OBJECT-TOKEN INFLUENCE IN VISION–LANGUAGE MODELS: WHERE IT LIVES AND HOW IT ADDS UP

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–Language Models (VLMs) extend language models with an image encoder, but how visual tokens covering objects operate inside the LM backbone remains less understood than in text-only settings. Prior work has typically examined different forms of object-token influence separately, without comparing them under a common depth statistic. We study object-token influence in seven adapter-style VLMs along three operationally distinct axes—erasure sensitivity, cross-instance transfer, and category readout. Placing all three on a common half-mass depth coordinate, we find the same order in every model: erasure sensitivity precedes cross-instance transfer, which in turn precedes category readout. The two intervention axes concentrate in shallow-to-mid layers, whereas the readout axis concentrates in mid-to-deep layers, yielding an intervention–readout gap. Whole-set object-token interventions exhibit supra-additive effects relative to a matched background contrast; this excess concentrates in shallow layers and is indistinguishable from zero at deeper layers. These findings identify candidate depths for targeted intervention and motivate visual-token pruning criteria that account for joint rather than purely tokenwise effects.

1 INTRODUCTION

Modern Vision–Language Models (VLMs) transform an image into a sequence of visual tokens that is processed by a pretrained language model (LM) together with text (Liu et al., 2023; 2024; Li et al., 2023a). These tokens arrive at the LM backbone already contextualized by the vision encoder, and are then repeatedly mixed with the text prompt across layers before contributing to the output. Recent interpretability studies suggest that visual tokens covering an object carry a disproportionate amount of object-relevant information (Neo et al., 2025; Basu et al., 2024; Palit et al., 2023; Kaduri et al., 2025). For brevity, we refer to these spatially localized visual tokens as *object tokens*.

However, an object token can be said to have "influence" in several operationally distinct senses. Replacing it tests whether the current output is susceptible to losing its representation (Neo et al., 2025; Basu et al., 2024); transferring its hidden state to another image tests whether it carries causally portable content (Meng et al., 2022; Wang et al., 2023; Palit et al., 2023); and comparing it with category-labeled text states tests whether its category is readable from the representation (nostalgebraist, 2020; Belrose et al., 2023; Vilas et al., 2023; Jiang et al., 2025; Krojer et al., 2026). These measurements define *erasure sensitivity*, *cross-instance transfer*, and *category readout*, respectively. Prior work has typically characterized these forms of object-token influence separately, using method-specific summaries such as layerwise peaks; the closest precedent (Huang & Chang, 2025) contrasts layerwise causal-effect and decodability curves in a single vision transformer, but only qualitatively rather than under a common depth statistic.

Beyond *where*, a separate question is whether intervention effects combine additively across the tokens composing an object. An object usually spans multiple visual patches, but token-level interventions are often interpreted individually or aggregated as though their effects combined additively. This assumption need not hold in a transformer: an intervention on the whole object-token set may have an effect different from the sum of interventions on its parts.

We study both questions across seven VLMs from five families. Because the three axes have different raw scales, we convert each layerwise signal into a normalized mass profile and locate the relative depth by which half of its mass has accumulated, giving a scale-free summary of where

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.

each effect concentrates. To test additivity, we compare the effect of intervening on the full object-token set with the summed effects on its two disjoint halves at the same depth. Our contributions are threefold:

1. We formulate erasure sensitivity, cross-instance transfer, and category readout as normalized layerwise mass profiles and compare their half-mass depths on a common relative-depth axis.
2. The common-support analysis places the three axes in a consistent depth order, with erasure sensitivity and cross-instance transfer concentrating in shallow-to-mid layers and category readout in mid-to-deep layers—including a separation between the two intervention axes, both of which replace hidden states and measure changes in the output answer.
3. We introduce a background-controlled measure of whole-set non-additivity and show that the resulting supra-additive excess concentrates in shallow layers and is indistinguishable from zero at deeper layers.

Together, these findings clarify where different forms of object-token influence concentrate and how their effects combine across the token set. Code and analysis artifacts are available at `https://anonymous.4open.science/r/object-token-influence-D31C`.

2 Background and Research Questions

2.1 Visual Tokens in the LM Backbone

Visual and object tokens. An image I is encoded by a CLIP-style vision transformer and mapped through a learned adapter into N *visual tokens* that share the LM's embedding space (Liu et al., 2024); the LM processes the concatenation of visual and text tokens through its residual stream. Given an annotated object instance in the image, we call the visual tokens whose image patches cover that instance the *object tokens*; the remaining visual tokens fall on background patches.

Answer score. Hidden states along the residual stream can be read as distributions over the vocabulary at each depth (nostalgebraist, 2020); R instead compares them against text states (§4.1.3), and interventions are scored at the output. For an input x whose prompt occupies n positions and a target string tokenized as $y = (y_0, \ldots, y_{m-1})$, the model's answer score is

$$M_y(x) = \frac{1}{m} \sum_{k=0}^{m-1} \log \operatorname{softmax}\big(\operatorname{logits}_{n-1+k}(x \oplus y)\big)[y_k], \tag{1}$$

the mean teacher-forced log-probability of y.

2.2 Open Questions and Hypotheses

The overarching question of this work is how object-token influence operates in the LM backbone of a VLM. We separate it into two sub-questions, one about *where* such influence is located across depth and one about *how* it aggregates over the token set that carries it.

2.2.1 Where object-token influence concentrates

We consider three operationally distinct axes: erasure sensitivity (Δ, does erasing an object token break the answer?), cross-instance transfer (C, does moving it into another image change the answer there?), and category readout (R, can the category be read from the object token's state?). Erasure and transfer provide two complementary intervention-based views of object-token influence—susceptibility to losing receiver content and causal portability across instances—whereas R provides a non-interventional view through the state's geometry. Each axis has been characterized across layers on its own, and causal effect and probe-based decodability have been separated in a single vision transformer on one task (§6), but as qualitative observations of separate curves rather than measurements under one location statistic. We place all three profiles on a common depth coordinate and ask where each one concentrates (§4).

2.2.2 WHETHER OBJECT-TOKEN EFFECTS COMBINE ADDITIVELY

Object-level interventions in the literature are reported either at the granularity of the full object mask or for single tokens alone; between these regimes sits the implicit assumption that the effect of intervening on the full object-token set is well summarized by the sum of the effects on its parts. This need not hold in a transformer: a whole-set intervention may differ from the sum of half-set interventions. We test for such non-additivity and characterize its depth profile at each fixed layer using a background contrast matched in size and intervened on in the same way (§5).

3 SETUP

3.1 MODELS

We evaluate seven adapter-style VLMs from five families: LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2024), Idefics3-8B (Laurençon et al., 2024), Gemma-3-4B (Team, 2025), Qwen2.5-VL-3B and Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3.5-8B (Wang et al., 2025). Their LM backbones have $L = 28$ to $L = 40$ layers, so all depth statistics are placed on the relative-depth axis $\rho_\ell = \ell/L \in [0, 1)$; measurements are taken on a fixed per-model grid $\mathcal{G}$ of eleven layers (Appendix A.1). An intervention at layer ℓ replaces residual states in place: we write $x[\ell; S \mapsto v]$ for a forward pass in which each position $i \in S$ has its residual state replaced by $v(i) \in \mathbb{R}^d$ at layer ℓ (Appendix A.2).

3.2 DATASETS

All images and instance masks come from COCO val2017 (Lin et al., 2014). Receiver and donor images are drawn from disjoint subsets of the source pool (Appendix A.3). Two prompts appear throughout, both applied to COCO images: a *naming prompt* asking the model to identify the object in the image (input x_I^{name}), and an *existence prompt* asking a yes/no question about a category y (input x_B^{exist} for the receiver's own category y_B); exact strings, adapted from the POPE format (Li et al., 2023b), are in Appendix A.4. Category readout (§4.1.3) additionally uses a bank of $20{,}000$ Visual Genome region descriptions (Krishna et al., 2017), each matched against COCO category expressions and encoded once per model as a text-only input (Appendix A.5).

3.2.1 PAIRS

An analysis pair $g = (A, B)$ consists of a donor image A and a receiver image B with instances of different categories $y_A \neq y_B$. Let $\mathcal{O}_A, \mathcal{O}_B$ be their object-token sets (defined below). A candidate pair enters the shared pool $\mathcal{P}_0$ only if, in every one of the seven models, both sets are non-empty and $\mathrm{IoU}(\mathcal{O}_A, \mathcal{O}_B) \geq 0.4$; a deterministic receiver round-robin fills $\mathcal{P}_0$ to $1{,}000$ pairs. Within a pair, receiver and donor tokens are matched rank by rank in raster order up to a per-pair cap of 60, giving the receiver token set $\mathcal{T}_g$ and a donor correspondence $\sigma_g : \mathcal{T}_g \to \mathcal{O}_A$. Every measurement in §4–§5 is keyed to $\mathcal{T}_g$: C transplants the state at $\sigma_g(i)$ into position i under the naming prompt, R reads that same state without intervention, and Δ overwrites position i together with the rest of $\mathcal{T}_g$ under the existence prompt (Appendix A.6).

3.2.2 PREPROCESSING

For each model, we project every COCO instance mask onto the model's grid of final visual tokens: a token is an object token of annotation a if the mask covers at least $\theta = 0.3$ of its cell and a is the argmax annotation there. We write $\mathcal{O}(a)$ for this set, and $\mathcal{U}$ for the background set of tokens that no annotation covers to θ; the receiver's background $\mathcal{U}_B$ supplies the replacement states of Δ and of the set-interaction interventions in §5. Full construction and sensitivity to θ are in Appendix A.7.

4 DEPTH ORGANIZATION OF OBJECT-TOKEN INFLUENCE

We compare three object-token profiles across depth—two intervention axes and one non-interventional readout axis (§2.2.1). We first construct their layerwise signals, then summarize where each concentrates using the same half-mass depth statistic.

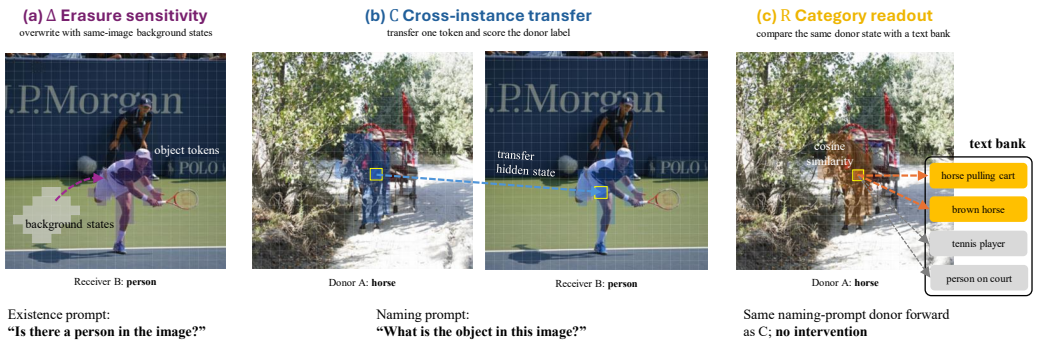


Figure 1: Illustration of the three operational axes for a single sample pair $g = (A, B)$, consisting of donor image A and receiver image B. The axes are keyed to the receiver object-token set $\mathcal{T}_g$ and its donor correspondence. (a) Erasure sensitivity Δ simultaneously overwrites the receiver object-token states with same-image background states under an existence prompt. (b) Cross-instance transfer C transplants a corresponding donor state into one receiver position per forward pass under a naming prompt and scores the donor label. (c) Category readout R uses the same donor state as in (b), without intervention, and compares it by cosine similarity against a category-labeled text bank. See §4.1.1–4.1.3 for formal definitions and baselines.

4.1 Layerwise Measurements and Baselines

Because the three axes measure different quantities, each layerwise signal is defined on the fixed layer grid relative to an axis-specific baseline.

4.1.1 Erasure Sensitivity (Δ)

Erasure sensitivity uses activation patching (Zhang & Nanda, 2024) to test whether overwriting the object-token states reduces the correct-answer margin (Figure 1a). Under the existence prompt, the answer is scored by the yes/no margin $m(x) = \max_{y \in \mathcal{Y}_{\text{yes}}} M_y(x) - \max_{y \in \mathcal{Y}_{\text{no}}} M_y(x)$, which is positive when the model favors yes (Appendix B.1). The receiver's category is present by construction, so the ground-truth answer to the existence question is yes; the model's clean answer is recorded but is not used to filter pairs. For each pair g, a deterministic map $\beta_g : \mathcal{T}_g \to \mathcal{U}_B$ assigns a background patch to each object-token position; the object-side arms of §5 reuse the map (Appendix A.2). Let h^{ℓ}_{B,β_g} denote the layer-ℓ background-state assignment, $i \mapsto h^{\ell}_{B,\beta_g(i)}$. We overwrite receiver object-token states simultaneously and record the margin shift:

$$\Delta_g(\ell) = m\big(x_B^{\text{exist}}[\ell;\, \mathcal{T}_g \mapsto h^{\ell}_{B,\beta_g}]\big) - m\big(x_B^{\text{exist}}\big). \tag{2}$$

Thus, $\Delta_g(\ell) < 0$ when the overwrite moves the answer toward "no". The baseline $b_\Delta \equiv 0$ is certified by an identity arm that rewrites each $i \in \mathcal{T}_g$ with its own state. The arm reproduces the clean forward bit-exactly, yielding $\Delta_g(\ell) = 0$ at all sampled layers for all seven models and showing that the patching procedure itself introduces no change (Appendix B.2). We orient the signal so that larger values indicate stronger erasure: $s^{\Delta}_g(\ell) = -\Delta_g(\ell)$, matching the sign convention of s^C and s^R. Using the measured background response $\delta^{\text{full}}_{\text{bg},g}(\ell)$ from §5 as an alternative baseline leaves the depth ordering unchanged in all seven models (Appendix B.6).

