FLIP: Final Layer Inference-Time Probing for Vision–Language Models

Drandreb Earl O. Juanico [1] **Rowel O. Atienza** [1 2]

Abstract

We present FLIP, a final-layer inference-time probe for testing whether a logit-facing intervention site in an open-weight vision–language model (VLM) supports structured, task-linked computation rather than generic perturbation. Behavioral change under internal intervention is otherwise mechanistically ambiguous: it may reflect improved use of visual evidence, generic output instability, or outright degradation. FLIP applies elementwise flooring to the final normalized hidden state before logit computation, leaving parameters, prompts, and decoding unchanged. On a controlled detection/counting probe, sweeping intervention strength reveals three regions: negligible change, a bounded interior regime in which detection recall at IoU 0.50 (R_{50}) improves while tolerant counting error ($\mathcal{E}_{\text{count}}$) falls, and over-suppression. We formalize a four-criterion probe-and-sweep protocol for disciplining the interpretation of intervention effects: regime structure, grounding-proxy alignment, feature-coherence dependence, and failure to reproduce the same positive regime on a performance-based negative control. The post-normalization state passed to the output head is the logit-facing instantiation of this test; under a non-targeted flooring sweep it satisfies the full protocol. Raw decoder-layer interventions—including the last-block output before final normalization—and the singleton-pair left/right control fail to reproduce the Final-site signature, while same-site operators and multiple VLMs replicate it. FLIP is therefore a validation step for intervention-based mechanistic interpretability, not a steering method.

[1]AI Graduate Program, University of the Philippines, Diliman, Quezon City, Philippines 1101 [2]EEE Institute, College of Engineering, University of the Philippines, Diliman, Quezon City, Philippines 1101. Correspondence to: Drandreb Earl O. Juanico <dojuanico@alum.up.edu.ph>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Open-weight vision-language models expose final-layer hidden states at inference time, but behavioral change under intervention is mechanistically ambiguous: it may reflect improved use of visual evidence, generic instability, or outright degradation. Single-point interventions cannot resolve this ambiguity. We therefore evaluate intervention sites under sweeps over intervention strength ϑ and ask whether they pass a four-criterion probe-and-sweep protocol for structured, task-linked computation rather than generic perturbation; a fuller definition of the latter is given in Appendix A.

FLIP (Final Layer Inference-Time Probing) applies elementwise flooring to the post-normalization Final-site hidden state immediately before logit computation (Section 3). We use the standard linearity of the output head as a baseline for interpreting this site: a hidden-state displacement can be useful only insofar as it projects onto task-relevant output directions. The empirical question is therefore not whether FLIP can perturb outputs, but whether a non-targeted flooring sweep reveals a bounded, task-linked response. On a controlled detection/counting probe, we observe three response regions: negligible change, an interior regime with improved R_{50} and reduced $\mathcal{E}_{\text{count}}$, and over-suppression.

The aim is not to identify a circuit, but to validate whether an intervention effect is structured enough to justify circuit-level follow-up. FLIP therefore treats behavioral change as credible only if it survives independent constraints: regime structure, grounding-proxy alignment, feature-coherence dependence, and task-specific negative control. Raw decoder-layer checks (Figure 4; Appendix G, Figure 8) provide site-specific contrasts under the implemented sweep, while cross-architecture and same-site operator replications (Figure 6; Appendices F, M.1) test the protocol-level signature beyond one model or operator. The delta-method product summary (Table 1) tests compatibility with a shared detection-to-counting grounding factor.

Contributions. (1) We propose a probe-and-sweep protocol that treats intervention effects as objects requiring validation, not as self-interpreting performance changes. (2) We instantiate it with FLIP, a minimal Final-site flooring probe grounded in the standard output-head baseline. (3) We show that primary-task gains are insufficient: accepted

effects must also satisfy grounding-proxy alignment, feature-coherence dependence, negative-control contrast, and logit-facing site specificity. (4) We replicate this protocol-level signature across same-site magnitude-bounding operators and multiple vision-language models.

2. Background and Related Work

Intervention-based interpretability trades off causal specificity and population coverage. Causal abstraction uses interchange interventions to test aligned causal variables (Geiger et al., 2021); causal mediation asks which internal components are implicated in behavior (Vig et al., 2020); activation patching and automated circuit discovery localize task-relevant components after a metric and dataset have been chosen (Zhang and Nanda, 2024; Conmy et al., 2023); and circuit-tracing methods aim to recover prompt-level computational graphs (Ameisen et al., 2025). These methods provide stronger mechanistic follow-up, but their conclusions can depend on the chosen behavior, metric, examples, and ablation design (Miller et al., 2024). Recent VLM work similarly analyzes visual-token processing, learns interpretable feature bases, or intervenes through latent and hidden-state steering to improve grounding or reduce hallucination (Neo et al., 2024; Pach et al., 2025; Yang et al., 2026; Liu et al., 2025; Su et al., 2025; Juanico et al., 2025; Li et al., 2025). FLIP targets the preceding selection problem: before assigning an effect to a circuit or using it for steering, it asks whether a candidate site exhibits a population-level, task-linked, control-sensitive signature that warrants mechanistic follow-up.

3. FLIP: Final Layer Inference-Time Probing

FLIP is an inference-time intervention applied to the post-normalization final hidden-state tensor. Figure 1 shows the operator: components below ϑ are clamped elementwise before logit computation, while all others are left unchanged.

3.1. Operational Definition

Let $\mathbf{z}^{(t)} \in \mathbb{R}^{n_t \times d}$ denote the post-normalization Final-site tensor passed to the output head during generation call t, with n_t equal to the prompt/image prefill length for the first next-token logits and $n_t = 1$ for cached decoding steps. For strength $\vartheta \in \mathbb{R}$, FLIP outputs

$$\widetilde{\mathbf{z}}^{(t)} = \max(\mathbf{z}^{(t)}, \vartheta), \tag{1}$$

element-wise, the replacement is applied at every next-token logit computation during generation. Thus, FLIP covers the prefill logits used for the first response token and each subsequent autoregressive decoding step, while leaving weights, prompts, tokenization, decoding rules, and KV caches unchanged.

FLIP (definition used in all experiments). Given the post-normalization Final-site hidden state $\mathbf{z}$ and strength ϑ, FLIP outputs $\widetilde{\mathbf{z}} = \max(\mathbf{z}, \vartheta)$ element-wise before logit computation. No model weights, prompts, or decoding rules are changed.

In the main experiments, FLIP targets the post-normalization final hidden state passed to the output head. We distinguish this *Final* site from raw decoder-layer outputs, including the pre-normalization last-block output. FLIP is a probe, not an optimization method.

4. Tasks and Vision-Centric Evaluation Setting

Unless otherwise stated, main-text experiments use Qwen3-VL-4B (Bai et al., 2025); additional model results are reported in the appendix.

4.1. Intermediate Visual Signals

Prompted localization. We query the model with `Detect <object>. Provide correct bounding boxes.` and parse outputs into a common pixel-space representation; model-specific details are given in Appendix B.

Dataset and supervision. Localization prompts are instantiated on a clustered subset of MS COCO val2017 (Lin et al., 2014) comprising 7,056 image–prompt pairs, constructed image-centrically with one query per present object category. Image-level clustering is used because prompts drawn from the same image are not independent; annotations are used only to instantiate `<object>` and score outputs.

4.2. Reasoning-Based Counting

Counting prompt. `How many <object>? Answer with integer only.` isolates errors attributable to visual instance discrimination rather than formatting.

4.3. Vision-Centric Benchmarks

We additionally evaluate MMStar, MindCube, and NaturalBench via LMMS-Eval as intervention probes rather than optimization targets.[1]

5. Theory

This section gives the simple operator and output-head facts used to motivate the probe. These facts are not themselves

[1] Third-party asset credits, licenses, and terms of use for Qwen3-VL, Kimi-VL-A3B, MS COCO val2017, and LMMS-Eval are documented in the supplemental repository.

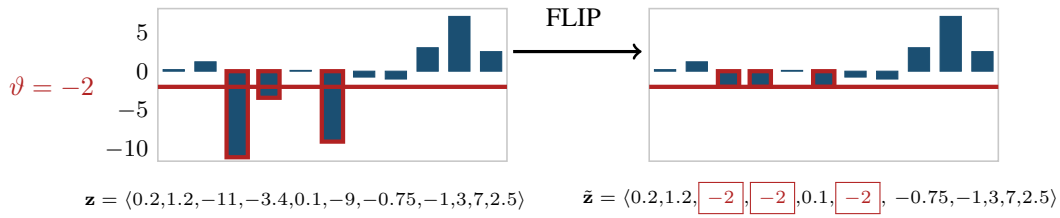

Figure 1. **Hidden-state flooring view of FLIP.** Components below ϑ are clamped elementwise at the post-normalization Final site before logit computation (here, indices 3, 4, and 6). FLIP uses $\widetilde{\mathbf{z}} = \max(\mathbf{z}, \vartheta)$ as the intervention; attention and rollout maps are used only as downstream diagnostics.

the main novelty: they provide the baseline against which the empirical protocol is interpreted. Since FLIP is not trained, task-conditioned, or optimized, any useful effect must come from alignment between the induced displacement and task-relevant directions already present in the logit-facing representation.

5.1. Componentwise Flooring and Perturbation Growth

For $x \in \mathbb{R}$ and threshold $\vartheta \in \mathbb{R}$, define

$$\phi_\vartheta(x) = \max(x, \vartheta), \qquad \delta_\vartheta(x) = (\vartheta - x)_+. \tag{2}$$

Applied elementwise to $\mathbf{z} \in \mathbb{R}^{n \times d}$,

$$\widetilde{\mathbf{z}} = \phi_\vartheta(\mathbf{z}), \qquad \mathbf{\Delta}_\vartheta = \widetilde{\mathbf{z}} - \mathbf{z}. \tag{3}$$

Define the active set

$$\mathcal{S}_\vartheta = \{(i,k) : z_{ik} < \vartheta\}, \qquad A(\vartheta) = \frac{|\mathcal{S}_\vartheta|}{nd}, \tag{4}$$

and mean clamp magnitude

$$M(\vartheta) = \frac{1}{nd} \sum_{i,k} (\vartheta - z_{ik})_+. \tag{5}$$

Proposition 5.1 (Monotone clamping)**.** *For every* (i,k),

$$\widetilde{z}_{ik} = z_{ik} + (\vartheta - z_{ik})_+, \tag{6}$$

and $\|\mathbf{\Delta}_\vartheta\|_p$ *is non-decreasing in* ϑ *for all* $p \in [1, \infty]$.

As ϑ increases, FLIP passes from near-inactivity, to a bounded intermediate region, to widespread clamping. This geometry defines the strength axis of the probe but does not by itself determine decoded task metrics.

5.2. Standard Output-Head Alignment Baseline

Because FLIP is applied immediately before the output head, its first-order interpretation follows from the standard linear logit map. For a final-token hidden state $\mathbf{z}_t$ with output matrix W,

$$\widetilde{\ell}_t - \ell_t = W^\top \Delta_{\vartheta,t}, \qquad \Delta_{\vartheta,t} = (\vartheta - \mathbf{z}_t)_+. \tag{7}$$

For a task-relevant token set Y^+ and contrast set Y^-, define the log-sum-exp margin

$$m(\mathbf{z}_t) = \log \sum_{y \in Y^+} \exp(w_y^\top \mathbf{z}_t) - \log \sum_{y \in Y^-} \exp(w_y^\top \mathbf{z}_t). \tag{8}$$

A first-order expansion gives

$$m(\widetilde{\mathbf{z}}_t) - m(\mathbf{z}_t) \approx \langle \bar{w}_{Y^+}(\mathbf{z}_t) - \bar{w}_{Y^-}(\mathbf{z}_t), \Delta_{\vartheta,t} \rangle, \tag{9}$$

where $\bar{w}_{Y^+}$ and $\bar{w}_{Y^-}$ are softmax-weighted output-head directions over the two token sets.

This calculation is not a new representation theory; it defines the baseline for interpreting FLIP. Since the clamp is object-agnostic, any useful change must come from projection onto task-relevant output directions. The protocol then tests whether that necessary condition is accompanied by grounding-proxy alignment, feature-coherence dependence, logit-facing site specificity, and negative-control contrast.

6. Grounding-Proxy Association Analysis

This section tests whether FLIP effects are compatible with a grounding-linked pathway, without claiming full causal identification.

Let $\mathbb{1}_{\text{FLIP}} \in \{0, 1\}$ denote intervention. Detection quality is

$$R_{50} = \frac{\text{TP}_{50}}{\text{gt_count}}, \qquad \Delta R_{50} = R_{50} - \mathbb{E}[R_{50} \mid \mathbb{1}_{\text{FLIP}} = 0]. \tag{10}$$

For counting, let y be the parsed model count and $y^\star = \text{gt_count}$ the ground-truth count; the tolerant error is

$$\mathcal{E}_{\text{count}} = \max(0, |y - y^\star| - 1). \tag{11}$$

Localization and counting are elicited by separate prompts and scored from independent outputs, so ΔR_{50} is not a literal within-pass mediator of $\mathcal{E}_{\text{count}}$. We instead use it as an observable proxy for an unobserved intervention-sensitive grounding state:

$$\mathbb{1}_{\text{FLIP}} \to G_\vartheta, \qquad G_\vartheta \rightsquigarrow \{\Delta R_{50}, \mathcal{E}_{\text{count}}\}.$$

We estimate the product summary from

$$\Delta R_{50} \sim \mathbb{1}_{\mathrm{FLIP}}, \qquad \mathcal{E}_{\mathrm{count}} \sim \mathbb{1}_{\mathrm{FLIP}} + \Delta R_{50}.$$

The first model is a Gaussian GLM with image-clustered SEs and supplies the a-term; the second is a Poisson GLM on strict counting responses and supplies the b-term. Confidence intervals are large-sample clustered intervals over image-level units, not intervals requiring normally distributed individual errors.

