THE ACCENT OF FEATURE DRIFT: WHERE RECONSTRUCTION ERROR IS NOT NOISE

Anonymous authors
Paper under double-blind review

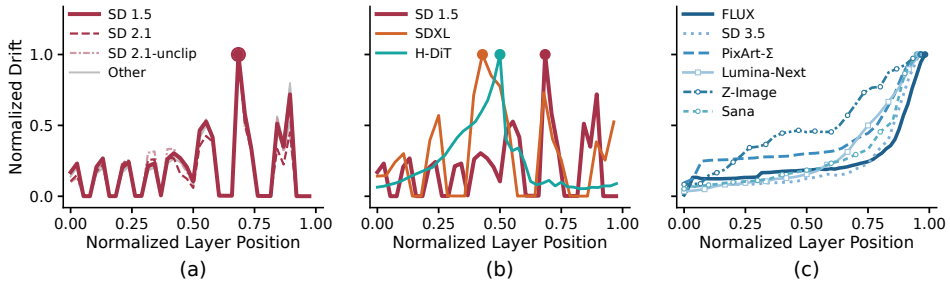

Figure 1: Normalized drift profiles. (a) SD 1.5 and its variants (gray: SD 1.4, LCM LoRA, RandText, and Realistic Vision), SD 2.1, and SD 2.1-unCLIP-small. (b) Interior-localized profiles. (c) Terminal-ramp profiles.

ABSTRACT

Reconstruction trajectories provide a readout of how diffusion architectures organize internal features. We study **feature drift**, the per-layer discrepancy between a reconstruction trajectory and an independently sampled forward reference of the same image, and define the **architecture accent** as its min–max normalized depth profile. Under a specified readout protocol, the accent is reproducible across tested variants within trained lineages and varies across architectures. Ten model instances spanning five backbone families exhibit two dominant forms, interior-localized and terminal-ramp, with a terminal-spike boundary in two further lineages. The dominant classes are recovered in leave-one-out evaluation within this registry. These forms have different measurement interpretations: terminal ramps largely track the reference features' norm schedule and lose their ramp label under relative L2, whereas the tested interior-localized profiles retain peaks before the final block under norm-free metrics. Controlled interventions show that drift magnitude does not uniquely determine reconstruction quality, and that site sensitivity in training-free correction differs across architectures. Random-initialization controls and retraining under two objectives further indicate that the observed organization develops with training. Together, these results show how reconstruction trajectories can be used to read out architectural organization and identify candidate sites for training-free correction.

1 INTRODUCTION

Diffusion inversion methods improve reconstruction through more accurate trajectories (Song et al., 2021a; Wallace et al., 2023), feature alignment (Mokady et al., 2023), and tighter solvers (Wang et al., 2025; Rout et al., 2025; Jiao et al., 2026; Wang et al., 2026). Other methods bypass inversion altogether (Kulikov et al., 2025). In all of these pipelines, the discrepancy between the reconstruction trajectory and a reference is treated as reconstruction error to be reduced. This paper asks a different question: we measure, at every layer, the discrepancy between reconstruction-trajectory features and an independently sampled forward reference of the same image, and we ask whether its organization along network depth provides a reproducible signature of architectural organization. Prior work on temporal dynamics across DiT feature dimensions (Zheng et al., 2026) and representation

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.

structure (Zhang et al., 2026) motivates asking whether this discrepancy has a reproducible organization across network depth. On 104 images, SD 1.5 lineage variants (a predecessor checkpoint, a distilled model, a fine-tuned model, and random text embeddings) share the same peak layer, with structural distances at most 0.0232, and the successor checkpoint SD 2.1 preserves that peak as well (Figure 1(a), §2.2). The measured discrepancy is defined against an independent forward reference and remains distinct from the solver's endpoint reconstruction error. We call its normalized profile along depth the **architecture accent**.

We ask (i) whether the accent is reproducible across variants of a trained model under a specified readout protocol, (ii) what it measures about architectural organization, and (iii) how useful it is for selecting correction sites. Our contributions are threefold.

1. **A reproducible profile that depends on the architecture, §2.** Ten model instances from five backbone families exhibit two dominant morphologies under the canonical readout, and lineage variants retain their peak positions. Leave-one-out prediction recovers all ten labels, including with features that exclude peak position and monotonicity, whereas the family-label baseline recovers at most nine. Norm-free analyses distinguish the two forms: the ramp label depends on the feature-norm schedule, whereas the three evaluated interior-localized instances retain peaks before the final block (Appendix C.9).
2. **Architecture-specific pathway sensitivity and emergence with training, §3.** Peak-skip interventions have opposite effects in SD 1.5 and SDXL, changing PSNR by $+2.20$ dB and -11.59 dB, respectively. In SD 1.5, interventions that move drift magnitude in opposite directions can both improve reconstruction, so magnitude alone does not explain the gain. Random-initialization controls and retraining under two objectives associate the organization with training (Appendix D).
3. **Architecture-dependent sensitivity to correction sites, §3.4.** On the canonical 104-image SD 1.5 evaluation we compare five-layer site-selection strategies, by highest drift, reference norm, even spacing, and random draws, at matched layer count, injection strength, and reconstruction steps, against a no-selection latent-correction baseline. Site sensitivity is a property of the architecture: two five-layer sites in HunyuanDiT differ by 3.15 dB, while SD 1.5 and FLUX are site-insensitive under the same protocol (Tables 3 and F.1; Appendix F).

2 The Architecture Accent of Feature Drift

Let M be a diffusion backbone with L ordered, hooked blocks. We measure the output of each residual and attention block in UNet backbones and each transformer block in DiT backbones, with $L = 38$ for SD 1.5, 24 for SD 3.5 Medium, and 57 for FLUX. Profiles are compared at normalized layer position l/L. An image x is encoded, inverted to z_T, and reconstructed to $\hat{x}$ along the reverse trajectory. We use DDIM for diffusion models and Euler discretization of the velocity field for flow-matching models. Unless otherwise stated, the main experiments follow the canonical P-multi protocol on a fixed set of 104 images. Model lists, hub versions, and protocol derivations appear in Appendices A.1–A.2.

2.1 Feature Drift and Architecture Accent

Feature drift is the measured feature discrepancy; the architecture accent is its normalized depth profile; and morphology is the profile class assigned by the rule in §2.3. Readout subtypes describe what happens beyond the final block.

Definition 1 (Feature drift)**.** The feature drift of layer l on image x is

$$d_l(x) = \mathbb{E}_{t \in K}\big[\|\mathcal{P}(f_l^{\mathrm{ref}}(x,t)) - \mathcal{P}(f_l^{\mathrm{recon}}(x,t))\|_2\big],$$

where f_l^{ref} and f_l^{recon} are the features of an independent forward reference and of the reconstruction trajectory, read at the same timestep. The fixed aggregation $\mathcal{P}$ is applied identically to both features: mean pooling over the token axis for sequence outputs $[B, N, C]$, including attention and transformer blocks, and the identity for convolutional outputs $[B, C, H, W]$. We use the same rule across architectures and layers. The set K contains nine indices on the N-step scheduler grid, chosen by the fixed relative rule in Table A.2.

The reference latent is $z_{\text{ref}}(t) = \alpha_t z_0 + \sigma_t \varepsilon$, with reference noise drawn independently of inversion under the seed-42 protocol. The coefficients α_t and σ_t specify the signal and noise scales. For variance-preserving DDPM backbones, $\alpha_t = \sqrt{\bar{\alpha}_t}$ and $\sigma_t = \sqrt{1 - \bar{\alpha}_t}$, where $\bar{\alpha}_t$ is the cumulative signal-retention coefficient determined by the training noise schedule. For rectified-flow backbones, σ_t is the sampler's interpolation noise level and $\alpha_t = 1 - \sigma_t$. Appendix A.1 gives the corresponding ODEs (Song et al., 2021b; Lipman et al., 2023) and backbone assignments. Reconstruction features are read at matching timesteps.

Two measurement conventions matter. First, $\mathcal{P}$ does not divide by the number of tensor elements, so absolute magnitudes are protocol-relative and our claims concern the organization along depth; dividing each layer's distance by the square root of its element count leaves the peak layer and morphology label unchanged in the nine evaluated instances (Appendix C.9(b)). Second, the reference uses the same z_0 and an independent noise draw, so drift measures discrepancy under a specified reference coupling rather than the solver's endpoint residual (Appendix A.1). The empirical organization survives removal of the endpoint inversion residual in the exact-inversion controls of Appendix C.10.

Definition 2 (Drift profile)**.** The drift profile of M is $D(M) = [\, d_1, \ldots, d_L \,]$, where $d_l = \mathbb{E}_{x \in \mathcal{D}}[d_l(x)]$ and $\mathcal{D}$ is the evaluation set.

Definition 3 (Architecture accent)**.** The architecture accent is the min–max normalized drift profile $\Phi(M) = \text{norm}(D(M)) \in [0, 1]^L$, an ordered vector over the canonicalized layer order, with normalization applied across layers within each model.

The min–max normalization removes the overall scale and offset while retaining the organization along depth. We evaluate three properties: stability under lineage variants, architecture dependence, and relation to architectural topology.

To compare accents we use a structural distance $D_s(A, B) = (D_{\text{pp}}^2 + D_{\text{mag}}^2)^{1/2}$, where D_{pp} is the difference between normalized peak positions and D_{mag} measures distributional shape through concentration and Gini statistics, with formulas in Appendix B.1. All morphology labels in the main analysis are defined under the canonical absolute L2 metric of Definition 1, at block level. The distance summarizes peak position, concentration, and Gini rather than the full ordering; full-profile comparisons additionally report pairwise Spearman correlation.

Two further readouts probe the metric itself. A diagnostic relative metric $\text{rel}_l(x, t) = \|f_l^{\text{ref}} - f_l^{\text{recon}}\|_2 / \|f_l^{\text{ref}}\|_2$ tests whether terminal drift is explained by readout truncation or feature-norm growth, and re-evaluates the morphology norm-free (Appendix C.9). For terminal-peaked models the readout extends beyond the final block to the post-block stages (norm out, projection out, velocity, latent, and pixel space), separating terminal block-localized peaks from post-block relative accumulation (Appendix C.6). Both refine the interpretation of the terminal-ramp morphology without changing the classification rule (§3.1).

2.2 Stability Across Model Variants and Protocols

We first test whether the accent survives perturbations within a trained lineage. Four SD 1.5 lineage variants (SD 1.4, LCM LoRA, RandText, Realistic Vision) were fully re-measured under the canonical readout. The peak positions are unchanged with $D_{\text{pp}} = 0$, and $D_s = D_{\text{mag}} \in [0.0045, 0.0232]$. The successor checkpoint of the same UNet family, SD 2.1, has the same peak layer `up_blocks.2.resnets.0` at pp $= 0.6842$ with $D_s = 0.0099$, extending the identity of the peak position to an independent checkpoint. The distilled SDXL-Turbo likewise has $D_{\text{pp}} = 0$ and $D_{\text{mag}} = 0.0022$, below the noise floor $p_{95} = 0.0042$. Under random half-splits the profiles correlate at Spearman ≥ 0.9987 with the same peak layer in 20/20 splits (Appendix C.5).

Within-lineage comparisons give structural distances of $D_s \in [0.0022, 0.0963]$, with the largest value arising from the PixArt-α–PixArt-Σ release pair, while between the two dominant morphology classes the minimum D_s is 0.2679, about $2.8\times$ the largest within-lineage distance. Morphology is assigned by the rule of §2.3 from peak position and monotonic trend, so the shape coordinates may vary within a class; the specific distances are given in Appendix C.5. Separately, random initialization and training yield a peak-position difference of $D_{\text{pp}} = 0.658$ (Appendix D.1).

A 23-cell confound matrix re-measures each perturbed side under the canonical protocol against a baseline of SD 1.5 or PixArt-Σ (Appendix C.7). Every lineage and protocol cell locks at its

baseline peak and is TOST-equivalent under the pre-registered structural-distance margin of 0.1245 ($p < 0.001$; Appendix B.2), while the cross-morphology cells have mutually distinct peaks.

Peak position and terminal topology are preserved across the tested lineage, sampler, data-split, precision, noise-level, reference-seed, and inversion-framework perturbations (Appendix C.2). Reseeding the reference ε leaves the peak fixed on four models at max $D_s = 0.0017$, and exact-inversion frameworks move no peak, at a cost of 10–19% amplitude (Appendix C.10). Driving the end-to-end residual to exactly zero leaves the per-image argmax in $104/104$ agreement on both backbones, and 81–90% of the drift magnitude survives with all of its peak structure (Appendix C.10). The control for model scale preserves a terminal peak but changes the discrete morphology label (§2.3). The readout protocol is therefore part of the measurement definition; single-step readouts, for example, move the peak (Appendix A.2).

Verdicts are unchanged across a wide margin interval around 0.1245, and the positive control behaves as required: random injection at $\eta = 1$ moves the SD 1.5 peak to 0.895, violating the margin (Appendix B.2).

2.3 Morphological Organization

Morphology labels follow a rule frozen before the extended-registry measurements (Appendix A). The fixed threshold is $\tau = 0.90$.

Interior-localized. The profile attains its maximum before the final block.

Terminal-ramp. The profile peaks at the final block, and the Spearman correlation between layer index and normalized drift satisfies $\rho \geq \tau$. Profiles satisfying neither criterion are labelled *other* and excluded from dominant-morphology centroid estimation.

In the measured registry, the labels of the ten dominant instances and the three terminal-spike boundary instances are unchanged throughout $\tau \in (0.6012, 0.9872)$, between the largest boundary-instance correlation and the smallest dominant-ramp correlation (Appendix C.1).

We report the continuous monotonicity value ρ alongside each label; τ defines a decision rule without assuming a natural gap in the observed spectrum. Morphologies supported by only a few independent lineages are reported as boundary morphologies. Both dominant morphologies are defined under the canonical absolute L2 readout, and leave-one-out prediction follows the same frozen design (§2.4).

The registry contains ten dominant instances plus the cross-lineage boundary pair and two extension controls (Table 1). Normalized profiles are in Figure 2, and readings for each model are in Table 1. The ten dominant architectures split into four interior-localized profiles, with an internal argmax and drift concentrated around a bottleneck, and six terminal-ramp profiles, with a final-block argmax and monotonic accumulation at $\rho \geq \tau$. The six include two families added after the freeze: Z-Image, a single-stream S3-DiT, and Sana, a linear-attention DiT. Extended measurement identifies a third, terminal-spike type: the Kandinsky family and Karlo all have a final-block argmax but are non-monotonic, with a localized spike at the terminal end and $\rho < \tau$; we report it as a boundary morphology, observed in two independent lineages. SD 2.1-unCLIP-small remains interior-localized, differing from SD 2.1 by $D_s = 0.0109$, so unCLIP conditioning alone does not distinguish terminal-spike from interior-localized profiles (Appendix C.1).

The four localized members concentrate their drift around an internal bottleneck: SD 1.5 and SD 2.1 in the decoder upsampling segment at 0.684, SDXL in `mid_block` at 0.429, and HunyuanDiT (H-DiT) in the middle layers at 0.500. The six ramp members show a strong increasing trend toward the output with the argmax on the final block, the lower bound being FLUX at $\rho = +0.9872$. The scale-invariant part of the ramp is its final-block argmax, or terminal topology. Monotonicity varies along the $3.2\times$ scale axis, where the 8.1B SD 3.5-Large reads $\rho = +0.839$ on the same layers (Appendix C.3), below the frozen $\tau = 0.90$, so the rule assigns the 8.1B tier to *other*: the terminal topology survives the scale axis, the ramp label does not. Within the ramp class, a final-block argmax means that D_{pp} reflects only a $1/L$ quantization difference, and the informative coordinates are the shape parameters conc and Gini of Appendix C.1.

Table 1: The accent registry. Peak position, profile monotonicity, and resolved readout subtype for each measured architecture, grouped into the three drift morphologies. The readout column lists resolved subtypes only. †Extension-registry instance. ‡$n = 103$, with one reference image dropped for an fp16 overflow under a rule fixed before measurement.

Model	Family	pp	ρ	Readout
Interior-localized ($n = 4$)				
SD 1.5	UNet	0.6842	+0.11	—
SD 2.1	UNet	0.6842	+0.08	—
SDXL	UNet	0.4286	+0.07	—
HunyuanDiT	DiT	0.5000	−0.37	—
Terminal-ramp ($n = 6$)				
FLUX	MM-DiT	0.9825	+0.9872	post-block accumulation
SD 3.5	MM-DiT	0.9583	+0.9983	terminal block peak
PixArt-Σ	DiT	0.9643	+1.0000	post-block accumulation
Lumina-Next	DiT	0.9583	+0.9991	unresolved
Z-Image†	single-stream S3-DiT	0.9667	+0.9893	unresolved
Sana-1600M†‡	linear-attention DiT	0.9500	+1.0000	unresolved
Terminal-spike, boundary morphology, two independent lineages				
Kandinsky-2.2	unCLIP-UNet	0.9808	+0.6004	post-block accumulation
Karlo-v1-alpha	unCLIP-UNet	0.9808	+0.5718	post-block accumulation

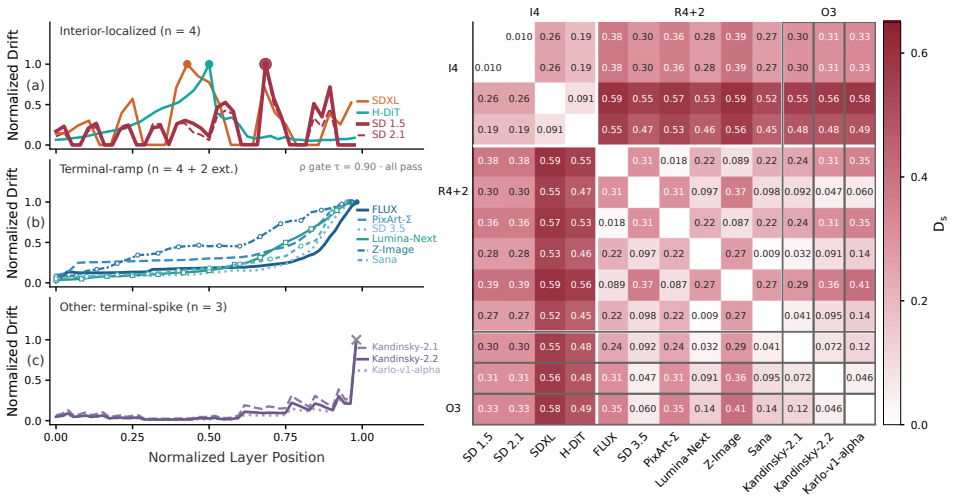

Figure 2: Normalized drift profiles and pairwise structural distances. (a) Interior-localized group. (b) Terminal-ramp group. (c) Terminal-spike group. The heatmap (right) gives pairwise structural distances among all 13 models.

Testing alternative explanations. Noise, family labels, and random initialization are examined in Appendices C.4, C.5, and D.1; readout truncation and feature-norm growth in §3.1 and Appendix C.9.

2.4 Registry-Level Separability of the Morphology Classes

We ask whether each dominant instance's morphology class is recoverable from the remaining dominant instances alone. The design was frozen before the data: each dominant instance is held out in turn and classified by nearest centroid, with centroids and feature normalization computed from training folds only and lineage and perturbation cells excluded to avoid leave-one-duplicate-out pseudo-replication.

Table 2: Leave-one-out prediction of the morphology class from each held-out dominant architecture, against the family-label baseline.

Feature space	Contents	LOO acc.	Exact p	margin	assign.	Verdict
64-pt profile	fold-specific resampled profile	10/10	1/210	0.95	$+1.45$	prediction holds
ortho-full	10-dim label-orthogonal scalars	10/10	1/210	1.15	$+2.08$	prediction holds
ortho-shape	9-dim, no positional quantity	10/10	1/210	0.88	$+1.76$	prediction holds
family-label	within-family majority baseline	$\leq 9/10$	—	—	—	below the accent

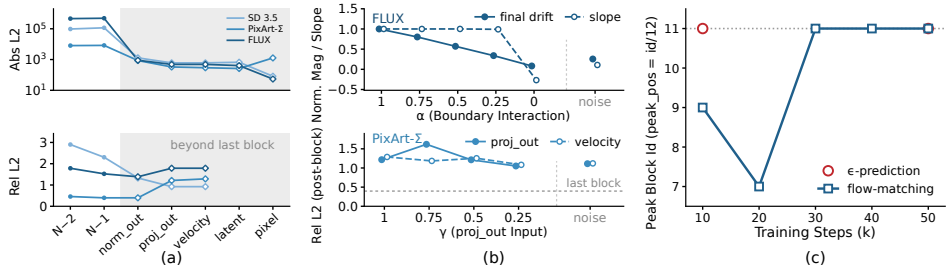

Figure 3: **Falsifying three trivial explanations.** (a) Stage-wise readout: the terminal morphology is not a hook-truncation artifact (§3.1; Figure C.1, Table C.6). (b) Two negative controls: local components do not govern morphology (§3.2). (c) Controlled retraining: peak block index along training steps for ϵ-prediction and flow matching, converging to the same peak layer (§3.3, Appendix D). Randomly initialized networks do not show the trained organization.

Leave-one-out prediction. All three feature spaces give $10/10$ (Table 2), including a 9-dimensional argmax-free set that drops the label-defining ingredients, peak position and monotonicity, so predictability does not rely on locating the peak; the score reaches the ceiling of the exact permutation null ($\binom{10}{4} = 210$ assignments). The effect size is carried by the separation margins in the table; margins, assignments, and the image bootstrap ($1000/1000$) are given in Appendix C.8. The object predicted is the morphology class under the canonical absolute metric (Appendix C.9).

What the design establishes. The result is leave-one-instance separability within the measured registry, whose closest pair, SD 1.5 and SD 2.1, is discussed in Appendix C.8.

Status of the boundary morphologies. Boundary decisions follow the monotonicity rule, which carries the order information an L2 profile distance cannot capture; the two cross-lineage instances share a post-block relative accumulation readout and differ only at the output end (Appendices C.6, C.8).

