THE UMBRA EFFECT: USE–ELICITATION DISSOCIATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Anthropic's J-lens study links verbalizable representations to a global workspace in language models. This raises a complementary question for vision-language models: when visual information influences an answer, can an intervention also elicit that information in words? We identify the **Umbra effect**, a dissociation between causal use and verbal elicitation of visual information, and characterize it using mirrored-image interchange, linear probes, and calibrated signed activation injection. Across five checkpoints from two model families, we characterize three findings: **1) Use–elicitation dissociation.** Four Qwen checkpoints show weak spatial pooled-direction elicitation in tested settings where interchange changes answer margins, despite effective color controls. **2) Vision-to-answer handoff.** A detailed Qwen2.5-VL-3B study localizes a crossing of vision- and answer-position interchange effects at L23, within a band comprising six tested layers with causally influential, linearly decodable spatial representations. **3) Payload-dependent elicitation.** LLaVA-OneVision responds to position-preserving spatial payloads where pooled payloads fail, identifying a boundary on the cross-model pattern. Photograph and clutter tests additionally extend the use–elicitation comparison within the 3B case study. Together, these results motivate auditing causal contribution, decodability, and controlled elicitation as distinct measurements, with payload and dose calibration guiding the interpretation of negative readouts.

1 INTRODUCTION

Anthropic's J-lens study identifies verbalizable representations with several functional properties of a global workspace in language models (Gurnee et al., 2026). It connects what a model can report with representations involved in internal reasoning. For vision-language models, this motivates a complementary question: how closely does the visual information used to answer a question align with what a specified intervention can elicit? A model can attend to relevant image regions and still answer incorrectly (Zhang et al., 2025a); spatial benchmarks expose related failures in identifying relations and composing visual evidence (Kamath et al., 2023; Liu et al., 2023; Fu et al., 2024). To interpret these failures, we need to distinguish information that contributes to an answer, information recoverable from activations, and information that can be elicited in a recipient prompt.

Internal analysis offers several complementary measurements. A linear probe tests whether a label can be recovered from activations. An interchange intervention tests how replacing activations changes a task outcome. Activation steering tests whether an added direction can change generation. Their answers need not agree because each operates on a different object: a predictive decision boundary, a replacement of an existing state, or a perturbation in a recipient context. The practical question is how to interpret disagreement among these measurements on the same visual relation.

We call the observed use–elicitation dissociation the **Umbra effect**: visual relations influence answers under interchange yet respond weakly to elicitation with the tested payloads. To characterize this phenomenon, we measure *causal use*, *linear decodability*, and *elicitation*. Mirrored image pairs reverse one spatial relation while preserving object identity. We interchange their vision-position residuals, fit pair-grouped probes, and inject relation payloads into a neutral-image prompt. Signed injections are compared through a direction differential, Δ, at a dose selected using color controls. We use *elicitation* specifically for generated relation words under this intervention protocol. This definition makes the tested pathway and recipient prompt part of the result.

