FROM PIXELS TO INITIAL STATES: GEOMETRY AND PROPAGATION OF SOURCE-INDUCED SENSITIVITY IN DIFFUSION AND FLOW MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Small source-image perturbations can substantially alter image-to-image diffusion and flow outputs. Output differences alone do not reveal how the induced initial-state displacement and its subsequent propagation contribute to the response. Contrary to a simple displacement-growth account, we argue that larger mean output differences need not accompany greater mean normalized trajectory displacement, even at approximately matched initial-state magnitudes. We use the source-conditioned initial state as the analysis interface, fixing stochastic inputs, conditioning, and execution within each clean–perturbed pair. Across SD1.5, SDXL, and FLUX, target-optimized (native) sources induce larger initial displacements than tested controls; direct state interventions in SDXL/FLUX show radial and directional dependence. Our primary 50-image FFHQ bounded-source comparison approximately matches initial-state radii for native perturbations and perturbations optimized with the other model (transferred). Native mean DINOv2 dissimilarity from paired clean outputs is larger in both SDXL and FLUX, showing that initial-state magnitude alone does not account for the observed responses. Despite this output ordering, native mean normalized trajectory displacement is smaller in SDXL and larger in FLUX than its transferred counterpart. Both output and trajectory-displacement orderings persist under fourfold refinement of the executed integration schedules. A separate 200-image non-face COCO cohort extends the unmatched source-to-state and output comparisons without repeating matched-radius interventions.

1 INTRODUCTION

Small source-image perturbations can substantially alter image-to-image diffusion and flow outputs under identical generation settings (Figure 1(a)) (Salman et al., 2023; Van Le et al., 2023; Xu et al., 2024), enabling protection against unauthorized editing (Choi et al., 2025) but complicating reliable editing. Output differences alone do not distinguish unequal initial-state magnitudes, direction-dependent response, or subsequent trajectory separation.

Prior work links protective perturbations to latent-representation distortion in customization (Peng et al., 2025) and documents error accumulation from training–inference mismatch (Ning et al., 2023) and quantization (Lee et al., 2025). These findings motivate examining initial displacement and propagation, but leave their relationship to source-induced output response under matched initial-state magnitude unresolved. A simple displacement-growth account would associate larger mean output changes with a larger initial displacement or greater subsequent trajectory separation.

In our configurations, the source-conditioned initial state z_{init} carries the source's influence into the solver (Figure 1(b)). With stochastic inputs, conditioning, and execution fixed within each clean–perturbed pair,

$$x \longrightarrow z_{\text{init}} \longrightarrow y, \qquad \delta x \longrightarrow \Delta z_{\text{init}} \longrightarrow \Delta y. \tag{1}$$

To distinguish these possibilities, we write each nonzero induced displacement in terms of its magnitude and direction:

$$\Delta z_{\text{init}} = ru, \quad r = \|\Delta z_{\text{init}}\|_2, \quad u = \frac{\Delta z_{\text{init}}}{\|\Delta z_{\text{init}}\|_2}. \tag{2}$$

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.

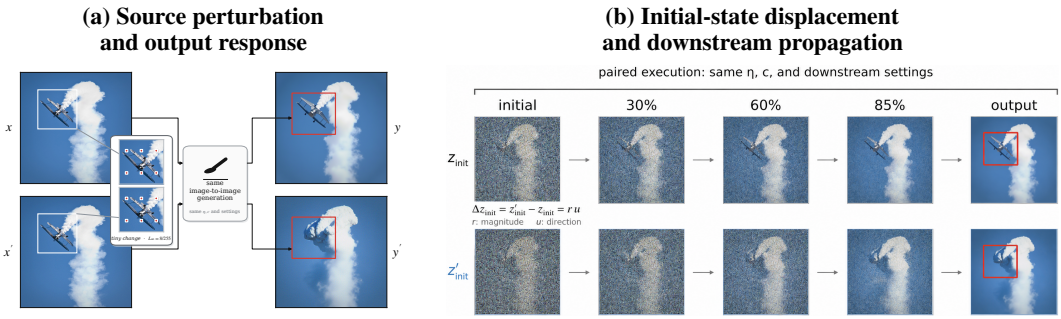

Figure 1: From source perturbations to initial-state propagation. (a) Visually negligible source-image changes can yield substantially different outputs under identical generation conditions. (b) We analyze this sensitivity at the source-conditioned initial state: paired executions share stochastic inputs, conditioning, and downstream settings, while the induced initial-state displacement differs in magnitude and direction. Decoded intermediate states are shown only for illustration and are not used for statistical inference.

We vary radius at fixed direction and compare directions at matched radius to test whether displacement magnitude alone accounts for the response. Because direct state edits need not have bounded-source preimages, we complement them with bounded source perturbations whose realized initial-state radii are approximately matched. Matching the initial radius leaves subsequent trajectory separation unconstrained, which we measure separately.

We construct source endpoints by optimizing clean-relative model prediction discrepancies, then measure initial-state geometry and use it alone to select matched candidates before output evaluation. Neither initial-state radius nor output metrics enter the constructor objective. Native sources use the target model; transferred sources use the other model and transfer in image space. These labels denote complete construction procedures.

Native sources induce larger initial displacements than tested controls across three models; direct SDXL/FLUX state interventions show radial and directional dependence. Our primary bounded-source comparison (P4; 50 FFHQ images) finds larger native mean DINOv2 responses than transferred in both SDXL and FLUX at approximately matched initial-state radii. Yet native mean normalized trajectory displacement is smaller in SDXL and larger in FLUX than transferred, challenging a shared greater-growth explanation. Both output and trajectory-displacement orderings persist under fourfold integration-schedule refinement. An exploratory same-target control on the reused cohort also yields larger mean DINOv2 responses for prediction-discrepancy (PD) than encoder-displacement-only (ENC) sources at approximately matched radii defined from their own endpoint pairs. Our contributions are:

1. An initial-state intervention design connects bounded source changes to downstream response, controlling magnitude and direction separately.
2. Approximately matched-radius bounded-source comparisons show response differences beyond initial-state magnitude, complemented by a separate same-target PD/ENC control.
3. The native-over-transferred output ordering coexists with opposite normalized trajectory-displacement orderings across the two targets; this pattern persists under fourfold integration-schedule refinement.

2 Related Work

2.1 Source-Induced Sensitivity in Image-to-Image Generation

SDEdit initializes generation from a noised source image (Meng et al., 2022). Protective perturbations show that small source changes can disrupt editing, personalization, or reconstruction, through distinct objectives. PhotoGuard targets encoder features or diffusion outputs (Salman et al., 2023); AdvDM maximizes a sampled diffusion loss (Liang et al., 2023); Mist combines semantic and textural losses (Liang & Wu, 2023). DiffusionGuard targets early denoising with mask augmentation (Choi et al., 2025), while personalization studies perturb training images or attention (Van Le et al., 2023; Xu et al., 2024; Zhao et al., 2024). CAT examines latent distortion and trains adapters to counter protective

perturbations (Peng et al., 2025). Their objectives and evaluation settings differ; they do not imply a common perturbation mechanism.

These studies primarily optimize protection or attack effectiveness. Our analysis uses bounded source constructions to examine how initial-state magnitude, direction, and subsequent propagation relate to output sensitivity. Its central comparison uses initial-state geometry to approximately match source-reachable radii before output evaluation. The prediction-discrepancy constructor supplies diagnostic sources for this response analysis. Appendix H expands the comparison.

2.2 Geometry and Propagation of Generative States

Errors introduced during iterative generation can accumulate across sampling. Training–inference mismatch exposes diffusion samplers to their own prediction errors (Ning et al., 2023); AccuQuant studies quantization-error accumulation across denoising steps (Lee et al., 2025). These observations motivate measuring downstream state separation, but do not establish the displacement-growth ordering examined here. Our controlled comparison measures separation from a paired clean trajectory within a fixed target execution; prediction error against ground truth and quantization error against a full-precision model are different estimands.

Semantic latent directions and NoiseCLR (Kwon et al., 2023; Li et al., 2024; Dalva & Yanardag, 2024), guidance-update angles (Jin et al., 2025), Hessian geometry (Lobashev et al., 2025), and probability-density geodesics (Yu et al., 2025) motivate separating displacement magnitude from direction. Geometry-Aware Distillation aligns Jacobian–vector responses to initial-noise changes (Huang et al., 2026). However, generic anisotropy alone does not determine how bounded, source-reachable directions behave when their realized initial-state magnitudes are approximately matched. Our experiments therefore combine controlled state-space comparisons with source-realizable matched-radius interventions and subsequent trajectory measurements. This combination tests response differences beyond initial-state magnitude within bounded source-reachable changes; the tested directions have no assigned semantic attributes.

2.3 Interventional Analysis of Generative Models

Attention and information-flow analyses (Tang et al., 2023; Kong et al., 2024), studies of generalization and composition (Kadkhodaie et al., 2024; Niedoba et al., 2025; Bradley et al., 2025), and knowledge localization/editing (Basu et al., 2024a;b) examine how generative behavior is implemented. Such studies often analyze or intervene on internal representations, attention, or model components to localize information or explain behavior. We instead use the source-conditioned initial state as a controlled interface and measure finite downstream response under fixed execution, without claiming neural-component localization or formal mediation. This scope supports comparisons of source-induced sensitivity across realized state changes, while leaving their implementation within the network unresolved.

3 An Interventional Framework for Source-Induced Sensitivity

3.1 The Initial State as an Intervention Interface

We intervene on the state supplied to the generative solver, after source encoding and noise initialization, rather than on the VAE code alone. Fix a model M, stochastic input η, and conditioning and execution settings c in an image-to-image diffusion or flow computation (Ho et al., 2020; Rombach et al., 2022; Lipman et al., 2023):

$$h_x = E_M(x), \qquad z_{\text{init}} = I_M(h_x, \eta, c), \qquad z_T = \Phi_{M,c}(z_{\text{init}}), \qquad y = D_M(z_T). \tag{3}$$

Here E_M encodes the source, I_M constructs the first state consumed by denoising or flow computation, and $\Phi_{M,c}$ evolves it to the final state z_T (also written z_f). The unchanged output map D_M includes decoding and postprocessing.

For any bounded change $x' = x + \delta$, define $F_M(x, \eta, c) := I_M(E_M(x), \eta, c)$ and $\Delta z_{\text{init}} := F_M(x + \delta, \eta, c) - F_M(x, \eta, c)$. We use *causal* for interventions on this realized state with (η, c) and downstream execution held fixed. The estimand is the response to a controlled state change; it does not require a unique source preimage or identify natural direct and indirect effects.

To obtain diagnostic source perturbations, a prediction-discrepancy objective selects images under an $8/255$ budget. Neither state displacement nor output metrics enter this objective; this exclusion

does not imply statistical independence from the evaluated outcomes. Let $p_{M,j}(\tilde{x})$ be model M's prediction at probe setting $j \in \mathcal{T}_M$, with $N_{M,j}$ scalar elements. We select

$$\begin{aligned} \mathcal{B}_\epsilon(x) &= \{\tilde{x} \in [0,1]^d : \|\tilde{x} - x\|_\infty \le \epsilon\}, \\ x_M^\star &= \mathcal{A}_M(x, J_M, \mathcal{B}_\epsilon(x)) \in \mathcal{B}_\epsilon(x), \\ J_M(\tilde{x}) &= \frac{1}{|\mathcal{T}_M|} \sum_{j \in \mathcal{T}_M} \frac{\|p_{M,j}(\tilde{x}) - \mathrm{sg}[p_{M,j}(x)]\|_2^2}{N_{M,j}}, \end{aligned} \tag{4}$$

Here sg detaches the clean prediction. $\mathcal{A}_M$ is projected sign-gradient ascent in the model's prediction parameterization, and $\delta_M^\star = x_M^\star - x$. We call this bounded, optimized source change the native variation. Both Γ_M and the semantic metrics are measured after construction.

3.2 Measuring Initial-State Displacement

To compare initial-state displacements under a common source budget, for $v \in \mathbb{R}^{d_v}$ let $\mathrm{RMS}(v) = \|v\|_2/\sqrt{d_v}$ and define $\Gamma_M(x, \delta) := \mathrm{RMS}(\Delta z_{\mathrm{init}})$. This dimension normalization is neither a source-normalized gain nor a cross-model sensitivity ranking. Because J_M and Γ_M share model computations, their relation characterizes probe-specific coupling; the source comparison alone does not identify an independent internal mechanism. Appendix A.1 gives the continuous-map interpretation.

3.3 Separating Displacement Magnitude from Direction

Larger displacements could explain larger output changes. Using $\Delta z_{\mathrm{init}} = ru$ from Eq. 2, an optional output embedding Ψ, and fixed dissimilarity d_Ψ, define

$$Y_M(z) := D_M(\Phi_{M,c}(z)), \qquad R_M(z_{\mathrm{init}}, r, u) := d_\Psi(Y_M(z_{\mathrm{init}} + ru), Y_M(z_{\mathrm{init}})). \tag{5}$$

R_M can depend on both r and u. Varying r at fixed u isolates radial sensitivity; varying u at matched r tests whether magnitude is sufficient. The primary d_Ψ is DINOv2 dissimilarity (Oquab et al., 2023): one minus cosine similarity between normalized features of the perturbed output and its paired clean generated output. Larger values mean greater feature difference, not higher output quality. Radius and direction are defined within each target's executed coordinates. Appendix A.1 gives finite-response identities under stated regularity assumptions.

Initial-radius matching controls the starting magnitude but not subsequent evolution; Section 4.5 separately tests whether the bounded-source output ordering accompanies systematically greater native trajectory displacement.

3.4 Restoring Source Feasibility

Directly edited states need not correspond to bounded source images. A state is *source-reachable under budget* ϵ if some δ satisfying $\|\delta\|_\infty \le \epsilon$ produces it: $z = F_M(x + \delta)$. This denotes budget feasibility, not natural-image typicality. Constructing both compared states through bounded images guarantees such preimages.

Native and transferred sources are executed by the same target. Their displacement ratio describes magnitude transfer; radius matching compares directions realized by selected bounded images, not guaranteed copies of the optimized endpoint directions.

Each branch follows a deterministic candidate path from the clean image toward its optimized endpoint.

Let $r_{\mathrm{native}}^{\mathrm{end}}$ and $r_{\mathrm{transfer}}^{\mathrm{end}}$ denote the seed-averaged endpoint radii of the two source families in the target model. We define

$$r_{\mathrm{common}} = \min\left(r_{\mathrm{native}}^{\mathrm{end}}, r_{\mathrm{transfer}}^{\mathrm{end}}\right), \qquad r_{\mathrm{target}}(\rho) = \rho r_{\mathrm{common}}, \quad \rho \in \{.5, .7, .9\}. \tag{6}$$

Before semantic evaluation, the specified matcher selects the nearest candidate on each discrete uint8 path using seed-averaged radii, then applies a 2% pairwise mismatch tolerance. Neither the reference nor a target fraction guarantees an exactly shared executed radius. See Appendix Algorithm 2.

Algorithm 1 summarizes source construction and two separate comparisons. Bounded-source matching uses separately constructed branch endpoints; the transferred endpoint is constructed using its source model's probe and optimization procedure.

