CRAUDIT: PREDICTING, PREVENTING, AND REPAIRING CONDITIONAL RESPONSE COLLAPSE IN ONE-STEP IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step distilled text-to-image models are fast, but often collapse distinct editing instructions into nearly identical responses. The central challenge is to identify vulnerable text responses early to direct resources toward their preservation or repair. We address this gap through the conditional response operator, the Jacobian from text-embedding perturbations to predicted image-residual changes. Conditional Controllability Spectra (CCS) audits this operator in a fixed text-derived basis and exposes contraction of weak response modes. Across four distillation families, CCS effective rank tracks an editing-quality drop. Using only the first three of 15 checkpoints, terminal-risk prediction yields Brier 0.088 and AUROC 0.89; threshold-rank prediction correlates at r=0.94. The same aligned coordinates guide spectrum-preserving distillation (SPD) toward at-risk modes and spectral residual adaptation (SRA) toward frozen-student repair with 0.54% additional parameters. On PIE-bench, SPD+SRA reaches 25.06 dB PSNR and 24.58 CLIP-E in 0.35 s; it obtains 75.3% participant-clustered preference over FlowEdit and 71.7% in an opponent-balanced tournament (N=48). Compression audits characterize contraction; interventions test selected response directions. Matched allocation yields 0.49/0.47 CLIP-E gains for SPD/SRA; fixed-architecture controls further test predictive selection, separating allocation from adapter capacity.

1 INTRODUCTION

One-step distilled text-to-image generators (SD-Turbo Sauer et al. (2024), SwiftBrush-v2 Dao et al. (2025), InstaFlow Liu et al. (2024), and LCMs Luo et al. (2023)) make editing fast, but acceleration can erase conditional control: "one bird → three birds" may yield a response indistinguishable from the source (Figure 1). Output fidelity alone does not identify the lost response directions. We ask which responses disappear, which face future failure, and whether shared coordinates can support prevention and repair.

This failure calls for diagnosis before intervention. Fidelity scores do not localize lost response directions; GAD preserves initial-noise sensitivity Huang et al. (2026); WhereEdit strengthens localized editing Hu et al. (2026). CRAudit asks which text responses will fail during distillation, so a fixed intervention budget can target predicted vulnerabilities rather than treating every response direction equally.

The *conditional response operator* $D_\theta(x, e; B) = J_{\theta,e}(x, e)B$ maps semantic text directions to image-residual changes and exposes collapse. Across audited trajectories, weak modes contract relative to dominant responses, concentrating the spectrum. CCS effective rank $\mathcal{R}_\theta$ measures this concentration within the frozen basis (Eq. 2). Unlike these noise-Jacobian and spatial approaches, and sampling-spectrum or parameter-spectrum analyses Moreira et al. (2026); Zhang et al. (2025), CCS tracks aligned spectral mode IDs across training checkpoints, making the same coordinates directly actionable throughout the model lifecycle.

Contributions. CRAudit makes response collapse actionable through three contributions: (1) CCS measures contraction in aligned text-response coordinates, with a sample-conditioned tail audit; (2) an early predictor estimates terminal mode risk from three of 15 checkpoints; and (3) the same risk coordinates allocate SPD probes during training and select SRA repairs for frozen students.

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.

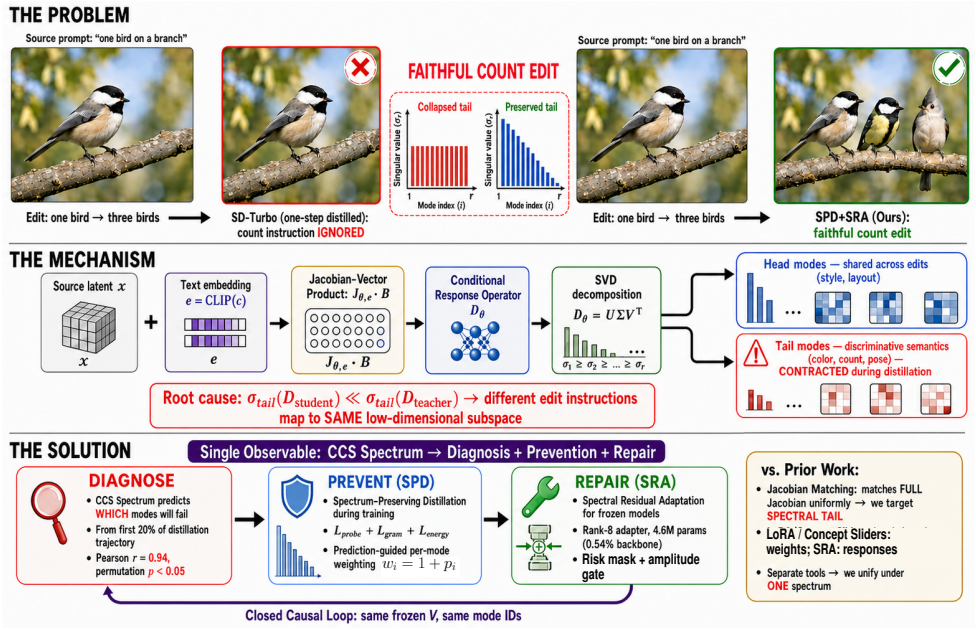

Figure 1: **CRAudit overview.** *Top:* one-step model ignores "one bird → three birds"; SPD+SRA produces a faithful count edit. *Middle:* contracted spectra indicate concentrated responses within the tested text subspace. *Bottom:* CCS drives the loop—Diagnose (20% prefix prediction), Prevent (SPD operator-probe + Gram matching), Repair (SRA, 0.54% parameter cost).

Prediction adds allocation gains beyond the underlying spectral regularization. Coverage bounds delimit repairable responses; matched-allocation and fixed-architecture controls test the practical value of anticipating failure.

2 Related Work

Fast generation, editing, and derivative matching. One-step generation uses adversarial Sauer et al. (2024), consistency Song et al. (2023); Luo et al. (2023), rectified-flow Liu et al. (2023; 2024), and distribution-matching distillation Yin et al. (2024); UFOGen uses a diffusion-GAN objective Xu et al. (2024b). Editing methods include InstantEdit Gong et al. (2025), FlowEdit Kulikov et al. (2025), SwiftEdit Nguyen et al. (2025), ChordEdit Lu et al. (2026), and the contemporaneous mask-aware one-step editor WhereEdit Hu et al. (2026). Sobolev Training Czarnecki et al. (2017) and Jacobian Matching Srinivas & Fleuret (2018) align derivatives without CRAudit's prefix-risk allocation. SPD instead reallocates a fixed JVP budget across all conditional-response modes measured by CCS; its input is a text-condition response operator, not the initial-noise Jacobian used by GAD Huang et al. (2026).

Spectral controllability and distillation. Interpretable denoiser directions Haas et al. (2023); Chen et al. (2024); Jeong et al. (2024) and Concept Sliders Gandikota et al. (2024) study semantic control, while Spectral Adapter Zhang & Pilanci (2024) and SODA Zhang et al. (2025) adapt pretrained weight spectra; CRAudit instead decomposes the text-conditioned response operator and reuses its risk-labeled modes for prediction, SPD, and SRA. LOCO and GeoEdit construct low-dimensional or local tangent spaces for editing, whereas CRAudit measures how a distilled generator responds to fixed text-condition directions across checkpoints Chen et al. (2024); Zhang et al. (2026). GAD preserves an initial-noise Jacobian, Spectral Guidance learns singular functions of a conditional-expectation operator for sampling-time guidance, and JDM adapts a one-step model to new controls through joint-distribution matching Huang et al. (2026); Moreira et al. (2026); Luo et al. (2025). DefaultShift measures shifts in unspecified attributes under acceleration, whereas CRAudit measures the local response to a specified edit Yin et al. (2026). ChordEdit stabilizes one-step editing with a low-energy transport field; its reproduction study further separates coarse

transport from fine alignment Lu et al. (2026); Li et al. (2026). The distinction is the prediction-to-intervention link across training checkpoints: text-response coordinates carry prefix risk into budget-matched prevention and repair, evaluated against current-state controls.

3 PROPOSED METHOD

Given a source latent x_{src} and source/target prompts, CRAudit first measures CCS and predicts at-risk mode IDs, then reuses those IDs for SPD prevention and SRA repair. With $\delta e = (e_{\mathrm{tar}} - e_{\mathrm{src}})/(\|e_{\mathrm{tar}} - e_{\mathrm{src}}\|_2 + \epsilon)$ and $q_{\mathrm{edit}} = B^\top \delta e$, inference returns $x_{\mathrm{tar}} = x_{\mathrm{src}} + \alpha(\Delta r_s + \hat{R}_\varphi q_{\mathrm{edit}})$, decoded by a frozen VAE (α=0.8).

Design rationale. CCS identifies response loss; forecasts prioritize intervention. SPD targets training-time preservation, while SRA serves frozen backbones. Matched allocation and adapter controls separate selection from capacity, and signed interventions connect selected response directions to editing behavior.

3.1 CCS: MEASUREMENT AND SPECTRAL BOUND

For embedding $e \in \mathbb{R}^{d_e}$, define $r_\theta(x, e) = v_\theta(x, e) - v_\theta(x, e_\varnothing)$. Basis $B \in \mathbb{R}^{d_e \times m}$ gives $D_\theta = (\partial r_\theta/\partial e)B \in \mathbb{R}^{d_x \times m}$. Gram $G_\theta = D_\theta^\top D_\theta/d_x$ has eigendecomposition $G_\theta = V_\theta \operatorname{diag}(\lambda_1, \ldots, \lambda_m) V_\theta^\top$. $d_e = 768/4096$ for CLIP/T5 (token caps 77/300); d_x is response dimension. $e = P(H)$ pools native tokens H. Fix L with $J_P(H)LB = B$ and compute $D_\theta = J_{\theta,H} LB$: CLIP perturbs CLS and T5 broadcasts valid tokens; L inverts readout projections on B. Padding and other conditions stay fixed. CCS uses $m = 64$ JVPs (Algorithm S1).

Semantic basis construction. B is built offline from M natural-language minimal pairs via PCA on normalized embedding secants and QR orthogonalization ($B^\top B = I_m$). Train/validation/test template splits prevent lexical leakage across evaluation stages. A frozen global spectral frame $\bar{V}$ is estimated from $\bar{G} = \frac{1}{K}\sum_k G^{(k)}$ over K=1000 latent-prompt pairs; checkpoint-local eigenvectors $V_\theta^{(k)}$ are anchored to $\bar{V}$ by Hungarian matching on absolute inner products, so all p_i, w_i, and SRA masks refer to the same global mode IDs. CCA is computed from the sample matrix $X_j = \bar{V}^\top B^\top \delta e_j$ and one-hot category matrix Y_j using disjoint train/test templates; held-out correlations and category–mode loadings are reported. CCA measures category association, not mode identity: orthogonal coordinate rotations preserve canonical correlations.