3 Mechanism and Diagnostic Implications

§2 established the two dominant morphologies across the ten dominant instances, the stability of their peak positions within lineages, and their registry-level separability under the frozen rule. This section asks what the accent measures: whether the terminal peak is a readout artifact (§3.1), whether local components control the morphology (§3.2), and how the organization relates to information pathways and to training (§3.3); §3.4 turns to its diagnostic uses (Figure 3).

3.1 Is the Accent a Readout Artifact?

Hypothesis. The terminal peak admits two trivial readings. *Hook truncation*: the final layer's absolute drift looks largest only because the readout stops there. *Norm inflation*: the absolute profile is a distortion caused by feature-norm growth. Both predict that no organization should survive beyond the final block or beyond the norm schedule.

Test. We extend the readout beyond the final block (Figure C.1, Table C.6; Appendix C.6). Five measured cases show two subtypes. SD 3.5 has a terminal block-localized peak: absolute drift falls by 99.5% after the final block and relative drift also declines monotonically, with the per-image argmax at block 23 (absolute) and block 22 (relative) in 100% of images. PixArt-Σ, FLUX, and the boundary pair instead show post-block relative accumulation. Their absolute drift also collapses by 96.5–99.9%, but relative drift overtakes the final block at proj_out, velocity, or latent readouts. Overtaking holds at all nine K indices for PixArt and on 87% of images for FLUX, whose relative drift rises from 1.52 to 1.79.

Verdict. The two readings separate cleanly, and only one of them is falsified. *Hook truncation: falsified.* Readout stages beyond the final block carry a different organization rather than a continuation of the same rise, so the terminal peak is not a boundary effect of where the hooks stop. The relative metric further shows that the absolute slope contains a norm-growth component: the within-block relative Spearman is ≈ 0 for SD 3.5 and -0.55 and -0.51 for PixArt and FLUX (Appendix C.6).

Norm inflation: confirmed as the carrier. Appendix C.9 quantifies the component: the absolute profile is rank-equivalent, not identical, to the feature-norm schedule across six models at $\rho \geq +0.929$, the norm term contributes 95%–146% of the ramp slope, and the terminal-ramp label does not persist under norm-free metrics, where the ramp profiles peak in the interior. The terminal-ramp class is therefore a readout of the feature-norm schedule, which is itself an architectural property stable within a lineage, and the readout is unbiased (Appendix C.9(d)). The relation is a carrier, not an identity: in Z-Image the reference norm declines terminally while the drift still rises (Appendix C.9(g)). The three evaluated interior-localized instances retain their internal peaks under all three metrics, so the distinction between the two forms survives norm-free measurement. Whether post-block relative overtaking reflects an active output mechanism is addressed by the probes of §3.2.

3.2 Is the Accent Controlled by Local Components?

A local-control account reads the morphology as a knob set by a single component, such as interaction strength or post-block readout, whose manipulation should scale or destroy the morphology rather than only the drift magnitude. We probed two candidate components in two terminal-ramp instances on $n = 12$ images with the same images and noise seed, interventions applied to both passes, and norm-matched controls (Appendix E). In FLUX, reducing interaction strength monotonically compresses drift magnitude, and the ramp survives a $4\times$ amplitude decay; full zeroing collapses the ramp, but so does norm-matched structural destruction, whereas norm-matched attenuation preserves it, implicating structural integrity rather than interaction strength alone. In PixArt-Σ, post-block relative overtaking does not vary with readout dose, and zeroing the terminal block leaves reconstruction essentially intact. Across 7–11 preset conditions on each of two independent architectures, neither component governs the morphology.

3.3 Information Pathways and Training Dynamics

The accent could reflect the architecture's information pathways. If it does, then in interior-localized architectures the mismatch should concentrate around the bottleneck containing the peak, the functional role of a same-named component can be opposite in different pathways, and the accent should form with training rather than being intrinsic to random initialization. The tests are run per architecture, on SD 1.5 and SDXL.

SD 1.5: skip connections as candidate carriers. In SD 1.5, each skip connection concatenates an encoder feature into the decoder block of the same resolution, so perturbing the skip changes what downstream blocks receive without changing the parameter count. We run four interventions and one correlation analysis on the skip tensor (Appendix E).

- **Cut A** cuts the peak skip, reducing drift by 27.7% and improving PSNR by $+2.20$ dB; **Cut B**, which cuts a low-drift skip, has almost no effect.
- **Noise A** replaces the skip content with noise, raising the L2 drift magnitude yet still improving PSNR; it is TOST-equivalent to Cut A at $\Delta = 0.5$ dB.

The interventions move the drift magnitude in opposite directions while both improving PSNR, so magnitude does not explain their shared gain: the gain tracks what the pathway carries, not how large

the drift is. The interventions change feature content, scale, and downstream computation together, so they identify intervention effects without identifying a mediation path. The bottleneck pathway itself shapes the accent: Cut A significantly changes $31/38$ layers (all 31 survive BH-FDR, $20/38$ under conservative Bonferroni), against $5/38$ for the low-drift skip ($4/38$ after BH-FDR).

SDXL: role reversal of a same-named component. The skip topologies of SD 1.5 and SDXL do not correspond one-to-one, so the cross-architecture comparison is limited to the sign level: the skip at the same structural position is a mismatch source in SD 1.5, where Cut A improves $+2.20$ dB, but a necessary information pathway in SDXL, where cutting it collapses PSNR by -11.59 dB. The role of a skip pathway is set by its position in the architecture's information pathways, which makes site effectiveness a property of the instance rather than of the morphology class.

Controlled retraining. Five random-initialization seeds of the same SD 1.5 architecture all show an "input-proximal" peak at 0.0263, far from the trained accent, with no trained organization; a DiT-S/2 (39.8M parameters, 12 blocks) trained from scratch under two objectives, ϵ-prediction and flow matching, converges from different starting points to the same peak layer, block 11 (Appendix D).

How the increments combine. Per-layer increments separate the two classes by accumulation behavior, not position. Against a matched-step random-direction baseline, the localized instance (H-DiT) cancels coherently, collapsing its net terminal displacement to $0.37\times$ the baseline ($0.05\times$ over the whole chain), while the ramp instances accumulate incoherently ($0.95\times$ PixArt-Σ, $1.08\times$ SD 3.5), with increments nearly orthogonal to the accumulated drift; the class contrast is large on all four endpoints (d 8.7–61.5; Appendix C.9(e)).

The interventions reveal pathway sensitivity that differs across architectures, and retraining under two objectives shows the measured organization emerging with training; the ramp class is accounted for by the norm-schedule decomposition of Appendix C.9.

3.4 Architecture-Dependent Sensitivity to Correction Sites

Site sensitivity is a property of the architecture. Injecting at the peak transition layer of HunyuanDiT gives $+5.65$ dB against $+2.50$ dB at the top-5 drift layers, a difference of 3.15 dB, with both sites fixed by the §2 peaks before any outcome is measured. Both sites are drift-derived and the selection images lie in the evaluation set, so this measures how much the choice among drift-informed sites matters, not an advantage over an information-free choice. SD 1.5 and FLUX are site-insensitive under the same protocol, and SD 1.5 shares HunyuanDiT's interior-localized label, so sensitivity does not follow the morphology class. A disjoint re-run reproduces the site list and verdict for SD 1.5 and FLUX; the HunyuanDiT leg retains the overlap caveat (Appendix A.2).

Matched-budget comparison of correction sites in SD 1.5. Both variants require neither training nor per-image optimization: latent correction, which uses no accent location, improves PSNR by $+3.67$ dB on the canonical 104 images, while feature correction at the accent's top-five layers yields $+3.48$ dB. At matched layer count, injection strength, and reconstruction steps, site selection changes the gain by only a few tenths of a dB against the ≈ 3.5 dB overall gain (Table 3; Appendix F). Accent selection exceeds the tested random sets at $\lambda = 0.3$ but falls within their range at $\lambda = 0.7$, and it trails norm-adjacent selection at both strengths.

Table 3: **Controlled comparison of correction sites (SD 1.5, P-multi@104).** ΔPSNR against the same harness's baseline. Every arm shares one inversion and one hook set, so only the corrected layer set differs. CIs and paired differences are in Appendix F.

Five-layer injection set	$\lambda = 0.7$	$\lambda = 0.3$
accent (drift top-5)	$+3.478$	$+3.415$
reference-norm ranks 6–10 (disjoint)	$\mathbf{+3.568}$	$\mathbf{+3.432}$
evenly spaced over canonical depth	$+3.418$	$+3.238$
uniformly random (mean of 10 seeds)	$+3.441$	$+3.217$

Editing fidelity. Accent-guided anchoring on 121 edit pairs reduces LPIPS (Zhang et al., 2018) by 85% against each pair's $\lambda = 0$ reconstruction, and $\lambda = 0.05$ already attains $> 90\%$ of that

improvement, so the anchoring strength needs little tuning (Figure F.1); this measures fidelity to the original image (Appendix F).

4 Related Work

Diffusion inversion and reconstruction error. Work on diffusion and score-based inversion (Song et al., 2021a;b) improves trajectory accuracy (Wallace et al., 2023), aligns features (Mokady et al., 2023), or uses more accurate inversion solvers (Rout et al., 2025; Wang et al., 2025), and recent flow-model methods add predictor–corrector inversion (Jiao et al., 2026) and velocity correction (Wang et al., 2026), while FlowEdit bypasses inversion (Kulikov et al., 2025). We study the organization of the feature discrepancy along depth. Edit-friendly DDPM noise spaces (Huberman-Spiegelglas et al., 2024) and Direct Inversion (Ju et al., 2024) eliminate the reconstruction residual by construction, while AIDI uses fixed-point refinement (Pan et al., 2023). The accent survives the removal of the discretization residual under all three frameworks (Appendix C.10). PIE-Bench and EditBench (Ju et al., 2024; Wang et al., 2023) address broader editing evaluation than our 121 controlled pairs.

Diffusion editing and architecture analysis. Attention and feature injection methods examine which internal states influence generation and editing (Hertz et al., 2023; Cai et al., 2025). Other studies analyze or manipulate diffusion representations (Luo et al., 2023a; Kwon et al., 2023; Deng et al., 2025), examine temporal heterogeneity across DiT feature dimensions (Zheng et al., 2026), or relate representation structure to memorization and generalization (Zhang et al., 2026), alongside comparisons across UNet, DiT, and MM-DiT backbones (Rombach et al., 2022; Peebles & Xie, 2023; Esser et al., 2024).

5 Discussion

Reconstruction–reference discrepancies are organized along depth reproducibly within a lineage and specifically to the architecture. Ten dominant instances across five backbone families fall into two dominant morphologies under the frozen rule, with peak positions stable across the tested weight, sampler, data-split, precision, noise-level, reference-seed, and inversion-framework perturbations, and leave-one-out prediction recovers each instance's class within the registry.

The two morphologies read out different things. The absolute profile of a terminal ramp is rank-equivalent to the feature-norm schedule of the reference features, and the ramp label does not survive norm-free metrics, whereas the tested interior-localized profiles retain interior peaks under all three metrics. The terminal topology of the ramp class is its scale-invariant part: the 8.1B tier keeps its final-block peak while its monotonicity reading moves below the frozen threshold, so the discrete label is the more fragile of the two quantities. Interventions and retraining locate the localized form: pathway perturbations change reconstruction quality in architecture-specific ways, and the organization emerges with training.

The same architecture dependence governs correction sites: cutting the peak skip improves SD 1.5 by $+2.20$ dB while the same-named component is a necessary pathway in SDXL, and site sensitivity ranges from a 3.15 dB spread across drift-informed sites in HunyuanDiT to site-insensitive behaviour in SD 1.5 and FLUX under the same protocol. In SD 1.5, matched-budget comparisons show the gain comes mostly from the correction operation rather than the choice among sites. The profiles therefore serve as a training-free generator of candidate sites whose value is established per architecture.

The claims are scoped to the measured registry: a near-census of public backbones under a fixed 104-image manifest, with leave-one-out separability stated within that registry and mechanism tests run per architecture. Site selection uses 19 images contained in the evaluation set, with disjoint re-runs completed for SD 1.5 and FLUX, and the fidelity result is measured by LPIPS on 121 controlled pairs, quantifying fidelity to the original image.

Architecture accents provide a reproducible, protocol-defined view of how reconstruction–reference feature discrepancies are organized across diffusion backbones. Norm-free measurements distinguish what the two dominant morphologies read out, while interventions show that pathway sensitivity differs across architectures. The profiles also offer candidate sites for training-free correction, whose value is evaluated for each architecture.

AI USE STATEMENT

We used generative AI tools throughout manuscript preparation to revise and polish the text and improve its organization and clarity. We also used these tools for execution, including developing and debugging experimental and analysis code, interpreting results, reviewing technical arguments, and identifying relevant literature.

The experimental measurements reported in this paper were obtained by running experiments and analysis code, including AI-assisted code. The authors retain responsibility for the research decisions and the final manuscript, including the validity of the methods, the accuracy of the results and references, and the conclusions drawn from the evidence.

ETHICS STATEMENT

This work uses only publicly released pretrained checkpoints, under their original licenses, and images from the COCO validation set. The training-dynamics experiments of Appendix D train small models on the same public data. No human subjects, personal data, or newly collected data are involved, and the editing measurements are run on a controlled 121-pair set of public images. The method improves the fidelity of training-free inversion, which like the editing pipelines it serves could be applied to manipulate real images. The release will cover measurement and correction code.

REPRODUCIBILITY STATEMENT

All experiments use publicly released pretrained checkpoints and images from the COCO validation set. Model versions, hub identifiers, and the implementation stack are listed in Appendix A.2, Table A.1. The canonical protocol, canonicalization of layer order, metric definitions, and statistical tests (paired t, Cohen's d_z, dual Pearson and Spearman reporting, BH correction, and TOST) are given in §2 and Appendices A.2 and B. The morphology classification rule, leave-one-out prediction, and OOD detector are given in §2.3 and §2.4. The pre-registration freeze, the extensions made after the freeze (the registry extension, the reference-noise-seed check, the norm-free re-judgment, the re-run with disjoint selection, and the exact-inversion re-measurement), and the data-integrity notes are recorded in Appendix A.2 and Appendix C. The repository containing the code will be released upon acceptance. It will include the measurement code, the per-image profiles, and the evaluation manifest.

REFERENCES

Y. Benjamini and Y. Hochberg. *Controlling the false discovery rate: A practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society: Series B, 57(1):289–300, 1995.

Black Forest Labs. *FLUX.1-dev* (model card). 2024. `https://huggingface.co/black-forest-labs/FLUX.1-dev`

M. Cai, X. Cun, X. Li, W. Liu, Z. Zhang, Y. Zhang, Y. Shan, and X. Yue. *DiTCtrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation.* CVPR, 2025.

J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li. *PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis.* ICLR, 2024.

J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li. *PixArt-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation.* ECCV, 2024.

Y. Deng, X. He, C. Mei, P. Wang, and F. Tang. *FireFlow: Fast inversion of rectified flow for image semantic editing.* ICML, 2025.

B. Efron. *Bootstrap methods: Another look at the jackknife.* Annals of Statistics, 7(1):1–26, 1979.

P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach. *Scaling rectified flow transformers for high-resolution image synthesis.* ICML, 2024.

A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. *Prompt-to-prompt image editing with cross attention control.* ICLR, 2023.

J. Ho, A. Jain, and P. Abbeel. *Denoising diffusion probabilistic models.* NeurIPS, 2020.

I. Huberman-Spiegelglas, V. Kulikov, and T. Michaeli. *An edit friendly DDPM noise space: Inversion and manipulations.* CVPR, 2024.

G. Jiao, B. Huang, K.-C. J. Wang, and R. Liao. *UniEdit-Flow: Unleashing inversion and editing in the era of flow models.* ICLR, 2026.

H. Jo, J. Maeng, J. H. Park, N. Ahn, and I. K. Park. *A plug-and-play approach for robust image editing in text-to-image diffusion models.* ICCV Workshops (MMFM), 2025.

X. Ju, A. Zeng, Y. Bian, S. Liu, and Q. Xu. *Direct inversion: Boosting diffusion-based editing with 3 lines of code.* ICLR, 2024.

T. Karras, M. Aittala, T. Aila, and S. Laine. *Elucidating the design space of diffusion-based generative models.* NeurIPS, 2022.

V. Kulikov, M. Kleiner, I. Huberman-Spiegelglas, and T. Michaeli. *FlowEdit: Inversion-free text-based editing using pre-trained flow models.* ICCV, 2025.

M. Kwon, J. Jeong, and Y. Uh. *Diffusion models already have a semantic latent space.* ICLR, 2023.

D. Lee, J. Kim, J. Choi, J. Kim, M. Byeon, W. Baek, and S. Kim. *Karlo-v1.0.alpha on COYO-100M and CC15M* (model card). 2022. `https://github.com/kakaobrain/karlo`

Z. Li, J. Zhang, Q. Lin, J. Xiong, Y. Long, X. Deng, Y. Zhang, X. Liu, M. Huang, Z. Xiao, et al. *Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained Chinese understanding.* arXiv:2405.08748, 2024.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. *Flow matching for generative modeling.* ICLR, 2023.

C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. *DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models.* Machine Intelligence Research, 22(4):730–751, 2025.

G. Luo, L. Dunlap, D. H. Park, A. Holynski, and T. Darrell. *Diffusion hyperfeatures: Searching through time and space for semantic correspondence.* NeurIPS, 2023.

S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao. *Latent consistency models: Synthesizing high-resolution images with few-step inference.* arXiv:2310.04378, 2023.

R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or. *Null-text inversion for editing real images using guided diffusion models.* CVPR, 2023.

Z. Pan, R. Gherardi, X. Xie, and S. Huang. *Effective real image editing with accelerated iterative diffusion inversion.* ICCV, 2023.

W. Peebles and S. Xie. *Scalable diffusion models with transformers.* ICCV, 2023.

D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. *SDXL: Improving latent diffusion models for high-resolution image synthesis.* ICLR, 2024.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. *Learning transferable visual models from natural language supervision.* ICML, 2021.

A. Razzhigaev, A. Shakhmatov, A. Maltseva, V. Arkhipkin, I. Pavlov, I. Ryabov, A. Kuts, A. Panchenko, A. Kuznetsov, and D. Dimitrov. *Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion.* EMNLP System Demonstrations, 2023.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. *High-resolution image synthesis with latent diffusion models.* CVPR, 2022.

L. Rout, Y. Chen, N. Ruiz, C. Caramanis, S. Shakkottai, and W.-S. Chu. *Semantic image inversion and editing using rectified stochastic differential equations.* ICLR, 2025.

D. J. Schuirmann. *A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability.* Journal of Pharmacokinetics and Biopharmaceutics, 15(6):657–680, 1987.

J. Song, C. Meng, and S. Ermon. *Denoising diffusion implicit models.* ICLR, 2021.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. *Score-based generative modeling through stochastic differential equations.* ICLR, 2021.

Stability AI. *Stable Diffusion 3.5 Medium* (model card). 2024. `https://huggingface.co/stabilityai/stable-diffusion-3.5-medium`

B. Wallace, A. Gokul, and N. Naik. *EDICT: Exact diffusion inversion via coupled transformations.* CVPR, 2023.

S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet, R. Soricut, J. Baldridge, M. Norouzi, P. Anderson, and W. Chan. *Imagen Editor and EditBench: Advancing and evaluating text-guided image inpainting.* CVPR, 2023.

J. Wang, J. Pu, Z. Qi, J. Guo, Y. Ma, N. Huang, Y. Chen, X. Li, and Y. Shan. *Taming rectified flow for inversion and editing.* ICML, 2025.

C. Wang, B. Zhu, and C. Zhang. *Free lunch for stabilizing rectified flow inversion.* ICLR, 2026.

E. Xie, J. Chen, J. Chen, H. Cai, H. Tang, Y. Lin, Z. Zhang, M. Li, L. Zhu, Y. Lu, and S. Han. *SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers.* ICLR, 2025.

Z-Image Team et al. *Z-Image: An efficient image generation foundation model with single-stream diffusion transformer.* arXiv:2511.22699, 2025.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. *The unreasonable effectiveness of deep features as a perceptual metric.* CVPR, 2018.

Z. Zhang, X. Li, X. Li, L. Shi, M. Wu, M. Tao, and Q. Qu. *Generalization of diffusion models arises with a balanced representation space.* ICLR, 2026.

S. Zheng, G. Chen, Q. Zhou, Y. Lin, L. He, C. Zou, P. Cai, J. Liu, and L. Zhang. *Let features decide their own solvers: Hybrid feature caching for diffusion transformers.* ICLR, 2026.

L. Zhuo, R. Du, H. Xiao, Y. Li, D. Liu, R. Huang, W. Liu, X. Zhu, F.-Y. Wang, Z. Ma, X. Luo, Z. Wang, K. Zhang, L. Zhao, S. Liu, X. Yue, W. Ouyang, Y. Qiao, H. Li, and P. Gao. *Lumina-Next: Making Lumina-T2X stronger and faster with Next-DiT.* NeurIPS, 2024.