Algorithm 1 Source construction, state intervention, and reachable matching

Require: source x, target M, probe J_M, fixed constructor randomness and settings (α, K_M), budget ϵ, fixed conditioning/execution settings c, target-execution seeds $\mathcal{S} = \{2025, 2026, 2027\}$
Require: direct-state plan (r', v): ℓ_2 radii r' and unit directions v; paired stochastic input η for each seed
Require: serialized clean q_i^{clean}, endpoint pair $\mathcal{E}_i = \{q_{i,b}^{\text{end}}\}_{b \in \{\text{native,transferred}\}}$, fractions $\mathcal{R} = \{.5, .7, .9\}$

Phase I · Source construction
$\tilde{x} \leftarrow \text{Uniform}(\mathcal{B}_\epsilon(x))$; $x_M^\star \leftarrow \tilde{x}$
for $k = 1, \ldots, K_M$ **do**
  $\tilde{x} \leftarrow \Pi_{\mathcal{B}_\epsilon(x)}[\tilde{x} + \alpha\,\text{sign}(\nabla_{\tilde{x}} J_M)]$
  retain the best-objective continuous iterate as $x_M^\star$
serialize $x_M^\star$ and capture $z_{\text{init}} = F_M(x, \eta, c)$ and $\Delta z_{\text{init}} = F_M(x_M^\star, \eta, c) - z_{\text{init}}$ for fixed η
$r \leftarrow \|\Delta z_{\text{init}}\|_2$, $u \leftarrow \Delta z_{\text{init}}/r$, and $\Gamma_M \leftarrow r/\sqrt{d_M}$
Phase II · State-space controls
for all (r', v) in the direct intervention plan **do**
  run $Y_M(z_{\text{init}} + r'v)$ with the same execution settings
Phase III · Source-feasible matching
for all $\rho \in \mathcal{R}$ **do**
  $(\mathcal{Q}_i^{(\rho)}, \mathcal{G}_i^{(\rho)}) \leftarrow \text{FrozenMatch}(F_M, q_i^{\text{clean}}, \mathcal{E}_i, c, \mathcal{S}, \rho)$ (Algorithm 2)
require source feasibility and the frozen cohort geometry gate (Appendix C.2); stop on failure
Phase IV · Matched-source evaluation
evaluate all matched sources $\mathcal{Q}_i^{(\rho)}$ against paired clean outputs using seeds $\mathcal{S}$
average semantic endpoints over $\mathcal{S}$ within each image
perform the specified image-level inference

3.5 Pre-specified Tests and Evidential Roles

Table 1 summarizes the four specified comparisons.

Label	Question	Comparison and prediction	Role
P1	Source-to-state transfer	Native Γ_M exceeds controls	Measure displacement
P2	Radial response	Response increases with r at fixed native u	Isolate magnitude
P3	Directional response	Responses differ at matched r	Test direction
P4	Bounded-source direction	Higher native mean DINOv2 dissimilarity at the primary radius	Restore source feasibility

Table 1: Prespecified comparisons and roles; predictions, not observed outcomes.

P4 uses a fresh holdout. Trajectory measurements examine downstream propagation; schedule refinement and the same-target encoder-only constructor control are exploratory follow-ups on the reused P4 cohort.

4 Experiments

4.1 Experimental Setup

Models, data, and controls. We evaluate SD1.5 (Rombach et al., 2022), SDXL (Podell et al., 2024), and FLUX.1-dev (Black Forest Labs, 2024) on FFHQ (Karras et al., 2019), plus SDXL/FLUX on non-face COCO (Lin et al., 2014). Sources use an $8/255$ budget; controls are target-model random-sign and pixel-RMS-matched Gaussian perturbations (Appendix B.1).

Cohorts and inference. Images are inference units; seeds and projections are averaged within image. For Γ_M and DINOv2 cosine dissimilarity (DINO), 10,000 image-cluster bootstrap resamples give simultaneous studentized max-t intervals. P4's primary family contains the two $\rho = .9$ DINO contrasts; lower radii are exploratory and other endpoints post-hoc. Refined DINO and refined trajectory reuse P4 images in separate exploratory two-target families, distinct from P4 and original trajectory; error control is within-family only (Appendix D.4). The same-target PD-versus-encoder-only control reuses the P4 images and forms a separate exploratory two-target simultaneous family (Appendices B.3, D.1, E.2).

Intervention fidelity. In SDXL/FLUX, z_{init} is the only downstream visual source input; ControlNet/IP-Adapter are disabled. A 10-image SDXL replay reproduces natural outputs exactly (pixel MSE and LPIPS $= 0$; mean DINO residual 2.384×10^{-8}; Appendix Figure A1).

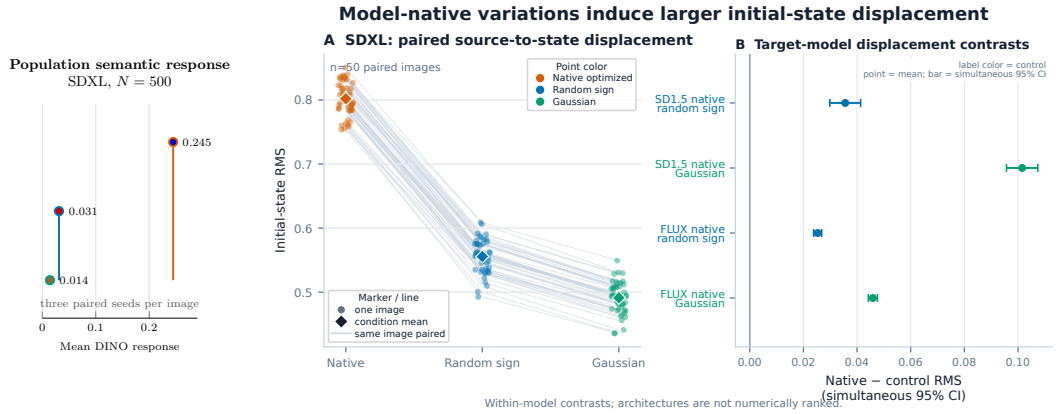

Figure 2: SDXL population response ($N = 500$, left) and source-to-state transfer (right). Paired SDXL displacements and SD1.5/FLUX contrasts against controls constructed for each target model are shown; markers denote means and bars simultaneous 95% intervals.

Model	Control	Native−control Γ_M	Within-study simultaneous 95% CI	Win
A. Generic controls constructed for the target model				
SD1.5	Random sign	0.035590	$[0.029852, 0.041329]$	–
SD1.5	Gaussian	0.101560	$[0.095717, 0.107402]$	–
FLUX	Random sign	0.025314	$[0.023800, 0.026828]$	–
FLUX	Gaussian	0.045897	$[0.044206, 0.047589]$	–
B. Optimized random prediction-space projections				
SD1.5	One-dimensional projection	0.057844	$[0.051538, 0.064150]$	100%
FLUX	One-dimensional projection	0.049740	$[0.045531, 0.053949]$	100%
SDXL	One-dimensional projection	0.073341	$[0.070454, 0.076228]$	100%

Table 2: Source-to-state contrasts. Generic-control contrasts use 50 images per model and separate within-target families (Appendix D.4). Optimized-projection contrasts use independent 24-image development cohorts with projections averaged per image and a separate three-model family.

4.2 Unequal Source-to-State Transfer

Unmatched source comparisons establish output and initial-displacement differences under a common pixel budget. On 500 FFHQ images, SDXL mean DINO cosine dissimilarity is 0.244767 for the native variation, 0.031356 for random sign, and 0.014423 for Gaussian. Native-minus-control contrasts are 0.213410 (95% CI $[0.205377, 0.221568]$) and 0.230344 ($[0.222159, 0.238713]$), with 100% image-level win rates.

The six ArcFace, CLIP-image, and LPIPS native-minus-control intervals also exclude zero, with 99.8–100% image-level win rates (Appendix Table 20 and Appendix Figure A3).

Γ_M quantifies initial displacement after source construction. Table 2 gives SD1.5/FLUX generic-control contrasts and three-model optimized-projection contrasts; every simultaneous 95% interval is positive. SDXL traces share the generic-control ordering. Gaussian controls match native pixel RMS; random-sign controls have higher source RMS but smaller state displacements, excluding source RMS alone as an explanation (Figure 2).

With the same optimizer and $8/255$ budget, native mean Γ_M exceeds every tested optimized one-dimensional random prediction-space projection: native-to-projection ratios are 1.101 for SDXL, 1.200 for SD1.5, and 1.275 for FLUX. This characterizes source-to-state coupling, not a mechanism's contribution; direct interventions next separate magnitude and direction.

4.3 Magnitude and Direction Have Distinct Downstream Effects

To test the effect of magnitude, we vary radius along each fixed native direction. Across the tested radii, SDXL mean DINO dissimilarity rises from 0.063275 and 0.077815 to 0.234008. The higher-minus-lower radius contrast for FLUX is 0.019080 (simultaneous 95% CI $[0.013963, 0.024197]$).

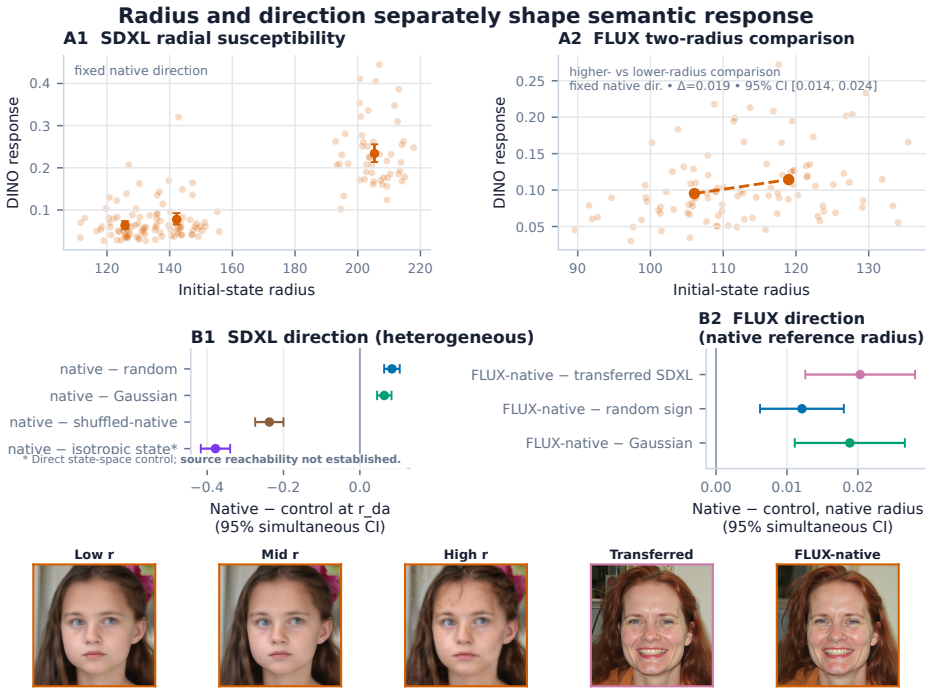

Figure 3: Radius sweeps at fixed native directions (A1–A2) and direction comparisons at reference radii (B1–B2); r_{da} is the image-specific native-induced initial-state L_2 reference radius. FLUX uses its native reference radius. The qualitative row shows paired examples. Three seeds are averaged within each of 50 images; inference families and the older FLUX execution-record limitation are in Appendix D.4.

Magnitude is nevertheless insufficient. At matched radius, SDXL's native direction exceeds random-sign and Gaussian directions but trails shuffled-native and isotropic directions. Direction matters, even though the native direction is not the strongest arbitrary state direction. At the FLUX native reference radius, the native direction exceeds the direction induced by an SDXL-optimized source by 0.020321 ($[0.012582, 0.028061]$). Figure 3 shows the distinct radial and directional responses.

Comparisons use target coordinates; unmatched rows combine magnitude and direction (Appendix Table 21; older FLUX lacks BF16 state receipts to recheck achieved radius equality, Appendix D.4). These ambient state edits need not have bounded-source preimages.

4.4 Directional Effects Persist under Bounded-Source Matching

To test directions with bounded source preimages, we construct both branches in source space and approximately match initial-state radii before output evaluation. The 50-image FFHQ holdout was fixed from 550 previously unused eligible images, independently of outcomes. Both SDXL↔FLUX transfer directions were pre-specified; SD1.5 was outside the scope.

Manipulation checks. The matcher uses the seed-averaged target-state radius. At $\rho = .9$, all 300 paired-seed mismatches are at most 0.5% (maximum 0.05143%; Table 3(a)). Median native/transferred direction cosines are 0.756050 for FLUX and 0.526988 for SDXL (Table 3(b)).

Semantic effect. At $\rho = .9$, FLUX favors native by 0.008095 DINO dissimilarity (simultaneous 95% CI $[0.004503, 0.011688]$; 88% wins), and SDXL by 0.008863 ($[0.003838, 0.013888]$; 72% wins). Paired effects are $d_z = 0.749$ and 0.586, respectively. These two intervals form the P4 primary family; $\rho = .5, .7$ are exploratory with descriptive pointwise intervals in Figure 4.

Endpoint robustness. Table 4 compares DINO with the post-hoc endpoints. ArcFace and LPIPS agree for both targets. CLIP-image agrees for FLUX, whereas its SDXL interval contains zero and its point estimate is slightly negative. The native-direction advantage varies across representations; DINO remains the sole primary endpoint. Appendix Figures A2–A9 show checks and examples.

(a) Seed-wise radius mismatch (%)

Scope	N	Median	Q95	Max.	Count $\leq 0.5\%$
$\rho = .9$	300	0.00623	0.02434	0.05143	300/300
All ρ	900	–	–	–	900/900

(b) Native–transferred direction similarity

Target	N	Q25	Median	Q75	Median angle
FLUX	50	0.739205	0.756050	0.768414	40.88°
SDXL	50	0.516326	0.526988	0.535266	58.20°

Table 3: Bounded-source matching checks. (a) Seed-wise relative radius mismatch (%) at primary $\rho = .9$, plus counts over $\rho \in \{.5, .7, .9\}$; N counts paired image–target–seed–radius cells. (b) Image-aggregated native–transferred direction-cosine quantiles and median angular separation at $\rho = .9$; $N = 50$ images per target.

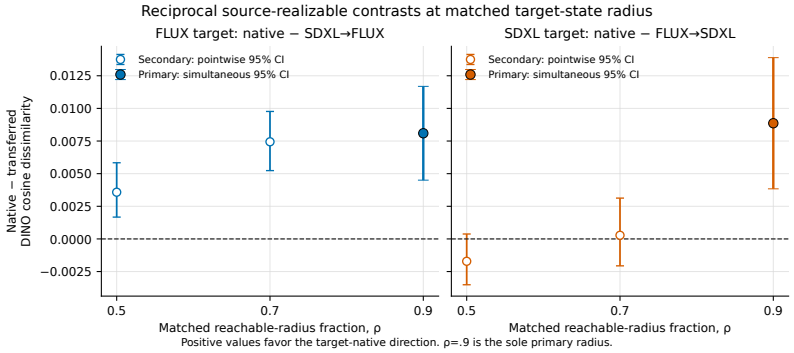

Figure 4: Bounded-source matched-radius DINO contrasts. Filled $\rho = .9$ markers form the simultaneous primary family; open $\rho = .5, .7$ markers are exploratory. All cells satisfy the 2% matching tolerance; $N = 50$ images per target with three paired seeds.

Endpoint	FLUX contrast [95% CI]	SDXL contrast [95% CI]
DINO (primary)	$0.008095\ [0.004503, 0.011688]$	$0.008863\ [0.003838, 0.013888]$
CLIP-image	$0.008210\ [0.006397, 0.010087]$	$-0.000984\ [-0.002458, 0.000488]$
ArcFace	$0.014612\ [0.011005, 0.018249]$	$0.017200\ [0.010705, 0.025658]$
LPIPS	$0.009267\ [0.006745, 0.012004]$	$0.002862\ [0.001822, 0.004092]$

Table 4: Bounded-source native-minus-transferred contrasts at $\rho = .9$ ($N = 50$ per target; seeds averaged within image). The DINO intervals form the primary simultaneous family; other endpoints are post-hoc secondary analyses with their reported intervals.