7. Experiments

FLIP is applied at inference time by replacing the post-normalization Final-site tensor $\mathbf{z}$ with $\widetilde{\mathbf{z}} = \max(\mathbf{z}, \vartheta)$ prior to logit computation. Implementation details are provided in Appendix E; code, prompts, sweep grids, run manifests, and analysis scripts are available in the anonymized repository at https://anonymous.4open.science/r/flip-qwen3vl-EBAC/.

7.1. Experimental Setup

Experiments use 7,056 clustered COCO image–query pairs, with localization and counting queries issued per sample under matched intervention strengths. Because multiple object queries may be drawn from the same image, statistical uncertainty is computed at the image-cluster level rather than by treating image–query pairs as independent. Reported 95% confidence intervals therefore rely on the large-sample approximation for the image-clustered estimator, not on normality of individual errors or residuals. We use Qwen3-VL-4B-Instruct as the anchor model; cross-architecture results are in Appendix F. Unless otherwise noted, the core intervention sweeps use a fixed, pre-specified ϑ grid and fixed random seeds for replicability; repeated-run dispersion is reported only where runtime-level nondeterminism cannot be controlled in the serving stack. We treat pointwise tests across ϑ as descriptive screening rather than as a family of confirmatory discoveries: no claim depends on declaring any single scanned ϑ significant. Regime identification is based on the full sweep and the FWHM window around the primary ΔR_{50} peak, then checked against grounding-proxy alignment, feature-coherence controls, negative controls, raw-layer site specificity, and image-cluster subsampling. Hardware details for the core sweeps are reported in Appendix E.

Evaluation Protocol (Behavioral Readout). All interventions are evaluated through final model outputs. For raw decoder-layer checks, we apply $h^{(\ell)} \to \tilde{h}^{(\ell)}(\vartheta)$ at the selected layer and run the remaining computation to logits and decoded predictions. For the *Final* site, we intervene on the post-normalization state passed to the output head. Thus, all layer checks measure downstream behavior, not intermediate-state change alone.

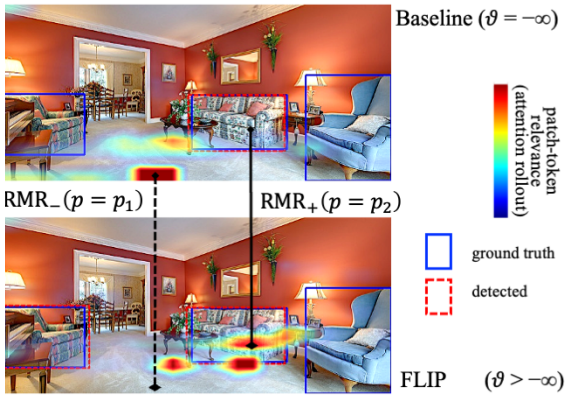

Figure 2. **Attribution redistribution under FLIP for a single image–prompt pair.** Top: baseline target-conditioned patch relevance. Bottom: relevance after intervention. The figure qualitatively illustrates the bounded redistribution associated with the interior regime; the saliency construction is described in Appendix H.1, and the aggregate RMR diagnostics are reported in Appendix H. Representative example only.

7.2. Metrics

Primary quantities. We use R_{50} / ΔR_{50} as the detection-side mediator and $\mathcal{E}_{\mathrm{count}}$ as the downstream outcome.

Attribution redistribution (RMR). Relevance mass reallocation (RMR) is an appendix-side redistribution diagnostic rather than a causal measure. We track RMR_+, RMR_-, and $\mathrm{RMR}_{\mathrm{obj}}$; construction and interpretation are given in Appendices H.1 and H.3.

7.3. Results

Primary regime evidence. Figure 3 shows a three-regime response in ΔR_{50}: negligible change, a bounded interior gain, and collapse. The interior regime is not declared from the single largest point in the sweep, and we do not interpret unadjusted pointwise p-values over ϑ as separate discoveries. Instead, the peak defines a descriptive FWHM window that must survive the full protocol: grounding-proxy alignment, feature-coherence disruption, negative-control contrast, logit-facing site specificity, and subsample stability. Appendix N (Figure 13) shows that the same interior window remains stable under image-cluster subsampling.

Grounding-proxy alignment. The interior regime aligns with downstream behavior: intermediate flooring increases ΔR_{50} while reducing $\mathcal{E}_{\mathrm{count}}$, and over-suppression reverses the effect. Table 1 shows a non-zero product summary in the same ϑ window where recall peaks. Because localization and counting are separate forward passes, this supports a shared grounding-proxy interpretation, not a claim that one

Table 1. **Delta-method grounding-proxy product summary on Qwen3-VL-4B.** a: effect of FLIP on ΔR_{50}; b: Poisson-GLM association of ΔR_{50} with $\log \mathbb{E}[\mathcal{E}_{\text{count}}]$ on strict counting responses; $a \times b$: product summary; and $\text{IRR}_{\text{indirect}} = \exp(a \times b)$. Values below 1 indicate that improved recall is associated with reduced tolerant counting error. The quantity is interpreted as grounding-proxy compatibility, not direct within-pass causal mediation.

ϑ	a	SE(a)	b	SE(b)	$a \times b$	SE($a \times b$)	p-value	$\text{IRR}_{\text{indirect}}$ [95% CI]
-50.0	0.00208	0.00054	-2.070	0.158	-0.00430	0.00117	0.0002	0.9957 [0.9934, 0.9980]
-5.0	0.02677	0.00214	-2.092	0.152	-0.05600	0.00606	< 0.0001	0.9455 [0.9344, 0.9568]
-4.0	0.02962	0.00219	-2.087	0.151	-0.06180	0.00638	< 0.0001	0.9401 [0.9284, 0.9519]
-2.0	0.03878	0.00242	-2.079	0.148	-0.08061	0.00765	< 0.0001	0.9226 [0.9088, 0.9365]
-1.0	0.04832	0.00277	-2.049	0.146	-0.09903	0.00905	< 0.0001	0.9057 [0.8898, 0.9219]
-0.5	0.05283	0.00294	-2.015	0.136	-0.10647	0.00932	< 0.0001	0.8990 [0.8827, 0.9156]
-0.3	0.05777	0.00299	-2.032	0.137	-0.11739	0.01001	< 0.0001	0.8892 [0.8720, 0.9069]
-0.2	0.05782	0.00304	-2.027	0.137	-0.11723	0.01002	< 0.0001	0.8894 [0.8721, 0.9070]
0.0	0.05900	0.00321	-2.074	0.133	-0.12236	0.01028	< 0.0001	0.8848 [0.8672, 0.9028]
0.5	0.02638	0.00355	-2.151	0.138	-0.05674	0.00847	< 0.0001	0.9448 [0.9293, 0.9607]
1.0	-0.531480	0.00590	-2.069	0.157	1.09974	0.08454	< 0.0001	3.0034 [2.5447, 3.5447]

output directly causes the other.

Logit-facing site specificity. Figure 4 shows that hidden-state flooring alone does not induce the Final-site signature. Raw decoder-layer interventions remain near baseline, collapse, or show only partial movement, while the post-normalization Final site exhibits the full interior regime. To test whether this contrast is a raw-threshold artifact, we also overlay percentile-matched sweeps for Layers 8, 16, and 32. These matched raw-layer sweeps do not recover the Final-site dose–response and instead attenuate or worsen the response. Thus, the raw-layer failures are not explained by applying too weak or poorly scaled a threshold. The full protocol—primary-task regime, left/right contrast, grounding-proxy alignment, and feature-coherence dependence—localizes to the post-normalization logit-facing state rather than to raw-layer depth or marginal intervention strength alone.

Specificity under controls. At the Final site, permutation controls (Figure 5) show that the positive interior regime is not preserved once feature coherence is disrupted. All ΔR_{50} and mediator quantities in this control are centered against the same unpermuted no-intervention baseline ($\lambda = 0, \vartheta = -\infty$), not against a λ-specific permuted baseline. Thus any baseline damage from permutation is retained rather than subtracted away. Relative to $\lambda = 0$, nonzero λ settings lose the positive window and collapse earlier, though their ordering is not strictly monotonic; the control is therefore interpreted as a feature-coherence stress test rather than as an isolated flooring effect on an already permuted model.

Operator comparison. Flooring, symmetric clipping, and additive shift each produce regime-shaped primary-task curves, so non-monotonicity or peak ΔR_{50} alone is not specific evidence. The full operator replication (Appendix M.1, Figure 11) therefore evaluates the joint signature: grounding-proxy alignment, feature-coherence dependence, logit-facing site specificity, and failure on the left/right performance control.

Cross-architecture replication and broader spatial control. Figure 6 shows the same qualitative three-phase structure in Qwen3-VL-8B, Qwen3-VL-4B, and Kimi-VL-A3B, with Qwen3-VL-8B the lower-amplitude case. The bottom-panel singleton-pair left/right control does not reproduce the coherent positive interior regime across models; instead, it stays near baseline or drifts downward before late collapse; task details are given in Appendix J.1. Appendix F adds the auxiliary indoor/outdoor response-switch diagnostic and mediation replications (Tables 2 and 3).

Appendix-side corroboration. Appendix H.3 shows that, for the lower-amplitude Qwen3-VL-8B case, RMR_{+} tracks the interior regime while RMR_{-} and RMR_{obj} mark the transition to over-suppression. Appendix I further shows structured sweeps on MMStar, MindCube, and NaturalBench. These analyses broaden the regime picture but remain secondary to the mediator-aligned protocol evidence.

8. Discussion

FLIP occupies the screening stage between broad behavioral probing and circuit-level identification. Primary-task improvement is not enough, and neither is proximity to the last decoder block. The relevant evidence is the joint signature: bounded grounding-proxy-aligned response, Final-site feature-coherence dependence, failure on a performance-based negative control, and specificity to the normalized logit-facing readout. The raw-layer and percentile-matched checks separate depth, dose, and validation: matching marginal intervention strength at Layers 8, 16, and 32 does not recover the Final-site response. Thus, FLIP does not claim that a circuit has been found; it nominates the Final-site interior regime as the most justified target for patching, circuit tracing, or causal-abstraction tests.

The controls define the scope of this interpretation. Table 1 supports a shared grounding-proxy account in the interior

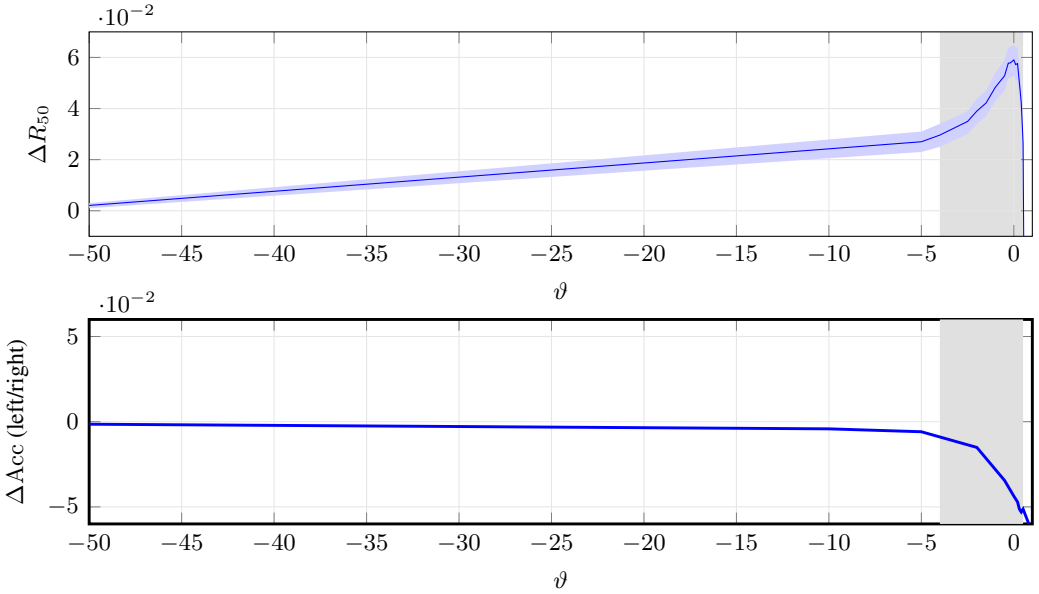

Figure 3. **Primary dose–response and negative control.** Top: FLIP on Qwen3-VL-4B-Instruct yields the expected three-region detection response in ΔR_{50}: near-baseline behavior, bounded interior gain, and collapse under strong flooring. Bottom: the singleton-pair left/right control, measured by baseline-centered ΔAcc, does not reproduce this positive interior regime; it remains near baseline over most non-collapse strengths and degrades under stronger intervention. The shaded band marks the FWHM window from the primary detection peak; bands denote 95% CI where available.

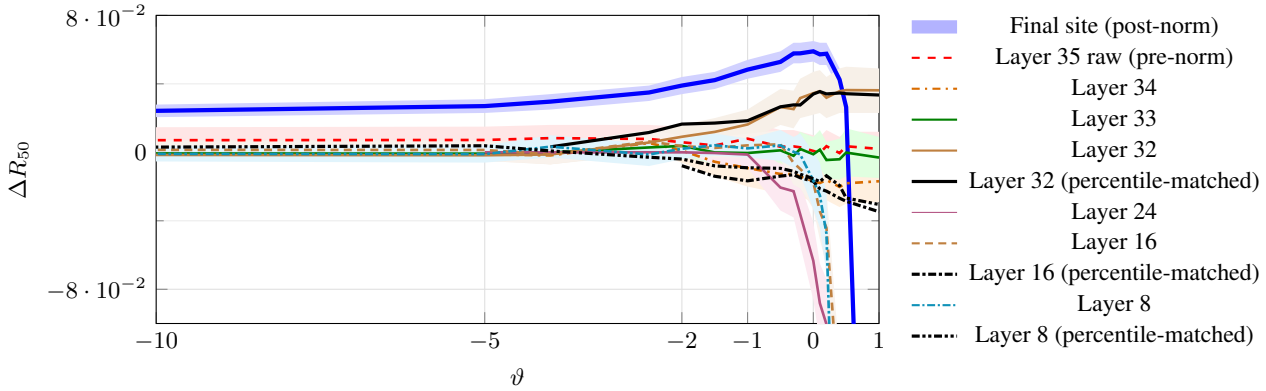

Figure 4. **Raw decoder-layer and percentile-matched interventions versus the post-normalization Final site.** FLIP is applied to selected raw decoder-layer outputs and compared with the Final site passed to the output head. The Final site exhibits the bounded interior regime, while raw-layer interventions remain near baseline, collapse, or show only partial movement. Percentile-matched sweeps for Layers 8, 16, and 32 do not recover the Final-site dose–response and instead attenuate or worsen the response, ruling out a simple raw-threshold calibration explanation. Bands denote 95% CI where available.

