CONDITIONAL GEOMETRIC FEEDBACK CONTROL FOR TRAINING-FREE IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-guided image editing requires realizing target changes while preserving non-target content in the source image. Existing work uses latent-space geometry to characterize generative representations and editing directions, but how to translate this geometry into preservation feedback adapted to current editing conditions remains less explored. We propose Conditional Geometric Feedback Control (CGFC), a training-free and inversion-free image editing method. We construct local Fisher geometry from the source–target condition-label posterior to characterize the sensitivity of condition discrimination to latent corrections. Under Gaussian linear interpolation and population-optimal conditional velocities, its nonzero sensitive axis aligns with the same-state conditional velocity difference, motivating the model-predicted difference as a practical proxy for the control axis. Using this shared axis, one CGFC update controls both the direction of the current prediction residual at fixed norm within dynamic spatial support and the accumulated source displacement orthogonal to the axis. On PIE-Bench with SD3, CGFC improves content preservation and round-trip consistency. These results support using condition-dependent latent geometry to design feedback control for step-wise image editing.

1 INTRODUCTION

Diffusion models reverse progressive noising (Ho et al., 2020), while flow matching learns continuous velocity fields (Lipman et al., 2023). Their latent space variants enable high-resolution text-to-image generation (Rombach et al., 2022; Liu et al., 2023; Esser et al., 2024). Text-guided editing reuses these dynamics to realize a target change while preserving the source background, identity, and other non-target content; unintended content drift remains common.

Earlier methods constrain different stages: SDEdit corrupts and denoises the input, Prompt-to-Prompt retains layout through cross-attention, and DiffEdit locates edit regions from conditional prediction differences (Meng et al., 2022; Hertz et al., 2023; Couairon et al., 2023). Null-text Inversion improves real-image reconstruction by optimizing the unconditional embedding (Mokady et al., 2023). These input, attention, spatial, and reconstruction constraints have been followed by direct control of the latent editing trajectory.

FlowEdit constructs an inversion-free ODE between source and target conditional distributions (Kulikov et al., 2025). FlowAlign adds stepwise feedback for source consistency (Kim et al., 2026); Prism adapts spatial guidance (Jeong et al., 2026); SAM-Flow pairs edit localization with source-trajectory anchoring (Cui et al., 2026); and FlowDC decomposes velocities into time-varying subspaces (Jiang et al., 2025). These methods regulate different aspects of a latent path, but specifying where or how strongly to intervene does not determine how a correction affects the current edit. Non-target content may still drift as updates accumulate.

Latent geometry makes this gap precise: generative representations are anisotropic and semantically structured (Wang & Ponce, 2021; Arvanitidis et al., 2022; Park et al., 2023; Chen et al., 2024). The local KL divergence of the posterior over the source and target condition labels assigns unequal costs to equal-norm corrections in distinct directions. Masks set spatial support; gains set strength; neither chooses direction. This motivates a condition-pair criterion for immediate correction and accumulated displacement, complementing prior decomposition and anchoring.

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.

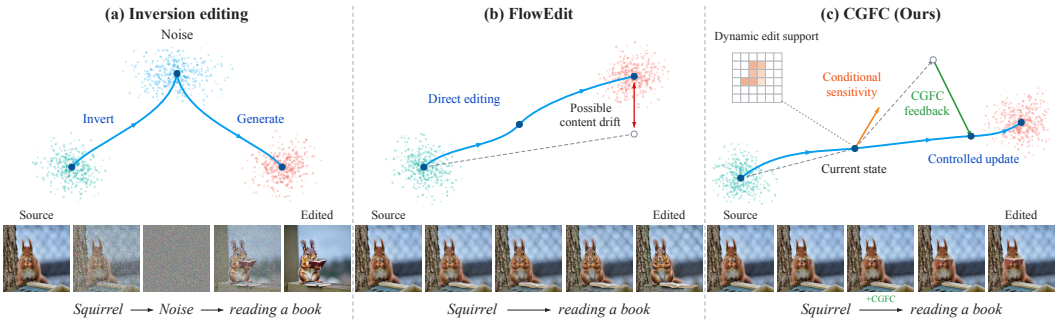

Figure 1: **Editing paths and CGFC.** (a) Inversion traverses noise; (b) FlowEdit can drift; (c) CGFC adjusts the residual and limits source displacement orthogonal to the condition axis. Bottom: illustrative process frames for a squirrel-to-book edit, without matched steps or budgets. Gray, green, and blue denote reference, correction, and controlled updates.

We probe drift in content coordinates using analytical Gaussian models with explicit edit and content coordinates. A spatial reference uses raw residual feedback without displacement feedback; geometric feedback adjusts the residual direction and accumulated source displacement. Both use the same spatial rule and three velocity evaluations per step. The spatial reference exhibits drift in content coordinates after matched perturbations and under state-dependent conditional directions. The comparison tests the joint effect of these two feedback settings on content drift.

We propose **Conditional Geometric Feedback Control (CGFC)**. The posterior over condition labels yields a local Fisher metric of rank at most one, whose nonzero eigenvector defines a condition-sensitive axis. Under Gaussian linear interpolation and population-optimal conditional velocities, it aligns with the same-state velocity difference, which the frozen model estimates. CGFC uses a single parameterized update to adjust the supported preservation residual at fixed norm and bound the source-relative displacement orthogonal to the condition axis. Because $r = -u + \sigma\Delta v$ can be small even when source displacement u is large, the update acts on both quantities. Both terms reuse existing predictions, requiring no training, inversion, or extra denoiser calls (Figure 1).

This view places existing methods along a common trajectory: direct transport supplies the path, while feedback, spatial guidance, and velocity decomposition regulate it. CGFC adds a condition-dependent criterion for preservation feedback. In analytical tests, mean drift in content coordinates falls from 0.244 to 0.164 in perturbation recovery and from 0.224 to 0.190 in state-dependent editing (Figures 2 and 3). Across 700 PIE-Bench tasks with SD3 Medium, full-image PSNR relative to the source rises from Prism's 27.381 to 28.490 dB and LPIPS falls from 0.06691 to 0.05765; round-trip reconstruction and fixed-parameter SD3.5 Medium/Large preservation also improve. Full-image target CLIP decreases from 24.569 to 24.335, reflecting the preservation–editing trade-off.

Our contributions are:

1. **Condition-pair geometry for feedback.** We derive the latent-input Fisher metric of the posterior over condition labels and, under Gaussian interpolation with population-optimal velocities, relate its sensitive axis to the same-state conditional velocity difference approximated by a frozen model.

2. **Joint feedback with local analysis.** A single CGFC control law adjusts the supported residual direction at fixed norm and controls accumulated source displacement through bounded feedback orthogonal to the condition axis. Its directional and displacement gains can each be set to zero to disable the corresponding adjustment. We characterize both operations at the current state under stated assumptions.

3. **Tests of drift and preservation.** Analytical Gaussian trajectories expose content drift; comparisons on 700 PIE-Bench tasks using SD3 and fixed-parameter SD3.5 transfer evaluations assess source preservation, round-trip consistency, target alignment, and their trade-off.

2 Related Work

From inversion to inversion-free editing. Inversion-based methods connect an image to a generative trajectory before changing its condition. Null-text Inversion optimizes unconditional embeddings; RF-Inversion uses dynamic optimal control; UniEdit-Flow couples predictor–corrector inversion with region-aware sampling (Mokady et al., 2023; Rout et al., 2025; Jiao et al., 2026). Inversion-free methods avoid explicitly estimating an input-specific noise endpoint: SDEdit denoises a corrupted input, InfEdit uses virtual inversion and consistency sampling, and FlowEdit constructs direct transport between conditional distributions (Meng et al., 2022; Xu et al., 2024; Kulikov et al., 2025). CGFC studies how to organize preservation feedback along such a direct editing path with frozen model weights.

Content preservation and spatial support. Attention control and reuse of source features provide content constraints in Prompt-to-Prompt, Plug-and-Play Diffusion Features, and MasaCtrl (Hertz et al., 2023; Tumanyan et al., 2023; Cao et al., 2023). DiffEdit estimates masks from conditional prediction differences; Prism derives dynamic guidance support from their magnitude hierarchy; SAM-Flow uses edit localization, differential flow, and source anchoring (Couairon et al., 2023; Jeong et al., 2026; Cui et al., 2026). A mask selects locations for conditional guidance or adjustment of the residual direction but does not determine the direction within those locations. CGFC uses the same selected locations for conditional guidance, the residual projection, and local norm restoration.

Latent geometry and semantic editing. Riemannian analyses characterize anisotropy and semantic directions in generative latent spaces (Wang & Ponce, 2021); Fisher–Rao pullbacks induced by the decoder distribution provide a statistical construction of latent geometry (Arvanitidis et al., 2022). Encoder pullback bases and the low-rank Jacobian structure of LOCO Edit support direction discovery and localized diffusion editing (Park et al., 2023; Chen et al., 2024). RemEdit learns a manifold module for editing paths, while Spacetime studies Fisher–Rao geometry over state and time (Adhikarla & Davison, 2026; Karczewski et al., 2026). CGFC instead uses the posterior over condition labels to define task-dependent local sensitivity. This geometry guides preservation feedback without requiring a metric for the entire image manifold.

