STRUCTURAL TEXT RECIPIENTS ROUTE SAME-PROMPT VISUAL REALIZATION IN DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention-based editing and guidance make internal attention a practical interface for controlling text-to-image generation. Yet high attention mass does not specify whether a recipient supports prompt semantics or influences how those semantics are visually realized. We investigate this distinction through paired interventions in Stable Diffusion 3, holding prompt and initial noise fixed. Across 553 prompts and four seeds, dynamically selected Contrastive Language–Image Pre-training (CLIP)-fused structural text recipients (padding and special-token positions) reproduce 98.9% of the all-sink visual displacement magnitude relative to a custom attention processor with intervention disabled (no-op), whereas event-count-matched structural-position and random-text suppression controls produce much smaller effects. Deleting the selected value contributions without renormalization retains 96.1% of this displacement: the effect does not require redistributing attention mass. Exact local identities distinguish contribution deletion from redistribution, while a trajectory expansion identifies the additional role of downstream propagation. Consistent with this distinction, later interventions can produce larger immediate changes but smaller final effects, and dose scaling produces graded mean displacement. Two vision-language model families find high prompt-semantics preservation even among the largest audited image changes relative to the implementation-matched no-op. In this setting, structural recipients therefore carry contributions that strongly affect visual realization without necessarily disrupting judged prompt semantics. Their function is better characterized by intervention consequences than by attention mass alone.

1 INTRODUCTION

Internal attention is an important interface for controlling text-to-image (T2I) diffusion. Prompt-to-Prompt edits cross-attention to guide image changes (Hertz et al., 2022), and attention-based guidance modifies generation without retraining (Chefer et al., 2023; Hong et al., 2023; Ahn et al., 2024). These uses make a practical question unavoidable: what does a high-attention recipient (a key/value position) actually contribute to the generated image? A large weight can identify a useful intervention site without identifying semantic content. Conflating these roles can obscure what an editing operation changes and what an importance-based removal discards. Our question is therefore not whether attention weights alone are explanations, but where high-impact contributions are concentrated and how changing them affects a denoising trajectory.

Attention sinks, positions receiving concentrated attention mass, motivate distinguishing mass from function. Language-model sinks have been related to long-context anchoring, softmax normalization, avoidance of increasingly similar token representations, and value geometry (Xiao et al., 2024; Gu et al., 2025; Barbero et al., 2025; Binkowski et al., 2026). In T2I models, padding-token analyses and denoising information-flow studies already expose structure beyond prompt-word attribution (Toker et al., 2025; Kaplan et al., 2025). Prior diffusion-sink work reports that perceptual trajectories can change while semantic alignment remains comparatively stable (Wu & Summa, 2026). The unresolved issue is the causal organization of this dissociation: which recipients carry the effect, whether it depends on renormalization, and when their contributions matter. None of these questions is answered by inspecting attention mass alone.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

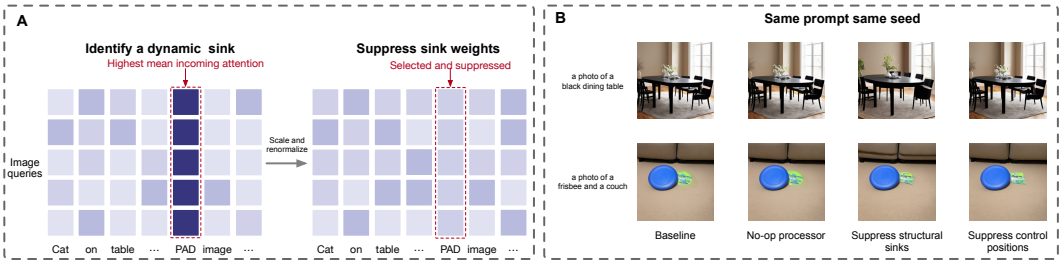


Figure 1: **Dynamic structural-recipient intervention changes same-prompt visual realization.** (A) Schematic of head-wise dynamic recipient selection and renormalizing suppression; shading illustrates attention weights. (B) Left to right: native baseline, custom processor with no intervention, CLIP-structural suppression, and an alternative structural-position control. Each row fixes prompt and noise; custom no-op isolates the intervention from processor changes. Examples are selected by displacement rank without semantic-success filtering (Appendix B.1). Tables 1 and 5 report aggregate displacement and semantic judgments.

We study this organization through controlled interventions in multimodal diffusion transformers (MM-DiTs), suppressing or scaling dynamically selected recipients while holding prompt and initial noise fixed. This measures same-prompt realization change, not an instructed semantic edit. We use Stable Diffusion 3 (SD3) (Esser et al., 2024) with explicit conditional and unconditional token maps under classifier-free guidance (CFG) (Ho & Salimans, 2022). A custom no-op processor provides the direct intervention reference; the native baseline separately reveals changes introduced by the processor.

The effect is strongly concentrated. Across 553 GenEval prompts (Ghosh et al., 2023) and four seeds, structural recipients in the fused Contrastive Language–Image Pre-training (CLIP) stream (Radford et al., 2021) produce 98.91% of the all-sink direct-no-op displacement magnitude. Alternative structural positions and random text targets produce much smaller changes despite matching selected-event counts. This localizes high-impact endogenous events. To distinguish contribution removal from mass redistribution, we then delete the selected value contributions without renormalizing the remaining weights. This retains 96.11% of the structural displacement magnitude: a large effect persists without redistribution. Exact local identities distinguish the two operations, and a trajectory expansion with an explicit remainder describes how local changes propagate. Structural identity remains entangled with mass and value geometry, so the result characterizes selected contribution sites rather than isolating the structural label as their cause.

The downstream effect depends on when and how strongly these contributions are changed: mean displacement varies with suppression strength, while a larger immediate perturbation need not produce a larger endpoint change. Two vision-language model (VLM) families, Qwen2.5-VL (Bai et al., 2025) and LLaVA-Next (Liu et al., 2024a), find high prompt-relevant preservation even among the largest no-op-relative image changes. The observed influence on realization is therefore not equivalent to a loss of judged prompt semantics.

Contributions. (1) We locate SD3's dominant dynamic-sink effect at CLIP-fused structural recipients using paired interventions and event-matched controls. (2) We distinguish contribution deletion from renormalizing suppression through exact local identities and a deletion diagnostic, with an explicit-remainder expansion for downstream propagation. (3) We establish graded mean dose responses and paired immediate/endpoint ordering reversals at the tested layer. (4) We evaluate image-level semantic consequences with two VLM families, including the highest-displacement subset of the audited no-op/suppression pairs.

2 Related Work

Prompt-to-Prompt controls cross-attention during prompt-based editing (Hertz et al., 2022), whereas null-text inversion optimizes unconditional embeddings to enable editing of inverted images (Mokady et al., 2023). Self-Attention Guidance (SAG) and Perturbed-Attention Guidance (PAG) instead construct guidance signals from perturbed attention computations (Hong et al., 2023;

Ahn et al., 2024). Recent MM-DiT methods address token imbalance, rare semantic emergence, key/value editing, and block-wise control (Lv et al., 2025; Han et al., 2026; Li, 2026; Shin et al., 2025). These methods motivate distinguishing an editing or guidance objective from the localization question studied here. Our controls test the choice of recipient and the role of weight renormalization, rather than success at a requested edit. The supplementary PAG comparison contextualizes displacement and alignment changes at the tested scales, not editing performance (Table 33).

Sink, register, and massive-activation studies examine related but distinct objects: attention concentration, dedicated computational slots, and activation magnitude (Gu et al., 2025; Sun et al., 2026; Li et al., 2026; Darcet et al., 2024; Jiang et al., 2026; Olah et al., 2020). We select recipients by incoming mass rather than activation norm and test their value contributions; register and outlier findings alone do not locate these sites. Diffusion-language and video comparisons appear in Appendix D.

Padding and information-flow analyses characterize what diffusion representations encode and how information is transmitted (Toker et al., 2025; Kaplan et al., 2025); our paired interventions measure the image consequences of changing selected contributions. T-GATE studies temporal redundancy in diffusion attention (Liu et al., 2024b), while Representation Alignment (REPA) improves generation through training-time representation alignment at early network blocks (Yu et al., 2024). Our timing study instead varies inference-time intervention along denoising steps, addressing the temporal organization of recipient contributions during sampling. Extended comparisons are provided in Appendix D.

3 Causal Setup

Study overview. The primary study (P) uses one SD3 protocol to examine contribution location, intervention operation, temporal dependence, and image-level consequences. All main findings below come from this study, with references specified for each measurement. Model-specific supplementary studies are reported separately; Table 6 records their execution-verification status and distinguishes exploratory observations from the primary evidence.

3.1 Dynamic Recipients and Interventions

For a prompt c, diffusion sampling maps initial Gaussian noise to an image through repeated denoising updates. We keep the prompt and initial noise fixed within each comparison. In SD3 joint attention, image-query rows attend to a shared image–text key/value table. Its text stream contains 77 CLIP positions, formed by fusing the CLIP-L and CLIP-G text encoders along the hidden dimension, followed by 256 Text-to-Text Transfer Transformer (T5) positions; the two CLIP encoders do not occupy separate position ranges.

At each intervention site and in each head, we average unmodified attention probabilities over image-query rows and select the top-k recipient. Our primary setting uses $k = 1$ at layer 12. Selection is head-wise, not per-query top-1 selection or a ranking pooled across heads. Family gating is applied after selection: *CLIP structural* accepts selected CLIP-fused padding or special positions; *all sinks* accepts any selected recipient. Each tokenizer labels its padding identifier (ID) as padding, other special IDs as special, and remaining IDs as content. A fused position is content if either encoder assigns content, padding if both assign padding, and otherwise special if either assigns special. Thus beginning-of-sequence/start-of-text (BOS/SOT), end-of-text (EOT), and padding positions can qualify, but a content/special disagreement is excluded. When padding and EOT share an ID, this coarse selector groups them rather than distinguishing their sequence roles; both belong to the accepted structural family. Finer role annotations are described in Appendix F.3. This operational sink definition is a dynamic rank criterion, not a claim that the same slot is a sink in all prompts, heads, or timesteps.

Suppression subtracts $\lambda = 10^4$ from selected image-query logits before softmax. Dose scaling instead adds $\log \gamma$ for $0 < \gamma \leq 1$; $\gamma = 0$ uses the saturated hard mask. Text-query rows are unchanged. The distinct *contribution-deletion* diagnostic sets selected weighted value contributions to zero without renormalizing other probabilities. Structural-position controls select alternative non-sink structural targets, and random-text controls relocate the same number of masking events to other text positions. These comparisons probe location and event budget, not a jointly mass- and

Table 1: **Branch-correct SD3 revalidation.** The primary measure d_0 is mean direct LPIPS to the custom no-op. The supplementary over-floor percentage is $100(\mathbb{E}d(B, I) - \mathbb{E}d(B, N))/(\mathbb{E}d(B, A) - \mathbb{E}d(B, N))$, where B, N, A, I are native baseline, no-op, all-sink, and intervention images. It is a difference-of-distances statistic, not a direct causal distance. Intervals are paired prompt-cluster bootstrap 95% confidence intervals (CIs), averaging four seeds within each prompt. All-sink is the normalization reference.

Intervention	d_0	Over-floor (%)
All sinks, both branches	0.1329	100.00
CLIP structural, both	0.1314	98.43 [96.99, 99.84]
CLIP structural, $\gamma = 0.5$	0.0800	45.40 [43.69, 47.08]
Structural-position control	0.0414	6.03 [4.59, 7.46]
Random-text control	0.0379	1.92 [0.57, 3.28]
CLIP structural, conditional only	0.1260	93.26 [90.00, 96.58]
CLIP structural, unconditional only	0.1291	96.72 [93.31, 100.41]

geometry-matched counterfactual. They follow a controlled intervention rationale (Vig et al., 2020; Meng et al., 2022), but are not activation-replacement or causal-mediation estimators.

3.2 Primary Protocol and Measurements

The primary SD3-medium runs use float16, 1024×1024 generation, 20 inference steps, CFG scale 7.5, and base seeds 1000, 2000, 3000, and 4000 on 553 GenEval prompts (Ghosh et al., 2023). The generator seed is the base seed plus prompt index. Unless stated otherwise, intervention occurs only at zero-based step 0; the processor observes without intervening at other steps. Conditional and unconditional CFG branches use token maps from the positive and empty negative prompts, respectively. Both branches are edited unless explicitly isolated. For guidance scale w, the guided prediction change is $\Delta v_{\mathrm{CFG}} = w\Delta v_c + (1 - w)\Delta v_u$, where subscripts c, u denote conditional and unconditional predictions. Nonempty negative prompts are not tested.

Our primary perceptual measure is Learned Perceptual Image Patch Similarity (LPIPS), using AlexNet at evaluation resolution 256×256, directly between the custom no-op and intervened images. The native/no-op difference is reported separately. An additional *over-floor* statistic subtracts native/no-op distance from native/intervention distance; it is not the direct no-op distance and does not exactly remove numerical error. Ratios compare mean perceptual magnitudes, not image identity or causal shares. CLIP-T is image–prompt embedding cosine similarity using CLIP ViT-B/32 (a base Vision Transformer with 32×32 patches); VLM judgments assess relative preservation and absolute prompt validity. Primary image comparisons use paired prompt-cluster bootstrap intervals, retaining all four seeds within each resampled prompt. Full configurations, source checks, and audit procedures appear in Appendices B and C.

4 Where the Effect Is Concentrated

We locate the suppression effect by recipient family, then test whether alternative structural positions or random text targets reproduce it under matched selected-event counts.

4.1 Structural Recipients Recover the All-Sink Magnitude

Table 1 gives the primary branch-correct comparison. All-sink suppression yields mean direct-no-op LPIPS 0.1329; restricting it to CLIP-fused structural recipients yields 0.1314. Their magnitude ratio is 98.91% [97.82, 99.95]. The structural family therefore reproduces nearly all of the displacement magnitude in this setting. The table also reports the separately defined over-floor statistic.

Magnitude recovery does not imply the same trajectory. Structural and all-sink outputs differ from each other by LPIPS 0.0559 [0.0529, 0.0589]. Nor is the native implementation a transparent substitute for custom no-op: their mean distance is 0.0357. Recovery ratios measure displacement magnitude, not image equivalence or preserved semantic information.

A matched 128-prompt, four-seed replication at layers 8 and 16 recovers 85.1% and 99.3% of the respective all-sink displacement magnitudes. Structural suppression exceeds both controls at each layer, and no-renormalization deletion retains 89.9% and 87.9% of its effect (Appendix B.4). Concentration and deletion effects therefore recur beyond layer 12, with depth-dependent magnitudes.

Both CFG branches contribute to the observed effect. Conditional-only and unconditional-only structural interventions give direct-no-op distances 0.1260 and 0.1291. Their difference, 0.0031 [-0.0007, 0.0071], does not resolve an ordering. In each single-branch arm, the untouched branch's first-step prediction matches no-op exactly. Thus the effect is not confined to positive-prompt computation. The final-image effects of these separate rollouts are not additive contributions to the both-branch result.

4.2 What the Controls Identify

Event-count-matched structural-position and random-text suppression controls produce direct-no-op LPIPS 0.0414 and 0.0379, respectively, well below the structural intervention. The paired structural-minus-structural-control difference is 0.0900 [0.0862, 0.0938], and the operational control-adjusted recovery is 92.40% [90.45, 94.36]. Event logs verify matched counts of 49,875 conditional and 49,333 unconditional selections, different control targets, and no recorded branch-map mismatches. Equal event budgets or membership in the same structural family therefore do not suffice to reproduce the effect.

The replacement targets differ in incoming mass and projected value geometry. P therefore localizes selected high-impact events without isolating structural identity from those properties.

5 From Contributions to Trajectory Effects

We distinguish contribution deletion from attention redistribution through exact local identities and a downstream trajectory expansion. Deletion, dose, and timing experiments then test how these operations affect final displacement, without assuming that larger immediate changes yield larger endpoint effects.

5.1 Exact Local Effects and Downstream Propagation

The local identities apply to any selected recipient set, independently of the preceding empirical localization. Suppress layer, head, and time indices below; all identities hold at a fixed local state and selected set.