CONTENTS OF THE APPENDIX

A Protocol Derivation and Experimental Setup 14
A.1 Protocol derivation: DDPM forward, DDIM inversion, and the consistency proposition 14
A.2 Experimental setup 15
B Metrics, Noise Floor, and Statistical Tests 20
B.1 Metric definitions 20
B.2 Noise floor and statistical tests 20
C Robustness, Resolution, and the Scale Axis 21
C.1 Full-model accent extension data 21
C.2 Robustness overview 22
C.3 Scale-axis test 23
C.4 Family-label control 23
C.5 Resolution, lineage supplements, and boundary distances 24
C.6 Terminal stage-wise readout: pre-registered decisions and per-stage values 25
C.7 The confound matrix: pre-registered and reference-noise-seed cells 26
C.8 Full statistics of leave-one-out prediction and the OOD detector (details of §2.4) 28
C.9 Decomposing the norm schedule: what the absolute profile reads out (details of §2.1) 29
C.10 Exact-inversion frameworks: re-measuring the accent with the residual removed 33
D Training Dynamics and Crystallization Experiments 36
D.1 Random-initialization exclusion 36
D.2 Dual-objective crystallization and data integrity 36
E Additional Mechanism Analyses 37
F Full Correction and Editing Results 39

A Protocol Derivation and Experimental Setup

A.1 Protocol derivation: DDPM forward, DDIM inversion, and the consistency proposition

The DDPM forward process (Ho et al., 2020) is $q(z_t \mid z_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\, z_{t-1}, \beta_t I)$, with marginals $q(z_t \mid z_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\, z_0, (1-\bar{\alpha}_t) I)$, where the training noise schedule β_t fixes $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, and the reparameterization $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$, which is precisely the theoretical origin of the external reference of §2.2. DDIM (Song et al., 2021a) gives the deterministic update $z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, \hat{z}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\, \epsilon_\theta(z_t, t)$ from $\hat{z}_0 = (z_t - \sqrt{1-\bar{\alpha}_t}\, \epsilon_\theta(z_t, t))/\sqrt{\bar{\alpha}_t}$. Inversion runs the same map in the reverse direction, $z_{t+1} = \sqrt{\bar{\alpha}_{t+1}}\, \hat{z}_0(z_t) + \sqrt{1-\bar{\alpha}_{t+1}}\, \epsilon_\theta(z_t, t)$, approximating $\epsilon_\theta(z_{t+1}, t)$ by $\epsilon_\theta(z_t, t)$ and thereby introducing a per-step local discretization error

$$\delta_t = \sqrt{1-\bar{\alpha}_{t+1}}\, [\epsilon_\theta(z_{t+1}, t) - \epsilon_\theta(z_t, t)]\,,$$

which is one of the two sources of the per-layer drift of §2.1. The two sources are separated cleanly at the level of the measurement. The first is the *coupling difference* between the reconstruction trajectory and the externally drawn reference, which is present even when the inversion is error-free, because the reference is a different sample of the same conditional law, as the analytic control below shows. The second is the accumulated local error $\sum_t \delta_t$ inside the inversion–reconstruction loop, which with a finite step size is unavoidable and which Appendix C.10 removes by construction. Everything this paper reports about the *organization* of drift concerns the first source as read out through the backbone, and removing the second entirely leaves the peak and the morphology unchanged. A unified view is available. DDIM is a discretization of the probability-flow ODE (Song et al., 2021b) $dz = [f(t)z - \frac{g^2(t)}{2} \nabla_z \log p_t(z)]\, dt$, with $\nabla_z \log p_t \approx -\epsilon_\theta/\sqrt{1-\bar{\alpha}_t}$. For flow-matching models such as FLUX, SD 3.5, and Lumina-Next, the trajectory is defined by a velocity field $v_\theta(z, t)$, and inversion and reconstruction are Euler discretizations of $dz/dt = \pm v_\theta(z, t)$ with per-step local error $O(h^2)$. All registry entries, boundary instances, and extension controls thus sit in one and the same ODE framework, giving the protocol design of "cross-architecture comparability" a mathematical basis.

Proposition 1 (Conditional consistency of ideal inversion)**.** *Assume (i) neither inversion nor reconstruction incurs discretization error ($\eta = 0$, the trajectory is exactly self-inverse), (ii) ϵ_θ is exact on the reference trajectory, and (iii) the reconstruction endpoint is coupled to the reference sample, $z_T^{\mathrm{recon}} = z_T^{\mathrm{ref}}$, with the reference drawn from the backbone's own forward law $z_t^{\mathrm{ref}} = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$ and held fixed. Then $z_t^{\mathrm{recon}} = z_t^{\mathrm{ref}}$ for all t, and hence $d_l(x) = 0$ for every layer. Assumption (iii) is a coupling condition on the reference path that has to be imposed; it is not implied by (i) or (ii).*

Proof (by induction). Assume $z_t^{\mathrm{recon}} = z_t^{\mathrm{ref}}$. Reparameterizing $\hat{z}_0$ together with the exactness of ϵ_θ at the reference point gives $\hat{z}_0 = z_0$. The deterministic update applies the same ϵ_θ at the same point, so $z_{t-1}^{\mathrm{recon}} = \sqrt{\bar{\alpha}_{t-1}}\, z_0 + \sqrt{1-\bar{\alpha}_{t-1}}\, \epsilon_\theta = z_{t-1}^{\mathrm{ref}}$. The base case $z_T^{\mathrm{recon}} = z_T^{\mathrm{ref}}$ is given by the idealization, and induction runs down to $t = 0$. Hence $f_l^{\mathrm{ref}} = f_l^{\mathrm{recon}}$ layer by layer, and $d_l(x) = 0$. □

The coupling assumption. Assumption (iii) does not follow from exact inversion or from a correct marginal score: exact inversion guarantees only that the trajectory integrates back to z_0, while the reference is an independently drawn sample. The frameworks of Appendix C.10 make the inversion–reconstruction pair mutually inverse but say nothing about the external reference, and without the coupling the implication fails, as the analytic control below shows.

Analytic control: what the metric is zero for. The zero point of the measurement is the reference *coupling*, not the absence of inversion error. Take Gaussian data and the VP process $dz = -\frac{1}{2}\beta(t) z\, dt + \sqrt{\beta(t)}\, dW$. The marginals are $\mathcal{N}(0, I)$ at every t, the exact score is $\nabla_z \log p_t(z) = -z$, and the probability-flow ODE $dz = [-\frac{1}{2}\beta(t)z + \frac{1}{2}\beta(t)z]\, dt$ has zero velocity. Exact inversion and reconstruction therefore both hold $z(t) = z_0$, they are mutually inverse, and the end-to-end residual is exactly zero. The measured drift against an independently drawn reference $\alpha_t z_0 + \sigma_t \varepsilon_{\mathrm{ref}} \neq z_0$ is nevertheless positive, with $d_l = \mathbb{E}\|\mathcal{P}(h_l(\alpha_t z_0 + \sigma_t \varepsilon_{\mathrm{ref}})) - \mathcal{P}(h_l(z_0))\|_2 > 0$

for any layer map h_l that separates the two states, and it would be zero only under the coupled reference z_0 itself. The empirical counterpart is the same statement: at an end-to-end residual of exactly zero, the frameworks of Appendix C.10 still read 81–90% of the canonical amplitude. The object of this paper is thus not inversion error but the backbone's response to a state drawn from its own forward law. The claim under test is that this response's organization depends on the architecture and is independent of the residual.

A.2 Experimental setup

Models. Registry models, lineage variants, and crystallization models are listed in Table A.1. The forward families of Definition 1 are the variance-preserving DDPM family for SD 1.5, SD 2.1, SDXL, HunyuanDiT, PixArt, Kandinsky, and Karlo, and the rectified-flow or linear-interpolant family for FLUX, SD 3.5, Lumina-Next (Zhuo et al., 2024), Z-Image (Z-Image Team et al., 2025), and Sana (Xie et al., 2025). Versions and hub identifiers are as follows. SD 1.5 is v1-5 (Rombach et al., 2022). SD 2.1 is stabilityai/stable-diffusion-2-1-base, the successor official checkpoint of the same SD 1.x UNet family. SDXL is base 1.0 (Podell et al., 2024). HunyuanDiT is v1.2 (Li et al., 2024). FLUX is FLUX.1-dev (Black Forest Labs, 2024). SD 3.5 is Medium (Stability AI, 2024). PixArt-Σ is the 1K model (Chen et al., 2024b). Lumina-Next is Alpha-VLLM/Lumina-Next-SFT-diffusers, a LuminaNextDiT2DModel with 24 layers, hidden size 2304, 2D-RoPE joint and cross attention, flow matching, and a Gemma-2 text encoder. Kandinsky-2.2 (Razzhigaev et al., 2023) is the kandinsky-community series, an SD-style UNet with 52 layers plus an unCLIP prior, run under prior conditioning with empty-prompt image_embeds, seed 42, and 25 steps. Lineage variants: SD 1.4, Realistic Vision, LCM LoRA (Luo et al., 2023b), and RandText form the SD 1.5 family, SDXL-Turbo the SDXL family, and PixArt-α (Chen et al., 2024a) the PixArt family. The boundary-morphology validation set and the extension controls are the following. Kandinsky-2.1, the predecessor checkpoint of Kandinsky-2.2, serves as the within-lineage replication. Karlo-v1-alpha (Lee et al., 2022), a UNet plus unCLIP prior, serves as the cross-lineage replication. SD 2.1-unCLIP-small, stable-diffusion-2-1-unclip-small, is the unCLIP-conditioning control, and it lands in the interior-localized class and shows no terminal spike. DiT-S/2 is trained from scratch with ϵ-prediction DDPM and with flow matching on 111 images for 50k steps for the crystallization experiment of Appendix D.2; its 39.8M parameter count is the measured count of the instance trained here, where the original paper reports 33M and the difference comes from implementation details, and all experiments use this instance. Two entries are registered by backbone structure rather than by a parameter budget: Lumina-Next and Kandinsky-2.2 do not list total parameter counts including text encoders and priors.

Table A.1: **All experimental models: architecture family, scale, and role.** Versions and hub identifiers are given above in this section.

Model	Family	Scale / structure	Role
SD 1.5	UNet	∼860M, 38 layers	backbone (localized)
SD 2.1	UNet	∼860M, 38 layers	backbone (localized)
SDXL	UNet	∼2.6B	backbone (localized)
HunyuanDiT	DiT (cross-attn)	∼1.5B	backbone (localized)
FLUX	MM-DiT (dual-stream)	∼12B	backbone (ramp)
SD 3.5	MM-DiT (dual-stream)	∼2.5B	backbone (ramp)
PixArt-Σ	DiT (cross-attn)	∼0.6B	backbone (ramp)
Lumina-Next	DiT (2D-RoPE joint/cross)	24 layers / hidden 2304	backbone (ramp)
Kandinsky-2.2	UNet + unCLIP prior	UNet, 52 layers	boundary morphology (label *other*)
SD 1.4 / RV / LCM / RandText	UNet	∼860M	lineage variants (SD 1.5 family)
SDXL-Turbo	UNet	∼2.6B	lineage variant (SDXL family)
PixArt-α	DiT (cross-attn)	∼0.6B	lineage variant (PixArt family)
DiT-S/2	DiT	39.8M	crystallization (Appendix D)

Data. The canonical evaluation set coco_val100 is a fixed manifest of 104 COCO val2017 images, released with the code upon acceptance as `outputs/coco_val100_manifest.txt`. Every script consumes it in sorted filename order, so readings for different models are index-aligned image by image. The set supports the morphology discovery of §2 and the quantitative evaluation of §3.4, and the 121 edit triplets of §3.4 span all 104 images.

Set overlap and the disjoint re-run. The 19-image set used for the site selection of §3.4 is a strict subset of these 104, with all 19 contained in it and verified by filename, so the site selection and the main evaluation share images. What is independent is only the site-selection *procedure*, with the peaks fixed before any outcome is measured. Because the shared images could in principle bias the comparison in favour of accent sites, the site selection is re-run on a selection set rebuilt to be disjoint from the evaluation set, applying the same sampling rule after excluding the 104, with the manifest released upon acceptance. On that disjoint set the SD 1.5 site ranking reproduces the ranking computed on the 104-image evaluation set, with $\rho = +0.9989$ and identical top-5 and top-10, and injection at the disjoint-derived sites scores $+3.470$ dB against $+3.302$ dB for the original site list on the same 104 images, a paired difference of $+0.168$ dB with CI95 $[0.139, 0.200]$. The disjoint selection does not favour accent sites. For FLUX the profile from disjoint selection likewise reproduces the canonical one, with $\rho = +0.9993$, identical top-5, peak position identical at 0.9825, and concentration 0.496 versus 0.495. The HunyuanDiT leg was not re-run for this version, so its numbers retain the overlap caveat. A 100-prompt generated-image validation of the correction uses a different, independent source and is unaffected. The original sampling rule of the 104-image set is not reconstructible from the records. The set is a fixed manifest, and reproducibility is carried by the manifest IDs rather than by the rule. Each section notes the set it uses, and dataset reuse is disclosed in §3.4.

Protocol. The canonical protocol is P-multi: DDIM with 50 steps, empty prompt, and $\eta = 0$. FLUX, SD 3.5, and Lumina-Next use Euler with 50 steps on the sigma grid (Karras et al., 2022). Kandinsky-2.2 is inverted under prior conditioning with image_embeds and seed 42 under the same P-multi protocol, so the K set and step grid follow the same rule. The time-step set takes 9 indices under a relative rule, the first three, middle three, and last three indices of the N-step scheduler grid, which at the canonical $N = 50$ are $\{0, 1, 2, 24, 25, 26, 47, 48, 49\}$, used for every model in this paper, with all indices taken when $N \leq 6$; the layer order is canonicalized and then aggregated layer by layer. The single-step protocols P-t0 and P-tT ($K = 1$) are degenerate approximations retained only as robustness controls and, for P-tT, as the source of early-stage values (Table A.2). Known protocol disagreements, for example the SDXL peak flipping $0.964 \rightarrow 0.429$ and the SD 3.5 peak moving from block 22 to block 23, are recorded, and the main text uses P-multi only.

Table A.2: **Protocols: P-multi / P-t0 / P-tT.**

Protocol	K	Reference	Status
P-multi	relative rule {first 3, middle 3, last 3}, 9 indices in total	independently sampled forward reference	**canonical**
P-t0	$\{t \approx 0\}$, one step	last inversion step	degenerate approximation ($K = 1$), robustness appendix
P-tT	$\{t \approx T\}$, one step	first inversion step	degenerate approximation ($K = 1$), source of early-stage values

Measurement-path specification. For each image x, the VAE encoder produces z_0. The measurement then follows three distinct paths. An independent forward-reference pass evaluates $z_t^{\text{ref}} = \alpha_t z_0 + \sigma_t \varepsilon_t$. Inversion maps z_0 to z_T^{inv}. Reconstruction starts from that endpoint and follows the matching reverse trajectory, producing z_t^{recon}. The drift compares only the reference and reconstruction features at the same image, timestep, and hooked layer, $\mathcal{P}(f_{l,t}^{\text{ref}})$ against $\mathcal{P}(f_{l,t}^{\text{recon}})$. Latent correction instead uses the paired states z_t^{inv} and z_t^{recon}, and feature correction uses the cached inversion feature and the current reconstruction feature at a matched layer and timestep. Thus the external reference is not the correction target. The intervention arms differ in scope: the SD 1.5 probes patch the inversion, reference, and reconstruction forwards across all 50 DDIM steps and the nine reference reads, latent and feature correction act on the reconstruction pass only at the

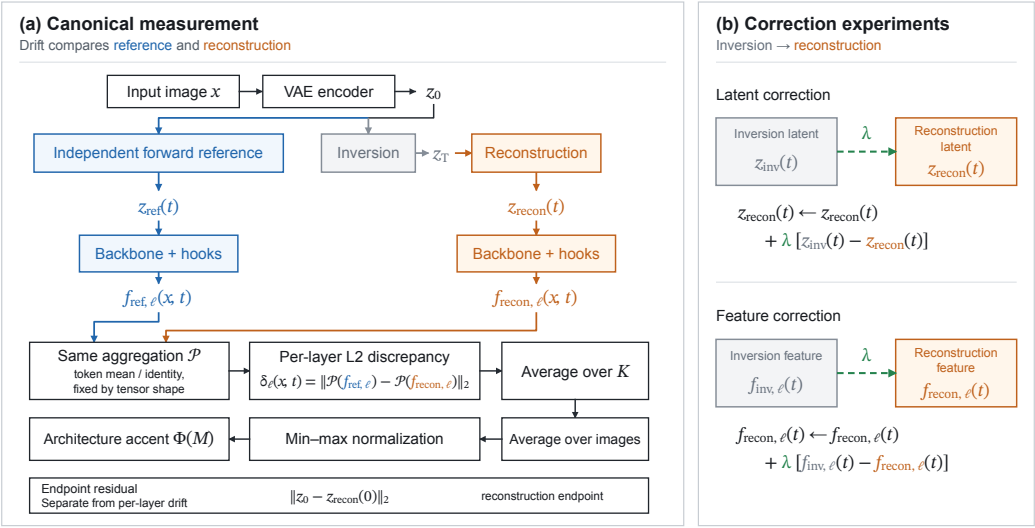

Figure A.1: **Measurement protocol and correction operations.** (a) The reference leg and the inversion–reconstruction leg, read through the same backbone at matched timestep and conditioning. (b) The two correction operations, which act on the reconstruction state only; no correction edge enters the measurement. Both panels are specified in the text above and in §2.2.

matched scheduler timestep of the 50-step trajectory, the FLUX endpoint correction is a single operation on the endpoint latent after the 28-step Euler pair with no intermediate timestep injection, and the two local-component probes, FLUX interaction scaling and PixArt-Σ readout scaling, stay registered for the whole condition run, so their reference reads receive the same patch. In the canonical P-multi measurements, matched conditions share the image manifest, conditioning, the 50-step scheduler grid, the nine relative timestep indices, the model-specific seed-42 noise protocol, and the architecture-specific hook list. The noise coupling is model-specific: for example, SD 1.5 and PixArt-Σ draw a separate ε_t at each selected index from a reset seeded stream, whereas Z-Image reuses one per-image ε across the nine indices. Within each image–timestep–layer pair, the protocol first applies $\mathcal{P}$, the token mean for $[B, N, C]$ outputs and the identity for $[B, C, H, W]$ outputs, and then computes the feature distance. It averages the nine timestep readings within image and averages images last. This ordering preserves the image as the pairing unit and is also the required order for the decompositions in Appendix C.9. Figure A.1 diagrams this measurement path together with the two correction operations, and Table A.3 records the hook path and the intervention operation of each probe and correction arm.

Implementation notes for Table A.3. Every arm runs the canonical P-multi protocol with 50 steps, the nine relative timestep indices, seed 42, and empty prompt. The SD 1.5 skip-pathway arms patch `unet.up_blocks[2].forward`, for Cut A and Noise A, and `unet.up_blocks[0].forward`, for Cut B, through the `res_hidden_states_tuple` argument, whose entries are runtime skip tensors of shape $[B, C_i, H_i, W_i]$ that the scripts do not serialize numerically. The patch is active in the inversion, reference, and reconstruction passes across all 50 DDIM steps and the nine reference reads. The FLUX arm registers a forward pre-hook on `transformer.single_transformer_blocks[0]` and rescales its `hidden_states` keyword, which in the diffusers 0.38 layout is the image-only token stream $[B, N_{\mathrm{img}}, 3072]$ with $N_{\mathrm{img}} = (H/16)(W/16)$ at the native image size. Its `wrong` control applies the same zeroing at `single_transformer_blocks[18]`, and its norm-matched control replaces the stream by Gaussian noise rescaled to its L2 norm in fp32. The PixArt-Σ arm registers a forward pre-hook on `transformer.proj_out`, on PixArt-Σ-XL-2-1024-MS with images at 1024×1024, so the hooked input is $[1, 4096, 1152]$, and rescales its first positional input, the modulated stream leaving `norm_out`. Since `norm_out` is an affine-free LayerNorm and therefore scale-invariant, the readout knob is the `proj_out` input. Both local-component hooks stay registered for the whole condition run, so the reference reads receive the same patch. Both are $n = 12$ pi-

Table A.3: **Intervention and correction implementation record.** Intervention target and update rule of each arm. Call sites, tensor dimensions, and shared protocol parameters are in the implementation notes following the table, and the feature-correction layer sets in Table A.4.

Experiment / model	Intervention target	Update rule	Cache / execution
(a) Mechanistic interventions			
SD 1.5 · Cut A	`unet.up_blocks[2]` skip inputs	$s_i \leftarrow 0$ for every tuple entry	No cache; all paths rerun per arm
SD 1.5 · Cut B	`unet.up_blocks[0]` skip inputs	$s_i \leftarrow 0$ for every tuple entry	No cache; all paths rerun per arm
SD 1.5 · Noise A	`unet.up_blocks[2]` skip inputs	$s_i \leftarrow \mathrm{randn_like}(s_i)\,\mathrm{std}(s_i) + \mathrm{mean}(s_i)$: per-tensor mean and standard deviation matched, not the L2 norm	Noise redrawn at every call, not cached; all paths rerun
FLUX · interaction scaling	image-token stream entering `single_transformer_blocks[0]`	$h \leftarrow \alpha h$, $\alpha \in \{0.75, 0.5, 0.25, 0\}$	Active in inversion, reconstruction, and the reference reads; noise control redrawn per call
PixArt-Σ · readout scaling	input to `proj_out` (post-`norm_out` stream)	$h \leftarrow \gamma h$, $\gamma \in \{0.75, 0.5, 0.25, 0\}$	Active in both passes; noise control replaces the same input
(b) Correction experiments			
SD 1.5 · latent correction	latent before each UNet call	$z_t^{\mathrm{recon}} \leftarrow z_t^{\mathrm{recon}} + \lambda\,(z_t^{\mathrm{inv}} - z_t^{\mathrm{recon}})$, $\lambda = 0.7$	Inversion latents cached on CPU by timestep; reconstruction rerun from the cached terminal state
FLUX · endpoint correction	unpacked VAE latent after flow reconstruction	$z_0^{\mathrm{corr}} \leftarrow z_0^{\mathrm{recon}} + \lambda\,(z_0 - z_0^{\mathrm{recon}})$, $\lambda = 0.7$, pulling the endpoint toward the original latent z_0	Single operation on the endpoint; only the corrected latent is decoded, no trajectory recomputation
SD 1.5, SDXL, HunyuanDiT · feature correction	selected residual and transformer blocks (Table A.4)	$f_{\ell,t}^{\mathrm{recon}} \leftarrow f_{\ell,t}^{\mathrm{recon}} + \lambda\,(f_{\ell,t}^{\mathrm{inv}} - f_{\ell,t}^{\mathrm{recon}})$, $\lambda = 0.7$	Inversion features detached to CPU and cached by timestep; reconstruction rerun with hooks

lots with no per-image inference, and their records are `flux_boundary_sweep_pilot12.json` and `pixart_output_field_sweep_pilot12.json` under `outputs/confound_matrix/runs/`. Latent correction is `ddim_reconstruct_with_editing` in `scripts/phase7_editing_benchmark.py` with `latent_correction` and `corr_lam` $= 0.7$. The endpoint variant calls `apply_correction_latent` from `scripts/flux_common.py` as `apply_correction_latent(z_0_recon_raw, z_0, lam=0.7)`, so its pull-back target is the original encoded latent and no intermediate timestep is injected. Feature correction hooks the selected blocks and restores the matching inversion feature $f^{\text{inv}}_{\ell,t}$ at each of the nine timesteps of the reconstruction pass. Per-layer feature-map shapes are not serialized by those scripts, so the hooked outputs are recorded by shape family, $[B, C, H, W]$ for the UNet blocks and $[B, N, D]$ for HunyuanDiT.