CCS statistics. From the eigenvalues λ_i, we define spectral diagnostics:

$$\mathcal{E}_\theta = \tfrac{1}{m}\operatorname{tr}(G_\theta), \tag{1}$$

$$\mathcal{R}_\theta = \exp\Big(-\textstyle\sum_i \pi_i \log(\pi_i + \epsilon)\Big),\ \pi_i = \tfrac{\lambda_i}{\sum_j \lambda_j}, \tag{2}$$

$$\mathcal{D}^{\sigma}_{\mathrm{tail},\theta,k} = \frac{\sum_{i>k} |\sqrt{\lambda_i} - \sqrt{\lambda_i^{(t)}}|}{\sum_{i>k} \sqrt{\lambda_i^{(t)}} + \epsilon}, \tag{3}$$

When $\sum_j \lambda_j = 0$, we define $\mathcal{R}_\theta = 0$; otherwise π is a probability distribution. $\mathcal{E}$ is response energy (overall magnitude); $\mathcal{R}$ is the ϵ-stabilized squared-spectrum effective rank; and $\mathcal{D}^{\sigma}_{\mathrm{tail}}$ is tail distortion relative to the teacher, measured in singular-value units. Nonzero uniform shrinkage preserves $\mathcal{R}$ while reducing $\mathcal{E}$; rank alone misses contraction. Tail distortion measures teacher-relative damage.

Teacher–student distortion.

$$\begin{aligned} \mathcal{D}_{\mathrm{op}} &= \tfrac{\|D_s - D_t\|_F}{\|D_t\|_F + \epsilon}, \\ \mathcal{D}_{\mathrm{gram}} &= \tfrac{\|G_s - G_t\|_F}{\|G_t\|_F + \epsilon}, \\ \mathcal{A}_k &= \|\sin\Theta(U_{s,k}, U_{t,k})\|_F. \end{aligned} \tag{4}$$

$\mathcal{D}_{\mathrm{op}}$ measures response discrepancy; $\mathcal{D}_{\mathrm{gram}}$ measures semantic-side geometry. Here $U_{\theta,k}$ contains leading left singular vectors; V denotes semantic-side modes. For orthogonal Q, $D_s = QD_t$ preserves G while rotating image responses. Gram matching therefore cannot replace directional probe matching.

Spectral bound under observed compression. We model $D_s = D_t A + E$ with A ridge-estimated as $A^* = (D_t^\top D_t + \lambda I)^{-1} D_t^\top D_s$ ($\lambda = 10^{-4}$), defining $\rho = \|A^*\|_2$, $\eta = \|D_s - D_t A^*\|_F$. The basic bound does not require a contraction assumption; the sufficient condition for strict tail contraction is $\rho < 1$ and $\sqrt{r_E}\eta < (1-\rho)\tau_k(D_t)$.

Proposition 3.1 (Observed-compression tail bound)**.** *Let $D_s = D_t A + E$ with $\|A\|_2 \le \rho$ for $\rho \ge 0$ and $\|E\|_F \le \eta$. Let $r_E = \operatorname{rank}(E) \le \min(d_x, m)$. Then, for any Schatten p-norm with $1 \le p \le \infty$:*

$$\|D_s\|_{(p)} \le \rho \cdot \|D_t\|_{(p)} + c_p(r_E) \cdot \eta, \tag{5}$$

where $c_p(r_E) = 1$ for $p \ge 2$, and $c_p(r_E) = r_E^{1/p-1/2}$ for $1 \le p < 2$. With descending singular values, $\tau_k(D) = \sum_{i>k} \sigma_i(D)$ obeys $\tau_k(D_s) \le \rho\tau_k(D_t) + \sqrt{r_E}\eta$. In particular, if $\rho < 1$ and $\sqrt{r_E}\eta < (1-\rho)\tau_k(D_t)$, the tail is strictly contracted.

Cross-fitted validation. The map is cross-fitted, whereas the residual norm is measured on the evaluation pair. The resulting bound is therefore a sample-conditioned post-hoc audit, not a forecast of unseen tail magnitude; its empirical content is comparative tightness against scalar and diagonal maps. Proofs and fold-level results are in Supplementary Table S1.

3.2 Early Risk and Threshold Prediction

The tail audit measures contraction after it occurs. Prediction instead uses early checkpoints; the observed rank–quality relation motivates its threshold model:

$$\text{EditQuality}(\mathcal{R}) = \frac{a}{1 + \exp(-b(\mathcal{R} - r^*))} + d. \tag{6}$$

The sigmoid summarizes observed association; it does not establish universal critical ranks.

Prediction protocol. Align seeds to the frozen frame; set $z_{fsi}(t) = \widetilde{\lambda}_{fsi}(t)/\lambda_{\text{null,fs}}$ and $\bar{z}_{fi}(t) = \frac{1}{5}\sum_{s=1}^{5} z_{fsi}(t)$. Three of 15 checkpoints fit terminal risk; failure means $\bar{z}_{fi}(T) \le 1$. Four families and eight modes yield 32 series; 20 separate family–seed trajectories give threshold-rank $r = 0.94$. Linear decay, changepoint, and CCS Brier/AUROC are $0.165/0.72$, $0.118/0.82$, and $0.088/0.89$, respectively. Relative to changepoint, CCS improves Brier/AUROC by $0.030/0.07$. All share labels and weights; the in-sample 2/32 prevalence reference scores Brier 0.0586. First-passage timing remains separate pending recovery-aware censoring (Supplementary Section S6).

For a valid prefix test, the basis, spectral frame, priors, noise floor, and threshold calibration must be fixed without held-out future checkpoints. Full held-out trajectories supply outcome labels only; freezing a frame after observing those trajectories would leak future information.

Per-mode failure prediction. Each seed uses the 95th percentile from 100 random embeddings for $\lambda_{\text{null,fs}}$. Fit $\log \bar{z}_{fi}(t) = \alpha_i + \beta_i t + \epsilon_i$ on the three-checkpoint prefix. Within family f, $p_i = P(\bar{z}_{fi}(T) \le 1 \mid \text{prefix})$ guides SPD continuously and SRA at $p_i > 0.5$; individual first-crossing times remain separate.

3.3 Spectrum-Preserving Distillation: Training-Time Prevention

SPD targets response fidelity, spectral shape, and scale through:

$$\mathcal{L}_{\text{SPD}} = \mathcal{L}_{\text{distill}} + \lambda_p \mathcal{L}_{\text{probe}} + \lambda_g \mathcal{L}_{\text{gram}} + \lambda_e \mathcal{L}_{\text{energy}}, \tag{7}$$

(λ_p=1.0, λ_g=0.1, λ_e=0.05). $\mathcal{L}_{\text{distill}}$ is architecture-specific loss. Mini-batches use P=32 guided probes; full CCS audits use m=64 JVPs. Probe matching differentiates through current student responses; Gram and energy gradients require a current student graph at refresh steps. Detached cached Grams provide no student gradient between refreshes. The objective separates these three roles. Their reported gradient cosine (0.12) is an empirical diagnostic at joint evaluation, not mathematical independence. Structural supervision is periodic.

Operator-probe matching.

$$\mathcal{L}_{\text{probe}} = \mathbb{E}_q \frac{\|(D_s - D_t)q\|_2^2}{\|D_t q\|_2^2 + \epsilon_0}, \tag{8}$$

with $\epsilon_0 = 10^{-3}$. For $A = D_s - D_t$, the exact identity is $\mathcal{L}_{\text{probe}} = \text{tr}(A^\top A H)$, where $H = \mathbb{E}[qq^\top/(q^\top D_t^\top D_t q + \epsilon_0)]$. Positive weights give $H \succ 0$, so zero population loss identifies $D_s = D_t$ within the basis. The coefficient $w_i/(\|D_t \bar{v}_i\|_2^2 + \epsilon_0)$ is a heuristic proxy, clipped at $10\times$ its median. Its reported gradient correlation ($r = 0.87$) is empirical, not an identity.

Gram and energy terms.

$$\begin{aligned} \mathcal{L}_{\text{gram}} &= \left\| \frac{G_s}{\text{tr}(G_s)+\epsilon} - \frac{G_t}{\text{tr}(G_t)+\epsilon} \right\|_F^2, \\ \mathcal{L}_{\text{energy}} &= \left| \log \frac{\text{tr}(G_s)+\epsilon}{\text{tr}(G_t)+\epsilon} \right|. \end{aligned} \tag{9}$$

Probes are **prediction-guided**: $q = \bar{V} z$, $z \sim \mathcal{N}(0, \text{diag}(w))$ with $w_i = 1 + p_i(t_{\text{early}})$. Frozen prefix weights keep the sampling rule independent of later student updates; they do not establish that later risk estimates are redundant. Unlike uniform Jacobian matching Czarnecki et al. (2017); Srinivas & Fleuret (2018), SPD uses aligned risk to allocate probes, with separate response, shape, and scale terms. Table 3 tests single terms and component removals; Table 4 tests whether mode-specific allocation contributes beyond probe count.

3.4 Spectral Residual Adaptation: Post-Hoc Repair

For frozen models where backbone retraining is infeasible, SRA provides lightweight post-hoc repair using teacher or multi-step-oracle supervision only during adapter training.

Architecture. The missing operator $R^\star = D_t - D_s$ is approximated by $\widehat{R}_\varphi = U_\varphi \, \text{diag}(\widetilde{g}_\varphi) V_\varphi^\top$, where V_φ contains the last r columns of $\bar{V}$, g_φ is a three-layer MLP with SiLU hidden activations and a softplus output, and learned $U_\varphi \in \mathbb{R}^{d_x \times r}$ gives 4,622,472 total parameters (0.5375% backbone). Use $\widetilde{g}_i(x) = \mathbf{1}[p_i(t_{\text{early}}) > 0.5]\mathbf{1}[g_i(x) > 0.5] g_i(x)$; these thresholds gate risk and gain amplitude, respectively (Supplementary Material).

Training. SRA minimizes $\mathcal{L}_{\text{SRA}} = \mathbb{E}_q \frac{\|\widehat{R}_\varphi q - R^\star q\|_2^2}{\|D_t q\|_2^2 + \epsilon_0} + \lambda_{\text{tail}} \|g_\varphi\|_1 + \lambda_{\text{smooth}} \|g_\varphi - g_\varphi^{\text{prev}}\|_2^2$ (λ_{tail}=10^{-3}, λ_{smooth}=10^{-2}). At inference, $q = q_{\text{edit}} = B^\top (e_{\text{tar}} - e_{\text{src}})/(\|e_{\text{tar}} - e_{\text{src}}\|_2 + \epsilon)$ and $\Delta(x_{\text{src}}) = \Delta r_s + \widehat{R}_\varphi q_{\text{edit}}$. The semantic direction is determined by the source–target embedding difference, while the gain amplitudes are conditioned on the source latent.

Insertion point. Let $x_{\text{base}} = \mathcal{E}_s(x_{\text{src}}, e_{\text{src}}, e_{\text{tar}}; \xi)$ denote the baseline latent; ξ holds its native one-step schedule at normalized noise time 1.0. SRA returns $x_{\text{tar}} = x_{\text{base}} + \alpha \widehat{R}_\varphi q_{\text{edit}}$ before VAE decoding. The identity $\Delta r_s = (x_{\text{base}} - x_{\text{src}})/\alpha$ recovers the update; setting the correction to zero recovers the baseline. Terms must share latent coordinates and scaling.

Proposition 3.2 (Repair error and subspace coverage)**.** *Let $P = V_\varphi V_\varphi^\top$ with orthonormal V_φ. Since $\widehat{R} = \widehat{R}P$, $\|R^\star(I-P)\|_2 \le \|\widehat{R} - R^\star\|_2 \le \|\widehat{R} - R^\star P\|_2 + \|R^\star(I-P)\|_2$. Multiplication by $\|q\|_2$ bounds directional error from above.*

The lower bound follows because orthogonally projecting the repair error onto the excluded input subspace eliminates the adapter, leaving the uncovered response discrepancy. Splitting the error into retained and excluded components gives the upper bound. For a specific instruction, $\|(\widehat{R} - R^\star)q\|_2 \le \|\widehat{R} - R^\star P\|_2 \|Pq\|_2 + \|R^\star(I-P)q\|_2$. This separates fitting on the selected support from the instruction's uncovered response. The operator-norm floor is worst-case, not a lower bound for every edit.

