CONE GEOMETRY IS A ROUTING-CONTROLLED STATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformer representations often concentrate along shared directions. We study how attention routing, the allocation of attention weights over value vectors, shapes this geometry under fixed inputs and model weights. We measure cosine-cone concentration by mean pairwise cosine similarity and intervene by suppressing attention to selected recipient positions. In Stable Diffusion 3, suppressing special text recipients opens the image-token cone, whereas suppressing padding recipients narrows it at the examined layer and denoising step. In language models, response signs vary with intervention and readout depth, including in a 14-billion-parameter model. A 16-model reanalysis finds matching signs across two datasets for 12 of 13 models with non-negligible responses on both. We derive the exact value-replacement update induced by suppression and its first-order geometric response. Diagnostics at four selected sites separate cosine-readout and propagation errors; component interventions reveal contrasting effects of shared and token-specific updates. Under the tested interventions, cosine-cone geometry responds to routing changes in ways that depend on recipient identity, network depth, and model.

1 INTRODUCTION

Transformers build contextual token representations through attention-based mixing of value vectors, residual connections, and feed-forward transformations (Vaswani et al., 2017). Directional concentration, in which representations align along shared directions, is a recurring observation in learned embeddings and contextual representations (Gao et al., 2019; Ethayarajh, 2019). This geometry reflects both learned representations and the computations that mix them during a forward pass. We examine the contribution of attention routing: the allocation of attention weights over recipient key positions and their associated value vectors.

Attention sinks bring the relationship between routing and geometry into focus (Barbero et al., 2025; Queipo-de Llano et al., 2025). These are positions that attract substantial attention despite carrying little semantic information; retaining initial sinks can stabilize streaming inference (Xiao et al., 2024). Their concentrated incoming attention provides a concrete intervention site. Suppressing a sink redistributes its attention weight among other recipients, changing the mixture of value vectors. The resulting geometric response can therefore depend on both the selected recipient and the transformations that follow the edit.

We ask how suppressing an existing route changes downstream directional concentration. Value-aware analyses show that the contribution of a route depends on the vectors being mixed (Kobayashi et al., 2020). We apply a negative bias to selected pre-softmax attention scores with inputs and pretrained weights fixed, then measure mean off-diagonal cosine similarity among valid tokens. We call a decrease in this statistic cone opening and an increase cone narrowing. This paired intervention measures the geometric effect of redistributing attention away from a specified recipient.

Stable Diffusion 3 (SD3) (Esser et al., 2024) provides a within-model comparison: suppressing special text recipients, including end-of-text (EOT)-like positions, opens the image-token cone, while suppressing padding positions used to fill the text sequence narrows it at the examined site (Figure 1). We call these token families structural recipients because they are identified by their sequence role. We then hold the first-token position fixed in language models (LMs) and vary the intervention layer and the readout layer, where geometry is measured; response signs can reverse during propagation (Section 4.2). A fixed-depth reanalysis of 16 models on WikiText-2 (Merity et al., 2016) and Open-

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.

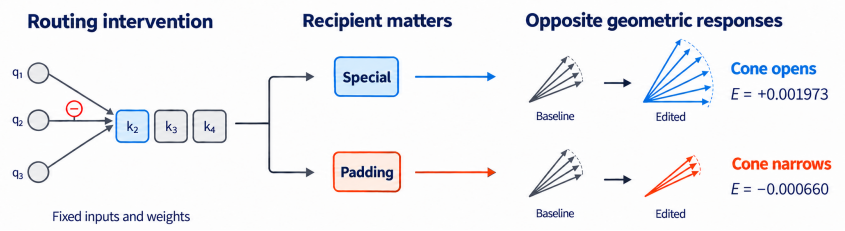

Figure 1: **Recipient-side routing suppression can produce opposite geometric responses.** With inputs and model weights fixed, suppressing attention to special text recipients, including end-of-text (EOT)-like positions, opens the image-token cosine cone, whereas suppressing padding-like recipients narrows it in Stable Diffusion 3 (SD3) at layer 12 and denoising step 0. The reported mean effects use $E = C_{\text{baseline}} - C_{\text{edited}}$, where C is mean off-diagonal cosine similarity; positive E denotes cone opening. The vector fans illustrate angular concentration schematically and are not measured projections or drawn to scale.

WebText (Gokaslan & Cohen, 2019) finds matching signs for 12 of 13 models with non-negligible responses on both, while response directions differ across models (Table 3). These comparisons establish recipient-, depth-, and model-dependent responses under the respective intervention protocols.

Our contributions are threefold. (1) We characterize signed geometric responses across recipient families, intervention and readout depths, and models, and assess their robustness to suppression strength and selected-token inclusion. (2) We derive the value-replacement update induced by routing suppression and its propagated first-order geometric response, linking a specified attention edit to changes in cosine-cone concentration. (3) At four selected diagnostic sites, we separate finite-update prediction errors into cosine-readout and propagation terms. Component interventions further reveal contrasting effects of shared and token-specific updates (Section 5.2).

2 Related Work

Prior work explains attention sinks through training dynamics (Gu et al., 2025) and connects them to massive activations, unusually large hidden-state coordinates (Sun et al., 2024). Vision-transformer registers provide dedicated tokens for internal computation (Darcet et al., 2024). Mixing and compression analyses relate these phenomena to representation geometry (Barbero et al., 2025; Queipo-de Llano et al., 2025). Our experiments complement these accounts of sink formation and function by measuring the signed geometric response to suppression of existing routes.

Representation studies characterize anisotropy, or directional non-uniformity (Ethayarajh, 2019), and rank collapse, where token representations approach a lower-dimensional subspace, in pure-attention networks (Dong et al., 2021). Rogue-dimension analyses identify coordinates that dominate cosine similarity (Timkey & Van Schijndel, 2021). Our paired cosine and centered spectral readouts distinguish changes in angular concentration from changes in how variance is distributed across directions.

Attention attribution studies examine how distinct attention weights can produce the same output under specified dimensional conditions (Brunner et al., 2019) and information flow across layers (Abnar & Zuidema, 2020); attention-control methods modify routing for image editing (Hertz et al., 2022). Our value-replacement identity specifies the update induced by suppression, and the propagation diagnostics test its geometric consequences. Appendix D provides the extended comparison with geometry, outlier, attribution, and intervention studies.

3 Measuring and Intervening on Cosine-Cone Geometry

To measure how routing shapes representation geometry, we pair a specified attention edit with a geometric readout under fixed inputs and model weights. We first define cosine-cone concentration,

then specify recipient selection and suppression. Three experimental protocols use these definitions to examine recipient identity, propagation across depth, and variation across models.

3.1 Geometry Readout

Let $H \in \mathbb{R}^{n \times d}$ denote the valid token representations at a measured layer. We use the mean off-diagonal cosine similarity as the main cone statistic:

$$\mathcal{C}(H) = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{\langle h_i, h_j \rangle}{\|h_i\|_2 \|h_j\|_2}. \tag{1}$$

For baseline H and intervened $\widetilde{H}$, we report cone opening as $\Delta\mathcal{C} = \mathcal{C}(H) - \mathcal{C}(\widetilde{H})$. Positive values denote opening and negative values denote narrowing. We reserve P for attention matrices below, so $\mathcal{C}$ always denotes a representation-geometry statistic.

The statistic has an exact centroid form. Let $\hat{h}_i = h_i / \|h_i\|_2$ and $\mu(H) = n^{-1} \sum_i \hat{h}_i$ be the normalized-token centroid. Since $\sum_{i \neq j} \hat{h}_i^\top \hat{h}_j = \| \sum_i \hat{h}_i \|_2^2 - n$,

$$\mathcal{C}(H) = \frac{n \|\mu(H)\|_2^2 - 1}{n - 1}. \tag{2}$$

Cone opening is therefore a statement about the norm of the mean normalized direction, not only about the centered covariance spectrum. A common translation can change cosine concentration while preserving pairwise Euclidean distances and the centered spectrum. We therefore also measure spectral entropy, effective rank, participation ratio, and maximum explained variance of centered hidden states (definitions in Appendix E). Section 4.4 tests whether these readouts agree with cone opening.

3.2 Recipient Selection and Routing Suppression

For a head with attention matrix $P \in \mathbb{R}^{q \times k}$, the incoming mass of key j is $m_j = |\mathcal{Q}|^{-1} \sum_{i \in \mathcal{Q}} P_{ij}$, where $\mathcal{Q}$ contains the valid query positions. Candidate sink recipients are the top-K keys by incoming mass, computed separately for each head and layer; sink-filtered audits use $K = 1$ unless stated otherwise. In SD3, $\mathcal{Q}$ is the image-query set inside the Multimodal Diffusion Transformer (MM-DiT). Selected keys are mapped to text-token families using the tokenizers of the Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) and Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020) text encoders. We call padding and special/EOT-like text positions structural recipients and split them into carrier families. In the LM depth and breadth audits, the first-token condition selects position zero only in heads where it belongs to the top-K incoming-mass keys. Content controls use the same top-K family filter; second-token and early-token controls use separate position-selection rules (Appendix A).

For pre-softmax logits z_{ij} and selected recipients $\mathcal{S}$, we apply $\widetilde{z}_{ij} = z_{ij} - M\mathbf{1}[j \in \mathcal{S}]$ with $M = 10^4$. The model then recomputes attention and downstream states. Masking is per-head, with recipient selection determined by the corresponding protocol; other attention weights renormalize. Hidden states are not directly edited in this routing intervention.

