p)$ as an independent local control variable.

We fix the model-predicted direction and restrict each local correction to the non-negative ray $\{m\,\mathbf{d}_t(p) : m \geq 0\}$. The resulting admissible guidance class is

$$\mathcal{G}_t = \{\widetilde{\mathbf{g}}_t \mid \widetilde{\mathbf{g}}_t(p) = m_t(p)\mathbf{d}_t(p),\ m_t(p) \geq 0\}, \tag{3}$$

so that, for every $p \in S_t$, $\widetilde{\mathbf{g}}_t(p) = \alpha_t(p)\mathbf{g}_t(p)$ with $\alpha_t(p) = m_t(p)/r_t(p) \geq 0$. Hence the controlled correction remains on the ray generated by the original guidance, with $\|\widetilde{\mathbf{g}}_t(p)\|_2 = m_t(p)$ on S_t. For $p \notin S_t$, the definition $\mathbf{d}_t(p) = \mathbf{0}$ gives $\widetilde{\mathbf{g}}_t(p) = \mathbf{0}$ independently of $m_t(p)$. Thus the correspondence between scalar magnitude and vector guidance is exact on the active support S_t.

This parameterization reduces the optimized degree of freedom at each active location from C directional components to one non-negative scalar. The pretrained model supplies $\mathbf{d}_t = \mathbf{d}(\boldsymbol{x}_t)$, while the controller optimizes the spatial field $r_t \mapsto m_t$; equivalently, $\mathbf{g}_t = r_t\mathbf{d}_t$ is replaced by $\widetilde{\mathbf{g}}_t = m_t\mathbf{d}_t$. For $m_t = r_t$, the original guidance is recovered exactly, $\widetilde{\mathbf{g}}_t = \mathbf{g}_t$, including the zero-guidance locations where both fields vanish. The resulting $H \times W$ magnitude field defines the scalar control parameterization used by the state-wise optimization in Sec. 3.3.

3.3 Optimal Magnitude Control

State-Wise Magnitude Formulation. The raw magnitude field $r_t = r(\boldsymbol{x}_t, c)$ represents the local amplitude of the conditional-unconditional velocity discrepancy and varies with the current latent state $\boldsymbol{x}_t$. Large values of $r_t(p)$ may encode useful conditional responses, whereas strong spatial variation $\|\nabla r_t\|_F$ or abrupt changes across neighboring sampling steps can reduce the spatial and temporal regularity of the control after scaling by w. A global rescaling $r_t \mapsto \kappa r_t$ cannot distinguish these effects, since the same factor κ is applied to every spatial location and preserves their relative magnitude ratios. We therefore keep the directional field $\mathbf{d}_t$ fixed and optimize only a scalar field $m_t \in \mathbb{R}^{H\times W}$. Under the parameterization $\widetilde{\mathbf{g}}_t = m_t \odot \mathbf{d}_t$, the local vector-field control reduces to scalar magnitude optimization while the model-predicted orientation remains fixed. For a fixed state $\boldsymbol{x}_t$ and direction $\mathbf{d}_t$, the state-wise magnitude problem is

$$m_t^\star = \arg\min_m \ \mathcal{J}_t(m) = \frac{1}{2}\|m - r_t\|_F^2 + \frac{\lambda}{2}\|m\|_F^2 + \frac{\gamma}{2}\|\nabla m\|_F^2 + \frac{\mu}{2}\|m - m_t^{\mathrm{ref}}\|_F^2, \tag{4}$$

where $\lambda, \gamma, \mu \geq 0$. The four terms impose fidelity to r_t, energy control through $\|m\|_F^2$, spatial regularity through $\|\nabla m\|_F^2$, and temporal coupling through $\|m - m_t^{\mathrm{ref}}\|_F^2$. The reference m_t^{ref} is inherited from the preceding sampling step, so the current optimum depends on both r_t and the previous magnitude solution. The fidelity term anchors the controlled magnitude to the current model response r_t; without the fidelity term, the update is determined only by the regularization terms and the preceding reference m_t^{ref}. Expanding Eq. (4) gives

$$\mathcal{J}_t(m) = \frac{1+\lambda+\mu}{2}\langle m, m\rangle + \frac{\gamma}{2}\langle \nabla m, \nabla m\rangle - \langle b_t, m\rangle + C_t, \tag{5}$$

where $b_t = r_t + \mu m_t^{\mathrm{ref}}$ and $C_t = \frac{1}{2}\|r_t\|_F^2 + \frac{\mu}{2}\|m_t^{\mathrm{ref}}\|_F^2$. Under periodic or homogeneous Neumann boundary conditions, $\langle m, -\Delta m\rangle = \|\nabla m\|_2^2$. Defining $\mathcal{A} = (1+\lambda+\mu)I - \gamma\Delta$ gives $\mathcal{J}_t(m) = \frac{1}{2}\langle m, \mathcal{A}m\rangle - \langle b_t, m\rangle + C_t$, with gradient operator $\mathcal{A}m - b_t$.

Elliptic Optimality and Spectral Solution. For any $h \neq 0$, the quadratic form satisfies

$$\langle h, \mathcal{A}h\rangle = (1+\lambda+\mu)\|h\|_2^2 + \gamma\|\nabla h\|_2^2 \geq (1+\lambda+\mu)\|h\|_2^2 > 0. \tag{6}$$

Hence $\mathcal{A} \succ 0$ and $\mathcal{A} \succeq (1+\lambda+\mu)I$. The objective $\mathcal{J}_t$ is therefore $(1+\lambda+\mu)$-strongly convex under the stated inner product, which guarantees a unique state-wise optimum. For an admissible perturbation ϕ, let $m_\tau = m + \tau\phi$. The first variation is $\delta\mathcal{J}_t(m;\phi) = \langle (1+\lambda+\mu)m - b_t, \phi\rangle + \gamma\langle \nabla m, \nabla\phi\rangle$. By Green's identity, $\langle \nabla m, \nabla\phi\rangle = -\langle \Delta m, \phi\rangle + \int_{\partial\Omega} \phi\, \partial_n m\, dS$, where the boundary term vanishes under the stated boundary conditions. Therefore, $\delta\mathcal{J}_t(m_t^\star;\phi) = 0$ for every admissible ϕ yields

$$[(1+\lambda+\mu)I - \gamma\Delta]\, m_t^\star = r_t + \mu m_t^{\mathrm{ref}}, \tag{7}$$

which defines the screened elliptic optimality system $\mathcal{A}m_t^\star = b_t$. Since $\mathcal{A} \succ 0$, $m_t^\star = \mathcal{A}^{-1}b_t$ exists uniquely. For an eigenbasis $\{\psi_j\}$ of $-\Delta$ with eigenvalues $\nu_j \geq 0$, $\mathcal{A}\psi_j = (1+\lambda+\mu+\gamma\nu_j)\psi_j$. Therefore, $\sigma(\mathcal{A}) \subset [1+\lambda+\mu, \infty)$ and $\|\mathcal{A}^{-1}\|_2 \leq (1+\lambda+\mu)^{-1}$. The zero-order term separates every spatial mode from zero, while $\gamma\nu_j$ increases the penalty with spatial variation.

For computation, we use the periodic finite-difference realization of the same quadratic problem. Let D denote the discrete spatial gradient and $L = D^\top D$ the corresponding positive semidefinite Laplacian. The discrete optimality system is $[(1+\lambda+\mu)I + \gamma L]m_t^\star = b_t$. On an $H \times W$ periodic grid, we use unit grid spacing $h_x = h_y = 1$ in all experiments. The discrete Fourier transform diagonalizes L, with eigenvalues

$$\Lambda_{q,\ell} = 4\sin^2(\pi q/H) + 4\sin^2(\pi\ell/W). \tag{8}$$

The state-wise optimum used for computation is therefore

$$\widehat{m}_{t,q,\ell}^\star = \frac{\widehat{r}_{t,q,\ell} + \mu\widehat{m}_{t,q,\ell}^{\mathrm{ref}}}{1+\lambda+\mu+\gamma\Lambda_{q,\ell}}, \tag{9}$$

which solves each spatial frequency independently without iterative optimization. This finite-difference spectral form is used in Algorithm 1. In the continuous spectral limit, $\Lambda_{q,\ell}$ corresponds to $\|\omega\|^2$, recovering the spectral representation of Eq. (7).

The discrete optimum also preserves non-negativity. Since $r_t \geq 0$ and the reference is initialized and propagated non-negatively, $b_t = r_t + \mu m_t^{\mathrm{ref}} \geq 0$. Let $a = 1+\lambda+\mu > 0$ and write the discrete system as $(aI + \gamma L)m_t^\star = b_t$. If $m_t^\star(i) < 0$ were a global minimum, the nearest-neighbor finite-difference Laplacian would satisfy $(Lm_t^\star)_i = \sum_j w_{ij}(m_t^\star(i) - m_t^\star(j)) \leq 0$. Consequently, $(aI + \gamma L)m_t^\star(i) < 0$, contradicting $b_t(i) \geq 0$. Hence $m_t^\star \geq 0$. The discrete optimum therefore belongs to the non-negative magnitude space required by Eq. (3), and the reconstruction $m_t^\star \odot \mathbf{d}_t$ remains on the admissible ray at every active guidance location.

Spatial–State Consistency Control Analysis. The spectral representation characterizes the spatial and state-consistency behavior of the controller. Define

$$H_r(q,\ell) = (1+\lambda+\mu+\gamma\Lambda_{q,\ell})^{-1}, \qquad H_{\mathrm{ref}}(q,\ell) = \mu H_r(q,\ell), \tag{10}$$

so that $\widehat{m}_t^\star = H_r\widehat{r}_t + H_{\mathrm{ref}}\widehat{m}_t^{\mathrm{ref}}$. The derivative $\partial_\Lambda H_r = -\gamma(1+\lambda+\mu+\gamma\Lambda)^{-2} \leq 0$ shows that larger Laplacian eigenvalues receive stronger attenuation for $\gamma > 0$. Moreover, $H_r(0) = (1+\lambda+\mu)^{-1}$ and $H_r(\Lambda) \to 0$ as $\Lambda \to \infty$. For $\Lambda_2 > \Lambda_1$ and $\gamma > 0$,

$$\frac{H_r(\Lambda_2)}{H_r(\Lambda_1)} = \frac{1+\lambda+\mu+\gamma\Lambda_1}{1+\lambda+\mu+\gamma\Lambda_2} < 1. \tag{11}$$

Algorithm 1 CtrlGen Sampling

Require: v_θ, c, $\{t_k\}_{k=0}^{K}$, w, λ, γ, μ

Initialize $\boldsymbol{x}_{t_0}$ and $m^{\mathrm{ref}} \leftarrow \mathbf{0}$; precompute $\{\Lambda_{q,\ell}\}$ ▷ *State, reference, and spectrum*

for $k = 0$ **to** $K-1$ **do**

  $\mathbf{g}_k \leftarrow v_\theta(\boldsymbol{x}_{t_k}, t_k, c) - v_\theta(\boldsymbol{x}_{t_k}, t_k, \emptyset)$ ▷ *Raw guidance field*

  $r_k(p) \leftarrow \|\mathbf{g}_k(p)\|_2, \quad \mathbf{d}_k(p) \leftarrow \begin{cases} \mathbf{g}_k(p)/r_k(p), & r_k(p) > 0, \\ \mathbf{0}, & r_k(p) = 0 \end{cases}$ ▷ *Direction–magnitude decoupling*

  $\widehat{m}^{\star}_{k,q,\ell} \leftarrow \dfrac{\widehat{r}_{k,q,\ell} + \mu \widehat{m}^{\mathrm{ref}}_{q,\ell}}{1 + \lambda + \mu + \gamma \Lambda_{q,\ell}}, \quad m^{\star}_k \leftarrow \mathcal{F}^{-1}[\widehat{m}^{\star}_k]$ ▷ *State-wise spectral optimum*

  $\widetilde{\mathbf{g}}^{\star}_k \leftarrow m^{\star}_k \odot \mathbf{d}_k$ ▷ *Direction-preserving reconstruction*

  $\widehat{\mathbf{v}}_k \leftarrow v_\theta(\boldsymbol{x}_{t_k}, t_k, \emptyset) + w \widetilde{\mathbf{g}}^{\star}_k$ ▷ *Controlled velocity field*

  $\boldsymbol{x}_{t_{k+1}} \leftarrow \mathrm{SolverStep}(\boldsymbol{x}_{t_k}, \widehat{\mathbf{v}}_k, t_k, t_{k+1})$ ▷ *State transition*

  $m^{\mathrm{ref}} \leftarrow m^{\star}_k$ ▷ *State-consistency update*

end for

return $\boldsymbol{x}_{t_K}$

The optimal magnitude controller therefore retains a finite response for slowly varying components while progressively suppressing components with larger spatial variation, as induced directly by $\gamma\|Dm\|_F^2$. The spectral representation further separates the roles of the three coefficients: λ produces global magnitude shrinkage through the uniform spectral shift, γ strengthens attenuation according to $\Lambda_{q,\ell}$, and μ couples the current solution to the preceding magnitude reference. For each spatial mode, the reference recursion has coefficient $q_{q,\ell} = \mu/(1+\lambda+\mu+\gamma\Lambda_{q,\ell})$, with $0 \le q_{q,\ell} \le \mu/(1+\lambda+\mu) < 1$. For a fixed input sequence, the influence of the preceding reference therefore decays geometrically across sampling steps, providing state-consistency regularization without introducing an additional filtering rule. This contraction characterizes only the state-wise magnitude recursion and does not imply trajectory-level stability of the nonlinear generative dynamics.