Editing dynamics and conditional geometry. FlowAlign derives inversion-free updates from an optimal control objective with terminal source similarity; FlowDC reorganizes editing velocities in time-varying subspaces (Kim et al., 2026; Jiang et al., 2025). Prism relates expected squared guidance magnitude to Fisher divergence between conditional scores (Jeong et al., 2026). Our posterior input Fisher is a different object: it supplies a local directional criterion for feedback. CGFC uses this geometry to set the direction of residual feedback and constrain accumulated source displacement within one parameterized control law.

3 Method

We derive a conditional axis from the local geometry of the source–target condition-label posterior and use it to design feedback on the preservation residual and accumulated source displacement (Figure 1(c)).

3.1 CGFC geometric control law

Given an image I_s and conditions c_s, c_t, we seek a target-consistent edit preserving non-target content. Fix encoder $\mathcal{E}$, decoder $\mathcal{D}$, and velocity model v_θ, with $z_s = \mathcal{E}(I_s) \in \mathbb{R}^D$. Starting at $x_0 = z_s$, follow $\sigma_0 > \cdots > \sigma_N = 0$ and return $\mathcal{D}(x_N)$. Write $\Delta\sigma_k = \sigma_{k+1} - \sigma_k < 0$, omitting k within each step.

CGFC controls the edit through a source-relative state. At noise level σ, shared reference noise $\epsilon \sim \mathcal{N}(0, I)$ defines the noisy source, corresponding edit state, and accumulated source displacement (Kulikov et al., 2025):

$$q = (1-\sigma)z_s + \sigma\epsilon, \qquad p = x + q - z_s, \qquad u = x - z_s. \tag{1}$$

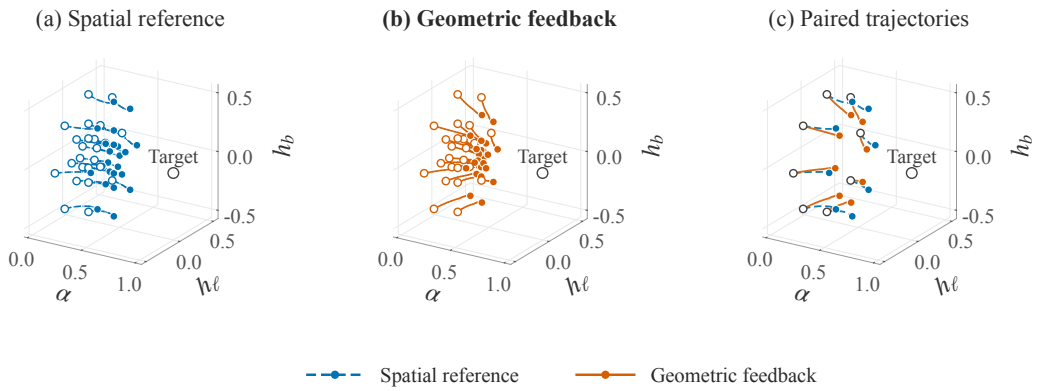

Figure 2: **Recovery from matched perturbations.** (a–b) All 24 conditions; (c) eight pairs at $\delta = 0.5$. Blue dashed: spatial reference; orange solid: geometric feedback. Open/filled circles mark starts/endpoints; gray marks targets. Projections are fixed; errors are full-dimensional means.

Thus q is the noisy source reference, p is the corresponding edit state, and u records displacement from z_s; reference noise may be resampled between steps. Three predictions determine the conditional response at p and the source velocity at q:

$$v_t = v_\theta(p, c_t, \sigma), \quad v_s = v_\theta(p, c_s, \sigma), \quad v_q = v_\theta(q, c_s, \sigma), \quad \delta v = v_t - v_s. \tag{2}$$

The same-state response δv supplies the edit-sensitive direction. CGFC converts its spatial channel magnitudes and the task-protocol edit type τ into a dynamic binary gate $M = M^\top = M^2$. Broadcast over latent channels, M couples conditional transport to geometric feedback; Appendix C.1 specifies its empirical rule and conditional-sensitivity interpretation. The gated target velocity and its difference from source-reference velocity define the transport signal:

$$\bar{v} = v_s + \gamma M \delta v, \qquad \Delta v = \bar{v} - v_q, \tag{3}$$

where $\gamma > 0$. Here δv compares conditions at the *same state*; Δv compares guided target and source velocities at *different states*. The discrepancy between clean-state predictions for the source reference and current edit defines the instantaneous preservation error (Kim et al., 2026):

$$r = (q - \sigma v_q) - (p - \sigma \bar{v}) = -u + \sigma \Delta v. \tag{4}$$

Because $r = -u + \sigma\Delta v$, a small instantaneous error can coexist with a large accumulated displacement when the two terms cancel. CGFC therefore uses one conditional axis d for both signals: A_M, controlled by β, changes the direction of r within M; $a = -\eta Q_d u$, controlled by λ, acts on source displacement. With $\widetilde{r} = A_M r$ and discrete coefficient $\zeta > 0$, the geometric control update is

$$\boxed{x^+ = x + \Delta\sigma\Delta v + \zeta\widetilde{r} + a = x + B\Delta v - K(x - z_s)}, \tag{5}$$

$$B = \Delta\sigma I + \zeta\sigma A_M, \qquad K = \zeta A_M + \eta Q_d. \tag{6}$$

Equation (5) is CGFC's state-dependent geometric control law. Its residual-direction gain is $1 - \beta$: at $\beta = 1$, $A_M = I$; at $\lambda = 0$, $a = 0$. The same M enters transport and residual geometry, while A_M contributes to both B and K. Because A_M and η depend on the current predictions and state, B and K are state-dependent maps, and the second form is an algebraic regrouping rather than a fixed linear controller. The positive $\zeta\sigma A_M$ term can offset part of the negative-step transport coefficient, linking this feedback to the measured preservation–editing trade-off. Conditional sensitivity supplies the axis used to design both feedback terms.

3.2 Conditional posterior Fisher and the control axis

To assess how a correction changes source–target discrimination, fix $0 < \sigma < 1$ and conditional marginals $p_\sigma(z \mid c_s), p_\sigma(z \mid c_t)$. Introduce a label $Y \in \{s, t\}$ with fixed positive priors $\pi_s + \pi_t = 1$.

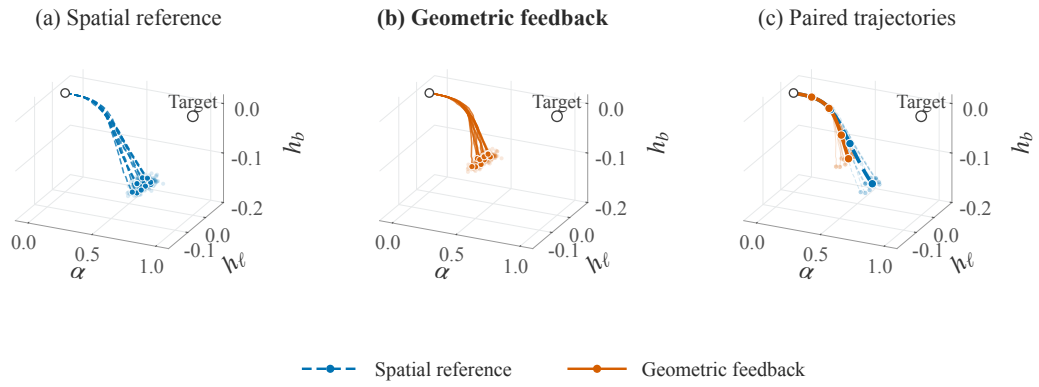

Figure 3: **State-dependent editing.** (a–b) Twelve fixed paths and all 256 endpoints. (c) Paired paths (thin) and the mean of all 256 source-aligned projections (thick). Blue dashed: spatial reference; orange solid: geometric feedback. Errors use the full latent space.

Its posterior is

$$\ell(z) = \log \frac{\pi_t p_\sigma(z \mid c_t)}{\pi_s p_\sigma(z \mid c_s)}, \qquad \omega(z) = \Pr(Y = t \mid z) = \mathrm{sigmoid}(\ell(z)). \tag{7}$$

Writing $g = \nabla_z \ell(z)$, the posterior Fisher information with respect to the latent input is

$$F(z) = \mathbb{E}_{Y|z}\left[\nabla_z \log \Pr(Y \mid z) \nabla_z \log \Pr(Y \mid z)^\top\right] = \omega(1-\omega) g g^\top. \tag{8}$$

For a small perturbation h,

$$D_{\mathrm{KL}}(\Pr(Y \mid z) \,\|\, \Pr(Y \mid z+h)) = \tfrac{1}{2} h^\top F(z) h + o(\|h\|^2). \tag{9}$$

This rank-at-most-one, positive-semidefinite structure measures local sensitivity of condition discrimination. It is the pullback of the label posterior's Fisher–Rao geometry to the latent input (Arvanitidis et al., 2022).