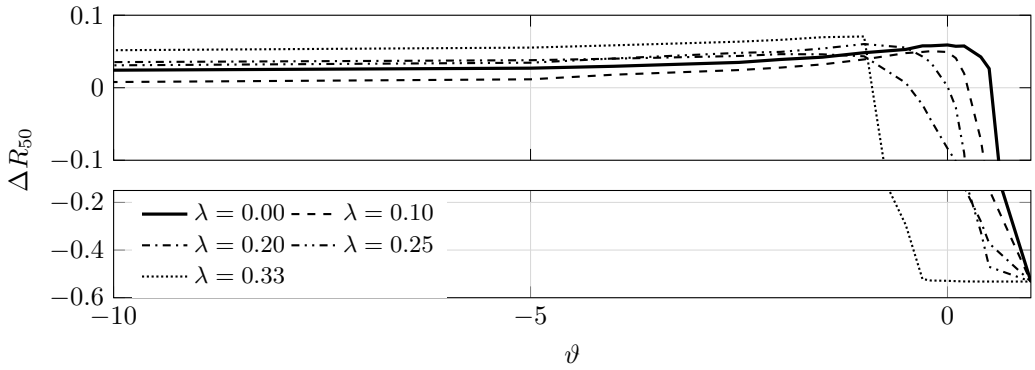

Figure 5. **Graduated permutation control via mixing parameter λ.** Both panels show ΔR_{50} under feature-coherence mixing, with all curves centered against the same unpermuted no-intervention baseline ($\lambda = 0, \vartheta = -\infty$) rather than against λ-specific permuted baselines. The upper panel zooms the near-zero/interior-response range, while the lower panel zooms the collapse regime. Relative to the unpermuted case ($\lambda = 0$), all permuted conditions lose the positive regime earlier and enter collapse sooner, although the ordering among nonzero λ values is not strictly monotonic. This indicates that feature coherence is necessary for the stability of the interior regime, and that the observed loss is not an artifact of subtracting an already degraded permuted baseline.

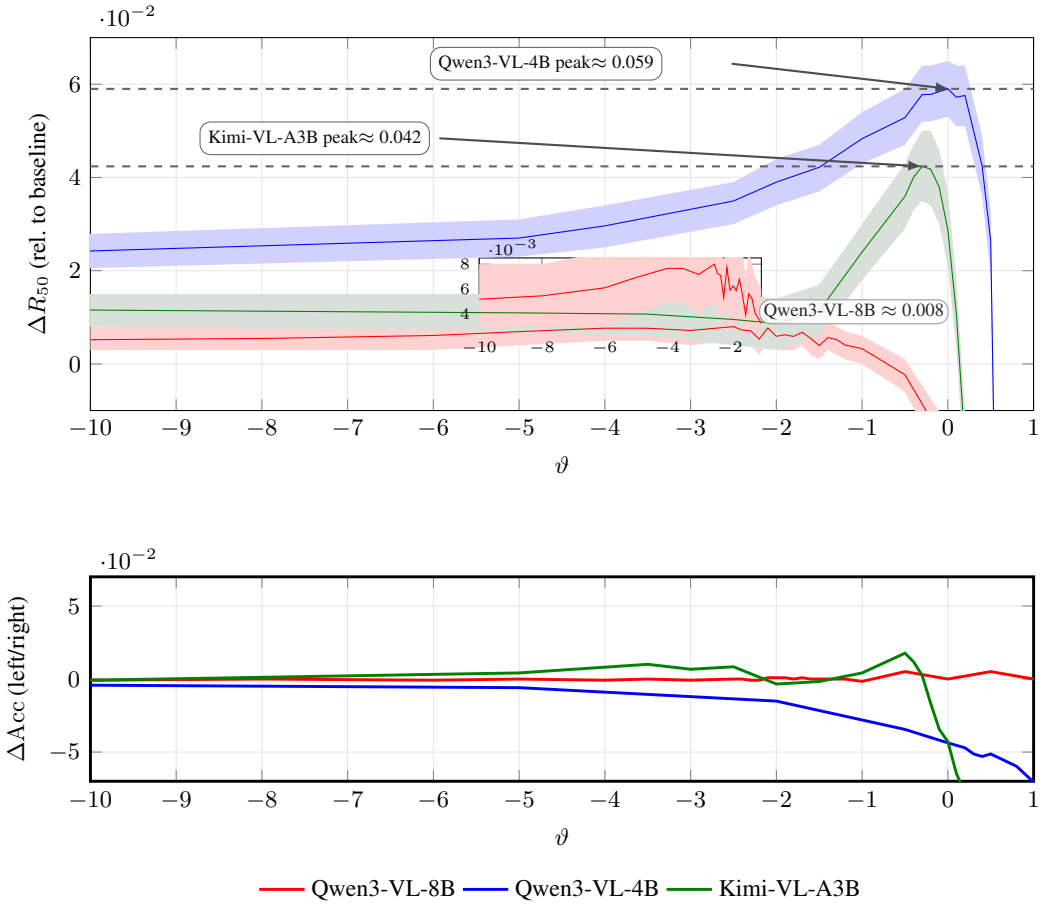

Figure 6. **Multi-model dose–response and broader spatial control. Top:** ΔR_{50} (relative to baseline ($\vartheta = -\infty$)) for Qwen3-VL-8B, Qwen3-VL-4B, and Kimi-VL-A3B across intervention strengths. Qwen3-VL-4B and Kimi-VL-A3B exhibit substantially larger interior peaks than Qwen3-VL-8B. **Bottom:** baseline-centered accuracy change on the singleton-pair left/right spatial control for the same models. Unlike the main detection/counting probe, this broader spatial control does not exhibit a reproducible bounded interior optimum across models; responses remain near baseline over much of the sweep and then diverge into model-dependent mild drift or late collapse.

regime and reversal under over-suppression, while Figure 6 shows cross-model persistence despite effect-size variation. The left/right task rules out a broad performance-gain account, and the indoor/outdoor diagnostic is retained only as a baseline-relative instability check. FLIP is therefore a diagnostic triage tool, not a deployment-safety certificate, optimization method, circuit identification result, or proof of model understanding.

Limitations. FLIP validates intervention effects; it does not identify the circuit producing them or close the gap from population-level triage to causal mechanism. The grounding-proxy analysis is scoped to detection/counting, where ΔR_{50} is observable but not a literal within-pass mediator. Within this scope, we test feature-coherence disruption, performance-based negative controls, same-site operator replication, image-cluster subsampling, multiple VLMs, and raw-layer site specificity. The raw-layer checks are conducted under the implemented threshold sweep and do not replace percentile- or norm-matched cross-site interventions. The raw-layer result is therefore best read as a readout-space contrast under this sweep, not as a claim that task-relevant information is absent from earlier layers. Bridging FLIP-positive sites to circuits through matched sweeps, patching, causal abstraction, or circuit tracing remains future work.

9. Conclusion

FLIP reframes final layer inference-time intervention as population-level effect validation for mechanistic interpretability. Rather than treating a primary-task gain as self-interpreting, the protocol requires grounding-proxy alignment, feature-coherence dependence, performance-based negative-control contrast, and specificity to the logit-facing readout. Raw-layer and percentile-matched checks further rule out a simple threshold-calibration explanation. This separates structured, task-linked sensitivity from generic perturbation and provides a lightweight screen for deciding when downstream patching, circuit tracing, or causal-abstraction analysis is warranted.

Acknowledgement

DE Juanico acknowledges the funding and technical support of the ERDT and the UP Diliman Artificial Intelligence Graduate Program.

References

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34: 9574–9586, 2021.

Drandreb Earl Juanico, Rowel Atienza, and Jeffrey Kenneth Go. Interpretable open-vocabulary referring object detection with reverse contrast attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 723–732, 2025.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Forty-Second International Conference on Machine Learning*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LBl7Hez0fF. ICLR 2025 Spotlight.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit evaluation metrics are not robust. In *First Conference on Language Modeling*, 2024.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In

The Thirteenth International Conference on Learning Representations, 2024.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.634. URL https://aclanthology.org/2025.acl-long.634/.

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.

Jingcheng Yang, Tianhu Xiong, Shengyi Qian, Klara Nahrstedt, and Mingyuan Wu. Circuit tracing in vision-language models: Understanding the internal mechanisms of multimodal thinking. *arXiv preprint arXiv:2602.20330*, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024.

A. Generic Perturbation: Definition and Scope

We use the term *generic perturbation* to denote behavioral changes induced by non-specific disruption of the representation, rather than alignment with task-relevant intermediate computations.

Formally, let $h(\mathbf{z})$ denote a task-relevant readout and $\widetilde{\mathbf{z}} = T(\mathbf{z})$ an intervention. We say that the induced change $\Delta h = h(\widetilde{\mathbf{z}}) - h(\mathbf{z})$ reflects generic perturbation if it is not systematically mediated by a task-relevant intermediate quantity, such as localization recall in the detection/counting setting studied here.

Under this definition, generic perturbation includes: (i) uniform magnitude suppression, (ii) random feature misalignment, and (iii) global shifts that do not preserve feature semantics. The distinction is operationalized in the main text through grounding-proxy alignment, feature-coherence controls, and negative-control tasks.

B. Detection Matching and Mediator Construction

Given ground-truth boxes $\{g_i\}_{i=1}^{n_g}$ and predictions $\{d_j\}_{j=1}^{n_d}$: IoU matrix L with $L_{ij} = \mathrm{IoU}(g_i, d_j)$; class-consistency matrix $N \in \{-1, +1\}$; confidence-weighted overlap $C_{ij} = s_j L_{ij}$. In our implementation, s_j is the normalized predicted-box area, used only to break ties in matching. A one-to-one assignment under IoU threshold 0.50 produces per-sample true-positive counts TP_{50}, from which recall R_{50} is computed as in Section 6. Mean IoU is used only for matching eligibility and not as the primary mediator in the main analysis.

C. Output-Head Alignment: Scope and Interpretation

The output-head alignment baseline used in the main text is a standard pre-decoding calculation, not a new representation theory. Its purpose is to make explicit why useful behavior under FLIP is non-trivial: the flooring operator is not task-conditioned, so any beneficial effect must arise from how the induced displacement projects onto task-relevant output-head directions.

For a final-token hidden state $\mathbf{z}_t$ and output matrix W, FLIP induces

$$\widetilde{\ell}_t - \ell_t = W^\top \Delta_{\vartheta,t}.$$

Thus the sign and size of any useful pre-decoding change depend on how the flooring displacement projects onto output-head directions. For a log-sum-exp margin between task-relevant tokens Y^+ and contrast tokens Y^-,

$$m(\mathbf{z}_t) = \log \sum_{y \in Y^+} \exp(w_y^\top \mathbf{z}_t) - \log \sum_{y \in Y^-} \exp(w_y^\top \mathbf{z}_t),$$

the first-order change is governed by

$$\langle \bar{w}_{Y^+}(\mathbf{z}_t) - \bar{w}_{Y^-}(\mathbf{z}_t), \Delta_{\vartheta,t} \rangle .$$

A systematic improvement therefore requires positive alignment between the non-targeted flooring displacement and task-relevant output-head directions.

This does not imply that R_{50} or $\mathcal{E}_{\mathrm{count}}$ are continuous or monotone functions of ϑ. The empirical pipeline includes autoregressive decoding, string parsing, bounding-box matching, IoU thresholding, and integer count-error scoring. These operations can introduce discontinuities. Accordingly, the paper does not claim a theorem about monotonic behavior of decoded metrics.

The role of the theory is instead to identify what a useful FLIP effect would require at the logit-facing representation. Because FLIP does not encode the object category, box coordinates, or count target, a stable interior gain is unexpected under a generic perturbation account. The empirical protocol then tests whether the behavioral signature is consistent with task-linked alignment: the gain should align with the detection-to-counting mediator, weaken when feature/output-head coherence is disrupted, localize to the normalized logit-facing site, and fail to reproduce on negative-control tasks.

D. Grounding-Proxy Interpretation and Limitations

The product summary

$$\mathbb{1}_{\mathrm{FLIP}} \to \Delta R_{50}, \qquad \mathcal{E}_{\mathrm{count}} \sim \mathbb{1}_{\mathrm{FLIP}} + \Delta R_{50}$$

is interpreted as grounding-proxy evidence, not proof of direct within-pass mediation. Localization and counting are elicited by separate prompts, so the observed localization output cannot literally cause the separately generated counting output. Instead, ΔR_{50} is used as an observable proxy for an unobserved intervention-sensitive grounding state that may be expressed as both improved localization and reduced counting error.

This analysis, therefore, supports the narrower claim that the interior regime is compatible with a shared grounding factor. It does not identify the specific circuit, exclude other unobserved mediators, or establish causal sufficiency.

A potential concern is that ΔR_{50} and $\mathcal{E}_{\mathrm{count}}$ may share a common computational ancestor through predicted bounding boxes, making the mediation relationship tautological. In the present design, localization and counting are elicited by separate prompts and scored from independent model outputs, so the observed association is not attributable to a shared output-level pipeline. The remaining alternative is a common representational cause within the model, in

which the intervention jointly affects both localization and counting. This possibility is not fully excluded, but is constrained by the Final-site permutation control and raw-layer checks, which localize the full signature to the normalized logit-facing site rather than raw decoder-layer outputs.