Table A.4: **Feature-correction layer sets.** The top-five drift layers per architecture, in the canonical layer order used throughout; HunyuanDiT selects the terminal five transformer blocks.

Model	Selected layers	
SD 1.5	`up_blocks.2.resnets.0`	`mid_block.resnets.1`
	`up_blocks.0.resnets.0`	`up_blocks.1.resnets.1`
	`mid_block.resnets.0`	
SDXL	`mid_block.resnets.1`	`up_blocks.0.resnets.0`
	`up_blocks.0.resnets.1`	`mid_block.resnets.0`
	`down_blocks.2.resnets.1`	
HunyuanDiT	`blocks.16–blocks.20`	

Statistical testing setup. The noise floor is the distribution of D_{mag} between profiles resampled by $B = 1000$ bootstraps (Efron, 1979) of the same model, with median $= 0.0014$ and $p_{95} = 0.0042$, and the peak position is invariant under resampling, so $D_s = D_{\text{mag}}$ at the floor. The unified tests are paired t (two-sided), Cohen's d_z, dual reporting of Pearson r and Spearman ρ, BH correction (Benjamini & Hochberg, 1995), and TOST equivalence (Schuirmann, 1987), with the equivalence margins set as in Appendix B.2.

Pre-registration record and post-freeze extensions. All pre-registered parameters (the morphology classification rule, the TOST equivalence margin 0.1245, the leave-one-out feature spaces, and the stage-wise readout decision rules) were frozen before the extended registry and the confound-matrix analyses made after the freeze. The frozen record and the configuration hashes will be released with the code upon acceptance. The permutation-test p values were recomputed after the freeze by exact enumeration instead of the $B = 1000$ Monte Carlo estimate, with the null hypothesis, test statistic, and design unchanged and only the inference made exact. The registry extension admitted Z-Image and Sana under the same frozen rule, and the leave-one-out prediction, exact enumeration, separation margins, and OOD-detector statistics were recomputed on the ten-instance registry under the same procedure, replacing the original 8-instance report of Appendix C.8 with the classification rule unchanged. The two dominant morphologies, the leave-one-out prediction, and the confound-matrix decisions are therefore confirmatory analyses, while the boundary instances and extension controls are extended evidence measured after the freeze under the same canonical protocol, which is why the boundary morphology is reported to the scope of its evidence in §2.3. The acceptance rule of the reference-noise-seed robustness check, peak locking and a ρ range below 0.02, was pre-registered and executed under that rule, with all four representative models passing. The check re-runs the reference leg only and reuses the inversion and reconstruction trajectories for every seed, so the canonical artifacts are not modified. The decomposition by norm schedule and the norm-free re-judgment of Appendix C.9 are extension measurements made after the freeze, as are the element-count re-normalization re-judgment of Appendix C.9(b), the re-run with disjoint selection described under Data above, and the exact-inversion re-measurement of Appendix C.10, which reuses the canonical hook set, reference noise, and z_0 and reports the residual of each framework as an implementation gate. The pooling correction of the norm probe, which made the token-mean pooling consistent with the probe family, was completed inside that extension and affects none of the reported numbers of the frozen analyses. Measurement scripts, per-image profiles, and the evaluation manifest will be released with the code upon acceptance.

B METRICS, NOISE FLOOR, AND STATISTICAL TESTS

B.1 METRIC DEFINITIONS

Explicit definitions of concentration and Gini, identical across the paper, are

$$\mathrm{conc}(d) = \frac{\sum_{l \in \mathrm{top}20\%(d)} d_l}{\sum_{l=1}^{L} d_l},$$
$$\mathrm{Gini}(d) = \frac{\sum_{i=1}^{L} \sum_{j=1}^{L} |d_i - d_j|}{2L \sum_{i=1}^{L} d_i},$$

where $\mathrm{top}20\%(d)$ is the index set of the top 20% drift layers, with $k = \lceil 0.2L \rceil$, and the Gini coefficient uses the mean-difference normalization, where 0 means uniform and 1 means concentrated in a single layer.

The two profiles A, B are compared through three coordinates. The peak-position distance is a grid distance,

$$D_{\mathrm{pp}}(A, B) = |pp_A - pp_B|\,,$$
$$pp = \frac{1}{L} \arg\max_l d_l \;\in\; \left\{0, \tfrac{1}{L}, \ldots, \tfrac{L-1}{L}\right\},$$

where l is a zero-based layer index in canonical order, so pp is the index of the peak layer divided by the layer count, giving $23/24 = 0.9583$ for SD 3.5 and $56/57 = 0.9825$ for FLUX. Min–max normalization of the profile does not move its argmax, so peak position is a property of the raw profile. Peak positions are quantized in units of $1/L$, and comparing pp across architectures of different depth carries that $1/L$ resolution. The shape-coordinate distance and the structural distance are

$$D_{\mathrm{mag}}(A, B) = \sqrt{(\mathrm{conc}_A - \mathrm{conc}_B)^2 + (\mathrm{Gini}_A - \mathrm{Gini}_B)^2},$$
$$D_s = \sqrt{D_{\mathrm{pp}}^2 + D_{\mathrm{mag}}^2}.$$

A structural distance is read against the noise floor below: under resampling of one and the same model, $D_{\mathrm{pp}} \equiv 0$ since the peak is invariant, so $D_s = D_{\mathrm{mag}}$ at the floor.

D_s compares two profiles through three statistics (peak position, concentration, and Gini) and is not a metric on the full shape of a curve. Where full-shape agreement matters we report the profile Spearman ρ alongside, which is also the quantity that carries the order information D_s cannot capture and that the morphology rule of §2.3 uses for exactly that reason.

Measurement implementation note. Layer-order canonicalization corrects two classes of problems: lexicographic ordering, where `block_10` sorts before `block_2`, and UNet within-block ordering, where attentions and resnets are sorted alphabetically, the reverse of execution order. The correction is implemented with a sort key computed per architecture and verified against structural hashes. This canonical metric, combined with $D_s = \sqrt{D_{\mathrm{pp}}^2 + D_{\mathrm{mag}}^2}$, is used for all models and all subsequent analyses, including peak position, shape distance, leave-one-out prediction, and the confound matrix, all of which come from the same metric, and it depends on no tunable shape parameter left undefined in this paper.

B.2 NOISE FLOOR AND STATISTICAL TESTS

Under P-multi, the noise floor has median $= 0.0014$ and $p_{95} = 0.0042$. Under the single-step P-tT protocol with $B = 100$, the median is 0.0071 and $p_{95} = 0.0163$. The equivalence margins are set as follows. A margin of 0.2 dB is about 5% of the between-group PSNR standard deviation of 3.0–4.0 dB and far below the correction gain of $+2.75$ dB, so by the SESOI convention it counts as a negligible difference. A margin of 0.5 dB is used for the qualitative equivalence claim of Noise A, since both cutting and noise replacement improve PSNR by more than $+2.2$ dB and only a loose margin is needed. The TOST equivalence margin for structural distance is 0.1245, a pre-registered value set to half of the smallest cross-morphology $D_s = 0.249$ in the existing cross-architecture

summary. After the registry extension, the smallest cross-morphology pair is SD 1.5–Sana at $D_s = 0.2679$, reported in Appendix C.5, still above $2\times$ the margin (0.249). The two *a priori* requirements admit any margin that is at least an order of magnitude above the measurement noise, $\delta \geq 0.042$, and below half the smallest between-class distance, $\delta < 0.134$. The chosen margin, 0.1245, satisfies both design constraints and lies within the verdict-stability interval $(0.0989,\ 0.2679)$ over which all confound-matrix verdicts are unchanged, at $\approx 30 \times p_{95}$. As a precision ablation, the per-step drift trajectories comparing the full fp16 and bf16 pipelines give Pearson $r = 0.9982$ with a systematic bias of 5.8%, far below the dynamic range of cross-layer drift magnitudes of $1000\times$. Numerical precision does not change the accent structure.

C Robustness, Resolution, and the Scale Axis

C.1 Full-model accent extension data

Table C.1: **Full-model accent extension data (P-multi@104; boundary instances and extension controls included).** pp is the normalized peak position, conc the share of the top 20% drift layers ($k = \lceil 0.2L \rceil$), Gini the concentration field, and ρ the Spearman correlation between layer order and profile; formal definitions and the peak layers are given in Appendix B.1 and the text below.

	Localized backbones				Ramp backbones				Boundary	
	SD 1.5	SD 2.1	SDXL	H-DiT	FLUX	SD 3.5	PixArt-Σ	Lumina-Next	Kandinsky -2.2	Karlo -v1-alpha
pp	0.6842	0.6842	0.4286	0.5000	0.9825	0.9583	0.9643	0.9583	0.9808	0.9808
conc	0.5830	0.5771	0.5672	0.6095	0.4952	0.6942	0.4949	0.6060	0.6628	0.6991
Gini	0.5876	0.5956	0.5703	0.6076	0.3757	0.6077	0.3768	0.5675	0.6350	0.6630
ρ	$+0.11$	$+0.08$	$+0.07$	-0.37	$+0.9872$	$+0.9983$	$+1.0000$	$+0.9991$	$+0.6004$	$+0.5718$

	Extension controls		Scale	Lineage variants					
	Kandinsky -2.1	SD 2.1-unCLIP-small	SD 3.5-Large	PixArt-α	SDXL-Turbo	SD 1.4	RV	LCM	RandText
pp	0.9808	0.6842	0.9737	0.9643	0.4286	0.6842	0.6842	0.6842	0.6842
conc	0.6061	0.5674	0.7244	0.5520	0.5690	0.5890	0.5980	0.5870	0.6020
Gini	0.5908	0.5906	0.5986	0.4545	0.5703	0.5913	0.5978	0.5894	0.6003
ρ	$+0.6012$	$+0.13$	$+0.839$	$+0.99^{\dagger}$	—	—	—	—	—

Peak layers. SD 1.5 and SD 2.1 peak at `up_blocks.2.resnets.0` ($L = 38$), SDXL at `mid_block.resnets.1` (28), HunyuanDiT at blocks.20 (40), FLUX at `single_37` (57), SD 3.5 at block 23 (24), PixArt-Σ at T27 (28), Lumina-Next at layers.23 (24), Kandinsky-2.2, Kandinsky-2.1, and Karlo at `up_blocks.3.resnets.3` (52), and SD 2.1-unCLIP-small at `up_blocks.2.resnets.0` (38). The integer in parentheses is that backbone's hooked-layer count L, so a printed layer name maps onto the shared l/L axis as pp $\times L$. For the ramp instances the peak is the final layer by construction, so the pp column is a degenerate argmax and carries no information about them.

The table collects the accent statistics of the full registry together with the boundary instances, the extension controls, and the lineage variants. For the localized backbones the ρ column is weak or negative, which is expected rather than problematic: an interior-peak profile is non-monotonic, so $\rho \approx 0$ is a covariate of the class decision, not a failure of it. The four ramp members give $\rho \geq +0.9872$, with FLUX the minimum, all above the decision threshold $\tau = 0.90$. In the lineage-variant columns only the PixArt-α ρ is recorded from the existing summary; the remaining variants share the peak position of SD 1.5 / SDXL, so their ρ values are not listed separately.

The boundary instances (Kandinsky-2.2 and 2.1, Karlo) have a final-layer argmax but $\rho < \tau$, so their rule label is *other* and their morphological name is the terminal-spike boundary morphology (§2.3); SD 2.1-unCLIP-small is the unCLIP conditioning control and lands in the localized class

($D_{\mathrm{pp}} = 0$ against SD 1.5). The monotonicity bound is calibrated on Medium-tier models: the 8.1B SD 3.5-Large still peaks on the final layer but reads $+0.839 < \tau$, so under the same frozen rule that tier is *other*, with mass compressed into the final third (Appendix C.3). The terminal topology is therefore scale-invariant, while the ramp label is not. Among the extended registry instances, Z-Image (S3-DiT, single-stream, pp 0.9667, $\rho + 0.9893$) and Sana-1600M (linear-attention DiT, pp 0.9500, $\rho + 1.0000$) are the fifth and sixth ramp members.

C.2 Robustness overview

The two dominant morphologies are robust in peak position and terminal topology along eight orthogonal dimensions (Table C.2). The discrete rule label is less stable than the profile it summarizes: the axis of model scale moves the label of the largest tier, whose monotonicity reading falls below the frozen threshold while its peak position and terminal topology are unchanged, as §2.3 and Appendix C.3 discuss. Among the perturbations scanned, the only one that displaces the peak itself is $\eta = 1$-level random noise injection, whose shift is consistent across protocols (Appendix C.5). Every row is measured under the P-multi readout; agreement between P-multi and the single-step protocols is not claimed here (Appendix A.2).

Table C.2: **Robustness overview: eight orthogonal dimensions and the single failure boundary.** ✓marks that peak position and terminal topology survive the perturbation, × that they do not; a perturbation that keeps the topology but changes the rule label says so in the reading column. Details for each dimension are in Appendix B.2 and this appendix.

Dimension	Perturbation	Result	Key reading
Lineage	checkpoint / LoRA / distillation / fine-tuning / text-encoder replacement / release pair	✓	$D_{\mathrm{pp}} = 0$, $D_s \leq 0.0232$ (SD 1.5 family) and ≤ 0.0963 (loosest, PixArt α–Σ); SD 2.1 same peak 0.0099
Sampler	DPM++ / Euler / ancestral Euler a × {20, 50} steps	✓	peak position unchanged in all cases
Data split	20 random half-splits (4 architectures)	✓	Spearman ≥ 0.9987, peak layer identical 20/20
Numerical precision	fp16 vs. bf16	✓	Pearson $r = 0.9982$
Noise level	per σ (FLUX, 3 points)	✓	peak position invariant per σ
Model scale	SD 3.5 2.5B→ 8.1B (×3.2)	✓ / ×	peak position and terminal topology unchanged; ramp label not retained at 8.1B ($\rho = +0.839 < \tau$)
Reference-noise seed	reference ε with new RNG seeds, 4 representative models	✓	peak locked $4/4$, $\max \Delta\rho = 0.0139 < 0.02$
Inversion framework	edit-friendly DDPM / Direct Inversion / AIDI, 104 images, 2 models	✓	peak unmoved, amplitude only −10–19% (Appendix C.10)
Randomness strength	$\eta = 1$ random DDIM	×	peak shift $0.684 \rightarrow 0.895$ ($D_{\mathrm{pp}} = 0.211$)

Reference-noise seed: the single random source of Definition 1. The canonical protocol draws reference features from a single RNG seed 42, so peak position, ρ, and morphology labels all inherit one and the same draw, which is the most directly exposed surface of the measurement protocol. This check re-runs the reference leg only: inversion and reconstruction trajectories and the features on the reconstruction side are reused per seed rather than recomputed, and $\varepsilon_{\mathrm{ref}}$ is drawn with 5 seeds per image from the set $\{42, 0, 1, 2, 3\}$, at a cost of $1.31\times$ the single-seed baseline of a full independent re-run. The four representative models cover both dominant morphologies and three families: SD 1.5 and H-DiT from the localized class and the UNet and DiT families, and PixArt-Σ and SD 3.5 from the ramp class and the DiT and MM-DiT families. The acceptance criterion is pre-registered as peak locking, meaning an identical argmax index across the 5 seeds, and a ρ range below 0.02.

All four models pass. The peak locks $4/4$, with D_{pp} across seeds identically 0.0000: SD 1.5 is locked at 0.684, H-DiT at 0.500, PixArt-Σ at 0.964, and SD 3.5 at 0.958. The largest ρ range is 0.0139 for H-DiT, while PixArt-Σ and SD 3.5 give 0.0000. The morphology label changes on none

of the 20 seed legs. Seed sensitivity of the shape coordinates is below the noise floor: the worst D_s across seeds is 0.0017 for SD 1.5, all 16 ref-seed cells, covering 4 models and 4 alternative seeds, fall inside the equivalence margin and are TOST-equivalent, with values for each cell in the confound matrix of Appendix C.7, and the worst bootstrap CI95 upper bound is 0.0023, below $p_{95} = 0.0042$. At the per-image level, the argmax over the 104 images and 4 alternative seeds agrees with seed 42 in 1.000 of cases, $416/416$ per model and $1664/1664$ summed over the four models, with relative profile deviations of 0.0029–0.0187, largest for SD 1.5. As a validity check, the seed-42 leg against the canonical JSON gives Spearman 0.9996–1.0000, identical argmax everywhere, and scale ratios 0.9936–1.0000. The 0.9936 of H-DiT is a global scale factor that does not enter any feature under min–max normalization.

C.3 SCALE-AXIS TEST

The within-family control is the pair of SD 3.5 tiers: Medium, with 2.5B parameters and 24 joint blocks, and Large, with 8.1B parameters and 38 joint blocks, share the MM-DiT topology, the same set of text encoders, and the same protocol, with the $\times 3.2$ scale difference as the only difference (Table C.3).

Table C.3: **Scale-axis control: SD 3.5 Medium vs. Large (P-multi@104, same family, same protocol, empty prompt).**

Metric	Medium (2.5B, 24 layers)	Large (8.1B, 38 layers)	Δ
pp	0.9583 (block 23)	0.9737 (block 37)	$+0.0154$ ($= 1/24 - 1/38$)
conc	0.6942	0.7244	$+0.0302$
sp (Gini)	0.6077	0.5986	-0.0091
Spearman (layer order, profile)	$+0.9983$	$+0.839$	scale-dependent bound

Three readings follow. First, the terminal topology and the peak position are strictly invariant across the $\times 3.2$ scale: both peaks sit on the final layer, block 23 for Medium and block 37 for Large, and Δpp matches the $1/L$ quantization difference exactly, $1/24 - 1/38 = 0.01535$, so the peak difference is fully explained by layer count quantization. The rule label is not invariant: Large's monotonicity reading falls to $\rho = +0.839 < \tau = 0.90$, so under the same frozen rule the 8.1B tier is *other* (Table C.1). Second, the shape coordinates depend weakly on scale at the lineage level, with $D_{\mathrm{mag}} = 0.0315$, composed from the conc and sp increments of the table, about $7.5\times$ the noise floor p_{95}, below the largest within-lineage variation 0.0963, an order of magnitude below the morphology-class separation, and self-consistent with the resolution framework of Appendix C.5. Third, the profile shows why the label crosses: the first $2/3$ of Large's layers form a shallow plateau with mass compressed into the final third, so the "monotone ramp" read off the Medium tier is at 8.1B a terminal-compressed ramp.

C.4 FAMILY-LABEL CONTROL

The two dominant morphologies cross-mismatch with architecture-family labels. The ramp class mixes MM-DiT (FLUX and SD 3.5) with cross-attention DiT (PixArt-Σ and Lumina-Next). The localized class mixes UNet (SD 1.5, SD 2.1, and SDXL) with cross-attention DiT (HunyuanDiT). The UNet family itself contains the terminal-spike boundary morphology, the Kandinsky family, which is a terminal spike rather than localized. Two controls make the point. First, same label but different morphology: HunyuanDiT is a DiT just as PixArt-Σ and Lumina-Next are, yet they fall into different classes, with $D_s = 0.53$ and 0.46, far above the noise floor. Second, different label but same morphology: SDXL, a UNet, and HunyuanDiT, a DiT, differ structurally yet are both localized. The extension instances add two families not present in the original registry, single-stream S3-DiT in Z-Image and linear-attention DiT in Sana, and both are terminal-ramp, so neither single-stream concatenation nor linear attention produces an exception. More sharply, Sana, a linear-attention DiT, and Lumina-Next, an MM-DiT, have a profile distance of $D_s = 0.0093$ with CI95 $[0.009, 0.023]$, TOST-equivalent (Appendix C.8), hugging the within-lineage noise floor: two families with different attention mechanisms and independent training pipelines give nearly indistinguishable drift

organization. If morphology followed from family labels, none of these controls should exist. What the morphology tracks is the organization of information pathways, not the "family lineage" of an architecture. §3.3 provides mechanism evidence on the localized class, and a pathway account of the ramp class remains a hypothesis at the level of observation.

C.5 Resolution, lineage supplements, and boundary distances

The morphology classes are resolvable. Cross-morphology pairs, meaning any localized and ramp pair, give $D_s \in [0.2679, 0.5922]$, with the minimum at SD 1.5–Sana and the maximum at SDXL–Z-Image, far above the noise floor and the equivalence margin.

The noise hypothesis, details of §2.3. If drift organization were merely unstructured noise, there would be no reason for lineage variants to share a profile. The measured within-lineage distance is ≤ 0.0232 for every checkpoint, LoRA, fine-tune, and text-encoder cell, and at most 0.0963 once the PixArt release pair is included, against ≥ 0.2679 for cross-morphology pairs, at least two orders of magnitude above the noise floor.

Small within-class distances are not guaranteed: morphology classes are products of rules. Within-class distances can reach 0.257 in the localized class, for the SD 2.1 to SDXL pair with peak positions 0.684 versus 0.429, and 0.306 for SD 3.5–FLUX in the ramp class, so within-class pairs can be far more distant than the closest cross-morphology pair. Class membership is given by the decision rule, the final-layer argmax plus the monotonicity criterion of §2.3, not by clustering in the shape coordinates. The 2D shape coordinates conc and Gini would place SD 3.5, at 0.694 and 0.608, in the neighborhood of the localized centroid, whereas the rule assigns it to the ramp class. This is consistent with the $10/10$ full-profile prediction of §2.4 and the stage-wise readout of §3.1: describing the ramp morphology requires the full profile and readout information, not just scalar coordinates.