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

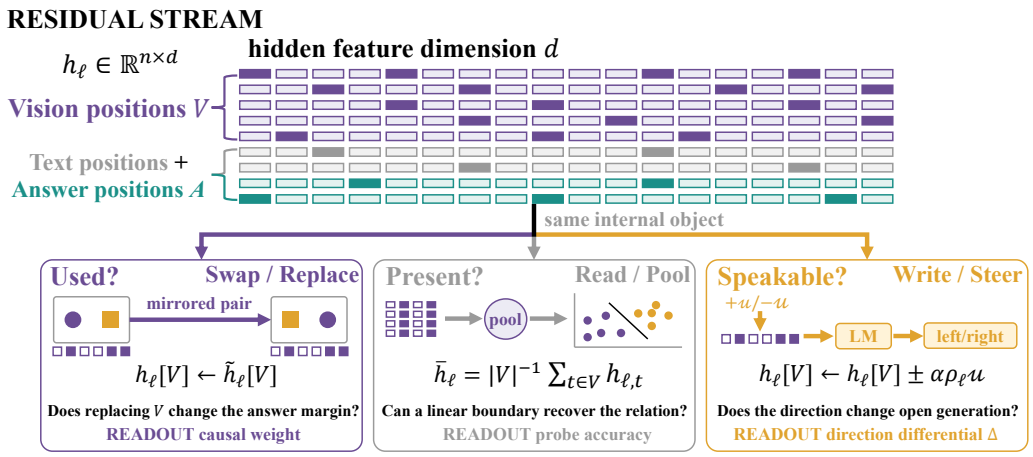


Figure 1: **Three questions about the same layer.** Mirrored-pair interchange measures the contribution of vision-position residuals to the answer margin; a pair-grouped linear probe measures relation decodability; signed activation injection measures whether a payload changes generated relation words. These are complementary measurements with different interventions and recipient contexts.

The recurring Qwen result is a use–elicitation dissociation: spatial interchange affects answer margins while pooled injection produces little direction-sensitive response. The full three-readout intersection depends on model and layer. Qwen2.5-VL-3B has an Umbra band with probe accuracy .510–.575; the 7B models cross the probe threshold through much more of their causal region; Qwen2-VL-2B remains below it. LLaVA-OneVision supplies the informative contrast: preserving a spatial payload's position structure yields successful relation generation in settings where pooling does not.

Contributions. 1) The Umbra effect: we identify a recurring use–elicitation dissociation across four Qwen checkpoints and a full three-readout band in the 3B case study. **2) Vision-to-answer handoff:** in that case study, answer-window interchange overtakes vision interchange at L23, while the vision-position probe peaks later. **3) Payload-dependent elicitation:** a five-model comparison identifies LLaVA's position-preserving response as a cross-family boundary; photograph and clutter tests examine image-domain transfer on the 3B model.

These distinctions change how an unsuccessful audit should guide the next analysis. A negative injection result calls for a control that verifies the chosen dose and site, followed by a payload comparison. A high probe score calls for an intervention that tests the representation's task contribution. A proposed correction should be evaluated without supplying the correct label. Our answer-window experiment provides a diagnostic upper bound under a supplied label; its weaker non-oracle alternatives make that distinction concrete. A separate J-lens audit tests whether a token readout is usable on the target decoder before it is assigned a representational interpretation.

2 Measuring Causal Use, Decodability, and Elicitation

The design separates two questions: which measurements disagree within one model, and which parts of that disagreement recur when the model or task changes. Qwen2.5-VL-3B is the detailed case study. The other checkpoints test model variation; photographs and clutter test the image domain on that same 3B model.

Models and task roles. We evaluate Qwen2.5-VL-3B/7B (Bai et al., 2025), Qwen2-VL-2B/7B (Wang et al., 2024), and LLaVA-OneVision-7B (Li et al., 2025). Within-generation size comparisons examine how decodability changes with scale; the two Qwen generations test checkpoint variation; LLaVA tests sensitivity to a different model family and visual interface. The resulting comparisons characterize these selected checkpoints across size, generation, and family. Cardinal and diagonal relations test answer-token format. Semantic and counting tasks test content dependence. COCO

photographs (Lin et al., 2014) and cluttered synthetic scenes test whether the clean-canvas result persists with more complex visual input.

Inputs and token positions. For input $x = (I, q)$, let $h_\ell \in \mathbb{R}^{n\times d}$ be a decoder residual stream at layer ℓ, $V(x)$ its vision positions, and $A(x)$ the last four prompt positions, called the *answer window*. The 3B decoder has 36 layers; we inspect 25 selected layers. The main set contains 32 mirrored pairs of two colored shapes with a reversed relation $r \in \{\text{left}, \text{right}, \text{above}, \text{below}\}$. The correct and opposite relation tokens, y_r and $y_{\bar{r}}$, define the answer margin

$$m(x) = \log p_f(y_r \mid x) - \log p_f(y_{\bar{r}} \mid x). \tag{1}$$

The main analysis uses 23 pairs for which both members have positive margins. This selection isolates interventions on initially correct comparisons. The remaining nine pairs support a separate diagnostic of failed predictions. Probe folds are grouped by mirrored pair so partners remain in the same fold.