Proposition 1 (Axis correspondence under an ideal conditional flow)**.** *Assume $Z_\sigma = (1-\sigma)Z_0 + \sigma E$, where $E \sim \mathcal{N}(0, I)$ is independent of (Z_0, Y), the conditional data have finite second moments, and the model gives the population-optimal velocity $v^*(z, c, \sigma) = \mathbb{E}[E - Z_0 \mid Z_\sigma = z, c]$. Then*

$$g = -\frac{1-\sigma}{\sigma}\left[v^*(z, c_t, \sigma) - v^*(z, c_s, \sigma)\right]. \tag{10}$$

If $F(z) \neq 0$, then $F / \operatorname{tr} F = P_{d^*}$, where $d^* = \sigma\gamma[v^*(z, c_t, \sigma) - v^*(z, c_s, \sigma)]$ and $P_b = bb^\top / \|b\|^2$ projects onto a nonzero axis b.

This links conditional information geometry to flow-matching velocities (Lipman et al., 2023); Appendix A gives the proof. At $z = p$, the practical controller uses the model proxy

$$d = \sigma\gamma\delta v, \qquad P_d = \frac{dd^\top}{\|d\|^2}, \qquad Q_d = I - P_d, \tag{11}$$

with $P_0 = 0$. Existing predictions provide this proxy without estimating ω. With reference noise fixed, $\partial p / \partial x = I$, so the axis also applies to local corrections of x. Under Proposition 1, spatial channel blocks satisfy $\operatorname{tr}(F_{jj}) \propto \left\|\delta v^*_{:,j}\right\|^2$, with a shared factor. The same response therefore informs direction and relative spatial sensitivity. The implementation uses fixed empirical thresholds to select spatial support (Appendix C).

Table 1: Round-trip reconstruction over 700 tasks using SD3 Medium. Metrics compare returned and original images; PSNR is in dB. Displayed LPIPS and DINO distances equal raw values $\times 10^3$; MSE equals raw MSE $\times 10^4$. Other columns are unscaled. Budgets follow Table 2.

Method	PSNR↑	LPIPS↓	DINO distance↓	MSE↓
SDEdit	14.206	411.73	74.87	415.82
DDIB	9.162	631.35	154.71	1340.22
RF-Inversion	12.618	485.01	91.40	583.43
FlowEdit	18.669	206.53	42.82	154.80
FlowAlign	22.950	127.25	23.76	67.67
UniEdit	22.541	171.48	28.43	67.04
Prism	25.647	78.81	14.25	37.90
SAM-Flow	20.067	140.18	37.52	116.86
CGFC (Ours)	**26.468**	**68.96**	**12.46**	**30.36**

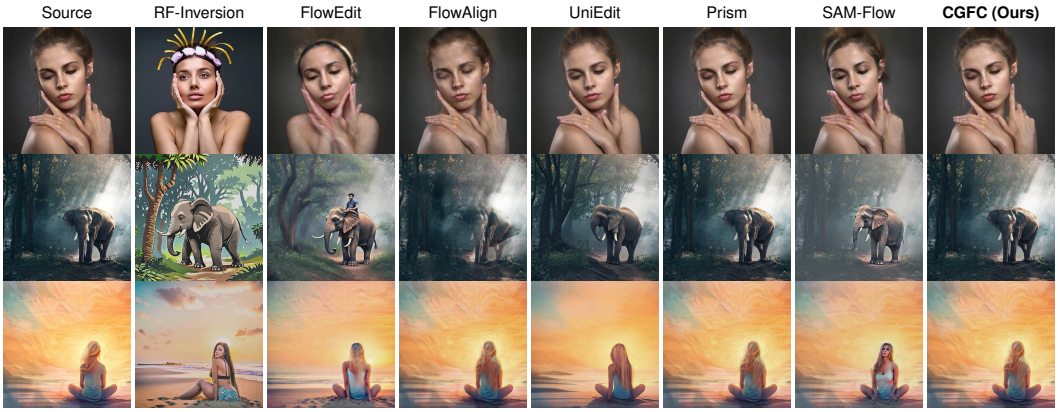

Figure 4: **SD3 round-trip reconstruction.** Original inputs (first column) and returns from each method's own edited image. These examples illustrate content consistency over the edit–return trajectory.

3.3 Conditional geometric feedback control

Residual feedback controlled by β. To distinguish directional adjustment from scalar weakening, CGFC preserves the supported residual norm. Let $r_M = Mr$ and $d_M = Md$. We attenuate the conditional-axis component and restore the norm within the same support:

$$w_M = \big[I - (1-\beta)P_{d_M}\big] r_M, \qquad s_M = \frac{\|r_M\|}{\|w_M\|}, \\ \widetilde{r} = (I-M)r + s_M w_M, \qquad 0 < \beta \le 1. \tag{12}$$

Set $s_M = 1$ when $r_M = 0$. The residual, axis, projection, and norm share M, aligning the geometric calculation with its spatial support. This operation leaves the outside-support residual and direct transport unchanged. Its operator is

$$A_M = (I-M) + s_M\big[I - (1-\beta)P_{Md}\big]M. \tag{13}$$

Proposition 2 (Local norm and restricted Fisher cost)**.** *For a fixed binary support M, Eq. (12) satisfies*

$$\|M\widetilde{r}\| = \|Mr\|, \qquad (I-M)\widetilde{r} = (I-M)r, \qquad \|\widetilde{r}\| = \|r\|. \tag{14}$$

If the model axis agrees with the ideal axis in Proposition 1, then, for $F_M = MFM$,

$$\widetilde{r}^\top F_M \widetilde{r} \le r^\top F_M r. \tag{15}$$

This establishes norm-preserving control of sensitivity under the supported metric F_M at the current state (Appendices B and D).

Table 2: Full-benchmark image editing over 700 tasks using SD3 Medium. Structure is DINO structure distance; PSNR and LPIPS compare full edited and source images. CLIP is computed on the full image (Full) and edited region (Region). Displayed Structure and Full LPIPS equal raw values $\times 10^3$; other columns are unscaled. PSNR is in dB; NFE is per direction.

	Content preservation			CLIP↑		Budget
Method	Structure↓	Full PSNR↑	Full LPIPS↓	Full	Region	NFE
SDEdit	45.10	17.730	304.24	26.827	23.549	66
DDIB	140.95	9.803	579.32	27.132	23.504	68
RF-Inversion	71.80	14.855	407.50	**27.493**	**24.204**	84
FlowEdit	24.18	21.483	143.89	26.502	22.936	132
FlowAlign	11.27	25.305	85.58	25.478	21.911	99
UniEdit	17.76	23.991	150.33	25.643	22.196	100
Prism	7.86	27.381	66.91	24.569	21.183	99
SAM-Flow	18.71	23.319	95.99	25.640	22.348	99
CGFC (Ours)	**6.54**	**28.490**	**57.65**	24.335	21.019	99

Displacement feedback controlled by λ. The β-controlled operation acts on the current residual rather than directly constraining accumulated displacement $u = x - z_s$. CGFC addresses the latter with feedback orthogonal to the *current global conditional axis* d. Its magnitude is capped relative to the reference increment n, computed before the β- and λ-controlled changes:

$$n = \Delta\sigma\Delta v + \zeta r, \qquad R = \rho \|n\|, \qquad a = -\eta Q_d u, \quad \eta = \min\left(\lambda, \frac{R}{\|Q_d u\|}\right), \tag{16}$$

where $\lambda, \rho \geq 0$. If $Q_d u = 0$, set $\eta = \lambda$ and $a = 0$. The λ-controlled feedback uses the unmasked P_d for global displacement; the β-controlled residual update uses P_{Md} within the editing support.

Proposition 3 (Local control characterization of displacement feedback)**.** *Equation* (16) *uniquely solves the convex projection problem*

$$a = \arg\min_b \tfrac{1}{2} \|b + \lambda u\|^2, \qquad d^\top b = 0, \quad \|b\| \leq R. \tag{17}$$

For $0 \leq \eta \leq 2$ *and* $V(u) = \frac{1}{2}\|u\|^2$*, the* λ*-controlled displacement substep alone satisfies*

$$V(u + a) - V(u) = -\eta(1 - \eta/2) \|Q_d u\|^2 \leq 0. \tag{18}$$

Under ideal axis correspondence, $d^\top a = 0$ implies $g^\top a = 0$: this is the feedback closest to $-\lambda u$ within the step budget that has zero first-order effect on conditional log-odds. Thus both corrections use the same conditional response, acting on instantaneous disagreement and accumulated displacement, respectively.

Implementation and scope. Algorithm 1 in Appendix C evaluates both terms at the same state. Projections use inner products and elementwise operations, with no extra denoiser calls beyond the three-prediction interface. Section 4 evaluates the resulting editing and preservation behavior, and Appendix G discusses the scope of the local guarantees.

4 Experiments

We evaluate CGFC through the following experiments. Bold scores indicate backbone-wise best values, including displayed ties.