Consider a joint-attention block in a text-to-image diffusion model. Let i index image-query rows, j index recipient positions in the joint key/value table, and j' denote a dummy normalization index over the same recipient set; we reserve s for selected recipients inside a set S. Here q_i is the image query, k_j and v_j are the recipient key and value vectors, a_{ij} is the attention probability from image query i to recipient j, W_O is the attention output projection, and o_i is the projected attention output for image-query row i. The usual key-dimension scaling is absorbed into q_i. The attention output is

$$o_i = \sum_j a_{ij} W_O v_j, \qquad a_{ij} = \frac{\exp(q_i^\top k_j)}{\sum_{j'} \exp(q_i^\top k_{j'})}. \tag{1}$$

A dynamic sink recipient is not merely a token identity or a high-norm activation; it is a recipient state that receives unusually large incoming mass from image queries at a particular denoising phase.

Let S be a selected recipient set for row i, with S^c denoting the complement of S within the recipient set. Define its incoming mass and within/outside value averages by

$$m_{i,S} = \sum_{s \in S} a_{is}, \qquad \bar{v}_{S,i} = \frac{1}{m_{i,S}} \sum_{s \in S} a_{is} v_s, \qquad \bar{v}_{S^c,i} = \frac{1}{1 - m_{i,S}} \sum_{j \notin S} a_{ij} v_j. \tag{2}$$

When $m_{i,S} = 0$, the selected set contributes no mass and the intervention has zero local effect on this row. Below we consider $0 < m_{i,S} < 1$, which is the nondegenerate case in our experiments.

Hard pre-softmax suppression of S sets the selected probabilities to zero and renormalizes all non-selected recipients:

$$\tilde{a}_{ij} = \begin{cases} 0, & j \in S, \\ a_{ij}/(1 - m_{i,S}), & j \notin S. \end{cases} \tag{3}$$

Therefore the local output change is an exact identity:

$$\Delta o_i^{\text{renorm}}(S) = \tilde{o}_i - o_i = m_{i,S}\, W_O \left(\bar{v}_{S^c,i} - \bar{v}_{S,i}\right). \tag{4}$$

The dose-response intervention has an equally explicit form. Multiplying the selected unnormalized attention weights by $\gamma \in [0, 1]$ gives denominator $d_{i,S}(\gamma) = 1 - (1 - \gamma)m_{i,S}$, and hence

$$\Delta o_i^{\gamma}(S) = \frac{(1 - \gamma)\, m_{i,S}(1 - m_{i,S})}{1 - (1 - \gamma)m_{i,S}}\, W_O \left(\bar{v}_{S^c,i} - \bar{v}_{S,i}\right). \tag{5}$$

When $\gamma = 0$, Eq. equation 5 reduces to $m_{i,S}W_O(\bar{v}_{S^c,i} - \bar{v}_{S,i})$, recovering the binary suppression case in Eq. equation 4. For the no-renormalization diagnostic, the selected contribution is deleted without changing the remaining probabilities:

$$\Delta o_i^{\text{delete}}(S) = -\sum_{s \in S} a_{is} W_O v_s = -m_{i,S} W_O \bar{v}_{S,i}. \tag{6}$$

Equations equation 4–equation 6 show that attention mass is only one factor. A selected recipient family has a large local effect only when incoming mass is paired with a large projected value contrast or contribution.

Local output change and final-image change are distinct: downstream network transformations can attenuate, rotate, or amplify the perturbation. For a smooth intervention path with strength η, Proposition 1 in Appendix E gives the exact decomposition

$$z_\eta(1) - z_0(1) = \eta \int_0^1 \Phi(1, t) g_S(z_0(t), t)\, dt + R_2(\eta).$$

Here time runs from initial noise at 0 to the endpoint at 1; z_η is the intervened trajectory, g_S the vector field's derivative with respect to strength at zero, Φ the unperturbed linearized state transition, and R_2 the remainder. Equation equation 8 gives an explicit second-order bound, with all constants and a proof in the appendix. It requires smooth dynamics and locally stable recipient selection; rank switches, discrete-sampler error, and a small remainder under full suppression are not guaranteed.

This motivates three empirical questions rather than predetermined outcomes: does deleting the contribution retain a large effect; does changing strength produce graded final displacement; and does an equally long late intervention have the same consequence? The framework admits all of these comparisons, but none follows from attention mass alone. In particular, a smooth local formula does not guarantee monotone final LPIPS. We address these questions in turn using the primary protocol.

5.2 Large Displacement Does Not Require Renormalization

Does the displacement require redistribution of the removed mass? We compare renormalizing suppression with contribution deletion at the same selected events on all 553 prompts and four seeds. Deletion retains 96.11% [93.76, 98.51] of the structural intervention's direct-no-op displacement magnitude (Table 2). Thus redistributing removed probability mass is not necessary for most of the measured magnitude. The result supports a substantial role for the selected value contributions, not an explanation based solely on forced attention redistribution.

The all-sink point estimate is 101.59%, but its interval [99.23, 104.04] includes 100%, so this comparison does not establish larger displacement under deletion. There is also no theoretical 100% ceiling: local deletion depends on the selected value contribution, whereas renormalizing suppression depends on its contrast with the remaining mixture (Eqs. equation 4 and equation 6). Neither local vector norm is necessarily larger, even before downstream propagation or the final-image metric is considered.

The corresponding image-query-only value-zero operation produces exactly the same pixels as no-renormalization deletion on all 2,212 pairs for each family. This is an implementation check of

Table 2: **Corrected deletion diagnostic.** No-renorm/renorm ratios of mean LPIPS displacement, with paired prompt-cluster bootstrap 95% CIs. Direct-no-op is the primary comparison; over-floor is supplementary. These magnitude ratios are not bounded above by 100%.

Recipients	Direct-no-op ratio (%)	Over-floor ratio (%)
All sinks	101.59 [99.23, 104.04]	102.47 [99.26, 105.84]
CLIP structural	96.11 [93.76, 98.51]	95.23 [92.09, 98.53]

Table 3: **Branch-correct structural dose response (553 prompts $\times$ 4 seeds).** Ratios use hard structural suppression ($\gamma = 0$), not all-sink suppression, as denominator. The $\gamma = 1$ endpoint aliases custom-noop; the auxiliary explicit-generation check is restricted to 16 prespecified prompts $\times$ 4 seeds.

γ	Direct LPIPS [95% CI]	Relative magnitude (%) [95% CI]
0	0.1314 [0.1263, 0.1363]	100.0 (reference)
0.25	0.1040 [0.0996, 0.1084]	79.14 [78.09, 80.18]
0.5	0.0800 [0.0762, 0.0837]	60.88 [59.61, 62.13]
0.75	0.0542 [0.0514, 0.0571]	41.27 [40.02, 42.53]
1	0 (no-op endpoint)	0

algebraically equivalent operations, not independent causal replication. The corrected native, no-op, and renormalizing anchors also match the preceding formal run pixel-for-pixel. The structural ratio is below one, but its gap is not an additive causal share of redistribution; final-image distance is nonlinear. The semantic audit below evaluates renormalizing images, not these deletion outputs.

5.3 Graded Mean Dose Response

We vary step-0 structural suppression strength across five doses (Table 3). Each nontrivial condition contains 2,212 prompt–seed pairs. Mean direct-no-op LPIPS decreases from 0.1314 under hard suppression to 0.1040, 0.0800, and 0.0542 as retained mass increases. All four adjacent-dose paired intervals exclude zero.

This is a graded population response, not a linear law or a per-image slider guarantee. Only 494/553 (89.3%) prompt-level four-seed averages and 1,427/2,212 (64.5%) individual pairs are monotone across the grid. Mean CLIP-T delta intervals include zero at each tested nontrivial dose, but non-rejection is not semantic equivalence. The hard-suppression VLM results cannot certify every dose. The $\gamma = 1$ endpoint aliases no-op; independent generation at that endpoint is restricted to 16 prespecified prompts and four seeds. Full contrasts and uncertainty definitions appear in Appendix B.6.

5.4 Timing Separates Immediate Perturbation from Final Consequence

The dose study fixes intervention time. Does an equally long intervention later in sampling have the same consequence? We select 128 prompts independently of outcomes and compare single-step suppression at zero-based steps 0, 5, 10, and 15. Each arm edits one step at layer 12. Structural suppression at step 5 gives only 5.87% [5.31, 6.46] of the step-0 final LPIPS; at steps 10 and 15 this falls to 0.53% and 0.13%. All-sink suppression follows the same ordering (Table 4). Paired early-minus-late intervals exclude zero. Editing steps 0–2 increases structural magnitude to 105.04% [102.58, 107.57] of the step-0 result, but uses three times the step budget and is not an additive decomposition of temporal contributions.

Late effects are not small because selection stops: every all-sink arm selects 24 events per branch, with mean summed head-wise mass between 2.13 and 2.87 across tested steps and branches. Structural arms retain roughly 20–23 events per branch. Trajectory summaries show zero divergence before intervention. For structural suppression at steps 0, 10, and 15, immediate post-step latent L2 (Euclidean distance to the paired no-op state) is 0.78, 0.87, and 1.34, whereas final L2 is 141.28, 8.44, and 4.19. A larger immediate late perturbation therefore need not yield a larger endpoint change.

Table 4: **Branch-correct timing comparison (128 prompts × 4 seeds).** Steps are zero-based. Single-step arms have equal step budgets; steps 0–2 use three intervention steps. LPIPS is measured directly against custom-noop.

Intervened steps	Structural LPIPS [95% CI]	All-sink LPIPS [95% CI]
0	0.13378 [0.12345, 0.14404]	0.13597 [0.12507, 0.14705]
5	0.00785 [0.00695, 0.00878]	0.00806 [0.00716, 0.00900]
10	0.00071 [0.00064, 0.00079]	0.00076 [0.00068, 0.00084]
15	0.00017 [0.00015, 0.00020]	0.00018 [0.00016, 0.00020]
0–2	0.14053 [0.13021, 0.15062]	0.14351 [0.13340, 0.15356]

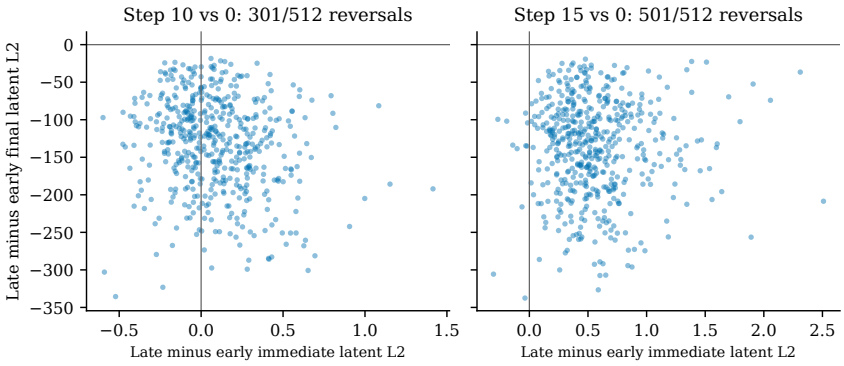


Figure 2: **Paired temporal ordering reversal.** Each point is a prompt–seed pair (128 prompts, four seeds) under the corrected SD3 protocol. Coordinates subtract the step-0 arm's latent distance to custom-noop from the late arm's distance, immediately after each arm's intervention (horizontal) and at the endpoint (vertical). The lower-right quadrant indicates larger immediate but smaller final displacement. Axes are latent L2 distances, not LPIPS or Jacobian estimates. Intervals in Table 9 cluster by prompt rather than treating seeds as independent prompts.

This ordering is not confined to condition means: step-15 structural suppression has larger immediate but smaller final distance than step-0 suppression in 501/512 pairs (97.85%; prompt-cluster 95% CI [96.09, 99.22]%). At step 10, the fraction is 301/512 (58.79%; [52.34, 65.04]%). All 512 pairs have smaller final distances at either late step; the joint reversal additionally requires a larger immediate distance (Figure 2; Appendix B.7). The joint observation shows that immediate perturbation norm alone does not order endpoint displacement: the rollout stage must also be considered.

These are finite-rollout comparisons, not direct estimates of a Jacobian. Exact mass, value geometry, remaining integration time, and local scheduler dynamics are not held fixed, so the finding concerns temporal sensitivity of the tested intervention rather than an isolated amplification factor or an all-layer law.

6 Image-Level Semantic Consequences

The preceding experiments establish displacement, but not its semantic consequence. We evaluate layer-12 both-branch structural suppression against the implementation-matched custom no-op using Qwen2.5-VL-7B (Bai et al., 2025) and LLaVA-Next-Mistral-7B (Liu et al., 2024a). Each judge evaluates the same 553 prompts with one outcome-independent seed per prompt, balanced across the four base seeds. The no-op reference matches the primary displacement measure. Relative preservation asks whether prompt-relevant content is maintained; absolute validity asks whether each image satisfies the prompt.

Relative preservation holds in 533/553 pairs for Qwen and 546/553 for LLaVA (Table 5). Absolute no-op/intervention validity is 97.83/96.38% for Qwen and 99.64/98.92% for LLaVA. The high raw agreement coexists with modest chance-corrected agreement: the dominant preservation class makes

Table 5: **No-op-relative cross-family semantic audit.** Preservation rates have Wilson 95% confidence intervals. Both judgments and LPIPS use custom no-op as reference. Q4 contains the 138 highest-displacement pairs; it is a subset of the full audit. Agreement and Cohen's κ concern the relative-preservation decision.

Subset	Qwen (%)	LLaVA (%)	Agreement (%)	κ
All 553 pairs	96.38 [94.48, 97.65]	98.73 [97.41, 99.39]	96.93	0.358
Q4, 138 pairs	94.20 [88.98, 97.03]	99.28 [96.01, 99.87]	94.93	0.212

agreement on rare failures less secure. Different VLM families probe evaluator sensitivity, not independent errors or human-level validity.

Aggregate preservation could be dominated by nearly unchanged images. We therefore rank the exact audited pairs by freshly measured no-op-relative LPIPS. In Q4 (LPIPS 0.183–0.505), Qwen judges preservation in 130/138 pairs and LLaVA in 137/138. High judged preservation is thus not confined to low-displacement pairs. The judges differ on seven Q4 pairs and both reject preservation on one. These descriptive rates do not establish semantic equivalence, and displacement stratification is not a causal test of semantic risk. Two missing LLaVA relative labels were obtained by a focused same-model re-query; neither belongs to Q4. Appendix B.1 documents the supplement and missingness bounds. Native-relative audits of both- and single-branch interventions are reported separately in Appendix B.3.

7 Discussion and Limitations

An importance claim must specify the outcome of interest. For pruning or compression, a small alignment-score change can coexist with substantial same-seed realization change. For attention modification, the deletion comparison distinguishes removal of a contribution from redistribution of its mass. Evaluations should therefore report realization displacement and prompt validity separately, using an implementation-matched no-op and explicit image references. These are evaluation implications; performance of editing or pruning methods must be assessed under their own objectives. Same-prompt displacement is not itself an editing or diversity objective: these require a target edit or budget-matched coverage evaluation. Supplementary diversity studies (Appendices G.4, G.5, and G.6) explore that distinction under their recorded settings, separately from the primary contribution-localization result.

The main dose, timing, and semantic studies use SD3 layer 12 and one sampling schedule; a matched 128-prompt replication extends contribution localization and deletion to layers 8 and 16 (Appendix B.4). Controls match event counts and target families, but not mass and value geometry jointly, leaving structural identity entangled with those properties. The trajectory expansion supplies an interpretive decomposition with explicit assumptions and remainder; propagation constants and Jacobian gains are not estimated. Pair-level timing and displacement-conditioned semantics reuse existing samples rather than form held-out predictions. VLM judgments assess layer-12 hard-suppression pairs against custom no-op; they do not validate every dose or deletion condition, and human assessment remains an important complement. Model-specific scope studies retain their own configurations and provenance status (Appendix C); observations with unresolved supplementary implementation identity are exploratory.