Same-target constructor control. To test whether maximizing source-encoder displacement alone reproduces the matched-radius response, we construct an encoder-only (ENC) bounded-source control with the same target-specific budget and optimization framework, then approximately match its realized initial-state radius to the prediction-discrepancy (PD) construction. PD retains the larger mean DINOv2 response in FLUX ($+0.004505$, simultaneous 95% CI $[0.001030, 0.007980]$) and SDXL ($+0.091660$, $[0.068013, 0.115307]$). Thus, encoder-displacement maximization alone does not reproduce the PD response under these conditions. This exploratory reused-cohort comparison distinguishes two complete constructor objectives, without localizing a neural mechanism. FLUX ENC required deterministic backward execution for repeatability; the historical PD backend setting was not recorded. Appendix E.2 gives construction and matching checks.

Initial-radius matching leaves subsequent trajectory evolution unconstrained. We next examine propagation for the native-versus-transferred P4 comparison.

4.5 Trajectory Displacement and Schedule Refinement

We measure same-target paired separation, not prediction error against ground truth or a different-precision model. For branch b, $G_t^b = \mathrm{RMS}(z_t^b - z_t^{\mathrm{clean}}) / \mathrm{RMS}(\Delta z_{\mathrm{init}}^b)$. The summary $\overline{G}^b$ averages all post-initial executed checkpoints, then paired seeds within image; target means average images.

At $\rho = .9$, FLUX native/transferred means are $2.236623/2.122521$, with $\Delta\overline{G} = +0.114102$ (simultaneous 95% CI $[0.108191, 0.120013]$); SDXL has $0.809012/0.864937$, with -0.055925 ($[-0.057526, -0.054324]$). The cohort-mean ordering holds at every post-initial executed checkpoint

($G = 1$ initially). Image-level summary win rates are 100% and 0%, respectively. These contrasts form a separate simultaneous family.

A shared greater-growth explanation would require positive native-minus-transferred trajectory contrasts in both targets; SDXL has the opposite sign. Greater mean normalized trajectory displacement is therefore not a shared explanation for the native output advantage. These cohort means do not describe individual trajectories; architecture remains confounded with target execution settings, and nonlinear magnitude–direction interactions remain possible.

Schedule-refinement robustness. We subdivide each executed interval fourfold, preserving P4 sources, initial states, paired stochasticity, conditioning, and noise horizon. Native-minus-transferred DINO remains positive for FLUX ($+0.011504$) and SDXL ($+0.009603$); trajectory contrasts retain opposite signs ($+0.093737$ and -0.046271, respectively). This exploratory follow-up establishes numerical-resolution robustness, not convergence (Appendix E.1, simultaneous intervals); conditioning strength, prompt, and execution horizon remain model-specific.

4.6 Initial-State and Output Comparisons beyond Portraits

On 200 non-face COCO images (three paired seeds), all eight unmatched contrasts against random-sign and Gaussian controls have positive intervals (Table 5). Geometry and DINO retain separate simultaneous families within target coordinates (annotation details: Appendix D.1).

Target	Endpoint	Native−random [95% CI]	Native−Gaussian [95% CI]
SDXL	Γ_M	$0.292960\ [0.287141, 0.298779]$	$0.347560\ [0.342048, 0.353072]$
	DINO	$0.198076\ [0.180078, 0.216073]$	$0.205069\ [0.186642, 0.223497]$
FLUX	Γ_M	$0.007217\ [0.006015, 0.008419]$	$0.023895\ [0.022511, 0.025279]$
	DINO	$0.032498\ [0.023070, 0.041925]$	$0.056777\ [0.045859, 0.067694]$

Table 5: COCO non-face source-to-state and output contrasts ($N = 200$; three paired seeds averaged within image). Separate simultaneous 95% intervals cover geometry and semantic families.

SDXL DINO wins on every image against both controls; FLUX win rates are 79.0% against random sign and 85.5% against Gaussian, so positive means coexist with image-level reversals. COCO repeats neither bounded-source matching nor trajectory comparisons. Appendix Table 22 gives all win rates and counts; Figure A6 shows examples.



Figure 5: COCO outputs for the first deterministically selected example in Appendix Figure A6. Clean, native, and random-sign outputs are paired within each target; no radius matching is used. Table 5 reports cohort-level results.

5 Conclusion

Bounded-source interventions reveal response differences at approximately matched initial-state radii. Initial-state magnitude alone does not account for the observed responses, and the same native-over-transferred output ordering persists under opposite normalized trajectory-displacement orderings, including under fourfold integration-schedule refinement. These findings concern the evaluated configurations, with architecture and execution confounded; they identify neither neural components nor formal mediation or universal architectural principles.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the development and refinement of the conceptual framework, mathematical claims and hypotheses, experimental methodology and protocol review, and interpretation of experimental results. We additionally used generative AI tools for manuscript drafting and editing, literature search and organization, LaTeX and cross-reference repair, software-code assistance, and scientific-figure labeling, layout, and presentation.

We did not use generative AI tools external to the evaluated systems to generate synthetic datasets, perform experimental measurements, select or exclude experimental outcomes, or replace the authors' statistical analyses. Images generated by the diffusion and flow models studied in this work are outputs of the evaluated experimental pipelines under documented settings and constitute experimental data rather than externally generated synthetic evidence.

All AI-assisted research ideas, mathematical statements, experimental protocols, code and analyses used in the study, literature references, result interpretations, and manuscript content were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses existing public image datasets and includes human annotation of image content for the COCO non-face cohort. It reports no intervention on human participants and makes no claim of IRB approval or exemption. The work is intended as an empirical analysis of source-induced sensitivity. The same insights could potentially be misused to manipulate generation pipelines, so conclusions are limited to the evaluated scientific scope.

REPRODUCIBILITY STATEMENT

An anonymized executable artifact, `core_p4_reproducible.tar.gz`, accompanies this submission. It contains the verified SDXL and FLUX source constructors, bounded-source matching implementation, P4 image-level manifest and source hashes, target-execution configurations, original and refined integration schedules, DINOv2 evaluation, trajectory analysis, statistical inference code, deterministic fixtures, and lightweight commands reproducing the reported aggregate P4, trajectory, and schedule-refinement statistics. Exact clean, perturbed, and generated images are external dependencies: neither the required historical clean inputs nor a verified reconstruction recipe is bundled. Appendix G distinguishes code, archived statistical validation, fixtures, and the remaining input and historical-provenance gaps; full model-backed replay remains unverified.

REFERENCES

Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Vlad I. Morariu, Nanxuan Zhao, Ryan A. Rossi, Varun Manjunatha, and Soheil Feizi. On mechanistic knowledge localization in text-to-image generative models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3224–3265. PMLR, 2024a. URL https://proceedings.mlr.press/v235/basu24b.html.

Samyadeep Basu, Nanxuan Zhao, Vlad I. Morariu, Soheil Feizi, and Varun Manjunatha. Localizing and editing knowledge in text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=Qmw9ne6SOQ.

Black Forest Labs. Announcing black forest labs: The FLUX.1 model family. Black Forest Labs, August 2024. URL https://blackforestlabs.ai/announcing-black-forest-labs/.

Arwen Bradley, Preetum Nakkiran, David Berthelot, James Thornton, and Joshua M. Susskind. Mechanisms of projective composition of diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5319–5354. PMLR, 2025. URL https://proceedings.mlr.press/v267/bradley25a.html.

William June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. DiffusionGuard: A robust defense against malicious diffusion-based image editing. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9OfKxKoYNw.

Yusuf Dalva and Pinar Yanardag. NoiseCLR: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24209–24218, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

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, 2026. URL https://arxiv.org/abs/2606.01651.

Cheng Jin, Zhenyu Xiao, Chutao Liu, and Yuantao Gu. Angle domain guidance: Latent diffusion requires rotation rather than extrapolation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28187–28212. PMLR, 2025. URL https://proceedings.mlr.press/v267/jin25j.html.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. PnP inversion: Boosting diffusion-based editing with 3 lines of code. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FoMZ4ljhVw.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ANvmVS2Yr0.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=X6tNkN6ate.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=pd1P2eUBVfq.

Seunghoon Lee, Jeongwoo Choi, Byunggwan Son, JaeHyeon Moon, Jeimin Jeon, and Bumsub Ham. AccuQuant: Simulating multiple denoising steps for quantizing diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/9f8265edac41c4f8b4445b0d0cb83b80-Abstract-Conference.html.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12006–12016, 2024.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models, 2023. URL https://arxiv.org/abs/2305.12683. Working paper.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20763–20786. PMLR, 2023. URL https://proceedings.mlr.press/v202/liang23g.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Alexander Lobashev, Dmitry Guskov, Maria Larchenko, and Mikhail Tamm. Hessian geometry of latent space in generative models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40227–40252. PMLR, 2025. URL https://proceedings.mlr.press/v267/lobashev25a.html.

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://openreview.net/forum?id=aBsCjcPu_tE.

Matthew Niedoba, Berend Zwartsenberg, Kevin Patrick Murphy, and Frank Wood. Towards a mechanistic explanation of diffusion model generalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46389–46411. PMLR, 2025. URL https://proceedings.mlr.press/v267/niedoba25a.html.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26245–26265. PMLR, 2023. URL https://proceedings.mlr.press/v202/ning23a.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Sen Peng, Mingyue Wang, Jianfei He, Jijia Yang, and Xiaohua Jia. CAT: Contrastive adversarial training for evaluating the robustness of protective perturbations in latent diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48872–48885. PMLR, 2025. URL https://proceedings.mlr.press/v267/peng25e.html.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

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*, pp. 10684–10695, 2022.

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

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918. PMLR, 2023.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.310.

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.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tran. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2116–2127, 2023.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen editor and EditBench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18359–18369, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_Imagen_Editor_and_EditBench_Advancing_and_Evaluating_Text-Guided_Image_Inpainting_CVPR_2023_paper.html.

Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24534–24543, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Qingtao Yu, Jaskirat Singh, Zhaoyuan Yang, Peter Henry Tu, Jing Zhang, Hongdong Li, Richard Hartley, and Dylan Campbell. Probability density geodesics in image diffusion latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27989–27998, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24398–24407, 2024.

APPENDIX

TABLE OF CONTENTS

A Mathematical Details . . . 15
A.1 Finite-Path Decomposition and Local Interpretation . . . 15
B Source Construction . . . 15
B.1 Model-Native Source Variations . . . 15
B.2 Optimized Random Prediction-Space Projections . . . 17
B.3 Encoder-Only Constructor Control . . . 17
C Bounded-Source Matching . . . 18
C.1 Matched-Radius Construction . . . 18
C.2 Manipulation Checks . . . 20
D Evaluation Protocol . . . 20
D.1 Experimental Scope and Cohort Accounting . . . 20
D.2 Semantic Endpoints . . . 21
D.3 Target Execution and Paired Stochasticity . . . 22
D.4 Statistical Inference . . . 23
E Additional Quantitative Results . . . 24
E.1 Trajectory Displacement and Schedule-Refinement Robustness . . . 24
E.2 Same-Target Encoder-Only Constructor Control . . . 25
E.3 Exploratory Terminal-State Readout (FR-2) . . . 27
E.4 Population Endpoint Robustness . . . 29
E.5 Directional Response and Broader Recurrence . . . 29
E.6 Bounded-Source Matching Diagnostics . . . 30
F Qualitative Results . . . 30
G Reproducibility . . . 36
H Extended Related Work and Literature Positioning . . . 38
H.1 Source Perturbation and Protection Methods . . . 38
H.2 Geometry and Propagation Analyses . . . 39
H.3 Editing Benchmarks and Evaluation Context . . . 40

A Mathematical Details

The finite-path identities below explain the radius/direction decomposition under stated regularity assumptions and distinguish its continuous-map interpretation from finite-precision execution.

A.1 Finite-Path Decomposition and Local Interpretation

Assume $s \mapsto F_M(x + s\delta)$ is absolutely continuous on $[0, 1]$, F_M is differentiable at $x + s\delta$ for almost every s, the chain rule holds along the path almost everywhere, and the matrix-valued map $s \mapsto J_{F_M}(x + s\delta)$ is integrable on $[0, 1]$. Then the source-induced displacement admits the exact finite-path identity

$$\Delta z_{\text{init}} = \int_0^1 J_{F_M}(x + s\delta)\,\delta\,\mathrm{d}s = \underbrace{\left(\int_0^1 J_{F_M}(x + s\delta)\,\mathrm{d}s\right)}_{\bar{J}_{F_M}(x,\delta)}\delta. \tag{7}$$

This general identity explains why equal pixel budgets may yield unequal state displacements without assuming $\Delta z_{\text{init}} \approx J_{F_M}(x)\delta$ over the finite range.

Let $H_M(z) = \Psi(Y_M(z))$ and $\Delta q(r, u) = H_M(z_{\text{init}} + ru) - H_M(z_{\text{init}})$. Assume the analogous conditions on $[0, r]$: absolute continuity of $s \mapsto H_M(z_{\text{init}} + su)$, differentiability of H_M at the path points for almost every s, the chain rule along the path almost everywhere, and integrability of the matrix-valued map $s \mapsto J_{H_M}(z_{\text{init}} + su)$. Then

$$\Delta q(r, u) = \int_0^r J_{H_M}(z_{\text{init}} + su)u\,\mathrm{d}s. \tag{8}$$

Thus response depends on radius, direction, and Jacobian variation along the path. If H_M is differentiable at z_{init}, then locally

$$\Delta q(r, u) = rJ_{H_M}(z_{\text{init}})u + o(r). \tag{9}$$

These identities provide a continuous-map interpretation under the stated regularity assumptions; the local expression motivates radius and direction comparisons. The empirical interventions use the executed finite-precision pipeline; these identities need not hold exactly after quantization.

B Source Construction

To measure initial-state displacement, we construct bounded source perturbations without using output metrics for selection. This design does not imply statistical independence between the construction objective and the evaluated endpoints. The settings below specify the construction and controls.

B.1 Model-Native Source Variations

For every model, the model-specific preprocessing produces a native-resolution RGB source $x \in [0, 1]^d$; the same source initializes the model-native variation, random-sign control, and Gaussian control. Only source pixels are updated. Model-facing normalization is applied on the probe branch, and visual encoding uses the deterministic VAE posterior mean. All network parameters are fixed, and clean and candidate branches share probe noise, conditioning, and schedule settings. With $\mathcal{B}_\epsilon(x) = \{\tilde{x} \in [0, 1]^d : \|\tilde{x} - x\|_\infty \le \epsilon\}$, the common construction is

$$x_M^\star = \mathcal{A}_M(x, J_M, \mathcal{B}_\epsilon(x)) \in \mathcal{B}_\epsilon(x), \qquad J_M(\tilde{x}) = \frac{1}{|\mathcal{T}_M|}\sum_{j\in\mathcal{T}_M}\frac{1}{N_{M,j}}\left\|p_{M,j}(\tilde{x}) - \mathrm{sg}[p_{M,j}(x)]\right\|_2^2. \tag{10}$$

Here $\mathcal{A}_M$ is projected sign-gradient ascent, $p_{M,j}$ is the fixed model prediction at probe setting j, and $N_{M,j}$ is its number of scalar elements. The three models use their native prediction parameterizations and schedules. Neither Γ_M nor any reported semantic metric enters J_M.

All model-native constructions impose $\|x_M^\star - x\|_\infty \le 8/255$ and valid $[0, 1]$ range constraints. Projected sign-gradient ascent uses step size $1/255$, a uniform random feasible start, projection

Model	Model identifier	Source domain and preprocessing	Serialization and model-facing normalization
SDXL	`stabilityai/stable-diffusion-xl-base-1.0`	1024×1024 RGB; EXIF transpose, then centered `PIL.ImageOps.fit` with LANCZOS at centering $(.5, .5)$.	Clamp $[0, 1]$, multiply by 255, round to uint8 HWC RGB, and save as Pillow PNG. Probe input is uint8$/255$ in FP32, then $2x - 1$ for the VAE.
SD1.5	Stable Diffusion 1.5	Native resolution and crop/resize operation unavailable.	FP32 master; $8/255$ cap; selected image serialized to uint8. Model-facing normalization unavailable.
FLUX	`black-forest-labs/FLUX.1-dev`	1024×1024 RGB; if needed, whole-image LANCZOS resize with no crop.	Same uint8 PNG rule as SDXL. Probe input is uint8$/255$ in FP32, then the VAE processor maps it to $[-1, 1]$.