4.1 Experimental setup

Data and metrics. We evaluate SD3 Medium on all 700 PIE-Bench tasks (Ju et al., 2024). CLIP reports image–target-text similarity on full images and annotated editing regions; regional scoring sets pixels outside the region to black without cropping. Its score alone does not determine overall editing quality and is interpreted alongside preservation metrics and qualitative results. DINO structure distance, full-image PSNR, and full-image LPIPS measure preservation. Editing metrics compare edited and source images; round-trip metrics compare returned and source images. The main SD3 comparison tables use all 700 tasks; smaller analyses specify their own task counts. Evaluation masks are used only for regional CLIP scoring.

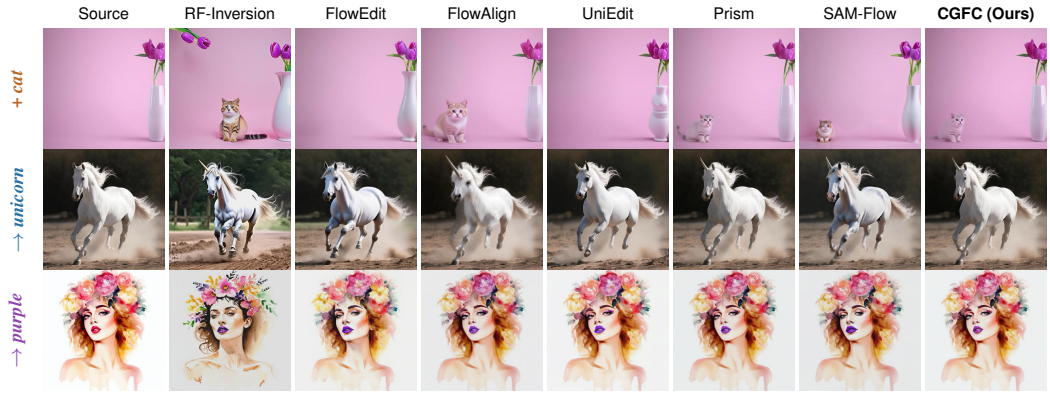

Figure 5: **Qualitative editing on SD3.** Methods share each row's input and instruction for object addition, modification, or local color editing. Outputs are unretouched; methods retain their registered settings and budgets.

Methods and budgets. We compare CGFC with SDEdit, DDIB, RF-Inversion, FlowEdit, FlowAlign, UniEdit, Prism, and SAM-Flow (Meng et al., 2022; Su et al., 2023; Rout et al., 2025; Kulikov et al., 2025; Kim et al., 2026; Jiao et al., 2026; Jeong et al., 2026; Cui et al., 2026), using frozen weights. CGFC reuses three predictions per update: 33 updates cost 99 denoiser function evaluations (NFE) per direction. Methods retain their registered guidance, schedules, and image pipelines; budgets appear in Table 2.

4.2 Feedback trajectories in analytical models

We compare two settings of the same feedback law in 128-dimensional Gaussian models with exact conditional velocities: the spatial reference has $(\beta, \lambda) = (1, 0)$, and geometric feedback has $(0.75, 0.02)$. Both use the same dynamic support; at $(1, 0)$, the residual direction is unchanged and displacement feedback is zero. Plots show editing progress α and local/background projections h_ℓ, h_b; errors use the full latent space (Appendix F).

Recovery from matched perturbations. After a shared editing prefix, both configurations receive the same content perturbation and recover for 32 steps. Across all 24 prespecified nonzero conditions, geometric feedback reduces mean content drift from 0.244 to 0.164 (Figure 2).

State-dependent editing from source samples. Different source and target covariances make the velocity contrast state dependent. Both configurations share 256 source starts and reference noise for 33 updates, without an imposed perturbation. Geometric feedback reduces mean content drift from 0.224 to 0.190 (Figure 3).

4.3 Round-trip reconstruction

We assess edit–return consistency using a source-to-target-to-source cycle, a construction also used in cycle-consistent image editing (Wang et al., 2026). Each task edits the input, swaps source and target prompts, and edits back. Table 1 measures full-image consistency over 700 SD3 edit–return tasks. Forward images follow each method's round-trip protocol.

CGFC has the best four round-trip means among the listed SD3 configurations. PSNR increases from Prism's 25.647 to 26.468 and LPIPS decreases from 0.07881 to 0.06896. These results support the controller's ability to retain source content over the edit–return trajectory.

4.4 Comparison with existing editors

CGFC attains the best three preservation scores among the listed SD3 methods (Table 2). Full-image editing PSNR rises from Prism's 27.381 to 28.490, and full-image LPIPS decreases from 0.06691 to 0.05765. Together with Figure 5, these results characterize the preservation benefit of conditional geometric feedback.

4.5 CGFC PARAMETER ABLATION

Table 3 compares six configurations on a randomly sampled subset of 100 images from PIE-Bench tasks using SD3 Medium. Here, χ toggles only the applied residual feedback; the cap in Eq. (16) retains $\zeta = 0.01$ throughout. Setting $1 - \beta = 0$ preserves the residual direction, while $\lambda = 0$ disables displacement feedback. When $\chi = 0$, directional adjustment is inactive. Removing residual feedback substantially reduces source preservation. With $\chi = 1$ and no directional adjustment, displacement feedback increases editing and round-trip PSNR by 1.504 and 1.063 dB, respectively. Directional adjustment provides further gains of 0.193 and 0.042 dB. The complete configuration achieves the best reported preservation and reconstruction means among the six settings.

Table 3: CGFC parameter ablation on a randomly sampled subset of 100 images from PIE-Bench. χ controls application of $\zeta\widetilde{r}$, $1 - \beta$ controls its direction, and λ controls displacement feedback; $\zeta = 0.01$ remains in the feedback cap for all rows. Displayed LPIPS and DINO equal raw values $\times 10^3$; MSE equals raw MSE $\times 10^4$.

			Editing				Round-trip reconstruction			
χ	$1-\beta$	λ	PSNR↑	LPIPS↓	MSE↓	DINO↓	PSNR↑	LPIPS↓	MSE↓	DINO↓
0	0.00	0.00	14.758	342.56	701.32	78.56	13.553	405.92	859.29	105.94
0	0.00	0.02	17.187	253.17	420.01	48.99	15.981	303.59	531.09	63.63
1	0.00	0.00	25.906	93.20	35.98	10.55	24.871	110.46	45.66	17.52
1	0.25	0.00	26.207	90.73	32.83	10.05	25.038	107.67	43.29	16.58
1	0.00	0.02	27.410	77.37	24.64	8.03	25.934	94.04	35.33	13.82
1	0.25	0.02	**27.603**	**76.46**	**23.20**	**7.97**	**25.976**	**92.95**	**34.26**	**13.66**

4.6 PRESERVATION ACROSS GUIDANCE SCALES

Table 2 shows a preservation–alignment trade-off: CGFC has lower full-image and regional target CLIP than Prism (24.335/21.019 versus 24.569/21.183), alongside better full-image PSNR and LPIPS. On 100 randomly across seven target CFG settings, CGFC has higher full-image editing PSNR than FlowEdit, FlowAlign, and Prism at each tested CFG. Appendix E.1 (Figure 6) reports the CLIP–PSNR curves; this small sweep does not establish Pareto dominance.

4.7 TRANSFER WITHIN THE SD FAMILY

With unchanged control parameters, we evaluate CGFC and five existing editors on all 700 tasks for each of SD3.5 Medium and Large. Relative to Prism, CGFC increases editing PSNR by 0.412 and 0.604 dB, respectively, and reduces full-image editing and round-trip LPIPS on both backbones. CGFC has the lowest editing and round-trip LPIPS among the compared methods on both backbones, and the highest editing and round-trip PSNR on both backbones. Appendix E.2 reports the six-method comparisons.

5 CONCLUSION

We presented CGFC, a training-free and inversion-free image editor guided by conditional Fisher geometry. A shared axis, estimated from same-state conditional velocity differences, organizes norm-preserving residual adjustment and bounded feedback on accumulated source displacement. The controller reuses existing predictions, with local properties established under stated assumptions.

Analytical experiments illustrate reduced content drift in the tested settings. On 700 PIE-Bench tasks, CGFC improves source preservation and round-trip consistency on SD3, with preservation gains transferring to SD3.5 under fixed controller parameters, alongside a preservation–alignment trade-off. We view CGFC as a heuristic exploration of geometric control in latent space, motivating task-appropriate geometric metrics, feedback strategies, and broader evaluation.

AI USE STATEMENT

Generative AI assisted with translation and language editing, manuscript and figure preparation, literature discovery, conceptual discussion and mathematical exposition, toy-experiment and visualization code, and experimental-result analysis. The authors remain responsible for the scientific claims, results, and final content.

REFERENCES

Eashan Adhikarla and Brian D. Davison. RemEdit: Efficient diffusion editing with Riemannian geometry. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5037–5046, 2026. URL https://arxiv.org/abs/2601.17927.

Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitrios Kalatzis, and Søren Hauberg. Pulling back information geometry. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4872–4894, 2022. URL https://proceedings.mlr.press/v151/arvanitidis22b.html.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22560–22570, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Cao_MasaCtrl_Tuning-Free_Mutual_Self-Attention_Control_for_Consistent_Image_Synthesis_and_ICCV_2023_paper.html.

Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspace in diffusion models for controllable image editing. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0859. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3018804d037cc101b73624f381bed0cb-Abstract-Conference.html.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3lge0p5o-M-.

Haowang Cui, Rui Chen, Tao Luo, Tao Guo, Zheng Qin, and Jiaze Wang. SAM-Flow: Source-anchored masked flow for training-free image editing. *arXiv preprint arXiv:2606.06228*, 2026. URL https://arxiv.org/abs/2606.06228v1. Version 1.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2208.01626.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Wootaek Jeong, Junghyo Sohn, Jee Seok Yoon, and Heung-Il Suk. Deconstructing guidance: A semantic hierarchy for precise diffusion model editing. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/623f3d87442bd220ecab3e8a5e4dcda6-Abstract-Conference.html.

Yilei Jiang, Zhen Wang, Yanghao Wang, Jun Yu, Yueting Zhuang, Jun Xiao, and Long Chen. FlowDC: Flow-based decoupling-decay for complex image editing. *arXiv preprint arXiv:2512.11395*, 2025. URL https://arxiv.org/abs/2512.11395v1. Version 1.

Guanlong Jiao, Biqing Huang, Kuan-Chieh Wang, and Renjie Liao. UniEdit-Flow: Unleashing inversion and editing in the era of flow models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/023d94f44110b9a3c62329beec739772-Abstract-Conference.html.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. PnP inversion: Boosting diffusion-based editing with 3 lines of code. In *International Conference on Learning Representations*, 2024. URL https://github.com/cure-lab/PnPInversion.

Rafał Karczewski, Markus Heinonen, Alison Pouplin, Søren Hauberg, and Vikas Garg. The spacetime of diffusion models: An information geometry perspective. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ad2b3b0963114adddbd2c97d6e74abb9-Abstract-Conference.html.

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*, 2026. URL https://arxiv.org/abs/2505.23145v3.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. FlowEdit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025. URL https://arxiv.org/abs/2412.08629.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2108.01073v2.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. NULL-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023. URL https://arxiv.org/abs/2211.09794.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of Riemannian geometry. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24129–24142, 2023. doi: 10.52202/075280-1048. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4bfcebedf7a2967c410b64670f27f904-Abstract.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2112.10752.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c07bc8b520aa87326b015b0da6a70b48-Abstract-Conference.html.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *International Conference on Learning Representations*, 2023.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023. URL https://arxiv.org/abs/2211.12572.

Binxu Wang and Carlos R. Ponce. The geometry of deep generative image models and its applications. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.06006v2.

Yanghao Wang, Zhen Wang, and Long Chen. FlowCycle: Target-aware image editing via cycle-consistent constraints. In *Computer Vision – ECCV 2026*, pp. 1–18, 2026. doi: 10.1007/978-3-032-37464-6_1. URL https://link.springer.com/chapter/10.1007/978-3-032-37464-6_1.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9452–9461, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xu_Inversion-Free_Image_Editing_with_Language-Guided_Diffusion_Models_CVPR_2024_paper.html.

A Conditional Fisher structure and ideal velocity correspondence

A.1 Binary conditional posterior and local KL expansion

Fix $0 < \sigma < 1$. The binary mixture density is

$$p_\sigma(z) = \pi_s p_\sigma(z \mid c_s) + \pi_t p_\sigma(z \mid c_t). \tag{A.1}$$

Bayes' rule gives $\Pr(Y = t \mid z) = \omega(z)$ and $\Pr(Y = s \mid z) = 1 - \omega(z)$. Since $\nabla\omega = \omega(1-\omega)g$, the posterior score vectors are

$$\nabla \log \omega = (1-\omega)g, \qquad \nabla \log(1-\omega) = -\omega g. \tag{A.2}$$

Taking their expected outer product under the posterior yields

$$F = \omega(1-\omega)^2 gg^\top + (1-\omega)\omega^2 gg^\top = \omega(1-\omega)gg^\top. \tag{A.3}$$

Equivalently, this pulls back the Bernoulli Fisher metric $1/[\omega(1-\omega)]$ through $z \mapsto \omega(z)$:

$$F = \frac{\nabla\omega\nabla\omega^\top}{\omega(1-\omega)}. \tag{A.4}$$

The geometry therefore arises from posterior sensitivity to the latent input, without assuming a known content manifold.

Let $L(h) = \sum_y \Pr(y \mid z) \log[\Pr(y \mid z)/\Pr(y \mid z+h)]$. For a smooth posterior with $0 < \omega < 1$, $L(0) = 0$. Posterior normalization gives $\nabla_h L(0) = 0$, and the information identity gives $\nabla_h^2 L(0) = F(z)$. Taylor expansion proves Eq. (9). This compares nearby posteriors at a fixed noise level, rather than probability laws over full editing trajectories.

A.2 Proof of Proposition 1

Under $Z_\sigma = (1-\sigma)Z_0 + \sigma E$, define $m_0(z,c) = \mathbb{E}[Z_0 \mid Z_\sigma = z, c]$ and $m_E(z,c) = \mathbb{E}[E \mid Z_\sigma = z, c]$. Then

$$z = (1-\sigma)m_0 + \sigma m_E, \qquad v^* = m_E - m_0, \tag{A.5}$$

so $m_0 = z - \sigma v^*$ and $m_E = z + (1-\sigma)v^*$. Differentiating the convolution density with Gaussian covariance $\sigma^2 I$ gives

$$\begin{aligned} \nabla_z \log p_\sigma(z \mid c) &= -\frac{z - (1-\sigma)m_0(z,c)}{\sigma^2} \\ &= -\frac{m_E(z,c)}{\sigma} = -\frac{z + (1-\sigma)v^*(z,c,\sigma)}{\sigma}. \end{aligned} \tag{A.6}$$

The fixed label priors have zero gradient with respect to z. Subtracting the two conditional scores proves Eq. (10). Substitution into the Fisher expression and cancellation of the scalar factor, when $F \neq 0$, yield

$$\frac{F}{\operatorname{tr} F} = \frac{\delta v^* (\delta v^*)^\top}{\|\delta v^*\|^2} = P_{d^*}, \qquad \delta v^* = v^*(z, c_t, \sigma) - v^*(z, c_s, \sigma). \tag{A.7}$$

The implementation replaces δv^* with the neural model's conditional velocity difference. This is a computable directional proxy; we do not assume that the network satisfies the exact density-ratio score identity at every state.

B Proofs of local control properties

B.1 Proof of Proposition 2

Let $b = P_{Md} Mr$ and $e = (I - P_{Md})Mr$. Both lie in the support of M, with $b \perp e$, and

$$Mr = b + e, \qquad w_M = \beta b + e, \qquad s_M^2 = \frac{\|b\|^2 + \|e\|^2}{\beta^2 \|b\|^2 + \|e\|^2}. \tag{B.1}$$

The expression $\widetilde{r} = (I - M)r + s_M w_M$ immediately gives invariance outside the support and norm preservation inside it. Orthogonality of the two supports also gives $\|\widetilde{r}\| = \|r\|$.

Under ideal axis correspondence, F_M is a nonnegative multiple of P_{Md}. For $Mr \neq 0$, let $\alpha_M = \|b\|^2 / \|Mr\|^2 \in [0, 1]$. Then

$$\widetilde{r}^\top F_M \widetilde{r} = \frac{\beta^2}{1 - (1 - \beta^2)\alpha_M} \, r^\top F_M r \leq r^\top F_M r. \tag{B.2}$$

If $Mr = 0$ or $Md = 0$, the zero-vector conventions give the same conclusion. Nonzero costs are equal when $\beta = 1$ or the input is entirely parallel to the conditional axis. If its sensitive component vanishes, both costs are zero. Thus the restricted Fisher cost is nonincreasing.

B.2 Proof of Proposition 3

First suppose $d \neq 0$. For $d^\top b = 0$, the objective decomposes as

$$\tfrac{1}{2} \|b + \lambda u\|^2 = \tfrac{1}{2} \|b + \lambda Q_d u\|^2 + \tfrac{1}{2}\lambda^2 \|P_d u\|^2 . \tag{B.3}$$

The second term is independent of b, so the unconstrained minimizer within the orthogonal subspace is $-\lambda Q_d u$. Projection onto the closed ball of radius R gives

$$a = \begin{cases} -\min(\lambda, R/\|Q_d u\|)\, Q_d u, & \|Q_d u\| > 0, \\ 0, & \|Q_d u\| = 0. \end{cases} \tag{B.4}$$

Strict convexity of the objective and the closed convex feasible set ensure uniqueness. If $d = 0$, the orthogonality constraint is vacuous and $Q_0 = I$, reducing the solution to ball projection of $-\lambda u$.