3.4 State-Dependent Guidance Control

Sec. 3.3 gives the unique magnitude solution $m_t^\star$ for a fixed latent state $\boldsymbol{x}_t$. The scalar optimum is lifted to the original guidance space through the directional field defined in Sec. 3.2. The admissible control at each location is restricted to the ray $\mathcal{R}_t(p) = \{m\,\mathbf{d}_t(p) : m \ge 0\}$, with $\|\mathbf{d}_t(p)\|_2 = 1$ whenever $p \in S_t$, where $S_t = \{p : r_t(p) > 0\}$ denotes the active guidance support. On S_t, the map $T_t : m \mapsto m \odot \mathbf{d}_t$ preserves the local amplitude, $\|T_t(m)(p)\|_2 = m(p)$ for $m(p) \ge 0$. The unique scalar optimum therefore determines the controlled guidance:

$$\widetilde{\mathbf{g}}_t^\star = m_t^\star \odot \mathbf{d}_t, \qquad \dot{\boldsymbol{x}}_t = v_\theta(\boldsymbol{x}_t, t, \emptyset) + w\widetilde{\mathbf{g}}_t^\star, \tag{12}$$

where $m_t^\star = \arg\min_m \mathcal{J}_t(m)$ and $m_t^\star \ge 0$ follows from Sec. 3.3. Thus Eq. (4) defines the magnitude optimization under the direction-preserving parameterization $\widetilde{\mathbf{g}}_t = T_t(m)$, with exact correspondence restricted to S_t. For $p \in S_t$, $\mathbf{d}_t(p) = \mathbf{g}_t(p)/r_t(p)$ and $\widetilde{\mathbf{g}}_t^\star(p) = \alpha_t^\star(p)\mathbf{g}_t(p)$, where $\alpha_t^\star(p) = m_t^\star(p)/r_t(p) \ge 0$. Moreover, $\|\widetilde{\mathbf{g}}_t^\star(p)\|_2 = m_t^\star(p)$, and whenever both vectors are nonzero,

$$\frac{\widetilde{\mathbf{g}}_t^\star(p)}{\|\widetilde{\mathbf{g}}_t^\star(p)\|_2} = \frac{\mathbf{g}_t(p)}{\|\mathbf{g}_t(p)\|_2} = \mathbf{d}_t(p), \tag{13}$$

establishing that the state-wise optimum modifies only the local control amplitude while preserving the semantic orientation supplied by the pretrained flow.

The optimal guidance remains state dependent because $r_t = r(\boldsymbol{x}_t)$ and $\mathbf{d}_t = \mathbf{d}(\boldsymbol{x}_t)$ are recomputed from the current latent state. Let $\mathcal{K}_{\lambda,\gamma,\mu} = [(1+\lambda+\mu)I - \gamma\Delta]^{-1}$ denote the continuous elliptic solution operator. From Eq. (7), $m_t^\star = \mathcal{K}_{\lambda,\gamma,\mu}[r(\boldsymbol{x}_t) + \mu m_t^{\mathrm{ref}}]$. Substituting $m_t^\star$ into Eq. (12) gives

$$\mathbf{u}_t^\star = \pi^\star(\boldsymbol{x}_t, m_t^{\mathrm{ref}}) = w\,\mathcal{K}_{\lambda,\gamma,\mu}\left[r(\boldsymbol{x}_t) + \mu m_t^{\mathrm{ref}}\right] \odot \mathbf{d}(\boldsymbol{x}_t), \tag{14}$$

which defines the state-wise feedback control induced by the optimal magnitude field. Equivalently, $\pi^\star(\boldsymbol{x}_t, m_t^{\mathrm{ref}}) = w\,T_t\big(\mathcal{K}_{\lambda,\gamma,\mu}[r(\boldsymbol{x}_t) + \mu m_t^{\mathrm{ref}}]\big)$. The current state therefore determines $r(\boldsymbol{x}_t)$ and $\mathbf{d}(\boldsymbol{x}_t)$, while the scalar magnitude remains the optimized degree of freedom. Algorithm 1 implements the corresponding discrete operator through the Laplacian spectrum $\Lambda_{q,\ell}$.

Table 1: **Quantitative evaluation of guidance methods.** Lower (↓) FID and higher (↑) CLIP, Aesthetic, PickScore, HPSv2, HPSv2.1, and MPS are preferred. Best guidance results are shown in **bold**, and second-best results are underlined. Our method is highlighted in gray.

Model	Guidance	FID↓	CLIP↑	Aesthetic↑	PickScore↑	HPSv2↑	HPSv2.1↑	MPS↑
SD3.5L	CFG (Ho & Salimans, 2022)	21.421	0.3681	5.5875	0.2190	0.2930	0.2842	7.2476
	APG (Sadat et al., 2024)	20.736	0.3693	5.6037	0.2204	0.2936	0.2857	7.4182
	CFG-Zero* (Fan et al., 2025)	20.317	0.3691	5.6124	0.2195	0.2942	0.2862	7.0430
	Rect-CFG++ (Saini et al., 2026)	20.550	0.3655	5.5663	0.2173	0.2887	0.2748	6.7854
	SMC-CFG (Wang et al., 2026)	20.044	0.3694	5.5790	0.2211	**0.2945**	0.2875	7.5719
	CtrlGen	**19.782**	**0.3712**	**5.6287**	**0.2310**	0.2911	**0.2978**	**9.0147**
FLUX.1-dev	CFG (Ho & Salimans, 2022)	27.323	0.3692	5.5397	0.2228	0.2917	0.2828	7.8387
	APG (Sadat et al., 2024)	26.847	0.3726	5.6774	0.2251	0.2968	0.2947	8.0615
	CFG-Zero* (Fan et al., 2025)	26.901	0.3742	5.7053	0.2262	0.2987	0.2992	8.1573
	Rect-CFG++ (Saini et al., 2026)	27.219	0.3728	5.6909	0.2252	0.2974	0.2963	7.9746
	SMC-CFG (Wang et al., 2026)	26.398	0.3743	5.7342	0.2268	0.2986	0.3021	8.2307
	CtrlGen	**26.338**	**0.3761**	**5.7618**	**0.2281**	**0.3012**	**0.3046**	**9.3475**
Qwen-Image	CFG (Ho & Salimans, 2022)	35.431	0.3815	5.5995	0.2260	0.2996	0.3038	8.1852
	APG (Sadat et al., 2024)	34.886	0.3831	5.6176	0.2267	0.3008	0.3064	8.2689
	CFG-Zero* (Fan et al., 2025)	35.391	0.3822	**5.6598**	**0.2279**	0.3019	0.3092	8.3739
	Rect-CFG++ (Saini et al., 2026)	34.371	0.3834	5.6007	0.2276	0.3017	0.3068	8.1026
	SMC-CFG (Wang et al., 2026)	**33.371**	0.3856	5.6289	0.2275	0.3026	0.3105	8.4320
	CtrlGen	34.137	**0.3865**	5.6312	0.2278	**0.3058**	**0.3201**	**9.5476**

After applying $\mathbf{u}_t^\star$, the flow advances $\boldsymbol{x}_t \mapsto \boldsymbol{x}_{t+\Delta t}$. The next model evaluation produces $\mathbf{g}_{t+\Delta t} = \mathbf{g}(\boldsymbol{x}_{t+\Delta t}, c)$ and hence $r_{t+\Delta t} = r(\boldsymbol{x}_{t+\Delta t}, c)$ and $\mathbf{d}_{t+\Delta t} = \mathbf{d}(\boldsymbol{x}_{t+\Delta t}, c)$. With the reference propagated from the preceding control state, the closed-loop recursion is $\boldsymbol{x}_t \mapsto \mathbf{g}_t \mapsto (r_t, \mathbf{d}_t) \mapsto m_t^\star \mapsto \mathbf{u}_t^\star \mapsto \boldsymbol{x}_{t+\Delta t}$, with $m_t^\star = \arg\min_m \mathcal{J}_t(m; \boldsymbol{x}_t, m_t^{\mathrm{ref}})$ resolved at each sampling state. Algorithm 1 summarizes the complete feedback-controlled sampling process. The optimality claim is state-wise and conditional on the model-predicted direction: strict convexity gives a unique $m_t^\star$, the map T_t constructs the corresponding direction-preserving guidance, and Eq. (14) applies $\mathbf{u}_t^\star$ to the flow dynamics. Repeated recomputation yields a sequence of state-wise optimal controls rather than a trajectory-level optimum of the generator.

4 Experiments

4.1 Experimental Setups

Datasets & Evaluation. We evaluate on 5,000 image-text pairs from MS-COCO (Lin et al., 2014) using three representative flow-based T2I models: Stable Diffusion 3.5 Large (Esser et al., 2024), FLUX.1-dev (Black Forest Labs, 2024), and Qwen-Image (Wu et al., 2025). We compare CtrlGen with standard Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) and recent training-free methods, including APG (Sadat et al., 2024), CFG-Zero (Fan et al., 2025), Rectified-CFG++ (Chung et al., 2025), and SMC-CFG (Wang et al., 2026). All methods use identical prompts, model checkpoints, sampling settings, and initial noise seeds for a controlled and fair comparison. We evaluate generation quality, text-image alignment, and preference-oriented quality using FID (Heusel et al., 2017), CLIP Score (Radford et al., 2021; Hessel et al., 2021), Aesthetic Score (Schuhmann, 2022), PickScore (Kirstain et al., 2023), HPSv2/HPSv2.1 (Wu et al., 2023), and MPS (Zhang et al., 2024). Lower FID and higher values otherwise indicate better performance.

Implementation Details. All experiments are conducted on a single NVIDIA RTX A6000 GPU with 48 GB of memory at 1024×1024 resolution. For FLUX.1-dev, which is guidance-distilled, we follow the SMC-CFG protocol and fix the embedded guidance to 1, while setting the guidance scale to $w = 3.0$ for the conditional–unconditional guidance discrepancy controlled by CtrlGen. The guidance scales are 7.5 for SD3.5L and 3.0 for Qwen-Image. The control coefficients λ, γ, and μ are fixed per model and shared across all evaluation settings. The default model-specific CtrlGen values are $\lambda = 5.0$, $\mu = 0.5$, and $\gamma = 0.1$ for FLUX.1-dev and Qwen-Image. For SD3.5L, the main evaluation uses $\lambda = 1.0$, $\mu = 0.1$, and $\gamma = 0.01$. All evaluations use 30 sampling steps with the default sampler and timestep schedule of each backbone. Magnitude control is applied directly on the native latent spatial grid at every sampling step without additional resizing.

4.2 Results and Analysis

Quantitative Comparison. Table 1 compares our method with CFG and recent training-free guidance methods across three representative flow-based T2I models. Our method achieves strong overall performance across different model architectures. On SD3.5L, it achieves the best results on six of the seven reported metrics, with MPS improving from 7.5719 to 9.0147 over SMC-CFG. On FLUX.1-dev, our method achieves the best performance on all reported metrics and improves MPS from 8.2307 to 9.3475. On Qwen-Image, it obtains the best CLIP, HPSv2, HPSv2.1, and MPS, while ranking second on FID, Aesthetic, and PickScore among guidance methods. The unguided Qwen-Image backbone yields a substantially lower FID of 24.894 while lacking explicit conditional guidance, so we report it only as a reference and exclude it from the guidance-method ranking. MPS further increases from 8.4320 to 9.5476. The consistent gains in preference-oriented metrics across all three backbones support the effectiveness and generality of magnitude regulation.