Table 6: Source-probe model identifiers and preprocessing. Missing SD1.5 fields are reported as unavailable rather than inferred.

Model	Prediction probe and conditioning	Optimization and paired probe noise	Encoding, iterate selection, and candidate policy
SDXL	Prompt "a faithful image-to-image reconstruction"; no negative prompt; ϵ-prediction at DDPM $t = 200$; positive embeddings only and classifier-free guidance disabled.	50 sign-gradient steps of $1/255$; iid uniform start in $[-8/255, 8/255]$ followed by range/L_∞ projection. One per-image keyed latent Gaussian is reused for clean and every iterate; the clean prediction is detached. Base seed 3407.	Fixed SDXL VAE posterior mean with scaling factor; BF16 probe. Maximum objective over the initial 50 iterates plus the final projected candidate, with the first exact tie retained. Continuous FP32 optimization; the selected best iterate is serialized once to uint8 PNG and not rescored.
SD1.5	ϵ-prediction discrepancy averaged over PNDM points $501, 251, 1$ with CFG $= 5$; prompt unavailable.	8 steps of $1/255$ from a uniform feasible start; projection after every step and shared clean/candidate probe noise. Random-start seed and fixed-noise construction unavailable.	Deterministic VAE posterior mean and best-objective continuous iterate, followed by one uint8 serialization. VAE configuration and tie implementation unavailable.
FLUX	Prompt "a faithful image-to-image reconstruction"; empty negative prompt; mean squared flow-velocity discrepancy at executed FlowMatch points 759.5109252929688, 512.8441162109375, and 310.90106201171875 (integration indices $[0, 2, 3]$); guidance embedding $= 3.5$, without true negative-prompt CFG.	8 sign-gradient steps of $1/255$; iid uniform start in $[-8/255, 8/255]$ followed by range/L_∞ projection. One per-image seeded encoded-latent Gaussian is reused for clean and every iterate at all three points. Base seed 2025.	Fixed FLUX VAE posterior mean with shift/scaling transform; BF16 probe. Maximum objective over the initial 8 iterates plus the final projected candidate, with the first exact tie retained. Continuous FP32 optimization followed by one uint8 PNG serialization.

Table 7: Source-probe objectives and optimization settings. The numerical FLUX points are the executed values, not nominal schedule fractions.

after every step, and retention of the best-objective continuous iterate. SDXL runs 50 steps with ϵ-prediction at DDPM $t = 200$ and no classifier-free guidance during construction. SD1.5 runs 8 steps and averages ϵ-prediction discrepancy over PNDM points 501, 251, and 1 with classifier-free guidance 5. In the verified P4 configuration, FLUX runs 8 steps and averages flow-velocity discrepancy over executed FlowMatch points 759.5109252929688, 512.8441162109375, and 310.90106201171875, corresponding to probe integration indices $[0, 2, 3]$, with guidance embedding 3.5.

The generic native controls begin from the same target-preprocessed source as the native perturbation. For the random-sign control, an independent uniform sign field is applied under the same $8/255$ cap and valid-range projection; it is deliberately *not* matched to the model-native variation's source RMS. For the capped, pixel-RMS-matched Gaussian control, a Gaussian field is scaled under the same cap and valid range until the uint8-serialized perturbation RMS matches that of the saved model-native image within 0.1 intensity level on the 0–255 scale. Thus the random-sign control shares the source and equal ℓ_∞ budget, whereas the Gaussian control also matches native serialized pixel RMS.

All Gaussian controls remain within the $8/255$ budget. The 0.1-level tolerance above is a design criterion on absolute serialized-RMS error; the available receipts do not establish a tighter achieved-error bound. The median random-sign-minus-native source RMS is $+2.604$ integer levels for SDXL, $+3.004$ for SD1.5, and $+2.339$ for FLUX. Thus source RMS alone cannot explain the larger native state displacement.

Controls for the cross-model comparisons. The cross-model experiment uses two sets of controls. Controls constructed for the target model (*target-model controls*) accompany native sources; controls constructed with the SDXL sources (*SDXL-source controls*) accompany transferred sources. For SD1.5, the former are constructed at 512×512, whereas the latter are constructed at 1024×1024 and resized to the target's 512×512 with LANCZOS. The SD1.5 Gaussian controls also differ in random realization and in the optimized-source RMS anchor. For FLUX, both sets are 1024×1024 but use different random realizations; native Gaussian controls match the FLUX-optimized source RMS, while transferred Gaussian controls match the SDXL-optimized source RMS. A shared pixel budget therefore implies neither a shared pixel RMS nor a shared control realization or source-processing path. These verified sizes apply here; other missing SD1.5 settings remain unavailable.

B.2 Optimized Random Prediction-Space Projections

The probe-specificity control asks whether optimizing an arbitrary prediction-space direction produces the same source-to-state transfer as the native full-discrepancy probe. It uses 24 development images per model, disjoint from the bounded-source holdout, and reuses native optimized sources. Each image undergoes eight direction runs with $K = 1$ projection per probe location. Budget, random start, optimizer, iteration count, step size, preprocessing, probe locations, conditioning, precision, best-iterate retention, and serialization match the corresponding native constructor. Γ_M is evaluated only after optimization; completion counts are reported in Appendix D.1.

The projection objective is

$$J_{\mathrm{proj}}(\tilde{x}) = \frac{1}{|\mathcal{T}_M|} \sum_{j \in \mathcal{T}_M} \langle v_j, \mathrm{vec}(p_{M,j}(\tilde{x}) - \mathrm{sg}[p_{M,j}(x)]) \rangle^2 .$$

For each run, one normalized direction $v_j \in \mathbb{R}^{N_j}$ is generated independently at each probe location, with independent Rademacher entries $\pm 1/\sqrt{N_j}$, where N_j is the number of scalar prediction elements. The vectors remain fixed during optimization, and clean predictions are detached. Predictions from different probe locations are not concatenated: each projection is squared before averaging over probe locations.

The eight direction runs are repeated measurements whose evaluated Γ_M values are averaged within image; images are the statistical units. Inference uses 10,000 image-level bootstrap resamples and a separate studentized max-t simultaneous 95% family over the three model-level native-minus-projection contrasts. The mean contrast is positive for every tested direction in all three models.

Model	Native Γ_M	Projection Γ_M	Native−projection [95% simultaneous CI]	Ratio	Win rate
SDXL	0.796667	0.723326	0.073341 [0.070454, 0.076228]	1.101	100.0%
SD1.5	0.347447	0.289603	0.057844 [0.051538, 0.064150]	1.200	100.0%
FLUX	0.230333	0.180592	0.049740 [0.045531, 0.053949]	1.275	100.0%

Table 8: Optimized random prediction-space projection control. Each image is optimized against eight fixed one-dimensional Rademacher prediction-space projections under the same $8/255$ source budget and optimizer as the native probe. Projection directions are averaged within image before inference; images are the statistical units. Intervals form one simultaneous three-model family.

All native and optimized-projection perturbations have source $L_\infty = 0.031373$. Native versus optimized-projection pixel RMS is 0.020967 versus 0.021667 for SDXL, 0.019336 versus 0.019340 for SD1.5, and 0.021936 versus 0.018828 for FLUX. Native sources therefore have no uniform source-energy advantage, although their perturbation geometry is not fully matched to the controls.

B.3 Encoder-Only Constructor Control

The same-target control compares the existing P4 prediction-discrepancy constructor (PD) with an encoder-displacement-only constructor (ENC). For target M, ENC maximizes

$$J_{\mathrm{enc},M}(\tilde{x}) = \frac{\|h_M(\tilde{x}) - \mathrm{sg}[h_M(x)]\|_2^2}{N_h}, \tag{11}$$

where h_M is the deterministic model-facing source-probe posterior-mean representation and N_h is its number of scalar elements. For SDXL, $h_M = 0.13025\,\mu_M$ has shape $[1, 4, 128, 128]$; for FLUX, $h_M = (\mu_M - 0.1159)\,0.3611$ has shape $[1, 16, 128, 128]$, where μ_M denotes the VAE posterior mean after the shared target preprocessing. The encoders execute in BF16 and objective reductions in FP32. This deterministic source-probe representation is distinct from the executed target initial state used for matching.

ENC performs zero denoiser, transformer, or flow prediction-backbone calls. Neither DINO nor any output endpoint, target z_{init} radius, or trajectory enters ENC optimization. The realized initial-state radius is measured only after source construction.

Within each target, PD and ENC share the source/preprocessing, $L_\infty \le 8/255$ cap, FP32 optimization master, projected sign-gradient updates, step size $1/255$, projection onto the valid range and source budget, and iteration count (SDXL: 50; FLUX: 8). They use the same historical random-start procedure and constructor seed where recoverable, retain the first iterate on objective ties, include the final candidate in best-iterate selection, and serialize the selected continuous candidate to uint8 once. PD endpoints are reused original P4 target-native artifacts after hash verification. Equal iteration count does not imply equal compute or equal convergence.

For FLUX ENC, deterministic algorithms were enabled because the default backward path failed a fixed-input gradient-repeat check. The corresponding historical PD backend-determinism setting was not preserved. This limits source-construction-runtime fairness: identical low-level backend flags and equal FLOPs are not established. Downstream P4 target execution is unchanged. Appendix E.2 reports the reused-cohort comparison after pair-specific radius matching.

C Bounded-Source Matching

Direct equal-radius state interventions do not guarantee source realizability. The matcher implements the corresponding source-space control by selecting bounded images using target-state geometry before semantic evaluation.

C.1 Matched-Radius Construction

Inputs. For each image i, target $M \in \{\text{SDXL}, \text{FLUX}\}$, and branch $b \in \{\text{native}, \text{transferred}\}$, the matcher receives the serialized clean PNG q_i^{clean}, the serialized probe-selected branch endpoint $q_{i,b}^{\text{end}}$, the target-state map F_M, the seeds $\mathcal{S} = \{2025, 2026, 2027\}$, the integer source budget 8, and $\rho \in \{.5, .7, .9\}$.

Three-seed target geometry. For every serialized candidate q and $s \in \mathcal{S}$, the executed state displacement and matching radius are

$$\begin{aligned} \Delta z_M(q, s) &= F_M(q, s, c) - F_M(q_i^{\text{clean}}, s, c), \\ r_M(q, s) &= \|\Delta z_M(q, s)\|_2, \\ \bar{r}_M(q) &= \frac{1}{3} \sum_{s \in \{2025, 2026, 2027\}} r_M(q, s). \end{aligned} \tag{12}$$

The native and transferred endpoint radii in Eq. 6 are $\bar{r}_M(q_{i,b}^{\text{end}})$. Candidate selection uses $\bar{r}_M(q)$, not a single-seed radius. The same three seeds generate the clean, native-selected, and transferred outputs; DINO dissimilarities are averaged over paired seeds within image before population inference.

Development before output evaluation. The initial scalar-grid matcher (V1) met the 2% tolerance in only 75/120 cells of a 20-image pilot (maximum mismatch 10.535%). A quantization-aware refinement (V2) reached 120/120 pilot cells (maximum 0.083%), but stopped on the fresh cohort when an aggregate radius decreased, violating its monotonicity assumption. No output generation or semantic evaluation occurred in that stopped run. The final construction (V3) uses fixed event prefixes and global nearest-radius selection without assuming monotonicity. It passed 140/140 development cells at each tested radius, excluded development images from the holdout, and was specified before final semantic generation. This history documents protocol development, not preregistration.

Candidate path. The clean-to-endpoint path is decomposed into adjacent canonical uint8 intervals. Within each interval, the matcher uses $B = 256$ prefix bins; the full path can therefore contain more than 256 candidates. Shared interval endpoints are deduplicated before global nearest-radius selection. Candidate generation is fixed before target-state measurement and uses no semantic score or target radius. No monotonicity assumption is made about the realized radius.

Exact recovered event rule. Decode uint8 HWC RGB and flatten in C order, so coordinate $c = ((\text{row}\, W) + \text{column})3 + \text{channel}$. Convert to NCHW FP32 and divide by 255. Evaluate the canonical interpolation grid $k/256$, $k = 0, \ldots, 256$, using the recovered FP32 interpolation and Q8 routine: clamp to $[0, 1]$, multiply by 255, apply `torch.round` (ties to even), and convert to uint8. Deduplicate serialized states retaining first occurrence; real-number threshold arithmetic is not an equivalent substitute.

For each adjacent distinct canonical pair $(q_\ell, q_{\ell+1})$, emit $|q_{\ell+1,c} - q_{\ell,c}|$ signed unit events at each changing coordinate. Sort coordinates by byte-lexicographic SHA256 of UTF8(salt), NUL, UTF8(image ID), NUL, and the ASCII decimal coordinate index, breaking hash ties by coordinate index. Within a coordinate, event ordinals are $0, \ldots, |q_{\ell+1,c} - q_{\ell,c}| - 1$; signs follow the coordinate difference. The unchanged protocol salt is stored in `configs/p4_matching.json` in the artifact. For interval event count m, use

$$\mathcal{P}(m) = \begin{cases} \{0, \ldots, m\}, & m \le 256, \\ \text{sorted unique}\{\lfloor jm/256 \rfloor : j = 0, \ldots, 256\}, & m > 256. \end{cases} \tag{13}$$

This yields up to 257 prefixes per interval before shared-endpoint removal, and does not enumerate every unit prefix when $m > 256$. Enumerate intervals and prefix counts in ascending order, retaining the first interval's zero prefix and omitting later zero prefixes. The latter must match the preceding high endpoint; a boundary hash disagreement or unexpected non-adjacent duplicate raises an error. Both clean and optimized endpoints are included.

Measure each candidate with scalar batch size one and a separately reset paired generator for each seed, preserving target posterior sampling, scheduler noise, and packing. Compute each displacement norm in FP64, then average the three norms as in Eq. 12; do not take the norm of a seed-averaged tensor. Selection uses the lexicographic key (absolute mean-radius error, zero-based canonical interval index, prefix count, RGB SHA256), with stable mergesort, no tolerance-based tie, and no semantic tie-break.

Algorithm 2 Paired-endpoint matched-radius selection: FrozenMatch

Require: F_M, q^{clean}, endpoint pair $\mathcal{E} = \{q_b^{\text{end}}\}_b$, fixed context c, paired seeds $\mathcal{S}$, fraction ρ; $B = 256$ bins per interval

for all branches b **do**
  $\mathcal{C}_b \leftarrow$ first-occurrence-unique FP32/Q8 canonical states at $k/256$
  $\mathcal{Q}_b \leftarrow \emptyset$
  for all adjacent states in $\mathcal{C}_b$, in interval order ℓ **do**
    form hash-ordered signed unit events; let m be their count
    for all $p \in \mathcal{P}(m)$ in ascending order, omitting $p = 0$ if $\ell > 0$ **do**
      append candidate $(q, \ell, p, \text{SHA256}_{\text{RGB}}(q))$ to $\mathcal{Q}_b$
  validate shared endpoints and absence of unexpected duplicates
  for all $q \in \mathcal{Q}_b$ **do**
    for all $s \in \mathcal{S}$ **do**
      compute $r_M(q, s) = \|F_M(q, s, c) - F_M(q^{\text{clean}}, s, c)\|_2$ in FP64
    $\bar{r}_M(q) \leftarrow |\mathcal{S}|^{-1} \sum_{s \in \mathcal{S}} r_M(q, s)$
$r^\star \leftarrow \rho \min_b \bar{r}_M(q_b^{\text{end}})$
for all branches b **do**
  stable-sort $\mathcal{Q}_b$ by $(|\bar{r}_M(q) - r^\star|, \ell, p, \text{SHA256}_{\text{RGB}}(q))$
  $q_b^\star \leftarrow$ first candidate; serialize using the recovered PNG routine
  validate coordinatewise path membership and $\|q_b^\star - q^{\text{clean}}\|_\infty \le 8$
return source pair $\mathcal{Q}^\star = \{q_b^\star\}_b$ and radii $\mathcal{G}^\star = \{r_M(q_b^\star, s)\}_{b,s}$

Here FrozenMatch names the paired operation in the pseudocode, not a literal code entry point. The supplied portable `core_p4.cli.match` orchestrates both paths and calls the recovered single-path `_select_global`; it is not the historical producing script. Algorithm 2 checks each selected

source's feasibility and returns both sources and measured radii. Algorithm 1 requires the cohort gate before any matched-source semantic evaluation; a failure stops progression without dropping, replacing, or rematching cells.