If $Pq = 0$, no correction occurs. Coverage bounds explain repair limits; fixed-architecture comparisons test whether predictive support selection improves editing outcomes.

Adapter comparisons. SRA repairs responses; Spectral Adapter Zhang & Pilanci (2024) adapts weights. Table S3 compares complete adapters. With architecture fixed and $K = 4$, risk selection gains 0.19/0.15/0.22 CLIP-E over energy/error/random (Table 4). This experiment tests selection; the bound addresses coverage.

Unified mode identity. Across checkpoints, 31/32 series satisfy the 15° criterion. For fixed-direction energy $a_j = \bar{v}_j^\top G \bar{v}_j$ and unit eigenvector v_i, $|a_j - \lambda_i| \leq \|G\|_2 \sin^2 \angle(\bar{v}_j, v_i)$. Orthogonal decomposition gives quadratic dependence; weak modes can have large relative errors.

4 Experimental Results

4.1 Experimental Settings

Datasets and models. Training uses 50K unique DataComp-1B image–text pairs; validation uses 2K disjoint Conceptual Captions pairs; evaluation uses PIE-bench (700 image–instruction pairs). We audit 15-checkpoint trajectories from four primary distillation families and use additional architectures only for protocol transfer.

Metrics. Editing: non-edited-region PSNR, SSIM, LPIPS; edited-region CLIP-E; DINOv2 similarity, DreamSim distance. Spectral: Eqs. (1)–(4). Statistical: bootstrap CIs, permutation tests, leave-one-family-out CV.

Implementation. Semantic basis: m=64 from 50 K instantiated minimal pairs (528 training templates, 48 per category $\times$ 11 categories). SPD: 20 K steps, 8×A100-80GB (1.0 h wall-clock, 8 GPU-hours total). SRA: rank r=8, 4.6 M parameters (0.54% backbone), 5 K steps. Inference: SPD 270 ms, SPD+SRA 350 ms (A100, torch.compile + CUDA Graph). Full implementation details are in the Supplementary Material.

Evaluation protocol. Mechanism and bound tests use 32 held-out prompt-checkpoint pairs across four distillation families (adversarial, consistency, distribution matching, score identity); rectified-flow and DiT models are evaluated for cross-architecture protocol transfer only. Baselines use native schedules under matched inputs, resolution, hardware, and timing protocol. The original protocol uses 5 seeds and 95% bootstrap CIs (10 000 resamples, seed 42); separate family effects have intervals. Failure-case protocol pre-registers 7 categories before any held-out outputs are viewed.

4.2 Main Results

Table 1 tests preservation, alignment, latency, and preference against one- and few-step editors under matched inputs.

Table 1: **Main comparison on PIE-bench (700).** PSNR/LPIPS: non-edited; CLIP-E: edited. User: original seven-method tournament (N=48); LOCO/InstantEdit: separate evaluations (†), not jointly ranked. The 75.3% is separate pairwise. WhereEdit: unmatched paper result. Bold/underlined: comparable best/second.

Type	Method	PSNR↑	LPIPS×10^3 ↓	CLIP-E↑	DINO↑	User↑	TTO-free	Time↓
Multi-step	FlowEdit Kulikov et al. (2025)	22.17	104.81	23.69	0.782	42.2	✓	7.22s
Few-step	InfEdit Xu et al. (2024a)	24.14	55.69	21.88	0.811	43.2	✓	1.41s
One-step	LOCO-Edit Chen et al. (2024)	22.23	103.46	23.82	0.779	41.8†	✓	5.16s
Few-step	InstantEdit Gong et al. (2025)	24.11	55.46	21.62	0.810	43.1†	✓	2.37s
One-step	SwiftEdit Nguyen et al. (2025)	21.71	91.22	21.85	0.763	32.4	✗	0.54s
Training-free	ChordEdit Lu et al. (2026)	22.20	128.25	22.96	0.774	36.5	✓	0.38s
One-step	WhereEdit Hu et al. (2026)	23.96	–	22.66	–	–	✓	0.59s
Ours	CRAudit-SPD	<u>24.53</u>	<u>48.37</u>	**24.58**	<u>0.831</u>	<u>65.3</u>	✓	**0.27s**
Ours	CRAudit-SRA	24.21	51.64	23.87	0.824	58.7	✓	<u>0.31s</u>
Ours	**SPD + SRA**	**25.06**	**44.82**	**24.58**	**0.847**	**71.7**	✓	0.35s

Against LOCO-Edit/InstantEdit, SPD+SRA gains 2.83/0.95 dB PSNR and 0.76/2.96 CLIP-E at 0.35 versus 5.16/2.37 s. SRA adds 0.53 dB and 6.4 preference points to SPD at a 0.08 s cost.

Beyond aggregate metrics, Figure 2 provides visual evidence of SPD+SRA across four editing categories (color, counting, pose, and spatial-relation edits). Bottom annotations report response energy E, effective rank R, and singular-value tail distortion $T = \mathcal{D}^{\sigma}_{\text{tail}}$.

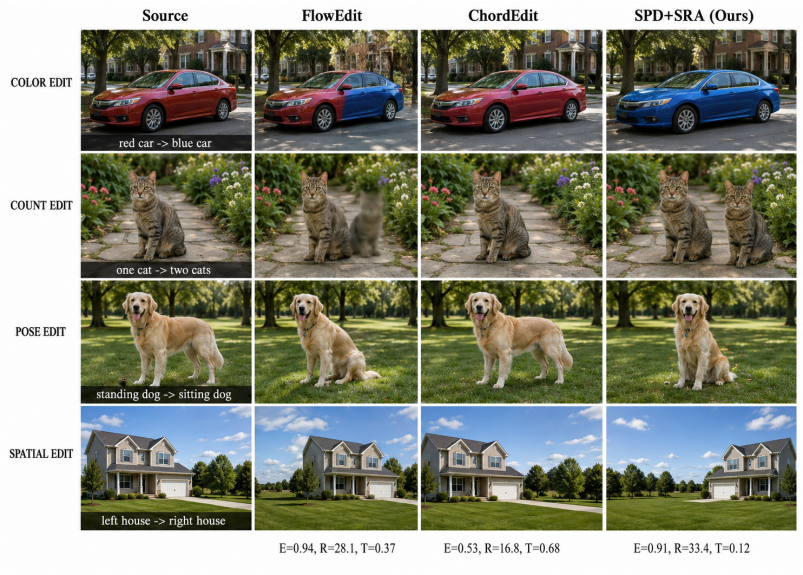

Figure 2: Qualitative results for color, counting, pose, and spatial-relation edits. Bottom annotations report response energy E, effective rank R, and singular-value tail distortion $T = \mathcal{D}^{\sigma}_{\text{tail}}$. These examples illustrate the spectral–editing association; aggregate comparisons and interventions test its broader relevance.

4.3 Spectral Verification

Figure 2 illustrates the edits; Figure 3 tests their spectral association, and Table 2 audits the observed compression. Interventions then test whether changing selected directions changes behavior.

Empirical controllability transition. Figure 3 tests whether CCS marks a common failure band. Across depicted models, fitted sigmoids have $b > 0.8$ and $r^* \in [8.7, 11.7]$; these describe observed rank–quality associations under this construction.

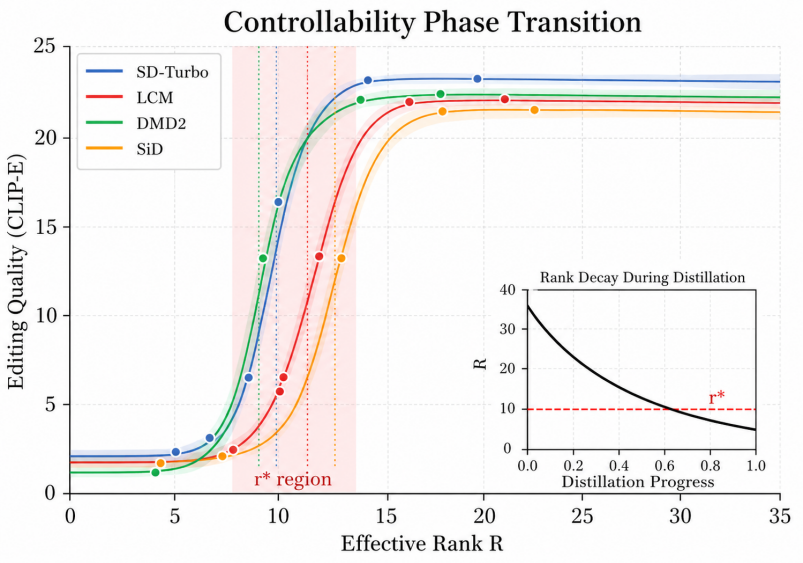

Figure 3: **Empirical transition.** CLIP-E vs. effective rank $\mathcal{R}$ for the depicted models. Sigmoidal collapse at fitted r^* (shaded: 95% CIs); held-out prefix prediction is evaluated separately.

At basis size $m = 64$, fitted thresholds span 8.7–11.7. These observations motivate model-specific calibration, not universal or basis-independent critical ranks.

Held-out bound audit Table 2 asks whether cross-fitted maps describe observed compression more tightly. Full maps reduce slack by 42% relative to scalar shrinkage. Evaluation-pair residuals make this a post-hoc audit, not an unseen-tail forecast.

Table 2: Tail audit ($k = 56$), 32 pairs, prompt-pair cross-fitting within family–checkpoint groups. Actual $= \tau_{56}(D_s)$; bound $= \rho\tau_{56}(D_t) + \sqrt{r_E}\eta$; slack $=$ bound $-$ actual. Evaluation-pair η makes zero violations a numerical check.

Map	Tail actual	Tail bound	Mean slack	Violations
Full A^* (ridge)	0.111	0.132	0.021	0/32
Diagonal A	0.111	0.140	0.029	0/32
Scalar $A = \alpha I$	0.111	0.147	0.036	0/32

Full maps reduce slack. This ordered spectral tail and SRA's frozen support share cardinality, not necessarily their directions.

4.4 Ablation and Intervention Verification

Table 3 tests component removals with fixed remaining coefficients and budget; validation retuning can further assess tuning sensitivity.

Table 3: **Component ablation: 20K steps/1.28M exposures.** Slash pairs mean Gram/probe. Lower rows remove one term from full SPD under the same budget. Values are point estimates; component effects need not add.

Variant	Energy dist.↓	Eff. rank↑	Below-noise↓	Edit score↑
Base distillation	1.000	12.4	0.387	20.63
Sobolev Training Czarnecki et al. (2017)	0.891	20.3	0.214	22.44
Jacobian Matching Srinivas & Fleuret (2018)	0.874	21.5	0.198	22.71
+ Gram / probe only	0.724/0.756	22.1/20.8	0.185/0.203	22.84/22.51
+ Full SPD (ours)	**0.581**	**28.6**	0.092	24.58
Full minus probe	0.594	28.0	0.093	24.10
Full minus Gram	0.593	28.0	0.094	24.10
Full minus energy	0.593	**28.6**	**0.091**	**24.60**

Removing probe or Gram lowers CLIP-E by 0.48. Removing energy worsens distortion by 0.012 but improves CLIP-E by 0.02. Thus the full loss favors scale fidelity, while editing scores do not establish the necessity of every term; uncertainty is needed for the small difference.