4.1.2 Cross-Instance Transfer (C)

Cross-instance transfer tests whether an object-token state can causally alter the donor-label score in another image context (Figure 1b). Under the naming prompt, we transplant donor state $h^{\ell}_{A,\sigma_g(i)}$ into receiver position i and normalize the resulting score change by the donor–receiver gap $D_g = M_{y_A}(x_A^{\text{name}}) - M_{y_A}(x_B^{\text{name}})$, where M_{y_A} scores only the first label token:

$$C_{g,i}(\ell) = \frac{M_{y_A}\big(x_B^{\text{name}}[\ell;\, i \mapsto h^{\ell}_{A,\sigma_g(i)}]\big) - M_{y_A}\big(x_B^{\text{name}}\big)}{D_g + \varepsilon}, \qquad \varepsilon = 10^{-6}. \tag{3}$$

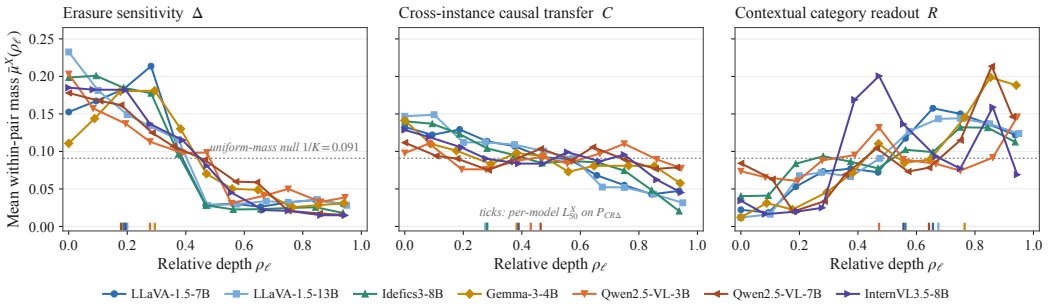


Figure 2: Cohort-mean layerwise mass profiles $\bar{\mu}^X(\rho_\ell)$ of the three influence axes on $\mathcal{P}_{CR\Delta}$ (Table 1). Dotted line: uniform-mass null $1/K \approx 0.091$ ($K = 11$). Bottom rug ticks: per-model L_{50}^X; values match Table 1. The plotted mean is a display device; L_{50}^X is a pair median (§4.2). In every model the three axes place their L_{50}^X in this order, though Δ and C ranges overlap across models.

We evaluate C only for pairs with $D_g > 0.5$, ensuring that the transplant has a positive score gap to close (Appendix B.4). Because state replacement perturbs the forward pass, raw C includes a generic replacement effect. We estimate this effect by repeating the transplant with a control donor A'_g whose label lies outside $\{y_A, y_B\}$. The receiver, intervention site, scored label y_A, and denominator D_g remain fixed between the target and control transfers. The controlled signal is $s_g^C(\ell) = \mathrm{med}_{i \in \mathcal{T}_g} C_{g,i}(\ell) - b_C(\ell)$, with $b_C(\ell)$ estimated from these control transfers (Appendix B.3).

4.1.3 Category Readout (R)

Category readout tests whether category information is recoverable from an object-token state (Figure 1c). Without intervening, we compare the same donor state $h^\ell_{A,\sigma_g(i)}$ used by C with text-token states from the phrase bank (Appendix A.5). The readout uses a projection π_k that removes at most one massive coordinate per text layer, a positive phrase set $\mathcal{S}^+(y_A)$, and $Q = 20$ token-count-matched negative sets $\{\mathcal{S}_q^-(y_A)\}_{q=1}^Q$. We define $\mathrm{sim}(h, \mathcal{S})$ as the maximum cosine similarity over phrase tokens and text layers because image and text depths are not aligned (Appendix B.5). The readout is the category-contrastive margin:

$$R_{g,i}(\ell) = \mathrm{sim}\big(h^\ell_{A,\sigma_g(i)},\ \mathcal{S}^+(y_A)\big) - \frac{1}{Q}\sum_{q=1}^{Q} \mathrm{sim}\big(h^\ell_{A,\sigma_g(i)},\ \mathcal{S}_q^-(y_A)\big). \tag{4}$$

We take $b_R \equiv 0$ as the principled null for the category contrast. Because the positive and negative sets have different category compositions, we also evaluate the slightly negative label-permutation null as an alternative baseline (Appendix B.5). Unlike C, R requires no pair-level prescreen and is measured for every pair with nonempty donor and receiver object-token sets. With $b_R \equiv 0$, the layerwise signal is $s_g^R(\ell) = \mathrm{med}_{i \in \mathcal{T}_g} R_{g,i}(\ell)$.

4.2 Locating Influence via Half-Mass Depth

Because the three layerwise signals s_g^C, s_g^R, and s_g^Δ arise from different measurements, we compare where their positive mass concentrates across depth rather than their magnitudes. For axis $X \in \{\Delta, C, R\}$, we define $W_g^X(\ell) = [s_g^X(\ell)]_+$ and $Z_g^X = \sum_{\ell \in \mathcal{G}} W_g^X(\ell)$. Among pairs eligible for axis X, those with $Z_g^X > 0$ form the axis population $\mathcal{P}_X$, and their profiles are normalized as $\mu_g^X(\ell) = W_g^X(\ell)/Z_g^X$ (Appendix C.1). The pair-level *half-mass depth* is the first grid depth at which half of this mass has accumulated:

$$h^X_{50,g} = \mathcal{H}\big[s_g^X\big] \equiv \min\Big\{\rho_\ell :\ \ell \in \mathcal{G},\ \sum_{\ell' \le \ell} \mu_g^X(\ell') \ge \tfrac{1}{2}\Big\}. \tag{5}$$

Applying the same operator $\mathcal{H}$ yields a common scale-free depth summary for the three axes, while the full profiles $\mu_g^X(\ell)$ characterize differences in shape (Figure 2).

Model	$\lvert\mathcal{P}_{CR\Delta}\rvert$	L_{50}^{Δ}	L_{50}^{C}	L_{50}^{R}
LLaVA-1.5-7B	815	0.188	0.281	0.656
LLaVA-1.5-13B	840	0.200	0.275	0.675
Idefics3-8B	843	0.188	0.281	0.562
Gemma-3-4B	702	0.294	0.382	0.765
Qwen2.5-VL-3B	776	0.278	0.431	0.472
Qwen2.5-VL-7B	768	0.179	0.464	0.643
InternVL3.5-8B	828	0.194	0.389	0.556

Table 1: Model-level half-mass depths on the common-support pair set $\mathcal{P}_{CR\Delta} = \mathcal{P}_C \cap \mathcal{P}_R \cap \mathcal{P}_\Delta$. Relative depth $\rho_\ell = \ell/L$ on the eleven-point grid (§3). $L_{50}^{\Delta} < L_{50}^{C} < L_{50}^{R}$ in every model.

At the model level, we report the pair median $L_{50}^{X}[\mathcal{P}] = \mathrm{median}_{g\in\mathcal{P}}\, h_{50,g}^{X}$. For $(X, Y) \in \{(\Delta, C), (C, R)\}$, we also report the median within-pair gap $G_{50}^{XY}[\mathcal{P}] = \mathrm{median}_{g\in\mathcal{P}}\big(h_{50,g}^{X} - h_{50,g}^{Y}\big)$ and the ordered fraction π^{XY} of pairs satisfying $h_{50,g}^{X} < h_{50,g}^{Y}$. Thus, $G_{50}^{XY} < 0$ places X earlier for the typical pair, while $\pi^{XY} > 0.5$ indicates that the ordering holds for a majority of pairs. All depth statistics are computed pair-first; 95% intervals for L_{50}^{X} and G_{50}^{XY} use the procedures in Appendix C.2.

4.3 A Consistent Depth Order Across Models

We examine the depth ordering through two adjacent contrasts before placing all three axes on common support. The C–R comparison evaluates the same donor states under two criteria—causal portability and category readability—whereas Δ–C compares the two intervention profiles.

An intervention–readout gap. The paired gap $G_{50}^{CR}[\mathcal{P}_{CR}]$ is negative in all seven models, ranging from -0.375 to -0.083, indicating that cross-instance transfer precedes category readout in depth for the typical pair. Consistently, 57–90% of pairs satisfy $h_{50,g}^{C} < h_{50,g}^{R}$ (Appendix D). The ordering is robust to the alternative readout baseline b_R^{perm}: paired G_{50}^{CR} remains negative and L_{50}^{R} is unchanged in all seven models (Appendix B.5). This intervention–readout gap qualitatively echoes Huang & Chang (2025), who find that object-token patching in a counting ViT is causally effective in middle layers, whereas count information is most decodable from object tokens in final layers, where patching is functionally inert. In our VLMs, after the projected visual tokens enter the LM backbone, we observe an analogous separation: cross-instance transfer concentrates in shallow-to-mid layers, while category readout concentrates at greater depth. Together, these findings are consistent with a module-relative organization of object-token influence rather than a single end-to-end depth trajectory.

Erasure sensitivity precedes cross-instance transfer. On the two-axis common support $\mathcal{P}_{\Delta C} = \mathcal{P}_\Delta \cap \mathcal{P}_C$, the paired gap $G_{50}^{\Delta C}$ is negative in all seven models, ranging from -0.194 to -0.088, and 59–75% of pairs satisfy $h_{50,g}^{\Delta} < h_{50,g}^{C}$ (Appendix F.3). Viewed as complementary interventions that erase or transplant object-token content, the two profiles differ in depth and follow a consistent $\Delta \prec C$ order. The ordering is preserved under the measured background baseline and on an LVIS-derived pool of 2,192 pairs spanning 385 donor categories (Appendices B.6 and F.3). Because R is not re-measured on LVIS, that extension supports only $\Delta \prec C$, not the full three-axis ordering.

A consistent three-axis ordering. We place all three axes on the common-support set $\mathcal{P}_{CR\Delta} = \mathcal{P}_C \cap \mathcal{P}_R \cap \mathcal{P}_\Delta$. Depending on the model, the common-support set retains 92.5–95.8% of the smallest axis-specific pair set (Appendix C.1).

$$\text{erasure sensitivity} \;\prec\; \text{cross-instance transfer} \;\prec\; \text{category readout}.$$

The corresponding values satisfy $L_{50}^{\Delta} < L_{50}^{C} < L_{50}^{R}$ in every model (Table 1). On the same common-support set, both paired gaps are negative in every model, while the ordered fractions range from 58.6% to 75.1% for $\Delta \prec C$ and from 56.7% to 91.0% for $C \prec R$ (Appendix D).

The full profiles reinforce this ordering: Δ-mass concentrates in the shallow band, C-mass in the shallow-to-mid band, and R-mass in the mid-to-deep band (Figure 2; per-model profiles in Figure 4).

The two intervention profiles also differ in shape: Δ-mass peaks at or before $\rho_\ell \approx 0.3$ and falls to a low plateau, whereas C-mass declines gradually and remains appreciable through the deepest sampled layers. Together, these results reveal a consistent depth organization in which the three signals concentrate in distinct but overlapping depth regions.

5 Set-Level Non-Additivity of Object-Token Influence

We directly test whether the effect of intervening on an object-token set is captured by the summed effects of intervening on its parts (§2.2.2). At each fixed layer, we compare the full-set effect with the summed effects of interventions on its two halves, using the corresponding comparison on a background set as a control.

5.1 Fixed-Layer Set Non-Additivity

For each pair, we split the receiver object-token set into two halves whose sizes differ by at most one, $\mathcal{T}_g = \mathcal{T}_g^{\mathrm{a}} \sqcup \mathcal{T}_g^{\mathrm{b}}$. For the background control, we construct an analogous set from annotation-free patches and partition it in the same way. We match its size to the object set whenever possible while keeping replacement sources and patched destinations disjoint; otherwise, we use the largest feasible set (Appendix E.1).

At each grid layer under the existence prompt, we evaluate six single-layer Δ-type intervention conditions (§4.1.1): the full object set and its two halves, and the full background set and its two halves. The replacement assigned to each position is held fixed across the relevant conditions. Thus, within the object or background comparison, the conditions differ only in which positions are patched (Appendix E.1).

For $X \in \{\mathrm{obj}, \mathrm{bg}\}$ and $\alpha \in \{\mathrm{full}, \mathrm{a}, \mathrm{b}\}$, let $\delta^{\alpha}_{X,g}(\ell)$ denote the clean-to-intervened reduction in the yes/no margin under the corresponding condition. A positive value indicates that the intervention moves the answer toward "no". The full-object condition is identical to the erasure intervention in §4.1.1, so $\delta^{\mathrm{full}}_{\mathrm{obj},g}(\ell) = -\Delta_g(\ell) = s^{\Delta}_g(\ell)$.