3.3 Experimental Protocols

We use three protocols to distinguish carrier identity, depth dependence, and model breadth (Table 15). The SD3 audit uses the 553 GenEval prompts (Ghosh et al., 2023) and four random seeds. Its headline carrier comparison measures image tokens at layer 12 and denoising step 0, where the signed split is clearest. A separate temporal audit tracks the persistence of an edit made at step 0. The LM layer-local audits use WikiText-2 sequences of length 512: 128 sequences for GPT-2 (Radford et al., 2019) and Qwen3-0.6B, and 64 for Qwen3-8B and Qwen3-14B (Yang et al., 2025). The 16-model breadth analysis additionally covers Pythia (Biderman et al., 2023), OLMo (Groeneveld et al., 2024), Mistral (Jiang et al., 2023), TinyLlama (Zhang et al., 2024), Phi-2 (Javaheripi et al., 2023), and StableLM-3B-4E1T (Tow et al.), using WikiText-2 and OpenWebText. Additional layer-local audits provide architecture checks.

Table 1: SD3 at layer 12, step 0. Positive effects denote cone opening; brackets give bootstrap 95% intervals over prompt/seed rows. Complete conditions appear in Table 16.

Recipient	Effect [95% CI]	Positive fraction
Special/EOT-like	$0.001973\,[0.001950, 0.001996]$	1.000
Padding-like	$-0.000660\,[-0.000678, -0.000643]$	0.056
Content control	$0.000118\,[0.000113, 0.000124]$	0.624

Geometry is computed over attention-mask-valid tokens. Positive fraction is the fraction of samples with $\Delta\mathcal{C} > 0$. SD3 confidence intervals (CIs) resample prompt/seed rows; LM intervals resample text-level rows after aggregation across four GPU shards. Layer-local examples include sites selected for large effects, so their intervals describe uncertainty conditional on selection. The breadth readout rule avoids choosing the largest observed effect within each run, but it is a reanalysis of existing runs rather than a prospectively registered experiment. Position/content controls, selection details, and robustness protocols are specified in Appendix A.

4 Signed Responses to Routing Interventions

We first compare different structural carriers in SD3, then examine how the response to a first-token routing edit varies across LM depth. A separate breadth analysis assesses model and dataset dependence under a common readout rule. Controls subsequently test whether these observations depend on the mask magnitude, inclusion of the selected token, or choice of geometry metric.

4.1 Opposite Responses across SD3 Carrier Families

Suppressing structural-recipient attention in SD3 does not have a single geometric direction. At layer 12 and step 0, special/EOT-like suppression opens the image-token cone, with $\Delta\mathcal{C} = 0.001973$ and bootstrap 95% interval $[0.001950, 0.001996]$. Padding-like suppression instead narrows it, with $\Delta\mathcal{C} = -0.000660$ and interval $[-0.000678, -0.000643]$. Table 1 contrasts these effects with a content control; the CLIP-specific conditions reproduce the sign split (Table 16). The combined structural-sink condition opens the cone but conceals the negative padding response. Content and image controls are much smaller (Table 16). These comparisons distinguish the specified edits without matching attention mass or update energy. Their absolute magnitudes are not directly comparable to LM effects because the architectures, readouts, and intervention timing differ.

The step-0 response decays over subsequent denoising steps (Figure 4); this persistence audit does not test independent later-step edits. Having established different signs across recipients, we next fix the recipient to the first token and vary depth.

4.2 Depth-Dependent Responses in Language Models

The layer-local LM protocol separates where a route is edited from where its geometric consequence is measured. In Fig. 2, rows index target layers and columns index measured layers. Each row traces the response to a single-layer first-token intervention through the remainder of Qwen3-14B. The heatmap contains signed depth bands rather than a uniformly positive response. A first-token label alone is therefore insufficient to specify whether an edit opens or narrows the measured cone. Table 2 includes positive sites across GPT-2 and Qwen3 scales, with Qwen3-14B opening at early, middle, and late locations. The full sweep retains negative bands that these selected positive examples would otherwise conceal.

Heterogeneous models further separate the direction of a response from the recipient's first-token identity. Mistral-7B has an early positive route at $L1 \to L2$ (0.028795), but a negative same-layer response at $L4 \to L4$ (-0.018341). The selected OLMo-7B and TinyLlama sites are also negative. Position/content controls are small relative to many of these responses, although their perturbation doses are not matched (Section 4.4). Mistral provides an architecture check on depth dependence, rather than simply another scale endpoint. Across these examples, the common observation is sensitivity to a specified route, with both signs represented. This depth dependence motivates a fixed readout rule for the broader model comparison.

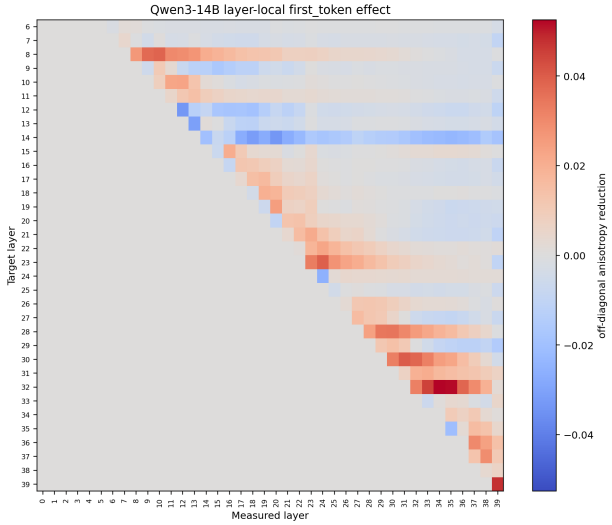

Figure 2: Qwen3-14B full layer-local first-token sweep. Rows are intervention target layers and columns are measured layers; color shows off-diagonal anisotropy reduction. The heatmap exposes signed depth-local bands rather than a uniform first-token effect. Representative positive sites include $L8 \to L10$, $L23 \to L24$, $L32 \to L34$, and $L39 \to L39$, while several early and mid target layers produce the opposite sign.

Table 2: Selected layer-local LM interventions. Ctrl. ratio is the strongest available position/content control divided by the route magnitude; norm ratio is the largest output-norm stress divided by the route magnitude when available. Positive and negative rows describe signed routing responses. Control doses are not matched; same-layer radial stresses preserve cosine by construction, so the norm ratios do not identify upstream norm mediation.

Setting	Route	Effect	Ctrl.	Norm	Role
GPT-2	$L7 \to L9$	0.031615	0.0036	9.92×10^{-4}	small LM positive
Qwen3-0.6B	$L25 \to L25$	0.034873	0.0005	9.33×10^{-6}	modern small LM positive
Qwen3-8B	$L25 \to L26$	0.031315	–	–	fixed-depth positive, deeper band
Qwen3-14B early	$L8 \to L10$	0.039213	0.0019	8.12×10^{-4}	large-model route
Qwen3-14B mid	$L23 \to L24$	0.039532	0.0014	1.57×10^{-4}	depth-local route
Qwen3-14B late	$L32 \to L34$	0.052856	0.0035	2.85×10^{-4}	strongest route
Qwen3-14B endpoint	$L39 \to L39$	0.047470	0.0483	1.06×10^{-5}	downstream endpoint
Mistral-7B early	$L1 \to L2$	0.028795	0.0364	5.57×10^{-4}	architecture-positive check
Mistral-7B signed	$L4 \to L4$	-0.018341	0.0098	1.04×10^{-5}	negative sign check
OLMo-7B signed	$L28 \to L28$	-0.118597	0.0011	1.09×10^{-6}	negative sign check
TinyLlama signed	$L4 \to L4$	-0.118547	0.0013	9.13×10^{-7}	negative sign check

Propagation can reverse the local response. In the full Qwen3-14B aggregate sweep, 24 of 34 target layers have a downstream response opposite to the same-layer response when both exceed 10^{-4} in absolute value. The count remains 24 at 10^{-5} and is 22 at 10^{-3}. These exploratory counts use correlated aggregate rows, not independent samples. Upstream effects remain below 4.95×10^{-7}. Appendix E gives baseline reconstruction, sample-count caveats, and the corresponding curves (Figure 5). The sign reversals motivate propagating the target update before predicting its geometric effect.

Table 3: Fixed-rule 16-model breadth matrix in two parallel panels. For each model, the measured layer is selected by the common rule $\mathrm{round}(0.75\,(L-1))$. This reanalysis uses existing breadth runs with simultaneous interventions across target layers; it is not a single-layer intervention or a prospectively registered new sample. Values report the first-token cone effect on WikiText-2 (WT2) and OpenWebText (OWT); † marks near-zero entries with $|\Delta\mathcal{C}| < 10^{-3}$ whose sign is not treated as determinate. Raw signs agree for 14 of 16 models; among the 13 models with determinate signs on both datasets, signs agree for 12. Signed dataset correlation is 0.656 and absolute-effect correlation is 0.469.