8 Conclusion

In the tested SD3 setting, dynamically selected structural text recipients carry nearly all of the all-sink visual displacement magnitude. Deleting their value contributions retains most of this effect without redistributing attention mass. Dose and timing experiments organize its trajectory-level consequences, while two VLM families find high semantic preservation even among the largest audited image changes. Influence on visual realization is therefore distinct from judged prompt semantics. Attention mass can locate candidates for intervention, but understanding their function requires measuring the consequences of the selected contributions.

AI USE STATEMENT

In this work, we used generative AI tools for grammar correction and rewriting. We have not used generative AI tools for generating synthetic numerical results to substitute for measured experimental outcomes, and AI-assisted recruitment or evaluation of human participants are not applicable to this work. Additionally, we used generative AI tools for literature discovery, manuscript organization and language editing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Jeongin Bae, Baeseong Park, Gunho Park, Minsub Kim, Joonhyung Lee, Junhee Yoo, Sunghyeon Woo, Jiwon Ryu, Se Jung Kwon, and Dongsoo Lee. Affine-scaled attention: Towards flexible and stable transformer attention. *arXiv preprint arXiv:2602.23057*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

Jakub Binkowski, Kamil Adamczewski, and Tomasz Kajdanowicz. Attention sinks as internal signals for hallucination detection in large language models. *arXiv preprint arXiv:2604.10697*, 2026.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4792–4808, 2024.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Harry Jake Cunningham and Nicola Muca Cirone. Contribution weights: A geometrical analysis of self-attention transformers, 2026. URL https://arxiv.org/abs/2606.07604.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Bo Gao, Michael W Spratling, and Letizia Gionfrida. Softplus attention with re-weighting boosts length extrapolation in large language models. *arXiv preprint arXiv:2501.13428*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025.

Woojung Han, Dayun Ju, Seong Jae Hwang, et al. Rare text semantics were always there in your diffusion transformer. *Advances in Neural Information Processing Systems*, 38:175066–175125, 2026.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

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

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7462–7471, 2023.

Nicholas Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. *Advances in neural information processing systems*, 38:56557–56595, 2026.

Guy Kaplan, Michael Toker, Yuval Reif, Yonatan Belinkov, and Roy Schwartz. Follow the flow: On information flow across textual tokens in text-to-image models. *arXiv preprint arXiv:2504.01137*, 2025.

Guandong Li. Dual-channel attention guidance for training-free image editing control in diffusion transformers. *arXiv preprint arXiv:2602.18022*, 2026.

Siquan Li, Kaiqi Jiang, Jiacheng Sun, and Tianyang Hu. The structural origin of attention sink: Variance discrepancy, super neurons, and dimension disparity. *arXiv preprint arXiv:2605.06611*, 2026.

Elon Litman and Gabe Guo. You need better attention priors. *arXiv preprint arXiv:2601.15380*, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Perez-Rua, and Jürgen Schmidhuber. Faster diffusion via temporal attention decomposition. *arXiv preprint arXiv:2404.02747*, 2024b.

Jingkun Liu, Yisong Yue, Max Welling, and Yue Song. Krause synchronization transformers. *arXiv preprint arXiv:2602.11534*, 2026.

Zhengyao Lv, Tianlin Pan, Chenyang Si, Zhaoxi Chen, Wangmeng Zuo, Ziwei Liu, and Kwan-Yee K Wong. Rethinking cross-modal interaction in multimodal diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5934–5943, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Seunghun Oh and Unsang Park. Attention frequency modulation: Training-free spectral modulation of diffusion cross-attention. *arXiv preprint arXiv:2603.28114*, 2026.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. *arXiv preprint arXiv:2510.06477*, 2025a.

Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. *arXiv preprint arXiv:2510.06477*, 2025b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Joonghyuk Shin, Alchan Hwang, Yujin Kim, Daneul Kim, and Jaesik Park. Exploring multi-modal diffusion transformers for enhanced prompt-based image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19492–19502, 2025.

Jaewon Sok, Jewon Yeom, Seonghyeon Park, Jeongjae Park, and Taesup Kim. Garbage attention in large language models: Bos sink heads and sink-aware pruning. *arXiv preprint arXiv:2601.06787*, 2026.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks, 2026. URL https://arxiv.org/abs/2603.05498.

Michael Toker, Ido Galil, Hadas Orgad, Rinon Gal, Yoad Tewel, Gal Chechik, and Yonatan Belinkov. Padding tone: A mechanistic analysis of padding tokens in t2i models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7618–7632, 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Fangzheng Wu and Brian Summa. Attention sinks in diffusion transformers: A causal analysis. *arXiv preprint arXiv:2605.09313*, 2026.

Xiaoyu Wu, Yifei Wang, Tsu-Jui Fu, Liang-Chieh Chen, Zhe Gan, and Chen Wei. Taming outlier tokens in diffusion transformers. *arXiv preprint arXiv:2605.05206*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Zihou Zhang, Zheyong Xie, Li Zhong, Haifeng Liu, Yao Hu, and Shaosheng Cao. One token is enough: Improving diffusion language models with a sink token. *arXiv preprint arXiv:2601.19657*, 2026.

Zayd MK Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. Softpick: No attention sink, no massive activations with rectified softmax. *arXiv preprint arXiv:2504.20966*, 2025.

A Evidence Register

Appendix guide. Appendix B documents the primary study and its paired audits; Appendix C specifies configurations; Appendix D expands the literature context; and Appendix E gives the trajectory derivation. Supplementary analyses are grouped into mechanism diagnostics (Appendix F), image and trajectory consequences (Appendix G), and comparative measurements (Appendix H). The register below specifies their execution-verification status.

Study role, exploratory design, and execution verification are separate attributes. Table 6 records the checks available for each group. P is the primary SD3 study; supplementary experiments retain their own configurations, controls, and measurement definitions. Entries with unresolved execution identity do not establish mechanism, robustness, or cross-model replication. Reusing the 553 prompts across settings does not create independent prompt samples. Effects and semantic judgments are never pooled across groups.

The supplementary family label "CLIP padding" and component label "CLIP PAD only" belong to separate analyses. Their recorded raw LPIPS/effect values are 0.1486/90.2% and 0.1473/89.1%, respectively. These are not rounding variants. The component analysis uses finer EOT/PAD classification; equality of event sets and run provenance has not been established. We preserve both records without merging them or attributing their difference to a single cause.

B Branch-Correct CFG Revalidation

We revalidate the central SD3 suppression comparison using 553 prompts, four base seeds (1000, 2000, 3000, 4000), and nine conditions, yielding 19,908 complete image-metric records. This run uses SD3-medium in float16, 1024×1024 generation, 20 inference steps, CFG scale 7.5, layer 12, top-1 incoming-mass selection per head, and intervention at step 0 only. The inactive processor remains in observe mode. LPIPS (AlexNet) is evaluated at 256×256; CLIP-T uses ViT-B/32. The conditional and unconditional branches use their own token-layout maps. Recorded hashes for the six generation-source files match the audited local sources. This does not recover the unmatched supplementary processor snapshot or retroactively verify its branch semantics.

The corresponding main-text results are reported in Table 1.

The native/no-op LPIPS floor is 0.0357. Structural suppression exceeds the structural-position control in direct no-op distance by 0.0900 [0.0862, 0.0938]; the operational control-adjusted ratio is 92.40% [90.45, 94.36]. Controls match selected-event counts (49,875 conditional and 49,333 unconditional events) and always choose a different recipient, but do not match the replacement recipient's mass or value geometry. Selected event labels have zero recorded branch-map mismatches. In single-branch conditions, the untouched branch's first-step output difference from no-op is exactly zero. The unconditional-minus-conditional final LPIPS-magnitude difference is 0.0031 [-0.0007, 0.0071], not a resolved ordering. Single-branch effects are not additive shares of the two-branch rollout.

Structural suppression changes mean CLIP-T relative to no-op by +0.000323 [-0.000147, +0.000804]. This is no detected mean decline, not an equivalence test or evidence of per-image semantic validity. We therefore evaluate the new images separately rather than reuse earlier VLM judgments.

B.1 No-op-Relative Semantic Audit

Figure 1 uses the primary SD3 layer-12, step-0, head-wise top-1 protocol with separate conditional and unconditional token maps. The dining-table and frisbee/couch examples occupy fixed rank quantiles 0.875 and 0.375 of direct-no-op LPIPS among the 553 audited pairs, without semantic-success filtering. Their distances are 0.2504 and 0.0890, respectively. These illustrative examples are not additional audit samples.

The primary semantic audit uses the same 553 prompt–seed identities as the both-branch native-relative audit, replacing the reference image with custom no-op and obtaining new judgments from both VLM families. This produces 1,106 pair–judge records, not an additional independent prompt sample. Image hashes and frozen manifest identities link both judges to the same pairs. LPIPS (AlexNet, evaluation resolution 256×256) is freshly computed against these no-op images. Rank quartiles break ties by prompt index and seed; Q1 has 139 pairs and Q2–Q4 have 138 each.

Two LLaVA responses omitted the relative-preservation boolean. We retained the original responses and all other labels and queried the same model on the same images with a missing-field-only rubric. Both replies explicitly returned true. The rubric asks about preservation of prompt-relevant categories, counts, relations, and attributes, rather than absolute prompt validity; it permits adding a prompt-required object absent from the reference. These are model-generated supplements, not hu-

Table 6: **Study roles and execution-verification status.** Verification refers to the available records, not merely to current code. Supplementary table captions specify each measurement's own reference and scale.

Group	Intervention and reference	Available checks / unresolved points	Evidence role
P: primary SD3	Branch-specific selection; direct no-op LPIPS and VLM; supplementary native-relative audit	Config/source hashes, branch/event checks and frozen audit pairs; no-op audit includes freshly measured LPIPS.	Main location, operation, timing and consequence results; L8/L16 localization replication
Supplementary SD3	Recorded family, layer, operation and sampling variants; native/no-op or latent references as captioned	Supplementary processor snapshot unmatched; earlier branch semantics not retrospectively certified	Supplementary observations; not primary robustness or mechanism support
Supplementary FLUX	Model-specific processor; schnell, four steps; per-block native/no-op comparisons	Inventory found configs/results, no source snapshots; intervention configs retain seed 4000, requiring shard-level reconciliation	Candidate scope observations; execution identity not yet certified
Supplementary U-Net / SD3.5	Architecture-specific processors and image-pair audits	Independent execution-identity verification incomplete; SD3.5 additionally floor-limited	Candidate boundary observations; not certified by P

man corrections or independent judgments. Their protocol, raw replies, original-response hashes, and pair identities are retained. Both pairs fall outside Q4. Treating the two originally missing labels as false/true bounds overall LLaVA preservation at [98.37, 98.73]%; these are missingness bounds, not confidence intervals. Qwen has no supplemented fields. Overall major-error rates are 20/553 (3.62%) for Qwen and 7/553 (1.27%) for LLaVA; in Q4 they are 8/138 (5.80%) and 1/138 (0.72%). The two judges both report preservation on 531 pairs, both reject it on five, and disagree on 17. Table 5 reports the primary rates and descriptive, unadjusted Wilson intervals.

B.2 Supplementary Native-Relative Semantic Audit

Each condition is evaluated on all 553 distinct prompts with one seed per prompt, assigned independently of outcomes and balanced across the four base seeds (139/138/138/138). The same pairs are used across conditions and judges. Qwen2.5-VL-7B and LLaVA-Next-Mistral-7B provide 3,318 judgments in total.

These supplementary judgments use native baseline, unlike the primary no-op-relative audit in Table 5; the rates below are kept separate by reference.

For the primary both-branch condition, the relative-preservation counts are 536/553 for Qwen (Wilson 95% CI [95.13, 98.07]%) and 543/553 for LLaVA ([96.70, 99.01]%). Absolute baseline/intervention prompt-validity rates are 97.29/96.93% for Qwen and 99.28/98.37% for LLaVA; major-error rates are 3.07% and 1.81%, respectively. We do not interpret the aggregate rates as a paired non-inferiority test or a guarantee of semantic preservation. Class imbalance contributes to high raw agreement, and agreement on rare errors is less secure than the overall agreement percentage suggests.

Seven LLaVA responses omitted the relative-preservation boolean (one for both branches and three for each single-branch condition). We retained the original responses and other labels and re-queried the same model on the same images using a focused missing-field rubric; four supplemental labels were true and three false. Some explanations still conflate shared baseline invalidity with relative semantic change. Treating all originally missing labels as false/true bounds LLaVA preservation at [98.01, 98.19]%, [98.19, 98.73]%, and [98.19, 98.73]%, respectively. These are missingness sensitivity bounds, not confidence intervals or manual corrections.

Table 7: **Semantic preservation by displacement quartile.** Both LPIPS and judgments compare intervention with native baseline. Entries are percentages with Wilson 95% CIs. Q1 is the lowest-displacement quartile; Q4 the highest. Bins are condition-specific, not shared displacement thresholds across branch conditions.

Structural intervention	Quartile	Qwen	LLaVA
Both branches	Q1	97.84 [93.85, 99.26]	99.28 [96.04, 99.87]
	Q2	97.10 [92.78, 98.87]	97.83 [93.80, 99.26]
	Q3	97.83 [93.80, 99.26]	97.83 [93.80, 99.26]
	Q4	94.93 [89.90, 97.52]	97.83 [93.80, 99.26]
Conditional only	Q1	98.56 [94.91, 99.60]	99.28 [96.04, 99.87]
	Q2	97.83 [93.80, 99.26]	99.28 [96.01, 99.87]
	Q3	94.20 [88.98, 97.03]	97.10 [92.78, 98.87]
	Q4	92.75 [87.17, 96.02]	98.55 [94.87, 99.60]
Unconditional only	Q1	98.56 [94.91, 99.60]	100.00 [97.31, 100.00]
	Q2	96.38 [91.80, 98.44]	97.83 [93.80, 99.26]
	Q3	97.10 [92.78, 98.87]	97.10 [92.78, 98.87]
	Q4	93.48 [88.07, 96.53]	98.55 [94.87, 99.60]

B.3 Supplementary Native-Relative Displacement Audit

We reuse the frozen 553 native-baseline/intervention pairs per condition, without generating images or obtaining new judgments. The join checks the formal configuration hash, manifest hashes, exact judgment/manifest fields, judge model identifiers, and image hashes. There are 3,318 distinct pair–judge records. The existing LPIPS CSV does not embed image hashes: the checks verify the current images against the frozen audit, but do not independently certify the history of metric computation.

Within each condition, pairs are sorted by native-baseline-relative LPIPS and split into rank quartiles, using prompt index and seed to break ties. Both judges use identical bins. Q1 contains 139 distinct prompts and Q2–Q4 138 each, with one audited seed per prompt. These bins describe realized displacement; they do not randomize displacement or identify its causal effect on semantic errors. Table 7 reports all bins, with descriptive, unadjusted Wilson 95% intervals.

For the primary both-branch Q4 subset, major-error rates are 5.07% for Qwen and 2.17% for LLaVA. Judge agreement in Q4 is 97.10% for both branches, 92.75% for conditional-only, and 94.93% for unconditional-only. The single-branch Q4 differences between judges reinforce that high overall agreement is not evidence of interchangeable error judgments. The seven documented missing-field supplements are retained, not counted as new independent judgments. These results concern hard structural suppression, not deletion or every point on the dose curve.

B.4 Matched Cross-Layer Replication

We repeat the branch-correct protocol at layers 8 and 16 on the same 128 prompts sampled from the 553-prompt set with fixed seed 20260921, using all four base seeds. Generation remains at 1024×1024, 20 steps, CFG 7.5, and step-0 head-wise top-1 selection with branch-specific token maps. Each layer has 3,584 images: 128 prompts $\times$ four seeds $\times$ seven arms (native baseline, no-op, all-sink suppression, structural suppression, structural-position control, random-text control, and structural deletion). All distances use that layer's custom no-op. Confidence intervals resample prompts with their four seeds and paired conditions together (5,000 bootstrap replicates); ratios are ratios of aggregate mean distances.

Structural suppression exceeds both controls at each layer, and deletion retains most of its displacement magnitude. These findings extend contribution localization and the deletion diagnostic beyond layer 12, while the 85% versus 99% structural/all-sink ratios show depth dependence. They do not establish a universal recovery ratio or a matched layer ranking against the full 553-prompt layer-12 study. This replication does not include a layer-8/16 semantic audit or a top-k sweep.