E. FLIP Deployment and Runtime Intervention

FLIP is implemented as an inference-time patch to logit computation: the post-normalization Final-site tensor $\mathbf{z} \in \mathbb{R}^{n \times d}$ is replaced by $\widetilde{\mathbf{z}} = \max(\mathbf{z}, \vartheta)$ before multiplication by the output projection $\mathbf{W}^\top$. This realizes the flooring operator directly at the readout site.

Strength sweep. ϑ is provided externally (e.g., a config file), enabling deterministic sweeps over a fixed grid without modifying parameters, tokenization, or decoding.

Compute resources. All COCO-based core sweeps, including the primary detection/counting runs, negative controls, raw-layer checks, same-site operator comparisons, feature-coherence controls, cross-architecture replications, and image-cluster subsampling runs, were executed as single-GPU inference jobs on one NVIDIA A100-SXM4-40GB GPU using NVIDIA-SMI/driver 570.195.03 and CUDA 12.8, with GPU memory utilization limited to 0.9. The corresponding wall-clock/GPU-hour requirements are reported in the supplemental run manifest; in our runs, a full model-level core sweep required approximately `5-6` GPU-hours, `52-53` GPU-hours, and `189` GPU-hours for Qwen3-VL-4B, Qwen3-VL-8B, and Kimi-VL-A3B, respectively. Post-hoc parsing, scoring, GLM/mediation analysis, subsample aggregation, and figure generation were CPU-side analyses and did not require GPU acceleration. Preliminary or failed runs outside the reported experiments required approximately `8 GPU-hours`.

Why the Final site. Final-site interventions are minimally invasive: they leave early visual feature extraction unchanged and perturb the normalized readout state where image/text evidence is mapped to logits.

F. Cross-Architecture Validation

Cross-architecture replication. Figure 6 shows cross-model dose–response for Qwen3-VL-8B, Qwen3-VL-4B, and Kimi-VL-A3B. Kimi-VL-A3B uses a different visual encoder, cross-modal fusion mechanism, and training recipe from the Qwen family, providing genuine cross-architecture replication. All three models exhibit the same qualitative three-phase structure. Importantly, the absolute gains are not uniformly tiny: the inset makes clear that Qwen3-VL-8B is the lower-amplitude case, while Qwen3-VL-4B and Kimi-VL-A3B show visibly larger interior peaks.

Cross-model instability diagnostic. The corresponding three-model indoor/outdoor response-switch diagnostic is shown in Figure 7. This diagnostic measures answer changes relative to baseline rather than task performance. We therefore use it only to check generic output instability; the performance-based negative-control contrast is supplied by the left/right accuracy results in the main text and Appendix J.1.

Cross-architecture mediation replications. To test whether the mediated detection-to-counting pathway extends beyond the anchor model, we report the corresponding delta-method product summary tables for Kimi-VL-A3B and Qwen3-VL-8B. These replications preserve the same qualitative interpretation as the main-text Qwen3-VL-4B table: an interior regime with $\mathrm{IRR}_{\mathrm{indirect}} < 1$ and reversal under over-suppression.

G. Raw Decoder-Layer Checks and Logit-Facing Specificity

This appendix reports raw decoder-layer checks under the same behavioral readout used in the main text. The goal is to test whether the full protocol is selectively satisfied at the normalized logit-facing Final site or appears trivially in raw decoder-layer spaces. All checks are evaluated through final outputs, so any effect must propagate to decoded behavior. The full result is shown in Figure 8.

These checks separate depth, dose, and validation. Early/middle layers remain near baseline before collapse, while late raw layers may show partial ΔR_{50} movement. We additionally include percentile-matched raw-layer sweeps for Layers 8, 16, and 32 to test whether the failure is explained by marginal threshold mismatch. These matched sweeps do not recover the Final-site response and instead attenuate or worsen the raw-layer dose–response. The accepted protocol, however, is not defined by nonzero ΔR_{50} alone: it requires negative-control contrast, grounding-proxy alignment, and feature-coherence dependence. The post-normalization Final curve is the logit-facing reference; the raw last-block output is a separate pre-normalization site.

Active-set and magnitude calibration. For each site, we track token-level active-set fraction $A_{\ell,t}(\vartheta)$ and mean clamp magnitude $M_{\ell,t}(\vartheta)$ (Section 5.1 and Table 4). A measures how many hidden dimensions are touched; M measures average displacement. Layers 8 and 16 are not inactive: at $\vartheta = 0$, their A values match the Final site, but their M values are roughly an order of magnitude smaller, followed by rapid saturation. Layer 32 is closer in A/M and can move ΔR_{50}, but still lacks the full feature-coherence-dependent

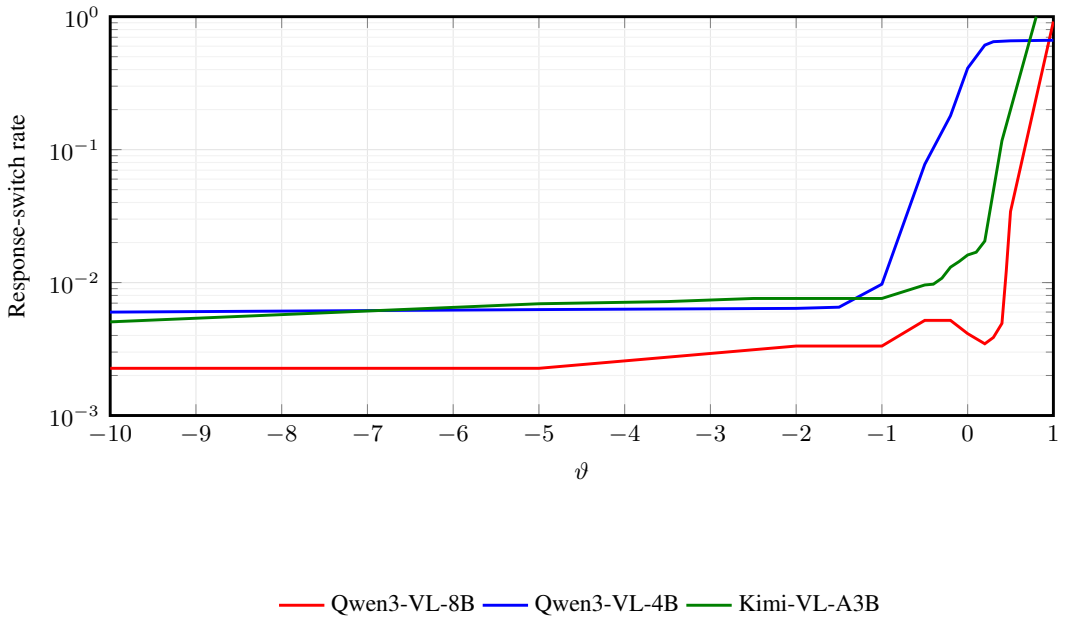

Figure 7. **Indoor/outdoor response switch-rate comparison.** Response switch-rate curves (defined in Appendix S) across Qwen3-VL-8B, Qwen3-VL-4B and Kimi-VL-A3B.

Table 2. **Delta-method grounding-proxy association for $\mathbb{1}_{\mathrm{FLIP}}$, ΔR_{50}, and $\mathcal{E}_{\mathrm{count}}$ on Kimi-VL-A3B.** a: effect of FLIP on the recall proxy ΔR_{50}; b: Poisson-GLM association of ΔR_{50} with $\log \mathbb{E}[\mathcal{E}_{\mathrm{count}}]$; $a \times b$: product summary; $\mathrm{IRR}_{\mathrm{indirect}} = \exp(a \times b)$, interpreted as a grounding-proxy association rather than direct within-pass mediation. Values below 1 indicate that improved recall is associated with reduced tolerant counting error.

ϑ	a	SE(a)	b	SE(b)	$a \times b$	SE($a \times b$)	p-value	$\mathrm{IRR}_{\mathrm{indirect}}$ [95% CI]
−3.5	0.00419	0.00231	−2.659	0.114	−0.01115	0.00615	0.0697	0.9889 [0.9771, 1.0009]
−3.0	0.00189	0.00241	−2.645	0.115	−0.0050	0.0064	0.4342	0.9950 [0.9826, 1.0075]
−2.5	−0.00020	0.00257	−2.668	0.110	0.00054	0.00685	0.9376	1.0005 [0.9872, 1.0141]
−2.0	0.00308	0.00269	−2.659	0.113	−0.00819	0.00715	0.2523	0.9918 [0.9780, 1.0058]
−1.5	0.00841	0.00292	−2.657	0.108	−0.02236	0.00780	0.0042	0.9779 [0.9630, 0.9930]
−1.0	0.02151	0.00312	−2.719	0.100	−0.05847	0.00875	< 0.0001	0.9432 [0.9272, 0.9595]
−0.5	0.03977	0.00357	−2.733	0.095	−0.10866	0.01047	< 0.0001	0.8970 [0.8788, 0.9156]
−0.4	0.04543	0.00368	−2.772	0.092	−0.12595	0.01102	< 0.0001	0.8817 [0.8628, 0.9009]
−0.3	0.04925	0.00374	−2.767	0.091	−0.13624	0.01129	< 0.0001	0.8726 [0.8535, 0.8922]
−0.2	0.05015	0.00382	−2.694	0.103	−0.13510	0.01153	< 0.0001	0.8736 [0.8541, 0.8936]
−0.1	0.04375	0.00383	−2.635	0.097	−0.11529	0.01095	< 0.0001	0.8911 [0.8722, 0.9104]
0.0	0.03282	0.00392	−2.624	0.099	−0.08613	0.01078	< 0.0001	0.9175 [0.8983, 0.9371]
0.1	0.01402	0.00408	−2.632	0.103	−0.03690	0.01083	0.0007	0.9638 [0.9435, 0.9844]
0.2	−0.01465	0.00407	−2.564	0.091	0.03755	0.01051	0.0004	1.0383 [1.0171, 1.0599]
0.3	−0.05500	0.00395	−2.572	0.094	0.14144	0.01140	< 0.0001	1.1519 [1.1265, 1.1780]
0.8	−0.23269	0.00535	−2.638	0.127	0.61381	0.03279	< 0.0001	1.8475 [1.7325, 1.9701]
1.0	−0.23269	0.00535	−2.638	0.127	0.61381	0.03279	< 0.0001	1.8475 [1.7325, 1.9701]

Table 3. **Delta-method grounding-proxy association for $\mathbb{1}_{\mathrm{FLIP}}$, ΔR_{50}, and $\mathcal{E}_{\mathrm{count}}$ on Qwen3-VL-8B.** a: effect of FLIP on the recall proxy ΔR_{50}; b: Poisson-GLM association of ΔR_{50} with $\log \mathbb{E}[\mathcal{E}_{\mathrm{count}}]$; $a \times b$: product summary; $\mathrm{IRR}_{\mathrm{indirect}} = \exp(a \times b)$, interpreted as a grounding-proxy association rather than direct within-pass mediation. Values below 1 indicate that improved recall is associated with reduced tolerant counting error.

ϑ	a	SE(a)	b	SE(b)	$a \times b$	SE($a \times b$)	p-value	$\mathrm{IRR}_{\mathrm{indirect}}$ [95% CI]
-60.0	0.00036	0.00024	-2.240	0.121	-0.00080	0.00053	0.1319	0.9992 [0.9982, 1.0002]
-40.0	0.00179	0.00054	-2.252	0.119	-0.00402	0.00124	0.0012	0.9960 [0.9936, 0.9984]
-20.0	0.00532	0.00106	-2.296	0.116	-0.01220	0.00251	< 0.001	0.9879 [0.9830, 0.9927]
-10.0	0.00519	0.00128	-2.272	0.119	-0.01180	0.00297	< 0.001	0.9883 [0.9825, 0.9940]
-5.0	0.00693	0.00155	-2.272	0.123	-0.01575	0.00362	< 0.001	0.9844 [0.9774, 0.9914]
-4.0	0.00766	0.00160	-2.279	0.115	-0.01747	0.00375	< 0.001	0.9827 [0.9755, 0.9899]
-2.5	0.00800	0.00158	-2.248	0.118	-0.01799	0.00368	< 0.001	0.9822 [0.9751, 0.9893]
-2.0	0.00592	0.00153	-2.254	0.112	-0.01335	0.00351	< 0.001	0.9867 [0.9800, 0.9935]
-1.5	0.00394	0.00153	-2.276	0.115	-0.00897	0.00350	0.010	0.9911 [0.9843, 0.9979]
-1.0	0.00327	0.00157	-2.249	0.119	-0.00735	0.00355	0.038	0.9927 [0.9858, 0.9996]
-0.5	-0.00232	0.00188	-2.222	0.113	0.00515	0.00418	0.218	1.0052 [0.9970, 1.0134]
0.0	-0.01799	0.00255	-2.312	0.110	0.04160	0.00622	< 0.001	1.0425 [1.0299, 1.0552]

signature. Thus raw-layer failures are not explained by too few clamped dimensions; the complete protocol localizes to the post-normalization logit-facing site. The percentile-matched overlays in Figure 4 further show that this localization is not removed by matching the marginal intervention scale at representative early, middle, and late raw-layer sites.

Table 4. **Token-level calibration for raw-layer checks.** Means of active-set fraction A and mean clamp magnitude M under unpermuted flooring. Early/middle layers show matched A but much smaller M at $\vartheta = 0$, then rapid saturation; Layer 32 is closer to the Final site but does not reproduce the full Final-site signature.

Layer	ϑ	Final A	Final M	Layer A	Layer M
8	0.0	0.500	0.968	0.500	0.091
8	0.5	0.602	1.122	0.979	0.504
8	1.0	0.662	1.453	0.999	1.001
16	0.0	0.498	0.963	0.497	0.102
16	0.5	0.591	1.039	0.980	0.508
16	1.0	0.631	1.317	0.998	1.006
32	0.0	0.454	0.821	0.498	0.639
32	0.5	0.535	1.000	0.643	0.926
32	1.0	0.608	1.192	0.758	1.277

H. Redistribution Breadth and Dispersion Diagnostics

H.1. Target-Conditioned Patch Relevance and RMR Computation

This appendix describes how target-conditioned patch relevance maps are computed and how redistribution diagnostics are derived from them. These quantities are used only as comparative diagnostics of redistribution geometry: they do not replace the primary mediator analysis based on ΔR_{50} and $\mathcal{E}_{\mathrm{count}}$.