Within-class resolvability and within-lineage unresolvability coexist. In the ramp class, FLUX–PixArt-Σ gives $D_{\mathrm{mag}} = 0.0011$, below the noise-floor median 0.0014: two different architectures are statistically indistinguishable in the shape parameters. Within a lineage, pairs separate down to 0.096, for the PixArt-α versus Σ pair. Shape parameters are properties of the instance, and resolvability is not guaranteed. Within the localized class, instances are resolved by peak position, with 0.4286, 0.5000, and 0.6842 mutually distinct and lineage-stable.

Distance profile of the boundary instances. Structural distances from Kandinsky-2.2 are as follows: only 0.0473 to SD 3.5, with a pp difference of 0.0225 and a shape difference of 0.0416, so its nearest neighbor is a ramp member, then 0.0910 to Lumina-Next, about 0.31 to FLUX, PixArt-Σ, SD 1.5, and SD 2.1, 0.485 to H-DiT, and 0.564 to SDXL. In L2 geometry it sits against the ramp class of §2.4 yet is excluded by the monotonicity criterion: the boundary status carries order information, not distance information.

Per-item values of the lineage variants, details of §2.2. In the SD 1.5 family, LCM gives 0.0045, SD 1.4 gives 0.0073, RV gives 0.0183, and RandText gives 0.0232, with the peak position locked at 0.6842 for all. SD 2.1 gives 0.0099 with the same peak. In the SDXL family, SDXL-Turbo gives $D_{\mathrm{mag}} = 0.0022$, inside the noise floor. The variation hierarchy is summarized in Table C.4. Its rows are separate comparison levels rather than one numeric scale: the untrained row reports a difference in peak position D_{pp}, not a structural distance D_s, and it is not comparable with the rows above it.

Sampler replacement and the $\eta = 1$ boundary. Sampler replacement, covering DPM++ (Lu et al., 2025), Euler, and ancestral Euler a, each at 20 and 50 steps, all leave the peak position unchanged. Random DDIM at $\eta = 1$ under the canonical protocol moves the SD 1.5 peak from 0.684 to 0.895, landing on `up_blocks.3.resnets.2`, with $D_{\mathrm{pp}} = 0.211$, conc moving $0.583 \to 0.713$, Gini $0.588 \to 0.699$, and $D_s = 0.271$. The shift is protocol-independent: under P-t0 the shift is exactly consistent at $D_{\mathrm{pp}} = 0.2105$, and the landing layer is the same under both protocols. The failure mode is the peak being deterministically pushed to a fixed position, not the disintegration of the profile.

Splits and blind classification. Under 20 random half-splits with seed 42 on 4 architectures, the profiles correlate at Spearman ≥ 0.9987, the peak layer is identical in 20/20 cases, and $D_s \leq 0.0062$. Nearest-centroid classification assigns every variant to the correct class at the prediction level.

Table C.4: **Variation hierarchy (P-multi@104).** Within-lineage, between-morphology, boundary, and untrained comparison levels; per-cell values are in the text above, and boundary instances are excluded from the between-morphology distances.

Level	Distance scale	Instances
Within lineage	$D_s = D_{\mathrm{mag}} \in [0.0022, 0.0963]$	SD 1.5 family: LCM 0.0045 / SD1.4 0.0073 / RV 0.0183 / RandText 0.0232; SD 2.1 same peak 0.0099; SDXL family: SDXL-Turbo 0.0022 (inside the noise floor); loosest cell PixArt α–Σ 0.0963 (release pair)
Between morphology classes	$D_s \in [0.2679, 0.5922]$	any localized × ramp pair (min SD 1.5–Sana; max SDXL–Z-Image)
Boundary instances	0.047–0.309 to the ramp class	Kandinsky-2.2 nearest neighbor SD 3.5 (0.0473)
Untrained	$D_{\mathrm{pp}} = 0.658$	random initialization vs. after training

C.6 Terminal stage-wise readout: pre-registered decisions and per-stage values

The stage-wise readout and the dual metrics are defined in §2.1, and the three decisions were fixed before the data. Terminal block-localized peak means that all three post-block stages fall below the final block by more than 10^{-4}. Post-block relative accumulation means that the relative drift at no fewer than two post-block stages is not below the final block. Mixed means an absolute ramp with a within-block non-monotonic relative profile, and no instance of it appears in this data set. Figure C.1 shows the ladder for the three models: the horizontal axis runs $\texttt{block}_{N-2} \to \texttt{block}_{N-1}$ (final)$\to$ `norm_out` $\to$ `proj_out` $\to$ `velocity` $\to$ `latent` $\to$ `pixel`, with the post-block stages shaded, the absolute L2 drift on a log axis above and the relative L2 drift on a linear axis below, the relative probe stopping at velocity. Per-stage values follow, covering all 104 images with 0 failures:

Table C.5: **Terminal stage-wise readout: per-stage values (P-multi@104).** abs is absolute L2 drift and rel is relative L2 drift, per stage of the readout ladder. Per-model readings and the three decisions are in the text below.

stage	SD 3.5		PixArt-Σ		FLUX	
	abs	rel	abs	rel	abs	rel
$\texttt{block}_{N-2}$	96719	2.902	8074	0.457	4.58×10^5	1.786
$\texttt{block}_{N-1}$ (final)	115851	2.303	8319	0.397	4.85×10^5	1.525
norm_out	1286	1.330	863	0.396	896	1.386
proj_out	608	0.922	330	1.217	484	1.788
velocity	608	0.922	294	1.288	484	1.788
latent (pre-VAE, K mean)	—	1.41	—	—	—	1.99
pixel (decoded at $t = 0$)	—	0.090	—	1.30	—	0.115

Per-model readings and decisions. SD 3.5: abs collapses by 99.5% after the final layer and rel declines monotonically in step, with per-image argmax block 23 under the absolute metric (100%) and block 22 under the relative metric (100%); decision, terminal block-localized peak. PixArt-Σ: abs collapses by 96.5% and rel overtakes the final layer at proj_out/velocity (pointwise at all 9 K indices); decision, post-block relative accumulation. FLUX: abs collapses by 99.9% (decline $> 95\%$ on 104/104 images) and rel overtakes (1.79 vs. 1.52, holding on 87% of images), with the within-block rel peak on the penultimate layer; decision, post-block relative accumulation. Five cases are resolved, one terminal block-localized peak plus four post-block relative accumulations, two of which are the boundary instances of Table C.6. Lumina-Next and Kandinsky-2.1 were not probed, so their readout subtype is unresolved.

The latent and pixel rows report relative values only, since they are not directly comparable across scales, and the absolute magnitudes of FLUX exceed those of the other two models because its final-layer features have a larger scale, so collapse ratios are compared within a model column. The per-K

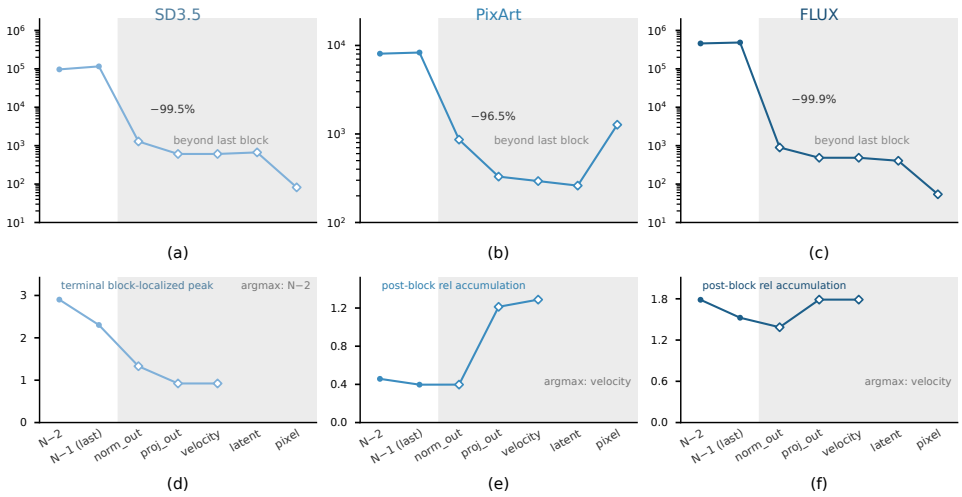

Figure C.1: **Terminal stage-wise readout: terminal block-localized peak vs. post-block relative accumulation.** Columns are SD 3.5, PixArt-Σ, and FLUX (P-multi@104). Decisions and per-stage values are in Table C.5.

decomposition (Figure C.2) shows the orderings hold across the noise range: at SD 3.5, block$_{23}$ rel exceeds norm_out rel at 7 of the 9 K indices (the exceptions are the two highest-noise readouts, $K = 48, 49$, where block rel falls below norm_out while staying above proj/vel at all 9), and norm_out rel exceeds proj/vel rel at all 9; at all 9 K indices of PixArt-Σ, velocity rel (1.17–1.38) exceeds T27 rel (0.21–0.72), so the overtaking is not a low-K artifact; for FLUX the overtaking is established by the full-pool mean and the per-image 87%, without a per-K decomposition. The within-block relative profiles separate the three models: the within-block rel Spearman is ≈ 0 for SD 3.5, -0.55 for PixArt-Σ with a rel peak at block 6, and -0.51 for FLUX with a rel peak at the single block 36. The absolute slope contains a norm-inflation component whose quantitative decomposition and norm-free re-judgment are given in Appendix C.9, while the morphology classification and D_s use the canonical absolute metric (§2.1, §2.3).

Boundary instances. Both boundary instances are post-block relative accumulation, with norm_out falling and proj_out and latent overtaking, and both terminal absolute spikes contain a marked norm-driven component (abs collapse ratio ≈ 0.0042, a 99.6% collapse). They differ in where the post-block stages end: Kandinsky-2.2 tops out at proj_out (rel argmax 100% proj_out), while Karlo extends to latent (90% latent, 10% proj_out).

Table C.6: **Terminal stage-wise readout of the boundary morphology (P-multi@104).** Relative L2 drift at each post-block stage over the final block, so > 1 means overtaking, for the two boundary instances. Both read as post-block relative accumulation; details in the text.

stage / metric	Kandinsky-2.2	Karlo
norm_out (rel / final)	0.506×	0.511×
proj_out (rel / final)	3.22×	2.08×
latent (rel / final)	1.50×	4.31×
abs collapse ratio (after final / final)	0.004224	0.004240
rel argmax (per image)	100% proj_out	90% latent / 10% proj_out

C.7 The confound matrix: full table of 19 pre-registered cells and 4 reference-noise-seed extension cells

Each cell re-measures the perturbed side under the canonical protocol and compares it with the baseline, SD 1.5 or PixArt-Σ, and all rows use the canonical 104-image pool; the fine-tuned pair carrying the † marker is measured on a 37-image subset, and the split-half rows compare paired

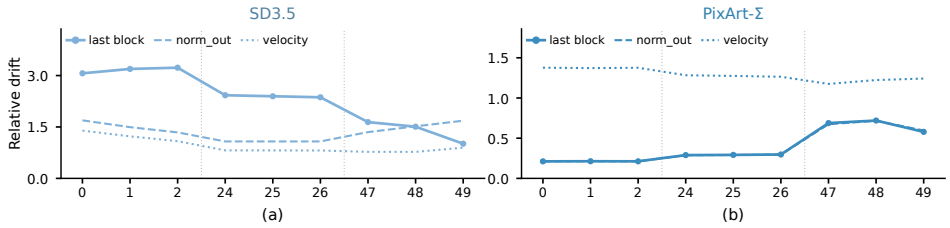

Figure C.2: **Per-K decomposition: the subtype decisions are not averaging artifacts (6-image probe).** The terminal stage-wise readout split by the 9 readout steps of 50-step inversion: relative drift at the final block (solid circles), norm_out (dashed), and velocity (dotted). The per-step orderings are stated in the text.

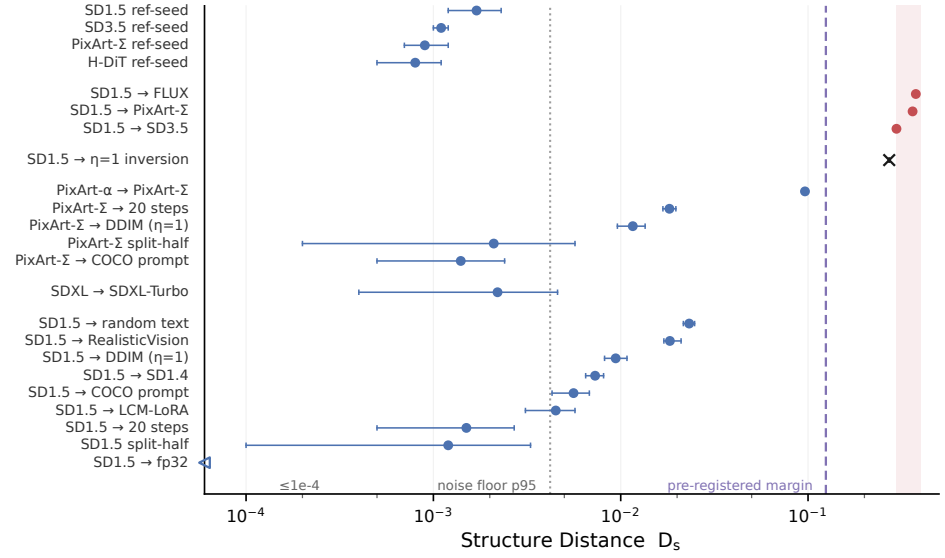

Figure C.3: **Confound-matrix forest plot of 23 cells (19 pre-registered cells plus 4 reference-noise-seed cells, canonical P-multi@104).** The purple dashed line is the margin 0.1245. Values for each cell are in Table C.7.

half-pools. D_s is based on the absolute L2 min–max profile, CI95 is an image-resampling bootstrap with $B = 1000$, and TOST is one-sided equivalence testing with H_0 given by $D_s \geq 0.1245$, where $p < 0.001$ records equivalence and $p = 1.000$ rejects it. The two η=1 rows differ in kind: DDIM sampling with η=1 is a protocol perturbation and equivalent, whereas the η=1 variance-preserving inversion is the positive control and breaks the margin. The registry-extension pair SD 1.5–Sana ($D_s = 0.2679$, Appendix C.5) is computed outside the matrix under the same canonical protocol.

The table also lists the reference-noise-seed cells of the Appendix C.2 design: baseline and perturbed side are the same architecture, the same images, and the same trajectory, differing only in the RNG seed of the reference ε. For each of the four models, the worst value over the 4 alternative seeds $0/1/2/3$ against seed 42 is reported. Since both sides share the image-resampling indices, the bootstrap is paired. All 16 cells of the four models lock the peak position at $D_{\text{pp}} = 0.0000$ and are judged equivalent; the values for each seed and the pre-registered acceptance are recorded in Appendix C.2. The forest plot in Figure C.3 shows all 23 cells, the 19 pre-registered confound cells plus the 4 reference-noise-seed cells, one cell per row with the point at D_s and the whiskers at the image-resampling CI95. Cells that sit at exactly zero are censored at the bottom axis line and annotated "$\leq 10^{-4}$"; blue marks an equivalent cell, red a separated one, black a break, and the purple dashed line is the margin. Per-cell values are in Table C.7.

Table C.7: **Full confound matrix: 19 pre-registered cells and 4 reference-noise-seed extension cells (canonical P-multi@104).** $D_s = \sqrt{D_{\text{pp}}^2 + D_{\text{mag}}^2}$; CI95 is an image-resampling bootstrap ($B = 1000$). Verdicts: equivalent (inside the margin 0.1245), separated, or break (the η=1 positive control).

Cell	D_s	95% CI	D_{pp}	D_{mag}	Peak pp	TOST p	n	Verdict
SD1.5 → fp32	0.0000	[0.0000, 0.0001]	0.0000	0.0000	0.684 → 0.684	< 0.001	104	Equivalent
SD1.5 split-half	0.0012	[0.0001, 0.0033]	–	–	–	< 0.001	–	Equivalent
SD1.5 → 20 steps	0.0015	[0.0005, 0.0027]	0.0000	0.0015	0.684 → 0.684	< 0.001	104	Equivalent
SD1.5 → LCM-LoRA	0.0045	[0.0031, 0.0057]	0.0000	0.0045	0.684 → 0.684	< 0.001	104	Equivalent
SD1.5 → COCO prompt	0.0056	[0.0043, 0.0068]	0.0000	0.0056	0.684 → 0.684	< 0.001	104	Equivalent
SD1.5 → SD1.4	0.0073	[0.0065, 0.0081]	0.0000	0.0073	0.684 → 0.684	< 0.001	104	Equivalent
SD1.5 → DDIM (η=1)	0.0094	[0.0082, 0.0108]	0.0000	0.0094	0.684 → 0.684	< 0.001	104	Equivalent
SD1.5 → RealisticVision	0.0183	[0.0170, 0.0210]	0.0000	0.0183	0.684 → 0.684	< 0.001	37†	Equivalent
SD1.5 → random text	0.0232	[0.0216, 0.0248]	0.0000	0.0232	0.684 → 0.684	< 0.001	104	Equivalent
SDXL → SDXL-Turbo	0.0022	[0.0004, 0.0046]	0.0000	0.0022	0.429 → 0.429	< 0.001	104	Equivalent
PixArt-Σ → COCO prompt	0.0014	[0.0005, 0.0024]	0.0000	0.0014	0.964 → 0.964	< 0.001	104	Equivalent
PixArt-Σ split-half	0.0021	[0.0002, 0.0057]	–	–	–	< 0.001	–	Equivalent
PixArt-Σ → DDIM (η=1)	0.0116	[0.0096, 0.0135]	0.0000	0.0116	0.964 → 0.964	< 0.001	104	Equivalent
PixArt-Σ → 20 steps	0.0182	[0.0168, 0.0197]	0.0000	0.0182	0.964 → 0.964	< 0.001	104	Equivalent
PixArt-α → PixArt-Σ	0.0963	[0.0938, 0.0989]	0.0000	0.0963	0.964 → 0.964	< 0.001	104	Equivalent
SD1.5 → η=1 inversion	0.2712	[0.2701, 0.2723]	0.2105	0.1710	0.684 → 0.895	1.000	104	Break
SD1.5 → SD3.5	0.2965	[0.2957, 0.2974]	0.2741	0.1130	0.684 → 0.958	1.000	104	Separated
SD1.5 → PixArt-Σ	0.3615	[0.3597, 0.3632]	0.2801	0.2285	0.684 → 0.964	1.000	104	Separated
SD1.5 → FLUX	0.3762	[0.3745, 0.3780]	0.2982	0.2294	0.684 → 0.982	1.000	104	Separated
SD1.5 ref-seed	0.0017	[0.0012, 0.0023]	0.0000	0.0017	0.684 → 0.684	< 0.001	104	Equivalent
PixArt-Σ ref-seed	0.0009	[0.0007, 0.0012]	0.0000	0.0009	0.964 → 0.964	< 0.001	104	Equivalent
H-DiT ref-seed	0.0008	[0.0005, 0.0011]	0.0000	0.0008	0.500 → 0.500	< 0.001	104	Equivalent
SD3.5 ref-seed	0.0011	[0.0010, 0.0012]	0.0000	0.0011	0.958 → 0.958	< 0.001	104	Equivalent

C.8 Full statistics of leave-one-out prediction and the OOD detector (details of §2.4)

Statistical implementation note. The pre-registration fixed the classifier, the feature spaces, and the permutation design. The computation of p values was changed after the freeze from a Monte Carlo estimate with $B = 1000$ to exact enumeration, with the null hypothesis and test statistic unchanged. After the registry extension, in which Z-Image and Sana were admitted by the frozen rule, all derived statistics in this subsection were recomputed on the ten-instance registry under the same procedure, and the enumeration space changed from $\binom{8}{4} = 70$ to $\binom{10}{4} = 210$.

The null space of the exact permutation. The observed $10/10$ sits at the ceiling of the permutation null: among all $\binom{10}{4} = 210$ fixed-margin label assignments under the unbalanced $4/6$ margins, $k = 1$ achieves the observed accuracy, so $p = 1/210 \approx 0.00476$. The complement assignment of a global label swap carries six interior labels and is not in the enumeration space, so the floor is $1/210$ rather than $2/210$. The original 8-instance registry had $\binom{8}{4} = 70$ and $p = 2/70$, its complement assignment being inside the space. The bound $p \geq 1/210$ is the combinatorial design floor of the assignment, independent of any Monte Carlo budget: increasing B cannot lower it and only estimates the same exact quantity more precisely.

Separation margins. Beyond the saturated accuracy endpoint, the effect size is carried by the separation margins. Let $m_i = d_{\text{other}} - d_{\text{self}}$ be the signed distance difference of instance i to the other-class and own-class centroids, where positive means correct classification. The worst margins of the observed assignment are $+0.95$, $+1.15$, and $+0.88$ for the 64-pt, ortho-full, and ortho-shape spaces, exceeding the worst margin of the best of the 209 nontrivial competing assignments by $+1.45$, $+2.08$, and $+1.76$. Every other label assignment puts at least one instance on the wrong side, and

the gap has widened further relative to the original registry. The nearest neighbors of the new instances are all of the same class: Z-Image↔PixArt-Σ gives $D_s = 0.0868$ with CI95 $[0.081, 0.092]$, and Sana↔Lumina-Next gives $D_s = 0.0093$ with CI95 $[0.009, 0.023]$, TOST-equivalent. Under a paired bootstrap over images with 103 shared-stem images resampled jointly and $B = 1000$, the $10/10$ score reproduces in $1000/1000$ runs. The exact binomial 95% CI at $n = 10$, $[0.69, 1.00]$, is a registry-scale quantity and is not used for population inference. Margins for each instance of the original 8-instance registry are in Table C.8.

Table C.8: **Per-instance separation margins (P-multi@104, original 8-instance registry).** Margins $m_i = d_{\text{other}} - d_{\text{self}}$ per feature space, with centroids and z-normalization computed from the training folds only; the worst and mean margins are the last two rows. Per-instance values for the extended 10-instance registry will be released with the code.