Causal use. We replace vision residuals with those of the mirrored partner and compare the resulting margin with an unrelated-item interchange. If $m^{\text{swap}}_{i,\ell}$ and $m^{\text{ctl}}_{i,\ell}$ are the two edited margins, the control-adjusted interchange effect is

$$\text{cw}_\ell = \frac{1}{N}\sum_{i=1}^{N}(m^{\text{ctl}}_{i,\ell} - m^{\text{swap}}_{i,\ell}), \qquad \mathcal{C} = \{\ell : \text{cw}_\ell > 0.2,\ p_\ell < .01\}. \tag{2}$$

Here p_ℓ is a one-sided paired sign-flip permutation value with 10^4 resamples. The unrelated-item control estimates generic replacement effects. The causal weight cw measures this layer-specific intervention contrast in nats.

Linear decodability. A four-way logistic probe predicts the relation from pooled vision residuals $\bar{h}_\ell = |V|^{-1}\sum_{t\in V} h_{\ell,t}$, using standardized features, $C = 1$, and five pair-grouped folds (Alain & Bengio, 2017; Pedregosa et al., 2011). The reported accuracy is compared with the nominal four-class chance level .25 and a fixed .50 decision threshold. It measures relation recovery from the pooled representation on the pair-grouped evaluation folds.

Elicitation and payload forms. Into a neutral gray-canvas prompt we add $\pm\alpha\rho_\ell u$ at selected positions, where ρ_ℓ is the mean residual norm. Three forms test distinct possible encodings: a pooled class-contrast direction, a position-preserving class-contrast pattern, and a class direction from the linear probe transformed back to activation coordinates. The first compresses the relation into one vector; the second retains token-position structure; the third tests a direction optimized for classification. Color payloads and token-derived directions serve as positive controls. All base-model weights remain frozen; only the probes are fitted.

Signed response and dose selection. We score a generation g as $o(g) = +1$ for target-only relation words, -1 for opposite-only words, and 0 for neither. The direction differential is

$$\Delta_\ell(\alpha) = \frac{1}{N'}\sum_i \frac{o_i^+(\alpha) - o_i^-(\alpha)}{2}, \qquad \alpha^* \in \arg\max_\alpha \Delta^{\text{color}}(\alpha). \tag{3}$$

A direction-independent scored response cancels. The zero null concerns the expected signed difference under no directional effect; a finite sample can still vary. Color-based dose selection keeps spatial outcomes out of calibration. It verifies an effective dose for that control, while leaving the spatial content's potentially different dose dependence to the sweep.

Umbra band and Umbra Index. Let s^{pool} and s^{pat} denote their raw elicitation success rates. Define

$$\mathcal{M} = \{\ell : \text{acc}_\ell \geq .50,\ s^{\text{pool}}_\ell \leq .35,\ s^{\text{pat}}_\ell \leq .35\}, \quad \text{UI} = \frac{\sum_{\ell\in\mathcal{C}\cap\mathcal{M}} \text{cw}_\ell}{\sum_{\ell\in\mathcal{C}} \text{cw}_\ell}. \tag{4}$$

Within this phenomenon, an *Umbra band* comprises at least three consecutive tested grid layers in $\mathcal{C} \cap \mathcal{M}$. The index summarizes the share of summed layerwise interchange effects meeting these thresholds. Raw success rates define band membership; calibrated Δ measures directional response at the evaluated doses. We report both quantities because they answer different questions about elicitation.

Table 1: **Three measurements across the Qwen2.5-VL-3B layer grid.** Causal weights and probe accuracies are from Table 3 (23 valid pairs; 46 probe items). Injection columns are the original success-rate summaries; calibrated Δ is reported separately. Ranges refer to tested layers.