Target-conditioned saliency maps. For each image–prompt pair, we compute a baseline patch relevance map and a FLIP-intervened patch relevance map conditioned on a specified target token sequence, such as the first `bbox_2d` field or a generated detection answer. The relevance pipeline first identifies the target token positions in the full sequence (prompt plus generated response). We then define y_{target} as a scalar target score for that sequence, computed as the sum of the logits assigned to the identified target tokens, and compute layerwise gradient-weighted attention relevance maps of the form

$$R^{(\ell)} = \mathrm{ReLU}\left(\frac{\partial y_{\mathrm{target}}}{\partial A^{(\ell)}} \odot A^{(\ell)} \right), \tag{12}$$

averaged over heads. These per-layer maps are combined into a joint relevance matrix by augmenting with the identity, row-normalizing, and multiplying cumulatively across layers.

To sharpen target specificity, the implementation also computes token-level Grad $\times$ Hidden importance from the final hidden state,

$$g_i = \left\langle h_i, \frac{\partial y_{\mathrm{target}}}{\partial h_i} \right\rangle, \tag{13}$$

with non-target positions masked to zero. The masked token-importance vector is propagated through the joint relevance matrix and restricted to the visual patch-token indices, yielding one scalar relevance value per image patch. For Qwen3-VL, these values are arranged on the model's 24×24 patch grid. Qualitative heatmaps are produced by interpolation and overlay, but all reported statistics are computed directly from the underlying patch relevance vectors.

Baseline and FLIP conditions. The same target-conditioned relevance procedure is run twice: once under the baseline condition and once under FLIP. Let $s_{-\infty} \in \mathbb{R}^P$ denote the baseline patch saliency map and $s_\vartheta \in \mathbb{R}^P$ the FLIP-intervened map, where P is the number of image patches. We define the patchwise saliency shift

$$\delta = s_\vartheta - s_{-\infty}. \tag{14}$$

All redistribution diagnostics are computed from δ.

Directional RMR diagnostics. Let $\mathcal{O}$ denote the set of object patches under the patch mask induced by the parsed bounding box. We report three directional redistribution statistics:

$$\mathrm{RMR}_{+}(\vartheta) = \sum_{p \in \mathcal{O}} \max(\delta_p, 0), \tag{15}$$

$$\mathrm{RMR}_{-}(\vartheta) = \sum_{p \in \mathcal{O}} \max(-\delta_p, 0), \tag{16}$$

$$\mathrm{RMR}_{\mathrm{obj}}(\vartheta) = \sum_{p \in \mathcal{O}} \delta_p - \sum_{p \notin \mathcal{O}} \delta_p. \tag{17}$$

Thus, RMR_{+} measures saliency mass shifted toward object patches, RMR_{-} measures saliency mass shifted away from object patches (reported as a positive magnitude), and $\mathrm{RMR}_{\mathrm{obj}}$ measures the resulting object-versus-background contrast. The object mask is derived from parsed bounding boxes projected onto the patch grid, so $\mathrm{RMR}_{\mathrm{obj}}$ is a relative contrast measure rather than a raw object saliency total.

Interpretive role. These quantities are used to refine the interpretation of the FLIP sweep. In the negligible regime, directional redistribution remains weak. In the interior regime, bounded positive redistribution toward object patches becomes detectable. Under over-suppression, object-aligned gains weaken while off-object loss and redistribution breadth increase. The diagnostics therefore complement the primary behavioral analysis without being treated as formal mediators or faithful causal explanations.

H.2. Comparative role of rollout

The target-conditioned saliency maps above are constructed using gradient-weighted attention relevance propagation with rollout. Their most reliable use in this paper is comparative: differences across ϑ for the same model, prompt, and target are informative, whereas absolute saliency magnitudes should not be over-interpreted. For coordinate-token targets, tokenization positions can shift across interventions, so target indices are anchored to baseline outputs and reused across FLIP runs. If a target token is absent at a given ϑ, that sample is marked missing for RMR computation at that ϑ.

H.3. Regime interpretation of the RMR diagnostics

The RMR statistics are used to relate redistribution geometry to the regime structure observed in the main dose–response analysis. They are not formal mediators and do not replace the primary evidence based on ΔR_{50} and $\mathcal{E}_{\mathrm{count}}$. Their role is interpretive: to indicate how target-conditioned relevance mass is reallocated across the FLIP sweep.

Negligible regime. At very negative ϑ, both ΔR_{50} and RMR_{+} are small, indicating that FLIP is effectively inactive at the redistribution level. This is the redistribution-side signature of the negligible-change regime. Consistently, RMR_{-} remains comparatively modest and $\mathrm{RMR}_{\mathrm{obj}}$ stays near zero in this region.

Interior regime. In the bounded interior regime, RMR_{+} rises and co-varies with ΔR_{50}, as shown directly in Figure 9(b). This indicates that the positive detection-side response is accompanied by bounded object-aligned redistribution of relevance mass. In this regime, RMR_{-} increases only gradually, and $\mathrm{RMR}_{\mathrm{obj}}$ remains comparatively modest. The resulting picture is one of constructive redistribution rather than generic suppression.

Over-suppression regime. As ϑ approaches the collapse region, RMR_{+} declines, RMR_{-} rises sharply, and $\mathrm{RMR}_{\mathrm{obj}}$ spikes late. This transition indicates that the regime has changed qualitatively. The late rise in $\mathrm{RMR}_{\mathrm{obj}}$ should not be read as evidence of a stronger constructive interior regime; instead, it is more consistent with suppression-dominant contrast enhancement, in which object-relative saliency rises because off-object mass is removed broadly.

Interpretive role. Taken together, the four-panel analysis refines the regime picture as follows: the negligible regime shows little redistribution; the interior regime is the point of strongest object-aligned reallocation, indexed by co-movement of RMR_{+} and ΔR_{50}; and the over-suppression regime is marked by rising RMR_{-} and late spikes in $\mathrm{RMR}_{\mathrm{obj}}$. Thus, the RMR diagnostics provide a redistribution-side corroboration of the main regime structure without elevating curve shape alone into a grounding-specificity claim.

To connect redistribution geometry more directly to the regime structure, we plot RMR_{+}, RMR_{-}, and $\mathrm{RMR}_{\mathrm{obj}}$ against ϑ for Qwen3-VL-8B, together with a direct overlay of RMR_{+} and ΔR_{50}.

I. Vision-Centric Benchmark Consistency

This appendix reports additional results on vision-centric benchmarks to assess whether the regime structure observed in the main text extends beyond the detection-based evaluation setting.

Scope and role. These benchmarks are not designed to isolate the detection-derived mediator used in the main analysis. As such, they do not constitute mediator-aligned evidence. Instead, they serve as *consistency checks* on whether intervention sweeps produce structured, non-monotonic response patterns in broader evaluation settings.

Benchmarks and setup. We evaluate performance under the same intervention sweep across MMStar, MindCube, and NaturalBench. Each benchmark emphasizes different aspects of vision–language performance, including reasoning, compositionality, and general visual understanding. Unlike the core intervention sweeps, which use fixed random seeds for replicability where supported, these benchmark evaluations are executed through the third-party LMMS-Eval stack. We therefore replicate the benchmark runs multiple times under independent vLLM initializations of the same Qwen3-VL-8B model and report standard deviations across those runs as the error bars in Figure 10. These LMMS-Eval benchmark data were generated on an NVIDIA GB10 system with 128 GB unified memory, using NVIDIA-SMI/driver 580.142 and CUDA 13.0, with GPU memory utilization set to 0.3; the supplemental run manifest reports the corresponding wall-clock/GPU-hour requirements for the MMStar, MindCube, and NaturalBench sweeps.

Observed regime patterns. Figure 10 shows that all three benchmarks exhibit structured response curves under increasing intervention strength ϑ.

MMStar remains elevated across a relatively wide range of ϑ, suggesting a broad plateau rather than a sharply localized interior optimum. MindCube exhibits the clearest three-phase structure, with an identifiable interior regime separating negligible and collapse regions. NaturalBench shows a peak in the interior regime followed by degradation under stronger suppression.

While the precise shape and width of the regimes vary across benchmarks, all three display departures from monotonic degradation, consistent with the existence of structured response regimes.

Interpretation. These results provide supporting evidence that the emergence of regime structure is not confined to the detection-recall metric used in the main analysis. However, because these benchmarks do not isolate the detection-derived mediator, they do not establish causal compatibility in the sense of Section 6.

Accordingly, we interpret these findings as *regime-consistency evidence* rather than primary evidence of grounding-proxy alignment. Because the error bars reflect run-to-run variability in the LMMS-Eval serving stack rather than uncertainty in the core fixed-seed intervention pipeline, they should be read as corroborative robustness information rather than as the primary inferential basis of the paper.

Relation to redistribution diagnostics. The regime patterns observed here are consistent with redistribution diagnostics reported in Appendix H, where intermediate intervention strengths correspond to structured changes in attention-derived measures. These diagnostics are presented as complementary geometric indicators rather than causal explanations; Appendix H.3 shows that, for the 8B hard case, RMR_+ tracks the interior regime while RMR_- and $\mathrm{RMR}_{\mathrm{obj}}$ characterize the transition to over-suppression.

J. Negative-Control Task Design

The primary negative control is the singleton-pair left/right spatial task in Appendix J.1. Unlike the indoor/outdoor response-switch diagnostic, this task is scored by accuracy and therefore tests whether the same intervention sweep produces a genuine performance gain on a distinct vision-language control task. The intended contrast is not that every control task must degrade monotonically, but that it should not reproduce the coherent positive interior regime observed for the detection/counting pathway.

We retain the indoor/outdoor experiment only as an auxiliary instability diagnostic: its response-switch rate measures whether answers change relative to baseline, not whether performance improves.

J.1. Singleton-Pair Left/Right Spatial Control

The singleton-pair left/right task is used as a broader spatial negative control. Its purpose is not to instantiate the main detection-derived mediator, but to test whether FLIP produces a reproducible bounded interior optimum on a distinct vision-language task that depends on coarse relational judgment rather than full localization/counting structure.

Dataset construction. The dataset is built from MS COCO val2017 in clustered-by-image form so that downstream analysis can still use image-level cluster-robust inference. For a candidate image to qualify, it must satisfy the same image-centric filtering logic used in the main clustered construction, including minimum category-richness requirements. Candidate left/right pairs are then formed only from *singleton categories*: the image must contain exactly one instance of object_1 and exactly one instance of object_2. Additional filters require sufficient horizontal separation between the two objects and may also impose minimum area and maximum-overlap constraints. By default, at most one pair is sampled per image to avoid overweighting visually busy scenes.

Prompt and label construction. Each retained pair yields a yes/no spatial-relation question of the form `Is the <object1> to the left of the <object2>? Answer only yes or no.` To maintain near-balanced labels, prompts are generated by alternating the true ordering and the swapped ordering across qualifying pairs, producing an overall dataset that is approximately $50/50$ yes/no.

Querying. The left/right control is queried directly from the stored text field rather than being rebuilt from extracted object names at inference time. The querying infrastructure otherwise matches the local chat-completions pipeline used for the other intervention experiments, including image-path resolution, payload construction, response extraction, and threshold bookkeeping.

Interpretive role. Because this control uses unambiguous singleton pairs, errors are less likely to come from multiplicity or ambiguous reference than in general localization prompts. At the same time, it does not instantiate the detection-to-counting mediator used in the main analysis. We therefore use $\Delta\mathrm{Acc}$ on this task as the performance-based negative control: it tests whether the intervention produces a comparable positive interior regime on a distinct spatial judgment, rather than merely causing answer switching.

K. Baseline Centering, Paired Designs, and Heterogeneity

This subsection clarifies the role of baseline-centered recall in the total-effect model and its relationship to an equivalent paired-difference formulation under image-clustered inference.

Point-estimate equivalence. Let $Y_i(\vartheta)$ denote the detection-side mediator for unit i under intervention strength ϑ, with baseline $Y_i(-\infty)$ corresponding to the no-intervention condition. In the present study, each image–query pair is evaluated under both conditions, so the same units appear in both arms.

In a binary-treatment regression with an intercept, the coefficient on the treatment indicator equals the difference in sample means:

$$\widehat{\alpha}_1 = \overline{Y}(\vartheta) - \overline{Y}(-\infty). \tag{18}$$

If instead we define per-sample paired differences

$$\delta_i = Y_i(\vartheta) - Y_i(-\infty), \tag{19}$$

then the sample mean satisfies

$$\overline{\delta} = \frac{1}{N}\sum_{i=1}^{N} \delta_i = \overline{Y}(\vartheta) - \overline{Y}(-\infty). \tag{20}$$

At the population level, the same equality follows from linearity of expectation:

$$\mathbb{E}[Y_i(\vartheta) - Y_i(-\infty)] = \mathbb{E}[Y_i(\vartheta)] - \mathbb{E}[Y_i(-\infty)]. \tag{21}$$

Thus, when both conditions are evaluated on the same set of units, global centering and paired subtraction target the same estimand and yield the same point estimate.

Clustered inference for the paired design. The equivalence above concerns the treatment contrast, not a claim that baseline centering is more conservative than paired subtraction. Because the pooled Gaussian GLM uses image-level cluster-robust standard errors, baseline and intervention observations from the same image contribute jointly to the cluster score. The resulting variance estimate therefore accounts for the within-cluster covariance induced by evaluating matched image–query units under both conditions. Under the balanced paired design used here, a pooled treatment-indicator regression with image-clustered standard errors and an equivalent paired-difference regression with image-clustered standard errors target the same large-sample uncertainty for the treatment contrast, up to finite-sample and implementation differences. We use the baseline-centered pooled specification because it keeps the reported ΔR_{50} scale aligned with the dose-response figures, mediation summaries, and stratified GLM table, not because it is inherently more conservative.