Model	Layer	WT2	OWT	Sign	Model	Layer	WT2	OWT	Sign
GPT-2	8	0.092747	0.069674	+/+	Qwen3-0.6B	20	0.021668	0.036008	+/+
GPT-2-medium	17	0.103678	0.071492	+/+	Qwen3-1.7B	20	0.008799	0.026980	+/+
Pythia-70M	4	-0.002494	-0.001159	-/-	Qwen3-4B	26	0.014408	0.023605	+/+
Pythia-410M	17	-0.032251	-0.019592	-/-	Qwen3-8B	26	0.022398	0.011125	+/+
Pythia-1B	11	-0.010269	-0.004797	-/-	Qwen3-14B	29	$-0.000194^{\dagger}$	0.004869	0/+
OLMo-1B	11	-0.042375	$-0.000567^{\dagger}$	-/0	Phi-2	23	-0.006361	-0.002233	-/-
OLMo-7B	23	-0.061684	-0.049068	-/-	StableLM-3B	23	-0.191025	0.011335	-/+
Mistral-7B	23	-0.002990	$-0.000528^{\dagger}$	-/0	TinyLlama-1.1B	16	-0.083320	-0.054064	-/-

4.3 Cross-Model and Cross-Dataset Breadth

To compare models without selecting their largest measured effect, we apply the common readout rule $\mathrm{round}(0.75(L-1))$, where L is the number of layers and indices start at zero, to the existing breadth runs. Unlike Section 4.2, these runs intervene simultaneously across target layers. Their response is therefore a different quantity from a single-layer propagation curve. Table 3 reports all 16 models on both datasets so that positive, negative, and near-zero entries remain visible.

Using raw signs, six models have positive effects on both datasets and eight have negative effects on both; overall sign agreement is 14/16. Some entries are sufficiently small that their signs should not carry the same weight as larger responses. Treating $|\Delta\mathcal{C}| < 10^{-3}$ as a practical near-zero region leaves 13 models with determinate signs on both datasets, of which 12 agree. StableLM-3B is the remaining determinate sign reversal. The signed-effect correlation across datasets is 0.656, while the absolute-effect correlation is 0.469. These summaries support mostly stable directions across the two text distributions under this protocol, with substantial model dependence and explicit exceptions.

The conclusion does not depend uniquely on choosing the 0.75 depth fraction. Repeating the readout at fractions 0.50 and 0.90 yields determinate-sign agreement of $10/12$ and $12/14$, respectively, compared with $12/13$ at 0.75. The changing set of sign reversals reinforces the depth dependence seen in the layer-local study. This check reduces reliance on one readout choice, but it neither turns the reanalysis into prospective validation nor makes the near-zero threshold a significance test. The breadth evidence complements the layer-local results: response directions can recur across datasets while varying across models and measurement depths.

4.4 Controls and Robustness

The selected-site response signs survive changes in mask strength and token inclusion. Across 18 first-token sites, 17 text-bootstrap 95% intervals exclude zero; the remaining OLMo site is near zero. Sweeping $M \in \{10, 100, 10^3, 10^4\}$ preserves all 18 signs, with mean absolute effect 0.031615 at $M = 10$ and 0.031653 for $M \geq 100$. Excluding the selected recipient from both baseline and intervened readouts also preserves all 18 signs, with median magnitude ratio $|\Delta\mathcal{C}_{\mathrm{exclude}}|/|\Delta\mathcal{C}| = 1.005$. The response persists among other tokens and does not require the default 10^4 bias. The intervals remain conditional on site selection, and the mask sweep does not establish a small-perturbation linear regime.

Position/content controls and radial stresses answer narrower questions about alternative edits. The former target second-token, early-token, or content-like recipients, and their responses are small relative to many selected first-token effects in Table 2. However, selected-head count, incoming mass, and update energy can differ. The radial stresses scale decoder-block outputs; positive scaling

preserves same-layer cosine by construction, and downstream normalization can attenuate its consequences. Their small ratios cannot exclude an upstream massive-activation mechanism. Appendix A gives the protocols and measurements needed to interpret these comparisons.

Finally, cosine opening and spectral decompression are related but separable outcomes. In the Qwen3-14B audit, $L32 \rightarrow L34$ increases effective rank by 0.391 and decreases maximum explained variance by 0.00818. Yet $L8 \rightarrow L10$ opens the cosine cone without a matching spectral decompression signature. Across the 160-row spectral sweep, cone effects correlate with spectral entropy change ($r = 0.410$) and decreased leading-direction dominance ($r = 0.487$), but unthresholded sign agreement is only 49/160, with many near-zero cone rows. The main claim consequently concerns the measured cosine response; spectral or information-loss claims require their own evidence.

5 Interpreting the Routing Response

5.1 From Routing Suppression to Cone Geometry

The signed responses motivate examining what the routing edit actually changes. For one suppressed recipient s, let a_{ij} be the baseline attention weights, v_j the value vectors, and $o_i = \sum_j a_{ij} v_j$. For $a_{is} < 1$, define the alternative mixture $\bar{o}_{i,-s} = (1 - a_{is})^{-1} \sum_{j \neq s} a_{ij} v_j$. With $\beta = e^{-M}$, the exact attention-output difference is

$$\widetilde{o}_i - o_i = \kappa_M(a_{is})(\bar{o}_{i,-s} - v_s), \qquad \kappa_M(a) = \frac{a(1-a)(1-\beta)}{1 - a + \beta a}. \tag{3}$$

In the infinite-mask limit, this becomes $a_{is}(\bar{o}_{i,-s} - v_s)$. Suppression therefore replaces the selected value contribution with a renormalized mixture of alternatives. Incoming mass weights this contrast but does not specify its direction. A causal query with no unselected accessible key is unchanged by a finite bias and is excluded from the replacement formula. The finite-mask remainder is provided in Appendix B.

The attention-output contrast must propagate to the measured representation before it can determine a cosine response. Let B collect the infinite-mask updates after the attention output projection, and let Φ be the downstream map to the measured hidden states. Projection biases cancel in the increment. Writing $U^{(1)} = D\Phi_0(B)$ for the first-order propagated update, where $D\Phi_0$ is the derivative of the downstream map at the unmodified state, its contribution to cone opening is

$$-\frac{2n}{n-1} \left\langle \mu(H), \frac{1}{n} \sum_i \frac{(I - \hat{h}_i \hat{h}_i^\top) u_i^{(1)}}{\|h_i\|_2} \right\rangle. \tag{4}$$

The token-normalization derivative removes each update's radial component before its alignment with the normalized centroid is evaluated. Both the value contrast and downstream propagation can therefore matter for the response direction. Appendix B derives the normalization, propagation, and finite-mask remainders and states the condition under which the first-order term determines the sign. We test this derivative by reconstructing the target update, evaluating its propagated Jacobian-vector product (JVP), and measuring finite-update errors. The historical raw-alignment probe omits the normalization derivative and is reported separately in the appendix; it does not validate Eq. 4.

5.2 Testing the Local Approximation at Finite Amplitude

A correct local derivative need not predict a full routing intervention. We examine four previously selected diagnostic sites: GPT-2 L7→L9, Qwen3-14B L23→L24, TinyLlama L4→L4, and OLMo-7B L28→L28. For these diagnostics, the intervention targets position zero in every head of the target layer. The experiment uses the same 64 articles across sites, yielding 256 article-site pairs rather than 256 independent articles. The collection pipeline verifies native attention-update reconstruction and checks the baseline hidden-state JVP against centered finite differences. The numerical audit accepts all 512 finite-difference checks. These implementation checks validate the derivative calculation, not its finite-amplitude predictive accuracy. Site selection is historical; the four sites are not a random sample of model families.

Site	$E(1)$	Linear MAE	Full-cosine MAE	Reduction
GPT-2	$+0.027178$	0.004477	0.011524	-157.4%
Qwen3-14B	$+0.040000$	0.001800	0.002540	-41.1%
TinyLlama	-0.146305	0.069527	0.044983	35.3%
OLMo-7B	-0.153018	0.064790	0.016777	74.1%

Table 4: Finite-amplitude diagnostics on 64 shared articles. MAE is averaged over the four specified nonzero amplitudes; $E(1)$ is the endpoint effect. Negative reduction denotes increased error. Only OLMo meets the original 50% improvement criterion with its multiplicity-adjusted interval.

The diagnostic separates readout nonlinearity from nonlinear propagation. Let $H(t) = \Phi(tB)$, $H_0 = H(0)$, and $D = H'(0)$. Along the fixed target-update path we evaluate

$$E_{\text{lin}}(t) = -t\, DC(H_0)[D], \tag{5}$$

$$E_{\cos}(t) = C(H_0) - C(H_0 + tD), \tag{6}$$

$$E(t) = E_{\text{lin}}(t) + \underbrace{[E_{\cos}(t) - E_{\text{lin}}(t)]}_{R_{\text{readout}}(t)} + \underbrace{[E(t) - E_{\cos}(t)]}_{R_{\text{propagation}}(t)}. \tag{7}$$

Here "full cosine" evaluates the nonlinear readout on the linear hidden path. The scale t multiplies the reconstructed update, not the logit bias; $E_{\cos}$ therefore need not equal the actual response E.

Correcting only the readout is not uniformly beneficial (Table 4, Figure 3). The frozen criterion requires at least a 50% reduction in prediction mean absolute error (MAE), aggregated within each article over $t \in \{0.25, 0.5, 0.75, 1\}$. Article bootstrap intervals use 20,000 draws and correction over the original family of 20 diagnostic comparisons. Only OLMo meets this criterion. TinyLlama improves by 35.3%, below the criterion, while GPT-2 and Qwen3-14B become less accurate. The percentages summarize mean errors; the criterion is assessed using the paired article score and its adjusted interval.