We define the within-set non-additivity residual and its background-controlled contrast as

$$\begin{aligned} I_g^X(\ell) &= \delta^{\mathrm{full}}_{X,g}(\ell) - \delta^{\mathrm{a}}_{X,g}(\ell) - \delta^{\mathrm{b}}_{X,g}(\ell), \quad X \in \{\mathrm{obj}, \mathrm{bg}\}, \\ J_g(\ell) &= I_g^{\mathrm{obj}}(\ell) - I_g^{\mathrm{bg}}(\ell). \end{aligned} \tag{6}$$

$I_g^X(\ell) > 0$ indicates that the full-set effect exceeds the sum of the two half-set effects. Accordingly, $J_g(\ell) > 0$ indicates excess non-additivity for the object set relative to the background control. Because all six conditions intervene at the same layer ℓ, $J_g(\ell)$ is a fixed-layer measure of set-effect non-additivity.

The population $\mathcal{P}_J$ contains pairs with all six intervention outcomes available at every grid layer. At each layer, we report the pair mean $\bar{J}(\ell) = |\mathcal{P}_J|^{-1} \sum_{g \in \mathcal{P}_J} J_g(\ell)$ with 95% two-way cluster-bootstrap intervals over donor and receiver images (Appendix C.2).

5.2 Set-Level Non-Additivity Concentrates at Shallow Depths

Figure 3 shows $\bar{J}(\ell)$ for each model. In every model, at least one of the first four sampled depths ($\rho_\ell \le 0.30$) has a positive $\bar{J}(\ell)$ whose interval excludes zero. We summarize this excess over the shallow region as

$$\bar{J}^{\mathrm{band}} = \frac{1}{|\mathcal{P}_J|} \sum_{g \in \mathcal{P}_J} \frac{1}{|\mathcal{G}_{\le 0.3}|} \sum_{\ell \in \mathcal{G}_{\le 0.3}} J_g(\ell), \tag{7}$$

where $\mathcal{G}_{\le 0.3} = \{\ell \in \mathcal{G} : \rho_\ell \le 0.30\}$. The shallow-region mean is positive with an interval excluding zero in all seven models, ranging from $+0.185$ for Idefics3-8B to $+0.914$ for Gemma-3-4B (Appendix E.3). Nearby cutoffs yield the same conclusion (Appendix E.4).

Descriptively, the peak occurs at or before $\rho_\ell = 0.19$ in every model and at the first sampled layer in Qwen2.5-VL-3B and InternVL3.5-8B. At the deepest sampled layers, $\bar{J}(\ell)$ is indistinguishable from zero in every model. Because the intervention effects are scored at the output, $J_g(\ell)$ is an operational

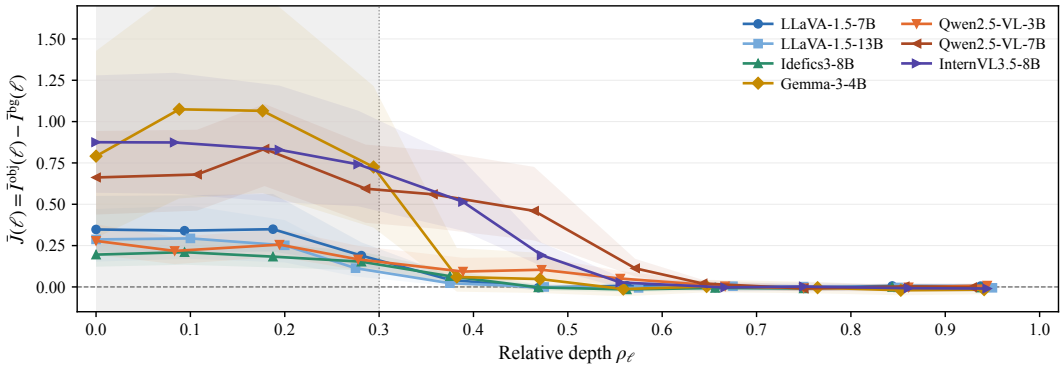


Figure 3: Per-layer background-controlled set non-additivity $\bar{J}(\ell) = \bar{I}^{\mathrm{obj}}(\ell) - \bar{I}^{\mathrm{bg}}(\ell)$ across seven VLMs. Shading marks the 95% two-way cluster-bootstrap interval on donor and receiver image; the grey band on the left marks the shallow region $\rho_\ell \leq 0.30$ used for the primary band summary $\bar{J}^{\mathrm{band}}$. Every model peaks in the shallow band and returns to values indistinguishable from zero at deeper layers.

measure of set-effect non-additivity. It may reflect non-additivity present at the intervention site or generated during downstream propagation, including through the nonlinear mapping from hidden states to the final answer score. Shallow interventions permit more downstream computation, so the observed concentration does not identify a specific token–token interaction mechanism at layer ℓ. Separating intervention-site contributions from downstream effects would require varying the intervention and readout depths independently. Likewise, the decay of $J_g(\ell)$ to zero at deeper layers does not by itself establish additivity at those layers.

Background non-inertness and the choice of contrast. The background intervention is not universally inert. In several models, $\delta^{\mathrm{full}}_{\mathrm{bg},g}(\ell)$ is non-zero, motivating the use of $J_g(\ell)$ rather than $I^{\mathrm{obj}}_g(\ell)$ alone. Subtracting $I^{\mathrm{bg}}_g(\ell)$ controls for non-additivity that an analogous intervention on background patches would already produce. The shallow-region sign is preserved across partition seeds, alternative aggregators (median and 10% trimmed mean), and the restriction to strictly size-matched pairs (Appendix E.4).

5.3 Relation to the Depth Profiles

Together with §4, these results distinguish where object-token influence is expressed from where its effects combine non-additively. The three influence axes span progressively deeper regions: Δ concentrates in shallow layers, C in shallow-to-mid layers, and R in mid-to-deep layers. By contrast, excess set-level non-additivity is more tightly concentrated in the shallow region $\rho_\ell \leq 0.30$. These observations are not redundant. Although shallow Δ-mass identifies depths at which erasure effects are concentrated, it does not determine whether the full-set effect exceeds the sum of the two half-set effects. Set-level non-additivity is therefore a distinct property of object-token influence that is more tightly localized to shallow depths.

6 Related Work

Prior work has studied erasure sensitivity, cross-instance transfer, and category readout, but has not jointly compared all three under a common depth statistic or examined how object-token effects compose over a set.

Intervention precedents. Neo et al. (2025) show that removing object-specific visual tokens substantially degrades object identification in LLaVA, while Basu et al. (2024) use multimodal causal tracing to localize early causal blocks and visual-token transfer on factual VQA. Kaduri et al. (2025) provide a complementary layerwise attention analysis. Unlike these approaches, our erasure-sensitivity profile overwrites a spatially defined object-token set independently at each fixed back-

bone layer (§4.1.1). Activation patching and related causal interventions were developed in text-only settings (Meng et al., 2022; Geva et al., 2023; Wang et al., 2023; Conmy et al., 2023) and extended to VLMs by Palit et al. (2023). Our cross-instance transfer axis (§4.1.2) instead transplants a hidden state from a *different* image's object token into the receiver. Related work characterizes staged image-to-text information flow (Zhang et al., 2025) and distinguishes direct image-token from text-mediated pathways using causal patching (Salazar et al., 2026). These studies analyze routing between token types; we compare three object-token influence profiles under a shared depth summary.

Readout precedents. The logit lens (nostalgebraist, 2020) and tuned lens (Belrose et al., 2023) have been applied to visual tokens (Neo et al., 2025), while multimodal-neuron work makes internal visual concepts interpretable (Goh et al., 2021; Schwettmann et al., 2023). The closest precedents to our category-contrastive readout are text-based decomposition of CLIP representations (Gandelsman et al., 2024), patch–segmentation-label contrasts (Wu et al., 2025), and text-bank similarity against Visual Genome descriptions (Krojer et al., 2026; Krishna et al., 2017).

Causality versus decodability. On a synthetic counting task, Huang & Chang (2025) find middle-layer object tokens to be causally influential but weakly decodable, while final-layer tokens are decodable but functionally inert. Our $C \prec R$ result is qualitatively aligned, but places the cross-instance transfer profile earlier, with $L_{50}^{C} \in [0.28, 0.46]$ across our seven VLMs (Table 1). Concurrent work in visual–linguistic conflict settings combines layerwise readout with full-sequence activation patching, finding early decodability of visual attributes and a causal effect carried primarily by image tokens (Nooralahzadeh et al., 2026). Its attribute-level readout differs from our category-contrastive text-bank construction, so the absolute depths are not directly comparable. Causal localization does not by itself identify an effective editing site (Hase et al., 2023), and interventions may activate fallback routes absent from the dominant clean computation (Salazar et al., 2026). Our Δ, C, and R ordering therefore characterizes the concentration of three operational profiles, rather than a unique clean-computation pathway or an optimal editing depth.

Token redundancy and set-level non-additivity. Visual-token pruning studies find substantial redundancy with limited accuracy loss (Chen et al., 2024; Lin et al., 2024; Xing et al., 2024; Zhang et al., 2024b; Yang et al., 2025; Zhang et al., 2024a), but do not directly test whether a whole-set intervention is predicted by interventions on its parts. Our fixed-layer analysis (§5) measures this decomposition residual and finds shallow-depth excess non-additivity relative to a matched background control.

7 Conclusion

Object-token influence in adapter-style vision–language models exhibits two consistent forms of depth organization. First, three operationally distinct measurements—erasure sensitivity, cross-instance transfer, and category readout—follow the same order, $L_{50}^{\Delta} < L_{50}^{C} < L_{50}^{R}$, under a common half-mass depth operator across seven VLMs. Second, whole-set object-token interventions exhibit supra-additive effects relative to a matched background contrast, with this excess concentrated at shallow depths. Both results are based on pair-first statistics and hold in every model we tested; in particular, the shallow-band non-additivity contrast clears zero in $7/7$ models. Placing the three axes on a shared depth coordinate makes their consistent ordering visible, while the set-level non-additivity analysis shows that the effect of intervening on the whole object-token set is not generally captured by summing the effects of interventions on its parts.

Limitations and future work. Our scope is limited to adapter-style VLMs up to 13B, object-existence and naming prompts, and images from COCO. Extension to larger models, cross-attention architectures, and multi-step reasoning tasks is left to future work. Moreover, the non-additivity contrast measures output-level non-additivity and does not identify whether that non-additivity arises at the intervention site, during downstream propagation, or from the scoring function. Two practical directions follow from our findings: the Δ–C–R ordering identifies candidate depths for targeted editing or steering—not a prescription, in view of the localization–editing gap noted in §6—while the shallow set-level non-additivity motivates visual-token pruning methods that account for joint rather than purely tokenwise effects.

AI USE STATEMENT

Generative AI tools were used for feedback on the methodology and the interpretation of results, implementation of parts of the pipeline, auditing the manuscript against the code, literature search and reference formatting, translation, and drafting and editing of selected passages. They were not used for research hypotheses, conceptual frameworks, mathematical claims or proofs, synthetic data, dataset cleaning, or qualitative analysis. The authors verified all AI-assisted code, citations, and text and take full responsibility for the content.

REFERENCES

Shuai Bai et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding Information Storage and Transfer in Multi-modal Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112*, 2023.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuang Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An Image Is Worth 1/2 Tokens After Layer 2: Plug-and-Play Acceleration for VLLM Inference. In *European Conference on Computer Vision (ECCV)*, 2024.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP's Image Representation via Text-Based Decomposition. In *International Conference on Learning Representations (ICLR)*, 2024.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030.

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lianghuan Huang and Yingshan Chang. Causality $\neq$ Decodability, and Vice Versa: Lessons from Interpreting Counting ViTs. In *NeurIPS Workshop on Cognitive Interpretability (CogInterp)*, 2025.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and Editing Vision-Language Representations to Mitigate Hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the Image? A Deep-Dive into the Vision of Vision Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. LatentLens: Revealing Highly Interpretable Visual Tokens in LLMs. *arXiv preprint arXiv:2602.00462*, 2026.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.

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 (ECCV)*, 2014.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting Multimodal Large Language Models with Visual Tokens Withdrawal for Rapid Inference. *arXiv preprint arXiv:2405.05803*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards Interpreting Visual Information Processing in Vision-Language Models. In *International Conference on Learning Representations (ICLR)*, 2025.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration Failure, Not Perceptual Blindness: How Vision-Language Models Resolve Visual-Linguistic Conflicts. *arXiv preprint arXiv:2604.09364*, 2026.

nostalgebraist. Interpreting GPT: The Logit Lens, 2020. LessWrong.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023.

Israfel Salazar, Stella Frank, Dan Oneata, Desmond Elliott, and Constanza Fierro. Pathways of Visual Information Flow in Vision-Language Models. *arXiv preprint arXiv:2607.03358*, 2026.

Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. Multimodal Neurons in Pretrained Text-Only Transformers. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023.