Table 8: **Branch-correct cross-layer replication.** Direct no-op-relative LPIPS and effect-magnitude ratios on 128 prompts × four seeds. Brackets are paired prompt-cluster bootstrap 95% CIs. Deletion removes selected contributions without renormalization.

Measurement	Layer 8	Layer 16
Native baseline LPIPS	0.0372	0.0328
All-sink suppression LPIPS	0.1048	0.1289
Structural suppression LPIPS	0.0891	0.1280
Structural-position control LPIPS	0.0367	0.0380
Random-text control LPIPS	0.0378	0.0352
Structural deletion LPIPS	0.0801	0.1125
Structural/all-sink (%)	85.05 [80.40, 89.75]	99.34 [97.54, 101.15]
Deletion/structural suppression (%)	89.88 [85.85, 94.42]	87.88 [83.11, 92.96]
Structural minus position control	0.0524 [0.0461, 0.0592]	0.0900 [0.0823, 0.0982]
Structural minus random control	0.0513 [0.0453, 0.0575]	0.0929 [0.0854, 0.1009]

B.5 Branch-Correct Contribution Deletion

We separately repeat contribution deletion and image-query value-zeroing on 553 prompts and four seeds under the same corrected protocol. Baseline, no-op, all-sink renormalizing and structural renormalizing anchors each match the preceding formal run pixel-for-pixel on all 2,212 pairs. Selected events match exactly across the three operations within each recipient family.

The corresponding main-text results are reported in Table 2.

For both families, no-renorm and value-zero images are identical on all 2,212 pairs. This verifies their local algebraic equivalence and implementation, not two independent sources of causal evidence. Structural deletion retains most, but measurably less than all, of the renormalizing displacement magnitude. The ratio gap is not an additive causal share attributable to redistribution. Its mean CLIP-T change relative to renorm is -0.000268 [-0.000749, +0.000215]; the VLM results above apply to renorm images and must not be transferred to deletion images.

These experiments revalidate hard suppression, two controls, CFG branch isolation, new-image semantic judgments, and contribution deletion. The extension below adds the full structural dose grid and a single-layer timing comparison. Supplementary studies retain their original protocol qualifications and are not silently pooled with the corrected evidence.

B.6 Branch-Correct Dose and Timing Extensions

Both extensions use the same recorded formal reference configuration: SD3-medium in float16, 1024×1024 generation, 20 inference steps, CFG 7.5, an empty negative prompt, layer 12, and per-head top-1 selection. Interventions use each CFG branch's own token map, with observation-only processing outside the selected steps. Seeds are 1000, 2000, 3000, and 4000. Dose uses all 553 prompts; timing uses 128 prompt indices sampled with fixed design seed 20260921, independently of measured outcomes. Reported distances are direct LPIPS to custom-noop, not floor-subtracted differences. Intervals are 95% paired prompt-cluster bootstrap intervals (10,000 resamples), retaining all four seeds within each sampled prompt. They are exploratory and not adjusted for multiple comparisons.

The corresponding main-text results are reported in Table 3.

The extension generates $\gamma = .25, .75$ and reuses formal-reference $\gamma = 0, .5$ and no-op images (Table 3). All four adjacent-dose LPIPS differences are positive: 0.02741 [0.02581, 0.02904], 0.02400 [0.02253, 0.02550], 0.02576 [0.02402, 0.02747], and 0.05424 [0.05143, 0.05707]. The mean response is graded but not linear. Monotonicity holds for 494/553 (89.3%) four-seed prompt averages and 1,427/2,212 (64.5%) individual prompt–seed pairs; it is not an image-level guarantee. Mean CLIP-T deltas relative to no-op have intervals containing zero at all four nontrivial doses. This is not an equivalence test, and the hard-suppression VLM audit does not validate semantic preservation throughout the dose curve.

The corresponding main-text results are reported in Table 4.

Table 9: **Temporal reversal within paired rollouts.** Each comparison includes 512 pairs (128 prompts, four seeds). Intervals bootstrap 128 prompt clusters with seeds retained (10,000 resamples; 95%, unadjusted). All late endpoint distances are smaller; reversal additionally requires a larger late immediate distance.

Recipients	Late vs early step	Reversals	Rate (%) [95% CI]
Structural	10 vs 0	301/512	58.79 [52.34, 65.04]
Structural	15 vs 0	501/512	97.85 [96.09, 99.22]
All sinks	10 vs 0	314/512	61.33 [55.27, 67.19]
All sinks	15 vs 0	504/512	98.44 [97.07, 99.41]

Relative to step-0 structural suppression, final LPIPS is 5.87% [5.31, 6.46] at step 5, 0.53% [0.48, 0.59] at step 10, and 0.13% [0.11, 0.15] at step 15 (Table 4). Paired step-0-minus-late differences all exclude zero. The early three-step arm reaches 105.04% [102.58, 107.57] of the single-step magnitude; its larger budget prevents an additive attribution of the increase to particular steps.

Late effects are not explained by absent selections: all-sink arms select 24 events per branch at every tested single step. Their mean summed incoming mass across selected head-wise events remains between 2.13 and 2.87 across steps and branches (a sum over heads, not a probability). Structural arms retain roughly 20–23 selected events per branch. The arms are not matched for value geometry or exact mass, so this is evidence against selection inactivity, not isolation of timing from every other mediator.

Trajectory summaries show zero pre-intervention divergence. For structural interventions at steps 0, 10, and 15, mean immediate post-step latent L2 is 0.78, 0.87, and 1.34, respectively, whereas final L2 is 141.28, 8.44, and 4.19. Thus a larger immediate late-step perturbation need not yield a larger final displacement. These are finite-perturbation rollout comparisons, not shared-state Jacobian measurements or proof of a universal amplification law. The timing conclusion is restricted to the tested layer and schedule.

B.7 Paired Temporal Ordering Reversals

We compare step 0 with steps 10 and 15 on the same 128 prompts and four seeds, using the complete recorded trajectories. Immediate distance is the latent L2 distance to custom-noop immediately after the intervened step; endpoint distance is measured after the final step. A reversal requires both a strictly larger late immediate distance and a strictly smaller late endpoint distance within the same prompt–seed pair. Step comparisons and the structural-primary/all-sink-supporting distinction were fixed before this pair-level analysis, after inspecting the earlier aggregate results; these are not held-out predictions.

For structural recipients, mean late-minus-early immediate L2 differences are +0.0875 [0.0506, 0.1255] at step 10 and +0.5615 [0.5186, 0.6043] at step 15. Corresponding endpoint differences are -132.84 [-140.81, -125.13] and -137.09 [-145.34, -129.13]. No exact ties or opposite-direction reversals occur in the observed pairs; this is not a population zero-probability claim. The results extend the mean ordering to individual rollouts, particularly at step 15. They do not isolate remaining integration time from scheduler dynamics, mass, or state-dependent value geometry, and the distance ratios are not Jacobian gains.

C Implementation Configurations

Models and access paths. The main SD3 experiments use `stabilityai/stable-diffusion-3-medium-diffusers` loaded through Hugging Face `diffusers` as `StableDiffusion3Pipeline` with `torch_dtype=float16`. SD3.5 anatomy and boundary experiments use `stabilityai/stable-diffusion-3.5-medium` with `bfloat16`. FLUX experiments use `black-forest-labs/FLUX.1-schnell` as `FluxPipeline` with `bfloat16`. U-Net cross-attention boundary experiments use `runwayml/stable-diffusion-v1-5`, `sd2-community/stable-diffusion-2-1-base`, and `stabilityai/st`

able-diffusion-xl-base-1.0. All model loading is via diffusers; local cache mode is used for models already present on the compute server.

Sampling. The primary branch-correct SD3 studies use 1024×1024 generation, 20 steps, and CFG 7.5, as specified in Section 3 and Appendix B. The following settings instead describe the supplementary supporting studies. Unless individually overridden, supplementary SD3 and SD3.5 runs use the default SD3 flow-matching scheduler in diffusers, 20 inference steps, classifier-free guidance (CFG) scale 7.5, and 512×512 output resolution. FLUX.1-schnell uses 4 inference steps, CFG scale 0.0, 512×512 resolution, maximum text sequence length 512, top-$k = 5$ recipient selection, and the first 5% of the denoising window. Stable Diffusion 1.5 (SD1.5), Stable Diffusion 2.1 (SD2.1), and Stable Diffusion XL (SDXL) use the default diffusers U-Net pipelines, 20 inference steps, CFG scale 7.5 for SD1.5/SD2.1 and 5.0 for SDXL, 512×512 resolution, top-$k = 1$, and the first 5% denoising window. The base protocol does not use negative prompts. The sampler/CFG robustness appendix varies CFG, step count, and swaps to a Heun flow-matching scheduler for SD3. These measurements are not pooled with the primary branch-correct studies.

Seeds and prompts. P uses 553 GenEval prompts and four base seeds $\{1000, 2000, 3000, 4000\}$, with paired conditions sharing prompt and initial noise. Its timing, cross-layer, and VLM subsets are specified separately. Supplementary designs also list four seeds, but this is not a completed audit of every supplementary pair. In particular, FLUX intervention configs retain seed 4000; shard-level seed assignment and image pairing remain to be reconciled. Supplementary sample counts are reported as recorded, not inferred from the surviving config alone.

Metrics. LPIPS is computed with the AlexNet backbone (lpips.LPIPS(net="alex")). CLIP-T uses openai/clip-vit-base-patch32 through Hugging Face transformers. The displacement–dispersion audit and mode-diversity pilot use features from the self-supervised visual encoder DINOv2 (facebook/dinov2-base). The primary corrected-image VLM audit uses Qwen2.5-VL-7B-Instruct and LLaVA-Next (llava-hf/llava-v1.6-mistral-7b-hf) on 553 no-op/suppression pairs for the primary both-branch condition. A separate native-relative audit covers 553 pairs for each of three branch conditions; neither its labels nor its displacement bins are pooled with the no-op-relative audit. Separately, supplementary carrier audits use two Qwen rubrics on 553 pairs per carrier and a 256-pair LLaVA cross-family spot-check on SD3/FLUX carriers. The phase-aligned semantic trajectory audit decodes intermediate SD3 latents at normalized denoising phases and computes CLIP-T, LPIPS, and DINOv2 similarity against the same-prompt, same-seed baseline decoded at the same phase.

Intervention implementation. SD3 interventions are implemented as a drop-in custom attention processor for the joint-attention blocks, matching the JointAttnProcessor2_0 interface. The processor computes dynamic top-k incoming-mass recipients per head from the unmodified image-query attention distribution, applies the selected pre-softmax logit mask or diagnostic value/no-renormalization operation, and then restores the original pipeline after generation. FLUX and U-Net experiments use architecture-specific processors with the same causal target: image-query rows attending to dynamically selected recipient positions.

Hardware. Generation and audits are sharded over four NVIDIA RTX A6000-class GPUs using Slurm array jobs with one GPU per task. The large formal runs are embarrassingly parallel over seeds and, for FLUX block probes, over block/seed shards.

D Extended Related Work

Attention sinks beyond language modeling. Language-model sink studies motivate a structural prior for high attention mass. Some works interpret sinks through long-context anchoring, softmax normalization, and over-mixing avoidance (Xiao et al., 2024; Gu et al., 2025; Barbero et al., 2025), while others connect sink scores to value-vector geometry (Binkowski et al., 2026), tail-spectrum or massive residual-stream components (Cancedda, 2024; Queipo-de Llano et al., 2025b), separate massive activations from attention sinks (Sun et al., 2026), or trace emergence to value aggregation and super-neuron amplification (Li et al., 2026). Concurrent large language model (LLM) analyses

further argue that attention-weight-only interpretation neglects value-vector geometry (Cunningham & Cirone, 2026), providing a parallel theoretical motivation for our routing-leverage decomposition. Architectural and mitigation work further shows that sink behavior can be removed or exploited: Softpick replaces sum-to-one softmax with a rectified attention rule and eliminates sinks/massive activations in language models (Zuhri et al., 2025), while BOS-sink pruning identifies some high-sink heads as redundant dumping grounds (Sok et al., 2026). Diffusion-language models extend the structural interpretation beyond autoregressive generation: a dedicated low-information sink token can stabilize moving sinks during denoising (Zhang et al., 2026). In video diffusion, Deep Forcing treats sink positions as useful cache structure for long-range temporal consistency (Yi et al., 2025); MotionStream further combines attention-sink chunks with rolling key–value caches to maintain stability under fixed-context, infinite-horizon video generation (Queipo-de Llano et al., 2025a). Our findings extend this "sink as routing resource" perspective from temporal continuity in video diffusion to same-prompt visual realization in T2I diffusion. The cited studies motivate testing the function of high-attention positions before removing them. Our primary SD3 study addresses their contribution to visual realization: deleting dynamically selected structural contributions changes the generated image even without redistributing their attention mass.

Diffusion attention control and MM-DiT internals. Diffusion editing and guidance methods often manipulate attention maps or attention logits to improve controllability (Hertz et al., 2022; Chefer et al., 2023; Hong et al., 2023; Ahn et al., 2024; Oh & Park, 2026). Prompt-to-Prompt preserves and edits cross-attention maps to localize token-level prompt edits, while null-text inversion optimizes unconditional embeddings to enable faithful inversion and subsequent source-image editing (Hertz et al., 2022; Mokady et al., 2023). These methods use attention as an editing interface conditioned on a desired prompt/image change. Our fixed-prompt intervention instead treats attention as a causal diagnostic: we suppress dynamically selected recipients under the same prompt and seed to identify routing variables that change visual realization without asking the model to perform a semantic edit. T-GATE identifies temporal phases in which attention is especially important (Liu et al., 2024b), Padding Tone shows that special/padding positions can carry nontrivial model-specific signal (Toker et al., 2025), and Follow the Flow studies information propagation through denoising (Kaplan et al., 2025). REPA studies a complementary depth-axis version of early specialization, showing that representation alignment for diffusion transformers can be supplied through early blocks rather than imposed uniformly through the network (Yu et al., 2024). In MM-DiT models, TACA rebalances cross-modal attention under visual/text token imbalance (Lv et al., 2025), ToRA amplifies text-side representations to surface rare semantics (Han et al., 2026), DCAG rescales key and value channels to control image-editing intensity (Li, 2026), and MM-DiT editing work decomposes joint attention into blockwise submatrices (Shin et al., 2025). DCAG is especially complementary because its value-channel result and our routing-leverage decomposition both argue that value geometry matters beyond raw attention mass. The distinction is intervention granularity and task: DCAG globally scales key/value channels for source-image editing, whereas our experiments identify dynamic recipient families that route same-prompt generation from noise. Our experiments use the same general attention object but target a different axis: dynamic recipient states that steer visual realization under an unchanged prompt.

Registers, outliers, and sink mitigation. Register and outlier-token work shows that special slots or high-norm tokens can organize transformer computation (Darcet et al., 2024; Jiang et al., 2026; Wu et al., 2026). Other work treats sinks as pathologies to be reduced through architectural priors, alternative attention mechanisms, or topological regularization (Litman & Guo, 2026; Bae et al., 2026; Gao et al., 2025; Liu et al., 2026). Our results suggest a more role-specific view in generative models: sinks can be pathological in some objectives, but dynamically selected structural recipients can expose high-impact contributions in the primary SD3 study. The supplementary fixed-slot diagnostic does not isolate alignment from mass concentration or intervention strength.

E Trajectory Expansion with Explicit Remainder

The local identities in Section 5 hold at a fixed state. The following result states the additional assumptions required for a continuous-time trajectory expansion. It is not a bound on the discrete sampler used in the experiments.