Component Ablation. Table 2 evaluates each component of our magnitude controller across three flow-based T2I models. The *w/o Decomp.* variant applies the same regularization directly to the vector guidance field instead of its scalar magnitude. Direct vector-field regularization can modify both the amplitude and orientation of the original guidance. In contrast, the decomposition explicitly confines the control to magnitude and preserves the model-predicted direction. Its performance drop consistently supports the benefit of magnitude control with a fixed guidance direction. Among the regularization terms, the spatial term γ has an effect on generation quality through control of local magnitude variation. The energy term λ and state-consistency term μ provide systematic, model-specific trade-offs across metrics. In particular, removal of λ retains alignment scores but reduces fidelity and preference quality. The full formulation achieves the best performance across all three backbones.

Table 2: **Component ablation across T2I models.** We evaluate the direction–magnitude decomposition and the effects of three regularization terms: the λ-, γ-, and μ-terms. Best results are shown in **bold**.

Model	Variant	FID↓	CLIP↑	Aes.↑	Pick↑	HPSv2.1↑	MPS↑
SD3.5L	w/o Decomp.	21.284	0.3667	5.5381	0.2237	0.2836	8.1763
	w/o λ-term	20.417	0.3706	5.5869	0.2294	0.2941	8.6817
	w/o γ-term	20.963	0.3697	5.5476	0.2272	0.2908	8.3574
	w/o μ-term	20.351	0.3691	5.6013	0.2265	0.2889	8.5246
	CtrlGen	**19.782**	**0.3712**	**5.6287**	**0.2310**	**0.2978**	**9.0147**
FLUX.1-dev	w/o Decomp.	27.816	0.3711	5.6417	0.2226	0.2948	8.4931
	w/o λ-term	26.714	0.3754	5.7068	0.2273	0.3017	9.0184
	w/o γ-term	27.403	0.3738	5.6631	0.2259	0.2991	8.7526
	w/o μ-term	26.926	0.3729	5.7196	0.2248	0.2976	8.8673
	CtrlGen	**26.338**	**0.3761**	**5.7618**	**0.2281**	**0.3046**	**9.3475**
Qwen-Image	w/o Decomp.	35.087	0.3826	5.5749	0.2241	0.3078	8.8037
	w/o λ-term	34.603	**0.3868**	5.6027	**0.2282**	0.3163	9.1965
	w/o γ-term	34.917	0.3843	5.5836	0.2267	0.3121	8.9148
	w/o μ-term	34.426	0.3835	5.6148	0.2253	0.3137	9.0732
	CtrlGen	**34.137**	0.3865	**5.6312**	0.2278	**0.3201**	**9.5476**

Hyperparameter Sensitivity. Table 3 studies the effects of λ, γ, and μ on FLUX.1-dev as a representative backbone. We vary one parameter at a time while keeping the other two fixed. The energy weight λ controls overall guidance magnitude. A small value preserves stronger conditional responses but provides weaker energy regularization, whereas an excessive value can suppress useful signals. The spatial weight γ controls local magnitude variation. Insufficient regularization leaves spatial fluctuations, while excessive smoothing can weaken fine-grained guidance. The temporal weight μ promotes consistency across adjacent sampling states, but strong reference coupling limits adaptation to the current latent state. Overall, the selected setting $\lambda = 5.0$, $\gamma = 0.10$, and $\mu = 0.50$ provides the best balance across generation quality, text alignment, and preference metrics under the evaluated sampling settings.

Table 3: **Hyperparameter sensitivity on FLUX.1-dev.** We vary one parameter while fixing the remaining two at their default values. Best results are shown in **bold**; the selected setting is highlighted in gray.

Group	λ	γ	μ	FID↓	CLIP↑	Aes.↑	Pick↑	HPSv2.1↑	MPS↑
λ	1.0	0.10	0.50	26.781	**0.3764**	5.7249	**0.2283**	0.3020	9.1068
	3.0	0.10	0.50	26.492	0.3762	5.7476	0.2280	0.3035	9.2741
	5.0	0.10	0.50	**26.338**	0.3761	**5.7618**	0.2281	**0.3046**	**9.3475**
	7.0	0.10	0.50	26.574	0.3753	5.7351	0.2272	0.3031	9.2164
γ	5.0	0.02	0.50	27.061	**0.3765**	5.6917	0.2274	0.3002	8.9648
	5.0	0.05	0.50	26.646	0.3763	5.7285	0.2278	0.3027	9.2016
	5.0	**0.10**	0.50	**26.338**	0.3761	**5.7618**	**0.2281**	**0.3046**	**9.3475**
	5.0	0.20	0.50	26.721	0.3749	5.7162	0.2266	0.3019	9.1087
μ	5.0	0.10	0.10	26.812	**0.3764**	5.7363	0.2276	0.3014	9.0789
	5.0	0.10	0.30	26.487	0.3762	5.7541	**0.2282**	0.3032	9.2367
	5.0	0.10	**0.50**	**26.338**	0.3761	**5.7618**	0.2281	**0.3046**	**9.3475**
	5.0	0.10	0.80	26.596	0.3755	5.7473	0.2274	0.3036	9.1826

Guidance-Scale Robustness. Figure 3 examines the effect of the guidance scale w on FLUX.1-dev. Standard CFG is sensitive to large guidance scales, with quality and preference scores declining beyond a moderate range. Competing guidance methods mitigate this behavior to different degrees. CtrlGen remains stable over a wider range of w and maintains competitive FID, CLIP, HPSv2.1, and MPS scores at larger guidance scales. This robustness is consistent with magnitude control, which suppresses excessive local responses while preserving the model-predicted guidance direction.

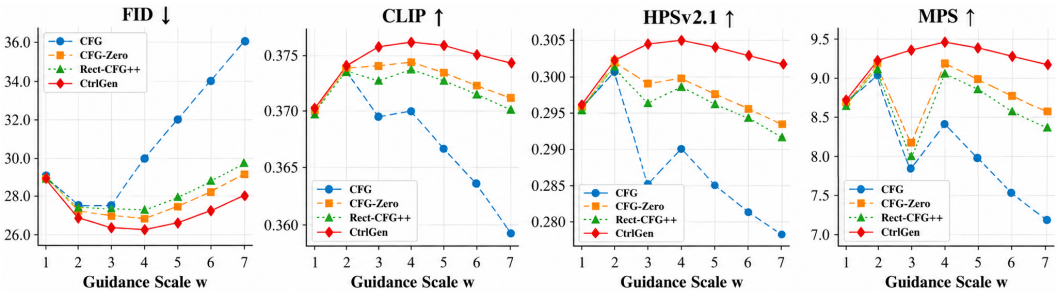

Figure 3: **Robustness to the guidance scale on FLUX.1-dev.** We vary the guidance scale w and report FID, CLIP, HPSv2.1, and MPS for different guidance methods. CtrlGen maintains a broader high-quality range as w increases and shows less degradation at large guidance scales.

Visual Analysis of Magnitude Control. Figure 4 examines the effect of magnitude control on local image quality in FLUX.1-dev. All variants use the same prompt and foreground and background regions for direct visual comparison. Without direction–magnitude decomposition, local structures become less distinct. Removing the spatial regularizer γ introduces visible texture fluctuations around the subjects and distant street structures. The variants without λ or μ preserve the overall scene but show less stable details and cross-region consistency. The full model retains clearer subject boundaries and a smoother, more coherent background across regions and scales.



Figure 4: **Visual analysis on FLUX.1-dev.** Enlarged foreground and background regions compare different CtrlGen ablations under the same prompt: "Two women waiting at a bench next to a street."

Computational Cost. Table 4 compares CtrlGen with CFG and SMC-CFG in terms of peak GPU memory and total FLOPs over 30-step generation. CtrlGen adds magnitude decomposition and elliptic control without additional model evaluations and requires only a spectral solution for the scalar magnitude field. At 1024^2 resolution, peak VRAM increases from 33.59 GB to 33.67 GB. The additional FLOPs remain computationally negligible across both image resolutions, with the number of sampling steps unchanged. Thus, CtrlGen incurs only marginal computational overhead relative to standard CFG and retains the original backbone sampling.

Table 4: **Computational cost comparison.** Peak VRAM and total inference FLOPs for 30-step generation at 512^2 and 1024^2 resolutions.

Size	Guidance	Peak VRAM (GB)	FLOPs (G)
512^2	CFG	31.99	1203370.06
	SMC-CFG	31.99	1203370.07
	CtrlGen	32.03	1203370.42
1024^2	CFG	33.59	3590870.89
	SMC-CFG	33.59	3590870.93
	CtrlGen	33.67	3590872.14

5 Conclusion

In this work, we identify the coupled regulation of guidance direction and magnitude as a limitation of flow-based generation. Our proposed **CtrlGen** decouples these components and regulates the magnitude while preserving the model-predicted direction. We formulate magnitude regulation as a state-wise control problem and derive an elliptic optimality condition with an explicit spectral solution. Experiments on SD3.5L, FLUX.1-dev, and Qwen-Image show consistent improvements in overall generation quality, text alignment, and preference-oriented metrics across diverse settings without model retraining or iterative optimization under practical inference-time constraints.

Limitation. CtrlGen cannot correct an inaccurate guidance direction, and fixed regularization weights may suppress fine-grained variations in more complex scenes. Future work will explore adaptive control of guidance direction and magnitude across diverse sampling states and broader conditions.

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

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Julius Berner, Lorenz Richter, and Karen Ullrich. An optimal control perspective on diffusion-based generative modeling. *arXiv preprint arXiv:2211.01364*, 2022.

Eric Tillmann Bill, Enis Simsar, and Thomas Hofmann. Focus: Optimal control for multi-entity world modeling in text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5915–5924, 2026.

Black Forest Labs. FLUX. GitHub repository, 2024. URL https://github.com/black-forest-labs/flux.

Arthur Earl Bryson. *Applied optimal control: optimization, estimation and control*. CRC press, 1975.

Angela Castillo, Jonas Kohler, Juan C Pérez, Juan Pablo Pérez, Albert Pumarola, Bernard Ghanem, Pablo Arbeláez, and Ali Thabet. Adaptive guidance: Training-free acceleration of conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1962–1970, 2025.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 30824–30850, 2025.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Carles Domingo i Enrich, Michal Drozdzal, Brian Karrer, and Ricky TQ Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *International Conference on Learning Representations*, volume 2025, pp. 53791–53846, 2025.

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.

Weichen Fan, Amber Yijia Zheng, Raymond A Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. *arXiv preprint arXiv:2502.02150*, 2025.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, volume 2025, pp. 34668–34684, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Jeongsol Kim, Yeobin Hong, Jonghyun Park, and Jong Chul Ye. Flowalign: Trajectory-regularized, inversion-free flow-based image editing. In *International Conference on Learning Representations*, volume 2026, pp. 121814–121846, 2026.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5392–5399. IEEE, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

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 *The twelfth international conference on learning representations*, 2023.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Pinelopi Papalampidi, Olivia Wiles, Ira Ktena, Aleksandar Shtedritski, Emanuele Bugliarello, Ivana Kajić, Isabela Albuquerque, and Aida Nematzadeh. Dynamic classifier-free diffusion guidance via online feedback. In *International Conference on Learning Representations*, volume 2026, pp. 127362–127378, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.

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

Shreshth Saini, Shashank Gupta, and Alan Bovik. Rectified cfg++ for flow based models. *Advances in Neural Information Processing Systems*, 38:149034–149074, 2026.

Christoph Schuhmann. LAION-Aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. Accessed: 2023-11-10.

Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9370–9379. IEEE, 2024.

Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. Conditional simulation using diffusion schrödinger bridges. In *Uncertainty in Artificial Intelligence*, pp. 1792–1802. PMLR, 2022.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025.

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

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Shikun Sun, Longhui Wei, Zhicai Wang, Zixuan Wang, Junliang Xing, Jia Jia, and Qi Tian. Inner classifier-free guidance and its taylor expansion for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 4360–4376, 2024.

Hanyang Wang, Yiyang Liu, Jiawei Chi, Fangfu Liu, Ran Xue, and Yueqi Duan. Cfg-ctrl: Control-based classifier-free diffusion guidance. *arXiv preprint arXiv:2603.03281*, 2026.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *arXiv preprint arXiv:2404.13040*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Mengfei Xia, Nan Xue, Yujun Shen, Ran Yi, Tieliang Gong, and Yong-Jin Liu. Rectified diffusion guidance for conditional generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13371–13380. IEEE, 2025.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8018–8027. IEEE, 2024.