Gemma Team. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025.

Martina G. Vilas, Timothy Schaumloeffel, and Gemma Roig. Analyzing Vision Transformers for Image Classification in Class Embedding Space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small. In *International Conference on Learning Representations (ICLR)*, 2023.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The Semantic Hub Hypothesis: Language Models Share Semantic Representations Across Languages and Modalities. In *International Conference on Learning Representations (ICLR)*, 2025.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. PyramidDrop: Accelerating Your Large Vision-Language Models via Pyramid Visual Redundancy Reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is Better but Not Necessary in Vision Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaofeng Zhang, Yihao Quan, Chaochen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From Redundancy to Relevance: Information Flow in LVLMs Across Reasoning Tasks. *arXiv preprint arXiv:2406.06579*, 2024a.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, et al. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. *arXiv preprint arXiv:2410.04417*, 2024b.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal Information Flow in Multimodal Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

A Setup Details

A.1 Per-model Layer Grids

The number of language-model layers L varies across the seven models, so every depth statistic is placed on the relative-depth axis $\rho_\ell = \ell/L \in [0, 1)$. The denominator is L and not ℓ_{11}, so ρ compares across models even though each grid stops short of the last block. Table 2 gives the eleven-layer observation grid $\mathcal{G} = \{\ell_1 < \cdots < \ell_{11}\}$ and the residual-stream dimension d per model.

Model	L	d	grid layers $\ell_1, \ldots, \ell_{11}$
LLaVA-1.5-7B	32	4,096	0, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30
LLaVA-1.5-13B	40	5,120	0, 4, 8, 11, 15, 19, 23, 27, 30, 34, 38
Idefics3-8B	32	4,096	0, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30
Gemma-3-4B	34	2,560	0, 3, 6, 10, 13, 16, 19, 22, 26, 29, 32
Qwen2.5-VL-3B	36	2,048	0, 3, 7, 10, 14, 17, 20, 24, 27, 31, 34
Qwen2.5-VL-7B	28	3,584	0, 3, 5, 8, 10, 13, 16, 18, 21, 24, 26
InternVL3.5-8B	36	4,096	0, 3, 7, 10, 14, 17, 20, 24, 27, 31, 34

Table 2: Backbone depth L, residual dimension d, and the eleven grid layers per model. $\ell_1 = 0$ is the embedded input in every model. Relative depth uses $\rho_\ell = \ell/L$, so the deepest grid layer ℓ_{11} has relative depth $\rho_{\ell_{11}} \in [0.929, 0.950]$ across models.

A.2 Intervention Notation and Determinism

For an input x, a target layer $\ell \in \mathcal{G}$, a set of visual-token positions S, and an assignment $v : S \to \mathbb{R}^d$, the notation $x[\ell; S \mapsto v]$ of Section 3.1 denotes the forward pass of x under a pre-hook on block ℓ that replaces the residual state at every position $i \in S$ by $v(i)$ before the block runs. The index ℓ refers to the state entering block ℓ, with $\ell = 0$ the embedded input.

Two properties are asserted at run time. The hook is *local*: every other residual entry at layer ℓ, and every residual entry at layers $< \ell$, is bit-identical to the clean forward. Locality is verified once per model by the identity arm (Appendix B.2), which asserts $m(x[\ell; \mathcal{T}_g \mapsto h^{\ell}_{B,\mathrm{id}}]) = m(x)$ bit-exactly at every grid layer. Downstream behavior is *deterministic*: under the seeds listed below, the same $(\text{model}, \ell, S, v)$ reproduces the residual state at every layer $\ell' > \ell$.

Precision and hardware. All forward passes and the text-bank encoding (§A.5) use bfloat16 on a single NVIDIA A100 40GB. We use bfloat16 rather than fp16 because Gemma-3-4B's residual stream regularly exceeds the fp16 range ($65{,}504$), which would cause clipping.

Random-number sources. Every random draw is seeded once at the top of the pipeline: (a) the image-split seed 7777 (Section 3); (b) the 60-token cap draw for $|\mathcal{T}_g| > 60$ (Section 3.2.1); (c) the balance seed for the readout's negative draws (Appendix B.5); (d) the balanced partition of the object-token set in Section 5.1; (e) the bootstrap random states (Appendix C.2); (f) the per-stratum shuffle of the text bank, seed 42 (Appendix A.5). The background assignments draw no random numbers. β_g of Section 4.1.1 enumerates $\mathcal{T}_g$ and $\mathcal{U}_B$ in raster order and sends the k-th object token to the $(k \bmod |\mathcal{U}_B|)$-th background patch, which is deterministic given the pair and injective whenever $|\mathcal{T}_g| \le |\mathcal{U}_B|$. β^{bg}_g of Section 5.1 likewise pairs the first $n = \min(|\mathcal{T}_g|, \lfloor |\mathcal{U}_B|/2 \rfloor)$ background patches in raster order with the last n, so its domain and image are disjoint by construction (Appendix E.1). Rerunning the pipeline from these seeds reproduces every reported number.

A.3 Dataset Split and Partitions

The source pool is

$$\text{source pool} \;=\; \big(\text{COCO val2017} \cap \text{LVIS v1 val}\big) \;\setminus\; \text{an earlier 200-image verification split},$$

yielding $4{,}609$ source images, which a single pass with seed 7777 partitions into five disjoint subsets (Table 3). The receiver and donor subsets supply every pair-based measurement of C, R, and Δ; the b_C-calibration subset supplies control donors for the C baseline (Appendix B.3).

Subset	$\lvert\cdot\rvert$	Use
receiver	200	receiver images B of pairs $g = (A, B)$
donor	1,500	donor images A of pairs $g = (A, B)$
b_C-calibration	300	control donors A'_g for the b_C baseline
tuning	300	hyper-parameter selection

Table 3: Four of the five disjoint subsets the seed-7777 pass draws from the $4{,}609$-image source pool; the fifth, 465 images, is held out and supplies no measurement reported here. The tuning partition was used only to select the object-token coverage threshold θ, the massive-coordinate threshold τ, the C prescreen threshold, and the shallow-band cutoff ρ_0; the other three supply every reported measurement.

A.4 Prompts and Answer Variants

Table 4 gives the two prompt templates used by the scorer M_y and margin m defined in the main text. Because y_B is the annotated receiver category by construction (Section 3.2.1), the ground-truth answer to the existence prompt is always `yes` in $\mathcal{P}_0$; an erasure effect therefore shifts the answer from `yes` toward `no`.

Prompt	Template	Used by
Naming	`USER: <image>\n What is the object in this image? Answer with one word.\n ASSISTANT:_`	C, b_C, and (through the donor forward) R
Existence (POPE)	`USER: <image>\n Is there {a|an} {`y_B`} in the image?\n ASSISTANT:_`	Δ and every arm of §5

Table 4: Prompt templates shown in a common chat format; each model's native chat template is applied. Underscores mark the scored continuation: the first token of y_A for the naming prompt and a yes/no variant for the existence prompt. In the existence prompt, a/an is chosen to match the surface form of y_B.

Yes/no variants. The margin $m(x) = \max_{y \in \mathcal{Y}_{\text{yes}}} M_y(x) - \max_{y \in \mathcal{Y}_{\text{no}}} M_y(x)$ takes the maximum over four surface variants each:

$$\mathcal{Y}_{\text{yes}} = \{\texttt{Yes},\ \texttt{Yes},\texttt{yes},\ \texttt{yes}\},$$
$$\mathcal{Y}_{\text{no}} = \{\texttt{No},\ \texttt{No},\texttt{no},\ \texttt{no}\}.$$

The leading-space variants cover cases where the ASSISTANT turn begins the completion at a whitespace token. A first-token rule cannot separate leading-space and non-space forms that share a first BPE token, so the margin scores every variant on the whole answer string. In five of seven models (Idefics3, Gemma-3, Qwen2.5-VL-{3B,7B}, InternVL3.5) the argmax yes/no variants remain single tokens throughout; in the two LLaVA-1.5 models the winning no-variant is multi-token, which the mean formulation of M_y accommodates (Appendix B.1).

A.5 Text Bank

The text bank used for category readout (R) contains $20{,}000$ Visual Genome region-description phrases (Krishna et al., 2017), matched against per-category regular expressions for the 80 COCO categories and encoded once per model at every grid layer.

Phrase mining. Phrases shorter than three characters are dropped. A per-category regular expression that accepts COCO plurals and common adjectival forms buckets the remainder into 80 category strata, and four non-category strata (color, material, spatial, general) are kept alongside them. Each stratum is shuffled with a fixed seed, and a deterministic round-robin interleave over the 84 strata yields the final bank, with 236–239 phrases per stratum. Category membership is recomputed from the phrase text rather than taken from the stratum, so a phrase may match several categories or none:

24.2% of phrases match two or more categories, 3.1% match none and enter only negative sets, and every COCO category has a nonempty positive set (Appendix B.5).

Per-model encoding. Each phrase p_j is passed through the model as a text-only input without a system prompt, and its per-token residual state at each of the eleven grid layers is stored in bfloat16. The stored tensor has shape [total tokens, d] with d as in Table 2; the token count reflects tokenizer-specific splits and ranges from $118{,}233$ to $159{,}448$ per layer across models.

A.6 Pair Construction and Token Correspondence

The shared pool contains $|\mathcal{P}_0| = 1{,}000$ pairs.

Candidate filter. A pair $g = (A, B)$ is a donor image A with an annotated instance a_A of category y_A and a receiver image B with an instance a_B of category $y_B \neq y_A$. Let $\mathcal{O}_A, \mathcal{O}_B$ denote their object-token sets (Appendix A.7), interpreted as subsets of the model's grid of N final visual tokens; N is model-specific (576 for LLaVA-1.5, 169 for Idefics3, 256 otherwise) and fixed across images, and the IoU is computed within each model's own grid. A candidate pair enters the shared pool only if, in every one of the seven models, both sets are non-empty and $\mathrm{IoU}(\mathcal{O}_A, \mathcal{O}_B) \geq 0.4$.

Deterministic receiver round-robin. Receiver images are enumerated in COCO image-id order, and for each receiver annotation all admissible donors are enumerated (donors by image id, annotations within an image by annotation id). A single round-robin pass draws one pair per receiver annotation in enumeration order, then repeats with the next admissible donor per receiver annotation, until the pool contains exactly $1{,}000$ pairs. This construction is deterministic given the source-pool split and needs no additional seed.

Token correspondence. Within a pair, receiver and donor object tokens are enumerated in raster order and matched rank by rank up to $n_g = \min(|\mathcal{O}_A|, |\mathcal{O}_B|)$: the r-th receiver object token is paired with the r-th donor object token. If $n_g > 60$ a seeded uniform subsample retains exactly 60 ranks (seed $= 42$); otherwise all n_g ranks are kept. This produces $\mathcal{T}_g$, the receiver token set on which all three axes are keyed, and the donor correspondence $\sigma_g : \mathcal{T}_g \to \mathcal{O}_A$. The cap was chosen from the observed distribution of $|\mathcal{O}_A|, |\mathcal{O}_B|$ in the tuning partition, where $\Pr\{n_g \leq 60\}$ exceeded 0.98; the tail of the longer mask beyond rank n_g enters no measurement, is small (median 0 ranks, 95th percentile below 10), and reintroducing it by extending the shorter side with rank repetition leaves every L_{50}^{X} unchanged to the printed digit.

A.7 Object-Token Definition and Threshold Sensitivity

For each model we reconstruct the exact image preprocessing and project every COCO instance mask onto the model's grid of N final visual tokens. LLaVA-1.5 resizes the shortest edge and center-crops; Qwen2.5-VL is center-cropped to a fixed 448×448 input (`min_pixels` = `max_pixels`); Idefics3, Gemma-3, and InternVL3.5 resize to a fixed square with image splitting and tiling disabled (`do_image_splitting` = `False`, `crop_to_patches` = `False`). After each model's token merging or pooling, a final token cell spans 14 px (LLaVA-1.5), 28 px (Idefics3, Qwen2.5-VL, InternVL3.5), or 56 px (Gemma-3). For token $t \in \{1, \ldots, N\}$ and annotation a, let $\mathrm{cov}(t, a) \in [0, 1]$ be the fraction of the area of t's cell covered by the mask of a. With $\theta = 0.3$ (calibrated on the tuning partition) applied once per final cell, the object-token set of annotation a is

$$\mathcal{O}(a) = \big\{t : \ \mathrm{cov}(t, a) \geq \theta \text{ and } a = \arg\max_{a'} \mathrm{cov}(t, a')\big\}, \tag{8}$$

and the background set of the image is $\mathcal{U} = \big\{t : \ \max_{a'} \mathrm{cov}(t, a') < \theta\big\}$.

Two consequences follow. Object-token sets of different annotations are disjoint by the argmax rule, so a token belongs to at most one object. And $\mathcal{U}$ contains only patches that no annotated instance covers to θ: patches covered above θ by an annotation other than the queried one belong to neither $\mathcal{O}(a)$ nor $\mathcal{U}$ and are never intervened on, so the background destination set $\mathcal{T}_g^{\mathrm{bg}} \subseteq \mathcal{U}_B$ of Section 5 never draws replacement material from another annotated object.