Region	cw (nats)	probe	pooled	pattern	classification
L0–L13	3.1–3.9	.445–.490	0.00	0.00–0.33	causal, sub-threshold probe
L14	3.27	.510	0.00	**0.50**	excluded by pattern rule
L16	3.51	.470	0.00	0.33	excluded by probe
L17–L24	4.35→0.39	**.510–.575**	**0.00**	0.00–0.17	**Umbra band**
L26–L30	≤0.19	.530–**.640**	–	–	below causal threshold
L31–L34	≈0	.465–.485	–	–	below threshold

3 THE UMBRA EFFECT: CAUSAL USE WITH WEAK ELICITATION

In Qwen2.5-VL-3B, spatial relations influence answers throughout the early and middle stack, but the pooled payload does not elicit them in the injection test. Six tested layers, L17, L18, L20, L21, L23, and L24, also cross the decodability threshold. This intersection is the detailed three-readout result.

Compare causal use, decodability, and elicitation at matched layers. Table 1 summarizes the full layer table. In the band, causal weights decrease from 4.349 nats at L17 to .387 at L24; probe accuracy ranges from .510 to .575. Pooled-injection success is 0 throughout the band, while position-preserving success ranges from 0 to .17. Color success is 1.00 at all 18 injected layers. The color contrast shows that generation can respond to content on the same vision-position pathway.

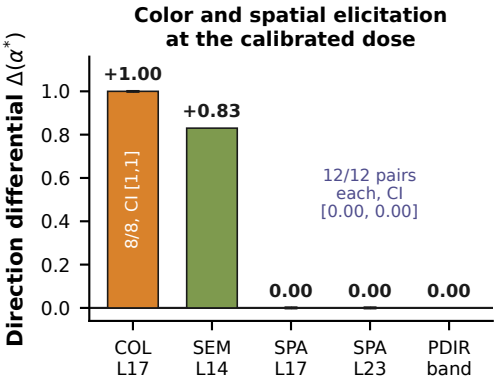


Figure 2: **Content-dependent elicitation on Qwen2.5-VL-3B.** At the color-selected dose $\alpha^* = 1$, compare color (COL, 8 pairs) with spatial pooled payloads (SPA, 12 pairs at each layer): $\Delta = 1$ versus 0. SEM denotes the semantic control; PDIR denotes probe-direction payloads. Intervals are empirical pair-bootstrap summaries of the sampled scores.

Figure 2 then compares signed responses at the color-selected dose, $\alpha^* = 1$. Color at L17 has $\Delta = 1.000$ on eight pairs; spatial pooled payloads at L17 and L23 have $\Delta = 0.000$ on 12 pairs each. Pair-resampling yields intervals $[1, 1]$ and $[0, 0]$ because the observed scores are constant within these small samples. A wider band check reports one pooled $+.08$ cell at L21 and otherwise zero responses; probe-direction Δ is zero at all six band layers.

Payload, site, and lexical comparisons. The probe-direction test also reports zero raw hits across the six band layers, four relations, and doses $\{1, 2, 4, 8\}$. Together with the position-preserving arm, this broadens the set of unsuccessful spatial edits beyond one pooled vector. The probe-direction arm tests an additional payload optimized for classification, alongside the class-contrast payloads.

The site atlas (Table 4) compares vision positions, post-image text positions, and the final prompt position. Spatial pooled and probe-direction rates are zero at all tested sites. Color reaches 1.00 at vision positions and at post-image text positions in L17–L18, making these settings informative content contrasts. Final-position color succeeds at most .17, so that site's spatial failure provides a weaker comparison. The site and payload results belong together: each specifies a different route by which a relation representation might affect generation.

Diagonal relations extend the test beyond single-token cardinal answers. Their reported index is .165, with a band at L18, L20, and L21 and zero pooled injection throughout the causal region (Table 6). The dissociation thus recurs across the two tested answer formats, directing attention to the relation representation and recipient context.