The two remainders can oppose each other, explaining why adding a readout correction can worsen the total prediction. At the GPT-2 endpoint, $E_{\text{lin}} = 0.036400$, $R_{\text{readout}} = 0.014666$, and $R_{\text{propagation}} = -0.023888$, yielding $E = 0.027178$. Keeping only the positive readout correction removes a cancellation that made the linear estimate closer to the observation. TinyLlama additionally shows a non-monotonic mean response: $E(0.1) = 0.001485$ but $E(1) = -0.146305$. Thus finite-update direction and magnitude cannot generally be extrapolated from the infinitesimal slope. This decomposition diagnoses prediction error; it does not localize the propagation remainder to a particular downstream module.

5.3 Common Displacement Is Not a Universal Mechanism

Readout reconstruction does not identify a universal target-layer component. We split the target update into a token-common component (the mean update broadcast to all tokens) and a centered component (each token's residual after subtracting that mean), and propagate each separately. In GPT-2, the full update opens the cone by 0.027178, but common-only narrows it by 0.008948 and centered-only opens it by 0.039510. Qwen3-14B instead supports a common-mode contribution: common-only produces 0.026399 opening versus 0.040000 for the full update and 0.016647 for centered-only, meeting all four original diagnostic criteria, including comparison with equal-norm random common directions. These contrasting cases limit the common-displacement explanation to particular sites. Common and centered patches are not energy-matched, and their propagated responses need not add to the full response.

A follow-up on 64 new shared articles retains the scalar cone-error rankings but rejects the stronger joint mechanism hypothesis: GPT-2's centered patch is worse at reconstructing the common hidden vector. The normalized-centroid audit confirms that smaller vector error need not imply smaller scalar cone error (Appendix E). Thus neither a readout reconstruction nor a favorable patch ranking alone establishes a universal component mechanism.

These edits establish localized causal responses, following the logic of activation interventions (Vig et al., 2020; Meng et al., 2022). Patching conclusions can depend on the corruption method and metric (Zhang & Nanda, 2024), and successful subspace edits may activate dormant pathways (Makelov

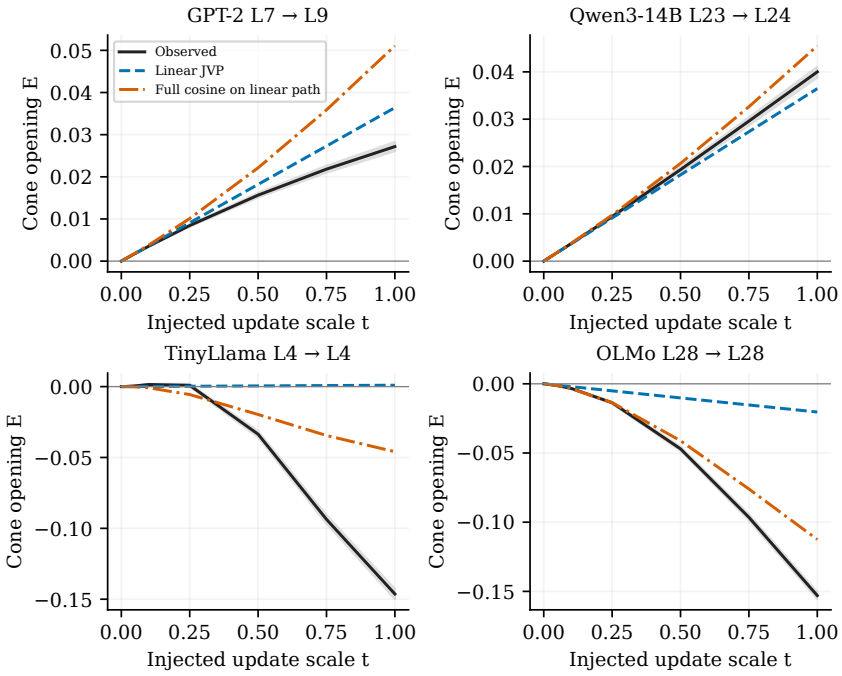

Figure 3: Observed and predicted effects along the injected-update path. Shading shows the original pointwise article-bootstrap intervals for the observed effect, not simultaneous confidence bands. The linear estimate and full cosine on the linear path share the same baseline JVP. These curves retain both opening and narrowing diagnostic sites.

et al., 2024). We therefore distinguish an intervention's effect from identification of the mechanism operating in the unmodified network. Section 6 consolidates the remaining causal and functional limits.

6 Discussion and Limitations

Routing edits produce context-dependent cosine responses, but the SD3 and LM experiments do not establish an identical internal mechanism. Three limits remain. First, position/content controls do not match intervention dose, and selected-site intervals do not correct for site selection or dependence. Second, four-site JVP checks validate the local calculation without establishing a general held-out predictor; component patches retain nonlinear interactions. Historical radial stresses do not test upstream sink formation. Third, concurrent changes in geometry and next-token distributions do not establish mediation, task-quality gains, or useful editing policies. The SD3 evidence concerns a selected layer and step. Its temporal audit tracks persistence after a step-0 edit, leaving independent later-step interventions and image-output effects untested. Further mechanism tests should match dose, specify the candidate downstream component, and evaluate predictions on new sites; extending model count alone would not resolve these identification gaps.

7 Conclusion

Recipient-side routing suppression produces both cone opening and cone narrowing. The SD3 recipient comparison and language-model depth and breadth audits show how response direction varies with token family, intervention layer, readout layer, and model. The value-replacement identity connects suppression to a concrete attention update, while finite-amplitude diagnostics separate readout and propagation errors. Component interventions further reveal distinct effects of shared and token-specific updates.

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

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

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.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*, 2019.

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.

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.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pp. 2793–2803. PMLR, 2021.

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.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Zizhuo Fu, Wenxuan Zeng, Runsheng Wang, and Meng Li. Attention sink forges native moe in attention layers: Sink-aware training to address head collapse. *arXiv preprint arXiv:2602.01203*, 2026.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019.

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.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 15789–15809, 2024.

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.

Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. Attention score is not all you need for token importance indicator in kv cache reduction: Value also matters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21158–21166, 2024.

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.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, 2020.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. Bert busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3392–3405, 2021.

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, 2026. URL https://arxiv.org/abs/2601.15380.

Chen Liu, Xingzhi Sun, Xi Xiao, Alexandre Van Tassel, Ke Xu, Kristof Reimann, Danqi Liao, Mark Gerstein, Tianyang Wang, Xiao Wang, et al. Dispersion loss counteracts embedding condensation and improves generalization in small language models. *arXiv preprint arXiv:2602.00217*, 2026.

Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. Positional artefacts propagate through masked language model embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5312–5327, 2021.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

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.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1286–1304, 2022.

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*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink. *Advances in Neural Information Processing Systems*, 38:123482–123521, 2026.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. *arXiv preprint arXiv:2603.05498*, 2026.

William Timkey and Marten Van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, 2021.

Jonathan Tow, Marco Bellagente, Dakota Mahan, and Carlos Riquelme. Stablelm 3b 4e1t. URL `[https://huggingface.co/stabilityai/stablelm-3b-4e1t](https://huggingface.co/stabilityai/stablelm-3b-4e1t)`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf`.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. In *International Conference on Learning Representations*, volume 2025, pp. 144–164, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

A Additional Audits and Protocol Details

A.1 Claim-to-Evidence Map

Table 5 distinguishes observed responses from additional causal interpretations that they do not establish.

Table 5: Claims, observations, and scope of inference.

Claim	Evidence	Scope
Carrier-dependent SD3 response	Special/padding split at layer 12, step 0	One diffusion model and selected site; does not identify a common LM mechanism.
Signed LM response	Fixed-rule breadth and layer-local examples	Fixed-depth reanalysis and selected-site effects are different estimands.
Weak radial-control response	Decoder-block output scaling	Same-layer cosine invariance and downstream normalization limit interpretation.
Local response decomposition	Replacement identity and normalized derivative	Historical raw-alignment accuracy does not validate the corrected derivative.
Output-distribution movement	Negative log-likelihood (NLL), Kullback-Leibler (KL) divergence, and top-1 prediction changes	Does not establish cone mediation or task improvement.
Persistence of initial SD3 edit	Later-step measurements following step-0 intervention	Not an intervention-time sensitivity sweep.