What this does and does not resolve about heterogeneity. The equivalence above establishes identical point estimates for the pooled treatment contrast, but it does not by itself rule out heterogeneous effects across subgroups. We therefore examine one interpretable source of heterogeneity directly: instance count. The updated Qwen3-VL-4B stratification shows that the small-count stratum closely mirrors the pooled pattern across the interior regime. The large-count stratum also remains positive and significant over the main interior window, but its gains are smaller and attenuate earlier, becoming non-significant at $\vartheta = 0.5$. The over-suppression reversal at $\vartheta = 1.0$ is significant in both strata, indicating that the regime structure—interior improvement followed by collapse under strong intervention—is preserved across this subdivision.

These findings argue against the concern that the main qualitative conclusions are an artifact of aggregation along this dimension. If global centering were masking strong opposing subgroup effects, one would expect sign instability across strata; instead, both strata preserve the positive interior direction and both exhibit collapse under over-suppression. The heterogeneity is primarily in effect size and attenuation: few-object scenes show stronger and more persistent gains, while many-object scenes weaken earlier. At the same time, this analysis addresses heterogeneity only along one interpretable axis. It does not exclude the possibility that other sources of heterogeneity—such as object category, occlusion, or image difficulty—could modulate effect size. A broader distributional heterogeneity analysis remains an important direction for future work.

L. Regression Specifications and Robust Inference

We estimate:

$$\Delta R_{50} = \alpha_0 + \alpha_1 \mathbb{1}_{\mathrm{FLIP}} + \varepsilon, \quad (22)$$

$$\log \mathbb{E}[\mathcal{E}_{\mathrm{count}} \mid \mathbb{1}_{\mathrm{FLIP}}, \Delta R_{50}] = \gamma_0 + \gamma_1 \mathbb{1}_{\mathrm{FLIP}} + \gamma_2 \Delta R_{50}. \quad (23)$$

Standard errors are clustered at the image level to account for dependence among object queries from the same image and for the covariance between matched baseline and intervention observations within each image cluster. The first equation is fit as a Gaussian GLM with image-clustered SEs and supplies the a-term; the second is fit as a Poisson GLM on strict counting responses and supplies the b-term. The auxiliary ZeroInflatedPoisson model used in the evaluation code belongs only to the two-part format-failure/tolerant-error decomposition for E_{combined}; it is not used for the reported product summary.

Cluster-robust stratified GLM. Table 5 reports cluster-robust Gaussian GLM estimates stratified by instance count (≤ 3 vs. > 3 objects) for the Qwen3-VL-4B anchor model. The pooled estimate is practically small at $\vartheta = -50$, then rises across the interior regime, remaining positive and significant through $\vartheta = 0.2$ before attenuating at $\vartheta = 0.5$ and reversing sharply at $\vartheta = 1.0$ ($p < 0.001$). The small-count stratum closely tracks the pooled pattern. The large-count stratum is also positive and significant over the main interior window, but attenuates earlier and is not significant at $\vartheta = 0.5$, consistent with stronger effects on cleaner few-object scenes and weaker gains as visual multiplicity increases.

M. Operator-Comparison Baselines and Feature-Coherence Control

This appendix separates two distinct comparisons that serve different roles in the paper. The first is a full four-criterion replication across same-site magnitude-bounding operators; the second is the graduated permutation control for the flooring instantiation itself. Together they test whether the protocol generalizes across operator choice while still isolating feature-coherence dependence.

M.1. Full Four-Criterion Replication Across Same-Site Operators

All operator-comparison baselines intervene at the same site as FLIP: the post-normalization Final-site tensor passed to the output head.

Symmetric clipping. The symmetric-clipping baseline is

$$\widetilde{\mathbf{z}}^{\mathrm{clip}} = \mathrm{clip}(\mathbf{z},\ -|\vartheta|,\ +|\vartheta|)\,. \quad (24)$$

Additive shift. The additive-shift baseline is a matched-mean translation:

$$\Psi_\vartheta(\mathbf{z}) = \mathbf{z} + \mu_\vartheta(\mathbf{z})\mathbf{1}, \qquad \mu_\vartheta(\mathbf{z}) = \frac{1}{nd}\sum_{i,k}(\vartheta - z_{ik})_+. \quad (25)$$

That is, the hidden state is shifted uniformly by the mean flooring effect that the corresponding flooring operator would induce at the same ϑ.

Interpretation of operator replication. Figure 12 provides the compact operator-baseline comparison at the level of primary-task dose–response alone, while Figure 11 extends the comparison to the full protocol for same-site magnitude-bounding interventions. Across flooring, symmetric clipping, and additive shift, the primary task exhibits a bounded positive regime under low or no permutation. By contrast, the singleton-pair left/right control, scored by $\Delta\mathrm{Acc}$ rather than response switching, remains near baseline or drifts downward over the non-collapse region and does not reproduce a comparable positive interior regime. The grounding-proxy pathway is likewise altered once feature coherence is disrupted. The operator-side result is therefore not that all operators behave identically, but that the protocol continues to distinguish structured, mediator-aligned regimes from generic perturbation effects across same-site bounded interventions.

Figure 11 shows the complete same-site operator-side replication of the four-criterion protocol.

M.2. Graduated Permutation Control

The graduated permutation control is not an alternative operator baseline. It keeps the flooring instantiation fixed and tests whether its positive interior regime depends on feature coherence.

Let $\mathbf{z} \in \mathbb{R}^{n\times d}$ denote the post-normalization Final-site tensor, with permutation acting along the feature dimension. For mixing parameter $\lambda \in [0,1]$, define a *partial random permutation operator* $S_\lambda \in \mathbb{R}^{d\times d}$ by

$$S_\lambda = \Pi_\lambda \oplus I_{1-\lambda}, \quad (26)$$

where: (i) Π_λ is a random permutation acting on a randomly selected subset of $\lceil \lambda d \rceil$ feature dimensions, (ii) $I_{1-\lambda}$ is the identity on the remaining feature dimensions, and (iii) $\oplus$ denotes the direct sum over the partition of $\{1,\ldots,d\}$ into permuted and unpermuted coordinates.

The graduated permutation control is then defined by

$$\widetilde{\mathbf{z}}^{(\lambda)} = \max(\mathbf{z}S_\lambda, \vartheta), \quad (27)$$

where the maximum is taken element-wise and no inverse permutation is applied. This omission is deliberate: because

elementwise flooring is permutation equivariant, an inverse-restored construction would satisfy

$$\max(\mathbf{z}S_\lambda, \vartheta)S_\lambda^{-1} = \max(\mathbf{z}, \vartheta),$$

and would therefore collapse back to standard flooring rather than stress feature–output-head alignment. Equivalently, a fraction λ of feature dimensions is permuted *without inverse restoration* before flooring, while the remaining $(1-\lambda)$ fraction is left unchanged.

This construction has clean boundary behavior:

$$\widetilde{\mathbf{z}}^{(0)} = \max(\mathbf{z}I, \vartheta) = \max(\mathbf{z}, \vartheta), \tag{28}$$

$$\widetilde{\mathbf{z}}^{(1)} = \max(\mathbf{z}P, \vartheta), \tag{29}$$

where P is a full random permutation on all d feature dimensions. Thus $\lambda = 0$ recovers standard flooring, while larger λ values progressively destroy feature alignment by increasing the fraction of permuted coordinates.

For all reported permutation-control curves, baseline centering uses the common unpermuted no-intervention run. Let $R_{50}^{(\lambda)}(\vartheta)$ denote recall after applying $\widetilde{\mathbf{z}}^{(\lambda)} = \max(\mathbf{z}S_\lambda, \vartheta)$. We plot

$$\Delta R_{50}^{(\lambda)}(\vartheta) = R_{50}^{(\lambda)}(\vartheta) - R_{50}^{(0)}(-\infty),$$

not $R_{50}^{(\lambda)}(\vartheta) - R_{50}^{(\lambda)}(-\infty)$. The a-term in the corresponding $\mathrm{IRR}_{\mathrm{indirect}}$ panels uses the same common-baseline $\Delta R_{50}^{(\lambda)}(\vartheta)$. Thus any degradation caused by permutation at $\vartheta = -\infty$ remains visible in the plotted level rather than being removed by λ-specific re-centering.

Figure 5 therefore tests whether the unpermuted interior window survives feature-coordinate misalignment. Relative to $\lambda = 0$, all nonzero λ settings lose the positive regime earlier and enter collapse sooner, although the ordering among nonzero λ values is not strictly monotonic. The key contrast is between $\lambda = 0$ and $\lambda > 0$: disrupting feature coherence destabilizes the regime-relevant window even when the intervention class is otherwise unchanged.

M.3. Related Operator Intuition: Symmetric Clipping and Additive Shift

The main text develops the output-head alignment baseline for flooring because FLIP is the primary intervention. Symmetric clipping and additive shift are used as same-site operator-comparison baselines. The relevant comparison is not that these operators are theoretically identical, but that each induces a displacement at the same logit-facing representation.

Symmetric clipping. The symmetric-clipping baseline is

$$\widetilde{\mathbf{z}}^{\mathrm{clip}} = \mathrm{clip}(\mathbf{z}, -|\vartheta|, +|\vartheta|).$$

It induces a displacement

$$\Delta_\vartheta^{\mathrm{clip}} = \widetilde{\mathbf{z}}^{\mathrm{clip}} - \mathbf{z}.$$

As $|\vartheta|$ shrinks along the sweep, more coordinates are clipped and the induced displacement becomes broader. Any useful effect must therefore arise from alignment between $\Delta_\vartheta^{\mathrm{clip}}$ and task-relevant output-head directions, not from clipping magnitude alone.

Additive shift. The additive-shift baseline applies

$$\Psi_\vartheta(\mathbf{z}) = \mathbf{z} + \mu_\vartheta(\mathbf{z})\mathbf{1}, \qquad \mu_\vartheta(\mathbf{z}) = \frac{1}{nd}\sum_{i,k}(\vartheta - z_{ik})_+.$$

Its displacement is global rather than support-local:

$$\Delta_\vartheta^{\mathrm{shift}} = \mu_\vartheta(\mathbf{z})\mathbf{1}.$$

For fixed $\mathbf{z}$, $\mu_\vartheta(\mathbf{z})$ is non-decreasing in ϑ, so the shift magnitude grows along the sweep. As with flooring and clipping, a useful effect requires the induced displacement to project favorably onto task-relevant output-head directions.

Interpretive role. The operator comparison therefore asks whether regime structure and feature-coherence dependence survive changes in the displacement geometry. A regime-shaped curve alone is not sufficient; the key question is whether the effect remains mediator-aligned, coherence-dependent, localized to the logit-facing site, and absent from negative-control tasks.

N. Subsample Stability of the Interior Regime

This appendix tests whether the interior regime is stable under reductions in effective sample size and under the finite granularity of the ϑ sweep. Because scanning many ϑ values can otherwise overstate small peaks, we use the grid only to define a descriptive FWHM regime window and require the resulting window to remain visible under image-cluster subsampling rather than claiming a separate confirmatory discovery at each grid point.

The subsampling procedure adds an argument `--n-subsamples` N (default 1). Each subsample uses seed $\texttt{base_seed} + s_{\mathrm{idx}}$, making the full subsampling pipeline reproducible. For each subsample, we draw an independent permutation of the image list without replacement. Fractions within a given subsample are then formed cumulatively from that single permuted ordering, so that $10\% \subset 20\% \subset 40\% \subset 60\% \subset 80\% \subset 100\%$ within the same subsample.

Figure 13 shows that the qualitative regime is preserved across fractions, the within-fraction means track the full-sample curve closely, and variability decreases with scale.

This supports the interpretation that the interior regime is a stable signal rather than a small-sample artifact. These image-cluster resampling results also serve as a robustness check on the large-sample clustered-inference interpretation used for the reported 95% confidence intervals.

O. Additional Qualitative Illustration

O.1. Illustrative Regime Visualization

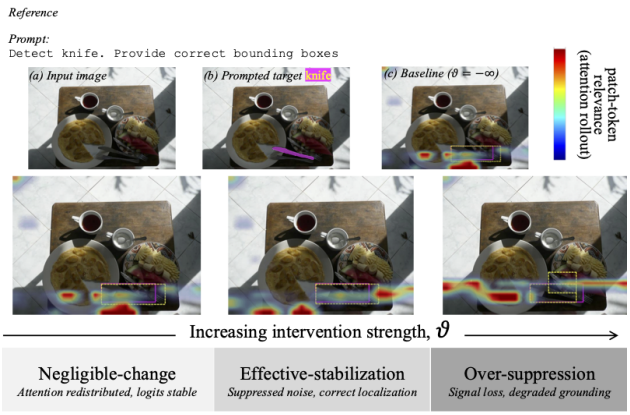

Figure 14. **Illustrative regime structure under increasing intervention strength ϑ.** As ϑ increases, behavior shifts from negligible change to an interior regime with qualitatively improved detections, and then to over-suppression. Heatmaps show patch-token relevance from attention rollout; solid boxes denote ground truth and dashed boxes denote model detections. Representative example only.

P. Proofs and Supporting Analysis

This appendix supports the operator and output-head statements used in Section 5. We state only hidden-state and pre-decoding facts; no claim is made that decoded metrics are continuous or monotone in ϑ.

P.1. Componentwise Flooring Identity and Monotonicity

For a scalar $x \in \mathbb{R}$ and threshold $\vartheta \in \mathbb{R}$, define

$$\phi_\vartheta(x) = \max(x, \vartheta), \qquad \delta_\vartheta(x) = \phi_\vartheta(x) - x = (\vartheta - x)_+.$$

Applied elementwise to $\mathbf{z} \in \mathbb{R}^{n\times d}$,

$$\widetilde{\mathbf{z}} = \phi_\vartheta(\mathbf{z}), \qquad \mathbf{\Delta}_\vartheta = \widetilde{\mathbf{z}} - \mathbf{z}.$$

For each coordinate (i, k),

$$\widetilde{z}_{ik} = z_{ik} + (\vartheta - z_{ik})_+.$$

Proof of Proposition 5.1. The identity follows immediately from the definition of max. If $\vartheta_1 \le \vartheta_2$, then

$$(\vartheta_1 - z_{ik})_+ \le (\vartheta_2 - z_{ik})_+$$

for every coordinate. Hence each entry of $\mathbf{\Delta}_\vartheta$ is coordinate-wise non-decreasing in ϑ, the active support expands monotonically, and $\|\mathbf{\Delta}_\vartheta\|_p$ is non-decreasing for any $p \in [1, \infty]$.