Proposition 1 (Routing Leverage Expansion with Remainder)**.** *Consider a smooth one-parameter intervention family on the denoising vector field, $\dot{z}_\eta(t) = f_\eta(z_\eta(t), t)$, with shared initial latent $z_\eta(0) = z_0(0)$ and $\eta \in [0, 1]$. Define the first variation $g_S(z, t) = \partial_\eta f_\eta(z, t)|_{\eta=0}$. Let $A(t) = \partial_z f_0(z_0(t), t)$, and let $\Phi(t, s)$ be the state transition matrix satisfying $\partial_t \Phi(t, s) = A(t)\Phi(t, s)$ and $\Phi(s, s) = I$. Assume both trajectories and the line segments joining them remain in a tube around the unperturbed trajectory throughout the interval, and that on this tube $\|\partial_z f_0\| \le L$, $\|\partial_z^2 f_0\| \le H$, $\|g_S\| \le G$, $\|\partial_z g_S\| \le B$, and*

$$\|f_\eta(z, t) - f_0(z, t) - \eta g_S(z, t)\| \le \frac{K}{2}\eta^2.$$

Let $D = G + K/2$. Then

$$z_\eta(1) - z_0(1) = \eta \int_0^1 \Phi(1, t) g_S(z_0(t), t)\, dt + R_2(\eta), \tag{7}$$

with the explicit second-order remainder bound

$$\|R_2(\eta)\| \le \eta^2 e^L \left(BDe^L + \frac{H}{2} D^2 e^{2L} + \frac{K}{2} \right). \tag{8}$$

At an attention intervention site, the finite local perturbations are given exactly by Eq. equation 4, equation 5, or equation 6. The first variation entering g_S is the derivative of the chosen local perturbation path at $\eta = 0$. For the dose path with $\eta = 1 - \gamma$, this derivative is

$$\partial_\eta \Delta o_i^{\gamma=1-\eta}(S)\big|_{\eta=0} = m_{i,S}(1 - m_{i,S}) W_O(\bar{v}_{S^c,i} - \bar{v}_{S,i}).$$

After multiplying by the local downstream Jacobian $J_{\mathrm{down}}(t)$ from the attention output to the latent-velocity update, the first-order term in Eq. equation 7 is large only when three factors align: the dose-path mass gate $m_{i,S}(t)(1 - m_{i,S}(t))$ (or the finite hard-mask mass $m_{i,S}(t)$ in Eq. equation 4), projected value contrast or contribution, and downstream flow sensitivity $\Phi(1, t) J_{\mathrm{down}}(t)$.

The local identities hold for fixed local queries, keys, values, and selected set. The trajectory expansion additionally requires the stated smoothness assumptions. Dynamic top-k selection and family gating need not remain smooth when recipient ranks switch; applying the expansion locally therefore requires a stable selected set (for example, a positive selection margin throughout the perturbation neighborhood). We do not prove a switching-event extension or a discrete-sampler error bound. Nor do we estimate the remainder constants: the bound does not ensure a small remainder for full-strength suppression.

Proof. Let $y(t) = \int_0^t \Phi(t, s) g_S(z_0(s), s)\, ds$, so $\dot{y} = A(t)y + g_S(z_0(t), t)$ and $y(0) = 0$. Set $\delta(t) = z_\eta(t) - z_0(t)$. Gronwall's inequality gives $\|\delta(t)\| \le \eta D e^L$ for $t \in [0, 1]$. Now write $r(t) = \delta(t) - \eta y(t)$. Taylor's theorem with remainder gives

$$\dot{r} = A(t)r + \rho_f(t) + \eta\rho_g(t) + \rho_\eta(t), \qquad \|\rho_f(t)\| \le \frac{H}{2}\|\delta(t)\|^2, \quad \|\rho_g(t)\| \le B\|\delta(t)\|, \quad \|\rho_\eta(t)\| \le \frac{K}{2}\eta^2.$$

Applying variation of constants and the bound $\|\Phi(t, s)\| \le e^{L(t-s)}$ yields

$$\|r(1)\| \le e^L \left(\frac{H}{2}\eta^2 D^2 e^{2L} + \eta^2 BDe^L + \frac{K}{2}\eta^2 \right).$$

Since $R_2(\eta) = r(1)$, Eq. equation 8 follows. □

F Supplementary Mechanism Diagnostics

The following supplementary analyses retain their original measurement definitions and are grouped by diagnostic target. Their execution status is given in Table 6.

F.1 ROUTING-LEVERAGE DECOMPOSITION DETAILS

This appendix gives additional detail for Table 27. The value-side probe runs the SD3 attention processor in observe-only mode and records, for layers 8/12/16/20 in the first denoising window, the top-1 incoming-mass recipient for each attention head. Each recipient is mapped to a family using the SD3 text layout: CLIP-fused structural positions, padding positions, content text positions, or image positions. For a selected top-1 recipient r, we report the value-side leverage proxy

$$a_r \left\| W_O \left(v_r - \bar{v}_{r^c} \right) \right\|_2 ,$$

where a_r is incoming attention mass and $\bar{v}_{r^c}$ denotes the attention-weighted value average over non-selected recipients. The downstream sensitivity proxy is measured separately by suppressing one recipient family in one layer during the first denoising window and recording the final-latent L2 displacement relative to the paired baseline.

Table 10: **Dynamic top-1 sink selection and value-side leverage by layer.** Top-1 columns are percentages of dynamic top-1 sink events in that layer with Wilson 95% confidence intervals. Value leverage is the descriptive mean $a_r \| W_O(v_r - \bar{v}_{r^c}) \|_2$ over selected top-1 recipients r.

Layer	CLIP top-1 [95% CI]	Pad top-1 [95% CI]	Content top-1 [95% CI]	Image top-1	CLIP value lev.	Pad value lev.	Content value lev.
8	69.0 [68.8, 69.3]	8.5 [8.3, 8.7]	10.0 [9.8, 10.2]	11.4	0.482	0.804	0.249
12	91.8 [91.6, 91.9]	0.6 [0.5, 0.6]	5.6 [5.5, 5.7]	0.04	0.978	0.972	0.305
16	86.4 [86.2, 86.7]	1.8 [1.8, 1.9]	0.01 [0.0, 0.0]	11.7	1.396	1.447	0.497
20	39.5 [39.2, 39.8]	51.7 [51.4, 52.0]	2.7 [2.6, 2.8]	5.7	0.961	0.479	0.212

Table 11: **Downstream sensitivity proxy by layer.** Values are final-latent L2 displacement after single-layer early-window suppression, averaged over 128 prompts and four seeds ($n = 512$ per cell), with bootstrap 95% confidence intervals.

Layer	CLIP structural	Padding	Content	Image-only
8	95.9 [90.9, 100.2]	84.5 [80.0, 88.8]	70.8 [66.7, 75.3]	59.2 [55.7, 62.1]
12	129.8 [123.8, 135.3]	125.6 [119.4, 131.5]	74.8 [71.0, 78.9]	55.5 [52.6, 58.4]
16	125.0 [119.2, 130.7]	124.5 [118.8, 130.0]	55.0 [51.8, 58.1]	56.3 [53.0, 59.5]
20	103.8 [99.1, 108.7]	102.3 [97.4, 107.4]	54.7 [51.5, 57.8]	54.2 [51.0, 57.7]

The supplementary probes record different layer orderings for value leverage, selection frequency and finite-displacement sensitivity. These associations are retained descriptively. Their execution identity is unresolved, so they are not used to explain P's main effect or validate the routing-leverage expansion. Finite rollout distances are not independent Jacobian measurements.

Table 12: **Exploratory in-sample associations for the routing-leverage proxy.** In-sample fits use 5 or 12 condition means sharing prompts/seeds, not held-out prediction. We report Pearson correlation, Spearman rank correlation, and univariate linear R^2. The structurality panel fits LPIPS over the no-op floor across SD3 structural and non-structural matched controls. The layer-family panel fits the single-layer latent sensitivity proxy; there, value contrast is the strongest single factor and adding selection frequency improves the fit.

Panel	Predictor	Pearson r	Spearman ρ	R^2
Structurality controls ($n = 5$)	mask mass	0.950	1.000	0.903
Structurality controls ($n = 5$)	projected value contrast	-0.508	-0.368	0.258
Structurality controls ($n = 5$)	mass $\times$ value	0.968	1.000	0.938
Layer/family sensitivity ($n = 12$)	selection frequency	0.547	0.420	0.299
Layer/family sensitivity ($n = 12$)	projected value contrast	0.826	0.825	0.682
Layer/family sensitivity ($n = 12$)	selection $\times$ value	0.584	0.573	0.341
Layer/family sensitivity ($n = 12$)	selection + value	–	–	0.774

The supplementary in-sample association analysis is reported in Table 12. In the structurality disambiguation probe, where the outcome is LPIPS over the mask-disabled floor across structural and non-structural matched controls, the mass–value product is the best univariate predictor ($R^2 = 0.938$), slightly above mass alone ($R^2 = 0.903$) and far above projected value contrast alone ($R^2 = 0.258$).

This is consistent with a joint role for mass and value geometry, but does not establish a necessary condition from five condition means. In the layer/family sensitivity probe, the outcome is a finite downstream response rather than an independent Jacobian measurement, so the appropriate question is not whether a product including sensitivity predicts itself. Instead, we ask whether selection and value-side contrast jointly organize sensitivity across 12 layer/family cells. Projected value contrast is the strongest single factor ($R^2 = 0.682$), selection frequency alone is weaker ($R^2 = 0.299$), and a two-factor linear model using selection frequency plus value contrast improves the fit ($R^2 = 0.774$). These are exploratory, in-sample fits on small sets of condition means, some of which share identical reported outcomes. Adding a predictor can improve in-sample fit without improving generalization. Moreover, the selection–value product has $R^2 = 0.341$, below value contrast alone. Thus these checks describe associations relevant to the decomposition; they do not validate a full mass–value–downstream product predictor or demonstrate out-of-sample predictive superiority.

F.2 Selection Sanity and Controls

For every intervention family we log the dynamically selected sink recipients before applying the mask. This verifies that the requested intervention is actually acting on the intended recipient family. In the SD3 main run, CLIP structural recipients dominate the layer-12 dynamic top-1 sink distribution, whereas image recipients almost never appear as dynamic sinks. Matched controls preserve the token family while replacing selected sinks with non-sink positions, and random-text controls preserve the number of selected events while relocating the intervention to alternative text positions. Both control families collapse near the mask-disabled floor, showing that neither token identity nor mask budget alone explains the effect.

F.3 Intervention Specification

Renormalizing suppression applies a pre-softmax logit mask to image-query rows. Contribution deletion acts after softmax without renormalization. Let $\mathcal{I}$ be the set of image-query rows and let R be the joint image–text recipient set; r' below is a dummy normalization index over R. For each selected layer/head/step, incoming mass is computed from the unmodified attention probabilities as

$$\alpha_{ir}^{\ell ht} = \frac{\exp(\ell_{ir}^{\ell ht})}{\sum_{r' \in R} \exp(\ell_{ir'}^{\ell ht})}, \qquad m_j^{\ell ht} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \alpha_{ij}^{\ell ht}. \tag{9}$$

The selected dynamic sink recipient is the top-k recipient under $m_j^{\ell ht}$; the main experiments use $k = 1$. This selection is made separately for each head h; selected recipients are masked independently within their own heads, with no cross-head score aggregation or shared mask target. A selector then determines whether this selected recipient belongs to a target family such as CLIP structural, padding, content, image, or all sinks. If selected, the intervention modifies the corresponding image-query logits by -10^4, followed by the ordinary softmax over all remaining recipients. Matched and random controls use the same pre-softmax operation but replace the mask target with a deterministic alternative text position; the dynamic sink recipient used for logging remains the original top-k recipient. The custom no-op condition in the released code runs the identical processor and selector bookkeeping while applying no logit delta; in the paper we call its image-level reference point the mask-disabled floor.

CLIP-fused structural selector. SD3 provides two CLIP text encoders whose 77-token hidden states are fused into a single 77-position CLIP segment before concatenation with the T5 segment. For each prompt, we tokenize with both CLIP tokenizers using max-length padding. At fused CLIP position p, we label the recipient as *padding* if both underlying token ids are padding ids, as *content* if either tokenizer has a non-special content token at p, and as *special* if no side is content and at least one side is a tokenizer special id such as start-of-text or another reserved token. The *CLIP structural* selector is then `text_source = clip_fused` and `token_type` $\in \{$`padding`, `special`$\}$. This definition is applied after dynamic top-k selection; it classifies the recipient selected by image-query attention rather than scanning all structural tokens statically. For the finer component ablation in Appendix F.4, we additionally use each tokenizer's attention mask to distinguish true non-padding EOT positions from padded tail positions that may share the same token id in CLIP-style tokenizers. CLIP-L and CLIP-G are fused along the hidden dimension in SD3, so CLIP-L/CLIP-G compo-

Table 13: **Renormalizing versus no-renormalization interventions.** Renorm masks selected recipients before softmax and redistributes mass; no-renorm deletes the selected contribution after softmax without redistributing mass. Effects are LPIPS over the custom-processor no-op floor; brackets are paired-bootstrap 95% confidence intervals.

Family	Renorm LPIPS−floor	No-renorm LPIPS−floor	No-renorm / renorm	Renorm CLIP Δ	No-renorm CLIP Δ
All sinks	0.1144 [0.1106, 0.1182]	0.1171 [0.1133, 0.1210]	102.4 [99.6, 105.4]	+0.0003 [-0.0002, +0.0008]	+0.0002 [-0.0003, +0.0007]
CLIP structural	0.1121 [0.1084, 0.1159]	0.1072 [0.1035, 0.1108]	95.6 [92.7, 98.6]	+0.0005 [+0.0000, +0.0010]	+0.0003 [-0.0002, +0.0008]
CLIP padding	0.1032 [0.0996, 0.1068]	0.0888 [0.0853, 0.0923]	86.0 [83.0, 89.2]	+0.0003 [-0.0002, +0.0008]	+0.0002 [-0.0003, +0.0006]
Content	0.0225 [0.0203, 0.0247]	0.0229 [0.0208, 0.0251]	101.7 [95.3, 108.9]	+0.0000 [-0.0003, +0.0004]	+0.0001 [-0.0002, +0.0004]
Image	0.0001 [-0.0001, 0.0003]	0.0002 [-0.0000, 0.0005]	169.0 [-846.7, 1079.9]	-0.0001 [-0.0004, +0.0002]	-0.0001 [-0.0004, +0.0002]

Table 14: **Value-side diagnostic.** On 553 prompts × 4 seeds, no-renormalization contribution deletion and direct value-zeroing give identical image-level metrics for the selected image-query contribution. Effects are normalized over the custom no-op floor and the all-sink no-renormalization reference.

Family	No-renorm effect	Value-zero effect	No-renorm CLIP Δ	Value-zero CLIP Δ
All sinks	100.0	100.0	+0.0002	+0.0002
CLIP structural	91.5	91.5	+0.0004	+0.0004
CLIP padding	75.7	75.7	+0.0003	+0.0003

nent ablations select fused recipient positions marked structural by the corresponding underlying tokenizer; they do not remove one half of the fused value vector.

For a fixed selected set, setting its logits to negative infinity is algebraically equivalent to post-softmax zeroing followed by renormalization. Our finite penalty approximates this limit, with saturation checked separately. In contrast, post-softmax deletion without renormalization removes the selected contribution without reallocating its mass, and is locally equivalent to zeroing the corresponding values for the affected query rows. These are two distinct diagnostics: renormalizing routing suppression and contribution deletion, not three independent causal targets.

The following contribution-deletion and value-editing tables are supplementary results, not measurements from the branch-correct revalidation. Their exact supplementary processor provenance remains unresolved. A separate corrected replication is now reported in Table 2; its percentages are not substituted into these supplementary tables. As a diagnostic, we compare the main renormalizing intervention against a no-renormalization variant that deletes the selected recipient contribution after softmax without redistributing its probability mass. The two variants select exactly the same dynamic events: for CLIP structural, both select 97,441 events, corresponding to 91.8% of rows and 98.4% of selected incoming mass, with the mask target equal to the original sink in every selected event. The no-renormalization intervention preserves most of the effect: 102.4% of the renormalizing all-sink effect, 95.6% of the CLIP-structural effect, and 86.0% of the padding effect. These are supplementary nonlinear endpoint-distance ratios, not additive components of a causal effect. The primary deletion conclusion uses the independently documented P comparison in Table 2.

We additionally run a direct value-side diagnostic on 553 prompts and four seeds. For the selected image-query contribution, no-renormalization deletion and value-zeroing are mathematically equivalent ways to remove $a_{is}W_O v_s$: one zeros the selected attention weight and the other zeros the selected value vector before the image-query matrix product. The identical metrics in Table 14 therefore serve as an implementation and causal-object check rather than a new independent effect estimate. They support the interpretation that the intervention is exposing selected value contributions, not only changing the softmax denominator.