Sweeping $\theta \in \{0.20, 0.25, 0.30, 0.35, 0.40\}$ on the tuning partition moves the per-model L_{50}^{C} and L_{50}^{R} by at most one grid step, and at the primary $\theta = 0.30$ both match the pair-first estimates of the main text.

B Scoring and Axis Details

B.1 Answer Margin and the Scorer M_y

For a target string $y = (y_0, \ldots, y_{m-1})$ (BPE-tokenized) and an input x whose prompt occupies n positions, the output score of Eq. (1) is the mean teacher-forced log-probability of the target:

$$M_y(x) = \frac{1}{m} \sum_{k=0}^{m-1} \log \mathrm{softmax}\big(\mathrm{logits}_{n-1+k}(x \oplus y)\big)[y_k], \tag{9}$$

where $x \oplus y$ appends the target tokens to the prompt. The completion is scored with the model in inference mode; appending tokens does not change earlier prompt positions or their hidden states, so every hook and every measured hidden state remains identical to the clean forward.

Why the mean for the yes/no margin. Under the first-token log-probability $\log P(y_0 \mid x)$ the argmax yes/no variants agree with those of Eq. (9) in five of seven models. The two LLaVA-1.5 models are the exception: their winning no-variant is multi-token, and its mean log-probability is stronger than any single-token no (Appendix A.4). The mean formulation handles this case uniformly across models. The naming score of C does not take a maximum over surface variants and scores the first token of y_A only, i.e. Eq. (9) with $m = 1$ (§4.1.2).

Variant agreement across arms. On every (g, ℓ) row, in every model, the winning yes-variant and the winning no-variant are the same across the clean forward and all six arms of Section 5, so m is a difference between two fixed variants and neither I^{obj} nor I^{bg} in Eq. (6) can be produced by variant switching.

B.2 Erasure Sensitivity (Δ): Instrumentation Floor

The instrumentation floor $b_\Delta \equiv 0$ of Section 4.1.1 is a measured property of the patching apparatus, not an assumption. The identity arm rewrites every $i \in \mathcal{T}_g$ with its own clean state, $h^{\ell}_{B,\mathrm{id}} : i \mapsto h^{\ell}_{B,i}$: the pre-hook writes back the state that was already there, so every subsequent block receives the clean tensor bit-exactly. We assert $m(x_B^{\mathrm{exist}}[\ell; \mathcal{T}_g \mapsto h^{\ell}_{B,\mathrm{id}}]) = m(x_B^{\mathrm{exist}})$ at run time for every pair and every grid layer of every model. The observed number of failures is 0 across 7 models $\times$ 11 layers $\times$ 1,000 pairs $=$ 77,000 checks, so $\Delta_g^{\mathrm{identity}}(\ell) = 0$ bit-exactly throughout.

B.3 Cross-Instance Transfer (C): Baseline b_C Construction

Raw C conflates category-specific transfer with the generic disturbance of overwriting any hidden state. The baseline $b_C(\ell)$ removes that disturbance by measuring the same intervention with a category-irrelevant donor.

Third-donor mining. For each pair $g = (A, B)$ we mine a control donor A'_g from the b_C-calibration partition of 300 images under two constraints: (i) $y_{A'_g} \notin \{y_A, y_B\}$; (ii) the pair-eligibility filter of Appendix A.6 applied to (A'_g, B), with $|\mathcal{O}_{A'_g}| \geq n_g$ so that the correspondence $\sigma'_g : \mathcal{T}_g \to \mathcal{O}_{A'_g}$ can be built. Ties are resolved by b_C-partition image-id order.

Control transfer. The control transfer uses the same first-label-token score M_{y_A} and denominator D_g as the target transfer, scoring y_A from the real donor rather than $y_{A'_g}$; thus, the only change is whether the injected content matches the label being scored:

$$C^{\circ}_{g,i}(\ell) = \frac{M_{y_A}\big(x_B^{\mathrm{name}}[\ell;\ i \mapsto h^{\ell}_{A'_g, \sigma'_g(i)}]\big) - M_{y_A}\big(x_B^{\mathrm{name}}\big)}{D_g + \varepsilon}. \tag{10}$$

$b_C(\ell)$ is the pair median of the token medians of $C^{\circ}_{g,i}(\ell)$ across the b_C-calibration pair set $\mathcal{P}_b \subseteq \mathcal{P}_C$:

$$b_C(\ell) = \operatorname*{med}_{g \in \mathcal{P}_b} \operatorname*{med}_{i \in \mathcal{T}_g} C^{\circ}_{g,i}(\ell), \qquad \ell \in \mathcal{G}. \tag{11}$$

The double-median aggregation matches how $s^C_g(\ell)$ is formed in Section 4.1.2, so baseline and target signal sit on the same aggregation level. Per-model $|\mathcal{P}_b|$ ranges from 254 to 285 pairs.

B.4 Cross-Instance Transfer (C): Prescreen Threshold Sensitivity

The transfer denominator $D_g = M_{y_A}(x^{\text{name}}_A) - M_{y_A}(x^{\text{name}}_B)$ uses the first-label-token score M_{y_A} and must be positive and large enough for the ratio $C_{g,i}(\ell)$ of Eq. (3) to have a defined and stable direction. The prescreen is on the signed gap, $D_g > 0.5$, not on $|D_g|$: a signed threshold restricts the measurement to pairs where the donor category has room to gain in the receiver, which is the regime in which successful transfer has a defined positive reference direction. Pairs with $D_g < -0.5$ would give C a negative reference direction and are excluded by the same signed rule; admitting them under an $|D_g|$ rule would require re-measuring 3–17% additional pairs per model (Gemma-3-4B the outlier at 17.3%).

Sweeping the signed threshold through $\theta_D \in \{0.3, 0.5, 0.7, 1.0\}$ leaves L^C_{50} unchanged in all seven models, at $|\mathcal{P}_C|$ within 2% of the primary $\theta_D = 0.5$ counts (Table 7). Pair-first L^C_{50} is lattice-valued on the eleven-point grid, and the pairs added or removed by threshold changes are a small minority whose half-mass depths resemble the rest.

B.5 Category Readout (R): Projection, Text-Bank Contrast, and Label Permutation

Massive-coordinate projection π_k. For each text layer $k \in \mathcal{G}$ and each coordinate $c \in \{1, \dots, d\}$ of the model's residual stream, let

$$\phi_{k,c} = \operatorname*{med}_{(j,t)} \frac{\left(u^k_{j,t}\right)^2_c}{\left\|u^k_{j,t}\right\|^2_2}, \tag{12}$$

the median over every text-bank token state at layer k. The set of *massive coordinates* at layer k is $\mathcal{D}_k = \{c : \phi_{k,c} > \tau\}$ with $\tau = 0.5$, and the projection π_k zeros exactly these coordinates before every cosine in Eq. (4). Because the shares $\phi_{k,\cdot}$ of one state sum to one, $|\mathcal{D}_k| \leq 1$: at most a single coordinate is zeroed. The projection fires at ten of the eleven grid layers in Gemma-3-4B ($\max_{k,c} \phi_{k,c} = 0.99$) and at the single deepest grid layer in Qwen2.5-VL-7B (0.55); in the other five models $\max_{k,c} \phi_{k,c} \leq 0.39$ and π_k is the identity everywhere. It remains part of the protocol so that a re-run under any regime that induces a new massive coordinate applies the correction automatically.

Positive and negative phrase sets. The positive token set for a target category y is $\mathcal{S}^+(y) = \{(j,t) : p_j \text{ matches the expression of } y\}$, every token of every text-bank phrase matching y (Appendix A.5); these sets may overlap across categories. For $q = 1, \dots, Q$ with $Q = 20$, the q-th negative set $\mathcal{S}^-_q(y)$ is a seeded uniform draw of exactly $|\mathcal{S}^+(y)|$ token states from the tokens of phrases that do not match y: the draw balances at the *token* level, not the phrase level, so phrase-length variation cannot move the token count. The balance seed depends on the category alone, so the same $\mathcal{S}^-_q(y)$ serves every measurement of category y at every layer. $Q = 20$ was chosen from a stability analysis on the tuning partition: sweeping $Q \in \{5, 10, 20, 40\}$ shows the standard error of $R_{g,i}(\ell)$ decrease as $1/\sqrt{Q}$ up to $Q = 20$ and plateau thereafter.

Free-layer maximum sim**.** The similarity of an image state h to a token set $\mathcal{S}$ is

$$\operatorname{sim}(h, \mathcal{S}) = \max_{(j,t) \in \mathcal{S}} \max_{k \in \mathcal{G}} \cos\left(\pi_k h,\ \pi_k u^k_{j,t}\right), \tag{13}$$

the outer maximum running over the eleven text-bank layers. Image depth and text depth are not on a common coordinate: pinning $k = \ell_{\text{img}}$ adds a model-dependent shift to L^R_{50} that is uncorrelated with the depth ordering of interest, so the free-layer maximum is used. Section 4's comparison does not depend on this choice.

	L_{50}^R		G_{50}^{CR}			
Model	$b_R \equiv 0$	b_R^{perm}	$b_R \equiv 0$	b_R^{perm}	95% CI	π^{CR}
LLaVA-1.5-7B	0.656	0.656	-0.375	-0.375	$[-0.375, -0.281]$	89.5%
LLaVA-1.5-13B	0.675	0.675	-0.375	-0.400	$[-0.400, -0.375]$	91.6%
Idefics3-8B	0.562	0.562	-0.281	-0.281	$[-0.281, -0.281]$	85.0%
Gemma-3-4B	0.765	0.765	-0.353	-0.382	$[-0.382, -0.382]$	86.2%
Qwen2.5-VL-3B	0.472	0.472	-0.083	-0.083	$[-0.083, -0.083]$	53.9%
Qwen2.5-VL-7B	0.643	0.643	-0.214	-0.214	$[-0.286, -0.196]$	76.7%
InternVL3.5-8B	0.556	0.556	-0.194	-0.194	$[-0.278, -0.194]$	76.9%

Table 5: Readout depth and the C–R paired gap under the principled baseline $b_R \equiv 0$ and under the label-permutation baseline b_R^{perm}. L_{50}^R is unchanged in all seven models. G_{50}^{CR} is computed on $\mathcal{P}_{CR}$ under each baseline; the 95% interval and π^{CR} refer to b_R^{perm} (pair-level percentile bootstrap, Appendix C.2). $G_{50}^{CR} < 0$ with the interval upper bound below zero in every model under both baselines.

Label-permutation baseline b_R^{perm}. The principled baseline is $b_R \equiv 0$: R is a category contrast, so its expected value is zero when no category information is present. An empirical baseline $b_R^{\text{perm}}(\ell)$ is computed by permuting the mapping from object token to category label and recomputing R under the same $\pi_k, \mathcal{S}^+, \mathcal{S}_q^-$, and pair-token aggregation; only the category-to-token mapping is permuted, and every other element of Eq. (4) is held fixed. Across all models and grid layers $b_R^{\text{perm}}(\ell)$ lies in $[-0.035, 0.0001]$, with the larger magnitudes tending to occur at greater depth. Substituting it for $b_R \equiv 0$ and reapplying the pair-first half-mass estimator gives Table 5: L_{50}^R is unchanged in all seven models, and the paired G_{50}^{CR} stays negative with its interval upper bound below zero in every model. The principled $b_R \equiv 0$ is therefore safe for the reported depth ordering.

B.6 Erasure Sensitivity (Δ): Controlled Signal Sensitivity

The identity arm (Appendix B.2) certifies that the patching apparatus contributes nothing bit-exactly, but it does not remove the generic component that overwriting any hidden state induces on the forward pass—the Δ-side analogue of the replacement shock that motivates b_C for C (Appendix B.3).

Controlled signal. Section 5 already runs a size-matched background arm $\delta_{\text{bg},g}^{\text{full}}(\ell)$ on the same pairs at every grid layer, so the identity floor can be replaced with a measured control on pre-existing data:

$$s_g^{\Delta,\text{ctrl}}(\ell) \;=\; \delta_{\text{obj},g}^{\text{full}}(\ell) \;-\; \delta_{\text{bg},g}^{\text{full}}(\ell). \tag{14}$$

Both arms overwrite receiver positions with same-image background states under the existence prompt: the object arm overwrites $\mathcal{T}_g$, and the background arm the disjoint set $\mathcal{T}_g^{\text{bg}} \subseteq \mathcal{U}_B$, size-matched to $\mathcal{T}_g$ whenever $\mathcal{U}_B$ allows (Appendix E.1). Their difference isolates the category-specific component of the erasure effect from the generic replacement shock that a size-matched intervention on arbitrary background patches already exhibits.