A.2 Layer-Local and Control Protocol

The LM layer-local audits suppress recipient-side attention to a selected carrier at one target layer and evaluate hidden-state cone geometry at all measured layers. For each target layer, the reported positive site is the measured layer with the largest off-diagonal anisotropy reduction. Same-target rows are also inspected because they can reveal opposite-signed local effects. These layer-local

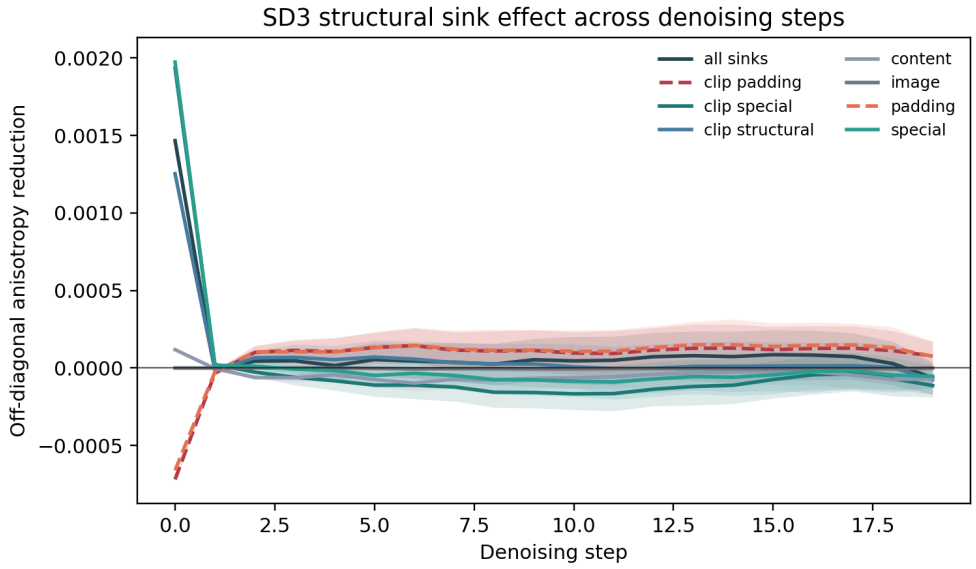

Figure 4: Persistence of an SD3 step-0 intervention at layer 12. The edit is active only at denoising step 0; later measurements track its propagated effect after the edit has been switched off. The signed split decays subsequently. This curve does not compare independent interventions at different denoising steps.

rows are used as controlled examples and mechanism-localizing checks; the fixed-depth breadth table is a common-rule reanalysis of existing all-target-layer runs, not a prospective single-layer confirmation. Position controls suppress second-token or early non-structural recipients, and content controls suppress content-like recipients selected by the same sink-recording machinery. These controls need not match selected-head count, attention mass, or update norm; differences cannot isolate structural identity alone. The hidden-state geometry is computed over attention-mask-valid nonpadding tokens. In the exclude-selected audit, the selected key/recipient position is removed from both the baseline and intervened valid-token sets before computing pairwise cosines, so the reported effect must arise among the remaining tokens rather than from the ablated vector's own cosine entries.

Mistral-7B is used as an architecture check rather than as a larger-scale endpoint. Its strongest first-token route is $L1 \rightarrow L2$; later same-target first-token interventions at $L4$ and $L31$ are negative. This is why the paper treats Mistral as evidence for signed, depth-local carrier behavior instead of as a monotone first-token story.

A.3 Radial-Control Protocol and Identification Limits

Both historical *output-norm* and *residual-norm* scripts attach their scaling hook to a decoder block's output. Replay matches the selected token's block-output norm from a routing run. Fixed stresses use radial factors $0.9, 0.95, 1.05, 1.1$ or $0.9, 1.1, 1.25, 1.5$. These are related interventions at the same type of location, not independent attention-output and residual-stream causal channels.

A positive radial scaling leaves same-layer cosine exactly unchanged in exact arithmetic, including several same-layer rows in the table below. Spectral changes under scaling do not remove that invariance. At downstream layers, normalization and the absence of a perturbation to upstream outlier formation remain alternative explanations for a weak response. Expected hook-scale changes are not a substitute for directly recorded post-hook and downstream activations. The ratios are retained as descriptive measurements, not evidence excluding all norm-mediated pathways.

Table 6: Historical radial-control measurements. Ratios divide a radial control effect by routing-effect magnitude. Both fixed-stress scripts scale decoder-block outputs. Same-layer ratios are subject to cosine invariance and do not test an upstream norm mechanism.

Model/site	Route	Routing	Replay	Output stress	Residual stress
GPT-2	$L7 \to L9$	0.031539	2.26×10^{-6}	9.92×10^{-4}	1.30×10^{-3}
Qwen3-0.6B	$L25 \to L25$	0.034873	9.33×10^{-6}	9.33×10^{-6}	–
Qwen3-14B early	$L8 \to L10$	0.039213	6.51×10^{-6}	8.12×10^{-4}	–
Qwen3-14B mid	$L23 \to L24$	0.039532	9.80×10^{-6}	1.57×10^{-4}	4.91×10^{-4}
Qwen3-14B late	$L32 \to L34$	0.052856	8.62×10^{-6}	2.85×10^{-4}	2.18×10^{-4}
Qwen3-14B endpoint	$L39 \to L39$	0.047470	2.36×10^{-5}	1.06×10^{-5}	3.3×10^{-5}
Mistral-7B early	$L1 \to L2$	0.028795	7.26×10^{-6}	5.57×10^{-4}	1.14×10^{-3}
Mistral-7B signed	$L4 \to L4$	-0.018341	1.04×10^{-5}	1.04×10^{-5}	1.04×10^{-5}
OLMo-7B signed	$L28 \to L28$	-0.118597	1.09×10^{-6}	1.09×10^{-6}	1.09×10^{-6}
TinyLlama signed	$L4 \to L4$	-0.118547	9.13×10^{-7}	9.13×10^{-7}	9.13×10^{-7}

A.4 Fixed-Depth Reanalysis

The common depth rule $\mathrm{round}(0.75(L-1))$ selects a measured layer from existing all-target-layer breadth runs. It avoids selecting the maximum effect within each run but does not establish prospective preregistration. The 10^{-3} threshold marks a practical near-zero region, not a confidence interval or significance test. Historical site choices and repeated measurements from one model are not independent held-out trials.

As a sensitivity check, we repeat the fixed-depth analysis at depth fractions 0.50, 0.75, and 0.90. Raw dataset sign agreement is $13/16$, $14/16$, and $14/16$, respectively; after down-weighting $|\Delta\mathcal{C}| < 10^{-3}$ rows, determinate-sign agreement is $10/12$, $12/13$, and $12/14$. This supports the dataset-stability claim as a fixed-depth phenomenon rather than a unique property of the 0.75 rule, while the changing flip set reinforces that signs are depth-banded rather than globally invariant.

A.5 Robustness Audits

Table 7 collects the lightweight robustness checks that are used to guard the main interpretation. These checks are deliberately not presented as additional discoveries. They assess sensitivity to statistical uncertainty, mask magnitude, selected-token inclusion, breadth readout choices, and scalar summaries of SD3 carriers. They do not remove site-selection uncertainty or identify carrier identity at a matched perturbation dose.

A.6 Historical Value-Patch Protocol

Historical value patches subtract selected contributions, remove their projection onto a measured-layer axis, or replace them with norm-comparable content values. The donor search is not query-causal, and copying a measured-layer axis into target-output coordinates does not account for the downstream derivative. Affine projection biases also need to be excluded from increment projections. These controls require reconstruction, causal-visibility, and perturbation-dose checks before they can identify a minimal component.

A.7 Audit Artifact Availability

The manuscript reports experimental protocols and aggregate statistics rather than enumerating file paths. For reproducibility, the analysis pipeline retains the corresponding raw rows, aggregate tables, decision summaries, and plotting inputs used to produce the reported tables and figures. These artifacts record the prompt/text indices, model identifiers, target and measured layers, intervention condition, and hidden-state metrics for each audit.

Table 7: Robustness checks supporting the main routing-control claim.

Concern	Audit	Result
LM uncertainty	Bootstrap over text rows on 18 confirmatory first-token rows	17/18 confidence intervals exclude zero; the remaining OLMo row is near zero.
Extreme mask value	$M \in \{10, 100, 10^3, 10^4\}$ routing-suppression sweep	Mean first-token magnitude is already 0.031615 at $M = 10$ and signs are preserved for all 18 sites.
Selected-token self-effect	Recompute cone geometry after excluding the selected recipient from baseline and intervened token sets	Signs are preserved for all 18 sites; median $\lvert\Delta\mathcal{C}_{\text{exclude}}\rvert/\lvert\Delta\mathcal{C}\rvert = 1.005$.
Fixed-depth freedom	Repeat breadth at depth fractions 0.50, 0.75, and 0.90	Determinate-sign agreement is $10/12$, $12/13$, and $12/14$, respectively.
SD3 temporal localization	20-step SD3 cross-step audit at layer 12	The edit is active at step 0 only; the curve measures subsequent decay, not comparative sensitivity to later-step edits.
SD3 mass/leverage shortcut	Clean-forward carrier-leverage audit with $m_s\lVert W_O v_s\rVert \cos(W_O v_s, \mu)$	Padding carriers have larger mass than special carriers but the opposite sign; the leverage scalar is positive for all carriers and therefore cannot explain the signed split. Paired special-vs-padding contrasts recover the split for write/leverage contrasts, but not for mass contrasts (Table 8).

Table 8: SD3 carrier-contrast diagnostic. Paired special-vs-padding contrasts recover the signed structural split for value/write quantities but not for incoming attention mass. Same-sign pairs count whether the contrast has the same sign as the observed cone-effect contrast across the four structural pairs {special:padding, CLIP-special:CLIP-padding, CLIP-special:padding, special:CLIP-padding}.

Contrast variable	Same-sign pairs	Pearson with observed contrast
Incoming mass	0/4	-0.642
Write norm	4/4	-0.003
Centroid alignment	4/4	-0.519
Mass $\times$ write norm $\times$ centroid alignment	4/4	-0.381

B Local Response Derivation

B.1 Finite-Mask Replacement Remainder

For the single-recipient intervention in Eq. equation 3, the infinite-mask update and finite-mask correction satisfy the following identity. The alternative mixture requires $a_{is} < 1$.