For the potential, $u^\top Q_d u = \|Q_d u\|^2$ implies

$$\begin{aligned} V(u + a) - V(u) &= u^\top a + \tfrac{1}{2} \|a\|^2 \\ &= -\eta \|Q_d u\|^2 + \tfrac{1}{2}\eta^2 \|Q_d u\|^2 . \end{aligned} \tag{B.5}$$

This is nonpositive for $\eta \in [0, 2]$. The proof freezes the current axis and considers only the displacement-feedback substep, excluding subsequent conditional transport and changes in the state-dependent operators.

B.3 Unified feedback form

Substituting $r = -u + \sigma\Delta v$ expands the complete increment as

$$\Delta\sigma\Delta v + \zeta A_M r - \eta Q_d u = \big(\Delta\sigma I + \zeta\sigma A_M\big)\Delta v - \big(\zeta A_M + \eta Q_d\big)u. \tag{B.6}$$

This proves Eqs. (5)–(6). The identity requires neither an exact density-score model nor an operator A_M that is fixed across steps; s_M itself depends on the current residual.

C Spatial support and numerical implementation

C.1 Dynamic spatial support

At a fixed latent state, δv changes only the condition; its channel norm $m_j = \|\delta v_{:,j}\|_2$ summarizes the response at spatial position j. Under Proposition 1, the corresponding principal channel block F_{jj} satisfies

$$\operatorname{tr}(F_{jj}) = \omega(1 - \omega) \left(\frac{1 - \sigma}{\sigma}\right)^2 \left\|\delta v^*_{:,j}\right\|_2^2 . \tag{C.1}$$

All positions share the same factor at a fixed state and noise level. Squared response magnitudes therefore describe relative sensitivity across spatial blocks, with the model's m_j^2 as a computable proxy.

Algorithm 1 Conditional Geometric Feedback Control (CGFC; one editing direction)

Require: Image I_s, conditions c_s, c_t, schedule $\{\sigma_k\}_{k=0}^{N}$
Require: Prespecified edit type τ, parameters $\gamma, \beta, \zeta, \lambda, \rho$
Ensure: Edited image $\widehat{I}$

$z_s \leftarrow \mathcal{E}(I_s)$, $x \leftarrow z_s$
for $k = 0, \ldots, N-1$ **do**
  Sample ϵ_k; form q, p, u by Eq. (1), then evaluate v_t, v_s, v_q by Eq. (2)
  $\delta v \leftarrow v_t - v_s$, $M \leftarrow \mathrm{DynamicMask}(\delta v, \tau)$, $\Delta v \leftarrow v_s + \gamma M \delta v - v_q$
  $d \leftarrow \sigma_k \gamma \delta v$, $r \leftarrow -u + \sigma_k \Delta v$
  β-controlled residual direction: $r_M \leftarrow Mr$, $w_M \leftarrow r_M - (1-\beta) P_{Md} r_M$
  $s_M \leftarrow \|r_M\| \,/\, \|w_M\|$ (set $s_M = 1$ if $r_M = 0$); $\widetilde{r} \leftarrow r - r_M + s_M w_M$
  λ-controlled displacement feedback: $a \leftarrow \mathrm{ProjectBall}(-\lambda Q_d u, \rho \,\|\Delta\sigma_k \Delta v + \zeta r\|)$
  $x \leftarrow x + \Delta\sigma_k \Delta v + \zeta \widetilde{r} + a$
end for
return $\widehat{I} = \mathcal{D}(x)$

CGFC uses this computable sensitivity proxy as a dynamic spatial gate for geometric feedback. At every step, we standardize m_j by its spatial mean $\bar{m}$ and sample standard deviation s_m. The prespecified edit type then selects the binary support:

$$z_j = \frac{m_j - \bar{m}}{s_m + 10^{-8}}, \qquad M_j = \begin{cases} \mathbf{1}[z_j \geq 3], & \text{object edits,} \\ \mathbf{1}[z_j < 0.6], & \text{background/style edits.} \end{cases} \tag{C.2}$$

The task protocol specifies the edit type τ. The cutoffs 3 and 0.6 are empirical values used in prior magnitude-based spatial guidance (Jeong et al., 2026); Proposition 1 motivates the sensitivity proxy but does not derive these cutoffs. Object edits select prominent responses; background and style edits select lower responses while excluding the strongly responding core. The resulting support gates CGFC's geometric feedback and is broadcast over channels. It reuses current predictions, without ground-truth evaluation masks or an additional segmentation model.

The same M controls guidance in Eq. (3) and the residual, axis, and norm restoration in Eq. (12). Only conditional guidance and the change to residual direction are selected by M: the residual outside the support remains in the update, and the λ-controlled displacement feedback uses the unmasked global axis. The local propositions hold for the current binary support and do not differentiate through mask selection.

C.2 Computational budget and execution order

Algorithm 1 requires three conditional predictions, v_t, v_s, v_q, per update; its additional control operations cost $O(D)$. The evaluated configuration uses 33 updates and 99 denoiser calls per editing direction, with $\beta = 0.75$, $\zeta = 0.01$, $\lambda = 0.02$, and $\rho = 0.25$. Guidance strength follows the configuration for each edit type. The control operations reuse this three-prediction interface.

The implementation computes projections, norm ratios, and the planned displacement correction in FP32 before casting to execution precision. For SD3, the β-controlled update uses projection denominator $\max(\|Md\|^2, 10^{-8})$, and the λ-controlled update uses $\max(\|d\|^2, 10^{-8})$. Squared norms in norm restoration have floor 10^{-12}, and the correction's ball projection also uses a numerical floor. The adjusted residual is cast first; displacement feedback is then injected through the residual interface:

$$\begin{aligned} \widetilde{r}_{\mathrm{exec}} &= \mathrm{cast}(\widetilde{r}_{\mathrm{FP32}}), \\ r_{\mathrm{exec}} &= \mathrm{cast}\left(\mathrm{float32}(\widetilde{r}_{\mathrm{exec}}) + \frac{a}{\zeta}\right), \\ x^{+} &= x + \Delta\sigma\Delta v + \zeta r_{\mathrm{exec}}. \end{aligned} \tag{C.3}$$

Displacement feedback is applied once through residual injection. The theoretical identities use exact arithmetic and the stated zero-vector conventions; execution uses denominator safeguards and finite-precision casting, so orthogonality, norm restoration, and the clipping bound are subject to numerical deviations. Equation (5) explains the algorithm without changing the evaluated operation order.

C.3 Endpoint conventions

The statistical propositions assume $0 < \sigma < 1$. At $\sigma = 1$, both conditional marginals are standard Gaussian and the true label Fisher is zero; an arbitrary nonzero prediction difference cannot be identified as a nonzero Fisher axis there. The score identity is singular at $\sigma = 0$. The final update advances from a positive noise level to zero without reevaluating that identity at zero. We retain the backbone's noise schedule and endpoint handling.

D Scope of the local geometric properties

The restricted cost $F_M = MFM$ measures conditional sensitivity for perturbations supported within the current editing region. It differs from the global cost F. For example, take

$$d = (1, 0, 1)^\top, \quad M = \mathrm{diag}(1, 1, 0), \quad r = (-1, 1, 1)^\top, \quad \beta = \tfrac{3}{4}. \tag{D.1}$$

Initially $d^\top r = 0$. After the β-controlled residual adjustment, with $s_M = \sqrt{2}/\sqrt{\beta^2 + 1}$, $\widetilde{r} = (-\beta s_M, s_M, 1)^\top$ and $d^\top \widetilde{r} = 1 - \beta s_M \neq 0$. Nonincrease of restricted Fisher cost therefore does not imply nonincrease of global Fisher cost.

Likewise, the λ-controlled feedback satisfies $d^\top a = 0$, so ideal axis correspondence gives

$$(\zeta\widetilde{r} + a)^\top F(\zeta\widetilde{r} + a) = \zeta^2 \widetilde{r}^\top F \widetilde{r}. \tag{D.2}$$

This cannot be combined with Eq. (15), which concerns F_M, to establish global contraction of the combined controller. Norm preservation also applies only to the preservation residual, not the total increment containing transport and displacement feedback.

Conditional Fisher information measures local sensitivity of source–target label discrimination, while input distance provides a feedback reference during editing. The theoretical results describe directions and constraints of fixed-state control suboperations. In particular, conditional orthogonality is a first-order posterior property; target semantics, structure, and background preservation in the resulting images require joint empirical evaluation.

E Supplementary results

E.1 Guidance-scale sensitivity

We compare CGFC, FlowEdit, FlowAlign, and Prism at target CFG $\{5, 7.5, 10, 12.5, 15, 20, 30\}$, with fixed CGFC control parameters. Figure 6 shows descriptive CLIP and editing PSNR means at each setting. CGFC has higher PSNR at each tested target CFG, while the CLIP–PSNR curves cover only the displayed alignment range.