Recomputed statistics. Applying the pair-first half-mass operator $\mathcal{H}$ of Section 4.2 to $s_g^{\Delta,\text{ctrl}}$ gives Table 6, on the controlled common-support population $\mathcal{P}_{CR\Delta,\text{ctrl}} = \mathcal{P}_C \cap \mathcal{P}_R \cap \mathcal{P}_{\Delta,\text{ctrl}}$ with $\mathcal{P}_{\Delta,\text{ctrl}} = \{g : Z_g^{\Delta,\text{ctrl}} > 0\}$. The controlled criterion admits slightly more pairs than the identity one, so $|\mathcal{P}_{CR\Delta,\text{ctrl}}|$ exceeds its identity-baseline counterpart by 0.1–2.1% depending on the model. Six of seven models yield L_{50}^{Δ} identical to three decimal places under the two baselines; in Qwen2.5-VL-3B the controlled value is one grid step shallower ($0.278 \to 0.194$). The paired statistics are computed on $\mathcal{P}_{\Delta,\text{ctrl}} \cap \mathcal{P}_C$ rather than on $\mathcal{P}_{CR\Delta}$ and are therefore not compared cell by cell with Table 9; under the controlled signal $G_{50}^{\Delta C}$ is negative and $\pi^{\Delta C}$ exceeds 50% in every model. The ordering $L_{50}^{\Delta} < L_{50}^{C} < L_{50}^{R}$ is preserved in $7/7$ models under the controlled signal.

Sign of the controlled signal. $s_g^{\Delta,\text{ctrl}}(\ell)$ is negative at some mid-depth grid layers in three models (Idefics3-8B $6/11$ layers, Qwen2.5-VL-3B $5/11$, LLaVA-1.5-13B $2/11$), where the size-matched background arm is more destructive than the object arm. This does not affect L_{50}^{Δ}, which is computed

	L_{50}^{Δ}		$G_{50}^{\Delta C}$ (controlled)		
Model	identity	controlled	median	95% CI	$\pi^{\Delta C}$
LLaVA-1.5-7B	0.188	0.188	-0.094	$[-0.094, -0.094]$	0.560
LLaVA-1.5-13B	0.200	0.200	-0.100	$[-0.100, -0.075]$	0.571
Idefics3-8B	0.188	0.188	-0.094	$[-0.094, -0.094]$	0.558
Gemma-3-4B	0.294	0.294	-0.088	$[-0.118, -0.088]$	0.622
Qwen2.5-VL-3B	0.278	**0.194**	-0.194	$[-0.194, -0.194]$	0.643
Qwen2.5-VL-7B	0.179	0.179	-0.250	$[-0.250, -0.179]$	0.794
InternVL3.5-8B	0.194	0.194	-0.194	$[-0.194, -0.194]$	0.713

Table 6: Δ under the identity baseline $b_{\Delta} \equiv 0$ and under the controlled baseline of Eq. (14). The two L_{50}^{Δ} columns are computed on $\mathcal{P}_{CR\Delta}$ and $\mathcal{P}_{CR\Delta,\mathrm{ctrl}}$ respectively; bold marks the one cell that differs. $G_{50}^{\Delta C}$ and $\pi^{\Delta C}$ are on $\mathcal{P}_{\Delta,\mathrm{ctrl}} \cap \mathcal{P}_C$, with 95% pair-level percentile bootstrap intervals; this population differs from the $\mathcal{P}_{CR\Delta}$ of Table 9, so the two tables are not compared cell by cell. The ordering $L_{50}^{\Delta} < L_{50}^{C} < L_{50}^{R}$ holds in $7/7$ models under both baselines, and $G_{50}^{\Delta C} < 0$ and $\pi^{\Delta C} > 50\%$ in every model under both.

Model	$\lvert\mathcal{P}_C\rvert$	$\lvert\mathcal{P}_R\rvert$	$\lvert\mathcal{P}_{\Delta}\rvert$	$\lvert\mathcal{P}_{CR}\rvert$	$\lvert\mathcal{P}_{CR\Delta}\rvert$	$\lvert\mathcal{P}_J\rvert$
LLaVA-1.5-7B	867	994	946	861	815	965
LLaVA-1.5-13B	899	988	945	890	840	965
Idefics3-8B	896	991	951	889	843	930
Gemma-3-4B	738	986	954	725	702	955
Qwen2.5-VL-3B	839	993	934	833	776	955
Qwen2.5-VL-7B	806	998	957	804	768	955
InternVL3.5-8B	864	997	964	863	828	955

Table 7: Per-model analysis population sizes. $\lvert\mathcal{P}_C\rvert$ excludes pairs failing the signed prescreen $D_g > 0.5$ (Appendix B.4) or with $Z_g^C = 0$; $\lvert\mathcal{P}_R\rvert$ and $\lvert\mathcal{P}_{\Delta}\rvert$ exclude pairs with $Z_g^R = 0$ and $Z_g^{\Delta} = 0$. $\lvert\mathcal{P}_{CR}\rvert$ and $\lvert\mathcal{P}_{CR\Delta}\rvert$ are the intersections used by the confirmatory C–R analysis and by the three-axis analysis. $\lvert\mathcal{P}_J\rvert$ is the separate six-arm population of Section 5, the pairs with all six decomposition arms observed at every grid layer.

from the positive part $W_g^{\Delta}(\ell) = [s_g^{\Delta,\mathrm{ctrl}}(\ell)]_+$ (§4.2), and it is the same non-inertness of the background arm that motivates defining J as a contrast in Section 5; per-arm shifts are in Appendix E.4.

C Half-Mass Estimator Details

C.1 Axis Populations and Their Intersection

The three axes $X \in \{\Delta, C, R\}$ are defined on different subsets of the raw pair pool $\mathcal{P}_0$: a pair contributes to an axis only if its axis-specific signal exceeds baseline at some layer. Prior to per-pair normalization (Section 4.2) we require

$$\mathcal{P}_X = \big\{g \in \mathcal{P}_0 : \exists\, \ell \in \mathcal{G},\ s_g^X(\ell) > 0\big\}, \tag{15}$$

and drop pairs with $Z_g^X = \sum_{\ell \in \mathcal{G}} [s_g^X(\ell)]_+ = 0$. The three-axis comparison of Section 4.3 is made on the common-support pair set $\mathcal{P}_{CR\Delta} = \mathcal{P}_C \cap \mathcal{P}_R \cap \mathcal{P}_{\Delta}$, whose per-model size ranges from 92.5% to 95.8% of the smallest per-axis population and forms the pair-count column of Table 1. Per-model populations are in Table 7.

C.2 Bootstrap Procedures

Every reported interval is a 95% interval, but three distinct resampling procedures are used, one per family of statistics.

- *Marginal half-mass depths* (L_{50}^X): two-way cluster bootstrap. Each pair is indexed by (donor image id, receiver image id); the multiset of donor ids and the multiset of receiver

Model	marginal medians L_{50}^{C}	L_{50}^{R}	paired estimand on $\mathcal{P}_{CR}$ G_{50}	95% CI	π^{CR}
LLaVA-1.5-7B	0.281	0.656	-0.375	$[-0.375, -0.375]$	84.9%
LLaVA-1.5-13B	0.275	0.675	-0.375	$[-0.375, -0.300]$	90.1%
Idefics3-8B	0.281	0.562	-0.281	$[-0.281, -0.281]$	84.7%
Gemma-3-4B	0.382	0.765	-0.353	$[-0.382, -0.294]$	84.3%
Qwen2.5-VL-3B	0.472	0.472	-0.083	$[-0.083, -0.083]$	56.8%
Qwen2.5-VL-7B	0.464	0.643	-0.214	$[-0.214, -0.179]$	74.8%
InternVL3.5-8B	0.389	0.556	-0.194	$[-0.194, -0.167]$	68.9%

Table 8: Confirmatory C–R analysis. Marginal half-mass depths are medians over each axis's own population; $G_{50} = \mathrm{med}_g(h_{50,g}^{C} - h_{50,g}^{R})$ and π^{CR} are the paired statistics on $\mathcal{P}_{CR}$ (725–890 pairs per model, Table 7). Intervals are pair-level percentile bootstraps (Appendix C.2); several collapse to point estimates because per-pair half-mass differences are lattice-valued. $G_{50} < 0$ in $7/7$ models, with the 95% CI upper bound below zero in every row. The marginal tie in Qwen2.5-VL-3B ($0.472 = 0.472$) is a property of the marginal summary: the paired estimand on that model is strictly negative.

ids are resampled independently, and the variance is combined as $V = V_{\mathrm{donor}} + V_{\mathrm{receiver}} - V_{(\mathrm{donor,receiver})}$. Intervals are point $\pm\, 1.96\sqrt{V}$ over 1,000 draws with seed 606060.

- *Paired half-mass differences* (G_{50}^{CR}, $G_{50}^{\Delta C}$, and the ordered fractions π^{XY}): pair-level percentile bootstrap. Pairs are resampled with replacement from the relevant population and the pair median of the per-pair gaps is recomputed. Table 9 uses 10,000 draws with seed 42; Table 8 uses 2,000 draws with the same seed, which gives identical intervals because the per-pair gaps are lattice-valued. The controlled-baseline $G_{50}^{\Delta C}$ of Table 6 uses 2,000 draws with seed 20260916.
- *Set-level non-additivity statistics* ($\bar{J}(\ell)$ and $\bar{J}^{\mathrm{band}}$): two-way cluster bootstrap over the same donor and receiver index sets. Layerwise intervals for $\bar{J}(\ell)$ use percentile bounds over 2,000 draws with seed 20260825, whereas intervals for $\bar{J}^{\mathrm{band}}$ use 1,000 draws with seed 99. The six-arm population $\mathcal{P}_J$, the pairs with all six arms observed at every grid layer, contains 930–965 pairs per model drawn from 419–440 donor and 177–182 receiver clusters.

The first family uses a normal interval and the second percentile bounds because the paired gaps are supported on a small lattice, where a normal approximation is a poor fit and percentile bounds often collapse onto a single grid value.

Pair-first principle. Every interval is attached to a pair median or a pair mean, never to a layerwise cohort mean.

C.3 Pair-First Half-Mass Depth

All reported L_{50}^{X} values are medians of pair-level half-mass depths, $L_{50}^{X} = \mathrm{med}_{g\in\mathcal{P}}\, h_{50,g}^{X}$ with $h_{50,g}^{X}$ from Eq. (5). The cohort-mean profiles $\bar{\mu}^{X}(\ell)$ plotted in Figure 2 are used only for display: the depth at which such a curve first crosses one half can differ from L_{50}^{X}, because averaging over pairs and taking a median do not commute on a finite layer grid. No reported number is read from the plotted curves.

D Paired Support for the Three-Axis Ordering

Section 4.3 states the ordering $L_{50}^{\Delta} < L_{50}^{C} < L_{50}^{R}$ at the model-median level. This appendix gives its pair-level counterpart: for each pair g and each axis pair $(X, Y) \in \{(\Delta, C), (C, R)\}$ we form the half-mass gap $d_g^{XY} \equiv h_{50,g}^{X} - h_{50,g}^{Y}$, and report its pair median G_{50}^{XY} together with the ordered fraction $\pi^{XY} \equiv |\mathcal{P}|^{-1} \sum_g \mathbf{1}[d_g^{XY} < 0]$.

Model	$\lvert\mathcal{P}_{CR\Delta}\rvert$	$G_{50}^{\Delta C}$ on $\mathcal{P}_{CR\Delta}$ median	95% CI	$\pi^{\Delta C}$	G_{50}^{CR} on $\mathcal{P}_{CR\Delta}$ median	95% CI	π^{CR}
LLaVA-1.5-7B	815	-0.094	$[-0.094, -0.094]$	60.2%	-0.375	$[-0.375, -0.375]$	85.3%
LLaVA-1.5-13B	840	-0.100	$[-0.100, -0.100]$	58.6%	-0.375	$[-0.375, -0.300]$	91.0%
Idefics3-8B	843	-0.094	$[-0.094, -0.094]$	63.7%	-0.281	$[-0.281, -0.281]$	84.9%
Gemma-3-4B	702	-0.088	$[-0.118, -0.088]$	59.7%	-0.382	$[-0.382, -0.294]$	84.2%
Qwen2.5-VL-3B	776	-0.194	$[-0.194, -0.194]$	70.5%	-0.083	$[-0.083, -0.083]$	56.7%
Qwen2.5-VL-7B	768	-0.179	$[-0.179, -0.179]$	75.1%	-0.214	$[-0.286, -0.179]$	75.3%
InternVL3.5-8B	828	-0.194	$[-0.194, -0.111]$	67.9%	-0.194	$[-0.194, -0.167]$	69.6%

Table 9: Paired within-pair half-mass differences on $\mathcal{P}_{CR\Delta}$. $G_{50}^{\Delta C}$ is the pair median of $h_{50,g}^{\Delta} - h_{50,g}^{C}$; G_{50}^{CR} is the pair median of $h_{50,g}^{C} - h_{50,g}^{R}$. 95% intervals are pair-level percentile bootstraps (10,000 draws, seed 42); several collapse to point estimates because per-pair half-mass differences are lattice-valued on the 11-point grid. Every G is negative in every model, and $\pi^{\Delta C}$ and π^{CR} exceed 50% in every model. The main text reports the depth ordering at the model-median level and cites this table as the within-pair sensitivity that supports the same ordering.