P.2. Output-Head Margin Expansion

For a final-token hidden state $\mathbf{z}_t$, logits are

$$\ell_t = W^\top \mathbf{z}_t + b.$$

Under FLIP,

$$\widetilde{\mathbf{z}}_t = \mathbf{z}_t + \Delta_{\vartheta,t}, \qquad \Delta_{\vartheta,t} = (\vartheta - \mathbf{z}_t)_+,$$

so the logit change is exactly

$$\widetilde{\ell}_t - \ell_t = W^\top \Delta_{\vartheta,t}.$$

For token sets Y^+ and Y^-, define

$$m(\mathbf{z}_t) = \log \sum_{y\in Y^+} \exp(w_y^\top \mathbf{z}_t) - \log \sum_{y\in Y^-} \exp(w_y^\top \mathbf{z}_t).$$

The gradient of this margin is

$$\nabla m(\mathbf{z}_t) = \bar{w}_{Y^+}(\mathbf{z}_t) - \bar{w}_{Y^-}(\mathbf{z}_t),$$

where each $\bar{w}$ is the softmax-weighted average of output-head vectors within the corresponding token set. A first-order Taylor expansion gives

$$m(\widetilde{\mathbf{z}}_t) - m(\mathbf{z}_t) = \langle \bar{w}_{Y^+}(\mathbf{z}_t) - \bar{w}_{Y^-}(\mathbf{z}_t), \Delta_{\vartheta,t}\rangle + o(\|\Delta_{\vartheta,t}\|).$$

Thus a useful pre-decoding margin increase requires positive alignment between the FLIP displacement and task-relevant output-head directions.

For raw decoder-layer sites, the analogous first-order effect must pass through the remaining computation:

$$\widetilde{\ell} - \ell \approx J_{\ell\to\text{logits}} \Delta_\vartheta^{(\ell)}.$$

The raw-layer analysis therefore tests whether such propagated alignment reproduces the signature observed at the normalized logit-facing representation.

Q. FLIP versus Prompt-Based Control

Prompt engineering operates exclusively at the level of the input distribution and relies on the model's learned priors to steer generation. By contrast, FLIP is an inference-time intervention applied after tokenization and cross-modal fusion, operating directly on the post-normalization Final-site hidden state before logit computation. FLIP does not introduce new linguistic content, cannot inject information absent from the model's internal representations, and fails gracefully when evidence is weak or over-suppressed—unlike prompt-based control.

R. Model Choice, Scope, and Claims

The core detection/counting experiments are instantiated on Qwen3-VL-4B-Instruct to enable controlled, auditable inference-time patching. Claims for the mediator-based regime analysis are therefore scoped to: (i) this model family instance, (ii) the post-normalization Final intervention site, and (iii) the evaluation protocol. The contribution is the *methodology*—regime-structured intervention sweeps, a mediator-based probe, and redistribution diagnostics. Cross-architecture validation is reported in Appendix F.

S. Auxiliary Instability Diagnostic: Global Scene Classification

For the indoors/outdoors diagnostic, answer stability is quantified via a *baseline-referenced response-switch rate*. Let $y_i^{(-\infty)}$ and $y_i^{(\vartheta)}$ denote the baseline and intervention outputs for sample i. Outputs are canonicalized before comparison by lowercasing, extracting the first alphabetic token, and mapping surface variants to the binary label set `indoors`/`outdoors` (e.g., `indoor`, `indoors`, `indo` $\mapsto$ `indoors`; `outdoor`, `outdoors`, `out` $\mapsto$ `outdoors`). Defining $c(\cdot)$ as this canonicalization map, the response-switch rate at strength ϑ is

$$\mathrm{RSR}(\vartheta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big\{c\Big(y_i^{(\vartheta)}\Big) \neq c\Big(y_i^{(-\infty)}\Big)\Big\}. \tag{30}$$

This metric measures response instability relative to baseline rather than task accuracy.

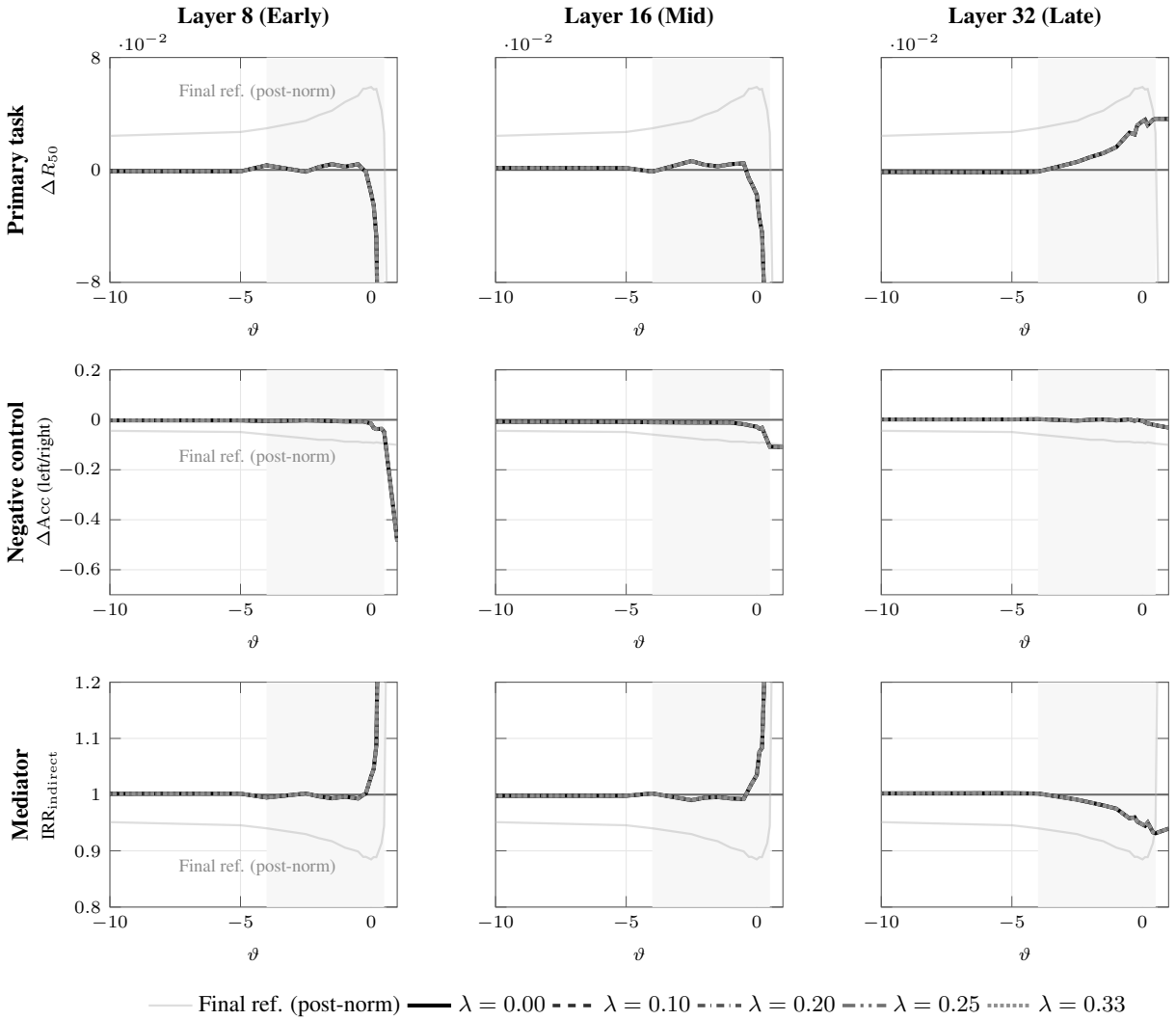

Figure 8. **Full protocol checks for raw decoder-layer sites.** Columns show representative raw decoder-layer outputs; rows show primary-task ΔR_{50}, left/right $\Delta\mathrm{Acc}$, and $\mathrm{IRR}_{\mathrm{indirect}} = \exp(a \times b)$ under permutation mixing. The post-normalization Final curve is the logit-facing reference. Raw layers may show partial ΔR_{50} movement, but they do not reproduce the full feature-coherence-dependent Final-site signature.

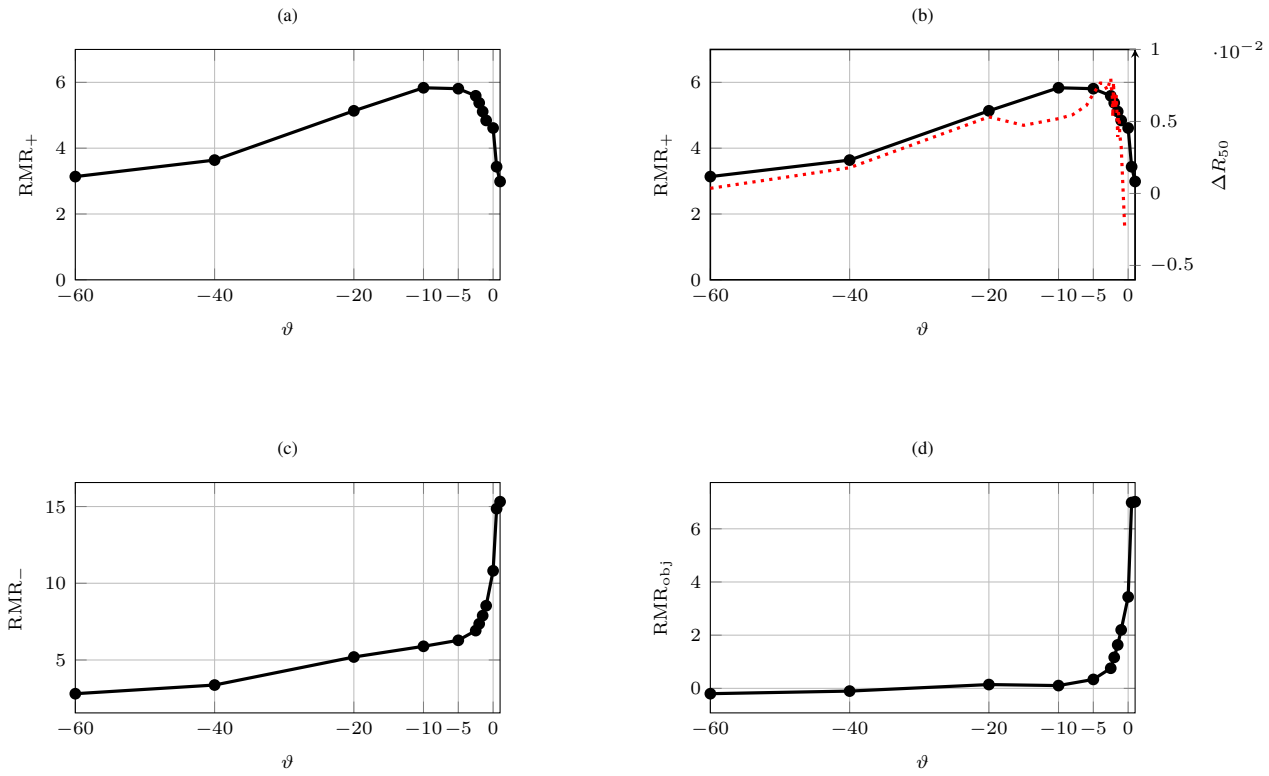

Figure 9. **Qwen3-VL-8B redistribution diagnostics across the FLIP sweep.** (a) ϑ versus RMR_+, which peaks in the same broad interior region where the detection-side response is positive. (b) Dual-axis overlay of ϑ versus RMR_+ (left axis) and ϑ versus ΔR_{50} (right axis), showing redistribution-side co-variation with the behavioral dose–response. (c) ϑ versus RMR_-, which grows toward collapse and marks increasing off-object saliency loss. (d) ϑ versus $\mathrm{RMR}_{\mathrm{obj}}$, which remains modest through the interior regime but spikes late, consistent with suppression-dominant contrast enhancement rather than constructive redistribution. Together, the panels link the appendix RMR diagnostics to the three-regime structure of the main FLIP sweep.

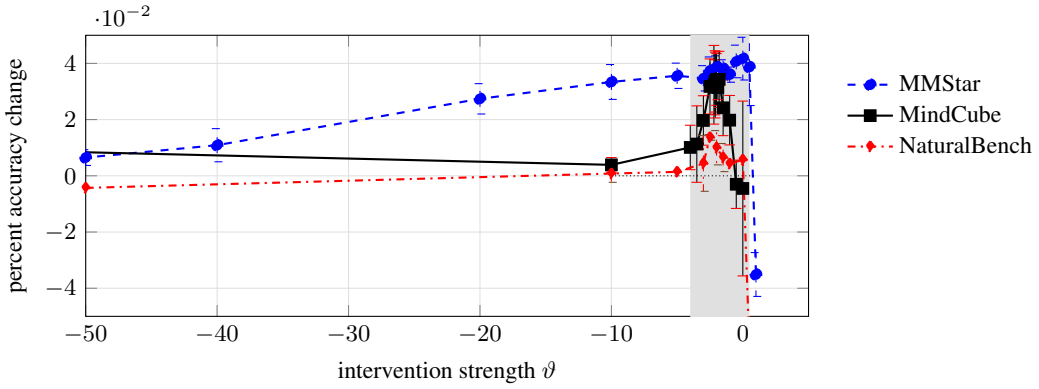

Figure 10. **Dose–response of FLIP on vision-centric benchmarks.** Curves show benchmark-specific performance changes relative to baseline. Error bars denote standard deviation across repeated LMMS-Eval runs using independent vLLM initializations of the same Qwen3-VL-8B model. The shaded band indicate the regime partition defined from the primary detection/counting sweep under the current R_{50} mediator. MMStar, MindCube, and NaturalBench show structured, non-monotonic responses consistent with the main regime picture.