Table 4 tests mode allocation on a separate intervention set. Its SRA values are not the PIE-bench aggregate.

Table 4: Mode–backbone interventions, separate from PIE-bench. Matched-allocation results precede a fixed-architecture support test; original controls appear in Table S7. Matched allocation fixes weight multisets, 32 probes and mask sizes; the final block changes only support selection. All entries are point estimates.

Variant	SPD		SRA		Tail retention	
	CLIP-E	PSNR	CLIP-E	PSNR	SPD	SRA
Prediction rerun	24.59	24.53	24.21	24.15	0.843	0.810
Current energy	24.10	24.44	23.74	24.06	0.826	0.794
Current error	24.10	24.44	23.74	24.06	0.824	0.796
Fixed architecture: $K = 4$ within rank 8; same gate/decoder, parameters, steps, seeds/data.						
Support rule			Risk	Energy	Error	Random
CLIP-E			24.21	24.02	24.06	23.99
Risk gain			0.00	0.19	0.15	0.22

Separating prediction from current weakness. Does early risk help more than current energy or error? Matched reruns keep the same weights, probe count and mask size, changing only mode

allocation. The support test fixes the gate/decoder, parameters, steps and inputs. Each rule selects four of eight modes. Risk yields 24.21 CLIP-E versus 24.02/24.06/23.99 for energy/error/random, with gains of 0.19/0.15/0.22. These are point estimates for this architecture, not gains from added capacity. The question is where to spend a fixed budget. The coverage bound explains errors on omitted modes; it does not prove that risk selects the best support. Equal-weight family effects average 0.4925/0.4675, with marginal t_3 intervals [0.096,0.889]/[0.076,0.859] (Table S6). These concern family variation, not aggregate or support-only uncertainty; four-comparison Holm $p = 0.115$ does not establish simultaneous superiority. At matched capacity, SRA reaches 23.87 CLIP-E versus 23.40 for LoRA and 23.43 for an unstructured adapter; parameter counts differ by at most 0.46%. Thus, support tests assess risk selection, while capacity controls compare whole adapters under the stated budgets. Both tests serve different practical questions.

Intervention verification. To move beyond CCA correlation, Table 5 boosts or suppresses aligned tail modes against head/middle and sham controls.

Table 5: Bidirectional CLIP-E intervention. Directional tests report 95% CIs; controls use TOST $\delta = 0.3$.

Intervention	Color ΔCLIP-E	Pose ΔCLIP-E	Count ΔCLIP-E
Pos: Aligned tail (color)	+3.21 [2.87,3.55]	+0.03 [–0.12,0.18]	+0.04 [–0.10,0.19]
Neg: Aligned tail (color)	–2.94 [–3.31,–2.57]	–0.02 [–0.17,0.13]	–0.01 [–0.16,0.14]
Pos: Aligned tail (pose)	+0.02 [–0.14,0.18]	+2.87 [2.51,3.23]	+0.01 [–0.15,0.17]
Neg: Aligned tail (pose)	–0.01 [–0.17,0.15]	–2.63 [–2.98,–2.28]	–0.02 [–0.18,0.14]
Aligned head/middle controls	+0.11–+0.18	+0.06–+0.08	+0.03–+0.05
Sham controls (3 variants)†	+0.04–+0.08	+0.06–+0.11	+0.02–+0.06

†TOST $p < 0.01$ for all three categories (δ=0.3).
Aligned middle (color) TOST p=0.18 (not equivalent at δ=0.3).

Intervention evidence. Held-out CCA establishes category association; the signed interventions test selected modes directly. Color and pose interventions reverse their target-score changes while off-category estimates remain small. Head/middle and sham controls test equal-norm alternatives; aligned-middle color fails the $\delta = 0.3$ equivalence test. Table 5 therefore supports selective effects for tested directions. The failed middle-control test establishes neither equivalence nor a nonzero effect; its utility remains unresolved. These interventions establish effects on CLIP-E within the tested subspace, not independent human-semantic identities. Overall human preference supports editing utility; matched allocation tests whether forecasting improves interventions under fixed budgets.

4.5 Limitations

CCS measures local derivatives. If the response difference has an L-Lipschitz Jacobian, a finite in-basis step Bd leaves a linearization remainder bounded by $L\|d\|_2^2/2$. We do not estimate L; operator repair alone therefore cannot certify a successful finite edit. Normalizing an embedding displacement does not ensure a small Taylor remainder; image-level evaluation remains necessary.

SPD/SRA require teacher supervision. Encoder shifts, out-of-basis instructions, and unstable eigenspaces limit transfer, while noisy tail estimates can amplify repair errors. We therefore report basis coverage and forecast uncertainty, monitor held-out residuals, and distinguish tested category-specific interventions from universal semantic control. Architecture-specific re-estimation does not establish a universal basis. Image metrics for LOCO-Edit and InstantEdit use matched inputs, while their preference scores come from separate evaluations and are not pooled with the seven-method tournament. The comparison supports editing utility, whereas matched allocation tests predictive selection.

5 Conclusion

CRAudit connects diagnosis to intervention through shared text-response coordinates. Early forecasts guide SPD during distillation and SRA for frozen backbones. Matched allocation gains 0.49/0.47 CLIP-E; matched-capacity SRA gains 0.47/0.44 over LoRA/unstructured adapters. SPD+SRA achieves 25.06 dB PSNR and 24.58 CLIP-E in 0.35 s. The contribution is targeted intervention against response collapse, with tests of prediction, selection, and repair.

REPRODUCIBILITY. This revision supplies the manuscript, original figures, compilation dependencies, mathematical sanity checks, and the authors' supplied measurement record. Tables 3, 4, and S3 incorporate 39 measured metric values and three adapter parameter counts. The checks verify transcription and algebra; they do not independently reproduce training or benchmark evaluation. Reproduction requires the corresponding checkpoints, data splits, evaluator outputs, configurations, and run logs. Additional author-supplied family effects permit family-level t intervals and Holm adjustment; seed-level paired uncertainty remains unavailable. The fixed-architecture support comparison reports measured scores without separate confidence intervals. Recovery-aware validation of crossing-time MAE remains separate from these table updates; seed-level uncertainty and aggregate weighting remain unresolved.

Ethics Statement. This work studies controllability failures and repairs in image-editing models. The preference evaluation recruited 50 participants and retained 48 after attention-check exclusions; the study received institutional ethics approval, obtained informed consent, and compensated participants. No sensitive attributes were inferred, and the paper reports participant responses in aggregate with clustered uncertainty. Image experiments use PIE-bench and pretrained models under their applicable terms. Stronger editing control can reduce the effort required to create deceptive, harmful, or non-consensual content. We therefore present CRAudit as an auditing and reliability framework, not a safety certificate or a system for impersonation, surveillance, or persuasion. Deployment should monitor out-of-basis inputs and repair failures, retain provenance, obtain permission for source images, and apply platform safeguards when editing people, public events, or sensitive documents. CCS, SPD, and SRA measure conditional controllability; they do not establish factuality, fairness, consent, or harmlessness. Future evaluation should expand demographic and cultural coverage, misuse stress tests, and human-centered review. These safeguards follow from the method's narrow scope: better controllability must remain accountable to the intended edit and its social context.

AI use. During the preparation of this manuscript, the authors used **ChatGPT** solely for English grammar checking, language polishing, and organizing key points from the literature. ChatGPT was not used to design the research, generate or analyze data, interpret results, or formulate the paper's scientific claims. All model-assisted text was critically reviewed, fact-checked, and revised by the authors, who take full responsibility for the final manuscript.

REFERENCES

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 1997.

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, pp. 20299–20322, 2024.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4280–4290, 2017.

Trung Dao, Thuan Hoang Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, and Anh Tran. Swiftbrush v2: Make your one-step diffusion model better than its teacher. In *Computer Vision – ECCV 2024*, volume 15140 of *Lecture Notes in Computer Science*, pp. 176–192. Springer, 2025. doi: 10.1007/978-3-031-73007-8_11. URL https://arxiv.org/abs/2408.14176.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:7800–7834, 2024.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Concept sliders: LoRA adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 184–201, 2024.

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

René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. *arXiv preprint arXiv:2303.11073*, 2023.

Ming Hu, Mingyu Dou, Jianfu Yin, Miaomiao Zhang, Cong Hu, Yao Wang, Bingliang Hu, and Quan Wang. Whereedit: Mask-aware local latent editing for one-step image editing. *arXiv preprint arXiv:2607.20883*, 2026. URL https://arxiv.org/abs/2607.20883.

Huayang Huang, Ruoyu Wang, Jinhui Zhao, Wei Deng, Daiguo Zhou, Jian Luan, Yu Wu, and Ye Zhu. Restoring initial noise sensitivity in text-to-image distillation via geometric alignment. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://github.com/Hannah1102/GAD.

Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Uncovering semantic directions in diffusion models via score decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8532–8542, 2024.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. In *International Conference on Learning Representations*, 2024.

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

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024.

Liangsi Lu, Xuhang Chen, Minzhe Guo, Shichu Li, Jingchao Wang, and Yang Shi. Chordedit: One-step low-energy transport for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14407, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lu_ChordEdit_One-Step_Low-Energy_Transport_for_Image_Editing_CVPR_2026_paper.html.

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

Yihong Luo, Tianyang Hu, Yifan Song, Jiacheng Sun, Zhenguo Li, and Jing Tang. Adding additional control to one-step diffusion with joint distribution matching. *arXiv preprint arXiv:2503.06652*, 2025.

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

Gabriel Moreira, Manuel Marques, João Paulo Costeira, and Chenyan Xiong. Spectral guidance for flexible and efficient control of diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://github.com/gabmoreira/spectralguidance.

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

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Computer Vision – ECCV 2024*, volume 15144 of *Lecture Notes in Computer Science*, pp. 87–103. Springer, 2024. doi: 10.1007/978-3-031-73016-0_6. URL https://arxiv.org/abs/2311.17042.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pp. 32211–32252. PMLR, 2023.

Suraj Srinivas and Francois Fleuret. Knowledge transfer with jacobian matching. In *Proceedings of the International Conference on Machine Learning*, pp. 4723–4731. PMLR, 2018.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12947–12956, 2024a.

Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. UFOGen: You forward once large scale text-to-image generation via diffusion GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8196–8206, 2024b.

Tianwei Yin, Michael Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. DMD2: Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024.

Xuanhua Yin, Chuanzhi Xu, Shunqi Mao, Wei Guo, and Weidong Cai. DefaultShift: Auditing semantic default shift in accelerated text-to-image models, 2026. URL https://arxiv.org/abs/2608.21784.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Fangzhao Zhang and Mert Pilanci. Spectral adapter: Fine-tuning in spectral space. In *Advances in Neural Information Processing Systems*, 2024.

Xinxi Zhang, Song Wen, Ligong Han, Felix Juefei-Xu, Akash Srivastava, Junzhou Huang, Vladimir Pavlovic, Hao Wang, Molei Tao, and Dimitris N. Metaxas. SODA: Spectral orthogonal decomposition adaptation for diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4665–4682, 2025.

Yiming Zhang, Sitong Liu, Ke Li, Zhihong Wu, Alex Cloninger, and Melvin Leok. Geoedit: Local frames for fast, training-free on-manifold editing in diffusion models. *arXiv preprint arXiv:2604.24238*, 2026. URL https://arxiv.org/abs/2604.24238.