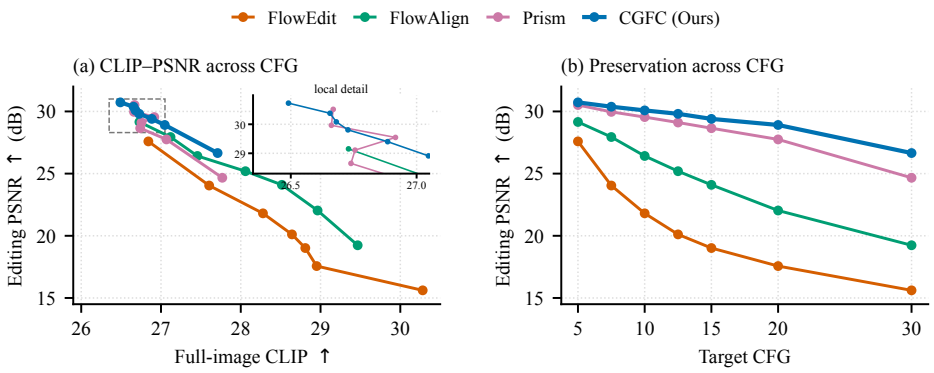

Figure 6: **CLIP scores and preservation across CFG on SD3.** (a) CLIP versus editing PSNR; the inset enlarges the dashed region. All seven settings are connected in CFG order. (b) Full-image editing PSNR versus target CFG. Curves show descriptive means over the same 100 randomly tasks; CGFC is blue.

E.2 SD3.5 TRANSFER WITH FIXED CONTROLLER PARAMETERS

We evaluate the same 700 PIE-Bench tasks on SD3.5 Medium and SD3.5 Large. Both transfer backbones use the same full-image source-comparison metrics and CLIP scoring as the SD3 direct-editing evaluation.

Table 4: Editing transfer to SD3.5 over 700 tasks per backbone. Structure, PSNR, and LPIPS use full images. CLIP is scored on the full image (Full) and edited region (Region). Displayed Structure and LPIPS equal raw values $\times 10^3$; other columns are unscaled. PSNR is in dB; NFE is per direction. CGFC control parameters are fixed.

Method	Structure↓	PSNR↑	LPIPS↓	CLIP Full↑	CLIP Region↑	NFE
SD3.5 Medium						
FlowEdit	1.93	33.398	19.40	24.137	20.776	100
FlowAlign	2.24	32.749	20.40	24.108	20.710	99
UniEdit	12.28	25.531	135.13	**26.221**	**22.478**	100
Prism	1.26	35.453	11.68	23.533	20.303	99
SAM-Flow	1.98	32.869	18.56	24.158	20.849	99
CGFC (Ours)	**1.18**	**35.865**	**11.03**	23.459	20.263	99
SD3.5 Large						
FlowEdit	3.03	31.232	29.26	24.850	21.316	100
FlowAlign	3.43	30.966	30.28	24.824	21.291	99
UniEdit	15.67	24.370	154.63	**26.243**	**22.696**	100
Prism	1.42	34.706	13.43	23.847	20.527	99
SAM-Flow	2.97	30.522	28.12	24.798	21.408	99
CGFC (Ours)	**1.27**	**35.310**	**12.38**	23.701	20.435	99

Table 5: Round-trip reconstruction on the same 700 SD3.5 tasks per backbone. All four metrics use full images. Displayed LPIPS and DINO distances equal raw values $\times 10^3$; MSE equals raw MSE $\times 10^4$. Other columns are unscaled; PSNR is in dB. Method configurations match Table 4.

Method	PSNR↑	LPIPS↓	DINO distance↓	MSE↓
SD3.5 Medium				
FlowEdit	30.992	23.55	4.15	9.18
FlowAlign	31.499	20.12	4.13	7.91
UniEdit	24.979	145.73	17.18	39.13
Prism	32.341	15.31	3.18	6.35
SAM-Flow	29.228	30.07	4.46	14.20
CGFC (Ours)	**32.422**	**14.97**	**3.16**	**6.22**
SD3.5 Large				
FlowEdit	29.333	34.26	5.58	13.99
FlowAlign	30.316	27.40	5.42	11.20
UniEdit	23.648	158.57	19.26	52.82
Prism	32.054	16.31	3.33	6.84
SAM-Flow	27.191	44.93	6.23	22.29
CGFC (Ours)	**32.223**	**15.71**	**3.27**	**6.53**

Each metric is computed per task and averaged arithmetically over all 700 tasks. PSNR, LPIPS, MSE, and DINO distance compare full images with their sources; full-image and edited-region CLIP assess target-text alignment. Return edits swap the prompts and start from each method's own forward result. Forward outputs are reused for the return direction.

Across both SD3.5 backbones, CGFC achieves the highest PSNR and lowest LPIPS for both editing and round-trip reconstruction among the six methods. Relative to Prism, editing PSNR improves by 0.412/0.604 dB on Medium/Large, and round-trip PSNR by 0.081/0.169 dB. Full-image structure distance, round-trip DINO distance, and MSE also decrease on both backbones. Figures 7 and 8 show reconstruction and editing examples, respectively.

E.3 Qualitative reconstruction across SD3.5 backbones

Figure 7 compares returned images for the same cat, fruit, Bernese mountain dog, and portrait examples on SD3.5 Medium and Large. Each method first edits the source and then swaps the source and target prompts for the return. These selected reconstruction examples complement the full-benchmark results in Table 5; they illustrate source-content consistency over the edit–return trajectory.

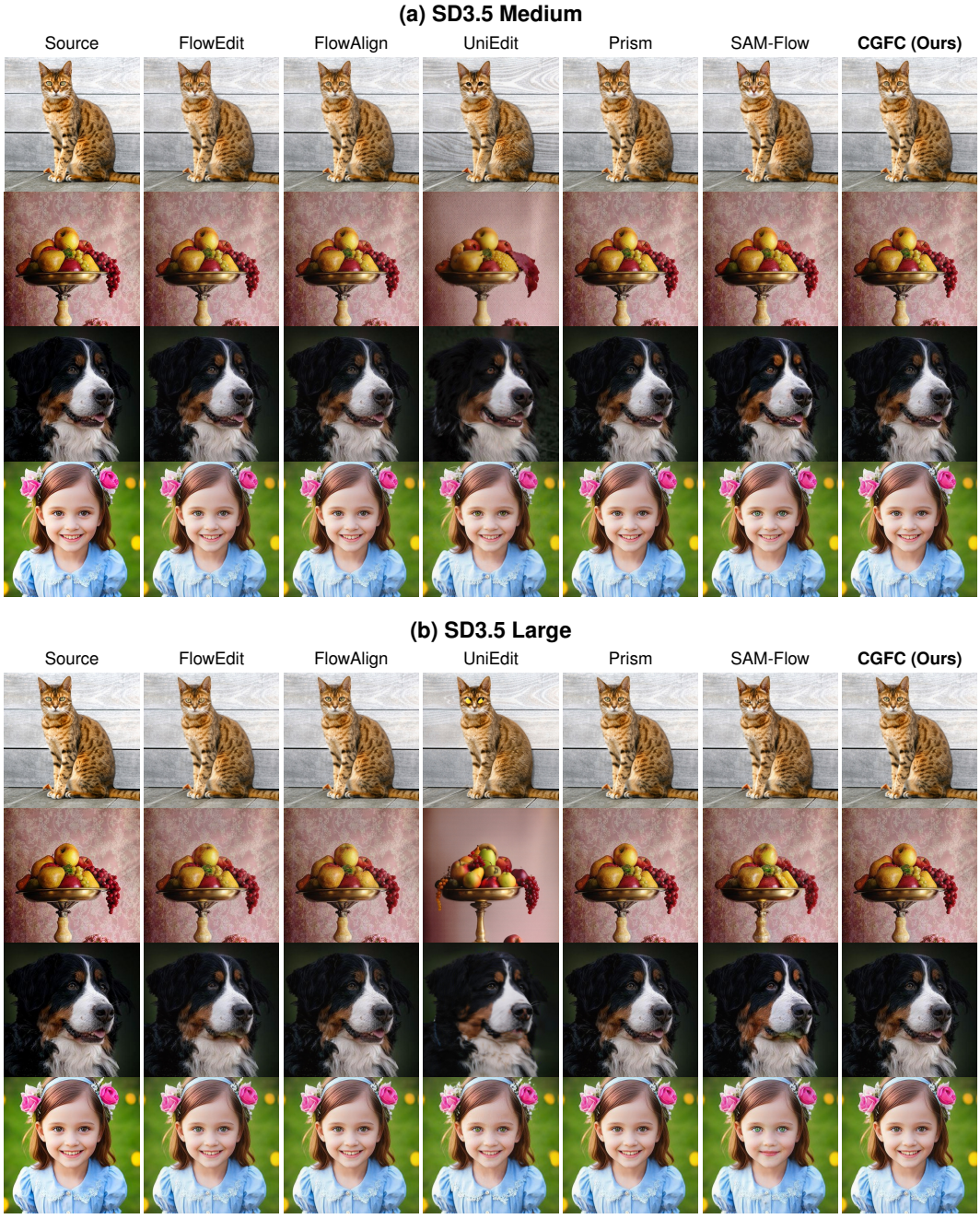

Figure 7: **Round-trip reconstruction on SD3.5 Medium and Large.** Source images (first column) and returns from each method's own forward edit. Rows show a cat, fruit, a Bernese mountain dog, and a portrait, with identical task IDs across the two backbones. Original outputs are displayed using the same layout as the main-text reconstruction comparison; CGFC retains its fixed controller parameters.