Equivalently,

$$\widetilde{o}_i - o_i = a_{is}(\bar{o}_{i,-s} - v_s) + R^{\text{mask}}_{i,M}, \qquad R^{\text{mask}}_{i,M} = -\frac{\beta a_{is}}{1 - a_{is} + \beta a_{is}}(\bar{o}_{i,-s} - v_s). \tag{8}$$

Thus the infinite-mask value-replacement term has an explicit finite-M remainder, with $\lVert R^{\text{mask}}_{i,M}\rVert_2 \le \beta a_{is}\lVert\bar{o}_{i,-s} - v_s\rVert_2/(1 - a_{is})$ whenever $a_{is} < 1$.

B.2 Normalized Readout Expansion

We will use the following local expansion with an explicit remainder. For a perturbation field $U = \{u_i\}_{i=1}^n$, define $h_i(t) = h_i + tu_i$, $r_i = \lVert h_i\rVert_2$, and $\Pi_i = I - \hat{h}_i\hat{h}_i^\top$. Assume $t \ge 0$ and $t\lVert u_i\rVert_2 \le \rho r_i$

for all i and some $\rho < 1$. Then

$$\widehat{h_i(t)} = \hat{h}_i + t\,\frac{\Pi_i u_i}{r_i} + R_i(t), \tag{9}$$

$$\|R_i(t)\|_2 \le \frac{3t^2\|u_i\|_2^2}{(1-\rho)^2 r_i^2}. \tag{10}$$

Let

$$\eta(U) = \frac{1}{n}\sum_i \frac{\Pi_i u_i}{r_i}, \qquad \bar{R}(t) = \frac{1}{n}\sum_i R_i(t). \tag{11}$$

Combining Eq. equation 2 with $\mu(H+tU) = \mu(H) + t\eta(U) + \bar{R}(t)$ gives the exact decomposition

$$\mathcal{C}(H+tU) - \mathcal{C}(H) = \frac{2nt}{n-1}\langle \mu(H), \eta(U)\rangle + \mathcal{R}_{\mathcal{C}}(t; H, U), \tag{12}$$

$$\mathcal{R}_{\mathcal{C}}(t; H, U) = \frac{2n}{n-1}\langle \mu(H), \bar{R}(t)\rangle + \frac{n}{n-1}\|t\eta(U) + \bar{R}(t)\|_2^2. \tag{13}$$

Thus the first-order sign is governed by the projection of the normalized update centroid $\eta(U)$ on the current cone axis, and the truncation is valid whenever $|\mathcal{R}_{\mathcal{C}}|$ is smaller than the linear term.

B.3 Propagation and the Response Remainder

The expansion above gives a local description of the geometric response. Let Φ denote the downstream map from the target attention-output tensor to the measured hidden-state tensor, with the rest of the network state fixed at the baseline forward pass. For the infinite-mask main term in Eq. equation 8, define the target-block perturbation

$$b_{i,s} = W_O\left[a_{is}(\bar{o}_{i,-s} - v_s)\right]. \tag{14}$$

If $D\Phi_0$ is the Fréchet derivative of Φ at the baseline attention output, the propagated first-order update is

$$u_{i,s}^{(1)} = \left[D\Phi_0(B_s)\right]_i, \qquad B_s = \{b_{i,s}\}_i. \tag{15}$$

For any twice differentiable downstream map whose second derivative is bounded by L_Φ on the segment between the baseline and intervened target outputs, the true hidden update induced by the infinite-mask term satisfies

$$\widetilde{h}_i - h_i = u_{i,s}^{(1)} + R_{i,s}^{\mathrm{prop}}, \qquad \sum_i \|R_{i,s}^{\mathrm{prop}}\|_2^2 \le \frac{L_\Phi^2}{4}\|B_s\|_F^4. \tag{16}$$

The finite-M mask adds the explicit output-space remainder in Eq. equation 8, propagated through the same downstream map.

The "low-rank broadcast" statement is then an empirical property of $U_s^{(1)} = \{u_{i,s}^{(1)}\}_i$, not an algebraic assumption. For any unit direction d_s we can write the exact decomposition

$$u_{i,s}^{(1)} = \gamma_{i,s} d_s + \epsilon_{i,s}, \qquad \gamma_{i,s} = \langle u_{i,s}^{(1)}, d_s\rangle, \qquad \langle \epsilon_{i,s}, d_s\rangle = 0. \tag{17}$$

The route is coherent when a direction d_s chosen from the data leaves small residual energy $\sum_i \|\epsilon_{i,s}\|_2^2$ relative to $\sum_i \|u_{i,s}^{(1)}\|_2^2$. One choice of d_s is the leading right singular direction of the tokenwise update matrix. Whether sink suppression produces such a coherent value-replacement update must be measured; high incoming mass alone does not establish this property.

A first-order sign diagnostic with remainder. Combining Eq. equation 12 with the propagated update gives the local cone-opening decomposition

$$\Delta\mathcal{C}_s = \mathcal{C}(H) - \mathcal{C}(\widetilde{H}_s) = \frac{2n}{n-1}B_s^{\mathrm{cone}} + R_s^{\mathrm{total}}, \tag{18}$$

where

$$\begin{aligned} B_s^{\mathrm{cone}} &= -\left\langle \mu(H), \eta(U_s^{(1)})\right\rangle \\ &= -\frac{1}{n}\sum_i \left\langle \mu(H), \frac{\Pi_i u_{i,s}^{(1)}}{\|h_i\|_2}\right\rangle. \end{aligned} \tag{19}$$

Write $H_1 = H + U_s^{(1)}$, let H_∞ be the fully propagated infinite-mask counterfactual, and let H_M be the finite-mask counterfactual. Define $R_s^{\text{prop}\to\mathcal{C}} = \mathcal{C}(H_1) - \mathcal{C}(H_\infty)$ and $R_s^{\text{mask}\to\mathcal{C}} = \mathcal{C}(H_\infty) - \mathcal{C}(H_M)$. The total remainder is the sum of three explicit sources:

$$R_s^{\text{total}} = -\mathcal{R}_{\mathcal{C}}(1; H, U_s^{(1)}) + R_s^{\text{prop}\to\mathcal{C}} + R_s^{\text{mask}\to\mathcal{C}}, \tag{20}$$

where $\mathcal{R}_{\mathcal{C}}$ is the normalization/cosine remainder in Eq. equation 13, $R_s^{\text{prop}\to\mathcal{C}}$ is induced by the downstream Taylor remainder in Eq. equation 16, and $R_s^{\text{mask}\to\mathcal{C}}$ is induced by the finite-M remainder in Eq. equation 8. The sign diagnostic is therefore valid under the explicit dominance condition

$$|R_s^{\text{total}}| < \frac{2n}{n-1}|B_s^{\text{cone}}|, \qquad \text{sign}(\Delta\mathcal{C}_s) = \text{sign}(B_s^{\text{cone}}). \tag{21}$$

In words, suppressing a recipient opens the cone when the selected route's propagated value-replacement update decreases the normalized centroid norm, after accounting for token normalization. Written in value-contrast form, the main term compares the selected value against the replacement mixture. Let $L_{i\ell}$ be the linear block of $D\Phi_0$ that maps a perturbation at target-query position ℓ to measured token i, so that $u_{i,s}^{(1)} = \sum_\ell a_{\ell s} L_{i\ell} W_O(\bar{o}_{\ell,-s} - v_s)$. Then

$$B_s^{\text{cone}} = \frac{1}{n}\sum_{i,\ell} a_{\ell s}\left(\langle \mu(H), G_{i\ell} v_s\rangle - \langle \mu(H), G_{i\ell}\bar{o}_{\ell,-s}\rangle\right), \qquad G_{i\ell} = \frac{\Pi_i L_{i\ell} W_O}{\|h_i\|_2}. \tag{22}$$

This formula is a local response identity, not an empirical guarantee of sign accuracy or a sink-specific low-rank theorem. For a fixed target recipient s and target layer ℓ_t, define

$$\ell^\star(s, \ell_t) = \min\{\ell_m : \text{sign}(B_s^{\text{cone}}(\ell_t, \ell_m)) \neq \text{sign}(B_s^{\text{cone}}(\ell_t, \ell_m - 1))\}, \tag{23}$$

when such an adjacent sign change exists. This defines a potential test of crossing locations, not a reported validation: the current experiments do not provide held-out crossing-location errors. The coordinate formulas are written for one head; for multiple heads, concatenate head updates before applying W_O. These identities do not exclude an upstream activation mechanism that forms the attention route.

C Supplementary Measurements

C.1 LM Uncertainty, Mask Strength, and Spectral Readouts

The LM uncertainty audit mirrors the SD3 bootstrap analysis. Across the 18 first-token confirmatory rows, 17 have bootstrap 95% confidence intervals excluding zero; the only exception is the near-zero OLMo $L1 \to L2$ row. The signs are deliberately mixed, with 10 positive and eight negative first-token routes. The headline positive routes are tight: GPT-2 $L7 \to L9$ is $0.031615\,[0.031202, 0.032022]$, Qwen3-0.6B $L25 \to L25$ is $0.034873\,[0.034276, 0.035464]$, and the four Qwen3-14B routes range from 0.039213 to 0.052856 with all confidence intervals excluding zero. Mistral also shows a signed early/late split: $L1 \to L2$ is positive at $0.028795\,[0.027562, 0.029979]$, whereas $L4 \to L4$ is negative at $-0.018341\,[-0.022756, -0.014551]$. Thus the signed landscape is not an artifact of point estimates.