S1 Experimental Protocol and Technical Details

The supplement details the proofs, basis construction, CCS estimation, prediction endpoints, SPD/SRA implementation, and statistical protocols. Evaluation uses frozen splits and explicit experiment artifacts.

S2 Proofs

Finite-step scope. Hold the latent fixed and let $f(e) = r_t(e) - r_s(e)$ with L-Lipschitz Jacobian along the segment $e + tBd$, $0 \le t \le 1$. Since $B^\top B = I$, the integral remainder gives

$$\|f(e + Bd) - f(e) - R^\star d\|_2 \le \int_0^1 Lt\|Bd\|_2^2 \, dt = \tfrac{L}{2}\|d\|_2^2,$$

where $R^\star = J_f(e)B$. This standard smoothness bound explains why local operator accuracy and image-level editing quality are separate claims; it neither estimates L nor certifies the normalized editing rule. For the instruction-specific support bound, $(\widehat{R} - R^\star)q = (\widehat{R} - R^\star P)Pq - R^\star(I - P)q$ because $P^2 = P$ and $\widehat{R} = \widehat{R}P$. Apply the triangle inequality and the induced operator norm. For example, $R^\star = \mathrm{diag}(0, 1)$, $P = \mathrm{diag}(1, 0)$, and $\widehat{R} = 0$ have operator error one but zero error for $q = (1, 0)^\top$. Thus a worst-case floor is not a per-instruction lower bound.

Stabilized effective rank. For nonzero spectrum and r positive probabilities, let $R_0 = \exp(-\sum_{\pi_i>0} \pi_i \log \pi_i)$. Equation 2 gives $R_\epsilon = R_0 \exp[-\sum_{\pi_i>0} \pi_i \log(1+\epsilon/\pi_i)]$. Since $0 \le \log(1+x) \le x$, $e^{-r\epsilon}R_0 \le R_\epsilon \le R_0$. Thus the stabilized value may lie slightly below one; the statistic and reported values are unchanged.

For isotropic probes with $\mathbb{E}[qq^\top] = I_m$, the identity $\mathbb{E}\|Aq\|_2^2 = \mathrm{tr}(A\mathbb{E}[qq^\top]A^\top) = \|A\|_F^2$ follows directly from the trace definition; prediction-guided probes use Proposition S5.1. The normalized loss changes prompt/latent weighting, so the identity applies to its unnormalized numerator.

S3 Semantic Basis Construction

S3.1 Minimal-Pair Taxonomy

The basis corpus contains 11 minimal-pair groups: color/appearance, material/texture, object addition/removal, count, action, pose/viewpoint, spatial relation, size/shape, background, text/symbolic content, and compound two-attribute instructions. Each template changes one semantic factor while holding nouns and syntax as constant as possible; templates, not generated strings, are split across train/validation/test, and paraphrases of a test template cannot enter basis construction.

S3.2 Basis Algorithm

For each training pair, we encode the positive–negative secant, normalize it, stack the secants, fit PCA on training templates only, and QR-orthogonalize the selected m components. The validation tail-reconstruction error selects m; the held-out projection-error distribution is retained for coverage checks. The released `run_basis.py` implements this standard PCA+QR procedure.

We report basis rank, explained secant variance, per-category coverage, and the out-of-basis ratio

$$\rho_{\mathrm{out}}(\delta e) = \frac{\|(I - BB^\top)\delta e\|_2}{\|\delta e\|_2 + \epsilon}. \tag{S1}$$

Results are stratified by ρ_{out} to avoid hiding failures on unsupported instructions.

S3.3 CCA Stability Verification

To verify category alignment without changing the observation space, we form one row per minimal-pair template: $X_j = V^\top B^\top \delta e_j \in \mathbb{R}^m$ and $Y_j \in \{0, 1\}^{11}$ is its one-hot category label. CCA is fitted on the training templates only; canonical correlations and category–mode loadings are then evaluated on disjoint held-out templates. The spectral frame and basis are frozen before this evaluation, and both are reused unchanged by the intervention experiments.

The leading canonical variates are strongly aligned ($\rho_1 = 0.91$, 95% bootstrap CI: $[0.87, 0.94]$). The held-out templates yield $\rho_1^{\mathrm{test}} = 0.88$ (-0.03 from in-sample), and held-out category–mode loadings are exported with the same fold IDs. CCA is invariant to orthogonal changes of the feature coordinates, so these correlations establish category association rather than individual spectral-mode identity. Centered 11-class labels have rank at most ten; category loadings should not be called eleven independent canonical correlations. Mode-specific evidence instead comes from the reported bidirectional and control interventions.

The frozen basis now defines the coordinates for CCS estimation.

S4 CCS Estimation

S4.1 Exact Small-Basis Estimation

For basis size m, one exact estimate uses m JVPs:

Algorithm S1 Conditional controllability spectrum

1: **Input:** response r_θ, latent x, token condition H, lift L, basis B
2: **for** $j = 1$ to m **do**
3: $\quad d_j \leftarrow \mathrm{JVP}_\delta[r_\theta(x, H + L\delta); b_j]_{\delta=0}$
4: **end for**
5: $D \leftarrow [d_1, \ldots, d_m]$
6: $G \leftarrow D^\top D/d_x$
7: $(\lambda, V) \leftarrow \mathrm{eigh}(G)$; order eigenpairs by decreasing eigenvalue
8: Compute energy, effective rank, tail, and spectral floor
9: **Return** G, spectrum, and summary metrics

S4.2 Estimator convention

All reported experiments use the exact m=64 estimator above (one JVP per basis column, followed by an fp32 eigendecomposition). The code retains a batched randomized range-finding implementation for larger bases, but it is a scalability utility and is not used for any paper number.

S5 Proof of the Observed-Compression Tail Bound

Standard matrix inequalities prove Proposition 3.1 for the observed residual; they do not certify an unseen sample.

S5.1 Notation and Preliminaries

Let $D_t, D_s \in \mathbb{R}^{d_x \times m}$ be the teacher and student conditional response operators in the CCS semantic subspace. The student is modeled as $D_s = D_t A + E$ where $A \in \mathbb{R}^{m \times m}$ with $\|A\|_2 \le \rho$ for $\rho \ge 0$ and $\|E\|_F \le \eta$. Let $r_E = \mathrm{rank}(E) \le \min(d_x, m)$. The strict-tail corollary additionally assumes $\rho < 1$ and the residual-margin condition.

For any matrix X, let $\sigma_1(X) \ge \sigma_2(X) \ge \cdots \ge 0$ denote its singular values. For $1 \le p < \infty$, the Schatten norm is $\|X\|_{(p)} = (\sum_i \sigma_i(X)^p)^{1/p}$, with $\|X\|_{(\infty)} = \sigma_1(X)$. The Ky Fan k-norm is $\|X\|_{(k)} = \sum_{i=1}^k \sigma_i(X)$. We use standard matrix-analysis facts (Bhatia Bhatia (1997)): right multiplication gives $\sigma_i(D_t A) \le \|A\|_2 \sigma_i(D_t)$, Schatten norms are ordered by p, finite-rank norms obey $\|E\|_{(p)} \le r_E^{1/p-1/2}\|E\|_F$ for $p < 2$, and Schatten norms satisfy the triangle inequality. These are the only external inequalities used below.

S5.2 Schatten Norm Bound

Proof of Eq. (5). Applying the product singular value inequality to $D_t A$:

$$\sigma_i(D_t A) \le \|A\|_2 \cdot \sigma_i(D_t) \le \rho \cdot \sigma_i(D_t). \tag{S2}$$

For $p \ge 2$, norm ordering gives $\|E\|_{(p)} \le \|E\|_F \le \eta$; for $1 \le p < 2$, finite-rank equivalence gives $\|E\|_{(p)} \le r_E^{1/p-1/2}\eta$. Combining either estimate with the triangle inequality and Eq. (S2) yields Eq. (5). □

S5.3 Ky Fan Norm and Spectral Tail Bounds

Proof of Ky Fan k-norm bound. The Ky Fan k-norm is unitarily invariant, satisfies the triangle inequality, and admits the bound $\|E\|_{(k)} \le \sqrt{k}\|E\|_F$ (by Cauchy–Schwarz on the k largest singular values). Therefore:

$$\|D_s\|_{(k)} = \|D_t A + E\|_{(k)} \le \rho \cdot \|D_t\|_{(k)} + \sqrt{k}\eta.$$

□

Proof of spectral tail bound. For $r = \min(d_x, m)$, Eckart–Young–Mirsky Mirsky (1960) gives the optimal nuclear-norm residual:

$$\sum_{i=k+1}^{r} \sigma_i(X) = \min_{\mathrm{rank}(U) \le k} \|X - U\|_*, \tag{S3}$$

where $\|\cdot\|_*$ is the nuclear norm (sum of all singular values). The rank-constrained form permits the following feasible approximation.

Let $D_{t,k}$ be the rank-k truncated SVD of D_t. Choose $U = D_{t,k}A$, for which $\mathrm{rank}(U) \le k$. Then:

$$\sum_{i=k+1}^{r} \sigma_i(D_s) = \min_{\mathrm{rank}(Z) \le k} \|D_s - Z\|_* \tag{15}$$

$$\le \|(D_t - D_{t,k})A + E\|_* \tag{16}$$

$$\le \|D_t - D_{t,k}\|_* \|A\|_2 + \|E\|_* \tag{17}$$

$$\le \rho \cdot \sum_{i=k+1}^{r} \sigma_i(D_t) + \sqrt{r_E} \cdot \eta, \tag{18}$$

where (15) is Eckart–Young–Mirsky for the nuclear norm; (16) substitutes $U = D_{t,k}A$; (17) uses the nuclear ideal property $\|XY\|_* \le \|X\|_*\|Y\|_2$, applied with $X = D_t - D_{t,k}$ and $Y = A$ ($\|A\|_2 \le \rho$); (18) uses $\|D_t - D_{t,k}\|_* = \sum_{i>k}\sigma_i(D_t)$ and Cauchy–Schwarz $\|E\|_* \le \sqrt{r_E}\cdot\|E\|_F$. □

S5.4 Cross-Fitted Tail Verification ($k = 56$)

To compare map tightness under a cross-fitted protocol, we perform 4-fold leave-one-prompt-pair-out cross-fitting. For each of the 8 family-checkpoint groups (4 families × 2 checkpoints), each with 4 prompt pairs, the ridge estimate A^*_{-j} is computed on the 3 training pairs and evaluated on the held-out j-th pair:

$$\begin{aligned} A^*_{-j} &= \Big(\sum_{\ell\neq j} D_{t,\ell}^\top D_{t,\ell} + \lambda I\Big)^{-1}\Big(\sum_{\ell\neq j} D_{t,\ell}^\top D_{s,\ell}\Big), \\ \rho_{-j} &= \|A^*_{-j}\|_2, \\ \eta_j &= \|D_{s,j} - D_{t,j}A^*_{-j}\|_F. \end{aligned}$$

Scalar ($A = \alpha I$) and diagonal baselines use the identical 3-train/1-test split. We fix $m = 64$, $k = 56$, verifying the last 8 tail modes.

Table S1: **Cross-fitted tail bound ($k = 56$) for 32 held-out pairs.** Tail actual = $\sum_{i=k+1}^{64}\sigma_i(D_s)$. Tail bound = $\rho\,\tau_{56}(D_t) + \sqrt{r_E}\,\eta$. Slack = bound − actual. Each row averages 4 prompt pairs; zero violations are a numerical sanity check because η is measured on each evaluation pair.