F.4 CLIP Component Ablation

This ablation answers whether the CLIP-structural carrier is specifically an EOT effect. It is not. True non-padding EOT positions are almost never selected as dynamic sinks in the layer-12 early-window protocol: they account for only 217 selected events out of 106,176 image-query/head records, 0.009% of selected incoming mass, and 0.4% of the all-sink image-level effect. Padding positions are the dominant single component, explaining 89.1% of the all-sink effect and 90.8% of the full CLIP-structural effect. BOS/SOT-like positions are a secondary carrier, explaining 56.8% of the

Table 15: **SD3 CLIP structural component ablation.** We separate true EOT, padding, BOS/SOT, and the CLIP-L/CLIP-G components of fused CLIP structural positions in the supplementary component run (553 prompts × 4 seeds). Effect is normalized by the custom-noop floor and all-sink effect; the last effect column normalizes by the full CLIP-structural effect.

Condition	Raw LPIPS	Effect/all (%)	Effect/CLIP struct. (%)	CLIP Δ	Selected mass (%)
All dynamic sinks	0.1598	100.0	102.0	+0.0003	100.0
CLIP structural	0.1576	98.0	100.0	+0.0005	98.4
CLIP EOT+PAD	0.1472	89.0	90.8	+0.0004	81.5
CLIP EOT only	0.0458	0.4	0.4	-0.0000	0.0
CLIP PAD only	0.1473	89.1	90.8	+0.0004	81.5
CLIP BOS/SOT only	0.1104	56.8	57.9	+0.0001	16.9
CLIP-L structural positions	0.1576	98.0	100.0	+0.0005	98.4
CLIP-G structural positions	0.1576	98.0	100.0	+0.0005	98.4

Table 16: **Pre-softmax penalty sensitivity.** A 50-prompt × 4-seed SD3 sweep varies the logit penalty λ used for hard recipient suppression. Effects are normalized within each λ variant over its custom-noop floor and all-sink reference.

λ	All sinks	CLIP structural	CLIP padding	CLIP Δ range
10^2	100.0	101.5	89.6	-0.0011 to -0.0001
10^3	100.0	101.5	89.6	-0.0011 to -0.0001
10^4	100.0	101.5	89.6	-0.0011 to -0.0001
10^5	100.0	101.5	89.6	-0.0011 to -0.0001
10^6	100.0	101.5	89.6	-0.0011 to -0.0001

all-sink effect despite only 16.9% of selected incoming mass. The CLIP-L and CLIP-G structural selectors give identical image-level metrics because SD3 fuses the two CLIP streams into the same 77 recipient positions; at the recipient level, the two components mark the same structural positions in this recorded prompt set. These component statistics describe the supplementary run; they are not used to subdivide P's structural-recipient effect into padding, BOS/SOT or EOT contributions.

F.5 Pre-Softmax Penalty Sensitivity

The intervention is insensitive to the exact large logit penalty. In the 50-prompt × 4-seed sensitivity sweep, $\lambda = 10^2$ already gives the same image-level effect and CLIP deltas as $\lambda = 10^3, \ldots, 10^6$: all-sink suppression is fixed at 100% by definition, CLIP-structural suppression recovers 101.5% of the local all-sink reference, and CLIP-padding suppression recovers 89.6%. Thus the default $\lambda = 10^4$ lies well inside the saturated hard-mask regime. It is a numerically conservative implementation of recipient exclusion, not a tuned continuous control parameter; continuous strength is instead studied separately through the γ dose-response experiment.

F.6 Sampler, CFG, and Step Robustness

We run a 50-prompt × 4-seed robustness sweep over classifier-free guidance (CFG), denoising step count, and an alternative Heun flow-matching scheduler. We use flow-matching schedulers rather than Denoising Diffusion Implicit Model (DDIM) samplers because SD3 is a flow-matching model. In image space, CLIP-structural recipients recover 95.2–102.0% of the all-sink effect and CLIP-padding recipients recover 87.7–94.9%, with CLIP deltas remaining near zero (Table 17). In latent trajectory space, CLIP-structural suppression explains 97.4–97.6% of the all-sink step-0 divergence and 97.3–100.7% at the final recorded step across the same variants (Table 18). The DPM flow-matching scheduler was not available in the server's diffusers version and is therefore left for future robustness sweeps.

Table 17: **Sampler/CFG/step robustness.** A 50-prompt × 4-seed SD3 robustness sweep shows that CLIP structural and padding recipients continue to recover most of the all-sink effect under CFG and step-count changes. Effects are normalized over each variant's custom no-op floor and all-sink reference.

Setting	All-sink LPIPS−floor	CLIP structural	CLIP padding	CLIP Δ range
CFG 5.0, 20 steps	0.0676	102.0	91.4	-0.0006 to +0.0002
CFG 7.5, 20 steps	0.1127	101.5	89.6	-0.0011 to -0.0001
CFG 10.0, 20 steps	0.1668	99.5	92.7	-0.0006 to +0.0004
CFG 7.5, 12 steps	0.1335	97.3	89.2	+0.0004 to +0.0023
CFG 7.5, 28 steps	0.1047	99.7	87.7	+0.0005 to +0.0023
Heun, CFG 7.5, 20 steps	0.1107	95.2	94.9	-0.0011 to -0.0004

Table 18: **Trajectory robustness.** Across CFG, step count, and Heun flow-matching scheduler changes, CLIP-structural suppression continues to explain most all-sink latent divergence at step 0 and at the final recorded step.

Setting	All-sink step-0 L2	CLIP step-0 (%)	CLIP final (%)
CFG 5.0, 20 steps	0.516	97.4	99.8
CFG 7.5, 20 steps	0.784	97.5	100.7
CFG 10.0, 20 steps	1.059	97.5	100.1
CFG 7.5, 12 steps	1.374	97.4	98.1
CFG 7.5, 28 steps	0.557	97.6	99.8
Heun, CFG 7.5, 20 steps	0.784	97.5	97.3

G Supplementary Image and Trajectory Consequences

These supplementary audits examine semantic judgments, temporal behavior, and candidate-set dispersion under their recorded configurations. Each audit applies to its own image pairs, rather than the primary study's outputs.

G.1 Second-VLM Semantic Spot-Check

To answer whether the semantic-preservation conclusion depends on a single VLM family, we additionally evaluate the clean SD3/FLUX carriers with LLaVA-Next on 256 paired images per carrier, using the same paired-image manifest as Qwen2.5-VL. LLaVA-Next reports 97.7–98.4% preservation and 1.2–2.3% major-error rates, closely matching Qwen2.5-VL on the same pairs. Agreement is 97.3–99.2% across carriers. Cohen's κ is moderate to high for three carriers and lower for SD3 CLIP padding because the failure class is rare and class-imbalanced; raw agreement remains 97.3%.

G.2 Image-Side Sink Later-Step Audit

The main SD3 result isolates the early layer-12 structural-text carrier, where image-side recipients are essentially absent from the dynamic top-1 set and image-only suppression explains only 0.8% of the all-sink final latent divergence. A full-phase trajectory audit over neighboring layers shows a more nuanced picture: image-side recipients can become load-bearing in other layers and later phases, explaining roughly 40–50% of the all-sink final latent effect at layers 8, 16, and 20. This does not overturn the main mechanism; instead it separates two regimes. The clean same-prompt realization axis in SD3 is an early layer-12 structural-text routing effect, while image-side recipient routing appears as a separate later/layer-specific trajectory handle.

G.3 Phase-Aligned Semantic Trajectory Audit

The final-image VLM and CLIP results show that structural-recipient suppression preserves prompt-level semantics at the end of sampling. A remaining possibility is delayed semantic recovery: the intervention might temporarily damage semantic alignment during denoising and only recover it by the final image. To test this, we decode intermediate SD3 latents at normalized denoising phases

Table 19: Second-VLM semantic spot-check. Qwen2.5-VL and LLaVA-Next evaluate the same 256 paired baseline/intervention images per clean carrier; preservation is the rate of prompt-relevant semantic preservation and major error is the rate of severe prompt-relevant failures.

Carrier	Judge	n	Preserve ↑	Major error ↓	Score ↑
FLUX dual-12 T5	llava_next_default	256	97.7	2.3	4.91
FLUX dual-12 T5	qwen_default	256	97.3	2.7	4.65
FLUX single-18 T5	llava_next_default	256	98.4	1.6	4.94
FLUX single-18 T5	qwen_default	256	97.3	2.7	4.84
SD3 CLIP structural	llava_next_default	256	98.4	1.2	4.94
SD3 CLIP structural	qwen_default	256	98.4	1.6	4.71
SD3 CLIP padding	llava_next_default	256	98.0	1.6	4.92
SD3 CLIP padding	qwen_default	256	97.7	2.3	4.71

Table 20: Second-VLM agreement on the same image pairs. Agreement and Cohen's κ are reported for prompt-relevant semantic preservation.

Carrier	Judge A	Judge B	n	Agreement	κ
FLUX dual-12 T5	llava_next_default	qwen_default	256	98.0	0.61
FLUX single-18 T5	llava_next_default	qwen_default	256	98.8	0.72
SD3 CLIP structural	llava_next_default	qwen_default	256	99.2	0.75
SD3 CLIP padding	llava_next_default	qwen_default	256	97.3	0.35

for 128 prompts and four paired seeds. Each intervention frame is compared to the same-prompt, same-seed baseline decoded at the same phase.

The supplementary audit reports small mean CLIP-T changes at sampled phases. These averages do not resolve per-image errors or unsampled transient changes and do not certify semantic preservation throughout denoising.

G.4 Displacement–Dispersion Decomposition

Section H.6.1 asks whether structural-recipient intervention broadens the output distribution or instead rearranges same-prompt realizations. We extract L2-normalized DINOv2 features for all 553 prompts and four paired seeds, then compare each intervention to the same-prompt, same-seed baseline. The Fréchet-style decomposition is computed after projecting the concatenated baseline/intervention features to 256 principal components. We also compute, for each prompt, the mean pairwise cosine distance among the four seed features; the reported Δ dispersion is paired by prompt and bootstrapped over prompts.

The covariance-dominated global feature shift should not be read as increased sampling diversity. For all three interventions, the within-prompt cross-seed dispersion confidence interval contains zero. Global Vendi scores are also nearly unchanged: baseline has Vendi score 213.31, whereas all-sinks, CLIP-structural, and padding/structural interventions have scores 213.14, 213.10, and 213.17, respectively. Thus the intervention changes how each prompt/seed trajectory is realized, but does not measurably widen the cross-seed distribution under this protocol.

G.5 Intervention-Augmented Candidate Sets

We additionally test whether sink-steered images can be combined with baseline images to improve user-facing candidate-set diversity. This evaluates a different claim from Section G.4: even if the intervention alone does not broaden the sampling distribution, a mixed set of baseline and steered images might still provide more valid visual options. For each prompt, we compare four baseline seeds, four same-seed steered samples, their eight-image union, and an eight-baseline reference using four additional baseline seeds. Diversity is mean pairwise LPIPS within each candidate set.

The result is negative for diversity enhancement under these sampling policies: mixed candidate sets have lower mean pairwise LPIPS than either baseline reference. This is consistent with redundant paired alternatives, but does not identify seed pairing as the cause; a deterministic transformation

Table 21: **Image-side sink later-step audit.** We run a 50-prompt × 4-seed full-phase trajectory audit and ask whether image-recipient dynamic sinks become important at later denoising steps. Image top-1 reports the largest fraction of selected top-1 recipients that are image-side at any step. Image-only effect is normalized by the corresponding all-sink latent divergence.

Layer	Max image top-1 (%)	Step	Final image L2	Final image/all (%)	Max image/all (%)
8	16.1	19	65.316	50.0	50.1
12	0.0	0	1.210	0.8	0.8
16	11.9	0	59.708	40.1	40.1
20	6.5	9	58.463	44.1	44.1

Table 22: **Phase-aligned semantic trajectory audit.** CLIP-T deltas are computed against the same-prompt, same-seed baseline decoded at the same denoising phase. Confidence intervals bootstrap prompt-seed pairs.

Condition	0%	10%	30%	100%
No-op	+0.0000 [-0.0000, +0.0000]	-0.0000 [-0.0000, +0.0000]	-0.0000 [-0.0002, +0.0001]	+0.0001 [-0.0003, +0.0006]
All sinks	+0.0000 [+0.0000, +0.0000]	+0.0000 [-0.0000, +0.0001]	-0.0003 [-0.0008, +0.0001]	-0.0001 [-0.0012, +0.0010]
CLIP structural	+0.0000 [+0.0000, +0.0000]	+0.0001 [-0.0000, +0.0002]	-0.0002 [-0.0007, +0.0002]	+0.0003 [-0.0006, +0.0014]
Padding	+0.0000 [+0.0000, +0.0000]	+0.0001 [-0.0000, +0.0001]	-0.0003 [-0.0008, +0.0001]	+0.0001 [-0.0008, +0.0010]
Content	-0.0000 [-0.0000, +0.0000]	+0.0000 [-0.0000, +0.0000]	+0.0001 [-0.0001, +0.0003]	+0.0000 [-0.0006, +0.0007]

can also increase dispersion. The established capability is same-seed counterfactual generation, not a demonstrated advantage in candidate-set diversity or instruction-following editing.

G.6 Randomized-Dose Diversity Pilot

To distinguish the fixed-dose null from the broader possibility of randomized steering, we evaluate 30 prompts and 16 seeds per prompt under five conditions. The pilot uses SD3 medium, layer-12 top-1 CLIP-structural selection, and the early intervention window. Each condition generates 480 images, for a total of 2,400. Base seeds are 1000, 1100, …, 2500, with the original prompt index added to each base seed for latent initialization. The same initial noise is used across conditions, and each condition receives 16 distinct seeds per prompt. The fixed dose is $\gamma = 0.5$; the randomized dose is drawn once per image from $\mathcal{U}(0.25, 0.75)$, independently of the latent-noise random number generator (RNG), and held fixed across eligible events. The random-text control applies that same dose to alternative text positions at the corresponding eligible events. It controls dose and event budget, not incoming mass or perceptual magnitude. Although the fixed and randomized policies share mean γ, their mean logit bias and nonlinear effects need not match; this is a policy comparison, not an isolation of randomness from all intervention-strength effects.

For each prompt and condition, diversity is the mean over the 120 unordered image pairs. We report AlexNet LPIPS and cosine distance between normalized DINOv2-base features and mean CLIP ViT-B/32 text–image score. Metrics use 256-square evaluation images. Differences are paired by prompt; 10,000 bootstrap resamples retain all seeds within each sampled prompt. The intervals are exploratory, unadjusted 95% intervals, not simultaneous confidence bounds or equivalence tests.

Random-dose mean LPIPS diversity is 0.661792, compared with 0.661876 for baseline and 0.661916 for fixed dose. Sixteen prompts increase relative to baseline and fourteen decrease. Selection logs record zero edited events for baseline and custom no-op, and 21,051 for each of the three intervention arms. These counts support successful triggering and equal event budgets; they do not by themselves verify gamma values, equal removed mass, or equal output perturbation. The two audit manifests contain the same 2,400 unique image paths, but all validity fields are empty. We therefore report neither valid-only diversity nor semantic noninferiority. This single randomization design provides no evidence of useful diversity improvement, without excluding other dose distributions, temporal policies, or outcomes beyond the tested metrics.

Table 23: **Displacement–dispersion decomposition in DINOv2 feature space.** Mean and covariance fractions decompose the global baseline-to-intervention feature shift. Δ disp. is the paired change in within-prompt cross-seed cosine distance with bootstrap 95% confidence intervals.

Intervention	n	Mean frac. (%)	Cov. frac. (%)	Δ disp.	Interpretation
All sinks	2212	0.9	99.1	-0.0016 [-0.0053, 0.0023]	no detected dispersion change
CLIP structural	2212	0.8	99.2	-0.0017 [-0.0055, 0.0020]	no detected dispersion change
Padding / structural	2212	0.8	99.2	0.0001 [-0.0035, 0.0036]	no detected dispersion change

Table 24: **Intervention-augmented candidate-set diversity.** Mixed-8 combines four baseline samples with four same-seed steered samples; fixed-4 averages all four-sample subsets from the mixed pool. Gains are relative to the four-seed baseline diversity; mixed-8/baseline-8 compares against four additional baseline seeds.