C.2 Manipulation Checks

Every selected candidate is therefore a valid serialized image satisfying an integer-level L_∞ budget of 8, equivalent to $8/255$ under the specified serialization convention. For manipulation checks, the aggregate and per-seed pairwise relative mismatches are

$$e_{\text{pair}} = \frac{|\bar{r}_{\text{native}} - \bar{r}_{\text{transferred}}|}{(\bar{r}_{\text{native}} + \bar{r}_{\text{transferred}})/2}, \qquad e^{(s)}_{\text{pair}} = \frac{|r_{\text{native}}(s) - r_{\text{transferred}}(s)|}{(r_{\text{native}}(s) + r_{\text{transferred}}(s))/2}. \tag{14}$$

Matching was optimized using the seed-averaged radius $\bar{r}_M(q)$, and the 2% tolerance applies to that matched manipulation. A separate paired-seed check uses $e^{(s)}_{\text{pair}}$ to assess individual executions; Table 3(a) reports its results without redefining the matcher as seed-wise.

The archived `configs/p4_matching.json` specifies the post-selection cohort gate: all 100 primary image–target pairs at $\rho = .9$ must have aggregate mismatch at most 2%; at each secondary radius separately, at least 95/100 must meet 2% and the maximum must be at most 5%. All 900 seed-wise pairs must be finite and within 2%, with defined directions. These checks do not change the selected candidates.

D Evaluation Protocol

The cohort and execution details below specify how initial-state displacement, directional response, and domain dependence are evaluated. P1–P4 identify the tests in Section 3.5; schedule refinement, the same-target encoder-only constructor control, and final-state radius matching (FR-2) are separate exploratory follow-ups on the reused P4 cohort.

D.1 Experimental Scope and Cohort Accounting

Study	Models	Domain	Images	Seeds	Evidential role
Population response	SDXL	FFHQ	500	3	Establish endpoint sensitivity at population scale
Source-to-state transfer	SD1.5, SDXL, FLUX	FFHQ	50/model	3	Test coupling from the native probe to Γ_M
Probe-specificity control	SD1.5, SDXL, FLUX	FFHQ development	24/model	8 directions	Test whether the optimized random prediction-space projections reproduce native source-to-state transfer
Initial-state replay	SDXL	FFHQ	10	fixed	Validate capture and reinjection of z_{init}
Direct radius/direction	SDXL, FLUX	FFHQ	50	3	Intervene on state magnitude and direction
Cross-model transfer	SD1.5, FLUX targets	FFHQ	50	3	Separate radius mismatch from directional response
Bounded-source test	SDXL, FLUX	fresh FFHQ	50	3	Test direction at matched reachable radius
Downstream amplification	SDXL, FLUX	fresh FFHQ	50/target	3	Test norm growth as an explanation; describe DINO relative to final RMS
Schedule-refinement robustness	SDXL, FLUX	FFHQ	same P4 50/target	3	Exploratory test of output and trajectory ordering persistence under fourfold integration refinement
Same-target constructor control	SDXL, FLUX	FFHQ	same P4 50/target	3	Exploratory test of prediction-discrepancy versus encoder-displacement-only source constructions at approximately matched realized initial-state magnitude.
FR-2 final-state readout	SDXL, FLUX	same P4 images	50/target	3	Exploratory comparison at approximately matched executed final-state radii
COCO external validity	SDXL, FLUX	COCO	200	3	Test recurrence beyond portraits

Table 9: Complete experimental scope of the currently reported studies. Images are the statistical units; paired seeds and random directions are repeated measurements averaged within image unless otherwise stated. The bounded-source holdout was fixed from 550 previously unused eligible images; FR-2, schedule refinement, and the constructor control reuse those images; none is a new cohort. All 100 constructor-control image–target units passed. All 2,100 planned P4 outputs completed.

The same-target constructor control reuses the original P4 50 images per target; it is not a new holdout. All 100/100 image–target units passed, with zero failures, exclusions, replacements, or rescue searches.

COCO annotation accounting. For the 200-image non-face COCO cohort, double-labeling covered 140 images (90.7% agreement, $\kappa = 0.8094$), with 13 adjudications. These annotation checks concern cohort selection; the response contrasts remain unmatched source-to-state and output comparisons.

Bounded-source holdout provenance. Eligibility was determined by prior experimental assignment, before semantic evaluation. The archived summary gives the counts below; geometry-development and earlier-semantic assignments form one ineligible union because separate category counts were not recorded.

Provenance stage	Count	Selection rule or check
Total FFHQ sampling frame	3,400	Frame established before the P4 holdout.
Previously assigned and ineligible	2,850	Union of geometry-development and earlier-semantic assignments.
Untouched eligible pool	550	No prior assignment represented in the ineligible union.
Final P4 holdout	50	Fixed from the eligible pool before P4 semantic evaluation.
Geometry-development overlap	0	Checked on the final holdout.
Prior semantic exposure	0	Checked on the final holdout.

Table 10: P4 sampling-frame provenance. Eligibility was determined independently of P4 outcomes; no semantic outcome was used for matching or exclusion.

Completed evaluations. All 576 prediction-space projection optimizations completed with zero failures or exclusions. The bounded-source experiment produced all 2,100 expected outputs (1,050 per target). The trajectory analysis used the primary-radius holdout, with 50 images per target, paired seeds $\{2025, 2026, 2027\}$, and clean, native, and transferred branches; all 900 trajectories completed with zero failures or exclusions and reproduced the archived output pixels exactly.

The completed schedule-refinement follow-up reuses the original P4 $\rho = .9$ cohort (50 images per target), with three paired seeds and clean, native, and transferred branches. All 900 branch-level original/refined initial states and first effective predictions match exactly, including the archived first-capture hashes. This is a reused, previously inspected cohort, not a new holdout.

D.2 Semantic Endpoints

DINOv2 dissimilarity. For the matched-radius experiment, ϕ is the DINOv2 evaluator (Oquab et al., 2023), implemented as `timm vit_small_patch14_dinov2.lvd142m`, a 384-dimensional Vision Transformer representation. RGB input is bicubically resized to 518×518 with antialiasing and crop fraction 1.0 (equivalently a 518×518 center crop), then normalized channelwise by mean $(.485, .456, .406)$ and standard deviation $(.229, .224, .225)$. The forward representation is the single CLS/global token pool, with no register tokens; its flattened FP32 vector is L_2 normalized. Writing $\widehat{\phi}(y) = \phi(y)/\|\phi(y)\|_2$, the endpoint is

$$D_{\mathrm{DINO}}(y, y_{\mathrm{ref}}) = 1 - \widehat{\phi}(y)^{\top}\widehat{\phi}(y_{\mathrm{ref}}). \tag{15}$$

Larger values mean greater feature difference, not output quality or an exhaustive measure of semantics. For every evaluated branch, y_{ref} is the paired clean-source output under the same target model, prompt, and seed. DINO is fixed as the primary semantic endpoint across both P4 targets; ArcFace, CLIP-image, and LPIPS provide secondary robustness analyses.

Secondary endpoints. ArcFace uses InsightFace `buffalo_l`, with the `w600k_r50` recognition model and `det_10g` detector. Detector input is 640×640 and recognition input is 112×112. The P4 endpoint is identity dissimilarity to the paired generated-clean output; P4 contains zero invalid identity rows.

CLIP-image uses OpenCLIP ViT-B/32 with `laion2b_s34b_b79k` weights and measures cosine dissimilarity to the paired generated-clean output. LPIPS uses an AlexNet backbone with 256×256 evaluation input and measures distance to the paired generated-clean output.

The 500-image population analysis and the 50-image P4 analysis use the same recorded secondary-evaluator configuration and model assets, while their endpoint references differ as specified above. Some library-internal alignment, preprocessing, and reduction details were not preserved. DINO remains primary; the other endpoints remain post-hoc secondary analyses.

D.3 Target Execution and Paired Stochasticity

Target	Model and source preprocessing	Captured z_{init}	Trajectory, conditioning, and paired stochasticity
SDXL	`stabilityai/stable-diffusion-xl-base-1.0`; 1024^2 RGB mapped to $[-1, 1]$.	Shape $[1, 4, 128, 128]$: VAE posterior sample with the model scaling factor and Euler noise at the first active timestep, captured immediately before denoising.	Strength .25; Euler scheduler; 20 requested steps with active timesteps $[201, 151, 101, 51, 1]$; CFG $w = 5$; prompt "a realistic portrait photograph of the same person"; BF16; 1024^2 output. The same seed is reset for clean, native, and transferred branches.
FLUX	`black-forest-labs/FLUX.1-dev`; 1024^2 RGB mapped to $[-1, 1]$.	Shape $[1, 4096, 64]$: VAE posterior sample with model shift/scaling, FlowMatch noise scaling, and native 2×2 packing, captured immediately before the flow loop.	Strength .5; FlowMatch Euler scheduler; 8 requested steps with active timesteps $[759.5109252929688, 654.5667724609375, 512.8441162109375, 310.90106201171875]$; guidance embedding 3.5; prompt "a faithful image-to-image reconstruction"; BF16; 1024^2 output. The same paired-seed rule is used.

Table 11: Target execution and captured initial state in bounded-source radius matching.

The SDXL scheduler configuration records 1,000 training steps, scaled-linear β from 0.00085 to 0.012, ϵ prediction, leading timestep spacing, steps offset 1, linear interpolation, no Karras sigmas, `set_alpha_to_one=false`, and `clip_sample=false`. The FLUX scheduler configuration records 1,000 training steps, dynamic shifting enabled, base/max sequence lengths $256/4096$, base/max shifts $.5/1.15$, and scheduler shift 3.0. In the executed 1024^2 P4 configuration, native 2×2 packing gives sequence length 4096 and dynamic-shift $\mu = 1.15$.

Under these SDXL and FLUX configurations, the captured z_{init} is the only source-derived visual input to the downstream trajectory; ControlNet and IP-Adapter inputs are disabled.

COCO execution. The separate COCO comparison uses 200 images per target at 1024×1024 resolution, BF16 precision, and seeds $\{2025, 2026, 2027\}$. Both targets use the prompt "a faithful photograph of the same scene, natural lighting, highly detailed" and an empty negative prompt. SDXL uses `stabilityai/stable-diffusion-xl-base-1.0`, guidance 5.0, strength 0.25, 20 requested inference steps, and `EulerDiscreteScheduler`. FLUX uses `black-forest-labs/FLUX.1-dev`, guidance 3.5, strength 0.5, 8 requested inference steps, and `FlowMatchEulerDiscreteScheduler`. The exact executed active timestep/sigma arrays for the historical COCO run were not preserved.

There were 2,400 unique branch executions per target and 4,800 pooled across the two targets, with zero recorded failures. Paired branches share the recorded stochastic inputs within each image/seed. Geometry and semantic results use $N = 200$ image-level units with seed averaging. COCO remains an unmatched source-to-state/output comparison, not a matched-radius or trajectory replication.

Fourfold integration-schedule refinement. The exploratory refinement reuses the exact original P4 $\rho = .9$ source PNGs and initial-state construction for all clean, native, and transferred branches. Original prompts, conditioning/guidance, model revisions, paired seeds $\{2025, 2026, 2027\}$, and start/end noise horizon are preserved. Each frozen, originally executed integration interval is subdivided fourfold: SDXL increases from 5 to 20 active updates, and FLUX from 4 to 16. This explicitly refines the frozen intervals rather than requesting a generic increase in `num_inference_steps`. All 900 original/refined initial states and first effective predictions match exactly, including archived first-capture hashes. There is no source reconstruction, re-optimization, or rematching. Each refined clean output is the paired DINO reference for the corresponding refined native and transferred outputs. The exact original and refined executed timestep/sigma arrays are provided in the accompanying executable artifact.

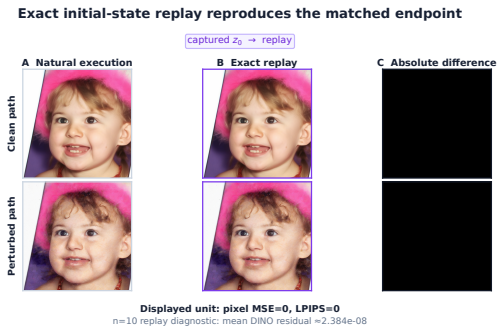

Figure A1: Initial-state replay check ($n = 10$). Natural and replayed executions are identical under fixed SDXL settings (pixel MSE $= 0$, LPIPS $= 0$); the mean DINO residual is $\approx 2.384 \times 10^{-8}$. One representative unit is shown.

D.4 Statistical Inference

Images are the statistical units, and the three paired seeds are averaged within image. Confidence intervals use 10,000 image-cluster bootstrap resamples. Simultaneous studentized max-t intervals are computed within each stated contrast family. For P4, only the DINO contrasts at $\rho = .9$ for the FLUX and SDXL targets form the primary simultaneous family. The exploratory final-state comparison, FR-2, uses the same images in a separate two-target family: seed-specific native-minus-transferred contrasts are first averaged within image, common bootstrap indices are matched by shared source-image ID, and the bootstrap seed is 20260726. Results at $\rho = .5$ and .7 are secondary and exploratory; their pointwise intervals are descriptive. CLIP-image, ArcFace, and LPIPS are post-hoc secondary robustness endpoints. Earlier cross-model intervals retain the seven original simultaneous families listed below; they do not form one family across the displayed rows. Family-wise control applies within each stated family; no global FWER claim is made across experiments, radii, endpoints, or the adaptive history that motivated FR-2.

Exploratory schedule-refinement families. Refined DINO uses one new two-target simultaneous studentized max-t family with 10,000 image-cluster bootstrap resamples. Refined trajectory displacement uses a separate two-target family with the same resample count and interval method. Images remain the statistical units; seeds $\{2025, 2026, 2027\}$ are averaged within image before inference. For trajectory displacement, G_t is first averaged over all post-initial refined checkpoints, then over seeds within image; no time-weighted or final-only replacement is used. Neither family shares error control with original P4, the original trajectory analysis, FR-2, the other refinement family, or any other experiment. These exploratory families do not enlarge the original P4 family or provide global FWER control. The executable artifact records frozen bootstrap seeds 2026092201 for refined DINO and 2026092202 for refined trajectory; these preserve the original separate families.

Exploratory same-target constructor family. The FLUX PD-minus-ENC and SDXL PD-minus-ENC DINO contrasts form one new exploratory two-target family on the reused P4 cohort. Images are the inference units; the three seeds are averaged within image before 10,000 image-cluster bootstrap resamples. Common resampling indices across targets, aligned by image ID, give studentized max-t simultaneous 95% intervals with frozen RNG seed 2026092207. This family is separate from P4, original trajectory, schedule-refinement DINO, schedule-refinement trajectory, FR-2, and every historical family, including the descriptive Q family. It does not enlarge P1–P4 or establish global FWER control.