We also sweep the attention-suppression mask strength $M \in \{10, 100, 10^3, 10^4\}$ on the confirmatory matrix. The mean absolute first-token effect is already 0.031615 at $M = 10$ and 0.031653 for $M \geq 100$, while matched position/content controls remain near 2.25×10^{-4}. First-token signs are preserved for all 18 sites across the sweep, and the historical raw-alignment accuracy remains unchanged (14/18 first-token rows; 50/62 joined rows). The sweep therefore rules out a dependence on the extreme 10^4 mask value, but it does not by itself show a distinct small-M linear-response regime. Finally, when the selected recipient is excluded from the cosine measurement itself, signs are again preserved for all 18 first-token sites and the median $|\Delta\mathcal{C}_{\text{exclude}}|/|\Delta\mathcal{C}|$ is 1.005. The cone movement is therefore not driven by including the ablated recipient vector in the measured token set.

The Qwen3-14B spectral audit sharpens the comparison to compression-valley analyses. Middle and late routing sites also increase spectral dispersion and reduce leading-direction dominance: for $L32 \to L34$, effective rank increases by 0.391 and maximum explained variance decreases by

Table 9: Theory diagnostic from existing runs. Correlations are computed over best condition/layer rows except for the 16-LM breadth matrix, where entries are medians over model/dataset runs. Higher coherence supports the shared-update account; more negative alignment indicates suppression deltas moving away from the baseline cone axis. Mass is the joined selected-recipient attention mass.

Setting	Rows	Corr(effect, coh.)	Corr($\|$effect$\|$, $\|$coh$\|$)	Corr(effect, align.)	Corr(effect, mass)
SD3 carrier split	16	0.794	0.864	-0.765	0.514
Qwen3-14B layer-local	136	0.798	0.799	-0.738	0.272
Mistral-7B layer-local	16	0.638	0.642	-0.722	0.349
16-LM breadth median	32	–	0.649	–	0.330

Table 10: Historical raw-delta alignment scores. These are not results for the normalized derivative in Eq. 4; site dependence and implementation limitations preclude treating the rows as independent trials.

Subset	Rows	Sign acc.	Corr(score, effect)
All joined rows	62	0.806	0.299
First-token rows	18	0.778	0.380
Position/content controls	44	0.818	0.368
$\|\text{effect}\| \geq 10^{-3}$	18	0.833	–
Qwen3-14B first-token	4	1.000	–
GPT-2 + Mistral-7B first-token	5	1.000	–

0.00818; for $L39 \rightarrow L39$, effective rank increases by 1.905 and maximum explained variance decreases by 0.00866. Across the 160-row Qwen3-14B layer-local spectral sweep, cone effects correlate positively with spectral entropy change ($r = 0.410$) and with decreased leading-direction dominance ($r = 0.487$), but the relation is not an identity: without thresholding, only 49/160 rows have matching cone/spectral signs, with many near-zero cone rows. The early $L8 \rightarrow L10$ site is informative in the other direction: it opens the cosine cone without a matching spectral decompression signature, and two such cone-opening/spectral-nondecompression rows appear in the sweep. We therefore treat CV-style spectral compression and our cosine cone statistic as related but separable readouts of routing-induced geometry.

C.2 Post-Hoc Update Geometry

Correlations below compare outcomes with summaries of the same intervention updates. They are descriptive and are not independent predictive validation.

C.3 Historical Alignment Predictor

Tables 10 and 11 retain the original raw-delta alignment results for traceability. That implementation omits token-normalization derivatives and needs corrected output projection and degenerate-row handling. These numbers do not validate Eq. 4. Multiple sites share models and texts; an independent-binomial interpretation is not used. Neither prediction cost nor held-out editing success was measured.

C.4 Update Singular Value Decomposition

Singular value decomposition (SVD) resolves the propagated update into orthogonal directions ranked by their squared singular values. The leading update direction explains 0.475 of energy on average and the first two explain 0.536. This does not show that the remaining directions are negligible or that low rank explains predictor failure.

C.5 Historical Value-Patch Measurements

These measurements precede causal-donor and target-coordinate corrections. Their ratios are not interpreted as identifying the minimal component.

Table 11: Simple sign-predictor baselines. Baselines are leave-one-out within each subset. Mass uses only selected attention mass; replacement magnitude uses only $\|\bar{o}_{-s} - o\|$ and is therefore a value-replacement magnitude proxy, not an activation-norm control.

Subset	Value-repl.	Majority	Position	Mass	Repl. magnitude
All rows	0.806	0.532	0.565	0.629	0.597
First-token	0.778	0.556	0.556	0.611	0.778
Controls	0.818	0.523	0.568	0.636	0.659
$\|\Delta\mathcal{C}\| \geq 10^{-3}$	0.833	0.000	0.000	0.667	0.889
First-token, $\|\Delta\mathcal{C}\| \geq 10^{-3}$	0.800	0.533	0.533	0.667	0.867

Table 12: SVD diagnostic of propagated update fields on confirmatory first-token sites. Top-1/top-2 are fractions of update energy explained by the first one or two singular directions of $U = H^{\text{int}} - H^{\text{base}}$. Failures are sign-predictor failures from Table 10.

Subset	Rows	Mean $\|\Delta\mathcal{C}\|$	Top-1 energy	Top-2 energy
All first-token sites	18	0.0317	0.475	0.536
Predictor successes	14	0.0213	0.506	0.574
Predictor failures	4	0.0680	0.365	0.403
Strong successes	12	0.0248	0.507	0.567
Strong failures	3	0.0907	0.475	0.522

C.6 Immediate Output Response

We measure next-token behavior under the same confirmatory layer-local interventions: mean token NLL, perplexity, logit KL from the baseline distribution, top-1 prediction changes, and mean absolute logit shift. First-token routes produce substantially larger functional changes than matched position/content controls (Table 14). Across the 18 first-token sites, the mean logit KL is 0.0935 and the mean top-1 change fraction is 0.096, compared with 0.00423 and 0.0162 for all controls. The median gap is also large: first-token median KL is 0.0380, while control median KL is 0.00082. Geometric effect magnitude also correlates with output changes: over first-token sites, $|\Delta\mathcal{C}|$ correlates with logit KL at 0.718, with top-1 change at 0.798, and with $|\Delta\text{NLL}|$ at 0.752. This is not a mediation claim: cone change and output-distribution movement are measured as parallel downstream consequences of the same routing suppression, and the correlations show covariation rather than proving that the cone state lies on the causal path to logits. Output distributions change for both cone-opening and cone-narrowing interventions. Cone-opening first-token routes have mean KL 0.0576 and top-1 change 0.100, cone-narrowing routes have mean KL 0.173 and top-1 change 0.127, while near-zero cone routes have mean KL only 0.00313 and top-1 change 0.0140. These output-distribution changes do not establish improved task performance or generation quality.

D Extended Related Work

Attention sinks, activations, and registers. Streaming LMs use initial attention sinks to preserve long-context generation (Xiao et al., 2024). Related studies characterize sink emergence, first-token bias, and resistance to over-mixing (Gu et al., 2025; Barbero et al., 2025), while massive-activation and register studies examine unusual tokens as computational resources (Sun et al., 2024; Darcet et al., 2024; Sun et al., 2026). Compression Valleys connects sinks, massive activations, and representation compression (Queipo-de Llano et al., 2025); other accounts study geometric reference frames, spectral subspaces, and architectural asymmetries (Ruscio et al., 2026; Cancedda, 2024; Li et al., 2026). Earlier masked-language-model studies provide complementary evidence about outlier dimensions. Disabling particular LayerNorm-associated outliers can strongly disrupt model behavior (Kovaleva et al., 2021); other studies connect persistent outliers to positional embeddings (Luo et al., 2021) and to token frequency and attention toward special tokens (Puccetti et al., 2022). These

Table 13: Value-direction versus routing decoupling.

Intervention	All median ratio	All sign agree	Positive median ratio	Negative median ratio
Remove selected value, no renorm	0.366	0.611	0.407	0.391
Remove cone-axis projection	0.151	0.556	0.195	0.154
Replace with matched non-structural value	0.765	0.778	0.363	2.464

Table 14: Functional audit under the same LM routing interventions. NLL is next-token negative log-likelihood; KL is from the baseline next-token distribution to the intervened distribution. First-token rows are the thesis-relevant interventions; controls aggregate second-token, early-token-control, and content rows.