Intervention added	n	Mixed-8 gain	Fixed-4 gain	Mixed-8/base-8	Novelty	CLIP Δ
All sinks	553	0.886 [0.885, 0.887]	0.886 [0.885, 0.887]	0.886 [0.881, 0.890]	0.1319	+0.0002
CLIP structural	553	0.886 [0.885, 0.887]	0.886 [0.885, 0.887]	0.886 [0.881, 0.890]	0.1301	+0.0002
Padding structural	553	0.884 [0.883, 0.886]	0.884 [0.883, 0.886]	0.884 [0.880, 0.888]	0.1225	+0.0002
CLIP structural $\gamma = 0.5$	553	0.874 [0.874, 0.875]	0.874 [0.874, 0.875]	0.874 [0.870, 0.878]	0.0794	+0.0001

G.7 Additional Qualitative Steering Examples

Figure 3 supplements the main-text teaser with additional condition comparisons. It illustrates changes in visual realization, rather than serving as primary evidence of effect size or semantic preservation. Quantitative claims are grounded in the corresponding protocol-specific evaluations; the corrected CFG evidence is reported separately in Appendix B.

H Supplementary Comparative Measurements and Scope

These supplementary comparisons consolidate family-level, anatomical, guidance, and model-specific measurements, retaining their original configurations and uncertainty definitions. Supplementary recovery ratios use their own native/no-op floor, and their semantic judgments concern different generated images. Execution status and evidence roles are specified in Table 6.

H.1 Model-Specific Scope Records

The following are supplementary observations with execution status specified in Table 6, not verified cross-model replications. The FLUX records report T5-special displacement ratios of 61.0% and 50.3% of the within-block all-sink over-floor effects at dual-12 and single-18, respectively. The associated 553-pair semantic audit finds 96.9% and 97.3% preservation under the default Qwen rubric and 94.6% and 96.7% under a stricter rubric. These are two prompts for one judge family; a separate 256-pair LLaVA spot-check supplies cross-family evidence, not a second full 553-pair audit of those supplementary outputs.

The U-Net records report lower semantic-preservation rates. Default/strict preservation rates are 75.8/65.3% for SD2.1, 66.4/56.2% for SD1.5, and 57.0/37.1% for SDXL, despite small mean CLIP changes in some cases. Thus CLIP stability alone cannot establish clean realization control across architectures. SD3.5 supplies anatomy evidence for shifted carriers, but its current processor floor prevents a clean image-level causal conclusion. Model configurations and complete supplementary comparisons remain in Appendices C, H.7, and G.1.

These supplementary observations motivate model-specific questions about carrier identity and semantic coupling: FLUX suggests block-dependent carriers, U-Net results suggest stronger semantic entanglement, and SD3.5 remains measurement-limited. Their supplementary provenance qualifications remain separate from the verified SD3 contribution, timing, dose, and semantic chain. Additional examples in Appendix G.7 retain their original generation settings.

Table 25: **Random-dose steering at equal image budget.** Each entry is random γ minus the reference, with prompt-paired bootstrap 95% CI. All values are multiplied by 10^3 for readability. No comparison detects a mean change; this does not establish equivalence. Absolute prompt-validity labels are unavailable for this pilot.

Reference	LPIPS diversity Δ	DINO dispersion Δ	CLIP-T Δ
Native baseline	-0.084 [-1.513, 1.340]	0.474 [-4.649, 5.526]	-0.281 [-1.050, 0.532]
Custom no-op	0.347 [-0.941, 1.635]	-1.503 [-7.240, 3.975]	-0.280 [-0.974, 0.370]
Fixed $\gamma = 0.5$	-0.124 [-0.891, 0.633]	-0.502 [-6.340, 4.833]	-0.252 [-0.900, 0.378]
Random-text, same dose	0.343 [-1.041, 1.690]	1.977 [-2.695, 6.661]	-0.469 [-1.116, 0.169]

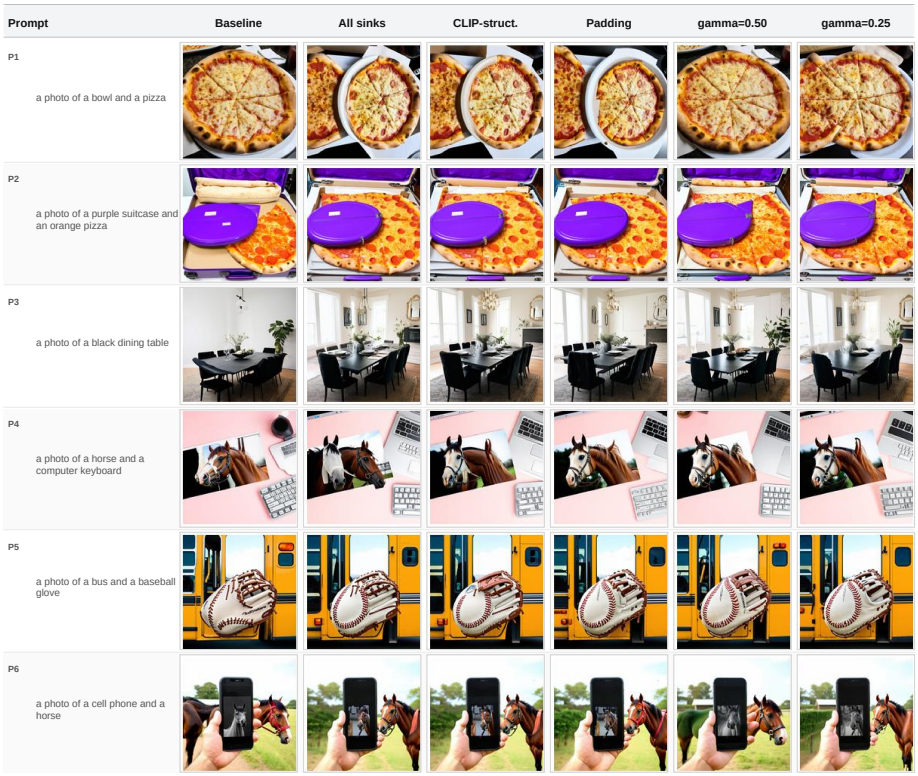


Figure 3: **Same-prompt realization steering under the supplementary protocol.** Columns show baseline, all-sink suppression, CLIP-structural suppression, padding-only suppression, and selected sink-mass dose settings, holding the prompt and seed fixed within each row. These illustrative examples are not regenerated branch-correct CFG results and do not establish semantic preservation for every displayed image.

H.2 Supplementary SD3 Localization Records

This section preserves the supplementary family, control and trajectory records. Their tabulated associations are not used to establish P's mechanism or robustness conclusions.

H.2.1 Which Recipients Carry the Dynamic-Sink Effect?

The following family breakdown retains the supplementary protocol. The primary branch-correct comparison is Table 1; its results are not pooled with this table.

We first ask which recipients carry the causal effect of dynamic sink suppression in SD3-medium. Across 553 GenEval prompts and four paired seeds, masking all selected sinks produces a large

Table 26: **Supplementary SD3 family breakdown.** Structural text sinks recover almost all all-sink visual-shift magnitude under this earlier protocol, while content and image sinks are weak. Raw LPIPS is reported against the same-prompt baseline; Effect and adjusted effect are normalized over the no-op floor (raw LPIPS $= 0.0454$). Brackets are paired-bootstrap 95% confidence intervals over prompt-seed pairs. VLM d/s reports supplementary default/strict Qwen2.5-VL semantic-preservation rates. The primary corrected comparison is Table 1.

Intervention	Carrier	Raw LPIPS	Effect [95% CI]	Adj. [95% CI]	Mass	CLIP Δ	VLM d/s	Role
All dynamic sinks	all recipients	0.1598	100.0 [100.0, 100.0]	–	100.0	+0.0003	–	reference
CLIP structural	structural text	0.1576	98.1 [96.7, 99.4]	91.8 [90.0, 93.8]	98.4	+0.0005	98.4/95.5	clean main carrier
CLIP padding	structural text	0.1486	90.2 [88.3, 92.2]	89.5 [87.2, 91.8]	81.9	+0.0003	97.6/94.9	clean secondary carrier
Content tokens	content text	0.0679	19.7 [17.9, 21.5]	19.3 [17.3, 21.2]	0.8	+0.0000	–	weak contrast
Image-side sinks	image	0.0455	0.1 [-0.1, 0.3]	–	0.0	-0.0001	–	negative control

Table 27: **Empirical routing-leverage decomposition in SD3.** Top-1 columns are dynamic sink selection frequencies with Wilson 95% confidence intervals. Sensitivity is final-latent L2 displacement from single-layer early-window suppression with bootstrap 95% confidence intervals.

Layer	CLIP top-1	Pad top-1	Content top-1	CLIP value	CLIP sens.	Pad sens.	Content sens.
8	69.0 [68.8, 69.3]	8.5 [8.3, 8.7]	10.0 [9.8, 10.2]	0.482	95.9 [90.9, 100.2]	84.5 [80.0, 88.8]	70.8 [66.7, 75.3]
12	91.8 [91.6, 91.9]	0.6 [0.5, 0.6]	5.6 [5.5, 5.7]	0.978	129.8 [123.8, 135.3]	125.6 [119.4, 131.5]	74.8 [71.0, 78.9]
16	86.4 [86.2, 86.7]	1.8 [1.8, 1.9]	0.0 [0.0, 0.0]	1.396	125.0 [119.2, 130.7]	124.5 [118.8, 130.0]	55.0 [51.8, 58.1]
20	39.5 [39.2, 39.8]	51.7 [51.4, 52.0]	2.7 [2.6, 2.8]	0.961	103.8 [99.1, 108.7]	102.3 [97.4, 107.4]	54.7 [51.5, 57.8]

same-prompt perceptual shift. However, this effect is highly localized: masking only CLIP-fused structural text recipients recovers nearly the full effect, while padding-only sinks recover most of it and content-token sinks recover only a small fraction. In the main early-window layer-12 protocol, image-side sinks collapse to the mask-disabled processor floor. This should not be read as a claim that image-side recipients are globally irrelevant: a full-phase appendix audit finds later/layer-specific image-side trajectory effects outside the clean early structural-text carrier (Appendix G.2). The all-sinks and CLIP-structural conditions are close but not identical: paired bootstrap gives a small LPIPS gap of -0.0022 [-0.0039, -0.0006] for CLIP structural relative to all sinks, while still recovering 98.1% [96.7, 99.4] of the floor-subtracted all-sink effect. A separate supplementary CLIP-component analysis records these ratios: within the CLIP-fused structural family, padding positions alone recover 89.1% of the all-sink effect and 90.8% of the full CLIP-structural effect, BOS/SOT-like positions recover 56.8%, while true EOT positions recover only 0.4% (Appendix F.4). These supplementary component ratios do not refine the carrier conclusion of P.

The supplementary controls compare alternative target-selection rules. Matched non-sink controls are specified to preserve token family and intervention budget but remove the dynamic recipient state; random-text controls preserve the number of masked events but relocate them to non-selected text positions. Both controls collapse near the mask-disabled floor. These comparisons show that token-family membership and masked-event count alone do not reproduce the effect. They do not equalize incoming mass or value geometry, and therefore support site specificity without isolating structural identity as an independent cause. This is compatible with, but not implied by, the local contribution identities in Section 5.

H.3 Ruling Out Alternative Explanations

H.3.1 Mass, Value, and Structurality Are Not Interchangeable

We next test whether the load-bearing carrier is explained by attention mass alone, value geometry alone, or their interaction with downstream flow sensitivity. For layers 8/12/16/20, we decompose early-window sink events into dynamic top-1 selection frequency, a value-side leverage term ($a_r \|W_O(v_r - \bar{v}_{r^c})\|$ for top-1 recipient r), and a downstream sensitivity proxy measured as final-latent displacement after single-layer suppression. This finite-effect proxy includes local perturbation magnitude and selection, and is not an independent, perturbation-normalized Jacobian gain. Layer 12 exhibits more concentrated structural selection in this audit: 91.8% of top-1 sink events land on CLIP-fused structural recipients, compared with 86.4% at layer 16, with non-overlapping Wilson intervals. Layer 16 has the largest value-side leverage (1.396 vs. 0.978 at layer 12), while downstream sensitivity at layers 12 and 16 is comparable within bootstrap uncertainty ([123.8,

Table 28: **Layer and top-k sensitivity.** Structural text recipients remain high-leverage beyond the default layer-12/top-1 setting, but top-1 gives the cleanest dynamic-recipient test. Values are LPIPS over the no-op floor normalized to all-sinks at the same layer/top-k setting.

Layer	k	All-sink LPIPS−floor	CLIP struct.	Padding	Content	Image
8	1	0.081	75.3	56.0	29.1	4.2
8	2	0.104	60.7	53.0	51.9	6.2
8	4	0.121	54.4	56.6	64.2	9.0
12	1	0.114	98.1	90.2	19.7	0.1
12	2	0.149	93.6	82.8	21.9	0.1
12	4	0.193	89.3	78.3	22.5	1.0
16	1	0.109	98.1	96.5	0.0	2.3
16	2	0.141	85.5	91.3	20.6	2.9
16	4	0.165	84.9	87.4	21.9	3.2
20	1	0.094	80.3	81.5	4.5	0.6
20	2	0.118	77.4	78.0	8.2	0.9
20	4	0.131	81.5	80.0	14.1	1.2

Table 29: **Structurality disambiguation in SD3.** For each CLIP-structural dynamic sink event, non-structural text controls mask the best available alternative recipient under mass, projected-value, leverage, or top-mass matching. Non-structural text recipients cannot match the structural sink mass/leverage and recover only a minority of the effect.

Intervention	LPIPS	Over floor	Effect %	Mask mass	Mask leverage
CLIP structural sink	0.1580	0.1120	100.0	0.1175	0.9776
Best non-structural mass	0.0828	0.0369	32.9	0.0076	0.1231
Best non-structural value	0.0485	0.0025	2.3	0.0012	0.0134
Best non-structural leverage	0.0811	0.0352	31.4	0.0075	0.1230
Top-mass non-structural text	0.0828	0.0369	32.9	0.0076	0.1231

135.3] and [119.2, 130.7], respectively). By layer 20, the carrier migrates toward padding-dominant structural recipients. Across all four layers, structural recipients have much larger downstream sensitivity than content or image-only sinks; at layer 12, CLIP structural and padding interventions produce latent-L2 shifts of 129.8 and 125.6, compared with 74.8 for content and 55.5 for image-only sinks. These observations are consistent with a multi-factor interpretation involving selection, projected value contrast, and downstream propagation; they do not independently validate a multiplicative predictor of final-image effects.

Table 28 records supplementary layer/top-k variations and their within-setting normalized ratios. It does not establish robustness of P. The separate branch-correct replication in Appendix B.4 tests layers 8 and 16 at top-1.

A sharper null hypothesis is that structurality is incidental: perhaps any early text recipient with comparable mass or value leverage would control the trajectory. We test this by anchoring every control to the same CLIP-structural dynamic sink event and replacing the mask target with the best available non-structural text recipient under four criteria: closest incoming mass, closest projected-value contrast, closest mass–value leverage, and highest non-structural incoming mass. The result favors the structural-carrier interpretation. Non-structural text recipients cannot match the structural sink mass or leverage in SD3: the structural events have mean incoming mass 0.1175 and mean mass–value leverage 0.9776, whereas the best non-structural mass/leverage alternatives reach only about 0.0076 mass and 0.123 leverage. Their image-level effect recovers only 31–33% of the CLIP-structural effect, and value-matched alternatives collapse to 2.3%. Thus the available non-structural alternatives do not reproduce the endogenous structural events. However, the large residual mass and leverage gaps limit identification: these data cannot determine whether structural identity would retain an effect under genuinely matched mass and geometry. Our conclusion concerns where SD3 hosts its endogenous high-leverage contributions, not the causal necessity of the structural label itself.