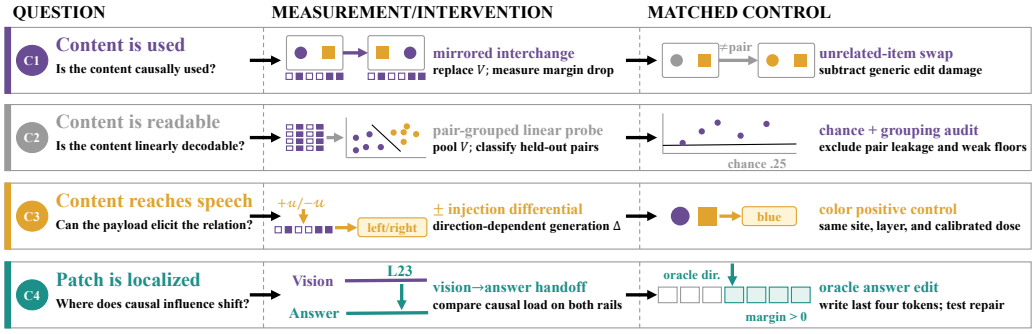


Figure 3: **Interventions and their controls.** Read each row as a question, its measurement, and a matched control. Rows C1–C3 implement the three readouts; C4 localizes where an intervention changes the answer margin. The unrelated-item swap controls for generic replacement effects; pair grouping keeps mirrored partners in the same probe fold; color tests whether an injection setting changes generation; the oracle answer-window edit tests downstream responsiveness. Each control addresses one source of ambiguity.

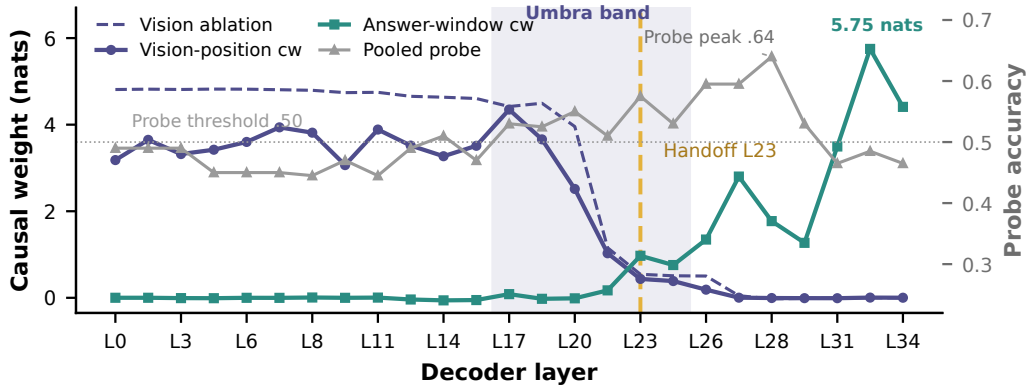


Figure 4: **Vision and answer positions exchange their relative influence.** On Qwen2.5-VL-3B, compare vision interchange (indigo), vision mean-ablation (dashed), and answer-window interchange (teal). The latter exceeds vision interchange at L23 and reaches 5.750 nats at L33. Probe accuracy (gray, right axis) peaks at L28, after vision interchange falls below the causal threshold. Table 7 gives the component values for this descriptive overlay.

Stability of the measured contrast. The six primary band permutation values remain below .05 after Bonferroni correction across 25 layers. A separate control-seed repeat gives positive bootstrap intervals for all six causal weights, and three stimulus seeds give positive band effects (Appendix E). The repeat's index, .2412 with interval [.2066,.2809], is distinct from the primary .2268. Threshold sensitivity is substantial: changing the probe cutoff from .50 to .45 or .55 changes the reported index to .7989 or .0541. The continuous measurements therefore carry more information than the band label alone.

4 Localizing the Gap: A Vision-to-Answer Handoff

The 3B band overlaps a transition in where interchange affects the answer margin. At L23, answer-window interchange first exceeds vision interchange; later, the pooled vision probe peaks while vision interchange is small. Figure 4 places these measurements on the same layer axis to distinguish where a relation remains readable from where the tested replacement affects the answer.