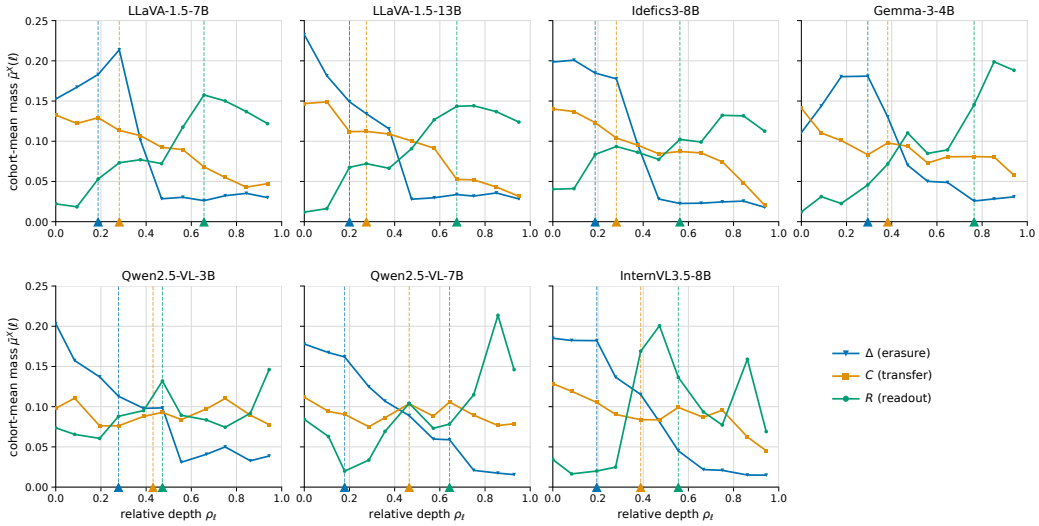


Figure 4: Per-model cohort-mean mass profiles $\bar{\mu}^X(\rho_\ell)$ for the three axes on the common-support pair set $\mathcal{P}_{CR\Delta}$. Triangles on the depth axis mark the model-level pair medians L_{50}^X, computed from the pair-level depths in Eq. (5), rather than the depths at which the cumulative plotted mean profiles reach one half. Every model places the three depths in the same order: $L_{50}^{\Delta} < L_{50}^{C} < L_{50}^{R}$.

The confirmatory G_{50}^{CR} analysis of Section 4.3 runs on the marginal $\mathcal{P}_{CR}$ (Table 8). Table 9 recomputes the same estimator on the smaller $\mathcal{P}_{CR\Delta}$ so that all three axes share the same pairs, and extends it to $G_{50}^{\Delta C}$ as an exploratory paired statistic; intervals are pair-level percentile bootstraps (Appendix C.2). Every model places the three half-mass depths in the same order, with $G_{50}^{\Delta C} < 0$, $G_{50}^{CR} < 0$, and $\pi^{\Delta C}, \pi^{CR} > 50\%$: each adjacent ordering stated at the model-median level in Table 1 also holds for a majority of pairs within each model. Figure 4 shows the same population one model at a time, with the three axes overlaid and their L_{50}^X marked, so the separation can be read within a model rather than across the panels of Figure 2.

Discreteness of the paired statistic. Per-pair gaps take only integer multiples of a grid step, so a sample median lands either on a grid value or on the midpoint of two adjacent ones. This affects one cell: restricting Qwen2.5-VL-3B from $\mathcal{P}_C$ to $\mathcal{P}_{CR\Delta}$ removes 63 of 839 pairs, leaving an even sample whose two central order statistics fall on the adjacent grid points 0.389 and 0.472, so the median is their midpoint 0.431. Either extreme is $\leq L_{50}^R = 0.472$ in that model, and the paired

estimand there is $G_{50}^{CR} = -0.083$, so neither the marginal nor the within-pair ordering depends on the midpoint convention.

E Set-Level Non-Additivity: Arms, Variant Selection, and Robustness

E.1 Six Arms and the Pinning Assertion

At each grid layer ℓ, six single-layer Δ-type interventions (Eq. (2)) are run under the existence prompt x_B^{exist}: three overwrite object tokens and three overwrite background tokens, all of them with same-image background states.

Sets and assignments. A seeded balanced partition splits the receiver object-token set into halves $\mathcal{T}_g = \mathcal{T}_g^{\text{a}} \sqcup \mathcal{T}_g^{\text{b}}$ of sizes differing by at most one, and the background destination set is split the same way into $\mathcal{T}_g^{\text{bg}} = \mathcal{T}_g^{\text{bg,a}} \sqcup \mathcal{T}_g^{\text{bg,b}}$. Each side has its own assignment supplying replacement states. The object assignment $\beta_g : \mathcal{T}_g \to \mathcal{U}_B$ is the one from Section 4.1.1. The background assignment β_g^{bg} enumerates $\mathcal{U}_B$ in raster order and pairs its first $n = \min(|\mathcal{T}_g|, \lfloor |\mathcal{U}_B|/2 \rfloor)$ patches with its last n, so $\mathcal{T}_g^{\text{bg}}$ and $\beta_g^{\text{bg}}(\mathcal{T}_g^{\text{bg}})$ are disjoint and the background arm never reads a state from a position it overwrites. The halving is what bounds n: when the receiver has fewer than $2|\mathcal{T}_g|$ annotation-free patches, $n < |\mathcal{T}_g|$ and the background arm patches fewer positions than the object arm (§E.4).

The six arms. Writing $x_B^{\text{exist}}[\ell; S \mapsto h_{B,\beta}^{\ell}]$ for the intervention that replaces each $i \in S$ by the state its side's assignment gives, the arms are

$$\begin{aligned} &\text{obj-full} : S = \mathcal{T}_g && \text{obj-a} : S = \mathcal{T}_g^{\text{a}} && \text{obj-b} : S = \mathcal{T}_g^{\text{b}} \\ &\text{bg-full} : S = \mathcal{T}_g^{\text{bg}} && \text{bg-a} : S = \mathcal{T}_g^{\text{bg,a}} && \text{bg-b} : S = \mathcal{T}_g^{\text{bg,b}}. \end{aligned}$$

Pinning. Each position receives the same state in every arm on its side that touches it. Consider $i \in \mathcal{T}_g^{\text{a}}$: both obj-full and obj-a overwrite it, and by pinning both write $h_{B,\beta_g(i)}^{\ell}$ there, so the residual $I_g^{\text{obj}}(\ell) = \delta_{\text{obj},g}^{\text{full}}(\ell) - \delta_{\text{obj},g}^{\text{a}}(\ell) - \delta_{\text{obj},g}^{\text{b}}(\ell)$ compares the arms only at positions that one of them leaves untouched. Without pinning the residual would instead report variation in the replacement material—what is written at each position—rather than set composition, which positions are overwritten. The same holds on the background side with β_g^{bg}.

E.2 Variant Selection Across Arms

The margin m takes a maximum over four yes-variants and four no-variants (Appendix A.4), so an arm could in principle move m by changing which variant wins rather than by changing the answer, and that movement would enter the residual I. Variant switching does not drive the estimate: the selection is identical across the clean forward and all six arms on every row in every model (Appendix B.1), so no part of I^{obj} or I^{bg} can come from it.

E.3 Per-Model $\bar{J}^{\text{band}}$

Table 10 reports the primary shallow-band summary per model together with two alternative aggregations, the pair median and a 10% trimmed mean. Every interval clears zero in every model, and the mean/median ratio ranges from 1.8 (Qwen2.5-VL-3B) to 5.8 (Gemma-3-4B): the shallow-band signal is a right tail on top of a positive median, not a tail alone.

E.4 Robustness

Band cutoff. Sweeping the cutoff ρ_0 that defines the band $\rho_\ell \leq \rho_0$ through $\{0.20, 0.25, 0.30, 0.35, 0.40\}$ leaves $\bar{J}^{\text{band}}$ strictly positive at every value in every model, within 19% of its value at the primary $\rho_0 = 0.30$, and preserves the ordering of the seven models across the sweep.

Model	mean over pairs (primary) $\bar{J}^{\text{band}}$	95% CI	median over pairs med J^{band}	95% CI	10% trimmed mean trim	95% CI
LLaVA-1.5-7B	$+0.307$	$[+0.188, +0.469]$	$+0.094$	$[+0.062, +0.094]$	$+0.127$	$[+0.086, +0.182]$
LLaVA-1.5-13B	$+0.236$	$[+0.143, +0.382]$	$+0.062$	$[+0.031, +0.094]$	$+0.082$	$[+0.052, +0.134]$
Idefics3-8B	$+0.185$	$[+0.123, +0.259]$	$+0.063$	$[+0.062, +0.094]$	$+0.105$	$[+0.069, +0.154]$
Gemma-3-4B	$+0.914$	$[+0.465, +1.541]$	$+0.156$	$[+0.094, +0.219]$	$+0.212$	$[+0.135, +0.363]$
Qwen2.5-VL-3B	$+0.230$	$[+0.161, +0.313]$	$+0.125$	$[+0.094, +0.156]$	$+0.164$	$[+0.117, +0.227]$
Qwen2.5-VL-7B	$+0.693$	$[+0.477, +0.972]$	$+0.281$	$[+0.188, +0.375]$	$+0.399$	$[+0.289, +0.530]$
InternVL3.5-8B	$+0.830$	$[+0.543, +1.222]$	$+0.219$	$[+0.156, +0.312]$	$+0.355$	$[+0.243, +0.525]$

Table 10: Three cohort aggregations of J^{band} per model, with 95% two-way cluster-bootstrap intervals over donor and receiver images. The mean is the paper's primary summary; median and trimmed mean are sensitivity. Every interval clears zero, and the mean/median ratio (1.8–5.8 across models) indicates a right tail on top of a positive median.

Background non-inertness. Table 11 reports the shallow-band margin reductions d^{obj} and d^{bg} per model. The background arm is nearly inert only in the two LLaVA-1.5 models; in the other five it is not, and in three of those it moves the answer away from "no". This is why J is defined as a contrast rather than as I^{obj} alone: subtracting I^{bg} removes the non-additivity that intervening on arbitrary patches already exhibits.

Model	d^{obj} (shallow band)	d^{bg} (shallow band)
LLaVA-1.5-7B	$+0.514$	$+0.003$
LLaVA-1.5-13B	$+0.406$	$+0.005$
Idefics3-8B	$+0.536$	$+0.102$
Gemma-3-4B	$+1.463$	-0.524
Qwen2.5-VL-3B	$+0.406$	$+0.169$
Qwen2.5-VL-7B	$+1.545$	-1.311
InternVL3.5-8B	$+1.455$	-0.221

Table 11: Per-arm shallow-band destructive shifts, averaged over pairs across the four shallow-band layers. The background arm is nearly inert only in the two LLaVA-1.5 models; the other five motivate defining J as the contrast $I^{\text{obj}} - I^{\text{bg}}$.

Strict size match. Because $|\mathcal{T}_g^{\text{bg}}| = \min(|\mathcal{T}_g|, \lfloor |\mathcal{U}_B|/2 \rfloor)$ (§E.1), a receiver with little annotation-free background gives a background arm smaller than the object arm, which would bias J upward. We therefore recompute on the strict size-match subset, the pairs whose two arms patch equally many positions, which retains 73–93% of $\mathcal{P}_J$ per model. Table 12 shows the interval clearing zero in $7/7$ models, and it does so at each of the four shallow-band layers as well. The direction of the bias is the opposite of the one we guarded against: six of seven models give a *larger* strict-match estimate, and the excluded pairs have a positive but small $\bar{J}^{\text{band}}$ of their own ($+0.02$ to $+0.29$), so they were pulling the primary estimate down.

E.5 Prompt Sensitivity of Shallow-Band Erasure Sensitivity

The main analysis uses a single existence prompt (Appendix A.4). We recompute Δ under three further prespecified templates—`Can you see {a|an}` y_B `in the image?`, `Does the image contain {a|an}` y_B`?`, and `Is {a|an}` y_B `visible in this image?`—holding the pair pool, object-token definition, intervention arms, and layer grid fixed; C and R are not rerun. The summary is $D_{P_k}^{\text{band}}(m)$, the mean over eligible pairs of the mean of $[-\Delta_{g,P_k}(\ell)]_+$ over the four shallow-band layers. Each model–prompt cell retains between 960 and 965 pairs, and clean-yes rates range from 0.904 to 0.998, so all four templates still elicit the intended positive existence judgment.

Table 13 gives $D_{P_k}^{\text{band}}(m)$ per model and template. The prespecified statistic was the mean of the six pairwise Spearman correlations between the seven-model rankings under the four prompts, with a threshold of 0.8; the observed mean is 0.899 and the smallest individual correlation is 0.857. Abso-

Model	$\lvert\mathcal{P}_J^{\text{strict}}\rvert$	Retained	$\bar{J}^{\text{band}}$	95% CI
LLaVA-1.5-7B	901	93%	$+0.324$	$[+0.202, +0.510]$
LLaVA-1.5-13B	901	93%	$+0.252$	$[+0.145, +0.415]$
Idefics3-8B	677	73%	$+0.185$	$[+0.115, +0.274]$
Gemma-3-4B	780	82%	$+1.069$	$[+0.535, +1.769]$
Qwen2.5-VL-3B	699	73%	$+0.276$	$[+0.192, +0.371]$
Qwen2.5-VL-7B	699	73%	$+0.840$	$[+0.561, +1.193]$
InternVL3.5-8B	778	81%	$+0.970$	$[+0.608, +1.424]$

Table 12: Shallow-band mean $\bar{J}^{\text{band}}$ on the strict size-match subset, the pairs whose background arm patches exactly $|\mathcal{T}_g|$ positions. Intervals are the same two-way cluster bootstrap as the primary estimate (Appendix C.2). Every interval clears zero, as does the interval at each of the four shallow-band layers in every model. Point estimates are 0.99 to 1.21 times the primary values of Table 10, so the excluded pairs lower $\bar{J}^{\text{band}}$ rather than inflate it.