Instance	64-pt profile	ortho-full	ortho-shape
SD 1.5	1.121	1.179	0.918
SD 2.1	1.203	1.667	1.347
SDXL	0.948	2.986	2.615
H-DiT	1.334	2.572	2.158
FLUX	1.482	1.019	0.888
SD 3.5	1.904	0.966	0.476
PixArt-Σ	1.926	1.212	1.073
Lumina-Next	2.156	2.353	1.874
Worst margin $\min_i m_i$	0.948	0.966	0.476
Mean margin	1.509	1.744	1.419

Family-label baseline. The baseline predicts the majority morphology within the family group, with ties counted as errors. Under the extended registry the baseline is at most $9/10$: H-DiT is the one necessarily wrong case, since its interior morphology is outvoted by the ramp majority of the DiT group. DiT-group ties, 1:1 within the group when PixArt-Σ or Lumina-Next is held out, and single-member families, S3-DiT and linear-attention DiT, can only lower the baseline through the fallback rule for held-out folds, giving $7/10$ when ties count as errors. The original 8-instance registry gives a baseline of $7/8$, or $6/8$ under the alternative convention that assigns Lumina-Next to MM-DiT. Every convention is below the $10/10$ of the three accent feature spaces.

Per-space behavior of the distance-based OOD detector. The frozen rule is "distance to the nearest class centroid $>$ the largest within-class held-out distance $\Rightarrow$ judged *other*". Under the extended registry of 10 instances, the profile space rejects the two cross-lineage boundary instances: Kandinsky-2.2 sits at 2.2027 from the ramp centroid, above the threshold 2.0649 (it gave 1.987, below the threshold, under the original registry), and Karlo remains rejected. The ortho-full and ortho-shape spaces do not trigger for any boundary instance, and the nearest class of the three instances is always terminal-ramp. Boundary decisions are given by the monotonicity rule, which depends on no other instance, as §2.3 states.

C.9 Decomposing the norm schedule: what the absolute profile reads out (details of §2.1)

This subsection answers the quantitative question raised in §2.1: to what extent is the morphological organization under the canonical absolute metric carried by the per-layer feature-norm schedule. Three groups of measurements follow: norm-free re-judgment, rank equivalence and ramp decomposition, and increment accounting. These are conclusions about measurement and do not constitute causal claims. All numbers are extended evidence measured after the freeze, with the freeze record in the Experimental setup subsection of Appendix A.2.

(a) A correction of the measurement metric. The early norm probe took the raw tensor norm of the hook, inconsistent with the token-mean pooling used by the probe family for 3D outputs. SD 1.5 attention layers showed direct-to-implied ratio deviations of 100–150×, while resnet layers are 4D tensors and were unaffected. After the pooling was made consistent and the measurements re-run, the old values of the 22 resnet layers reproduce to 3.3×10^{-7}, and the direct-to-implied ratio returns

to 0.95–1.19 for all models. All numbers in this subsection use the corrected metric. The correction affects decomposition-type numbers only and none of the readings already reported in the main text.

(b) Norm-free re-judgment. For all dominant instances, the §2.3 morphology rule is re-run under three metrics: canonical absolute L2 (ABS), relative L2 (rel), and cosine distance (cos). Results are in Table C.9. The three localized instances keep their interior label under all three metrics. The six ramp instances all lose the terminal-ramp label under the rel metric: the image-level bootstrap reproduces $1000/1000$, and the rel profiles peak in the interior. The flip of Z-Image is dominated by a small-denominator artifact (the rel peak falls on layers.0, whose norm percentile is 0), while Sana is a genuine flip (rel peak at blocks.18, whose norm ratio is 2.41). Under the cos metric, the three original ramp members measured under it likewise all lose the label. One boundary must be stated plainly: the exact rel argmax position of the localized class is not stable within a lineage (SD 1.5 at 0.132 versus SD 2.1 at 0.737, a difference of 0.6 within one family). The "persistence" under norm-free metrics is therefore persistence at the level of non-final-layer argmax, and it does not extend to the exact peak position. A fourth setting re-normalizes the canonical profile by element count instead, dividing each layer by $\sqrt{\text{numel}}$ on the stored profiles with no model forward, and re-applies the frozen rule. Peak layer, peak position, and morphology label are unchanged in all nine instances. The token-pooled families are unchanged to machine precision (their token count is uniform across layers, so the rescale is a global scale that the profile normalization removes), while the UNet instances move only in the concentration statistics: ρ shifts by -0.091, -0.114, and -0.030 for SD 1.5, SD 2.1, and SDXL, with SD 2.1's ρ crossing zero, and the concentration by at most 0.072. The label rule is therefore insensitive to element-count normalization, while the reported magnitudes are not.

Table C.9: **Re-judgment of the morphology rule under three metrics (P-multi@104).** A cell gives the §2.3 rule decision under that metric; parentheses give the normalized argmax position and the profile-monotonicity Spearman under that metric (resolved values only). ABS is the canonical absolute L2 metric, REL and COS are norm-free metrics, and dashes mark unmeasured cells. The per-model readings are discussed in the text above.

Model	ABS (canonical)	REL	COS
SD 1.5	interior (pp .684, $\rho + .11$)	interior (.132)	interior (.132)
SD 2.1	interior (.684, $\rho + .08$)	interior (.737)	interior (.737)
SDXL	interior (.429, $\rho + .07$)	—	—
HunyuanDiT	interior (.500, $\rho - .37$)	interior (.575)	interior (.575)
FLUX	ramp (.982, $\rho + .987$)	interior (.965, $\rho - .514$)	—
SD 3.5	ramp (.958, $\rho + .998$)	interior (.917, $\rho + .003$)	other ($\rho + .214$)
PixArt-Σ	ramp (.964, $\rho + 1.000$)	interior (.214, $\rho - .551$)	interior (.214, $\rho - .542$)
Lumina-Next	ramp (.958, $\rho + .999$)	interior (.000, $\rho - .486$)	interior (.042, $\rho - .758$)
Z-Image	ramp (.967, $\rho + .989$)	interior (layers.$0^{\dagger}$)	—
Sana-1600M	ramp (.950, $\rho + 1.000$)	interior (blocks.18)	—

(c) Rank equivalence and ramp decomposition. We compute the Spearman correlation between the absolute drift profile and the feature-norm schedule (per-layer $\|f_l^{\text{ref}}\|_2$, pooled consistently with the probe): SD 3.5 gives $+0.989$, PixArt-Σ $+0.929$, Lumina-Next $+1.000$, SD 1.5 $+0.984$, SD 2.1 $+0.986$, and H-DiT $+0.946$. Over these six models $\min \rho \geq +0.929$, and FLUX and SDXL are not included. Peak layers coincide $6/6$, and the shape label of the norm schedule itself agrees with the drift morphology label $6/6$. Under the ramp metric, the log decomposition $\log d_l = \log \|f_l\|_2 + \log \text{rel}_l$ holds exactly, where $\text{rel}_l := d_l / \|f_l\|_2$ is the aggregate relative drift of the K-averaged per-image profile. The decomposition is taken after the time average, so the identity is exact by construction, whereas the per-timestep identity $d_{l,t} = \|f_{l,t}\|_2 \, \text{rel}_{l,t}$ would not survive averaging in expectation. Since OLS is linear in the regressand, the slope against the centered layer index $l_c = l - \bar{l}$ is additive per image,

$$s(X) = \frac{\sum_l \log X_l \cdot l_c}{\sum_l l_c^2},$$
$$s(d) = s(\|f\|_2) + s(\text{rel}),$$

over the full profile at block level, with no layer sub-segment chosen. The norm share is therefore the exact ratio $\text{share}_{\text{norm}} = s(\|f\|_2)/s(d)$, reported as the median over images. It gives norm-term contributions of 95% for SD 3.5, 146% for PixArt-Σ, and 129% for Lumina-Next, with the extension instances Z-Image at 104% and Sana at 113% falling in the same interval. The relative term is negative for PixArt-Σ and Lumina-Next: the absolute ramp is entirely contributed by norm growth, and more than entirely.

(d) Lineage stability of the norm schedule itself. Comparing the norm profiles of SD 1.5↔SD 2.1 gives $D_{\text{pp}} = 0$, $D_{\text{mag}} = 0.0260$, $\rho = +0.967$, and an identical peak layer, `up_blocks.2.resnets.0`. This is the same order of magnitude as the within-lineage distance of the drift profiles ($D_s = 0.0101$, $\rho = +0.971$), both far below the equivalence margin 0.1245. "The norm schedule is a stable architectural property" therefore holds independently: the object that the absolute drift profile reads out stands on its own.

(e) Increment accounting, with n=104 coinciding pointwise with the n=12 pilot, so not underpowered. The cumulative behavior of per-layer drift increments is compared against a random-walk null. On the constant-shape segment of the hooked chain, write $\Delta_l = f_l^{\text{ref}} - f_l^{\text{recon}}$ for the drift vector at layer l and $\delta_l = \Delta_l - \Delta_{l-1}$ with $\delta_0 := \Delta_0$. The increments telescope exactly, $\Delta_l = \sum_{j \le l} \delta_j$, verified to $\le 10^{-10}$ relative on all three models, one line per image and K. For adjacent layers with a matched feature dimension and coordinate correspondence, the squared drift energy also obeys the exact recursion

$$\|\Delta_l\|_2^2 - \|\Delta_{l-1}\|_2^2 = \underbrace{\|\delta_l\|_2^2}_{\text{new increment energy}} + \underbrace{2\langle \Delta_{l-1}, \delta_l \rangle}_{\text{reinforcement or cancellation}} .$$

The first term is nonnegative, and the signed inner-product term shows whether the new increment reinforces the accumulated discrepancy or cancels it. This is a directly checkable accounting identity on the paired vectors, not a causal decomposition. In SD 1.5, 23 adjacent transitions have compatible pooled shapes. After averaging over K within image, the cross-term is negative for all 104 images at every such transition. The mean energy change is still positive at 9 of these 23 transitions, where the positive increment energy exceeds the cancellation term, and it is negative at the other 14. Both terms are computed for each image and timestep before averaging over K within image and then across images. The identity is evaluated only within contiguous hooked segments where pooled feature vectors share a coordinate dimension and the network defines a corresponding channel/token ordering. Scalar layer norms alone do not determine its inner product. With $m_l = \|\Delta_l\|$, $g_l = \|\delta_l\|$ and windows $q = \lceil L/4 \rceil$, ALL $= (0, L{-}1)$, QT $= (L{-}q, L{-}1)$, POST $= (\text{peak}{+}1, L{-}1)$, the five endpoints are

$$\begin{aligned}
E_1 &= S(\text{QT}), \\
E_2 &= \overline{\textstyle\sum_{l \in \text{QT},\, l>0}} \cos(\delta_l, \Delta_{l-1}), \\
E_3 &= \overline{\textstyle\sum_{l \in \text{QT},\, l<L-1}} \cos(\delta_l, \delta_{l+1}), \\
E_4 &= S(\text{ALL}), \\
E_5 &= \overline{\textstyle\sum_{l \in \text{POST}}} \cos(\delta_l, \Delta_{l-1}),
\end{aligned}$$

where the segment straightness is $S(a,b) = \|\Delta_b - \Delta_{a-1}\| / \sum_{l=a}^{b} g_l \in [0,1]$ with $\Delta_{-1} := 0$, and the per-image value is its mean over K before any averaging over images. E_1 and E_4 are straightness ratios, and the matched-step random-direction baseline follows from $\mathbb{E}\| \sum_j g_j u_j \|_2^2 = \sum_j g_j^2$ for independent, zero-mean random unit directions u_j. Thus $\sqrt{\sum_{l \in W} g_l^2} / \sum_{l \in W} g_l$ for window W is the normalized root-mean-square (RMS) endpoint displacement, not the mean endpoint displacement, and it reduces to $1/\sqrt{n}$ only for equal-length steps. The two ramp instances previously tested with this increment analysis lie near that null: terminal displacements of $0.95\times$ (PixArt-Σ) and $1.08\times$ (SD 3.5), with increments nearly orthogonal to the drift accumulated so far ($E_2 = -0.013$, $+0.028$). This is incoherent accumulation, with no coherent terminal direction. Over the whole chain the same reading holds (E_4 at $0.97\times$ and $1.06\times$ of the null), whereas the localized class (H-DiT) collapses its net displacement to $0.37\times$ the null on the terminal quartile and $0.05\times$ over the whole chain ($E_2 = -0.702$, $E_3 = -0.510$). This is coherent cancellation. Class-contrast distances (image-level bootstrap $B = 1000$, Cohen's d) are 8.7, 41.6, 61.5, and 9.9 on E_1–E_4, with narrow CI95 throughout. E_5 is defined only when the peak is not the final layer (POST nonempty), which

here means the interior-localized class only. It is reported per model for H-DiT at -0.740, the same cancellation signature, facing the post-peak segment rather than the terminal quartile. No class contrast for it exists by construction, since the terminal-peak models have no POST window. The Z-Image extension has mean whole-chain straightness 0.429 against its matched-step RMS baseline 0.238 over the 30-layer shape-compatible chain, and its mean energy change is positive at 24 of 29 adjacent transitions. At full-chain scale this instance therefore shows more net alignment than the random-direction baseline. This is a descriptive measurement, not a mechanism claim. Figure C.4 shows the per-layer energy decomposition behind these endpoints, with one column per model: the mean squared drift energy $\mathbb{E}\|\Delta_l\|_2^2$ per layer with its bootstrap CI95 over images and the terminal quarter shaded, then the exact split of that energy into the nonnegative increment term and the signed reinforcement term, the two closing the first panel by construction. The spanning panel gives R_W for the full chain and the terminal quarter.

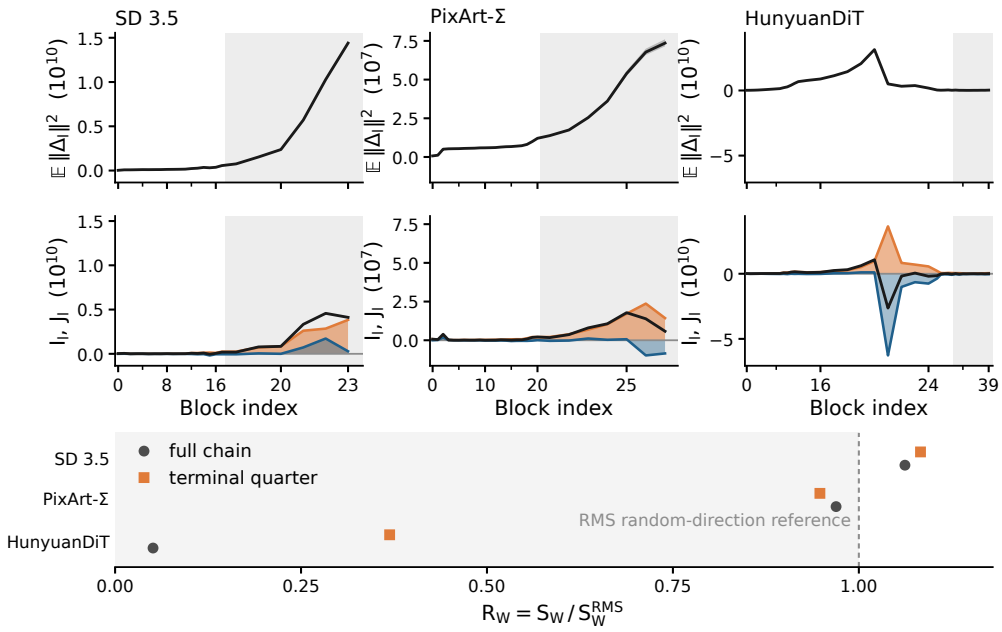

Figure C.4: **Increment accounting: reinforcement versus cancellation along the hooked chain (details of (e)).** Top row (black): mean squared drift energy per layer, terminal quarter shaded. Bottom row: increment energy $\|\delta_l\|_2^2$ (orange) and signed reinforcement $2\langle\Delta_{l-1}, \delta_l\rangle$ (blue). Right panel: measured segment straightness R_W against the matched-step random-direction reference (dashed line at 1). Values are in the text.

(g) Norm schedules of the extended-registry instances. The two new families give two sharply contrasting schedules. The reference norm of Sana-1600M rises monotonically in step with drift, from about 1.2×10^4 to 2.3×10^5, consistent with the rank-equivalence account of (c). The reference norm of Z-Image peaks in mid-network and then declines, from a peak of about 4.0×10^5 to about 3.2×10^5 at the final layer, while its drift profile still rises monotonically over the same segment ($\rho = +0.9893$, Table 1). Over those layers, drift increases while the norm of the reference features declines. This counterexample refines "the absolute profile is a rank-equivalent readout of the norm schedule" into a layered claim: the norm schedule is the main carrier, stable within a lineage, per (c) and (d), but it is not an identity, and the morphological dichotomy is stable on both carriers (Z-Image still satisfies the frozen monotonicity criterion under the drift metric). One point must be delineated precisely. For both instances, the overall ramp decomposition still has norm as the primary carrier (norm-term shares 104% and 113%, inside the frozen interval). The rel flip of Z-Image is dominated by the layers.0 small-denominator artifact (norm percentile 0 at that layer), so its terminal-segment counter-trend is neither an artifact of the rel metric nor covered by the overall share. It is a local fact in the layer-by-layer comparison of the absolute profile against the norm schedule.

(h) Sample-wise amplitude and direction decomposition. For one matched image, timestep, and layer with nonzero pooled vectors, let $a = \mathcal{P}(f_{l,t}^{\mathrm{ref}})$ and $b = \mathcal{P}(f_{l,t}^{\mathrm{recon}})$ be the identically pooled feature vectors. Their squared L2 discrepancy has the exact decomposition

$$\|a-b\|_2^2 = (\|a\|_2 - \|b\|_2)^2 + 2\|a\|_2\|b\|_2(1-\cos\theta),$$
$$\cos\theta = \frac{\langle a, b\rangle}{\|a\|_2\|b\|_2}.$$

The first term measures mismatch in feature amplitude, and the second measures directional disagreement, weighted by both feature norms. We report the reference and reconstruction norms and absolute drift after averaging the nine timesteps within each image, and the relative drift $r_l(x) = d_l(x)/\|f_l^{\mathrm{ref}}(x)\|_2$ follows the aggregate ratio used in (c). The two squared terms are computed for every image–timestep pair before averaging the nine timesteps within image and then averaging images. Squaring the paper's already averaged absolute drift would not recover either averaged component. This decomposition complements the reference-norm analysis in (b)–(g): for Z-Image, it tests whether the region where reference norms decline while drift rises is accompanied by a larger reconstruction norm, increasing directional disagreement, or both. Neither component can be inferred from the existing scalar norm and distance profiles alone. The paired pooled features, or their inner products, are required. These measurements are reported only for the layers and samples with both feature vectors available. For SD 1.5 ($n = 104$, 38 hooked layers), the direction term exceeds the amplitude term at every layer: the layerwise ratio of image-averaged direction energy to image-averaged squared distance ranges from 85.7% to 99.98%, with a median of 99.34%. Z-Image ($n = 104$, 30 layers) shows the opposite balance: its direction share is 0.9%–23.3% across layers, with a median of 1.7%, so the amplitude-mismatch term remains larger at every layer. From the reference-norm peak at layer 19 to the final layer, the mean reference norm falls by 20.6% ($6189 \to 4914$), while the mean reconstruction norm falls by 39.1% ($5483 \to 3338$). Absolute drift rises by 109.0% ($843 \to 1762$), and relative drift rises from 0.136 to 0.359. Over the same endpoints, the amplitude term's share of mean squared distance changes from 98.3% to 82.6%, while the direction term's share rises from 1.7% to 17.4%. Thus the counter-trend in Z-Image accompanies a larger decline in the norm on the reconstruction side and increasing directional disagreement. The amplitude mismatch remains the larger contribution. Figure C.5 reports three views per model, one model per row. The first column plots the reference norm, the reconstruction norm, and the absolute drift in raw units on a shared log axis. The second column decomposes the per-record squared distance into its amplitude and norm-weighted direction terms against their total, with the direction share of the area written in each panel. The third column puts the three metrizations of the same drift (absolute, relative, cosine distance) on one min–max normalized scale, with a dot at each curve's maximum.