Table 5. **Cluster-robust Gaussian GLM estimates of the intervention effect on ΔR_{50}, stratified by instance count for Qwen3-VL-4B.** Effects are reported pooled and by object-count bin; the coefficient of interest is the slope on $\mathbb{1}_{\mathrm{FLIP}}$.

ϑ	Setting	$\hat{\alpha}_1$	SE	p-value
-50.0	Pooled	0.0021	0.0010	< 0.001
	Small (≤ 3)	0.0023	0.0010	< 0.001
	Large (> 3)	0.0010	0.0010	0.075
-5.0	Pooled	0.0268	0.0020	< 0.001
	Small (≤ 3)	0.0283	0.0020	< 0.001
	Large (> 3)	0.0185	0.0030	< 0.001
-2.5	Pooled	0.0350	0.0020	< 0.001
	Small (≤ 3)	0.0366	0.0030	< 0.001
	Large (> 3)	0.0256	0.0040	< 0.001
-2.0	Pooled	0.0388	0.0020	< 0.001
	Small (≤ 3)	0.0412	0.0030	< 0.001
	Large (> 3)	0.0254	0.0040	< 0.001
-1.5	Pooled	0.0422	0.0030	< 0.001
	Small (≤ 3)	0.0444	0.0030	< 0.001
	Large (> 3)	0.0297	0.0040	< 0.001
-1.0	Pooled	0.0483	0.0030	< 0.001
	Small (≤ 3)	0.0508	0.0030	< 0.001
	Large (> 3)	0.0347	0.0050	< 0.001
-0.5	Pooled	0.0528	0.0030	< 0.001
	Small (≤ 3)	0.0539	0.0030	< 0.001
	Large (> 3)	0.0472	0.0050	< 0.001
0.0	Pooled	0.0590	0.0030	< 0.001
	Small (≤ 3)	0.0606	0.0040	< 0.001
	Large (> 3)	0.0500	0.0050	< 0.001
0.1	Pooled	0.0572	0.0030	< 0.001
	Small (≤ 3)	0.0594	0.0040	< 0.001
	Large (> 3)	0.0454	0.0050	< 0.001
0.2	Pooled	0.0576	0.0030	< 0.001
	Small (≤ 3)	0.0604	0.0040	< 0.001
	Large (> 3)	0.0419	0.0050	< 0.001
0.5	Pooled	0.0264	0.0040	< 0.001
	Small (≤ 3)	0.0324	0.0040	< 0.001
	Large (> 3)	-0.0068	0.0060	0.247
1.0	Pooled	-0.5315	0.0060	< 0.001
	Small (≤ 3)	-0.5736	0.0060	< 0.001
	Large (> 3)	-0.2992	0.0090	< 0.001

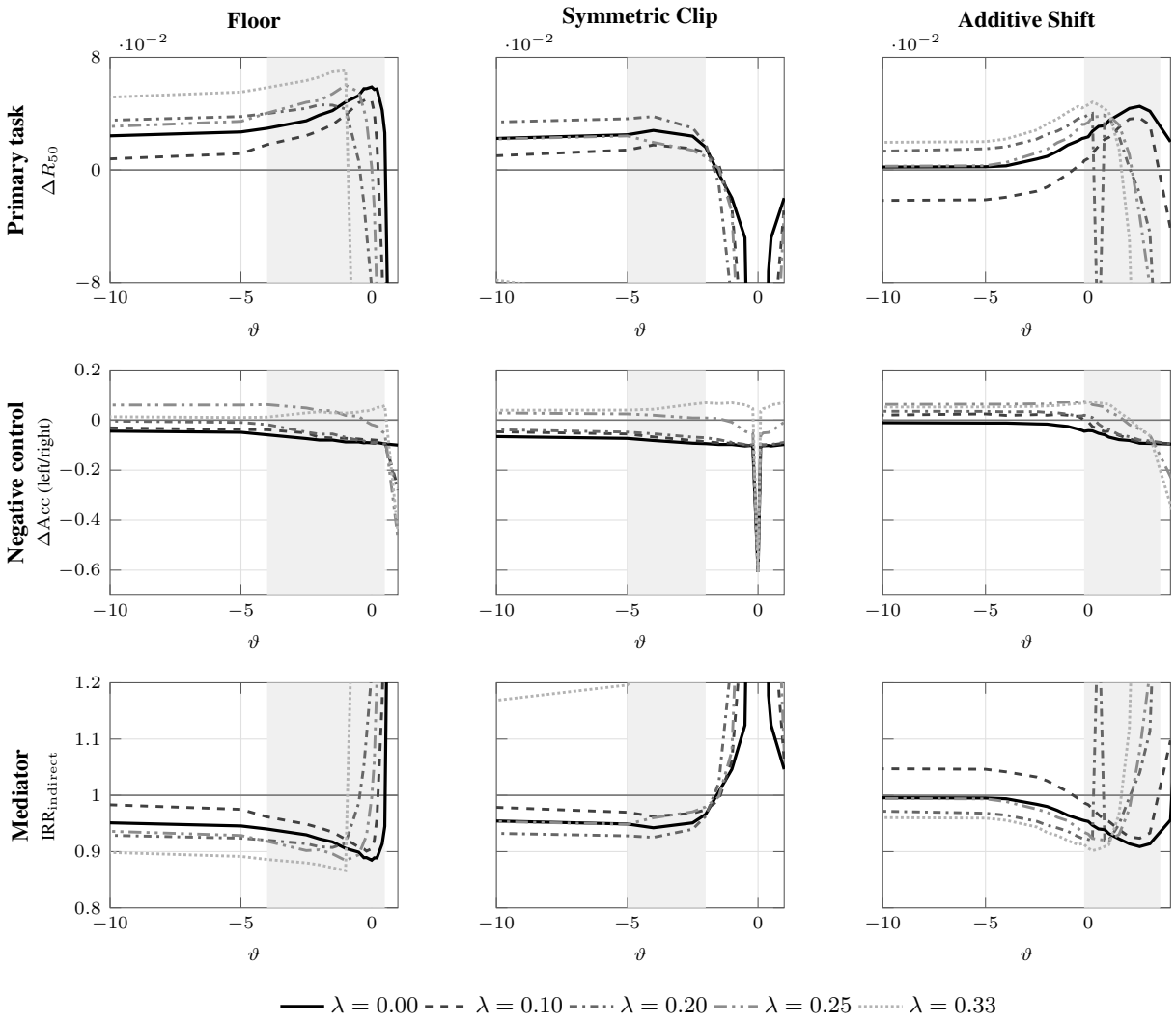

Figure 11. **Full four-criterion replication across same-site magnitude-bounding operators.** Columns correspond to flooring (FLIP), symmetric clipping, and additive shift. Rows show the primary-task dose–response ΔR_{50}, the singleton-pair left/right spatial control scored by $\Delta\mathrm{Acc}$, and the mediated indirect effect $\mathrm{IRR}_{\mathrm{indirect}} = \exp(a \times b)$ under graduated permutation mixing. Solid black denotes $\lambda = 0$; gray dash patterns denote increasing mixing. All baseline-centered quantities use the corresponding unpermuted no-intervention baseline, so nonzero-λ curves are not re-centered to already permuted baselines. Across operators, the primary task shows a bounded positive regime under low or no permutation, while coherence disruption destabilizes that regime. The left/right control remains near baseline or drifts downward over the non-collapse region and does not reproduce a comparable positive interior regime. The indirect-effect panels show that the mediated pathway is likewise altered under coherence disruption.

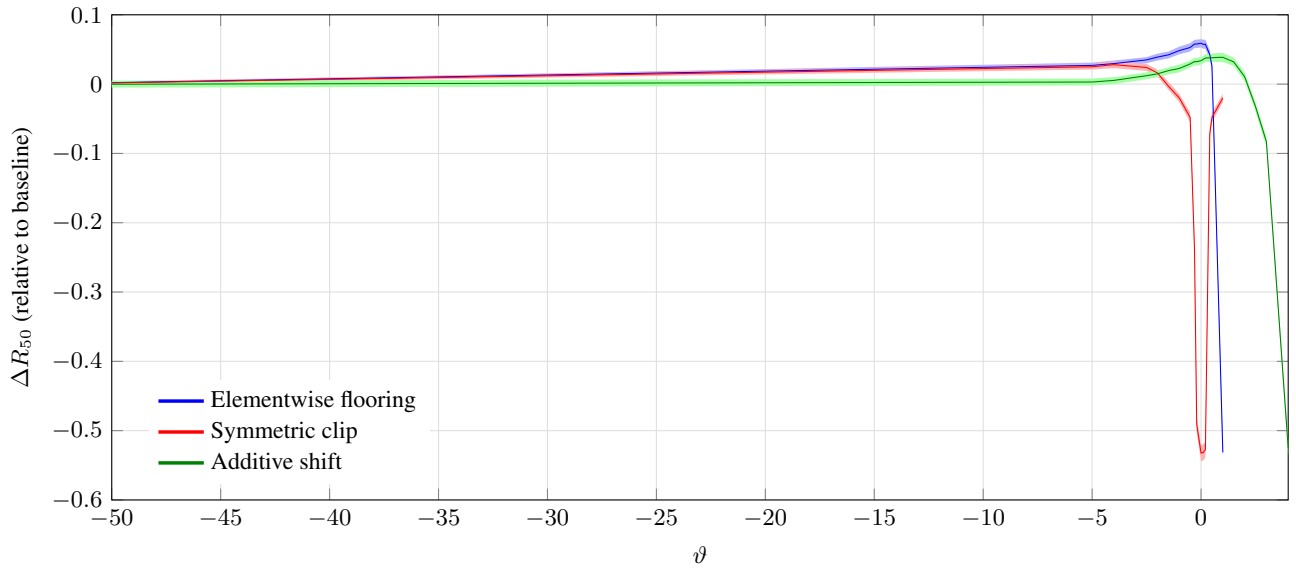

Figure 12. **Dose–response comparison across same-site operator baselines.** ΔR_{50} under flooring, symmetric clipping, and additive shift, with 95% confidence intervals. All three bounded interventions exhibit a bounded interior regime, but they differ in strength and collapse profile: flooring reaches the largest interior peak, symmetric clipping collapses earliest, and additive shift remains positive over a broader range before delayed collapse. The paper's specificity claim is therefore narrower than operator uniqueness: curve shape alone is insufficient, and the decisive tests are mediator alignment and feature-coherence dependence.

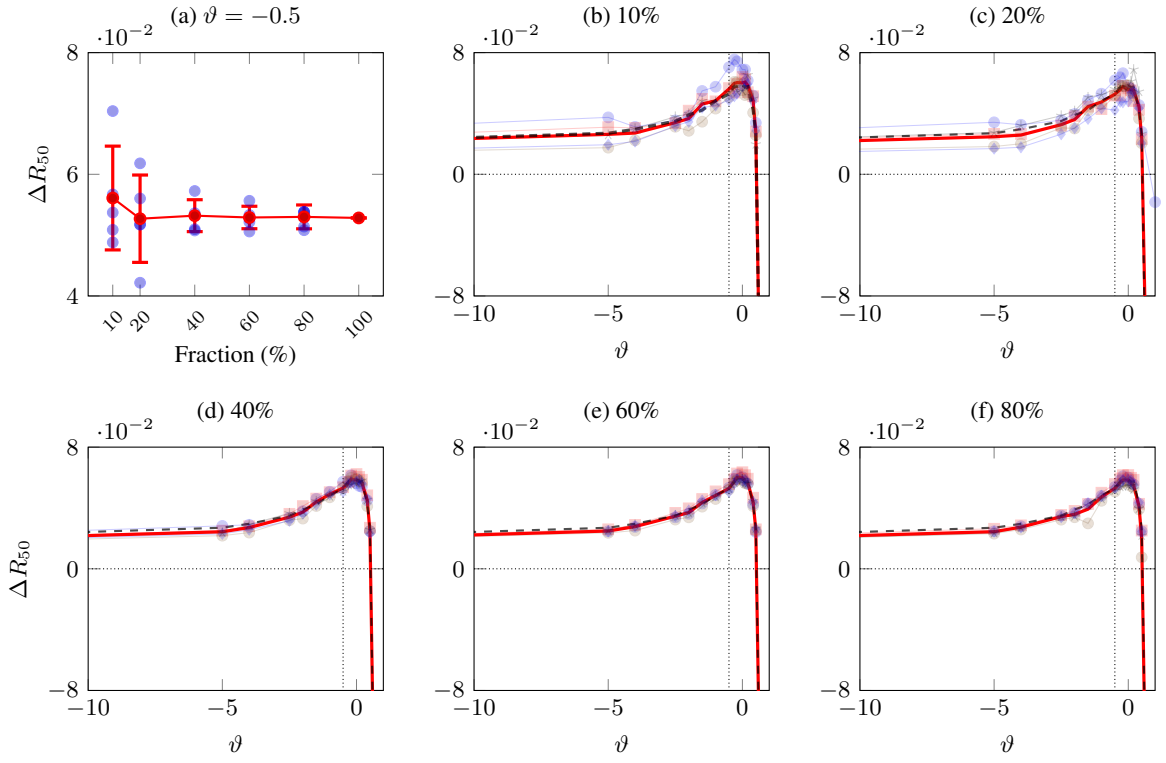

Figure 13. **Subsample stability of the interior regime. (a)** ΔR_{50} at fixed $\vartheta = -0.5$ across subsample fractions. Blue markers show individual resamples; red markers and error bars show the subsample mean and variability. Fractions are formed by subsampling image clusters. The 100% condition is the single full-dataset estimate and therefore has no error bar. **(b–f)** Dose–response curves for five resamples per fraction, with the red curve denoting the within-fraction mean and the dashed black curve the full-dataset reference. The qualitative regime is preserved across fractions, the means track the full-sample curve closely, and variability decreases with scale, indicating that the effect is a stable signal rather than a small-sample artifact.