E.4 Qualitative editing across SD3.5 backbones

Figure 8 compares selected edits on SD3.5 Medium and Large. The four instructions request closed eyes, blue eyes, purple lipstick, and blue eye makeup. Rows use the same source and target prompts across methods and backbones. These examples complement the full-benchmark editing results in Table 4.

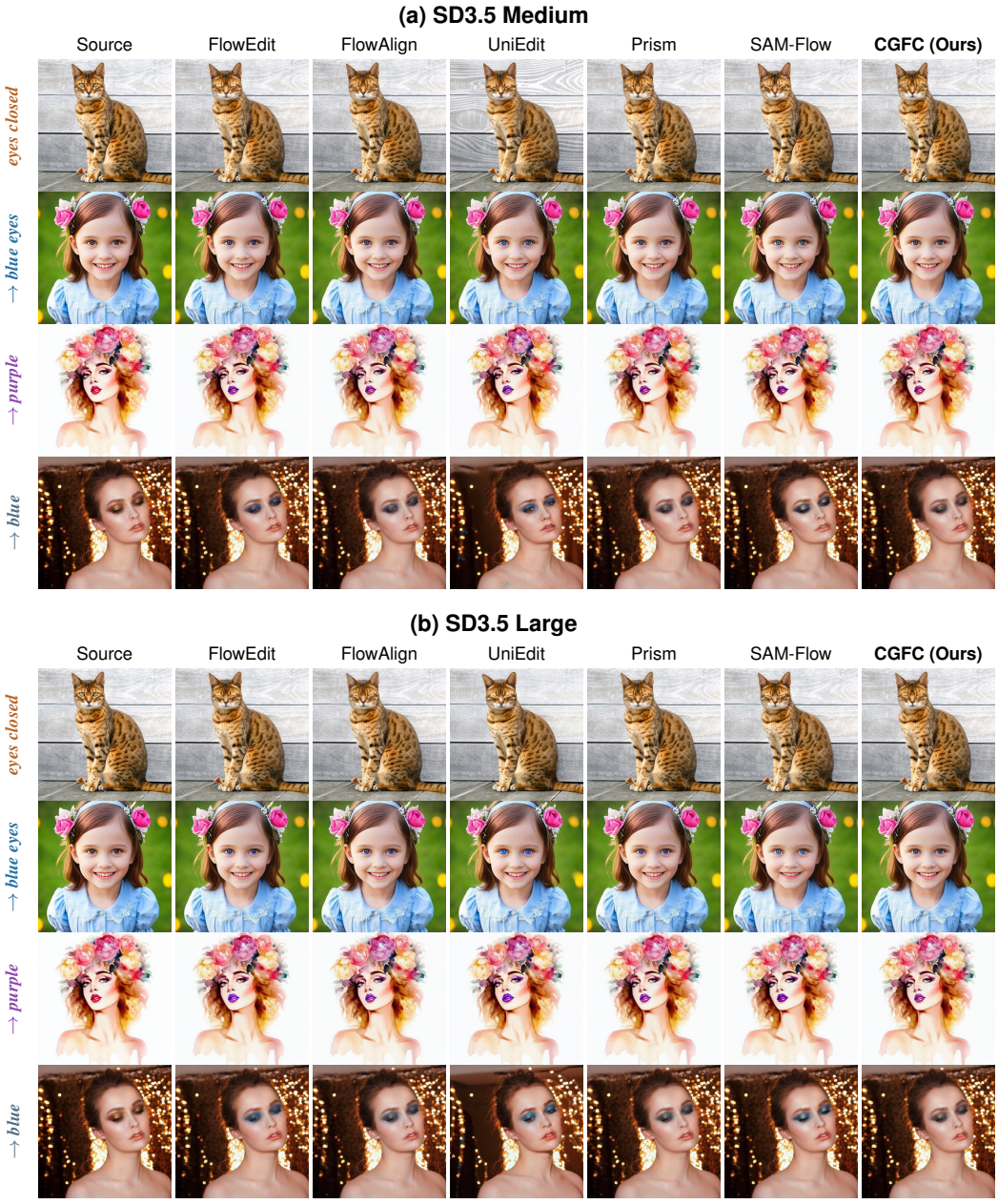

Figure 8: **Qualitative editing on SD3.5 Medium and Large.** Colored row labels indicate the requested changes. Source images and original method outputs are shown with matching task IDs across backbones. CGFC uses the same controller parameters as in the SD3 experiments.

F ANALYTICAL TOY MODELS AND VISUALIZATION

Section 4.2 presents the two main trajectory experiments and their results. This appendix gives their model definitions and evaluation details.

F.1 MODEL DEFINITIONS AND EVALUATION

Perturbation-recovery model. Source and target Gaussians have different means and the same nonuniform diagonal covariance. The first channel at four central positions defines the edit coordinates; their mean difference Δ is zero elsewhere. The other 124 coordinates evaluate content preservation. Both configurations share 32 reference-editing steps, receive a content perturbation at noise level $\sigma = 0.475$, and then recover separately for 32 steps. The complete time grid contains 64 Euler updates from 0.95 to 0. The perturbation is

$$u_{\text{kick}} = \delta(\cos\phi\, e_\ell + \sin\phi\, e_b), \qquad \delta \in \{0.125, 0.25, 0.5\}, \quad \phi \in \{0^\circ, 45^\circ, \ldots, 315^\circ\}, \tag{F.1}$$

where e_ℓ is the unit equal-weight direction in the second channel at the four edit positions, and e_b is its counterpart over the remaining 60 positions. All 24 nonzero deterministic interventions contribute to Figure 2; the zero-perturbation case serves as a diagnostic. In this shared-covariance model, the velocity contrast and support vary with noise level.

State-dependent editing model. The source covariance adds correlations between the edit direction and local/background directions to positive diagonal terms. The target covariance is $\Sigma_t = D\Sigma_s D$, where D differs from identity only at the four edit coordinates. Paired targets are $T(s) = \mu_t + D(s - \mu_s)$, preserving all 124 content coordinates. The covariance difference makes the conditional-velocity contrast affine in the current state.

Full-dimensional evaluation and projections. For perturbation recovery and state-dependent editing, let $\mathcal{J}_{\text{edit}}$ contain the four edit coordinates and $\mathcal{J}_{\text{content}}$ contain the remaining coordinates. We use

$$E_{\text{edit}}(x) = \frac{\|(x - T(s))_{\mathcal{J}_{\text{edit}}}\|}{\|T(s) - s\|}, \qquad E_{\text{content}}(x) = \|(x - s)_{\mathcal{J}_{\text{content}}}\|\,. \tag{F.2}$$

Recovery tasks use the fixed source mean s and $T(s) = s + \Delta$; state-dependent tasks use their sample-specific paired targets. In both settings, paired RMSE is computed as $\|x - T(s)\|/\sqrt{128}$ per endpoint and then averaged arithmetically over all prescribed conditions. Pairings are task-defined evaluation targets. Comparisons are restricted to the two configurations within each setting; absolute errors are not compared across settings.

The three-dimensional display coordinates in the trajectory figures are

$$\alpha = \frac{\langle x - s, T(s) - s\rangle}{\|T(s) - s\|^2}, \qquad h_\ell = \langle x - s, e_\ell\rangle\,, \qquad h_b = \langle x - s, e_b\rangle\,. \tag{F.3}$$

Here α denotes visualization progress and is distinct from the feedback vector a. Recovery uses a fixed projection; the state-dependent experiment uses sample-specific source-aligned projections. Paired panels share their view, ranges, and scales. The plots show the recorded trajectories and endpoints. The thick curves in Figure 3(c) average the projected coordinates at each step.

Endpoint behavior of complete configurations. In the recovery and state-dependent settings, geometric feedback reduces total content drift.

Numerical checks and interpretation. Both settings check the analytical score–velocity relation, residual identity, local norm preservation, restricted Fisher cost, and displacement-feedback orthogonality and bound, and save complete states for replay. In state-dependent editing, reduced total content drift accompanies a background-drift reduction from 0.161 to 0.078.These component-wise measurements locate the observed total-drift reduction primarily in the background coordinates.

G LIMITATIONS AND FUTURE WORK

CGFC represents an exploration of geometric feedback control in latent space. Its conditional Fisher geometry captures local sensitivity to condition discrimination; geometries better suited to editing semantics and content preservation remain to be explored. Our guarantees concern local feedback operations and do not establish full-trajectory contraction or universal endpoint improvement. Empirically, improved source preservation accompanies lower target CLIP scores than Prism. CLIP measures image–text alignment but does not fully characterize editing quality, so this trade-off requires joint assessment with preservation metrics and qualitative results. Image experiments are limited to PIE-Bench and SD3/SD3.5. Future work may extend latent-space geometric control through richer metrics, trajectory-level analysis, and broader generative models.