C.10 Exact-inversion frameworks: re-measuring the accent with the residual removed

Three lines construct an inversion whose reconstruction residual is zero or near zero by design rather than reduced by a better solver: edit-friendly DDPM noise spaces (Huberman-Spiegelglas et al., 2024), Direct Inversion (Ju et al., 2024) (Part II, whose two branches coincide under our reconstruction setting), and AIDI (Pan et al., 2023) (fixed-point iteration, $I = 6$). Each is re-measured under the full protocol: 104 images, the same hooks, the same reference noise (seed 42), and the same $K = 9$ timesteps, with all arms sharing z_0 and the reference features so that the only difference between arms is the inversion–reconstruction path. Two backbones cover one morphology class each: SD 1.5 (interior-localized, 38 hooked UNet modules, reference peak `up_blocks.2.resnets.0`, pp $= 0.6842$) and PixArt-Σ (terminal-ramp, 28 transformer blocks, peak `transformer_blocks.27`, pp $= 0.9643$). D_s is taken against the canonical profile of the same model (Table 1). Two DDIM arms bracket the comparison: one reproduces the registry protocol exactly (a pipeline self-check for PixArt that returns $D_s = 0.0000$, $\rho = +1.0000$), and one uses the same algebraic schedule as the framework arms. The residual gate $\mathrm{end_rel} = \|z_0^{\mathrm{recon}} - z_0\|_2/\|z_0\|_2$, measured per image at the end of the reconstruction leg, is the implementation check that the framework does what it is constructed to do. Three quantities are consequently reported separately: the endpoint residual (inversion against its own reconstruction), the measured drift itself (reconstruction against the *externally drawn* reference, the quantity of §2.1), and D_s against the same-run DDIM arm, which isolates the profile change caused by the inversion

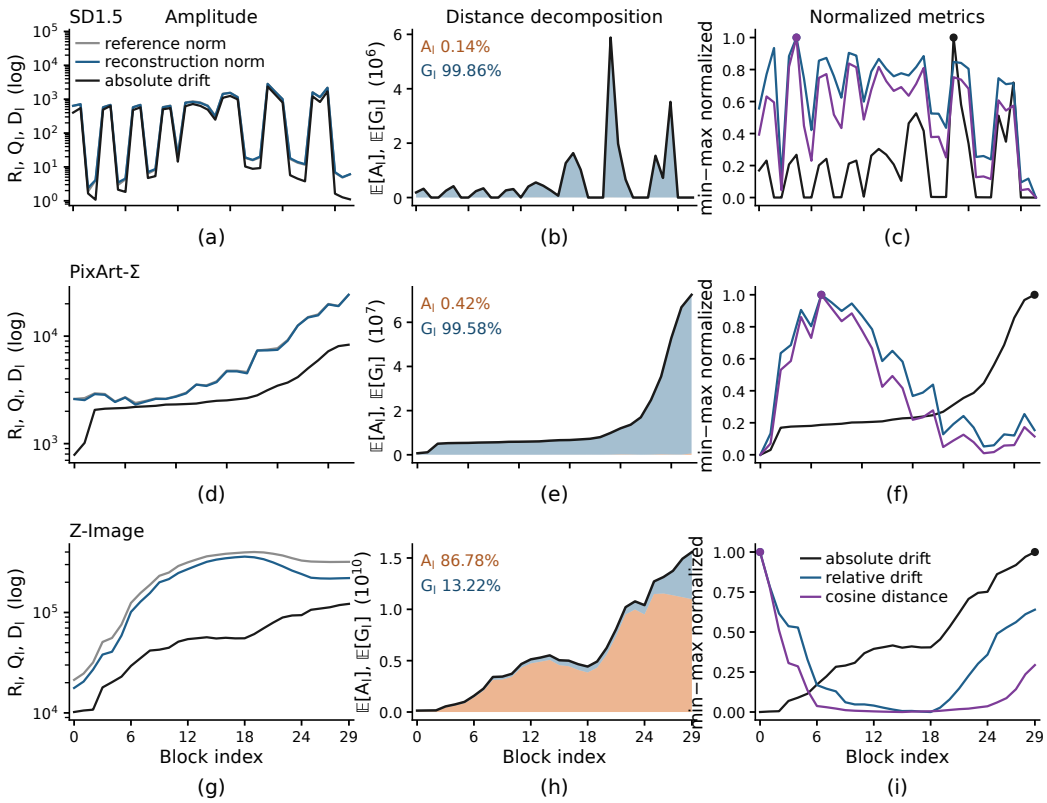

Figure C.5: **Amplitude and direction decomposition of the drift (details of (h)).** One model per row; panel letters run left to right. Gaps mark records whose inner product is undefined.

path alone with the reference held fixed. In Table C.10, "canonical" is the frozen registry profile this paper reports everywhere else, and D_s is the structural distance of Appendix B.1 measured against it, on the same noise floor ($p_{95} = 0.0042$) used throughout. Peak magnitude is the maximum of the aggregate drift profile, and because it scales with each backbone's feature norms it is not comparable across models; the magnitude ratio is therefore quoted relative to the model's own canonical profile. ρ is the per-image Spearman correlation with the canonical profile, summarized as median / minimum. All arms share z_0, the reference noise, and the hook set with the canonical run, so the profile differences isolate the inversion–reconstruction path.

Table C.10: **The accent under exact-inversion frameworks (P-multi@104, 104 images per arm).** Peak magnitude is not comparable across models. All arms share z_0, the reference noise, and the hook set with the canonical run, so the profile differences isolate the inversion–reconstruction path.

Arm	D_s vs. canonical (CI95)	D_s vs. own DDIM	peak magnitude	magnitude ratio	per-image argmax	ρ median / min
SD 1.5 (interior-localized; peak `up_blocks.2.resnets.0`*, pp* 0.6842*)*						
canonical (registry)	—	—	489.0	—	—	—
DDIM baseline	0.0000 (0.0000, 0.0018)	—	488.9	1.000	1.000	1.0000 / 0.9991
edit-friendly DDPM	0.0079 (0.0067, 0.0092)	0.0079	439.9	0.900	1.000	0.9943 / 0.9781
Direct Inversion	0.0093 (0.0080, 0.0108)	0.0093	395.9	0.810	1.000	0.9944 / 0.9761
AIDI	0.0082 (0.0066, 0.0097)	0.0082	415.4	0.850	0.981	0.9939 / 0.9786
PixArt-Σ (terminal-ramp; peak `transformer_blocks.27`*, pp* 0.9643*)*						
canonical (registry)	—	—	3243.8	—	—	—
DDIM, registry protocol	0.0000 (0.0001, 0.0030)	0.0122	3244.0	1.000	1.000	1.0000 / 1.0000
DDIM baseline	0.0121 (0.0100, 0.0142)	—	3081.3	0.950	1.000	1.0000 / 0.9995
edit-friendly DDPM	0.0507 (0.0495, 0.0520)	0.0386	2722.8	0.884	1.000	1.0000 / 0.9995
Direct Inversion	0.0573 (0.0558, 0.0587)	0.0451	2546.1	0.826	1.000	1.0000 / 0.9989
AIDI	0.0361 (0.0344, 0.0376)	0.0240	2643.7	0.858	1.000	1.0000 / 0.9989

Table C.11: **Implementation gate (criterion a): is the residual actually removed?** Edit-friendly DDPM and Direct Inversion return the initial latent to machine precision by construction; AIDI is approximate and its fixed-point iteration does not converge on every input. Every row covers the full 104 images per arm.

Arm	dtype	end_rel median	p90	max	residual drop (median)	images below 80% drop
SD 1.5						
edit-friendly DDPM	fp16	0.0000	0.0000	0.0000	100%	0/104
Direct Inversion	fp16	0.0000	0.0000	0.0000	100%	0/104
DDIM baseline	fp16	0.2263	0.4372	0.6222	(reference)	—
DDIM baseline	fp32	0.2249	0.4387	0.6222	(reference)	—
AIDI	fp16	0.0229	0.0888	0.5851	91.0%	26/104
AIDI	fp32	0.0062	0.0578	0.4066	96.8%	13/104
PixArt-Σ						
edit-friendly DDPM	fp16 / fp32	0.0000	0.0000	0.0000	100%	0/104
Direct Inversion	fp16 / fp32	0.0000	0.0000	0.0000	100%	0/104
DDIM, registry protocol	fp16	0.5086	0.5848	0.6704	(clipped, see (f))	—
DDIM baseline	fp16	0.3652	0.4633	0.6746	(reference)	—
DDIM baseline	fp32	0.3640	0.4587	0.6745	(reference)	—
AIDI	fp16	0.0473	0.0882	0.3661	86.3%	17/104
AIDI	fp32	0.0413	0.0648	0.3740	88.6%	10/104

(a) Exactness. Edit-friendly DDPM and Direct Inversion drive the residual to exact zero, not merely below the DDIM baseline but to 0.0000 on every image. Under these two arms, the features on the reconstruction side are the features on the inversion side, and Proposition 1's premise "the trajectory is exactly self-inverse" is realized. The canonical DDIM arm retains a median $\mathrm{end_rel}$ of 0.23 (SD 1.5) and 0.37 (PixArt-Σ): the discretization error the frameworks set out to remove is real and large. The self-check arm reproduces the registry profile to $\rho = +1.0000$ with $D_s = 0.0000$, so the measurement chain used here is the one used everywhere else in the paper. The endpoint residuals of each arm behind this reading are itemized in Table C.11.

(b) Peak position does not move. With the residual exactly zero, the aggregate argmax layer is unchanged in every arm of both models. The per-image agreement is 100% for edit-friendly DDPM and Direct Inversion on both backbones ($104/104$), while AIDI reaches 98.1% on SD 1.5 and 100% on PixArt-Σ. Per-image Spearman $\rho \geq 0.9761$ against the canonical profile throughout. Peak position is therefore not a property of the discretization residual: it is carried by the backbone's trajectory features, which the residual-free inversions leave untouched at that layer.

(c) Only the amplitude contracts. The structural distances stay small: $D_s \leq 0.0093$ for SD 1.5 and ≤ 0.0573 for PixArt-Σ, which is 2–$14\times$ the noise floor but 4.7–$34\times$ below the 0.2679 cross-morphology line. They are weak shifts of metric components rather than reorganizations. The peak magnitude falls by 10% (edit-friendly), 15% (AIDI), and 19% (Direct Inversion) on SD 1.5, and by 12%, 14%, and 17% on PixArt-Σ. The residual is thus a minor component of what the accent reads out: 81–90% of the drift magnitude and, more importantly, all of its peak structure, survives its complete removal. $D_{\mathrm{pp}} = 0$ and the morphology label is unchanged in every arm of both models.

(d) AIDI's fixed-point iteration does not converge on a subset of images. AIDI's median drop is 96.8% at fp32 (SD 1.5), but 13 of 104 images stay above the 80% gate, and on the three worst the residual is *higher* than the DDIM baseline's. This is a property of the method under the 50-step schedule, not of our harness: the per-layer fixed-point residual matches Algorithm 1 of Pan et al. (2023) update by update, larger iteration budgets do not fix the hard cases (the worst image moves from 0.231 at $I = 6$ only to 0.067 at $I = 12$), and the Anderson-accelerated variant ($m = 3$, $I = 8/15$) only partially reduces them. The accent is insensitive to this. On the SD 1.5 leg, the per-image profiles of the non-converged images agree with the canonical profile at least as well as the converged ones (ρ median 0.9963, min 0.9805), and dropping all 13 failed images changes the aggregate magnitude by 1.8%. On PixArt-Σ, the same holds for the 10 images above the gate at fp32 (ρ median and minimum 0.9995, argmax on the final block $10/10$, aggregate magnitude -0.6%).

(e) The gate itself needs fp32. Along the reconstruction chain, a latent perturbation is amplified by $\sim 1.14\times$ per layer. Measured on the canonical chain, a $k{=}1$ perturbation of 5.3×10^{-4} reaches 3.2×10^{-1} at $k{=}49$, about $600\times$ over 48 layers. fp16 rounding ($\sim 10^{-3}$ per layer) is therefore amplified

into a 0.1–0.3 end-to-end residual on its own. This is why AIDI's fp16 gate reads worse than its fp32 gate (91.0% versus 96.8% median drop), and why the residual gate has no discriminative power in fp16. The DDIM baseline is unaffected (0.2263 in fp16 versus 0.2249 in fp32), confirming that the amplification acts on rounding noise rather than masking real error. On PixArt-Σ, the same fp32 gate, complete on all 104 images, leaves the two exact arms at 0.0000 and the DDIM baseline essentially unmoved ($0.3652 \rightarrow 0.3640$), while AIDI improves by only $1.15\times$ ($0.0473 \rightarrow 0.0413$) against $3.7\times$ on SD 1.5. There, AIDI's residual gap is algorithmic rather than precision-limited, consistent with **(d)**. Neither effect reaches the profile: the fp32 re-run leaves every reading of Table C.10 unchanged at three decimals ($D_s = 0.0508$, 0.0573, and 0.0364 for edit-friendly DDPM, Direct Inversion, and AIDI, with the aggregate argmax $104/104$ at `transformer_blocks.27`).

(f) A reproducibility note on the canonical PixArt-Σ data. The registry's PixArt-Σ pipeline derives its DDIM scheduler from the released DPM configuration, which carries no `clip_sample` key, so `clip_sample=True` is in effect on its reconstruction leg: the reconstructed $\hat{z}_0$ is clipped to $[-1, 1]$, while it reaches magnitudes of 2.0–2.5 along the chain. The clipped and unclipped DDIM arms differ by $\sim$ 5–6% of the profile magnitude and by $D_s = 0.0121$, with $\rho = +1.0000$ and no peak movement, so the finding of this paper is not affected. A re-implementation that disables clipping should, however, expect that offset. The framework arms are compared against the unclipped DDIM arm, since clipping is not an error the frameworks are constructed to remove.

D Training Dynamics and Crystallization Experiments

This appendix gives the full tests of two suggestive pieces of evidence, the random-initialization exclusion (D.1) and dual-objective crystallization (D.2), both controlled-retraining observations. An impulse-response probe of the residual stream is the direct way to extend the profile measured here to the ramp mechanism.

D.1 Random-initialization exclusion

Five random-initialization seeds of the SD 1.5 architecture are measured under the P-multi protocol. Every seed gives peak position $= 0.0263$ at `down_blocks.0.resnets.1`, at the input end. The trained peak is $= 0.684$ in the decoder, at $D_{\text{pp}} = 0.658$ (reported in §2.2). The untrained network's peak is far from the trained accent, and all 5 seeds land on the same input-proximal peak, exhibiting no trained organization. This completes the test of hypothesis (iii) (the initialization hypothesis) of §2.3: training dynamics are part of accent formation.

D.2 Dual-objective crystallization and data integrity

The crystallization experiment is run on DiT-S/2 (39.8M, 12 blocks). DiT-S/2 is smaller than the registry architectures (860M–12B), and the conclusions of this section are stated as controlled comparisons within one architecture.

Design of the comparison. The claim being tested is whether two training objectives on one and the same architecture converge to the same peak layer. The comparison is held to a single architecture (same block layout, width, and depth), a single dataset of 111 images, and a single initialization seed, so the objective is the only degree of freedom; within a lineage the accent is likewise insensitive to weight-only changes, so the endpoint read here tracks the pathway topology rather than how well the model fits (Appendix C.7).

Under the same DiT-S/2 architecture we train from scratch with ϵ-prediction DDPM (Ho et al., 2020) and with flow matching (Lipman et al., 2023), using the same data (111 images), the same 50k steps, and the same initialization seed. We track the peak position along training steps (Figure 3(c)). The ϵ-prediction variant has already reached its final peak position at 10k, the first fifth of training. The P-multi measurement covers the 10k and 50k points, both at block 11 with pp$= 0.9167$. The flow variant moves from block 9 (10k) to block 7 (20k) to block 11, stable from 30k on and supported by the 30k/40k/50k points. The two variants converge to the same peak position, block 11.

Data integrity: the 20k/30k/40k checkpoints of the ϵ-prediction run were deleted before the P-multi measurement. Reaching the final state at 10k is therefore supported only by the two measured points, 10k and 50k. Figure 3(c) draws no line between the two points, since no intermediate measurements

exist. The flow trajectory is supported by five points (10k/20k/30k/40k/50k). The training data (111 images) and the main-experiment evaluation set coco_val100 are mutually independent.

E Additional Mechanism Analyses

Intervention effects on the skip pathway (referenced in §3.3). The skip enters the up_block as the concatenated tensor $[u; \alpha s]$, where u is the decoder's own representation at that resolution and s the encoder skip. The two are fused by concatenation, not by subtraction, and they differ in channel count. At the peak skip (`up_blocks.2.resnets.0`) u has 1280 channels and s has 640. The mismatch between them is a structural property of how the block combines the two paths, not a scalar distance we measure. The intervention patches `CrossAttnUpBlock2D.forward` in `diffusers` 0.38.0 and scales `res_hidden_states_tuple`, so Cut A is $\alpha : 1 \to 0$ with u held fixed. The concatenated tensor then passes through group normalization, a nonlinearity, and convolutions, so the perturbed computation is a general block map $F(u, \alpha s)$ whose response is measured rather than derived. Attenuating the peak skip over $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ (19 images, canonical protocol) gives mean PSNR 24.66 / 23.10 / 23.02 / 22.57 / 22.44 dB. The decrease is monotone in α with no interior optimum, so the best reconstruction is the full cut. The discriminating case is Noise A, which replaces the skip content with Gaussian noise matched to the skip's own mean and standard deviation. The scale is preserved while the content is destroyed, and PSNR still improves. The drift magnitude is therefore not a common explanation of the two gains. The skip topology and the three intervention sites are shown in Figure E.1.

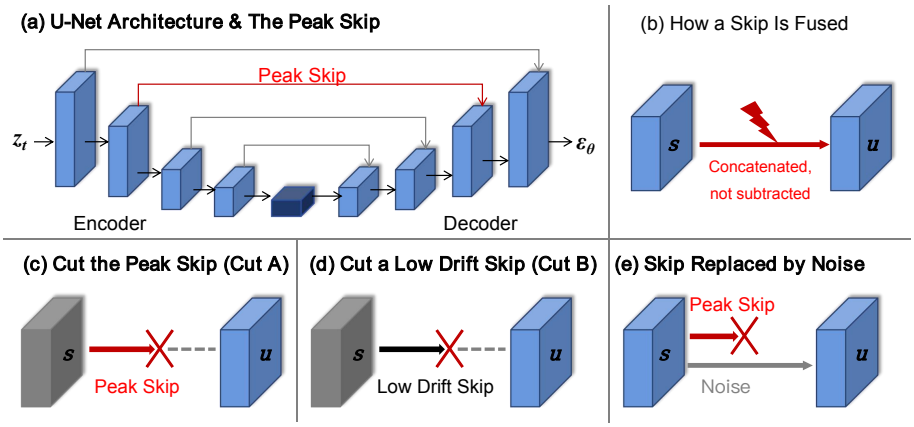

Figure E.1: **Peak-skip interventions in SD 1.5 (schematic).** (a) The four skips, with the peak skip carrying the drift peak highlighted. (b) How a skip is fused with the decoder's own representation: the encoder feature s is concatenated with the decoder tensor u before the block's convolutions. (c)–(e) Cut A cuts the peak skip, Cut B cuts a low-drift skip, and Noise A replaces the skip content with noise.

Layer-wise significance of the peak-skip intervention. The intervention reshapes the accent itself (Figure E.2). Cutting the peak skip significantly changes $31/38$ layers (all $31/31$ survive BH-FDR correction, $20/38$ survive the conservative Bonferroni correction). Cutting a low-drift skip changes far fewer layers ($5/38$ significant, $4/38$ survive BH-FDR).

Per-item values of the SD 1.5 skip interventions. Four interventions. Cut A ($\alpha = 0$, peak skip) reduces drift by 27.7% and improves PSNR by $+2.20$ dB. Cut B (low-drift skip) has no significant effect. Noise A replaces the skip content with noise. It raises the L2 drift magnitude yet still improves PSNR (paired $t = 4.97$, $p = 0.0001$, $n = 19$), and it is TOST-equivalent to Cut A at $\Delta = 0.5$ dB (one-sided $p = 0.0029$ and $p = 1.5 \times 10^{-6}$). Cutting sets the skip capacity to 0 while noise replacement leaves the capacity unchanged, and both improve PSNR alike. The gain therefore does not depend on the skip's content being preserved. The correlation analysis shows

that layers with higher drift respond more strongly to Cut A, $\rho(\text{drift}, \text{Cut A } \Delta\text{drift}) = -0.395$. The SDXL role-reversal control uses 30 steps and 20 generated images, cutting the same-named structural component rather than its global peak layer.

Controlled probes of the local-component control hypothesis (details of §3.2). Each of the two candidate hypotheses scans a preset set of doses. The pilot probe uses $n = 12$ with no per-image statistical inference.

FLUX: joint→single interaction strength. We scale the image-token stream entering the boundary by $\alpha \in \{0.75, 0.5, 0.25, 0\}$. Partial decay monotonically compresses the drift magnitude (final-layer drift $5.16 \times 10^5 \to 1.75 \times 10^5$), but the ramp survives a $4\times$ amplitude reduction ($\alpha = 0.25$, slope $\geq +0.99$). Only a full zeroing collapses the morphology, and that collapse site is nonspecific. Zeroing a mid-network single layer triggers a collapse of the same scale (ΔPSNR $+28.8$ vs. $+34.7$ dB). This is a symmetrization artifact: destroying the computation makes inversion and reconstruction degenerate into the same deterministic function. A norm-matched noise control separates magnitude from structure further. An equal-magnitude structural destruction destroys the ramp (slope $\to +0.11$), whereas an equal-magnitude decay preserves it ($+0.99$). Interaction strength is not a morphology knob.

PixArt-Σ: post-block relative escape. We apply readout decay $\gamma \in \{0.75, 0.5, 0.25\}$ to the proj_out input. The magnitude of post-block relative overtaking is dose-invariant (proj/vel rel ~ 1.1–1.6, never dropping to the T27 level of 0.40–0.56). Norm-matched noise at the proj_out entry reads out a rel indistinguishable from the base (1.112 vs. 1.217). Any decorrelated stream through a rank-reducing linear map reads out rel $\sim O(1)$, so part of the overtaking magnitude is a geometric fact of the rel-L2 metric. Zeroing the terminal block (T27) reduces mean PSNR from 17.97 to 17.03 dB, whereas zeroing T26 reduces it to 9.82 dB. Correcting the inversion endpoint z_T does not change the profile shape (slope $+1.000$, peak position 0.964) and makes reconstruction worse. Accumulation is a recomputation process inside the trajectory, not something actively amplified by the post-block stages.

What ΔW is here. ΔW denotes a *cross-checkpoint* weight difference, not a fine-tuning update. For each UNet weight tensor W_l (normalization layers excluded) it is the relative difference $\Delta W_l = \|W_l^{(1.4)} - W_l^{(1.5)}\|_F / \|W_l^{(1.4)}\|_F$ between two official checkpoints of the same architecture, CompVis/stable-diffusion-v1-4 and runwayml/stable-diffusion-v1-5. It is therefore a specific, external, and reproducible weight perturbation, the closest available stand-in for the same architecture with genuinely different weights. The statistic $\rho(\text{drift}, \Delta W)$ asks where the drift profile and the direction of that weight change agree per layer. We did not run this analysis under a fine-tuning update.

Functional asymmetry and rank decorrelation (Correlation). Drift and cross-checkpoint weight change concentrate in different functional sites. Drift concentrates in the ResNet residual stream, where residual-stream interventions have $2.1\times$ the ΔPSNR effect of attention interventions. ΔW concentrates in cross-attention K/V. The global rank correlation is $\rho(\text{drift}, \Delta W) = 0.24$. Up/attn layers show none at all ($\rho = 0.053$, n.s.), while down/attn layers partially correlate ($\rho = 0.303^*$, significant, Figure E.3(b)). This rank decorrelation is confined to the decoder side. It states that the drift magnitude of a layer and the magnitude of the weight change at that layer are not monotone in one another. It is a statement about the rank ordering of the two profiles, not about the geometric angle between value subspaces, which we did not measure. Together with the intervention asymmetry it offers one possible mechanistic explanation for checkpoint invariance. The weight-perturbation strength curve gives a further check. For $\varepsilon \leq 10^{-4}$ every D_s is below the noise floor, so the weight L2 distance does not predict accent distance.