Original cross-model families. Table 21 retains the original inference for all 13 contrasts. F1–F7 below identify statistical families, not additional hypotheses. The table's visual groups A–D do not define new families. In particular, F3 and F6 each jointly cover an initial-state RMS contrast and a DINO contrast, while the displayed F7 row belongs to an eight-comparison family.

ID	Original family	Rows	Full membership	Size	RNG seed
F1	Stage2C B1	01, 02	SD1.5 native versus target-model random-sign and Gaussian controls	2	20260727
F2	Stage2C A1	03, 04	SD1.5 transferred versus SDXL-source random-sign and Gaussian controls	2	20260727
F3	Stage2D recovered E1	05, 11	SD1.5 native–transferred RMS and recovered DINO contrasts, jointly	2	20260730
F4	Stage2C B2	06, 07	FLUX native versus target-model random-sign and Gaussian controls	2	20260727
F5	Stage2C A2	08, 09	FLUX transferred versus SDXL-source random-sign and Gaussian controls	2	20260727
F6	Stage2C E2	10, 12	FLUX native–transferred RMS and DINO contrasts, jointly	2	20260727
F7	Stage2D D_native	13	Native versus Gaussian, random-sign, reachable-random, and SDXL directions, each at native/control reference radii	8	20260739

Table 12: Original simultaneous families for the cross-model results. Row IDs refer to Table 21; F7 includes seven comparisons not displayed there. The RNG seed 20260739 is an integer, not a date.

All seven families use 10,000 valid image-cluster bootstrap resamples, paired indices within the original family, and equal image weighting after averaging three execution seeds. Point estimates are observed sample means. Intervals use the two-sided maximum absolute centered bootstrap t statistic, replicate-specific standard errors with `ddof=1`, and the linear 0.95 quantile. Coverage does not extend jointly across the 13 displayed rows. Historical Track B/Stage2C B rows use controls constructed for the target model; Track A/Stage2C A rows use controls constructed with the SDXL sources. The Stage2D recovered E1 and D_native identifiers retain the selected historical versions.

Why control-relative contrasts do not cancel. Let R_N, R_T be clean-relative RMS displacements for native and transferred sources, and C_N, C_T those of their respective controls. Then

$$\mathrm{mean}(R_N - C_N) - \mathrm{mean}(R_T - C_T) = \mathrm{mean}(R_N - R_T) + \mathrm{mean}(C_T - C_N).$$

The two controls differ, so their offset cannot be dropped. The archived scalar records satisfy this identity for all 200 image/control records. The two constructions share 50 image IDs, three execution seeds, weights, paired clean states, and noise, but their control-source and state hashes differ in all 150 image–seed pairs per target/control type. Appendix B.1 gives the construction differences. The audit does not separate causal contributions from resizing, random realization, and optimization.

E Additional Quantitative Results

We first report trajectory displacement and schedule refinement, then the same-target encoder-only constructor control and the supplementary exploratory terminal-state readout (FR-2). We then give the complete population endpoint family (Table 20), cross-model contrasts (Table 21), and broader COCO recurrence (Table 22). Intervals retain their stated families, and comparisons remain within each target's coordinates.

E.1 Trajectory Displacement and Schedule-Refinement Robustness

Propagation of matched initial displacements. The primary bounded-source result leaves open whether larger native output responses accompany greater trajectory separation on the $\rho = .9$ holdout. This paired separation under a fixed target differs from prediction-error accumulation associated with train–test input mismatch (Ning et al., 2023) or quantization relative to a full-precision model (Lee et al., 2025). For branch $b \in \{\mathrm{native}, \mathrm{transferred}\}$, define

$$\Delta z_t^b = z_t^b - z_t^{\mathrm{clean}}, \qquad G_t^b = \frac{\mathrm{RMS}(\Delta z_t^b)}{\mathrm{RMS}(\Delta z_{\mathrm{init}}^b)}. \tag{16}$$

The trajectory summary $\overline{G}^b$ is the arithmetic mean of G_t^b over all post-initial executed checkpoints. Paired seeds are averaged within image before inference; target means average over images. The two

native-minus-transferred target contrasts form one 10,000-resample simultaneous studentized max-t family.

Target	Native $\overline{G}$	Transferred $\overline{G}$	$\Delta\overline{G}$	Simultaneous 95% CI	Win rate
FLUX	2.236623	2.122521	$+0.114102$	$[0.108191, 0.120013]$	100%
SDXL	0.809012	0.864937	-0.055925	$[-0.057526, -0.054324]$	0%

Table 13: Downstream Euclidean amplification. Win rates are image-level native-greater-than-transferred rates after seed averaging. The two intervals form one simultaneous target family.

At every post-initial executed checkpoint, the cohort-mean normalized displacement is larger for native FLUX and transferred SDXL. At initialization, $G = 1$ by construction. The table reports image-level win rates for the trajectory summary, not per-sample ordering at every checkpoint. Despite positive native-minus-transferred P4 DINO contrasts in both targets (Section 4.4), mean normalized trajectory displacement is smaller for native in SDXL and larger in FLUX. Systematically greater native Euclidean growth is therefore not a shared explanation in these configurations. This does not isolate architecture from schedule and conditioning or exclude nonlinear magnitude–direction interactions; recorded checkpoints do not establish continuous-time monotonicity.

Schedule-refinement robustness. Table 14 reports the completed exploratory follow-up under the frozen interval subdivision in Appendix D.3. The trajectory summary uses the same normalized displacement definition as Eq. 16, averaging every post-initial refined checkpoint before seeds within image.

Target	Updates	Native	Transferred	Native−transferred	Simultaneous 95% CI	Win
A. Refined DINOv2 dissimilarity						
FLUX	16	0.070255	0.058751	$+0.011504$	$[0.007068, 0.015940]$	90%
SDXL	20	0.054397	0.044794	$+0.009603$	$[0.003925, 0.015281]$	72%
B. Refined mean normalized trajectory displacement $\overline{G}$						
FLUX	16	2.014843	1.921106	$+0.093737$	$[0.088901, 0.098574]$	100%
SDXL	20	0.846216	0.892487	-0.046271	$[-0.047602, -0.044940]$	0%

Table 14: Exploratory schedule-refinement robustness on the same original P4 $\rho = .9$ 50-image cohort per target, with three paired seeds averaged within image. Each original integration interval is subdivided fourfold; refined clean outputs are the paired DINO references, with no rematching. DINO and $\overline{G}$ use separate two-target 10,000-resample simultaneous studentized max-t families. $\overline{G}$ averages all post-initial refined checkpoints before seeds; win rates count positive image-level native-minus-transferred contrasts. Differences and intervals were computed before publication rounding.

The native-over-transferred DINO ordering remains positive in both targets under the refined schedules, while the trajectory-displacement contrast remains negative in SDXL and positive in FLUX. The cross-target dissociation between output ordering and propagation regime therefore persists under finer numerical integration. This analysis tests a fixed refinement of integration resolution on an already-inspected cohort; it does not assess other strengths, conditioning settings, schedules, domains, or architectures. It establishes neither unchanged effect magnitude nor a causal mechanism; no original/refined effect-magnitude comparison is performed.

E.2 Same-Target Encoder-Only Constructor Control

Scope and design. This completed discriminating control reuses the previously inspected original P4 50-image FFHQ cohort for SDXL and FLUX.1-dev, seeds $\{2025, 2026, 2027\}$, $\rho = .9$ only, and the original P4 target execution schedules. It is exploratory, not a new holdout, independent replication, primary confirmatory test, or additional P1–P4 family. Appendix B.3 specifies ENC and constructor fairness; the two-target inference family is defined in Appendix D.4.

The primary comparison is between matched sources, not the two optimized endpoints. For each image and target, seed-averaged endpoint radii define

$$r_{\text{common}}^{\text{PD/ENC}} = \min(r_{\text{PD,end}}, r_{\text{ENC,end}}), \qquad r_{\text{target}}^{\text{PD/ENC}} = 0.9\, r_{\text{common}}^{\text{PD/ENC}}. \tag{17}$$

The exact recovered P4 event-prefix matcher (Appendix C.1) independently selects candidates along clean-to-PD and clean-to-ENC paths. Semantic endpoints are evaluated only after matching; no DINO enters matching. This experiment recomputes its own PD/ENC pair-specific common endpoint radius. Thus $\rho = .9$ does not imply the same absolute initial-state radius as $\rho = .9$ in the original PD-versus-transferred P4 comparison; absolute DINO means from the two experiments must not be compared as if measured at a common radius.

Target	PD DINO	ENC DINO	PD−ENC	Simultaneous 95% CI	Win rate	d_z
FLUX	0.093580	0.089076	+0.004505	[0.001030, 0.007980]	78.0%	0.550
SDXL	0.199645	0.107985	+0.091660	[0.068013, 0.115307]	100.0%	1.645

Table 15: Same-target constructor control on the original P4 50-image cohort, with three seeds averaged within image. Realized PD/ENC initial-state radii are approximately matched. The intervals form one separate exploratory two-target family. Here $\rho = .9$ is relative to the PD/ENC pair-specific common endpoint radius and therefore does not imply the same absolute radius as original P4. Differences were computed before rounding.

Under approximately matched realized initial-state magnitude, the prediction-discrepancy constructor yields larger mean DINOv2 response than the tested encoder-displacement-only constructor in both targets. Maximizing encoder displacement alone does not reproduce the PD response under these conditions. Win rates count positive image-level PD-minus-ENC differences; d_z divides their mean by their image-level standard deviation.

Matching audit. All 100 image–target units passed the frozen 2% matching gate, with zero failures, exclusions, replacements, or rescue searches. Every serialized source remains within $8/255$. Table 16 reports seed-mean pair mismatch, seed-wise mismatch, branch error relative to the requested radius, and descriptive direction similarity. All reported radius errors are far below 2%.

Target	Mean pair mismatch (%)	Max pair mismatch (%)	Max seed-wise mismatch (%)	Max target error (%)	Mean direction cosine	Mean angle (degrees)
SDXL	0.006130	0.021917	0.023461	0.022558	0.513810	59.074
FLUX	0.006006	0.031052	0.031775	0.044790	0.911461	24.132

Table 16: PD/ENC manipulation checks. Pair mismatch uses the seed-mean radii; the next maximum checks individual paired seeds, and target error checks each branch against its requested radius. Direction cosine and angle are descriptive and are not explanatory mediators.

Descriptive construction diagnostics. Mean endpoint radii (PD/ENC) are $207.5694/260.5638$ for SDXL and $121.5595/123.9402$ for FLUX. Matched-source RMS is $3.7645/2.6070$ and $3.7083/3.5064$, respectively, in 0–255 intensity levels; maximum matched-source L_∞ is $6/4$ and $6/6$ levels. The reported ENC best-iterate summaries are median $49\,[46, 50]$ for SDXL and median $8\,[8, 8]$ for FLUX. Historical PD/ENC runtimes are $24.743/15.375$ seconds and $87.677/16.096$ seconds, respectively. These runtimes come from different runs and do not establish speedup. Source RMS is not matched: the scientific manipulation matches realized target initial-state magnitude.

Interpretation and limits. This same-target control narrows the explanation of the P4 directional response by showing that the tested encoder-displacement maximization alone is insufficient under the evaluated conditions. It compares two complete constructor objectives; it does not show that encoder geometry is irrelevant, identify a unique prediction-backbone mechanism or the geometric property responsible for response, or establish superiority over all source constructors. The difference in direction cosine between targets is descriptive, not a causal explanation of their effect-size difference.

FLUX ENC required deterministic backward execution for repeatability, whereas the historical PD backend setting was not recorded (Appendix B.3); equal iterations establish neither equal compute nor equal convergence. Cohort reuse and unchanged target-specific execution also preclude claims of independent replication or model identity alone causing the effect.

E.3 Exploratory Terminal-State Readout (FR-2)

FR-2 is a supplementary exploratory intervention on terminal states from the reused P4 cohort, motivated by the original trajectory and descriptive ratio analyses. It tests readout after approximate executed final-state magnitude matching; it is not independent confirmation of P4 or a neural-mechanism localization. The original FR-2 estimates, calibration, inference family, and motivating Q analysis are retained below.

For final-state matching, 300/300 geometry pairs passed. The available records include 900/900 original terminal tensors, 36/36 original-state decoder checks with exact RGB and PNG bytes, 600/600 cached matched non-clean states, 600/600 manipulated outputs, 600/600 DINO endpoints, 300/300 reused paired clean references, 100/100 image–target aggregates, and 2/2 simultaneous target contrasts. Failures, exclusions, new-output retries, and image-level ties are zero. Clean references were reused; exact reproduction applies to original-state decoder checks, not manipulated outputs.

Controlling final-state magnitude. With preceding trajectories unchanged, fix an image, seed, and target; write $z_f = z_T$, let C denote clean and $b \in \{N, \mathrm{Tr}\}$ index native and transferred, and define in FP32

$$d_f^b = z_f^b - z_f^C, \qquad R_f^b = \mathrm{RMS}(d_f^b), \qquad u_f^b = \frac{d_f^b}{\|d_f^b\|_2}, \qquad R_\star = \sqrt{R_f^N R_f^{\mathrm{Tr}}}. \tag{18}$$

$R_\star$ is shared only within this image, seed, and target. R_f^b is a terminal RMS radius, whereas the initial-state r is an L_2 radius; for n_f terminal elements, $\mathrm{RMS}(v) = \|v\|_2/\sqrt{n_f}$. The ideal continuous edit is $z_f^C + (R_\star/R_f^b)d_f^b$. The evaluated state instead promotes both original operands to FP32 before subtraction and uses

$$\widetilde{z}_f^b = \mathrm{cast}_{\mathrm{BF16}}\big(z_f^C.\mathrm{float}() + \beta_b u_f^b\big), \qquad \beta_0 = \sqrt{n_f}\, R_\star. \tag{19}$$

Geometry-only calibration evaluates β_0, both endpoints of $[0.9\beta_0, 1.1\beta_0]$, and exactly 40 bisection midpoints. It selects the closest executed RMS radius, breaking ties toward smaller β; neither decoder output nor DINO enters calibration. Calibration and tolerances were pre-specified before semantic evaluation of this intervention. The numerical criteria require (A) per-branch relative target-radius error $\le 10^{-4}$, (B) pairwise relative mismatch $\le 2 \times 10^{-4}$, (C) intended/executed direction cosine ≥ 0.999, (D) finite tensors of the expected native BF16 shape with unchanged clean center and positive radii $> 10^{-8}$, and (E) $\beta_b/\|d_f^b\|_2 \in [0.5, 2.0]$.

The native terminal-entry shapes are $[1, 4, 128, 128]$ for SDXL and packed $[1, 4096, 64]$ for FLUX. Exact model revisions are provided in the accompanying artifact. No solver step follows the edit. The unchanged native decoder tail includes the original inverse latent transformation or unpacking, VAE decoding, and postprocessing, and consumes the exact calibrated cached BF16 tensor. With the paired clean reference from the same target, image, and seed,

$$S_b(i,s) = D_{\mathrm{DINO}}\big(D_{\mathrm{native}}(\widetilde{z}_f^b), y_{\mathrm{clean}}(i,s)\big), \qquad \Delta_{\mathrm{readout},i} = \frac{1}{3}\sum_s \left[S_N(i,s) - S_{\mathrm{Tr}}(i,s)\right]. \tag{20}$$

The evaluator is the same `timm vit_small_patch14_dinov2.lvd142m` path specified in Appendix D.2: bicubic antialiased 518×518 input, normalized 384-dimensional CLS/global representation, and one minus cosine similarity. The new endpoints are not divided by radius. Inference uses $N = 50$ images per target; seeds are repeated measurements, not independent samples. The two contrasts form a separate 10,000-resample simultaneous studentized max-t family with seed 20260726 and common resampling indices matched by source-image ID. Paired standardized effects use the image-level paired-difference standard deviation.