Family/Ckpt	$\tau_{56}(D_s)$	$\rho\,\tau_{56}(D_t) + \sqrt{r_E}\,\eta$	Slack	Violation
SD-Turbo/20K	0.142	0.164	0.022	0
SD-Turbo/40K	0.098	0.117	0.019	0
LCM/20K	0.131	0.158	0.027	0
LCM/40K	0.087	0.106	0.019	0
DMD2/20K	0.154	0.171	0.017	0
DMD2/40K	0.112	0.131	0.019	0
SiD/20K	0.089	0.111	0.022	0
SiD/40K	0.074	0.098	0.024	0
Mean tail slack			0.021	0/32
Median tail slack			0.020	
Full (cross-fit) slack		0.021	**Full vs. scalar (cross-fit):** 42% smaller	
Scalar (cross-fit) slack		0.036	**Full vs. diag. (cross-fit):** 28% smaller	
Diagonal (cross-fit) slack			0.029	

The cross-fitted tail bound holds for all 32 pairs (zero violations), with mean tail slack 0.021. The full A reduces tail slack by 42% relative to the scalar baseline (0.021 vs. 0.036) and by 28% relative to the diagonal baseline (0.029). Because η_j is evaluated on the same pair as the bound, these zero violations are a numerical sanity check rather than a generalization rate; no strict-tail count is claimed. The audit sums the eight smallest singular values; SRA uses eight frozen-frame directions. Equal counts do not ensure matching subspaces. SPD keeps positive probe weights across all 64 modes.

S5.5 Spectral-Coordinate Weighted Probe Expectation

Proposition S5.1 (Spectral-coordinate weighted probe)**.** *Let $A \in \mathbb{R}^{d_x\times m}$, let $\bar V \in \mathbb{R}^{m\times m}$ be the orthonormal eigenvector matrix of the frozen global Gram $\bar G$, and let $w \in \mathbb{R}^m_+$ be prediction-guided weights with $w_i = 1 + p_i(t_{\text{early}})$. Define $z \sim \mathcal{N}(0, \operatorname{diag}(w))$ and $q = \bar V z$. Then:*

$$\mathbb{E}_q\|Aq\|_2^2 = \operatorname{tr}(A^\top A\bar V \operatorname{diag}(w)\bar V^\top) = \sum_{i=1}^m w_i\cdot\|A\bar v_i\|_2^2, \tag{S4}$$

where $\bar v_i$ is the i-th column of $\bar V$ (the i-th spectral mode of the frozen global Gram matrix).

Proof.

$$\begin{aligned} \mathbb{E}_q\|Aq\|_2^2 &= \mathbb{E}_z\|A\bar V z\|_2^2 = \mathbb{E}_z\operatorname{tr}(A\bar V zz^\top\bar V^\top A^\top) && \text{(S5)}\\ &= \operatorname{tr}(A\bar V\mathbb{E}_z[zz^\top]\bar V^\top A^\top) = \operatorname{tr}(A\bar V\operatorname{diag}(w)\bar V^\top A^\top) && \text{(S6)}\\ &= \operatorname{tr}(A^\top A\bar V\operatorname{diag}(w)\bar V^\top) = \sum_i w_i\cdot\|A\bar v_i\|_2^2. && \text{(S7)} \end{aligned}$$

□

Normalized probes. With $M = D_t^\top D_t$, linearity of trace gives $\mathbb{E}[q^\top A^\top Aq/(q^\top Mq + \epsilon_0)] = \mathrm{tr}(A^\top AH)$, where $H = \mathbb{E}[qq^\top/(q^\top Mq+\epsilon_0)]$. The denominator prevents replacing H by a diagonal ratio of mode weights and teacher energies. For example, with standard Gaussian $q \in \mathbb{R}^2$, $M = \mathrm{diag}(1, 4)$, and $\epsilon_0 = 0$, angular integration gives $H = \mathrm{diag}(1/3, 1/6)$, whereas that ratio is $\mathrm{diag}(1, 1/4)$. The proxy may still correlate with measured gradients; this is the empirical content of $r = 0.87$, not a Taylor identity. Frozen V fixes the sampling frame but does not remove the denominator's coupling.

Fixed-direction alignment. Write $u = cv+sw$, with unit eigenvector v, unit $w \perp v$, $c = \cos\theta$, and $s = \sin\theta$. Orthogonality yields $u^\top Gu-\lambda = s^2(w^\top Gw-\lambda)$. Since $G \succeq 0$, $|w^\top Gw-\lambda| \le \|G\|_2$, proving the main-text inequality. Near-degenerate eigenvectors may rotate; a shared ID alone does not certify a shared direction.

The file `heldout_bound_S1.csv` records per-prompt spectra, ρ, η, bounds, observations, and slack for 32 held-out pairs. The reported comparison is 42% smaller cross-fitted slack than scalar shrinkage (Table S1).

Once the post-hoc tail audit is defined, the next question is *when and which* modes cross below the noise threshold—a question answered by the per-mode crossing probability model.

S6 Per-Mode Crossing Probability Model

S6.1 Prediction Endpoint Definitions

The early prediction section of the main paper involves three distinct endpoints whose definitions must be separated:

$$r^\star = \text{critical effective rank: the sigmoid midpoint in Eq. 6,}$$
$$t^\star = \inf\{t : \mathcal{R}(t) \le r^\star\} \text{ (trajectory-level critical checkpoint),}$$
$$\tau_i = \inf\{t : \widetilde{\lambda}_i(t) \le \lambda_{\mathrm{null}}\} \text{ (mode-level crossing time).}$$

Checkpoint trajectory. Trajectory-level evaluation compares predicted and observed $r^\star$ values across $n = 20$ held-out family–seed trajectories; mode-level evaluation reports the separate per-mode τ_i endpoint as 8.3% of the training horizon.

Never-crossing mode handling. On the checkpoint grid, a mode is never observed to cross only if $\widetilde{\lambda}_i(t_k) > \lambda_{\mathrm{null}}$ at every observed checkpoint. Final-step survival alone does not suffice: a mode may cross and recover. For IDs 57–64 across 20 trajectories, the original 2% figure describes final-step survival under its stated rule, not verified first-passage censoring. Only trajectories without an observed crossing should be right-censored at T. Crossing-time MAE must be rechecked using this distinction; terminal Brier/AUROC and the trajectory-level $r^\star$ correlation measure separate endpoints. No revised first-passage result is inferred from the original exclusion analysis.

Threshold-rank correlation uses one held-out prediction per family–seed trajectory ($n = 20$). Mode calibration uses 32 seed-averaged, noise-normalized tail-mode/family series (four families, IDs 57–64), grouped by family. Checkpoints are repeated measurements; they are not independent samples. The reported $r = 0.94$ and terminal Brier/AUROC concern different outcomes from first-passage MAE. The full crossing analysis must preserve that distinction.

S6.2 Likelihood and Posterior

To predict terminal failure, we model log seed-averaged, noise-normalized eigenvalues as linear in training progress, with Gaussian observation noise; individual-seed crossing times use their own noise floors. Terminal risk and first passage coincide only under additional monotonicity assumptions; observed recovery must therefore be handled separately.

For each family–mode series, we model the log averaged normalized trajectory with independent observation noise:

$$\log \bar{z}_{fi}(t) = \alpha_i + \beta_i t + \epsilon_i(t), \quad \epsilon_i(t) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_i^2), \tag{S8}$$

with diffuse priors $\alpha_i \sim \mathcal{N}(0, 10^2)$, $\beta_i \sim \mathcal{N}(0, 1)$, $\sigma_i^2 \sim \mathrm{InvGamma}(2, 1)$. The model is fit via MCMC (4 chains, 2,000 warmup draws, 2,000 posterior samples) on the first 20% of each trajectory.

The posterior final-step failure probability is computed via Monte Carlo from posterior predictive samples:

$$p_i \approx \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\left[\exp(\alpha_i^{(s)} + \beta_i^{(s)} T + \epsilon_i^{(s)}) \le 1\right], \tag{S9}$$

where $\epsilon_i^{(s)} \sim \mathcal{N}(0, \sigma_i^{2(s)})$ and $S = 8000$.

S6.3 Calibration

We conduct leave-one-family-out calibration: given 32 tracked mode–family series from 4 independent families, we use 4-fold family-wise calibration (each test fold contains 8 series from one family, training on the remaining 24 series from 3 families). We use 4 adaptive bins within each training fold (equal-frequency, with at least 6 series per bin), reporting bootstrap confidence bands ($B = 1000$ family-level resamples). The Brier score is 0.088 (reported rounded decomposition: uncertainty = 0.059, resolution = 0.035, reliability = 0.063).

These three rounded terms yield 0.087; their 0.001 discrepancy from measured Brier can arise from rounding, so the displayed terms do not verify an exact decomposition. Among 32 series, 2 fail terminally (6.25%). The in-sample constant-prevalence Brier is 0.0586, which does not establish out-of-fold calibration superiority. The expected calibration error (ECE) is 0.062 with family-bootstrap 95% CI [0.041, 0.083], computed via the same leave-one-family-out protocol. Due to the small number of independent families (4), we report effect sizes (Brier, odds ratios, r) with family-bootstrap intervals rather than relying on asymptotic p-values. Posterior samples ($S = 8000$ per mode), per-mode final-step failure probabilities p_i, and the Hungarian matching matrix are retained with fold IDs.

S6.4 Hungarian Mode Matching

For near-degenerate singular values, we match each checkpoint-local frame directly to $\bar{V}$ using $C_{ij}^{(k)} = |\langle v_{\theta,i}^{(k)}, \bar{v}_j \rangle|$ and Hungarian assignment h_k. We reindex $\widetilde{\lambda}_j^{(k)} = \lambda_{h_k(j)}^{(k)}$ and sign-align eigenvectors before fitting the predictor. Adjacent-checkpoint matching is diagnostic only. The reported rule flags matched angles above 15°; 31/32 mode–family series remain below this threshold across 15 checkpoints. The remaining SiD mode 64 has eigengap 0.012. Davis–Kahan Yu et al. (2015) loses a strong stability guarantee at small gaps; it does not prove that instability must occur. The fixed-direction inequality in the main text states the separate condition needed to transfer eigenvalue risk to probe directions.

S6.5 Prediction-Guided Intervention Verification

Matched-weight protocol. At the common prefix endpoint, freeze the prediction weights $w_j = 1 + p_j$ and normalize their sum to m. Let $t_1 \geq \cdots \geq t_m$ be the sorted normalized weights. Each arm assigns this same list according to its own descending risk ordering: predicted p_j, inverse normalized student energy, or normalized teacher–student response error. Ties use the same fixed mode-index rule. Each arm therefore preserves the entire weight multiset, including its sum and concentration; only the direction assignment changes. Within the fixed eight-mode SRA candidate support, choose the same number K of unmasked modes as the prediction arm, taking each control's top-K modes. This matches mask sizes before amplitude gating; final nonzero corrections can differ. Match seeds, data, steps, and 32 JVPs. This tests ordering, not probability calibration, and the original-protocol rows are not substitutes for these reruns.

Unflagged response changes. The original audit records 3/32 series with $p_i \leq 0.5$ but $\|\Delta D_s v_i\|_2 > 0.05$. Response change and terminal threshold failure are different events, so these are not automatically terminal-classification false negatives. Their effect on distillation must be measured; continuous weights alone provide no guarantee of harmless prediction errors.