Read the crossing and the late probe peak together. Vision mean-ablation drops from 3.957 nats at L20 to 1.156 at L21 and .543 at L23. Answer-window interchange is close to zero through L20, rises to .971 at L23 ($p = .0057$), and reaches 5.750 at L33 ($p = .0002$). We define the vision-to-answer handoff as $\ell^* = \min\{\ell : \mathrm{cw}_\ell^A > \mathrm{cw}_\ell\}$; its observed value is L23. Pooled probe accuracy peaks at .640 at L28, where vision interchange is $-.005$. The layer profiles support a shift in intervention sensitivity toward answer positions, with decodable information persisting at vision positions after their direct effect has diminished.

This localization separates a probe's recovery of information from the sites at which interchange changes the answer. Interchange acts on the model's existing image-question state; injection tests transfer into a neutral-image recipient. The fact that these tests disagree is consistent with context- or format-dependent access. The handoff localizes this change at the level of token-position groups.

An answer-window diagnostic with the correct label supplied. On nine baseline-failed pairs, the reported band intervention summary is 8.33 nats versus .85 in the early comparison. Injecting the true-relation probe direction at the last four prompt positions, at layers in $\{23, 26, 28\}$, is reported to recover the answer margin on all evaluated failed items. This experiment asks whether the downstream answer state can respond when the relation is supplied. The outcome is recovery of the correct-versus-opposite token margin under a supplied label, motivating a comparison of intervention sites.

The attempted automatic alternatives are less effective: the failed-item probe has accuracy approximately .33, a band-probe arbitration baseline is net negative, and handoff amplification recovers no failures. Their contrast with the oracle intervention identifies label selection as a practical obstacle.

Position structure has a limited role in the 3B case. Intact patterns outperform shuffled patterns in some small-sample cells, but the repeat with 12 pairs and three seeds reduces the differences to .25 versus .14 at L14 and .33 versus .25 at L16. A per-position PCA probe improves over pooling by only .006 on average across L0–L13. These results provide little support for pooling alone explaining the early probe deficit. The cross-family experiment below tests the same payload distinction where its effect is larger.

5 Cross-model recurrence and payload-dependent elicitation

The four Qwen checkpoints share weak spatial pooled-direction elicitation in the tested calibration cells. They differ in decodability. LLaVA differs in the response to position-preserving payloads. These are distinct axes of variation, and they determine which part of the three-readout result transfers.

Within Qwen, decodability changes the band. On Qwen2-VL-2B, every tested causal layer falls below the .50 probe cutoff, giving $\mathrm{UI} = 0$ despite zero pooled injection in the causal region. Qwen2.5-VL-3B has the mid-stack band described above. Qwen2.5-VL-7B and Qwen2-VL-7B have reported indices 1.000 and .9506, respectively, with broad overlap between causal effects and decodability. The band therefore distinguishes how much of each checkpoint's tested causal region also meets the probe and elicitation thresholds.

Qwen2.5-VL-7B vision effects fall below the causal criterion by L16, earlier in relative depth than in the 3B model. Its band occupies the tested causal region through L15. This comparison extends the layerwise analysis of vision-position influence to 7B. The matched vision/answer-position crossing is measured in the 3B study.

LLaVA exposes payload-dependent elicitation. In LLaVA-OneVision, the position-preserving arm often yields .58–.83 success while the pooled arm is zero, with .08 pooled success at L0 (Table 12). A focused six-pair comparison gives pattern success .83 at L10 and .67 at L13, against pooled success 0; the recorded generations are coherent at doses 1–2. This is the clearest family boundary: a negative pooled edit can coexist with a successful structured edit.

Content type changes the reported overlap. Figure 5A compares semantic, spatial, and counting tasks on Qwen2.5-VL-3B. The reported indices are 0, .2268, and .3345. Semantic causal weights reach 6.797 nats and raw injection succeeds at .42–.75 over most of the causal region. Counting probes are stronger (.775–.975), while pooled injection declines to zero at many later layers. The continuous measurements distinguish the task conditions: semantic payloads often elicit their content,

Table 2: **Model comparisons separate pooled elicitation from the full band criterion.** Indices use the reported raw-rate criterion. Calibration entries summarize tested cells. LLaVA is compared using its pooled and position-preserving responses; its index is omitted.