Final-state results and numerical fidelity. Mean native DINO dissimilarity remains larger after approximate final-state radius matching in both targets, although FLUX has the smaller contrast and an interval lower bound closer to zero. Table 17 reports all semantic results, including d_z.

Target	Native DINO	Transferred DINO	Δ_{readout}	Simultaneous 95% CI	Win	d_z
SDXL	0.050350	0.037203	+0.013148	[0.007894, 0.018401]	80%	0.831533
FLUX	0.067820	0.064158	+0.003662	[0.000274, 0.007051]	64%	0.359109

Table 17: FR-2 final-state readout results. Means and native-minus-transferred contrasts use 50 reused P4 images per target after seed averaging. The two intervals constitute the separate exploratory FR-2 family. Win rates count positive image-level native-minus-transferred contrasts. Differences were computed before publication rounding.

Target	Max. target error	Max. pair mismatch	Min. direction cosine	Effective scale range
SDXL	0.001786930%	0.001796255%	0.999989808	[0.942243859, 1.06132087]
FLUX	0.003719873%	0.003723773%	0.999993563	[0.956144995, 1.04584751]

Table 18: Final-state matching geometry (FR-2). Errors are percentages; all values satisfy the specified approximate-matching and direction criteria.

At approximately matched final-state magnitude, the native direction produces larger mean DINO dissimilarity in both models. This comparison evaluates the full readout, including decoding, output postprocessing, and the representation metric; it does not identify the VAE as the origin of the effect. Completion and original-state reproduction checks are reported at the start of this subsection.

These ambient edited states have no verified bounded source preimages. The intervention does not localize the neural implementation of the directional effect or establish mediation of the bounded-source effect. Approximate radius matching and high direction cosine also leave open a semantic contribution from residual numerical error.

DINO dissimilarity relative to final-state displacement. To describe DINO dissimilarity relative to final-state displacement, we also summarize the original, unedited bounded-source executions at their natural last pre-decode checkpoint. The calculation uses archived states and the same DINO endpoint, without rerunning generation or selecting a timestep. Let

$$R^b_{\text{final}} = \text{RMS}(z^b_{\text{final}} - z^{\text{clean}}_{\text{final}}), \qquad S^b_{\text{final}} = D_{\text{DINO}}(y^b, y^{\text{clean}}), \qquad Q^b = \frac{S^b_{\text{final}}}{R^b_{\text{final}}}. \tag{21}$$

Q is calculated per seed before averaging within image and is used only as a descriptive ratio. Its table forms a distinct two-target 10,000-resample simultaneous studentized max-t family. It is not recomputed from FR-2 outputs.

Target	Native Q	Transferred Q	ΔQ	Simultaneous 95% CI	Win rate
SDXL	0.114840	0.084877	+0.029963	[0.018153, 0.041773]	84%
FLUX	0.109461	0.102159	+0.007302	[0.001637, 0.012966]	72%

Table 19: Original-output DINO dissimilarity divided by final-state RMS displacement. Win rates are image-level native-greater-than-transferred rates after seed averaging. The two intervals form one simultaneous target family.

Because Q contains the original bounded-source semantic endpoint in its numerator, it provides descriptive context rather than an independent causal explanation or mediation estimate; it is neither a derivative nor a Lipschitz gain.

E.4 Population Endpoint Robustness

Metric	Native vs Random	Native vs Gaussian	95% CI (R; G)	Win rate (R; G)
DINO cosine dissimilarity	0.213410	0.230344	[0.205377, 0.221568]; [0.222159, 0.238713]	100.0%; 100.0%
ArcFace identity drop	0.136518	0.152258	[0.131401, 0.141782]; [0.147049, 0.157578]	99.8%; 99.8%
CLIP-image drop	0.101462	0.109340	[0.097180, 0.105766]; [0.104959, 0.113742]	99.8%; 100.0%
LPIPS excess	0.114876	0.125897	[0.111797, 0.118018]; [0.122600, 0.129248]	100.0%; 100.0%

Table 20: Population-level endpoint robustness on 500 FFHQ images. Values are paired model-native-minus-control contrasts; DINO is primary and the remaining endpoints are secondary. Intervals are image-cluster bootstrap intervals, and R/G denote random-sign/Gaussian controls.

E.5 Directional Response and Broader Recurrence

The representation used to measure output difference affects the bounded-source comparison. Main Table 4 compares all four endpoints at $\rho = .9$ for both targets. DINO cosine dissimilarity is the sole primary endpoint, with its two contrasts forming the simultaneous primary family; CLIP-image cosine dissimilarity, ArcFace identity dissimilarity, and LPIPS are post-hoc secondary analyses.

ID	Target	Endpoint	Comparison	Estimate [95% CI]	Family
A. Native sources versus controls constructed for the target model					
01	SD1.5	State RMS	Native − target-model random sign	0.035590 [0.029852, 0.041329]	F1
02	SD1.5	State RMS	Native − target-model Gaussian	0.101560 [0.095717, 0.107402]	F1
06	FLUX	State RMS	Native − target-model random sign	0.025314 [0.023800, 0.026828]	F4
07	FLUX	State RMS	Native − target-model Gaussian	0.045897 [0.044206, 0.047589]	F4
B. Transferred sources versus controls constructed with the SDXL sources					
03	SD1.5	State RMS	Transferred − SDXL-source random sign	0.025316 [0.022690, 0.027942]	F2
04	SD1.5	State RMS	Transferred − SDXL-source Gaussian	0.077624 [0.075013, 0.080236]	F2
08	FLUX	State RMS	Transferred − SDXL-source random sign	-0.008576 [-0.009164, -0.007988]	F5
09	FLUX	State RMS	Transferred − SDXL-source Gaussian	0.014958 [0.014723, 0.015194]	F5
C. Direct native–transferred contrasts without radius matching					
05	SD1.5	State RMS	Native − transferred	0.122461 [0.117025, 0.127897]	F3
10	FLUX	State RMS	Native − transferred	0.033913 [0.031532, 0.036293]	F6
11	SD1.5	DINO	Native − transferred (recovered)	0.140316 [0.114083, 0.166549]	F3
12	FLUX	DINO	Native − transferred	0.041601 [0.031533, 0.051668]	F6
D. FLUX direction comparison at the native reference radius					
13	FLUX	DINO	Native − transferred (native radius)	0.020321 [0.012582, 0.028061]	F7

Table 21: Cross-model comparisons with separately constructed controls. Native-versus-control rows use target-model controls; transferred-versus-control rows use SDXL-source controls. The image identities and target executions are paired, but the control perturbations differ, so control-relative contrasts cannot be subtracted to recover a direct native–transferred contrast. Each interval retains its original simultaneous family (Appendix D.4); coverage is not joint across all rows. DINO is relative to the paired clean generated output. The final FLUX row uses the prescribed native reference radius. SD1.5 retains the selected recovered results and checker-independent direct decoding.

Model	Endpoint	Native vs Random	Native vs Gaussian	Decision
SDXL	Geometry	0.292960 [0.287141, 0.298779] (100.0%)	0.347560 [0.342048, 0.353072] (100.0%)	Supported
SDXL	Semantic	0.198076 [0.180078, 0.216073] (100.0%)	0.205069 [0.186642, 0.223497] (100.0%)	Supported
FLUX	Geometry	0.007217 [0.006015, 0.008419] (91.5%)	0.023895 [0.022511, 0.025279] (100.0%)	Supported
FLUX	Semantic	0.032498 [0.023070, 0.041925] (79.0%)	0.056777 [0.045859, 0.067694] (85.5%)	Supported

Table 22: COCO non-face initial-state displacement and output dissimilarity ($N = 200$ images per target; three seeds averaged within image). Geometry and semantic families use separate simultaneous intervals. Coverage is 2,400/2,400 unique branch executions per target and 4,800/4,800 pooled across the two targets, with zero recorded failures. This cohort does not repeat matched-radius interventions.

E.6 Bounded-Source Matching Diagnostics

The 3,400-image FFHQ frame was partitioned by prior experimental use before P4. Images assigned to geometry development or earlier semantic analyses were ineligible, leaving 550 previously unused images. The 50-image holdout was fixed from this pool before P4 semantic evaluation. For each image and target, both branches use the common target radius in Eq. 6; candidate radii are averaged over $\mathcal{S} = \{2025, 2026, 2027\}$ as in Eq. 12. The same seeds are used for paired semantic generation and averaged within image. The run produced all 2,100 expected outputs (1,050 per target). DINO cosine dissimilarity was not used for matching, no source was rematched after semantic generation, and the primary radius, contrast family, and statistical family were fixed before evaluation.

Across the full radius sweep, every seed-averaged matched pair passes the 2% tolerance. The maximum relative mismatch is 0.077274% at $\rho = .5$, 0.082043% at $\rho = .7$, and 0.050913% at $\rho = .9$. At the primary radius, residual signed radius imbalance is not positively associated with the native-minus-transferred DINO response (FLUX Pearson $r = -0.0693$, Spearman $\rho = -0.0484$; SDXL Pearson $r = -0.0909$, Spearman $\rho = -0.2119$). These associations are descriptive and are not adjusted causal estimates.

Main Figure 4 reports the semantic contrast at all three reachable radii, and Section 4.4 reports the corresponding matching checks. Figure A2 supplies the remaining manipulation checks. Every image–target–radius cell is source-reachable and direction-defined. Source L_∞ remains within the 8-integer-level budget. The direction-cosine and path/pixel preservation panels describe the realized constructions; they are validation evidence for the matched intervention, not semantic endpoints.

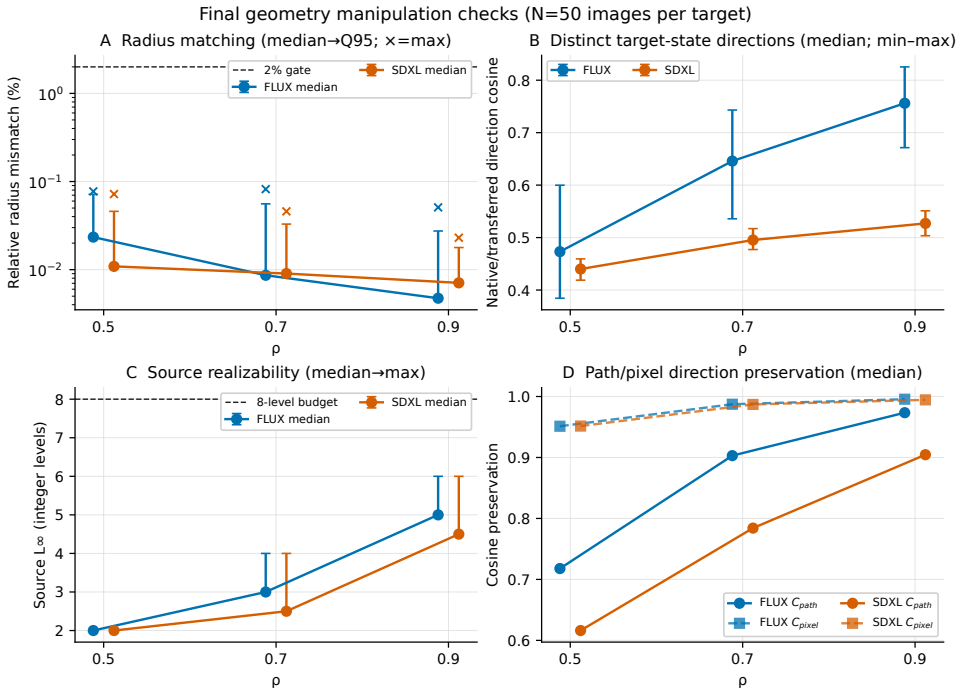

Figure A2: Geometry manipulation checks for the matched-radius bounded-source test ($N = 50$ images per target). Panel A reports target-state radius mismatch against the 2% tolerance; panel B verifies distinct native and transferred directions; panel C reports the source L_∞ budget; panel D summarizes direction preservation.

F Qualitative Results

Figures A3–A7 show source-condition, cross-model, radius, cross-domain, and direction comparisons. Statistical conclusions use the complete cohorts in Section 4.

How to read the grids. Within each row, columns share the same source image and should be compared horizontally. Rows marked Q25, Q50, and Q75 are different images selected nearest to image-level effect quantiles after paired seeds were averaged within image; they are not successive states of one image. Figure A6 uses COCO examples selected by ascending image ID with distinct supercategories and no candidate-image inspection. All selections are deterministic and post-hoc.

Intervention sequence. Figure A3 compares the model-native variation with random-sign and Gaussian controls; Figure A4 extends the display across models. Figure A5 holds direction fixed while changing radius; Figure A6 extends the display across domains. Figure A7 compares SDXL directions at the stated reference radius and FLUX directions at the prescribed native reference radius; achieved BF16 equality cannot be rechecked for this earlier FLUX run (Appendix G).



Figure A3: FFHQ outputs under the clean, model-native, random-sign-control, and Gaussian-control conditions at Q25, Q50, and Q75 of the image-level model-native-minus-random-sign DINO contrast. "Model-native" denotes source selection using J_M. Compare conditions within each row.

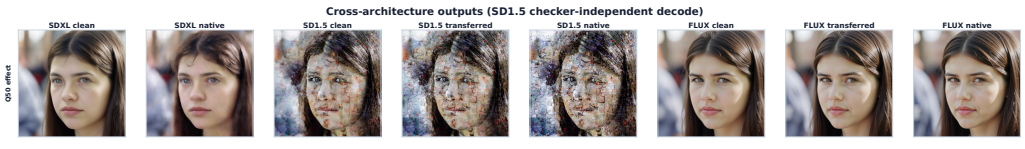

Figure A4: Cross-model outputs for one Q50 FLUX DINO contrast. A common source is evaluated with SDXL, SD1.5, and FLUX, with native and transferred conditions shown within each target. SD1.5 uses checker-independent direct decoding.



Figure A5: Fixed SDXL target-native direction at three radii for Q25, Q50, and Q75 of the high-minus-low DINO contrast. Here r_g, r_{rs}, and r_{da} are the Gaussian-, random-sign-, and target-native-induced initial-state L_2 reference radii. Each row fixes one image and seed 2025; only the radius changes.



Figure A6: COCO non-face examples selected by ascending image ID with distinct supercategories. Each row shows the common source and SDXL/FLUX outputs under clean, model-native, and random-sign-control conditions. Cohort inference uses all 200 images.



Figure A7: SDXL controls and FLUX native-versus-transferred directions at Q25, Q50, and Q75 of the DINO contrast. The SDXL rows compare directions at the stated reference radius. The FLUX rows use the prescribed native reference radius; the available records do not permit rechecking achieved BF16 radius equality for this earlier experiment (Appendix G). The SDXL isotropic condition is an ambient state-space control.

Cross-model comparison. Figure A4 uses a common source image across SDXL, SD1.5, and FLUX. Within each target, compare the native and transferred outputs with that target's clean output. This keeps the visual reference consistent with the paired clean-relative endpoints used in the quantitative analysis. Native and transferred denote the source-construction procedure relative to the target; these labels do not establish a common initial-state displacement magnitude across targets. Selection at the Q50 FLUX DINO contrast identifies one illustrative image, without making it a median example for every model. Table 21 reports the cohort-level contrasts, with their original inference families detailed in Appendix D.4.

Non-face domain coverage. Figure A6 extends the display to COCO examples selected by ascending image ID with distinct supercategories. Each row fixes the source and shows clean, model-native, and random-sign-control outputs for both targets. Compare conditions horizontally within SDXL or FLUX, using the corresponding clean output as the reference. The displayed images are selected by identity, independently of the size of their output response. Table 5 reports inference over all 200 images, with three paired seeds averaged within image. This cohort measures unmatched source-to-state and output contrasts; it does not repeat bounded-source radius matching or trajectory comparisons. Figures A8 and A9 next return to the bounded-source matched-radius comparison.