Candi Zheng and Yuan Lan. Characteristic guidance: Non-linear correction for diffusion model at large guidance scale. *arXiv preprint arXiv:2312.07586*, 2023.

CtrlGen: Optimal Control for Text-to-Image Generation Supplementary Material

CONTENTS

A Detailed Theoretical Derivations **14**

A.1 State-Wise Reduction of Guidance Control 14

A.2 Elliptic Optimality Condition 16

A.3 Well-Posedness and Properties of the Optimal Magnitude 18

A.4 Spectral Characterization and Discrete Realization 19

B Additional Experimental Analysis **21**

B.1 Magnitude Regulation and Sampling Dynamics 21

B.2 Controlled Spatial–State Consistency Ablation 22

B.3 Statistical Reliability Analysis 23

B.4 Evaluation on T2I-CompBench 24

B.5 Target–Background Analysis 25

B.6 Additional Qualitative Results 26

C Limitations and Future Work **26**

A DETAILED THEORETICAL DERIVATIONS

Table S1: **Control formulation of CtrlGen.** Summary of the state-wise magnitude-control construction.

Component	Form	Operator / Constraint	Role
Raw guidance	$\mathbf{g}_t = \mathbf{v}_t^c - \mathbf{v}_t^u$	$r_t = \|\mathbf{g}_t\|_2$	Model response
Decomposition	$\mathbf{g}_t = r_t \mathbf{d}_t$	$\|\mathbf{d}_t(p)\|_2 = 1$	Direction–magnitude split
Admissible guidance	$\widetilde{\mathbf{g}}_t = m_t \mathbf{d}_t$	$m_t(p) \geq 0$	Ray constraint
Magnitude objective	$m_t^\star = \arg\min_m \mathcal{J}_t(m)$	λ, γ, μ	State-wise optimization
Elliptic condition	$\mathcal{A} m_t^\star = b_t$	$\mathcal{A} = aI - \gamma\Delta$	Unique optimum
Discrete solution	$\widehat{m}_{k,\ell}^\star = \widehat{b}_{k,\ell}/(a + \gamma\Lambda_{k,\ell})$	$\Lambda_{k,\ell} \geq 0$	Direct FFT solve
Reconstruction	$\widetilde{\mathbf{g}}_t^\star = m_t^\star \mathbf{d}_t$	$\widetilde{\mathbf{g}}_t^\star \parallel \mathbf{g}_t$ on S_t	Direction preservation
Feedback control	$\mathbf{u}_t^\star = w\widetilde{\mathbf{g}}_t^\star$	$\pi^\star(\boldsymbol{x}_t, m_t^{\mathrm{ref}})$	State-dependent control

A.1 STATE-WISE REDUCTION OF GUIDANCE CONTROL

Controlled dynamics. Let $\boldsymbol{x}_t \in \mathbb{R}^{C\times H\times W}$ denote the latent state and $\mathbf{g}_t = v_\theta(\boldsymbol{x}_t, t, c) - v_\theta(\boldsymbol{x}_t, t, \emptyset)$ the conditional-unconditional velocity discrepancy. The guided dynamics are $\dot{\boldsymbol{x}}_t = v_\theta(\boldsymbol{x}_t, t, \emptyset) + w\mathbf{g}_t$. At each spatial location p, we decompose

$$\mathbf{g}_t(p) = \underbrace{\|\mathbf{g}_t(p)\|_2}_{r_t(p)} \underbrace{\begin{cases} \mathbf{g}_t(p)/\|\mathbf{g}_t(p)\|_2, & r_t(p) > 0, \\ \mathbf{0}, & r_t(p) = 0, \end{cases}}_{\mathbf{d}_t(p)}, \qquad r_t(p) \geq 0, \tag{A.1}$$

where $r_t \in \mathbb{R}^{H\times W}$ describes local guidance strength, while $\mathbf{d}_t \in \mathbb{R}^{C\times H\times W}$ retains the model-predicted direction. On the active support $S_t = \{p : r_t(p) > 0\}$, the directional field satisfies $\|\mathbf{d}_t(p)\|_2 = 1$; outside S_t, $\mathbf{d}_t(p) = \mathbf{0}$.

Restricted control space. CtrlGen fixes $\mathbf{d}_t$ at the current state and replaces r_t by a scalar field m_t, i.e., $\mathbf{g}_t = r_t\mathbf{d}_t \mapsto \widetilde{\mathbf{g}}_t = m_t\mathbf{d}_t$. For $p \in S_t$, this gives $\widetilde{\mathbf{g}}_t(p) = [m_t(p)/r_t(p)]\mathbf{g}_t(p)$. Thus $m_t(p) \geq 0$ changes the local strength without reversing the guidance ray and reduces the optimized control dimension from $C \times H \times W$ to $H \times W$.

Exact reduction under direction preservation. The scalar formulation follows from the direction-preserving restriction on the active support. For any admissible field $\widetilde{\mathbf{g}}_t = m_t\mathbf{d}_t$, the unit-norm property in Eq. (A.1) gives

$$\begin{aligned}\sum_{p\in S_t} \|\widetilde{\mathbf{g}}_t(p) - \mathbf{g}_t(p)\|_2^2 &= \sum_{p\in S_t} \|[m_t(p) - r_t(p)]\mathbf{d}_t(p)\|_2^2 \\ &= \sum_{p\in S_t} [m_t(p) - r_t(p)]^2 \underbrace{\|\mathbf{d}_t(p)\|_2^2}_{=1} = \sum_{p\in S_t} [m_t(p) - r_t(p)]^2,\end{aligned} \tag{A.2}$$

where locations outside S_t have no directional component because $\mathbf{d}_t(p) = \mathbf{0}$. Likewise, $\sum_{p\in S_t} \|\widetilde{\mathbf{g}}_t(p)\|_2^2 = \sum_{p\in S_t} m_t(p)^2$. Hence vector guidance fidelity and control energy correspond exactly to their scalar magnitude counterparts on the active guidance support. The full-grid magnitude field additionally participates in spatial and reference regularization, so the scalar objective should not be interpreted as an unconditional isometric reduction over locations outside S_t.

More generally, define the direction-preserving admissible set $\mathcal{U}_t = \{m \odot \mathbf{d}_t : m \in \mathbb{R}^{H\times W},\ m \geq 0\}$. The mapping

$$\mathcal{R}_{\mathbf{d}_t} : m \longmapsto m \odot \mathbf{d}_t, \qquad \underbrace{\sum_{p\in S_t} \|\mathcal{R}_{\mathbf{d}_t}(m_1)(p) - \mathcal{R}_{\mathbf{d}_t}(m_2)(p)\|_2^2}_{\text{vector-space distance on } S_t} = \underbrace{\sum_{p\in S_t} |m_1(p) - m_2(p)|^2}_{\text{magnitude-space distance on } S_t}, \tag{A.3}$$

is therefore an isometry between the scalar magnitude field and the direction-preserving vector representation after restriction to the active guidance support. This correspondence identifies $m_t|_{S_t}$ as the control coordinate associated with the direction-preserving guidance field, while the full-grid m_t remains the variable of the regularized state-wise optimization.

Trajectory-level control perspective. The state-wise construction can be contrasted with a continuous trajectory-level control problem. Let m_τ denote a time-dependent magnitude field over an interval $[t, t + \Delta t]$. With the directional component constrained to the model-predicted field and the initial state $\boldsymbol{x}_t$ fixed, the controlled trajectory satisfies

$$\dot{\boldsymbol{x}}_\tau = \underbrace{v_\theta(\boldsymbol{x}_\tau, \tau, \emptyset)}_{\text{base transport}} + \underbrace{w\, m_\tau \odot \mathbf{d}(\boldsymbol{x}_\tau, \tau, c)}_{\text{direction-preserving control}}, \tag{A.4}$$

and a corresponding trajectory-level magnitude functional can be written as

$$\mathcal{J}_{[t,t+\Delta t]}[m] = \int_t^{t+\Delta t} \left[\underbrace{\frac{1}{2}\|m_\tau - r(\boldsymbol{x}_\tau, \tau, c)\|_F^2}_{\text{fidelity}} + \underbrace{\frac{\lambda}{2}\|m_\tau\|_F^2}_{\text{energy}} + \underbrace{\frac{\gamma}{2}\|\nabla m_\tau\|_F^2}_{\text{spatial}} + \underbrace{\frac{\mu}{2}\|m_\tau - m_\tau^{\text{ref}}\|_F^2}_{\text{reference}} \right] d\tau. \tag{A.5}$$

This formulation couples m_τ to future latent states because $r(\boldsymbol{x}_\tau, \tau, c)$ and $\mathbf{d}(\boldsymbol{x}_\tau, \tau, c)$ both depend on the controlled trajectory. Solving the full trajectory-level problem would therefore require joint optimization over the nonlinear state evolution and the time-dependent magnitude field. CtrlGen does not claim a global optimum for this coupled trajectory problem.

State-wise control. CtrlGen instead resolves a local control problem along the sampling trajectory. At the current state $\boldsymbol{x}_t$, we evaluate

$$\underbrace{r_t = r(\boldsymbol{x}_t, t, c)}_{\text{state-conditioned magnitude}}, \qquad \underbrace{\mathbf{d}_t = \mathbf{d}(\boldsymbol{x}_t, t, c)}_{\text{state-conditioned direction}}, \tag{A.6}$$

and hold these quantities fixed within the current magnitude optimization. Conditioning on the current state therefore defines a spatial variational subproblem over m_t, rather than an exact reduction of Eq. (A.5). After the controlled update advances $\boldsymbol{x}_t$, the model evaluates new conditional and unconditional velocities, which produce new $r_{t+\Delta t}$ and $\mathbf{d}_{t+\Delta t}$. Thus the state-wise construction avoids coupled trajectory optimization while retaining feedback dependence on the evolving latent state.

Magnitude objective. For a fixed state, the controlled magnitude should remain close to the model response while controlling amplitude, spatial variation, and deviation from the reference magnitude. We define

$$m_t^\star = \arg\min_m \left\{ \underbrace{\frac{1}{2}\|m - r_t\|_F^2}_{\text{fidelity}} + \underbrace{\frac{\lambda}{2}\|m\|_F^2}_{\text{energy}} + \underbrace{\frac{\gamma}{2}\|\nabla m\|_F^2}_{\text{spatial}} + \underbrace{\frac{\mu}{2}\|m - m_t^{\text{ref}}\|_F^2}_{\text{reference}} \right\}, \tag{A.7}$$

where $\lambda, \gamma, \mu \geq 0$. The fidelity term anchors the controlled magnitude to the current model response r_t; without this term, the update is determined only by the regularization terms and the preceding reference m_t^{ref}. On S_t, the fidelity and energy terms correspond exactly to the direction-preserving vector quantities in Eq. (A.2), while the spatial and reference terms impose structure directly on the full-grid magnitude control field.

Operator form. Expanding the quadratic terms and collecting all terms independent of m into C_t gives $\mathcal{J}_t(m) = \frac{1+\lambda+\mu}{2}\langle m, m\rangle + \frac{\gamma}{2}\langle \nabla m, \nabla m\rangle - \langle r_t + \mu m_t^{\text{ref}}, m\rangle + C_t$, where $C_t = \frac{1}{2}\|r_t\|_F^2 + \frac{\mu}{2}\|m_t^{\text{ref}}\|_F^2$. Under periodic or homogeneous Neumann boundary conditions, $\langle \nabla m, \nabla m\rangle = \langle m, -\Delta m\rangle$. Hence

$$\mathcal{J}_t(m) = \frac{1}{2}\langle m, \mathcal{A}m\rangle - \langle b_t, m\rangle + C_t, \qquad \underbrace{\mathcal{A}}_{\text{control operator}} = (1+\lambda+\mu)I - \gamma\Delta, \qquad \underbrace{b_t}_{\text{effective forcing}} = r_t + \mu m_t^{\text{ref}}, \tag{A.8}$$

which exposes the screened elliptic operator induced by the spatial regularization. The reduced problem separates the state-dependent forcing b_t from the linear operator $\mathcal{A}$ determined by λ, γ, μ and the spatial boundary conditions.