Subset	Rows	Mean $\lvert\Delta\text{NLL}\rvert$	Mean KL	Mean top-1 change
First-token routes	18	0.0758	0.0935	0.0963
Position/content controls	54	0.00212	0.00423	0.0162
Cone-opening first-token	8	0.0412	0.0576	0.100
Cone-narrowing first-token	7	0.148	0.173	0.127
Near-zero first-token	3	7.82×10^{-4}	0.00313	0.0140
Qwen3-14B first-token	4	0.0340	0.0336	0.0698
GPT-2 first-token	1	0.0365	0.0836	0.1710
Mistral-7B first-token	4	0.0224	0.0469	0.0789

results concern specific encoders and intervention settings, rather than a universal identification of outlier dimensions with sink tokens. Our question concerns the downstream effect of suppressing routes into already selected recipients. The signed response is compatible with an upstream activation mechanism that forms those routes. SD3 additionally allows a within-model comparison between multiple structural carrier families, whereas the LM first-token condition holds the structural position fixed while examining depth and model dependence.

Representation geometry. Directional concentration has been studied in learned word embeddings (Gao et al., 2019) and contextual hidden representations (Ethayarajh, 2019). Analyses of pure-attention networks identify a tendency toward rank collapse when residual connections and feed-forward blocks are absent (Dong et al., 2021); this architectural limit does not specify the response of a complete pretrained transformer to a local edit. Signal-propagation theory also relates rank collapse to gradient behavior and studies depth-dependent residual scaling (Noci et al., 2022). Our finite-update diagnostics instead concern downstream changes in fixed, pretrained networks. Dispersion regularization can modify concentrated embeddings and improve generalization in the studied training settings (Liu et al., 2026), whereas rogue-dimension analyses show why cosine-based similarity need not track behavioral importance (Timkey & Van Schijndel, 2021). Geometric measurements and behavioral effects address different questions. Our cosine statistic depends on the normalized-token centroid relative to the origin; centered spectral statistics measure a different aspect of the representation. We report both and retain cases in which their responses disagree. The present study characterizes an intervention response of cosine concentration without equating it with information loss or explaining the benefits of training-time dispersion. It also does not infer a common internal feature space merely from observing responses across model families.

Attention attribution and intervention. The relationship between attention and explanation has been debated through counterfactual attention distributions and tests of their connection to model predictions (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). These studies reach different conclusions under different definitions and testing assumptions. In transformers, attention identifiability (Brunner et al., 2019) and attention rollout or flow (Abnar & Zuidema, 2020) further distinguish a local weight from its relation to the network's input and output. Our experiment asks a narrower question: how a specified routing edit changes a geometric readout, without treating the baseline attention map as an explanation of task behavior.

Attention interventions and routing. Prompt-to-Prompt uses cross-attention control for image editing (Hertz et al., 2022), and DiT attention guidance distinguishes Key-space control from Value-space aggregation (Li, 2026). Differential Transformer modifies the attention operator to reduce irrelevant-context allocation (Ye et al., 2025); attention-prior and sink-aware training work studies alternative routing structures (Litman & Guo, 2026; Fu et al., 2026). Our experiments keep the pretrained model weights fixed and use selective recipient suppression to measure representation responses. Norm-based attention analysis explicitly incorporates transformed value vectors (Kobayashi et al., 2020), and value-aware token importance methods further examine contributions beyond attention scores (Guo et al., 2024; Cunningham & Cirone, 2026). The replacement identity makes a complementary distinction: the increment depends on the selected value relative to its replacement mixture, followed by output projection and downstream propagation. This is an analysis of a specified edit, rather than a token-ranking rule, new attention architecture, or demonstrated image-editing policy.

E Supplementary Protocols and Diagnostic Detail

E.1 Spectral Readout Definitions

A common translation can change this statistic while preserving pairwise Euclidean distances and the centered spectrum. To distinguish directional concentration from spectral compression, we also compute spectral entropy, effective rank, participation ratio, and maximum explained variance on centered hidden states. With normalized singular-value powers $\lambda_r = \sigma_r^2 / \sum_j \sigma_j^2$, these are respectively $-\sum_r \lambda_r \log \lambda_r$, $\exp(-\sum_r \lambda_r \log \lambda_r)$, $1/\sum_r \lambda_r^2$, and $\max_r \lambda_r$. Their agreement with the cosine readout is an empirical question examined in Section 4.4.

E.2 Protocol Overview

Table 15 summarizes the three protocols introduced in Section 3.3. The carrier comparison, single-layer propagation sweep, and simultaneous multilayer edits isolate different aspects of the routing response and should be interpreted within their respective settings.

Table 15: The three intervention protocols answer different questions. Target layers are edited layers; measured layers are readout locations.

Protocol	Intervention	Readout	Question
SD3 carrier split	Structural text-recipient families	Image tokens at layer 12, step 0	Does response sign differ across carriers?
LM layer-local	One target layer at a time	Same and downstream layers	How does the response vary with depth?
LM breadth	Simultaneous edits across target layers	Fixed rule $\mathrm{round}(0.75(L-1))$	How do responses compare across models and datasets?

E.3 Complete SD3 Carrier Comparison

Table 16 expands the comparison in Section 4.1 to all eight recipient conditions. The CLIP-specific conditions retain the opposite signs for special and padding recipients, while content and image controls produce smaller effects. These comparisons use the specified interventions without matching attention mass or update energy.

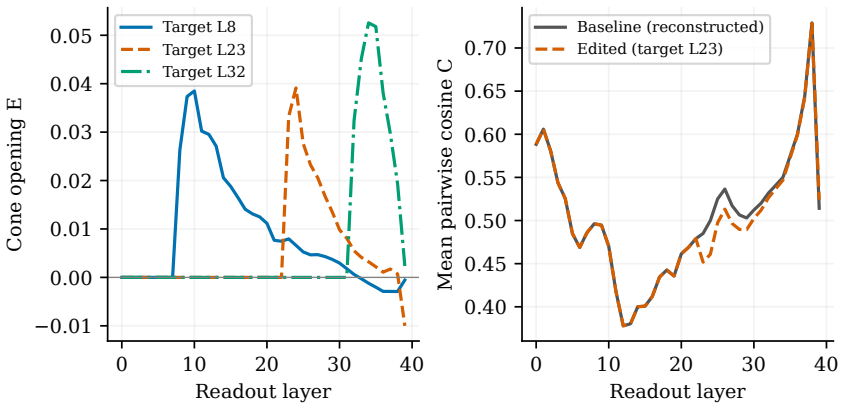

Figure 5: Depth propagation and absolute cosine context in Qwen3-14B. Left: three illustrative target layers from the full layer-local sweep. Right: reconstructed baseline and edited means for target L23. Curves are aggregate means; no article-level uncertainty is inferred from these merged summaries.

Table 16: SD3 Gate-1 carrier split at layer 12, step 0. Positive values denote off-diagonal anisotropy reduction, i.e. cone opening. Brackets report bootstrap 95% confidence intervals over prompt/seed rows. Padding-like carriers push the opposite signed direction.

Condition	Reduction [95% CI]	Pos. frac.	Interpretation
all_sinks	0.001466 [0.001431, 0.001500]	0.966	combined suppression opens cone
clip_structural	0.001253 [0.001219, 0.001287]	0.936	CLIP structural carrier
special	0.001973 [0.001950, 0.001996]	1.000	special/EOT-like carrier
clip_special	0.001933 [0.001911, 0.001956]	1.000	CLIP special carrier
padding	−0.000660 [−0.000678, −0.000643]	0.056	opposite signed carrier
clip_padding	−0.000718 [−0.000737, −0.000700]	0.056	opposite signed carrier
content	0.000118 [0.000113, 0.000124]	0.624	weak content control
image	−0.0000004 [−0.0000006, −0.0000002]	0.012	image-side floor

E.4 Baseline Context and Depth Reanalysis

To distinguish a local response from its later propagation, we reanalyze the full Qwen3-14B layer-local aggregate table: 34 target layers, 40 readout layers, and four recipient conditions. For each row, we recover the baseline mean as $C_0 = C_{\text{edited}} + E$, where E is the reported paired mean opening. Figure 5 shows both the signed response and its absolute cosine context. This avoids interpreting the same absolute change as an identical geometric state at every depth. The reconstruction supplies baseline means, not their variances or article-level distributions.

Propagation can reverse the direction observed at the edited layer. Defining a non-negligible aggregate response by $|E| > 10^{-4}$, 24 of the 34 target layers have at least one later readout with a sign opposite to their same-layer response. The count is 24 at 10^{-5} and 22 at 10^{-3}. These are exploratory threshold-sensitivity counts, not independent hypothesis tests or population frequencies. Upstream mean effects remain below 4.95×10^{-7} in absolute value. The source table combines runs with differing sample counts; reconstructed baselines vary by up to 0.00207 at a fixed readout. We therefore retain each row's own baseline and do not pool runs into a purported common article cohort. The observed reversals motivate explicitly propagating the attention increment before predicting a downstream cone response.

E.5 Follow-up Confirmation and Scalar and Vector Readouts

A subsequent two-site confirmation on 64 new shared articles preserves this distinction and rejects the stronger joint mechanism hypothesis. Although the predicted ordering of scalar cone errors reproduces at both sites, GPT-2's centered patch is worse at reconstructing the common hidden

vector. The follow-up normalized-centroid audit also finds that smaller vector error need not imply smaller scalar cone error. The distinction between scalar and vector errors follows from the centroid readout: for a patched normalized centroid $m_p = m_f + e$, the cone difference is $C_p - C_f = \frac{n}{n-1}(2\langle m_f, e\rangle + \|e\|^2)$. Vector-error magnitude alone does not determine this scalar difference. This identity explains a possible mismatch between evaluation criteria; it is not an independent prediction or a localization of its upstream cause.