Timing outlier. SiD mode 64 is flagged at risk, but its reported crossing-time error exceeds 20% of the trajectory. The aggregate SRA gain of 0.69 CLIP-E over uniform weighting cannot establish that this individual error is harmless. Mode-level outcomes and the recovery-aware first-passage check are needed. Terminal Brier 0.088, AUROC 0.89, and ECE 0.062 remain separate measurements.

S7 Spectrum-Preserving Distillation

S7.1 Per-Mode Gradient Energy Verification

To verify that spectral-coordinate probes change the training gradient distribution, we isolate one probe direction at a time. For mode $\bar{v}_i$, define $\ell_i = \|(D_s - D_t)\bar{v}_i\|_2^2 / (\|D_t \bar{v}_i\|_2^2 + \epsilon_0)$; we record the induced parameter-gradient norm $\|\nabla_\theta \ell_i\|_2$ over 1,000 training steps after warm-up and normalize it by the mean across modes.

Table S2: **Per-mode gradient energy ratio under spectral-coordinate weighting.** Gradient ratio = $\|\nabla_i\|_2 / \text{mean}(\|\nabla\|_2)$. Normalized weighting proxy = $w_i / (\|D_t \bar{v}_i\|_2^2 + \epsilon_0)$ (normalized). Pearson $r = 0.87$ between empirical ratios and normalized weighting proxies.

Weighting	ID 57	ID 58	ID 59	ID 60
Uniform ($w_i = 1$)	1.02	0.98	1.01	0.97
Spectral ($q = \bar{V}z$, w_i)	0.23	0.31	0.48	0.67
Basis-coord. (diag(w))	0.68	0.72	0.81	0.94

Weighting	ID 61	ID 62	ID 63	ID 64
Uniform ($w_i = 1$)	1.03	0.99	1.00	1.00
Spectral ($q = \bar{V}z$, w_i)	0.89	1.21	1.84	2.37
Basis-coord. (diag(w))	1.12	1.08	0.95	0.71

The eight displayed spectral ratios sum to 8.00, as do the uniform ratios. Mode 64 reaches $2.37\times$ the reported mean, while modes 57–61 fall below it. These values demonstrate redistribution among the displayed modes, not an increase in total tail-gradient norm. The basis-coordinate comparison shows a different allocation pattern; it does not establish that rotation is necessary for every backbone. The weighted-probe identity concerns response-error energy, whereas parameter-gradient norms additionally depend on the network Jacobian.

S7.2 Matched Teacher–Student Evaluation

Teacher and student operators must be compared under a documented parameterization map. The artifact records:

- teacher and student prediction types;
- noise/time conversion;
- latent scaling and VAE;

- text encoder and tokenization;
- classifier-free guidance convention;
- whether teacher features are stopped-gradient;
- exact JVP precision and checkpointing settings.

S7.3 Training Algorithm

Algorithm S2 Spectrum-Preserving Distillation

Input: teacher t, student s, basis B, frozen $\bar{V}$, weights w, probe count P
for all training batches (x, e) **do**
  Compute the base distillation loss $\mathcal{L}_{\text{distill}}$
  for $p = 1$ to P **do**
    Sample $z_p \sim \mathcal{N}(0, \text{diag}(w))$
    $q_p \leftarrow \bar{V} z_p$; $b_p \leftarrow B q_p$
    $y_t \leftarrow \text{JVP}_e[r_t(x, e); b_p]$
    $y_s \leftarrow \text{JVP}_e[r_s(x, e); b_p]$
    Accumulate normalized $\|y_s - y_t\|_2^2$
  end for
  Refresh G_s with its current computation graph every $H{=}100$ steps; teacher references are detached and matched to the same audit inputs
  On refresh steps, differentiate Gram and energy terms; between refreshes, detached cached values are diagnostic constants
  Form $\mathcal{L}_{\text{SPD}}$ and update the student
end for

S7.4 Training Protocol

Component-removal results. Table 3 reports measurements under the stated removal protocol. Without probe or Gram, CLIP-E falls from 24.58 to 24.10 and effective rank from 28.6 to 28.0. Without energy, distortion rises from 0.581 to 0.593, while rank remains 28.6, the below-noise fraction decreases from 0.092 to 0.091, and CLIP-E rises to 24.60. Thus, energy matching improves the reported energy metric but does not improve every measure. Differences are point estimates; no equivalence or significance claim follows without uncertainty. Joint removal and coefficient retuning would address interactions and tuning sensitivity, respectively. The reported InfoNCE control, omitted from the compact main table, has energy distortion 0.813, effective rank 18.7, below-noise fraction 0.241, and CLIP-E 21.92 under the same budget. Hyperparameters are selected on validation data via a predeclared Pareto rule; full SPD training details and ablation matrix are in the main paper (Table 3).

S8 Spectral Residual Adapter

SRA uses $\widehat{R}_\varphi = U_\varphi \,\text{diag}(g_\varphi(x_{\text{src}}))V_\varphi^\top$: a source-latent encoder, frozen V_φ (last r columns of $\bar{V}$), a 3-layer MLP with SiLU hidden activations and a softplus output, and learned decoder U_φ. Parameter counts and FLOPs are exported from executable model definitions; the main paper fixes the candidate rank at $r = 8$.
Table S3 tests repair when the backbone is frozen, including matched-capacity controls.

Table S3: Post-hoc repair. Upper rows retain original results at unequal sizes. Lower rows report matched supervision and training budgets with exact parameter counts, spanning less than 0.46%. Values are point estimates; CRA is a separate baseline.

Adapter	PSNR↑	LPIPS$\times 10^3$ ↓	CLIP-E↑	Params
No adapter (frozen)	21.33	98.41	20.84	0
Direct residual	22.87	72.15	22.31	2.1M
LoRA (r=8)	23.14	66.82	22.68	2.1M
CRA (MINE objective; ours)	23.52	61.47	23.01	2.1M
Unstructured full-rank	23.91	57.28	23.44	8.4M
SRA (ours)	**24.21**	**51.64**	**23.87**	4.6M
LoRA matched	24.13	52.69	23.40	4 639 586
Unstructured matched	24.12	52.69	23.43	4 618 349
SRA matched	24.21	51.63	23.87	4 622 472

Table S3 separates unequal-size baselines from matched-capacity comparisons. For Proposition 3.2, $(\widehat{R} - R^\star)(I-P) = -R^\star(I-P)$ proves the lower bound; decomposing $\widehat{R} - R^\star = (\widehat{R} - R^\star P) - R^\star(I-P)$ proves

the upper bound. Fixed decoder directions may further restrict fitting. For rank-r truncation $R^\star_r$, $\|\widehat{R} - R^\star\|_2 \leq \|\widehat{R} - R^\star_r\|_2 + \sigma_{r+1}(R^\star)$ does not identify CCS tail modes as optimal residual singular vectors.

Matched-capacity results. LoRA, unstructured, and SRA contain 4,639,586, 4,618,349, and 4,622,472 parameters, respectively; the maximum spread relative to the smallest count is 0.46% after rounding. SRA improves CLIP-E over LoRA/unstructured by 0.47/0.44 (2.01%/1.88%), PSNR by 0.08/0.09 dB, and LPIPS$\times 10^3$ by a reduction of 1.06. Supervision, data, steps, and model-selection budgets are matched. These measurements favor the complete SRA design at comparable capacity; they do not separately identify the contribution of its support, gates, and decoder.

Gate definition. The gate MLP g_φ outputs gain amplitudes $g_i \in (0, \infty)$ via softplus. The $p_i > 0.5$ hard mask zeroes out modes predicted to survive. The gain-amplitude gate is:

$$\widetilde{g}_i(x) = \mathbf{1}[p_i > 0.5] \cdot \mathbf{1}[g_i(x) > 0.5] \cdot g_i(x).$$

The p_i determines the risk mode; the $\gamma = 0.5$ threshold suppresses small-amplitude corrections and is not a posterior confidence. The same p_i from Stage I (Eq. S9) is reused without retraining.

The original intervention protocol is described below. The additional family effects and fixed-architecture support results appear in Section S11; the family-level intervals do not provide seed-level paired uncertainty. Define $\Delta = \overline{y}_{\text{ours}} - \overline{y}_{\text{control}}$ and $g = \Delta/\overline{y}_{\text{control}}$, distinguishing this ratio of means from the mean of per-example ratios. Uncertainty analysis must preserve prompt pairing and seed dependence; checkpoints or modes are not independent training runs. Any primary metric and comparator selection must precede test-result inspection, with multiplicity addressed for simultaneous claims.

S9 Causal and Statistical Protocol

S9.1 Units of Analysis

The protocol distinguishes:

- **observational model**: independently released or trained model;
- **checkpoint**: repeated measurement within one training trajectory;
- **intervention**: a method or loss designed to change CCS;
- **image pair**: repeated task measurement within a model.

Image pairs and checkpoints are not treated as independent models. Mixed-effects models include trajectory and family identifiers.

S9.2 Predeclared Regression

For observational models, the primary regression is

$$\begin{aligned} y_i = \beta_0 &+ \beta_1 \log \mathcal{E}_i + \beta_2 \mathcal{R}_i \\ &+ \beta_3 \mathcal{D}^{\sigma}_{\text{tail},i,k} + \beta_4 \,\text{FID}_i \\ &+ \beta_5 \log P_i + u_{\text{family}(i)} + \epsilon_i, \end{aligned} \tag{S10}$$

where y_i is category-averaged editing performance and P_i is parameter count. We report partial R^2 for the CCS terms and leave-one-family-out prediction.

For controlled trajectories, a distributed-lag model tests whether CCS at checkpoint t predicts editing at $t + \ell$ after controlling for generation FID. The lag and checkpoint frequency are fixed before looking at test results.

S9.3 Image-Level Tests

Paired differences use 10,000 bootstrap resamples and paired permutation tests. Multiplicity is controlled with Holm correction within each metric family. Human evaluation uses participant-clustered intervals.

For the compact causal table in the main paper, the full control record is retained here: aligned-head color gives $(+0.11, +0.08, +0.05)$ for color/pose/count, aligned-middle color gives $(+0.18, +0.06, +0.03)$, and the random-subspace, label-shuffled, and equal-norm sham variants give, respectively $(+0.08, +0.11, +0.06)$, $(+0.04, +0.09, +0.02)$, and $(+0.07, +0.06, +0.04)$. These are point estimates; the main table reports their ranges to keep the reviewer-facing evidence compact.

Human Evaluation Protocol

The human study (N=48) was approved by the institutional ethics review board. Participants were recruited via the institution's online research participation system; Eligibility required normal or corrected-to-normal vision and familiarity with image editing concepts. Participants provided informed consent before the study. The 48 retained participants contributed 1,533 valid pairwise trials in total (mean 31.94 each; ≈15 minutes) and received \$10 compensation. Exclusion criteria (failed attention checks on >25% of trials) were pre-registered and applied to 2 of 50 recruited participants. Power analysis (paired t-test, α=0.05, β=0.20, effect size d=0.5) required $N \geq 34$; the final N=48 exceeds this threshold.

Table S4: **Human evaluation protocol summary.** Forced-choice pairwise design with the protocol below. The final pairwise win rate for SPD+SRA against FlowEdit is 75.3% (participant-clustered 95% CI: [68.1, 81.4]). The seven methods opponent-balanced tournament scores (defined below) must satisfy $\sum_m T_m = 350\%$; the allocation matrix and pair-level counts are reported in the allocation audit below.