MM-DiT cross-modal boundary (Observation). At the removal of interaction (FLUX joint $\to$ single) a localized bimodal spike appears (image drift $1.55\times$, text drift $3.0\times$, Figure E.3(a)). This is consistent with the hypothesis of the cross-modal interaction boundary as a feature stabilizer. The boundary peak (s2) seen under P-tT appears under P-multi as a localized weak signal consistent across all σ (0.13–0.17). The two readings are two faces of one architecture at two protocol coordinates.

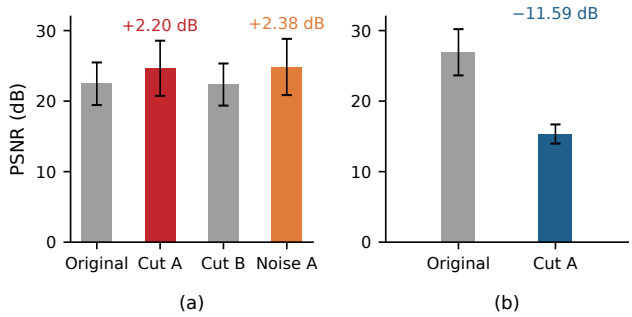

Figure E.2: **Mechanism: architectures localize the mismatch in different ways (summary of §3.3).** (a) The SD 1.5 skip-mismatch interventions: Cut A (peak skip), Cut B (low-drift skip), and Noise A (noise replacement), which doubly dissociates drift magnitude from reconstruction quality. (b) SDXL role reversal: cutting the same skip component destroys reconstruction rather than improving it. Values are in the text.

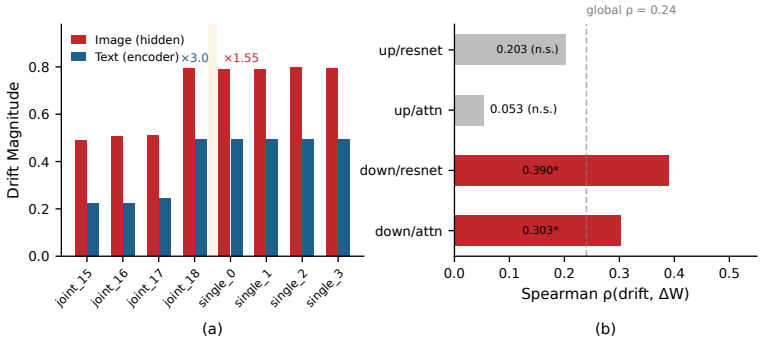

Figure E.3: **Additional mechanism analyses (P-multi@104).** (a) The FLUX joint → single cross-modal boundary: a localized bimodal spike at the removal of interaction, in image drift and in text drift. (b) Functional asymmetry and rank decorrelation: Spearman ρ (drift, ΔW) grouped by block type, with the gray dashed line marking the global $\rho = 0.24$ reference. Readings are in the text.

F Full Correction and Editing Results

Correction and injection strength. We evaluate latent and feature correction without training or per-image optimization (Table F.2). Figure F.1 shows the effect of injection strength on editing fidelity. LPIPS decreases sharply and then plateaus, with $\lambda = 0.05$ already achieving more than 90% of the improvement over the uncorrected setting.

For SD 1.5, latent correction interpolates the reconstruction latent with the inversion state at each matched timestep. FLUX instead uses a single endpoint correction. We selected $\lambda = 0.7$ on a five-image subset from $\{0.3, 0.5, 0.7\}$. Feature correction injects inversion features at the accent's top-five layers in SDXL, HunyuanDiT, and the 104-image evaluation. On the canonical 104 images at $\lambda = 0.7$, latent correction gives ΔPSNR $= +3.67$ dB, with Cohen's $d = 1.32$ and CI95 $[3.15, 4.21]$. Feature correction gives $+3.478$ dB, with CI95 $[3.01, 4.01]$ and dLPIPS $= -0.145$, using sites selected on the same 104 images. Latent correction performs better on these images without selecting layers.

Controlled comparison of correction sites. Each arm injects into five layers using the same λ and 50 reconstruction steps. All arms share one inversion and rerun only reconstruction, so the corrected layer set is the sole difference (Table F.1). At $\lambda = 0.7$, the accent's $+3.478$ dB falls inside the range of ten uniformly random five-layer sets: mean $+3.441$ dB and range $[+3.232, +3.554]$. Paired differences range from -0.076 to $+0.246$ dB, with five seeds outperforming accent selection. The accent arm also trails the next five layers by reference norm, disjoint from the drift top-five, by 0.090 dB. The paired CI95 is $[-0.114, -0.068]$, and norm selection is better on 91 of 104 images.

At $\lambda = 0.3$, accent selection exceeds all ten random sets by $+0.041$ to $+0.557$ dB. Each paired CI excludes zero, and accent selection is better on 96–103 of 104 images. It again trails the norm-adjacent layers by 0.017 dB, while exceeding five evenly spaced layers by 0.177 dB. Thus the layer set changes PSNR by a few tenths of a dB against an overall gain of ≈ 3.5 dB, with direction depending on λ and the comparator. A no-op arm with the same code path and an empty layer set is bit-identical to baseline. Drift and reference norm select the same top-five layers on six of seven measured models and overlap on $4/5$ layers in SD 2.1, with Spearman $\rho = 0.93$–1.00 and identical argmax throughout. The independent site-sensitivity experiments show a larger effect in HunyuanDiT, $+5.65$ versus $+2.50$ dB (§3.4). Figure F.2 gives the paired differences of each arm behind Table F.1: per image, $\Delta_i = \mathrm{PSNR}_i(\text{accent}) - \mathrm{PSNR}_i(\text{alternative})$, so positive means the accent site wins, with each row reporting the image mean and its paired bootstrap CI95 ($B = 1000$, one shared image-index resample across arms within a panel) and the ten random draws resolved individually above their seed mean. Panel (a) is $\lambda = 0.7$ and panel (b) is $\lambda = 0.3$, in the row layout of the table.

Residual Linear Interpolation (RLI) (Jo et al., 2025) is also training-free and acts on the decoder side: it interpolates self-attention value features to improve reconstruction and content preservation, with a manually tuned weight and no site selection. Our feature correction acts at the accent's top-five layers. Correction uses the inversion–reconstruction difference, $z_{\text{inv}} - z_{\text{recon}}$, or its feature analogue, which vanishes when the two trajectories coincide. Drift instead compares reconstruction with an independent forward reference through $f_{\text{ref}} - f_{\text{recon}}$. This discrepancy remains positive in the exact-inversion controls (Appendix A.1). The contribution is the measured regularity along depth and its use to identify candidate intervention sites. On the 19 selection images the latent-level arm is statistically equivalent to the attention-replay baseline (TOST $\Delta = 0.2$ dB). The interpolation form is not what distinguishes this correction from exact-inversion methods such as Direct Inversion (Ju et al., 2024), which derive a corrected trajectory from the inversion formula itself: here the correction is a diagnostic localization read off the accent, with the interpolation as its global corollary.

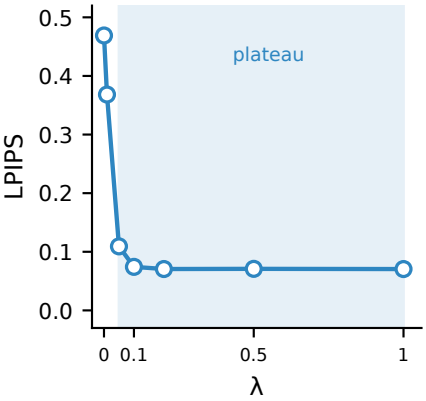

Figure F.1: **Injection-strength scan on 121 edit pairs.** LPIPS against the injection strength λ: it falls sharply and then plateaus. Values are in the text.

Table F.1: **Controlled comparison of correction sites (SD 1.5, canonical P-multi@104, 104 images per arm).** ΔPSNR against the same harness's baseline (22.05 dB), CI95 from an image-resampling bootstrap ($B = 1000$); every arm shares one inversion and one hook set, so only the corrected layer set differs between rows. †mean over seeds, with the seed range in brackets; paired accent-minus-alternative differences are in the text.

Five-layer injection set	$\lambda = 0.7$	$\lambda = 0.3$
accent (drift top-5)	$+3.478$ $[3.01, 4.01]$	$+3.415$ $[2.96, 3.92]$
reference-norm ranks 6–10 (disjoint)	$+3.568$ $[3.10, 4.08]$	$+3.432$ $[2.98, 3.94]$
evenly spaced over canonical depth	$+3.418$ $[2.93, 3.87]$	$+3.238$ $[2.81, 3.72]$
uniformly random (10 seeds)	$+3.441$ $[3.23, 3.55]^{\dagger}$	$+3.217$ $[2.86, 3.37]^{\dagger}$
no-op (empty layer set)	$+0.000$	—

Supplementary site-sensitivity results (details of §3.4). Site selection is done on 19 images and the main evaluation on the 104-image canonical set, which contains those 19. The overlap and the disjoint re-run are reported in Appendix A.2. For SD 1.5 and FLUX the disjoint selection reproduces the site list and the verdict, while the HunyuanDiT leg retains the overlap caveat. Each model compares two arms, accent-site injection and blind injection. SD 1.5 and FLUX tie and are site-insensitive. For FLUX the three injection granularities are exactly equivalent. HunyuanDiT is site-sensitive: injecting at the peak transition layer gives $+5.65$ dB, significantly better than $+2.50$

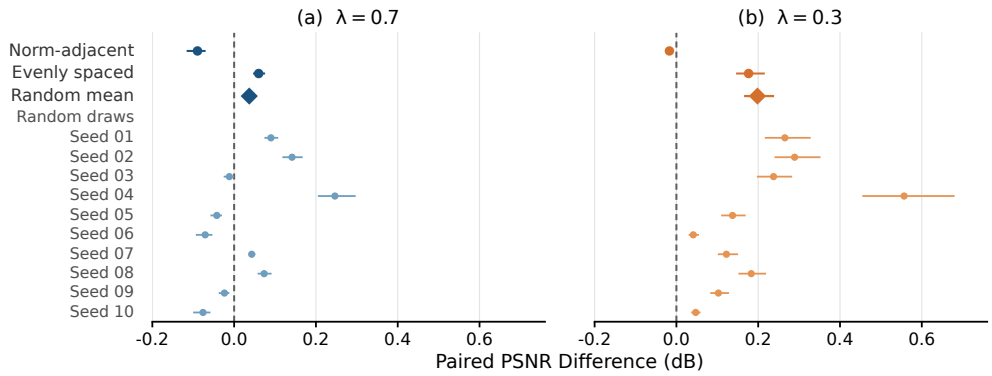

Figure F.2: **Paired gain of the accent site over alternative site sets under the same intervention settings.** Rows are the alternative site sets of Table F.1. (a) $\lambda = 0.7$; (b) $\lambda = 0.3$.

dB at the top-5 drift layers. Sensitivity tracks the pathway organization of each architecture rather than the morphology class. SD 1.5 and HunyuanDiT are both interior-localized yet fall on opposite sides, so the outcome is reported per architecture.

Provenance of the SD 1.5 correction sites. The reported sites are selected by the documented protocol on the canonical 104-image set ($+3.478$ dB, Table F.1). Because selection and evaluation share images, the selection is re-run on a set rebuilt to be disjoint from the 104. The disjoint run reproduces the documented ranking ($\rho = +0.9989$, top-5 and top-10 identical), hence the same five layers, and scores $+3.470$ dB. That number and the $+3.478$ dB of Table F.1 are the same layer set measured by two different harnesses (the correction-evaluation script and the controlled site comparison) on the same 104 images; the two baselines differ by 0.01 dB. The disjoint leg is recorded in Appendix A.2. The controlled site comparison above quantifies how much the choice of layers affects the correction gain.

We further validate correction across prompts on 100 diverse prompts with SD 1.5 under the canonical protocol, reconstructing generated images. This image source differs from the real-image inversion evaluation of §3.4. Correction gives $\Delta\text{PSNR} = +1.88 \pm 2.25$ dB ($p = 5.15 \times 10^{-13}$, $d = 0.835$). The FLUX validation uses 19 images with one-shot latent endpoint correction at $\lambda = 0.7$ and gives $\Delta\text{PSNR} = +3.94$ dB ($p = 6.4 \times 10^{-11}$, $d = 3.12$). The edit benchmark covers 121 edit pairs $\times$ 3 conditions, with mean PSNR gain $+7.4$ dB ($18.25 \rightarrow 25.65$). This is an editing-task convention and is not comparable with the reconstruction-task gain. The λ cliff scan over $\lambda \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0\}$ shows LPIPS dropping from the baseline 0.469 to 0.109 at $\lambda = 0.05$ and then reaching a plateau at 0.071 from $\lambda = 0.2$ on (Figure F.1). The range $\lambda \in [0.05, 1.0]$ covers $> 90\%$ of the LPIPS improvement, and the transition window from $\lambda = 0.01$ to 0.05 is only 0.04 wide. All values are measured against the $\lambda = 0$ reconstruction of each pair. Semantic agreement of the edit is not measured, so this is a fidelity result. CLIP-Dir uses CLIP ViT-L/14 (Radford et al., 2021).

NTI row. Lightweight reproduction configuration for SD 1.5: 50 sampling steps, 10 optimization iterations, guidance scale 7.5, learning rate 0.01, timed against same-batch DDIM (the attention-replay baseline) at 2.39 s/image, about $2.9\times$. Classic NTI with hundreds of iterations takes minutes per image.

Qualitative reconstruction cases. Four cases are fixed in advance by the per-image accent gain over the 104-image control run: its 10th, 50th, and 90th percentiles, and its minimum. Each case is one re-run of the 50-step DDIM protocol at $\lambda = 0.7$ in which all four arms share the same baseline reconstruction. Below each reconstruction is the per-pixel error map $e(p) = \text{mean}_c |I_c(p) - \hat{I}_c(p)|$ on the $[0, 1]$ display scale, drawn in magma on a colour scale set per case. The scale maximum vmax is the largest error over that case's four corrected arms, with no clipping. The maps are therefore

Table F.2: **Cost accounting (qualitative; proof-of-concept setting; supporting the closed-form claims of §3.4).** Storage is counted against a reconstruction loop that does not retain the inversion path, and no entry uses training. Latent correction adds no model calls; feature correction adds none when features are cached, or one inversion pass when recomputed.

Method	Per-image optim.	Extra storage	Extra computation	Applicable
Latent correction (ours)	none	one latent trajectory	none	inference
Feature correction, cached (ours)	none	one latent trajectory + inversion features at 5 layers × 50 steps	none	inference
Feature correction, recomputed (ours)	none	none	one inversion pass	inference
Attention-replay baseline	none	full cross-attn trajectory	none	inference
NTI-style per-image optimization	**yes (backprop)**	none	10 optimization iterations	editing
EDICT-style dual trajectory	none	second trajectory	2× sampler calls	reconstruction

comparable across arms within a case and not across cases. The original column has $e \equiv 0$ by construction and anchors the low end. The plate is split into two sheets, and the PSNR and LPIPS printed under each tile are measured on the rendered images themselves.

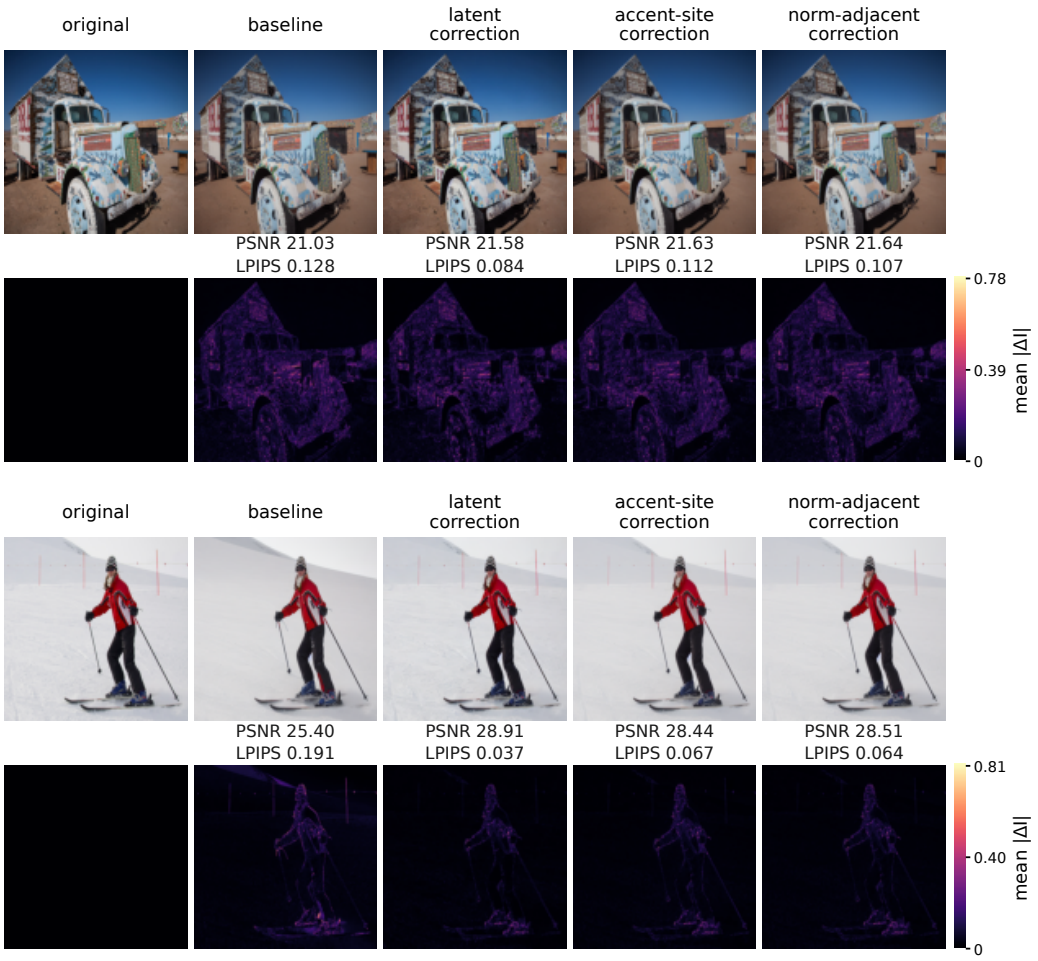

Figure F.3: **Reconstruction examples and per-pixel error maps, sheet (a) of one plate.** Cases at the 10th (top) and 50th (bottom) percentiles of the per-image accent gain. Columns: the original, the uncorrected baseline, and the three correction arms (site-free latent, accent-site, norm-adjacent), all at one shared inversion.

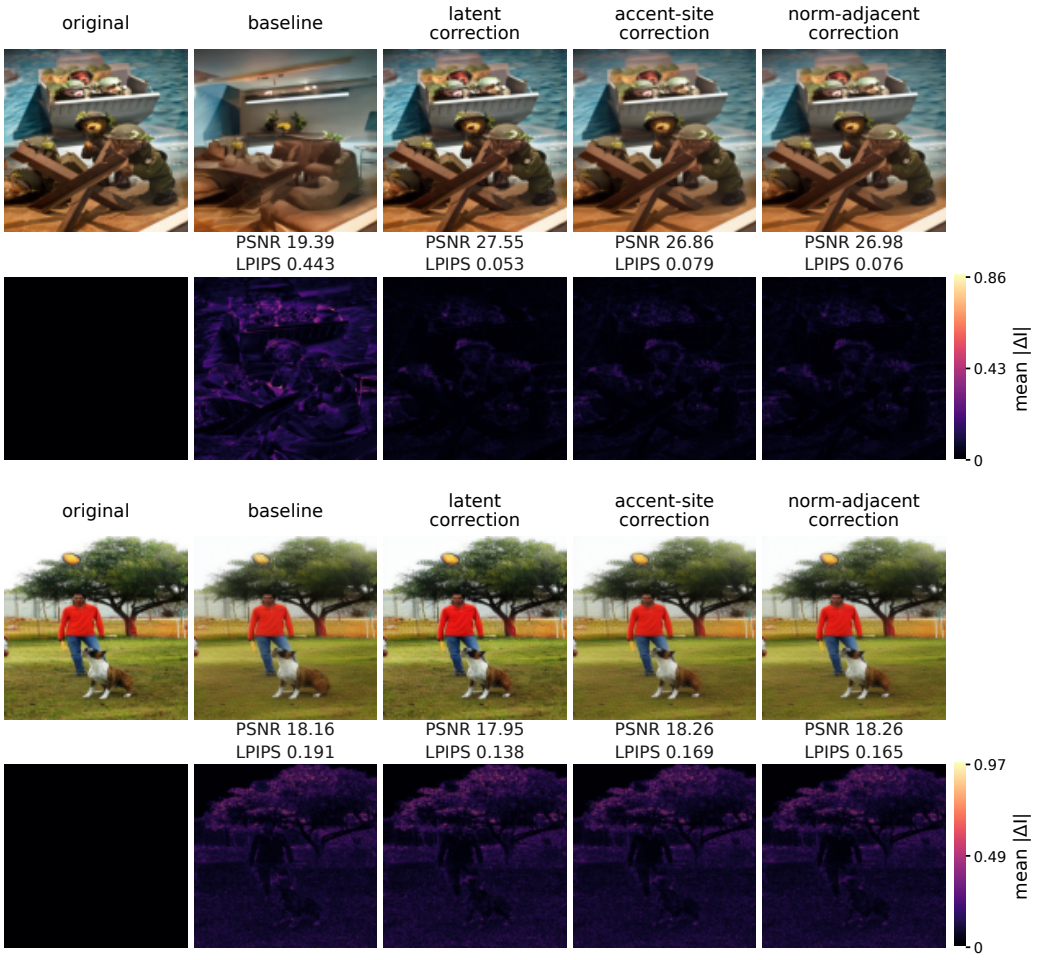

Figure F.4: **Reconstruction examples and per-pixel error maps, sheet (b) of one plate.** Cases at the 90th percentile (top) and at the minimum (bottom) of the per-image accent gain, in the layout of Figure F.3.