A.2 Elliptic Optimality Condition

Functional setting. Let $\Omega \subset \mathbb{R}^2$ denote the spatial domain of the magnitude field. For $\gamma > 0$, we consider $m, \phi \in H^1(\Omega)$ under homogeneous Neumann boundary conditions, or $m, \phi \in H^1_{\text{per}}(\Omega)$ under periodic boundary conditions. For a fixed sampling state, define

$$\underbrace{a = 1 + \lambda + \mu}_{>0}, \qquad \underbrace{b_t = r_t + \mu m_t^{\text{ref}}}_{\text{state-dependent forcing}}, \tag{A.9}$$

so that the objective in Eq. (A.8) can be written as

$$\mathcal{J}_t(m) = \frac{a}{2}\|m\|_2^2 + \frac{\gamma}{2}\|\nabla m\|_2^2 - \langle b_t, m\rangle + C_t, \tag{A.10}$$

where the optimization variable is a scalar spatial field, while r_t, m_t^{ref}, and hence b_t are fixed at the current control state. When $\gamma = 0$, no spatial derivative is required and the same objective can instead be posed directly over $L^2(\Omega)$.

First variation. For an arbitrary admissible perturbation ϕ, consider $m_\tau = m + \tau\phi$. For $\gamma > 0$, the directional derivative of the objective is

$$\begin{aligned} \delta\mathcal{J}_t(m;\phi) &= \frac{d}{d\tau}\mathcal{J}_t(m+\tau\phi)\Big|_{\tau=0} \\ &= \underbrace{\langle m - r_t, \phi\rangle}_{\text{fidelity}} + \underbrace{\lambda\langle m, \phi\rangle}_{\text{energy}} + \underbrace{\gamma\langle \nabla m, \nabla\phi\rangle}_{\text{spatial}} + \underbrace{\mu\langle m - m_t^{\text{ref}}, \phi\rangle}_{\text{reference}} \\ &= \langle am - b_t, \phi\rangle + \gamma\,\langle \nabla m, \nabla\phi\rangle\,, \end{aligned} \tag{A.11}$$

where each contribution follows directly from the corresponding quadratic term in Eq. (A.7). The spatial regularizer is the only term containing derivatives of the optimization variable and therefore generates the differential component of the optimality operator.

Integration by parts. For sufficiently regular m, the gradient term in Eq. (A.11) satisfies Green's identity,

$$\langle \nabla m, \nabla \phi \rangle = -\langle \Delta m, \phi \rangle + \int_{\partial\Omega} \phi \, \partial_n m \, dS, \tag{A.12}$$

where opposite boundary contributions cancel for periodic boundaries, while $\partial_n m = 0$ under homogeneous Neumann boundaries. In either case, the boundary contribution vanishes and, whenever Δm is defined with sufficient regularity, $\langle \nabla m, \nabla \phi \rangle = -\langle \Delta m, \phi \rangle$. Substitution into Eq. (A.11) then gives

$$\delta \mathcal{J}_t(m; \phi) = \left\langle \underbrace{am - \gamma \Delta m}_{\mathcal{A}m} - b_t, \phi \right\rangle, \tag{A.13}$$

so the spatial penalty $\|\nabla m\|_2^2$ induces the Laplacian contribution $-\gamma \Delta m$ in the strong optimality equation.

Weak optimality formulation. The optimality condition can be stated without requiring a classical second derivative of m. At the optimum $m_t^\star$, stationarity for every admissible ϕ requires

$$\underbrace{a\langle m_t^\star, \phi \rangle + \gamma \langle \nabla m_t^\star, \nabla \phi \rangle}_{\mathfrak{a}(m_t^\star, \phi)} = \underbrace{\langle b_t, \phi \rangle}_{\ell_t(\phi)}, \qquad \forall \phi \in H^1(\Omega), \tag{A.14}$$

with the corresponding periodic Sobolev space used for periodic boundaries. Here $\mathfrak{a} : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ denotes the bilinear form induced by the quadratic objective and $\ell_t : H^1(\Omega) \to \mathbb{R}$ denotes the forcing functional. Moreover,

$$\mathfrak{a}(m, m) = a\|m\|_2^2 + \gamma \|\nabla m\|_2^2 \geq a\|m\|_2^2, \tag{A.15}$$

so the positive zero-order coefficient removes the constant-mode degeneracy associated with the Laplacian. Equation (A.14) therefore provides the natural weak formulation of the magnitude optimality problem.

Stationarity and strong form. For a sufficiently regular solution, the weak condition admits the corresponding differential representation. Since $\delta \mathcal{J}_t(m_t^\star; \phi) = 0$ for every admissible ϕ, the weak optimality condition yields

$$\boxed{\underbrace{[(1 + \lambda + \mu)I - \gamma \Delta]}_{\mathcal{A}} m_t^\star = \underbrace{r_t + \mu m_t^{\mathrm{ref}}}_{b_t}}, \tag{A.16}$$

or equivalently $\mathcal{A} m_t^\star = b_t$. The zero-order term $(1 + \lambda + \mu)I$ regulates magnitude independently at each spatial location, while $-\gamma \Delta$ introduces coupling through neighboring spatial variations. Hence the pointwise magnitude problem obtained for $\gamma = 0$ becomes a spatially coupled differential problem for $\gamma > 0$.

Ellipticity of the optimality operator. For $\gamma > 0$, the differential operator $\mathcal{A} = aI - \gamma \Delta$ has principal part $-\gamma \Delta$. The corresponding principal symbol is

$$\underbrace{\sigma_2(\mathcal{A})(\xi)}_{\text{principal symbol}} = \gamma \|\xi\|_2^2 > 0, \qquad \xi \in \mathbb{R}^2 \setminus \{\mathbf{0}\}, \tag{A.17}$$

so $\mathcal{A}$ is a second-order elliptic operator for $\gamma > 0$. The positive zero-order coefficient $a = 1 + \lambda + \mu$ additionally shifts the operator spectrum away from zero. In the limiting case $\gamma = 0$, the differential component vanishes and Eq. (A.16) becomes the pointwise algebraic relation

$$m_t^\star = \frac{\underbrace{r_t}_{\text{current magnitude}} + \underbrace{\mu m_t^{\mathrm{ref}}}_{\text{reference contribution}}}{\underbrace{1 + \lambda + \mu}_{\text{uniform attenuation}}}, \tag{A.18}$$

so second-order ellipticity arises specifically from the spatial regularizer for $\gamma > 0$, whereas the remaining terms determine local magnitude attenuation and reference coupling.

Optimality-system interpretation. Combining Eqs. (A.14) and (A.16), the state-wise control problem admits the weak characterization

$$\underbrace{\mathfrak{a}(m_t^\star, \phi) = \ell_t(\phi) \quad \forall \phi \in H^1(\Omega)}_{\text{weak variational formulation}} \qquad \Longrightarrow \qquad \underbrace{\mathcal{A} m_t^\star = b_t}_{\text{strong differential form}} \quad , \tag{A.19}$$

whenever the solution has sufficient regularity for the strong form. Conversely, a sufficiently regular solution of the strong equation satisfying the stated boundary conditions satisfies the weak formulation. The weak form characterizes the variational optimum in the admissible function space, while the strong form exposes the screened elliptic operator used in the analytical derivation. This correspondence connects the regularized magnitude objective to the spectral solution derived in Sec. A.4.

A.3 Well-Posedness and Properties of the Optimal Magnitude

Continuity and coercivity. Recall the bilinear form $\mathfrak{a}(m, \phi) = a\langle m, \phi\rangle + \gamma\langle \nabla m, \nabla \phi\rangle$ from Eq. (A.14), where $a = 1 + \lambda + \mu > 0$. For $\gamma > 0$, the Cauchy–Schwarz inequality gives

$$|\mathfrak{a}(m, \phi)| \le \max\{a, \gamma\} \|m\|_{H^1} \|\phi\|_{H^1}, \tag{A.20}$$

so $\mathfrak{a}$ is continuous on $H^1(\Omega)$, or on $H^1_{\text{per}}(\Omega)$ under periodic boundary conditions. The corresponding quadratic form satisfies

$$\mathfrak{a}(m, m) = a\|m\|_2^2 + \gamma\|\nabla m\|_2^2 \ge \underbrace{\min\{a, \gamma\}}_{\alpha_{\mathfrak{a}} > 0} \|m\|_{H^1}^2, \tag{A.21}$$

which establishes coercivity for $\gamma > 0$. Equivalently, the quadratic form associated with $\mathcal{A} = aI - \gamma\Delta$ is strictly positive for every nonzero admissible field. When $\gamma = 0$, the problem is posed over $L^2(\Omega)$ and the operator reduces to the positive multiplication operator aI, with $\langle m, aIm\rangle = a\|m\|_2^2 > 0$ for $m \neq 0$.

Existence and uniqueness. The forcing functional $\ell_t(\phi) = \langle b_t, \phi\rangle$, with $b_t = r_t + \mu m_t^{\text{ref}}$, is bounded whenever $b_t \in L^2(\Omega)$. For $\gamma > 0$, continuity and coercivity of $\mathfrak{a}$ together with boundedness of ℓ_t allow the Lax–Milgram theorem to give a unique weak solution $m_t^\star \in H^1(\Omega)$, with the corresponding periodic space used under periodic boundary conditions. The optimization formulation gives the same conclusion because

$$\delta^2 \mathcal{J}_t(m)[h, h] = a\|h\|_2^2 + \gamma\|\nabla h\|_2^2 > 0, \qquad h \neq 0, \tag{A.22}$$

so $\mathcal{J}_t$ is strictly convex, while coercivity guarantees existence of a minimizer. The unique minimizer therefore coincides with the weak solution of Eq. (A.14) and, under sufficient regularity, with the strong solution of Eq. (A.16). For $\gamma = 0$, minimization over $L^2(\Omega)$ gives the unique solution $m_t^\star = b_t/a$ directly from $a > 0$.

Objective-gap characterization. The quadratic structure gives a quantitative characterization of the unique optimum. Let $e = m - m_t^\star$. Using the stationarity condition for $m_t^\star$, expansion of Eq. (A.8) cancels all terms linear in e and gives

$$\begin{aligned} \mathcal{J}_t(m) - \mathcal{J}_t(m_t^\star) &= \frac{1}{2}\langle e, \mathcal{A} e\rangle \\ &= \frac{a}{2}\|e\|_2^2 + \frac{\gamma}{2}\|\nabla e\|_2^2 \\ &\ge \frac{1 + \lambda + \mu}{2}\|m - m_t^\star\|_2^2, \end{aligned} \tag{A.23}$$

where the objective gap equals zero if and only if $m = m_t^\star$. The final lower-bound inequality becomes an equality whenever $\gamma\|\nabla e\|_2^2 = 0$, which can also occur for nonzero spatially constant perturbations under the stated boundary conditions. Thus every nonzero deviation from $m_t^\star$ incurs a strictly positive objective gap, while $(1 + \lambda + \mu)/2$ provides a frequency-independent lower bound on the quadratic growth around the optimum.

Non-negative forcing. By construction, $r_t \geq 0$. If the reference magnitude is initialized non-negatively, $m_t^{\text{ref}} \geq 0$, then $\mu \geq 0$ gives

$$\underbrace{b_t}_{\text{effective forcing}} = \underbrace{r_t}_{\geq 0} + \underbrace{\mu m_t^{\text{ref}}}_{\geq 0} \geq 0, \tag{A.24}$$

so every state-wise magnitude problem receives a non-negative forcing field. Since the reference update uses the preceding optimal magnitude, non-negativity propagates recursively once the initial reference is non-negative.

Non-negativity of the solution. For $\gamma > 0$, non-negativity can be established directly from the weak form without requiring pointwise second derivatives. Define the negative part $m_t^- = \max\{-m_t^\star, 0\} \in H^1(\Omega)$ and choose $\phi = m_t^-$ in Eq. (A.14). On the set where $m_t^\star < 0$, $m_t^\star = -m_t^-$ and $\nabla m_t^\star = -\nabla m_t^-$, which gives

$$-a\|m_t^-\|_2^2 - \gamma\|\nabla m_t^-\|_2^2 = \langle b_t, m_t^- \rangle \geq 0, \tag{A.25}$$

because $b_t \geq 0$ and $m_t^- \geq 0$. The left-hand side is non-positive and can equal a non-negative right-hand side only when

$$\|m_t^-\|_2^2 = 0, \qquad m_t^- = 0, \qquad m_t^\star \geq 0 \quad \text{a.e. in } \Omega, \tag{A.26}$$

which establishes positivity preservation at the level of the weak solution. For $\gamma = 0$, the same property follows immediately from $m_t^\star = b_t/a$ with $a > 0$ and $b_t \geq 0$. Consequently, for every active guidance location $p \in S_t = \{p : r_t(p) > 0\}$,

$$\widetilde{\mathbf{g}}_t^\star(p) = \underbrace{\frac{m_t^\star(p)}{r_t(p)}}_{\alpha_t^\star(p) \geq 0} \mathbf{g}_t(p), \tag{A.27}$$

so the optimized magnitude changes local guidance strength without reversing the model-predicted direction on the active support.