Model	P_0	P_1	P_2	P_3	Range
LLaVA-1.5-7B	0.5145	0.4983	0.4360	0.5071	0.0785
LLaVA-1.5-13B	0.4060	0.3856	0.3984	0.3892	0.0204
Idefics3-8B	0.5236	0.6028	0.5520	0.5723	0.0792
Gemma-3-4B	1.4518	1.4592	1.4677	1.3084	0.1593
Qwen2.5-VL-3B	0.4045	0.2503	0.4007	0.4605	0.2102
Qwen2.5-VL-7B	1.5383	1.3224	1.5570	1.8802	0.5578
InternVL3.5-8B	1.4426	1.8969	1.6268	1.3722	0.5247

Table 13: Shallow-band destructive susceptibility $D_{P_k}^{\text{band}}(m)$ for each model and existence-prompt template. The final column gives the within-model range across the four templates.

lute magnitudes remain prompt-dependent, with a largest within-model range of 0.558. The ablation therefore supports the stability of the cross-model *ranking* of shallow-band erasure sensitivity within this panel, not the invariance of its magnitude and not the $\Delta \prec C \prec R$ ordering, which we do not recompute here.

F LVIS Extension: A Larger Category Vocabulary

The object-token sets used throughout §4 are defined from COCO instance masks, whose category vocabulary is limited to 80 classes. This appendix re-measures Δ and C on LVIS v1 val (Gupta et al., 2019), whose 1,203 categories (rare, common, and frequent) are annotated over the same val2017 images, and reports what does and does not carry over. R and the set-level non-additivity analysis of §5 are not re-measured; §F.4 gives the reason for each.

F.1 Pair Construction

Instance masks are projected onto each model's patch grid exactly as in Appendix A.7, at the same coverage threshold $\theta = 0.3$, over the 4,752 LVIS images whose val2017 file is present. Candidate pairs pass the same filter as the main pool: different categories, non-empty object-token sets, and $\text{IoU}(\mathcal{O}_A, \mathcal{O}_B) \geq 0.4$ in all seven models simultaneously. The pool is filled by donor bucket under a round-robin over receiver annotations, as in Appendix A.6.

Two properties of the resulting pool bound what it can support. First, the rare-donor bucket is exhausted at 592 candidates after the seven-model IoU intersection, below the cap we had set; its size is a property of the data, not a design choice. Second, receivers are concentrated in the frequent bucket (2,091 of 2,192 pairs; 98 common, 3 rare), because few val2017 images carry a rare annotation that survives the intersection; any bucket-level statement below is therefore about the *donor* bucket only. The final pool contains 592 rare-donor, 800 common-donor, and 800 frequent-donor pairs, totalling 2,192.

Model	Data	$\lvert\mathcal{P}_\Delta\rvert$	$\lvert\mathcal{P}_C\rvert$	$\lvert\mathcal{P}_{\Delta C}\rvert$	L_{50}^{Δ}	L_{50}^{C}
LLaVA-1.5-7B	COCO	946	867	821	0.188	0.281
	LVIS	2124	1760	1702	0.188	0.375
LLaVA-1.5-13B	COCO	945	899	849	0.200	0.275
	LVIS	2121	1836	1780	0.200	0.375
Idefics3-8B	COCO	951	896	850	0.188	0.281
	LVIS	2113	1712	1646	0.188	0.375
Gemma-3-4B	COCO	954	738	714	0.294	0.382
	LVIS	2138	1549	1515	0.294	0.382
Qwen2.5-VL-3B	COCO	934	839	782	0.278	0.389
	LVIS	2077	1758	1660	0.278	0.472
Qwen2.5-VL-7B	COCO	957	806	770	0.179	0.464
	LVIS	2162	1726	1699	0.179	0.464
InternVL3.5-8B	COCO	964	864	829	0.194	0.389
	LVIS	2168	1805	1786	0.194	0.389

Table 14: Model-level half-mass depths on the two-axis common support $\mathcal{P}_{\Delta C} = \mathcal{P}_C \cap \mathcal{P}_\Delta$, for the COCO pool of the main text and the LVIS pool of Appendix F.1. Both columns use the operator $\mathcal{H}$ of §4.2 unchanged. The COCO rows are *not* those of Table 1, which is computed on $\mathcal{P}_{CR\Delta}$; dropping the $\mathcal{P}_R$ intersection changes which pairs contribute, and the two agree wherever it does not. $L_{50}^{\Delta} < L_{50}^{C}$ in 7/7 models on both datasets, and L_{50}^{Δ} is identical across datasets in every model.

LVIS category names carry parenthetical and underscored qualifiers. Before the first token of the donor label is taken, names are reduced to a display form by dropping parentheses and replacing underscores with spaces, so that `car_(automobile)` scores as *car*. The same reduction is applied when the transfer baseline b_C is calibrated (Appendix B.3), so target and baseline score the same string. The reduction changes the surface form of 44 of the 385 donor categories in the pool and discards a qualifier in 3–4 of them, where a narrower sense is scored as its head noun.

F.2 Shallow-Band Erasure Sensitivity

We additionally compare the shallow-band erasure sensitivity $D^{\mathrm{band}}(m)$—the same scalar used in Appendix E.5—between the COCO and LVIS pools. The seven-model rankings remain similar (Spearman $\rho = 0.893$), comparable to the mean $\rho = 0.899$ across prompt templates.

This auxiliary comparison concerns only the cross-model magnitude ranking of Δ; it carries no depth information and does not re-establish the $\Delta \prec C \prec R$ ordering.

F.3 The $\Delta \prec C$ Ordering

The depth ordering is assessed with the half-mass operator $\mathcal{H}$ of §4.2, unchanged. Because R is not measured here, the common-support population is $\mathcal{P}_{\Delta C} = \mathcal{P}_C \cap \mathcal{P}_\Delta$ rather than the $\mathcal{P}_{CR\Delta}$ of Table 1; the axis-population rule of Appendix C.1 is otherwise applied as stated. The COCO side is recomputed on $\mathcal{P}_{\Delta C}$ as well, so no column below is read off Table 1: the two share an estimator but may differ where dropping the $\mathcal{P}_R$ intersection changes which pairs contribute. The signed prescreen $D_g > 0.5$, the measured baseline b_C, and the instrumentation floor $b_\Delta \equiv 0$ are as in §4.1.1–4.1.2.

Table 14 places $L_{50}^{\Delta} < L_{50}^{C}$ in 7/7 models on both datasets. L_{50}^{Δ} is identical across the two datasets in every model. L_{50}^{C} is identical in three models and one grid step deeper in the other four—one step is the resolution of the eleven-point grid, so this is stability at the available resolution.

Marginal medians are not within-pair lags (§4.2), so Table 15 gives the paired form used in Table 9: the pair median $G_{50}^{\Delta C}$ of $h_{50,g}^{\Delta} - h_{50,g}^{C}$ and the ordered fraction $\pi^{\Delta C}$, with pair-level percentile bootstrap intervals (10,000 draws, seed 42; Appendix C.2).

$G_{50}^{\Delta C}$ is negative and $\pi^{\Delta C} > 50\%$ in every model on both datasets. Several intervals collapse to a point because per-pair gaps are lattice-valued, as in Table 9.

Model	Data	n	$G_{50}^{\Delta C}$	95% CI	$\pi^{\Delta C}$
LLaVA-1.5-7B	COCO	821	-0.094	$[-0.094, -0.094]$	60.4%
	LVIS	1702	-0.094	$[-0.188, -0.094]$	65.2%
LLaVA-1.5-13B	COCO	849	-0.100	$[-0.100, -0.100]$	58.7%
	LVIS	1780	-0.175	$[-0.175, -0.175]$	70.4%
Idefics3-8B	COCO	850	-0.094	$[-0.094, -0.094]$	63.6%
	LVIS	1646	-0.188	$[-0.188, -0.094]$	62.4%
Gemma-3-4B	COCO	714	-0.088	$[-0.118, -0.088]$	59.7%
	LVIS	1515	-0.088	$[-0.088, -0.088]$	56.2%
Qwen2.5-VL-3B	COCO	782	-0.194	$[-0.194, -0.194]$	70.5%
	LVIS	1660	-0.111	$[-0.194, -0.111]$	63.7%
Qwen2.5-VL-7B	COCO	770	-0.179	$[-0.179, -0.179]$	75.1%
	LVIS	1699	-0.179	$[-0.179, -0.179]$	73.5%
InternVL3.5-8B	COCO	829	-0.194	$[-0.194, -0.111]$	67.9%
	LVIS	1786	-0.194	$[-0.194, -0.194]$	69.8%

Table 15: Paired within-pair half-mass differences on $\mathcal{P}_{\Delta C}$. $G_{50}^{\Delta C}$ is the pair median of $h_{50,g}^{\Delta} - h_{50,g}^{C}$ and $\pi^{\Delta C}$ the fraction of pairs with $h_{50,g}^{\Delta} < h_{50,g}^{C}$. Intervals are pair-level percentile bootstraps (10,000 draws, seed 42; Appendix C.2); several collapse to point estimates because per-pair differences are lattice-valued on the eleven-point grid, as in Table 9. $G_{50}^{\Delta C} < 0$ and $\pi^{\Delta C} > 50\%$ in every model on both datasets.

	Δ axis			C axis		
Model	rare	common	frequent	rare	common	frequent
LLaVA-1.5-7B	3.0%	3.3%	3.0%	27.2%	20.5%	13.4%
LLaVA-1.5-13B	3.2%	3.3%	3.3%	21.3%	14.4%	14.4%
Idefics3-8B	2.9%	3.6%	4.1%	25.7%	23.5%	17.5%
Gemma-3-4B	1.9%	3.0%	2.4%	34.0%	25.3%	30.0%
Qwen2.5-VL-3B	7.1%	5.0%	4.1%	25.5%	19.5%	15.9%
Qwen2.5-VL-7B	1.0%	1.8%	1.3%	23.3%	21.4%	19.6%
InternVL3.5-8B	1.0%	1.4%	0.9%	20.3%	19.1%	14.3%

Table 16: Fraction of LVIS pool pairs that each axis does not retain, by donor bucket (592 rare, 800 common, 800 frequent). A pair is lost to an axis when its controlled signal never exceeds baseline ($Z_g^X = 0$, Appendix C.1), and to C additionally when it fails the signed prescreen $D_g > 0.5$. The Δ axis loses little in any bucket. The C axis loses rare-donor pairs at the higher rate in all seven models, so the population underlying L_{50}^{C} is shifted away from rare donors relative to the pool.

Substituting the controlled erasure-sensitivity signal of Appendix B.6 for the primary signal based on $b_\Delta \equiv 0$—which the LVIS sweep supports because it also ran the size-matched background arm—leaves $G_{50}^{\Delta C}$ negative in all fourteen cells (7 models $\times$ 2 datasets).

F.4 Scope and Caveats

The C-axis eligibility rule disproportionately filters rare-donor pairs. Pairs enter $\mathcal{P}_C$ only if they satisfy the signed-gap prescreen $D_g > 0.5$ (Appendix B.4) and have positive controlled mass, $Z_g^C > 0$ (Appendix C.1). Table 16 reports the resulting pair loss by donor bucket. The Δ axis loses at most 7.1% of pairs in every bucket of every model, whereas the C axis loses 20.3–34.0% of rare-donor pairs compared with 13.4–30.0% of frequent-donor pairs. The rare-donor loss rate is higher than the frequent-donor rate in all seven models. The $\Delta \prec C$ result above is therefore measured on a population in which rare donors are underrepresented relative to the original pool, and it should not be interpreted as evidence about rare categories specifically.

The preregistered rare-category hypothesis is not decided. Our preregistration for this extension stated a hypothesis about L_{50}^{R} being deeper for rare donors than for common and frequent ones.

Because R is not measured here, the hypothesis is undecided rather than supported or refuted. We draw no conclusion from the Δ-axis analogue, which is not the preregistered test.

R and set-level non-additivity are not measured. R compares object-token states against a bank of category-labeled text phrases encoded once per model (§4.1.3). Extending it to 1,203 categories would require rebuilding that bank in all seven models, which is a separate construction rather than a rerun. The set-level non-additivity analysis of §5 is excluded by the annotation protocol: LVIS is federated, so a patch left unannotated in a given image need not be free of annotated objects, and the background set $\mathcal{U}_B$ cannot be constructed under the guarantee §5 relies on.

What the extension supports. On an LVIS-derived pool of 2,192 pairs spanning 385 donor categories, the two-axis ordering $\Delta \prec C$ holds in all seven models under both the primary and controlled definitions of erasure sensitivity. Because R is not re-measured, this extension does not re-establish the full three-axis ordering of §4.3.