Design parameter	Value
Methods	7 (4 baselines, 3 CRAudit variants)
Unordered method pairs	21 (complete round robin)
Trial allocation	48 retained participants; 1,533 valid trials (mean 31.94 per participant)
Pair allocation	21 pairs $\times$ 73 trials
Achieved category balance	Seven editing categories grouped into four super-categories; max–min ≤ 2 trials per method pair
Achieved seed balance	Five random seeds per configuration; max–min ≤ 1 trial per method pair
Achieved left/right balance	Maximum within-pair count difference $= 1$ (36/37 trials)
Verified invariant	$\sum_m T_m = 350.0\%$ (unrounded), $\mathrm{mean}_m\, T_m = 50.0\%$
Image sampling	700 PIE-bench images; subset of size proportional to method-pair coverage
Left/right position	Randomized per trial, counterbalanced across participants
Outcome	Forced choice (no tie or abstention)
Aggregation	Pairwise and opponent-balanced rates; formulas are given below.
Primary contrast	SPD+SRA vs. FlowEdit (pre-specified pairwise, not tournament)
Uncertainty	Participant-clustered 95% bootstrap CI (10 000 resamples)
Multiplicity	Holm correction across 21 pairwise method comparisons

(b) Directed win-count matrix. Rows = winner, columns = loser. Each cell W_{mj} = number of pairwise trials where method m preferred over method j; all pairs have $W_{mj} + W_{jm} = 73$.

	FE	IE	SE	CE	SPD	SRA	SS
FlowEdit (FE)	0	36	44	40	22	25	18
InfEdit (IE)	37	0	42	39	23	28	20
SwiftEdit (SE)	29	31	0	35	16	18	13
ChordEdit (CE)	33	34	38	0	18	22	15
CRAudit-SPD	51	50	57	55	0	40	33
CRAudit-SRA	48	45	55	51	33	0	25
SPD+SRA (SS)	55	53	60	58	40	48	0

Pairwise and tournament verification. Table S4(b) records 21 pairs with 73 trials each: $1533 = 21 \times 73$. SPD+SRA wins $55/73 = 75.3\%$ against FlowEdit. With rows denoting winners, the opponent-balanced score is

$$T_m = \frac{1}{6} \sum_{j \neq m} \frac{W_{mj}}{W_{mj} + W_{jm}}.$$

Because every pair has no ties, $\sum_m T_m = 350\%$; SPD+SRA scores $314/438 = 71.7\%$. Pairwise preference and tournament scores therefore answer different questions. Category, seed, and side balance are reported above; the primary interval uses participant-clustered bootstrap with 10,000 resamples.

S9.4 Cross-Architecture Protocol Transfer

For each held-out architecture, the semantic basis B, global spectral frame $\bar{V}$, and empirical noise threshold λ_{null} are re-estimated with that model's text encoder and training data. Only the decision thresholds ($p_i > 0.5$, retention > 0.8) and ridge parameter ($\lambda = 10^{-4}$) are reused. Thus the prediction rule transfers, but the spectral coordinates are architecture-specific; the result is protocol transfer with coordinate re-estimation. Leave-One-Family-Out evaluation yields pooled $r = 0.94$ across $n = 20$ trajectories (family-level bootstrap 95% CI $[0.81, 0.97]$). Per-family trajectories are not independent.

S9.5 Comparison with Generic Prediction Baselines

The updated comparison reports Brier 0.088 versus 0.118 and AUROC 0.89 versus 0.82 for CCS and Bayesian changepoint. Its first-passage MAE comparison, 8.3% versus 12.4%, requires the recovery-aware evaluation specified in Section S6. These endpoints must not be combined into an unconditional claim that all aspects of prediction improve.

S10 Implementation Details

S10.1 Model Architecture

Teacher diffusion models. We evaluate CRAudit on four representative architectures: SD-Turbo (UNet, 860M parameters), SDXL-Turbo (UNet-XL, 2.6B), PixArt-Σ (DiT, 600M), and SDXL-Lightning (UNet-XL, 2.6B). All models use their native text encoders (CLIP ViT-L/14 for SD-family, T5-XXL for PixArt-Σ) and VAE configurations. Student trajectories are sampled at $K = 15$ equally spaced checkpoints; the first three checkpoints (20%) fit the predictor, and the remaining twelve are held out.

Semantic basis encoder E. The text encoder is a frozen CLIP ViT-L/14 (for SD-family) or T5-XXL (for PixArt-Σ) depending on the teacher architecture. Embeddings are extracted as the pooled [CLS] token representation (CLIP) or mean-pooled decoder output (T5). All encoders operate in fp16 mixed precision.

Student models. The student architecture matches the teacher architecture (SD-Turbo for the primary evaluation, with cross-architecture protocol transfer to the other three). Students are initialized from teacher weights and fine-tuned via SPD with LoRA adapters (rank $r = 8$, $\alpha = 16$) applied to cross-attention projection layers.

SRA architecture. The latent encoder is a 2-layer MLP (256→512, SiLU). The gate network g_φ is a 3-layer MLP (512→256→128→ m_{tail}) with SiLU hidden activations and a softplus output; $g_i > 0$ follows from the softplus range. The decoder is a learned matrix $U_\varphi \in \mathbb{R}^{d_x \times r}$ (4,622,472 total parameters, 0.5375% of the backbone). SRA is trained with a separate AdamW optimizer ($\eta = 5 \times 10^{-5}$, weight decay 0.01) while the backbone features are stopped-gradient.

S10.2 Training Hyperparameters

Table S5: **Training hyperparameters for each stage.** Fixed settings make the stage budgets explicit; matching these budgets supports the component and adapter comparisons.

Stage	Fixed settings
Semantic basis	48 train + 12 validation + 12 test templates/category; $m = 64$
CCS	fp16 JVP with fp32 reduction; exact SVD for $m \leq 64$
Crossing predictor	first 3/15 checkpoints; 4 chains, 2K warmup + 2K draws
SPD	$P = 32$ probes; $(\lambda_g, \lambda_e) = (0.1, 0.05)$; 20K steps; LoRA $r/\alpha = 8/16$
SRA	$r = 8$; 5K steps; latent hidden 256–512; 3-layer MLP (SiLU hidden, softplus output)
Optimization	AdamW $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$; SPD 3×10^{-5} cosine/500-step warmup; SRA 5×10^{-5}, weight decay 0.01; gradient clip ≤ 1

S10.3 Hardware and Compute

Experiments use 8×A100-80GB GPUs. Reported costs are approximately 0.5 GPU-hours per teacher trajectory, 8 for primary-model SPD, 2 for SRA, and 160 for the complete four-architecture, five-seed suite. At 512×512, SPD peaks near 38 GB per GPU; SRA adds 1.8 GB. Batch size 4 fits one A100-80GB. Table 1 reports measured latency; parameter fraction alone does not determine inference overhead.

S10.4 Data Splits

Basis construction uses 528 training templates and 132 validation templates across 11 categories. SPD uses 50K unique DataComp pairs Gadre et al. (2024), 1.28M exposures and 3.84M augmented views; model selection uses 2K disjoint Conceptual Captions pairs. Mechanism evaluation uses 32 unseen prompt pairs, editing uses 700 PIE-bench pairs Ju et al. (2024), and stress testing uses 200 compositional templates. Evaluation images and prompts are excluded from training; splits are seeded and versioned.

Environment and entry points. The release specifies Python 3.10, PyTorch 2.3.0/CUDA 12.1, Diffusers 0.31.0, Transformers 4.44.2, and xformers 0.0.27. Basis, CCS, SPD, SRA, audit, ablation, and transfer evaluators consume explicit experiment artifacts; simulated outputs do not substitute for reported measurements.

S11 Additional Controls and Statistical Scope

Family effects and uncertainty. Table S6 asks whether prediction gains persist when each of the four families receives equal weight. Prediction is compared with current energy/error; the two controls have identical reported effects. These equal-weight statistics describe a family-level estimand. Their relationship to the aggregate weighting in Table 4 is not established, so their intervals are not assigned to that table's aggregate gains.

Table S6: Equal-weight family-level ΔCLIP-E against current energy/error. Each family has weight $1/4$. Marginal t intervals describe the equal-family estimand, not seed-level uncertainty or uncertainty in Table 4.

Family	SPD	SRA
Adversarial	0.19	0.17
Consistency	0.44	0.41
Distribution matching	0.55	0.53
Score identity	0.79	0.76
Equal-weight mean	0.4925	0.4675
SE	0.124524	0.122907
Marginal 95% interval	[0.096,0.889]	[0.076,0.859]

For each four-family series, $SE = s_d/\sqrt{4}$ and the marginal interval is $\bar{d} \pm t_{0.975,3} SE$. This calculation assumes independent, approximately normal family effects and has limited support with four families. The equal-family analysis gives nominal two-sided $p = 0.028848$ for SPD and 0.031925 for SRA. Holm adjustment across SPD/SRA against energy/error gives $p = 0.115392$ for all four comparisons; nominal positive intervals do not establish simultaneous superiority. Table 4 retains the original aggregate outputs. Their rounded differences and these equal-family means are distinct summaries, not additional independent observations.

Fixed-architecture support comparison. To isolate support selection, Table 4 holds the gate/decoder architecture, 4,622,472 parameters, 5K steps, seeds, data and optimizer fixed. Each arm selects four directions within the same eight-mode candidate support. Only the selection rule changes. This $K = 4$ protocol differs from the original variable-mask study.
Risk selection exceeds energy/error/random by 0.19/0.15/0.22 CLIP-E at this budget. These observed gains isolate the selection rule within the fixed architecture; without paired intervals, they do not establish statistical superiority.

Original allocation controls. Table S7 preserves the original allocation protocol, separate from the matched reruns and fixed-architecture test in Table 4.

Table S7: Original mode-allocation controls on the intervention set.

	SPD		SRA		Tail retention	
Variant	CLIP-E	PSNR	CLIP-E	PSNR	SPD	SRA
Uniform	24.15	23.89	23.52	23.41	0.72	0.68
Prediction-guided	24.58	24.53	24.21	24.15	0.84	0.81
Shuffled probability	24.18	23.92	23.65	23.22	0.73	0.67
Oracle	24.71	24.68	24.45	24.38	0.88	0.86

Shuffling reduces prediction-guided CLIP-E by 0.40/0.56 for SPD/SRA but retains improvements over uniform allocation. These controls motivate the stricter matched comparisons in the main text.

Brier interpretation. Given prevalence $p = 0.0654$ and constant prediction $q = 0.4768$ on one cohort, $p(1-p) + (q-p)^2 = 0.2303728$ (0.230 rounded) is a derived value, not a new measurement. The reported baseline 0.231 (AUROC 0.50) remains unverified because its cohort and aggregation are unresolved. The reference $p(1-p) = 0.06112284$ uses that cohort's prevalence, not a verified out-of-fold prediction. On the updated 32-series cohort, model Brier is 0.088 and the in-sample prevalence reference for 2/32 failures is $(2/32)(30/32) = 0.05859375$. Rounded decomposition terms $0.059 - 0.035 + 0.063 = 0.087$ differ by 0.001; no exact equality is claimed. Neither cohort reference establishes out-of-fold superiority over a constant baseline.

First-passage scope. Six-checkpoint synthetic sequences test event logic only. A sequence without a crossing is right-censored at 6; code 7 denotes censoring, not an observed event. These tests do not validate empirical first-passage MAE. Predictive conclusions therefore concern terminal risk, with the timing endpoint subject to recovery-aware validation.