Stability with respect to the forcing. Consider two state-wise magnitude problems $\mathcal{A}m_t^{(1)} = b_t^{(1)}$ and $\mathcal{A}m_t^{(2)} = b_t^{(2)}$ under the same operator $\mathcal{A}$. Define $\delta m = m_t^{(1)} - m_t^{(2)}$ and $\delta b = b_t^{(1)} - b_t^{(2)}$. Subtracting the corresponding weak equations and choosing $\phi = \delta m$ gives

$$a\|\delta m\|_2^2 + \gamma\|\nabla \delta m\|_2^2 = \langle \delta b, \delta m \rangle \leq \|\delta b\|_2 \|\delta m\|_2, \tag{A.28}$$

where Cauchy–Schwarz was used in the final inequality. Dropping the non-negative gradient term and using $a = 1 + \lambda + \mu$ yields

$$\boxed{\|m_t^{(1)} - m_t^{(2)}\|_2 \leq \frac{\|r_t^{(1)} - r_t^{(2)}\|_2 + \mu\|m_{t,1}^{\text{ref}} - m_{t,2}^{\text{ref}}\|_2}{1 + \lambda + \mu}}, \tag{A.29}$$

where $b_t^{(i)} = r_t^{(i)} + \mu m_{t,i}^{\text{ref}}$. Hence the magnitude solution depends Lipschitz-continuously on the raw and reference magnitude inputs, with operator bound $\|\mathcal{A}^{-1}\|_{L^2 \to L^2} \leq (1 + \lambda + \mu)^{-1}$. This estimate concerns the optimized magnitude field only; no corresponding stability claim is implied for changes in the directional field $\mathbf{d}_t$, the complete controlled guidance $m_t^\star \odot \mathbf{d}_t$, or the nonlinear latent trajectory.

Scope of state-wise optimality. The results above establish optimality for the magnitude subproblem at a fixed latent state. Specifically, $m_t^\star$ is the unique global minimizer of Eq. (A.7) for the current r_t, $\mathbf{d}_t$, and m_t^{ref}. The result does not assert global optimality over the complete nonlinear sampling trajectory in Eq. (A.4). After each controlled update, the latent state changes and the pretrained model produces new magnitude and directional fields, so the subsequent state defines a new convex magnitude problem. CtrlGen therefore applies a sequence of state-wise optimal magnitude updates along the sampling trajectory rather than solving a single trajectory-level optimal control problem.

A.4 Spectral Characterization and Discrete Realization

Continuous spectral solution. The elliptic structure admits a direct frequency-domain characterization. Under periodic boundary conditions, $\mathcal{F}[-\Delta m](\omega) = \|\omega\|^2 \widehat{m}(\omega)$. Applying the spatial

Fourier transform to Eq. (A.16) gives $(1+\lambda+\mu+\gamma\|\omega\|^2)\widehat{m}_t^\star(\omega) = \widehat{r}_t(\omega) + \mu\widehat{m}_t^{\mathrm{ref}}(\omega)$. Since $1+\lambda+\mu+\gamma\|\omega\|^2 > 0$ for every frequency, each Fourier mode admits the unique solution

$$\boxed{\widehat{m}_t^\star(\omega) = \frac{\widehat{r}_t(\omega) + \mu\widehat{m}_t^{\mathrm{ref}}(\omega)}{1+\lambda+\mu+\gamma\|\omega\|^2}}, \tag{A.30}$$

which provides the continuous spectral representation of the unique state-wise magnitude optimum. The screened elliptic problem therefore reduces to independent mode-wise operations in the Fourier domain rather than iterative optimization over the complete spatial field. The denominator also shows directly that the zero-order term $1+\lambda+\mu$ shifts every Fourier mode away from zero, while the spatial term $\gamma\|\omega\|^2$ increasingly penalizes modes with larger spatial frequency.

Discrete spectral realization. The implementation operates on a finite spatial grid and uses the standard second-order periodic finite-difference Laplacian. Let $m_{i,j}$ denote the magnitude value on an $H \times W$ periodic grid with spacings h_x and h_y. For a two-dimensional discrete Fourier mode indexed by (k,ℓ), the corresponding eigenvalue of the positive semidefinite operator $L_h = -\Delta_h$ is

$$\underbrace{\Lambda_{k,\ell}}_{\text{discrete Laplacian spectrum}} = \frac{4}{h_x^2}\sin^2\left(\frac{\pi k}{H}\right) + \frac{4}{h_y^2}\sin^2\left(\frac{\pi \ell}{W}\right) \geq 0. \tag{A.31}$$

With $\mathcal{A}_h = (1+\lambda+\mu)I + \gamma L_h$, the discrete optimality equation $\mathcal{A}_h m_t^\star = b_t$ remains diagonal in the discrete Fourier basis. Replacing the continuous eigenvalue $\|\omega\|^2$ by $\Lambda_{k,\ell}$ therefore gives

$$\boxed{\widehat{m}_{t,k,\ell}^\star = \frac{\widehat{r}_{t,k,\ell} + \mu\widehat{m}_{t,k,\ell}^{\mathrm{ref}}}{1+\lambda+\mu+\gamma\Lambda_{k,\ell}}}. \tag{A.32}$$

Equation (A.32) is the discrete spectral update used in Algorithm 1. Because $1+\lambda+\mu+\gamma\Lambda_{k,\ell} > 0$ for every (k,ℓ), the discrete system has no spectral singularity. In particular, the constant mode satisfies $\Lambda_{0,0} = 0$ but retains the strictly positive coefficient $1+\lambda+\mu$. The finite-difference realization therefore preserves the direct FFT solution while remaining consistent with the discrete spatial regularizer.

Discrete positivity preservation. Positivity of the spectral denominator establishes invertibility but does not alone guarantee pointwise non-negativity after the inverse Fourier transform. For the standard nearest-neighbor finite-difference Laplacian above, non-negativity follows instead from the discrete maximum principle. Suppose $b_t \geq 0$ and a component $m_i^\star < 0$ is a global minimum. Using $L_h = D_h^\top D_h$ and the corresponding nearest-neighbor weights $w_{ij} \geq 0$, the minimum property gives

$$\begin{aligned}(L_h m_t^\star)_i &= \sum_{j\sim i} w_{ij}\left(m_i^\star - m_j^\star\right) \leq 0,\\ (\mathcal{A}_h m_t^\star)_i &= (1+\lambda+\mu)m_i^\star + \gamma(L_h m_t^\star)_i < 0.\end{aligned} \tag{A.33}$$

The second inequality contradicts $(\mathcal{A}_h m_t^\star)_i = b_{t,i} \geq 0$. Hence $m_t^\star \geq 0$ whenever $b_t = r_t + \mu m_t^{\mathrm{ref}} \geq 0$. The discrete FFT realization therefore preserves the non-negative magnitude property required by the direction-preserving admissible control space, provided that the spectral eigenvalues correspond to the finite-difference operator in Eq. (A.31).

Frequency response. For the continuous spectral representation, let $\rho = \|\omega\|$ and define $H_r(\rho) = (1+\lambda+\mu+\gamma\rho^2)^{-1}$. For $\gamma > 0$ and $\rho > 0$, the frequency response satisfies

$$\frac{dH_r(\rho)}{d\rho} = -\frac{2\gamma\rho}{(1+\lambda+\mu+\gamma\rho^2)^2} < 0, \qquad H_r(0) = \frac{1}{1+\lambda+\mu}, \qquad \lim_{\rho\to\infty} H_r(\rho) = 0. \tag{A.34}$$

The response therefore decreases monotonically with spatial frequency: low-frequency magnitude components retain a finite response, whereas higher-frequency components receive progressively stronger attenuation. For two frequencies $\rho_2 > \rho_1 \geq 0$, $H_r(\rho_2)/H_r(\rho_1) = (1+\lambda+\mu+\gamma\rho_1^2)/(1+\lambda+\mu+\gamma\rho_2^2) < 1$ when $\gamma > 0$. The discrete realization follows the same spectral interpretation with ρ^2 replaced by the non-negative eigenvalue $\Lambda_{k,\ell}$. The resulting frequency preference follows directly from the spatial regularizer $\gamma\|\nabla m\|_2^2$ rather than from an additional post-processing filter.

Parameter roles. The spectral representation also separates the contributions of the three regularization coefficients. For $H_r = (1+\lambda+\mu+\gamma\rho^2)^{-1}$, $\partial_\lambda H_r = -H_r^2 < 0$ and $\partial_\gamma H_r = -\rho^2 H_r^2 \leq 0$.

Increasing λ therefore attenuates all spatial modes through a uniform denominator shift, whereas increasing γ produces stronger attenuation as spatial frequency increases. The reference coefficient μ affects both the denominator and the reference contribution: $\widehat{m}_t^\star = H_r \widehat{r}_t + H_{\mathrm{ref}} \widehat{m}_t^{\mathrm{ref}}$, where $H_{\mathrm{ref}} = \mu H_r$. Thus λ regulates global magnitude shrinkage, γ regulates frequency-dependent spatial variation, and μ couples the current magnitude solution to the preceding sampling state through the same state-wise elliptic operator.

Limiting cases. Several limiting configurations clarify the structure of the magnitude controller. The continuous formulation gives

$$\begin{aligned} \gamma = 0: \quad & m_t^\star = \frac{r_t + \mu m_t^{\mathrm{ref}}}{1 + \lambda + \mu}, \\ \mu = 0: \quad & \widehat{m}_t^\star(\omega) = \frac{\widehat{r}_t(\omega)}{1 + \lambda + \gamma \|\omega\|^2}, \\ \gamma = \mu = 0: \quad & m_t^\star = \frac{r_t}{1 + \lambda}, \\ \lambda = \gamma = \mu = 0: \quad & m_t^\star = r_t, \qquad \widetilde{\mathbf{g}}_t^\star = r_t \odot \mathbf{d}_t = \mathbf{g}_t. \end{aligned} \tag{A.35}$$

The first case removes spatial coupling while retaining magnitude shrinkage and temporal reference coupling; the second removes the reference contribution while preserving spatial regularization; and the third retains only uniform magnitude shrinkage. The final case recovers the original guidance through the exact direction–magnitude decomposition on the active guidance support. Standard guidance is therefore contained as the unregularized limit of the proposed magnitude-control formulation.

State-dependent control. Define the continuous elliptic solution operator $\mathcal{K}_{\lambda,\gamma,\mu} = [(1 + \lambda + \mu)I - \gamma\Delta]^{-1}$. For a fixed latent state, substitution of the unique optimal magnitude into the controlled dynamics gives

$$\boxed{\mathbf{u}_t^\star(\boldsymbol{x}_t, m_t^{\mathrm{ref}}) = w\, \mathcal{K}_{\lambda,\gamma,\mu} \left[r(\boldsymbol{x}_t) + \mu m_t^{\mathrm{ref}} \right] \odot \mathbf{d}(\boldsymbol{x}_t)}. \tag{A.36}$$

Both $r(\boldsymbol{x}_t)$ and $\mathbf{d}(\boldsymbol{x}_t)$ are recomputed from the current latent state, while the discrete implementation replaces $\mathcal{K}_{\lambda,\gamma,\mu}$ by the finite-difference spectral operator represented by Eq. (A.32). After each controlled state transition, the next model evaluation produces new conditional and unconditional velocity fields and therefore new magnitude and directional fields. CtrlGen consequently solves a unique convex magnitude problem at every sampling state and applies the corresponding direction-preserving control, rather than solving a trajectory-level optimal control problem.

B Additional Experimental Analysis

B.1 Magnitude Regulation and Sampling Dynamics

Figure S1 provides a complementary analysis of the behavior of CtrlGen from parameter, frequency, and trajectory perspectives. The results are reported on FLUX.1-dev unless otherwise specified.

Effect of spatial regularization. Figure S1(a) shows the effect of γ on generation quality and preference. Moderate spatial regularization improves both criteria, while an overly large value reduces the benefit. According to Eq. (A.7), γ penalizes spatial variation of the magnitude field through $\|\nabla m\|_F^2$. The spectral response in Eq. (A.34) further shows that this penalty increases the attenuation of high-frequency components. The observed trend therefore reflects a balance between suppressing unstable local variation and retaining useful fine-scale guidance.