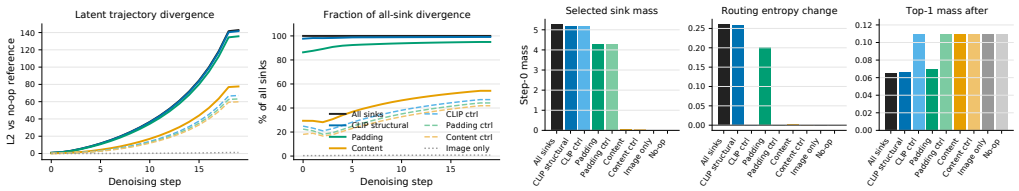


Figure 4: **Trajectory bifurcation under the supplementary protocol.** Structural-sink suppression causes an immediate first-step latent divergence and disrupts the dominant incoming-mass routing pattern at step 0.

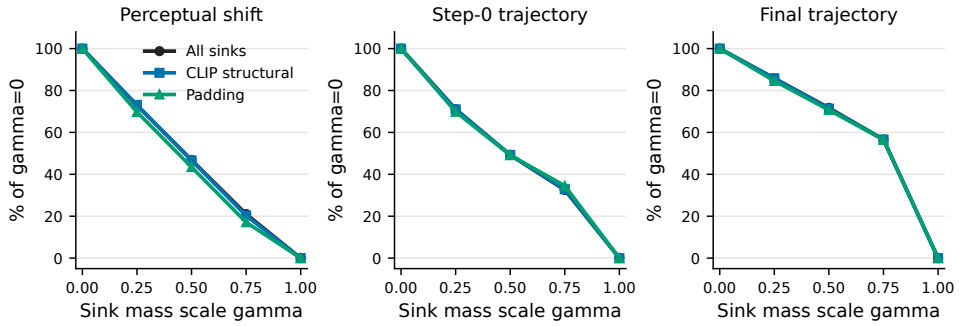


Figure 5: **Sink-mass dose response under the supplementary protocol.** Scaling selected sink mass with $\gamma \in \{0, .25, .5, .75, 1\}$ produces a smooth perceptual response. CLIP structural recipients closely track the all-sink curve, showing that the main all-sink effect is carried by a small structural-recipient family.

H.3.2 Responses to Tested Fixed-Slot Redirects

A supplementary diagnostic compares tested fixed-slot redirects with dynamic recipient suppression. Static-receptacle interventions force attention toward a fixed structural CLIP slot rather than toward the model's endogenous dynamic sink recipients. They do not reproduce the dynamic-sink trajectory. Instead, they overshoot the all-sink perceptual shift at strong redirect values (up to about $2.0\times$ the all-sink LPIPS effect in the formal ablation), create an artificial first-step latent attractor with more than $2.2\times$ the all-sink step-0 divergence, and reduce rather than redistribute attention entropy. These tested static receptacles do not reproduce endogenous routing merely by attracting mass. The result is consistent with sensitivity to the recipient's value and downstream context (a conditional interpretation), but does not isolate activation misalignment from concentration or intervention-strength effects. It does not rule out other trained or context-adapted receptacle designs.

H.4 Trajectory and Dose Control

H.4.1 First-Step Bifurcation and Continuous Sink-Mass Scaling

The figures and tables below document the supplementary trajectory and dose studies. Corrected results are reported in Section 5.3.

The routing-leverage view allows the downstream consequence of a local intervention to depend on its timing. The supplementary trajectory probe records an immediate first-step divergence. In the corrected timing probe, divergence is zero before intervention, appears at the intervened step, and subsequently propagates. For step-0 structural suppression, mean latent L2 grows from 0.78 immediately after intervention to 141.28 at the final step. Thus first-step intervention can initiate substantial later divergence; this does not mean that most of the final latent distance is already present at step 0.

Table 30: **Sink-mass dose-response uncertainty.** LPIPS is floor-subtracted and normalized within each family to $\gamma = 0$; intervals use paired bootstrap over prompt-seed pairs.

Family	γ	LPIPS % of $\gamma = 0$ [95% CI]	CLIP Δ [95% CI]
all_sinks	0.00	100.0 [100.0, 100.0]	+0.0002 [-0.0003, +0.0007]
all_sinks	0.25	72.9 [71.7, 74.1]	+0.0002 [-0.0002, +0.0007]
all_sinks	0.50	46.9 [45.3, 48.4]	+0.0001 [-0.0003, +0.0005]
all_sinks	0.75	21.0 [19.5, 22.5]	-0.0000 [-0.0004, +0.0003]
all_sinks	1.00	0.0 [0.0, 0.0]	+0.0000 [-0.0003, +0.0003]
clip_structural	0.00	100.0 [100.0, 100.0]	+0.0003 [-0.0002, +0.0008]
clip_structural	0.25	73.1 [71.8, 74.4]	+0.0003 [-0.0001, +0.0007]
clip_structural	0.50	46.7 [45.2, 48.2]	+0.0001 [-0.0002, +0.0005]
clip_structural	0.75	20.3 [18.7, 21.8]	+0.0001 [-0.0002, +0.0005]
clip_structural	1.00	0.0 [0.0, 0.0]	+0.0000 [-0.0003, +0.0003]
padding	0.00	100.0 [100.0, 100.0]	+0.0002 [-0.0002, +0.0007]
padding	0.25	69.8 [68.4, 71.1]	-0.0001 [-0.0005, +0.0003]
padding	0.50	43.5 [41.9, 45.1]	+0.0000 [-0.0003, +0.0004]
padding	0.75	17.1 [15.5, 18.7]	-0.0001 [-0.0004, +0.0002]
padding	1.00	0.0 [0.0, 0.0]	+0.0000 [-0.0003, +0.0003]

Table 31: **Supplementary additional-ablation statistics.** Exact distance numerator, denominator and recipient-exclusion rules require source verification. Recorded ratios are retained without causal-share interpretation.

Family	γ	Supplementary ratio
all_sinks	0.00	26.2%
clip_structural	0.00	25.0%
padding	0.00	22.9%
all_sinks	0.50	72.1%
clip_structural	0.50	74.0%
padding	0.50	70.2%

Binary masking could be dismissed as a destructive ablation. We therefore replace hard suppression with a sink-mass scaling parameter γ, where $\gamma = 0$ corresponds to hard suppression and $\gamma = 1$ recovers the mask-disabled endpoint. The response is smooth and monotonic. For all sinks, the normalized LPIPS effect decays from 100% to 72.9% [71.7, 74.1], 46.9% [45.3, 48.4], 21.0% [19.5, 22.5], and then 0% as γ moves from 0 to 1. CLIP structural recipients closely track the all-sink curve, while padding follows a slightly weaker but still monotonic curve. This turns the mechanism from a binary diagnostic into a continuous response across the tested dose grid in this supplementary run. Table 30 reports paired-bootstrap intervals for the complete dose curve.

The supplementary additional-ablation ratios in Table 31 require verification of their distance numerator, reference denominator and recipient-exclusion rule. They are not used as causal shares or mechanism evidence.

The supplementary phase-aligned audit reports small mean CLIP-T differences at sampled phases; these do not establish absence of transient semantic errors.

H.5 Semantic Judgments and Guidance Comparisons

H.5.1 VLM Judgments and Tested Guidance Scales

Large perceptual shifts are only useful for this paper if they do not simply reflect semantic drift. We therefore evaluate paired baseline/intervention images with two Qwen2.5-VL judge variants that separately score baseline prompt validity, intervention prompt validity, and relative semantic preservation. Relative preservation is not absolute prompt correctness: a pair can preserve an error already present in the baseline. Accordingly, the preservation rates below are not valid-image generation rates or guarantees of anatomical quality. A baseline-validity-stratified transition analysis is needed to distinguish retained errors from newly introduced or corrected errors; the aggregate

Table 32: **Multi-judge VLM semantic preservation audit.** Qwen2.5-VL evaluates 553 paired baseline/intervention images per carrier with the original prompt and a stricter prompt-semantics variant. Preserve is prompt-relevant semantic preservation; agreement and κ compare the two judge variants.

Carrier	Default preserve (%)	Strict preserve (%)	Strict major error (%)	Agreement (%)	κ
SD3 CLIP structural	98.4	95.5	4.5	96.7	0.46
SD3 CLIP padding	97.6	94.9	5.1	96.9	0.57
FLUX dual-12 T5	96.9	94.6	5.4	97.6	0.71
FLUX single-18 T5	97.3	96.7	3.3	99.1	0.84
SD2.1 structural	75.8	65.3	34.7	89.2	0.74
SD1.5 structural	66.4	56.2	43.8	89.5	0.78
SDXL structural	57.0	37.1	62.9	80.1	0.62

Table 33: **PAG scale comparison with paired-bootstrap confidence intervals.** LPIPS and CLIP Δ are computed against the same-prompt, same-seed baseline; 95% confidence intervals bootstrap prompt-seed pairs.

Method	LPIPS [95% CI]	CLIP Δ [95% CI]	Interpretation
Custom no-op	0.0381 [0.0324, 0.0441]	-0.0000 [-0.0006, +0.0006]	processor floor
Ours $\gamma = 0.5$	0.0900 [0.0780, 0.1026]	-0.0004 [-0.0013, +0.0004]	moderate sink-routing shift
Ours hard structural	0.1448 [0.1295, 0.1607]	-0.0003 [-0.0015, +0.0011]	hard structural shift
PAG scale=0.25	0.3304 [0.3115, 0.3486]	-0.0026 [-0.0048, -0.0006]	weakest PAG, already global
PAG scale=1.0	0.5857 [0.5641, 0.6072]	-0.0063 [-0.0101, -0.0024]	strong global guidance
PAG scale=3.0	0.8325 [0.8120, 0.8530]	-0.0460 [-0.0532, -0.0390]	large CLIP decrease

preservation rates alone do not establish that distinction. On the clean SD3 and FLUX carriers, relative semantic preservation is 96.9–98.4% under the default rubric and 94.6–96.7% under the stricter prompt-semantics rubric, with strict major-error rates between 3.3% and 5.4%. Agreement between the two judge variants is 96.7–99.1% on these clean carriers. A 256-pair spot-check with a different VLM family, LLaVA-Next, gives 97.7–98.4% preservation on the same clean carriers and agrees with Qwen2.5-VL on 97.3–99.2% of paired judgments (Appendix G.1). The same audit exposes a boundary: SD1.5, SD2.1, and SDXL show lower VLM preservation despite small CLIP deltas in some cases. We therefore use VLM preservation to separate clean realization control from architectures where structural routing is causally strong but more semantically entangled.

We do not position sink routing as a replacement for existing guidance or editing methods. Instead, we test whether it exposes a distinct control axis. At the tested settings, structural-recipient interventions have smaller perceptual shifts and smaller mean CLIP decreases than PAG. The paired-bootstrap CLIP confidence interval for hard structural suppression contains zero ([-0.0015, +0.0011]), as does the interval for $\gamma = 0.5$ sink-mass scaling ([-0.0013, +0.0004]); under this test we cannot reject no CLIP-detectable semantic drift. By contrast, even the weakest PAG setting tested has a CLIP interval strictly below zero ([-0.0048, -0.0006] for PAG 0.25), and stronger PAG scales drift further. The perceptual magnitude is also separated: hard structural suppression produces more than $2\times$ lower LPIPS than PAG 0.25, and their LPIPS intervals ([0.1295, 0.1607] vs. [0.3115, 0.3486]) do not overlap. However, a confidence interval containing zero for one method and excluding zero for another does not establish a significant between-method difference. Such a claim requires a direct paired contrast, and non-rejection of a zero delta is not a semantic-equivalence test. Because the weakest tested PAG setting already produces a larger perceptual shift, this table provides an operational comparison at the reported settings, not proof of orthogonal control axes or superior semantic preservation at matched perceptual magnitude.

H.6 Utility Boundary: Realization Control Without Demonstrated Diversity Gains

H.6.1 Deterministic Local Realization Shifts Rather Than Sampling Broadening

For fixed initial noise and deterministic sampling, recipient intervention changes the denoising path without adding an independent random source. Proposition 1 characterizes this response under its regularity assumptions, but does not predict whether the distribution across seeds broadens: even a deterministic map can increase dispersion. We therefore evaluate distributional broadening as a separate empirical question on the full 553-prompt, four-seed SD3 run using DINOv2 features (Appendix G.4).

The results do not provide evidence of increased dispersion in this setup. For all-sink, CLIP-structural, and padding interventions, the paired change in within-prompt cross-seed cosine dispersion is statistically indistinguishable from zero: -0.0016 $[-0.0053, 0.0023]$, -0.0017 $[-0.0055, 0.0020]$, and 0.0001 $[-0.0035, 0.0036]$, respectively. Global Vendi scores are also effectively unchanged (baseline 213.31 vs. 213.14, 213.10, and 213.17). At the same time, a Fréchet-style feature decomposition assigns nearly all global shift to covariance geometry rather than to a single feature-mean displacement. This global statistic does not establish prompt-conditioned covariance rearrangement. Individual realizations change without a detected increase in the tested dispersion measures; this does not establish distributional equivalence or rule out changes beyond the resolution of four seeds per prompt. Broadening or correcting a data-mode distribution remains an unestablished application, not a consequence of the local routing identities.

We also test the more application-oriented hypothesis that steered samples could be combined with baseline samples to enlarge the user-facing candidate set. This candidate-set interpretation is not supported as a diversity claim: combining four baseline samples with four same-seed steered samples produces lower pairwise LPIPS diversity than drawing additional baseline seeds (mixed-8/baseline-8 $= 0.886$ [0.881, 0.890] for CLIP-structural suppression; Appendix G.5). This result concerns the tested fixed-strength, paired sampling policy. A deterministic correspondence does not itself prohibit broader distributions, and we have not established that seed pairing causes the observed lack of diversity gain. The demonstrated use is fixed-prompt counterfactual visualization, not instruction-following image editing or improved sampling coverage.

We also test whether randomizing strength once per trajectory improves diversity at equal generation budget. A 30-prompt, 16-seed pilot compares baseline, custom no-op, fixed $\gamma = 0.5$, random $\gamma \sim \mathcal{U}(0.25, 0.75)$, and random-text targets receiving the same random dose. Random-dose LPIPS diversity differs from baseline by -0.000084 (95% CI $[-0.001513, 0.001340]$) and from fixed dose by -0.000124 ($[-0.000891, 0.000633]$); DINOv2 dispersion likewise shows no detected gain. Thus this randomized policy does not establish a diversity advantage, although it does not rule out other policies or metrics. Appendix G.6 reports the complete comparisons and execution checks. Absolute prompt-validity labels are not yet available for this pilot, so its CLIP results do not establish semantic equivalence.

H.7 Architectural Scope

H.7.1 Carrier Migration, Boundary Cases, and Qualitative Realization Steering

These supplementary measurements motivate model-specific hypotheses, rather than certify a cross-model mechanism. The SD3.5 processor floor additionally limits image-level interpretation. Padding Tone intervenes on padding representations during denoising (Toker et al., 2025); the distinction is selection rule and target, not static versus dynamic analysis.

Additional same-prompt, same-seed visual comparisons are provided in Appendix G.7 (Figure 3). They illustrate realization changes under the supplementary protocol rather than constitute new branch-correct or image-level semantic validation.

Table 34: **Supplementary model-specific measurements.** The SD3 row is supplementary, not study P. Execution status is listed in Table 6. VLM preservation reports default/strict Qwen2.5-VL judge variants on 553 paired images per carrier.

Model	Carrier	Effect	CLIP Δ	VLM preserved	Interpretation
SD3	CLIP structural	98.1	+0.0005	98.4 / 95.5	clean main carrier
FLUX dual-12	T5 special	61.0	near-flat	96.9 / 94.6	recorded T5 carrier
FLUX single-18	T5 special	50.3	near-flat	97.3 / 96.7	block-dependent carrier
SD2.1	structural cross-attn	high	+0.0001	75.8 / 65.3	lower recorded VLM rates
SD1.5	structural cross-attn	high	+0.0007	66.4 / 56.2	U-Net entanglement boundary
SDXL	structural cross-attn	high	small	57.0 / 37.1	stronger entanglement boundary
SD3.5	layer-16 T5 padding	floor-limited	–	–	anatomy shift, high floor