Figures A8 and A9 show deterministic Q25, median, and Q75 examples selected only after the primary statistics were fixed. Each row holds the source fixed and juxtaposes the clean target output, target-native source/output, and transferred source/output.

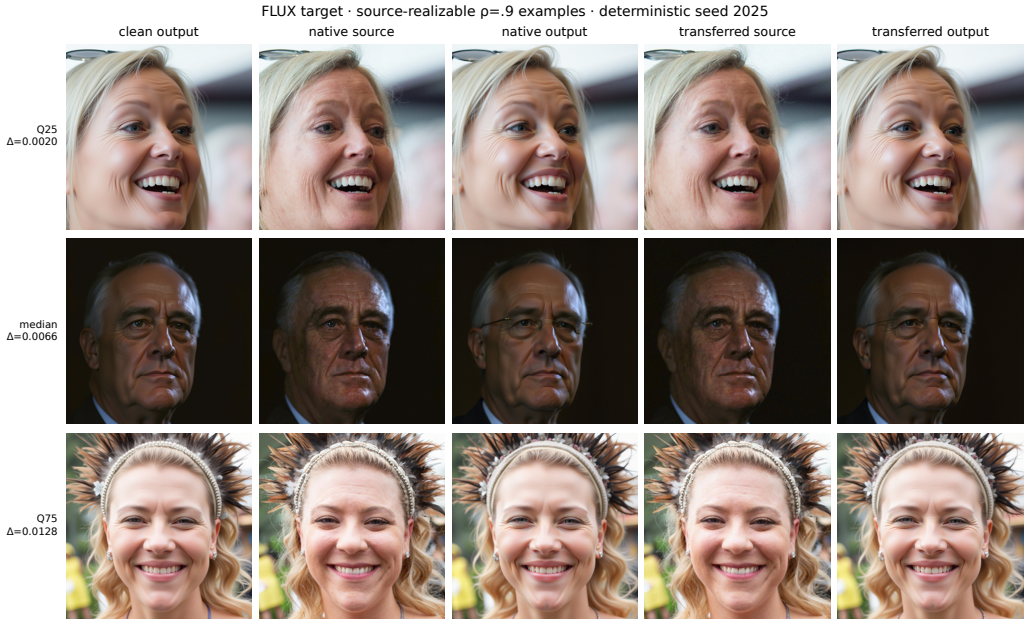

Figure A8: FLUX target examples at the primary source-realizable radius $\rho = .9$. Rows are deterministic Q25, median, and Q75 target-native-minus-SDXL→FLUX image-level effects after seed averaging. Compare outputs within a row; examples do not replace cohort-level inference.

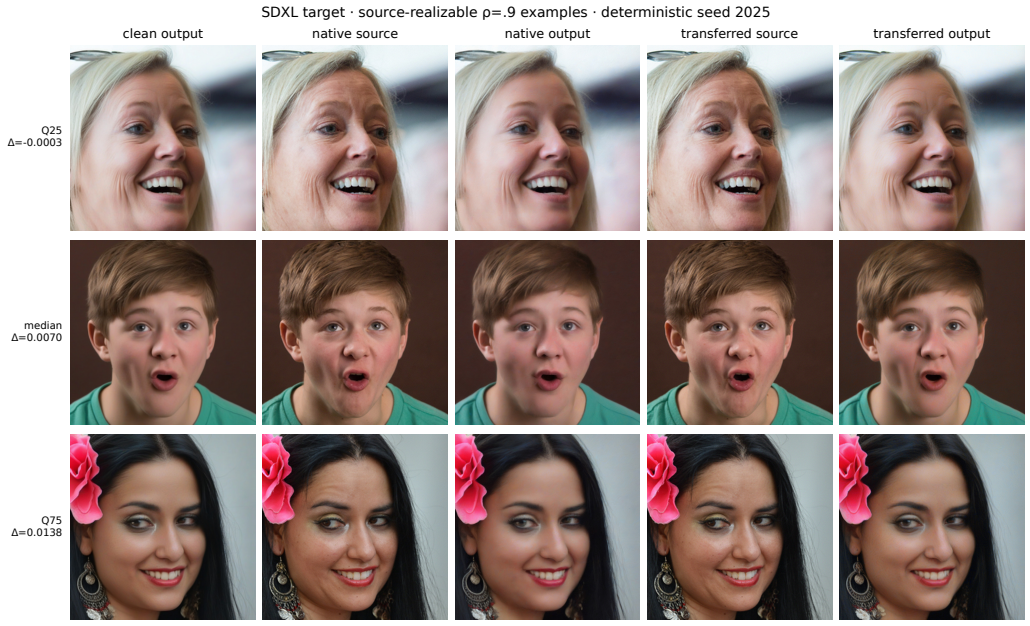

Figure A9: SDXL target examples at the primary source-realizable radius $\rho = .9$. Rows are deterministic Q25, median, and Q75 target-native-minus-FLUX→SDXL image-level effects after seed averaging. Compare outputs within a row; examples do not replace cohort-level inference.

G Reproducibility

Executable core-P4 implementation. The accompanying anonymous artifact provides the executable core-P4 implementation, including the SDXL and FLUX source constructors, bounded-source matcher, P4 image-level and source manifests, target-execution configurations, DINOv2 evaluation, trajectory analysis, and statistical inference code. Tables 6, 7, and 11 specify the model and execution settings. The separate PD/ENC experiment is outside the core-P4 package scope.

Statistical reproduction, fixtures, and schedules. The artifact provides image-level records and lightweight commands reproducing the aggregate P4, original-trajectory, and schedule-refinement statistics with their distinct frozen inference families. These statistical reproductions use recorded measurements. Deterministic fixtures check event-prefix enumeration, candidate selection, state capture, evaluation, and inference. The exact original and refined executed P4 timestep/sigma arrays are included. Hardware-independent bitwise replay is not established.

External inputs and remaining provenance limitations. The exact 50 historical clean P4 inputs remain available in the original experiment storage and match the recorded file-level and decoded-image checks. They are not bundled in the anonymous artifact. The historical public-data acquisition/export procedure that produced these 256×256 inputs from aligned FFHQ was not preserved.

Model weights and all clean, perturbed, and generated images are external to the artifact. End-to-end construction and downstream matching/generation require the external model weights and verified inputs; cached-source replay additionally requires the selected sources recorded in the manifests. Archived scalar statistics can be inspected independently of these images. GPU constructor, exhaustive-matching, and generation smoke replays were not rerun when preparing the artifact. Some historical producing code revisions remain unavailable, older auxiliary SD1.5/Stage2 provenance is partial, and the original BF16 executed-state/radius records for the older FLUX direct-state comparison remain unavailable. Missing SD1.5 source-probe fields remain marked unavailable. These limitations preclude full reproduction of every historical experiment.

Scope of the cross-model reconciliation. An audit of the archived scalar observations and the corresponding numerical and export code reproduced all 13 point estimates and 26 confidence limits in Table 21. This check resolved the control-label and statistical-family ambiguity; it did not independently remeasure the original state norms or DINO features. Original initialized-state tensors and DINO feature arrays are absent from the audited subtree, and the exact executed revision/environment remains unavailable for some Stage2D runs.

For the older FLUX direct-state comparison, the available record verifies the prescribed native reference radius, but the original BF16 executed-state/radius receipts are unavailable for checking achieved equality. This limitation is separate from the later FR-2 matching diagnostics. Some recovered SD1.5 DINO values differ by up to approximately 5.96×10^{-7} despite identical output hashes; the cause is not established, and the selected recovered estimates and intervals are retained. The verified 512×512 SD1.5 construction applies specifically to the audited cross-model experiment and does not fill missing preprocessing, prompt, VAE, or seed fields in unrelated runs. P4, trajectory, Q, and FR-2 analyses were unaffected in the identified archived-scalar aggregation/inference dependency; this finding does not certify all model executions or feature measurements. Historical measurement provenance therefore remains partial.

H Extended Related Work and Literature Positioning

This appendix expands the three literature axes in Section 2. The organizing question is how a bounded source change induces an initial-state displacement, how that displacement propagates, and how the resulting output responds. Prior protection, geometry, and interpretability studies motivate this question without supplying the particular source-realizable matched-radius comparison evaluated here.

H.1 Source Perturbation and Protection Methods

Source conditioning and task scope. SDEdit adds noise to a source image and denoises it with a generative prior (Meng et al., 2022). Other conditioning routes include feature injection (Tumanyan et al., 2023), subject-specific fine-tuning (Ruiz et al., 2023), spatial control (Zhang et al., 2023), and image-prompt adapters (Ye et al., 2023). These routes expose different interfaces to source information. Our initial-state interpretation applies to the evaluated executions, where no additional visual source input enters downstream generation; it is not an assertion about every image-conditioned architecture.

PhotoGuard. Salman et al. (2023) optimize bounded image perturbations to disrupt unauthorized diffusion editing. Their encoder attack moves the encoded source toward a target representation; their diffusion attack targets the generated image through the diffusion computation. Evaluation concerns the resulting edits and their departure from intended content. This explicit distinction between encoder and diffusion objectives precedes our study. We use the state consumed by the solver to compare finite responses, rather than claiming a new encoder/diffusion decomposition or comparing protection efficacy.

AdvDM and Mist. AdvDM optimizes source pixels by maximizing a diffusion training-loss surrogate with Monte Carlo sampling (Liang et al., 2023). Its purpose is to hinder extraction and imitation of image features, evaluated through downstream conditional generation. Mist combines a sampled noise-prediction loss with a target-directed encoder loss and examines transfer across image-to-image generation, textual inversion, and DreamBooth (Liang & Wu, 2023). The latter paper calls these semantic and textural terms and studies target-image choice and resistance to purification. Neither objective is identical to our clean-relative prediction-discrepancy probe: our probe compares paired model predictions and excludes both initial-state radius and the final output metric from optimization. Sharing a pixel constraint does not equate the optimized objective, the realized state direction, or the scientific estimand.

Editing protection and personalization robustness. DiffusionGuard optimizes adversarial image noise against early denoising behavior and augments masks to protect against previously unseen editing masks (Choi et al., 2025). Its evaluation addresses protection effectiveness, transfer, and noise-removal robustness. Anti-DreamBooth instead perturbs images supplied for personalization (Van Le et al., 2023); attention-targeting perturbations study customized diffusion models (Xu et al., 2024). Broader protective-perturbation analyses examine the resulting personalization failures (Zhao et al., 2024). These task differences matter: compromising adaptation to a subject and changing a fixed generator's response are different experiments.

CAT and the role of latent distortion. CAT investigates distorted latent representations in protected customization data, then uses contrastive adversarial training of lightweight adapters as an adaptive attack on protection (Peng et al., 2025). Thus it contributes both a latent-space analysis and a robustness test; it is not another protective-noise constructor. Its evidence makes it especially important to distinguish observing latent distortion from measuring downstream response at approximately matched initial-state magnitude.

Relation to our controlled comparison. The present study is not a protection leaderboard. Native and transferred labels identify complete source-construction procedures, not a ranking of published attacks. P4 constructs both compared states through bounded images, selects their realized radii before output evaluation, and measures paired clean-relative response within the same target. Approximately matched radius does not guarantee that every other geometric property is matched. The contribution

is evidence about source-reachable directions and their subsequent propagation, without asserting that our constructor outperforms any protection or attack method discussed above.

H.2 Geometry and Propagation Analyses

Semantic directions and anisotropic response. Work on diffusion semantic latent spaces (Kwon et al., 2023) and interpretable latent directions (Li et al., 2024) studies how state variation relates to image attributes. NoiseCLR learns editing directions from unlabeled images using a contrastive objective, without requiring attribute-specific text prompts (Dalva & Yanardag, 2024). These studies motivate direction-sensitive response and semantic control. Our directions are induced by bounded source constructions and are not assigned semantic labels; a DINOv2 difference need not correspond to one interpretable attribute.

Metrics, guidance, and differential sensitivity. Angle Domain Guidance analyzes norm amplification under classifier-free guidance and controls magnitude while changing angular alignment (Jin et al., 2025). Hessian-geometry work reconstructs a Fisher information metric through a learned log-partition function and studies changes in generative latent-space structure (Lobashev et al., 2025). Probability-density geodesics use a spatially varying metric that favors paths through high-density regions of diffusion latent space (Yu et al., 2025). These are distinct geometric constructions; none is identified with the Euclidean radii used in our experiments. Geometry-Aware Distillation aligns teacher and student Jacobian–vector products with respect to initial noise to preserve differential response (Huang et al., 2026). This provides a direct connection to initial-noise sensitivity, while changing the trained generator rather than measuring bounded-source effects in a fixed one.

Ambient directions versus source-reachable directions. An ambient state edit can vary radius or direction without demonstrating any bounded image that produces the edited state. Source reachability instead restricts the state to $F_M(x + \delta, \eta, c)$ for a feasible image perturbation (Section 3). General latent anisotropy therefore does not determine the responses of the directions accessible under a fixed source budget. P2/P3 characterize controlled ambient comparisons; P4 adds bounded-source preimages and approximate equality of realized initial-state radii. The discrete matching paths can change direction relative to their optimized endpoints, and matching radius does not constrain later trajectory separation. These distinctions motivate combining geometry controls with propagation measurements.

Iterative error accumulation versus clean-relative separation. Ning et al. (2023) study exposure bias from the difference between training inputs and inputs encountered during iterative sampling, and regularize training by perturbing inputs. AccuQuant explicitly accounts for accumulated quantization error by matching full-precision and quantized outputs over multiple denoising steps (Lee et al., 2025). These analyses motivate attention to propagation, but their reference computations and perturbation sources differ from ours. Our G_t measures separation between a perturbed-source branch and its paired clean-source branch in the same target execution, normalized by their initial separation. It is neither an error against ground truth nor a quantization discrepancy, and it need not grow monotonically.

The displacement-growth account considered here associates a larger mean output response with a larger initial displacement or greater subsequent trajectory separation. It is an explanatory comparator for this analysis, not a universal law supplied by the cited error-accumulation studies. Under approximate initial-radius matching, opposite SDXL/FLUX mean normalized trajectory-displacement orderings accompany the same native-over-transferred mean DINO ordering. Greater mean normalized trajectory displacement is thus not a shared explanation for the native output advantage. Fourfold schedule refinement preserves this pattern as a numerical-resolution robustness check, not proof of convergence; architecture remains confounded with model-specific conditioning and execution. FR-2 separately examines readout after exploratory final-state magnitude matching; its edited states have no verified bounded-source preimages. Together these experiments address finite response under stated controls, without establishing a universal propagation law, a neural-component attribution, or formal mediation.

H.3 Editing Benchmarks and Evaluation Context

PIE-Bench, introduced with PnP Inversion, contains 700 images with structured editing annotations and evaluates editability together with structure and background preservation (Ju et al., 2024). EditBench evaluates text-guided inpainting on natural and generated images, with prompts testing objects, attributes, and scenes and human assessment of text–image alignment (Wang et al., 2023). Both illustrate the value of evaluating what an edit changes and what it should preserve, rather than relying only on an aggregate image difference.

Our experiments answer a different question and do not evaluate either benchmark. DINOv2 dissimilarity measures deviation from a paired clean generated output; a larger value is not intrinsically a better edit or stronger protection. The FFHQ P4 holdout tests a pre-specified source-realizable response contrast under approximate initial-state magnitude matching. Its secondary endpoints reveal representation dependence, including the SDXL CLIP-image interval that contains zero. The separate COCO cohort tests recurrence of unmatched source-to-state and output contrasts beyond portraits. It repeats neither matched-radius intervention, trajectory comparison, nor final-state matching. These cohorts support the stated sensitivity comparisons, not coverage of structured editing instructions, inpainting masks, preservation criteria, or benchmark-level editing quality.