Frequency-domain behavior. Figures S1(b)–(c) examine the spectral characteristics of guidance magnitude. CtrlGen suppresses high-frequency components more strongly while retaining a larger fraction of the low-frequency response. This behavior agrees with the spectral solution in Eq. (A.30), where $H_r(\rho) = (1 + \lambda + \mu + \gamma\rho^2)^{-1}$ with $\rho = \|\omega\|$, and $\partial H_r(\rho)/\partial\rho < 0$ in Eq. (A.34). In Figure S1(c), γ parameterizes a shared spectral probe applied to the fixed magnitude outputs of all methods; it is not an inference hyperparameter of CFG or APG. Their variation therefore reflects sensitivity to the same frequency response rather than changes in their configurations. CtrlGen maintains the lowest high-frequency energy ratio across the sweep, consistent with the spectral attenuation observed in Figure S1(b).

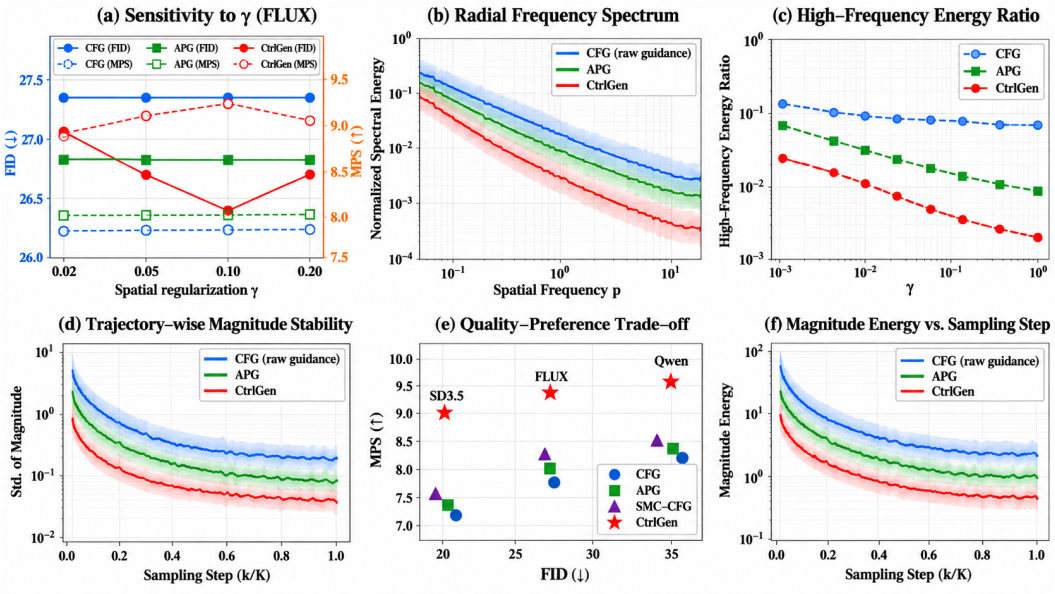

Figure S1: **Additional analysis of magnitude control.** (a) Sensitivity to the spatial regularization weight γ. (b) Radially averaged spectral energy of the raw and controlled guidance magnitudes. (c) High-frequency energy ratio under a shared spectral probe parameterized by γ. (d) Spatial magnitude variation along the sampling trajectory. (e) FID–MPS trade-off across guidance methods and T2I backbones. (f) Magnitude energy over sampling steps.

Trajectory-wise magnitude stability. Figures S1(d) and S1(f) analyze the magnitude field along the sampling trajectory. Raw guidance exhibits larger spatial variation and higher magnitude energy, especially at early sampling steps. CtrlGen reduces both quantities and yields a more stable evolution across subsequent states. This behavior is consistent with the state-dependent controller in Eq. (A.36), $\mathbf{u}_t = w\,\mathcal{K}_{\lambda,\gamma,\mu}[r(\boldsymbol{x}_t) + \mu m_t^{\mathrm{ref}}] \odot \mathbf{d}(\boldsymbol{x}_t)$. Both $r(\boldsymbol{x}_t)$ and $\mathbf{d}(\boldsymbol{x}_t)$ are recomputed after each state update. The observed trajectories therefore reflect repeated state-wise control rather than a fixed global rescaling of the guidance field throughout the evolving flow-based generation process.

Quality–preference trade-off. Figure S1(e) compares FID and MPS across guidance methods and model backbones. CtrlGen achieves a favorable quality–preference trade-off, with higher preference scores while maintaining competitive distributional quality. This result is consistent with the constrained guidance form $\widetilde{\mathbf{g}}_t = m_t^\star \odot \mathbf{d}_t$, where the controller modifies the local magnitude $m_t^\star$ while retaining the model-predicted direction $\mathbf{d}_t$. Together with the main quantitative results, the comparison indicates that the preference gains do not arise from a simple loss of distributional fidelity, but remain compatible with the original generative distribution across backbones.

B.2 Controlled Spatial–State Consistency Ablation

To distinguish spatial and state-consistency magnitude regulation from global magnitude attenuation, we conduct a controlled 2×2 ablation over γ and μ, while fixing the energy coefficient to $\lambda = 5$. This design preserves the same global magnitude regularization across all configurations and isolates the additional effects introduced by spatial regularization and state-consistency coupling.

When $\gamma = \mu = 0$, Eq. (A.16) reduces to $m_t^\star = r_t/(1+\lambda)$, and therefore $\widetilde{\mathbf{g}}_t^\star = \mathbf{g}_t/(1+\lambda)$. For a fixed CFG scale w, the magnitude-only configuration is thus equivalent to standard CFG with

$$\boxed{w_{\mathrm{eff}} = \frac{w}{1+\lambda}}, \tag{B.1}$$

providing a direct control for determining whether the observed gains can be explained by uniform guidance attenuation alone.

Table S2: **Controlled spatial–state consistency ablation of CtrlGen on FLUX.** The energy coefficient is fixed to $\lambda = 5$ in all configurations. The comparison separates uniform magnitude attenuation from spatial regularization and state-consistency coupling.

Variant	γ	μ	FID↓	CLIP↑	HPSv2.1↑	MPS↑
Magnitude only	0	0	27.218	0.3725	0.2965	8.5537
Spatial only	0.1	0	26.926	0.3729	0.2976	8.8673
State only	0	0.5	27.403	0.3738	0.2991	8.7526
Full CtrlGen	**0.1**	**0.5**	**26.338**	**0.3761**	**0.3046**	**9.3475**

Starting from this magnitude-only control, $\gamma > 0, \mu = 0$ introduces spatial regularization through the discrete elliptic operator, whereas $\gamma = 0, \mu > 0$ introduces coupling to the preceding control state m_i^{ref} without spatial interaction. The full configuration activates both mechanisms. Table S2 compares these four configurations under the same λ.

Controlled component analysis. Table S2 separates spatial regularization and state-consistency coupling while keeping the energy coefficient fixed. The magnitude-only configuration retains the uniform attenuation induced by λ but removes both additional mechanisms. Adding the spatial term introduces frequency-dependent regulation of the magnitude field, whereas the state-consistency term couples the current solution to the preceding control state. The complete configuration combines both mechanisms and provides the strongest overall performance across the reported metrics. These results indicate that the observed improvement cannot be attributed solely to uniform magnitude attenuation and instead benefits from joint spatial regularization and state-consistency coupling of the guidance magnitude.

B.3 Statistical Reliability Analysis

Paired evaluation protocol. To examine whether the performance gains of CtrlGen are robust to sample-level variation, we conduct a paired statistical analysis against SMC-CFG, the strongest competing guidance method in the main comparison. Both methods are evaluated using identical prompts, initial noise seeds, sampling configurations, and pretrained backbones, so that each evaluation sample forms a matched pair differing only in the guidance method. For a higher-is-better sample-wise metric s, let s_i^{CtrlGen} and s_i^{SMC} denote the scores obtained for the i-th image-text pair. The sample-wise paired difference is defined as

$$d_i = s_i^{\text{CtrlGen}} - s_i^{\text{SMC}}, \qquad i = 1, \dots, N, \tag{B.2}$$

where N denotes the number of evaluation samples. The mean paired difference is

$$\Delta s = \frac{1}{N} \sum_{i=1}^{N} d_i = \frac{1}{N} \sum_{i=1}^{N} \left(s_i^{\text{CtrlGen}} - s_i^{\text{SMC}} \right), \tag{B.3}$$

where $\Delta s > 0$ indicates a higher average score for CtrlGen under the corresponding metric.

Paired bootstrap confidence interval. To quantify the uncertainty of the observed mean difference without assuming a parametric distribution for the paired scores, we perform paired bootstrap resampling over the evaluation samples. For each bootstrap replicate b, we draw N indices with replacement from $\{1, \dots, N\}$ and apply the same resampled indices to both methods. Let $\mathcal{I}^{(b)} = \{i_1^{(b)}, \dots, i_N^{(b)}\}$ denote the resampled indices. The corresponding bootstrap estimate is

$$\Delta s^{(b)} = \frac{1}{N} \sum_{j=1}^{N} d_{i_j^{(b)}}, \qquad b = 1, \dots, B, \tag{B.4}$$

where $B = 10{,}000$ bootstrap replicates are used. The 95% percentile bootstrap confidence interval is computed as

$$\text{CI}_{95\%} = \left[Q_{0.025} \left(\{\Delta s^{(b)}\}_{b=1}^{B} \right), Q_{0.975} \left(\{\Delta s^{(b)}\}_{b=1}^{B} \right) \right], \tag{B.5}$$

where Q_α denotes the empirical α-quantile. Using matched resampling preserves the prompt- and seed-level correspondence between the two methods and controls for variation shared within each

Table S3: **Statistical reliability of CtrlGen over SMC-CFG.** Paired score differences, 95% bootstrap confidence intervals, and sample-wise win rates are reported using identical prompts and initial noise seeds.

Model	Metric	Δ Score	95% CI	Win Rate
SD3.5	CLIP	+0.0018	[+0.0011, +0.0025]	58.4%
	HPSv2.1	+0.0103	[+0.0084, +0.0122]	65.7%
	MPS	+1.4428	[+1.31, +1.57]	72.6%
FLUX.1-dev	CLIP	+0.0018	[+0.0010, +0.0026]	57.9%
	HPSv2.1	+0.0025	[+0.0012, +0.0038]	61.3%
	MPS	+1.1168	[+0.99, +1.24]	70.8%

Table S4: **Quantitative evaluation on T2I-CompBench.** We report compositional alignment on color binding, shape binding, texture binding, and spatial relationships across three flow-based text-to-image models. Higher values indicate better performance.

Model / Method	Color ↑	Shape ↑	Texture ↑	Spatial ↑
SD3.5L	0.6790	0.5915	0.7243	0.1625
w/ SMC-CFG	0.7461	0.6009	0.7406	0.2563
w/ CtrlGen	**0.7528**	**0.6097**	**0.7483**	**0.2716**
FLUX.1-dev	0.8172	0.5751	0.7432	0.2708
w/ SMC-CFG	0.8216	0.6199	0.7901	0.2939
w/ CtrlGen	**0.8274**	**0.6308**	**0.7996**	**0.3127**
Qwen-Image	0.7747	0.5621	0.6747	0.2968
w/ SMC-CFG	**0.8191**	0.5934	**0.7421**	0.4085
w/ CtrlGen	0.8146	**0.6028**	0.7387	**0.4219**

matched pair. A confidence interval lying entirely above zero provides evidence that the positive mean paired difference is robust to resampling of the evaluated samples.

Sample-wise win rate. The mean paired difference measures the magnitude of the aggregate improvement but does not characterize how broadly the improvement is distributed across individual samples. We therefore additionally report the sample-wise win rate for each higher-is-better metric,

$$\mathrm{WinRate}(s) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[s_i^{\mathrm{CtrlGen}} > s_i^{\mathrm{SMC}}\right] \times 100\%, \tag{B.6}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. A win is counted when CtrlGen obtains a higher score than SMC-CFG for the same prompt and initial noise seed. The win rate therefore complements the mean paired difference by quantifying the fraction of evaluation samples for which CtrlGen achieves the higher score. Together, Δs, the paired bootstrap confidence interval, and the win rate characterize the magnitude, statistical reliability, and sample-level prevalence of the observed performance differences.

B.4 Evaluation on T2I-CompBench

To further evaluate compositional generalization, we conduct experiments on T2I-CompBench using the same three flow-based text-to-image backbones as in the main experiments: SD3.5L, FLUX.1-dev, and Qwen-Image. Following the baseline evaluation protocol, we report four representative compositional dimensions: color binding, shape binding, texture binding, and spatial relationships. All methods use the same model checkpoints, prompts, sampling settings, and evaluation protocol for each backbone.

Compositional Generalization. Table S4 evaluates CtrlGen on compositional text-to-image generation beyond the MS-COCO evaluation in the main paper. CtrlGen improves compositional alignment in most evaluated settings, with particularly clear gains in shape binding and spatial relationships.

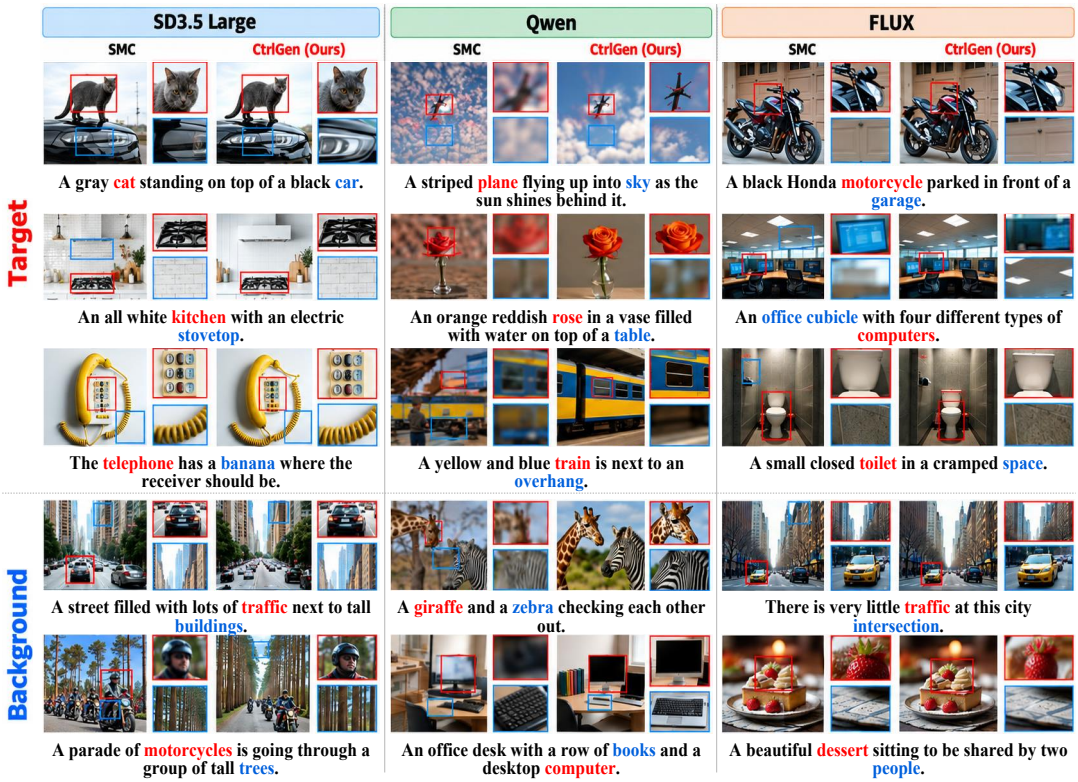

Figure S2: **Target–background comparison of SMC-CFG and CtrlGen.** Red and blue boxes highlight target and background regions, respectively, with enlarged crops. Across different flow-based models, CtrlGen generally improves target fidelity while preserving clearer contextual structure.

Across SD3.5L and FLUX.1-dev, the improvements cover both attribute-level binding and spatial composition. This trend is consistent with the spatial regularization of CtrlGen, which suppresses irregular local magnitude variations while preserving the model-predicted guidance direction. The larger gains on shape and spatial relationships suggest that such magnitude regulation is particularly useful when generation requires coherent object structure or relative spatial organization. On Qwen-Image, CtrlGen improves shape and spatial alignment, while color and texture binding remain slightly below SMC-CFG. One possible explanation is that color and texture depend more strongly on fine-grained local attributes, which may not uniformly benefit from spatial magnitude regularization. These results suggest that CtrlGen generalizes to compositional generation, although its benefits vary across compositional dimensions and model backbones.

B.5 Target–Background Analysis

Figure S2 compares SMC-CFG and CtrlGen at the region level on SD3.5 Large, Qwen-Image, and FLUX. Red and blue regions highlight prompt-relevant targets and surrounding backgrounds, respectively, complementing the global metrics with localized comparisons.

Target fidelity. CtrlGen generally produces more recognizable and prompt-consistent target content, with clearer fine structures across different object categories. In contrast, SMC-CFG occasionally yields blurred or incomplete target regions despite plausible global appearance. This behavior is consistent with Eq. (A.36), where CtrlGen optimizes the local magnitude $m_t^\star(p)$ while preserving the model-predicted direction $\mathbf{d}_t(p)$.

Background preservation. CtrlGen also tends to preserve clearer contextual structures and local geometry in the background regions. This agrees with the spatial regularization $\frac{\gamma}{2}\|\nabla m\|_F^2$ in Eq. (A.7), whose elliptic condition couples neighboring magnitude values through $-\gamma\Delta m_t^\star$ without

spatially smoothing the directional field $\mathbf{d}_t$. This separation suppresses excessive local guidance while preserving structural direction.

Target–background balance. The improvement is not uniform across all regions: SMC-CFG remains competitive in some smooth backgrounds, and CtrlGen does not improve every local crop. This is consistent with the frequency response $H_r(\rho) = (1 + \lambda + \mu + \gamma\rho^2)^{-1}$ in Eq. (A.34), which increasingly attenuates high-frequency magnitude variations. Rather than uniformly sharpening all regions, CtrlGen promotes a more stable balance between target fidelity and background structure, with consistent behavior across the three model families.

B.6 Additional Qualitative Results

Figure S3 provides additional qualitative comparisons between SMC-CFG and CtrlGen across SD3.5 Large, Qwen-Image, and FLUX, covering diverse objects, attributes, and scene structures.

SD3.5 Large. On SD3.5 Large, the main differences appear in local object structure and attribute realization. CtrlGen produces clearer prompt-relevant details, such as the cat body, kitchen layout, telephone components, and traffic elements, while preserving the overall scene appearance. This behavior is consistent with the direction-preserving control $\widetilde{\mathbf{g}}_t = m_t^\star \odot \mathbf{d}_t$, where only the local magnitude $m_t^\star$ is regulated while the model-predicted direction $\mathbf{d}_t$ is retained. SMC-CFG preserves plausible global composition, but some local attributes are less distinct or spatially consistent.

Qwen-Image. Qwen-Image shows stronger differences in object identity and multi-object relations. CtrlGen improves the visibility and separation of prompt-relevant entities, including the aircraft, rose, train, and animal pair, while reducing ambiguity in fine structures. This agrees with the state-dependent controller in Eq. (A.36), where the magnitude $r(\boldsymbol{x}_t)$ and direction $\mathbf{d}(\boldsymbol{x}_t)$ are recomputed after each state update rather than fixed over the sampling trajectory. SMC-CFG produces plausible scenes, but some prompt-specific entities are less clearly resolved.

FLUX. For FLUX, the difference is more apparent in geometric consistency and background organization. CtrlGen preserves clearer object boundaries and more coherent scene structure in the motorcycle, office, restroom, street, and dessert examples. This observation is consistent with the elliptic optimality condition in Eq. (A.16), where the spatial term $-\gamma\Delta m_t^\star$ couples neighboring magnitude responses and suppresses abrupt local variation. SMC-CFG remains competitive in overall realism, but some local regions exhibit weaker structural consistency or prompt correspondence.

C Limitations and Future Work

Limitations. CtrlGen regulates guidance magnitude while preserving the model-predicted direction, but several limitations remain. First, the method assumes that the original guidance direction is sufficiently reliable. Since CtrlGen optimizes only the magnitude field, it cannot correct an inaccurate direction produced by the underlying model. Second, the coefficients λ, γ, and μ are fixed for each backbone. This avoids sample-specific optimization, but a fixed regularization strength may not be optimal for scenes with different spatial structures or fine-grained attributes. Strong spatial regularization may also suppress useful local magnitude variations. Finally, our optimality result is state-wise rather than trajectory-level. The screened elliptic system has a unique optimum at each sampling state, but this result does not guarantee global optimality or stability of the full nonlinear generative trajectory.

Future Work. Future work can address these limitations in several directions. One extension is to control both guidance direction and magnitude without sacrificing the efficiency of the current formulation. Another is to use state-adaptive coefficients based on magnitude statistics or spatial structure, so that the control strength can vary across generation states. A trajectory-level objective could also account for the complete sampling dynamics and provide stronger theoretical guarantees. These extensions could improve the flexibility of CtrlGen while preserving its training-free and efficient inference-time design.

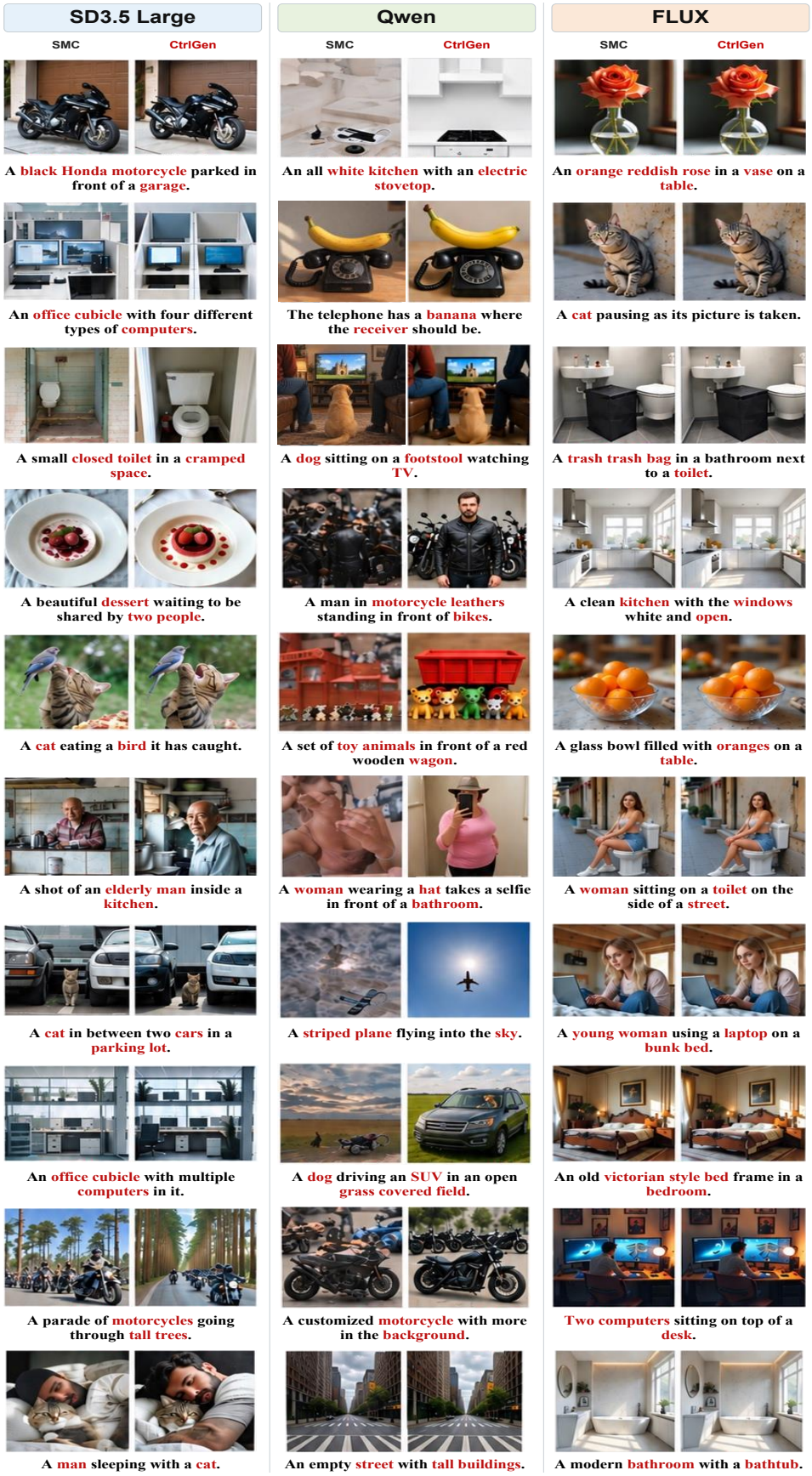

Figure S3: **Additional qualitative comparison across different flow-based text-to-image models.**