Model	Reported UI	Color Δ	Main contrast
Qwen2.5-VL-3B	.2268	1.00	mid-stack band
Qwen2.5-VL-7B	1.000	1.00	broad band
Qwen2-VL-7B	.9506	1.00	broad band
Qwen2-VL-2B	.0000	.875	probe below .50
LLaVA-OneVision-7B	–	1.00	pattern responds

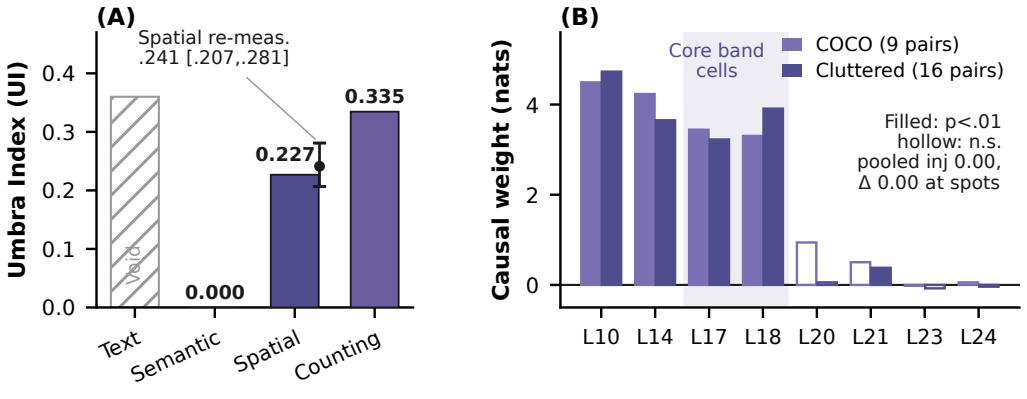


Figure 5: **Task and image conditions on Qwen2.5-VL-3B.** (A) Reported UI values are semantic 0.000, spatial 0.2268, and counting 0.3345. The spatial bootstrap interval belongs to a separate control-seed repeat (point 0.2412). (B) Compare causal weights on COCO photographs (9 valid pairs) and cluttered scenes (16). Hollow bars fail $p < .01$. At L10–L18, photo pooled-injection success is 0; photo probe accuracy is .40–.55. These data test image-domain transfer on one model.

whereas counting retains strong decodability as pooled elicitation weakens. A text-only control is excluded from the comparison because its relation label is present verbatim in the input.

Image-domain transfer has a narrower evidence scope. On nine baseline-valid COCO mirrored photo pairs, vision interchange is 3.304–4.491 nats at L10, L14, L17, and L18, with $p < .01$ and zero pooled injection. The photo probe ranges from .40 to .55 over these layers, so this arm supports the use–elicitation comparison without establishing a stable Umbra band. Sixteen cluttered synthetic pairs show the same contrast at L14, L17, and L18; L10 has .08 pooled success. Photo Δ spot checks are zero, whereas clutter checks include $-.10$. These tests extend the image conditions evaluated on Qwen2.5-VL-3B; the checkpoint comparison separately tests model variation.

6 Calibration determines what a negative readout means

A negative response is informative only relative to an operational measurement. The injection controls identify dose and output-scoring effects; the J-lens experiment supplies a separate example of a readout that fails before it can support a representation-level conclusion.

Separate baseline preference from directional response. An open-ended success rate depends on which words the model prefers to generate. A forced-choice prompt changes that baseline again: naming two options can yield roughly .5 target success with no useful response to the payload. Signed Δ subtracts responses to opposite payloads, removing a direction-independent scored tendency. Its estimand is the change in the signed relation-word score. Together with raw rates, it distinguishes an unchanged preferred answer from an answer that tracks the sign of an edit.

Compare doses while monitoring the control. Larger edits can reverse or collapse the color response. The cross-model dose curves (Figure 6) therefore report individual doses, with effective

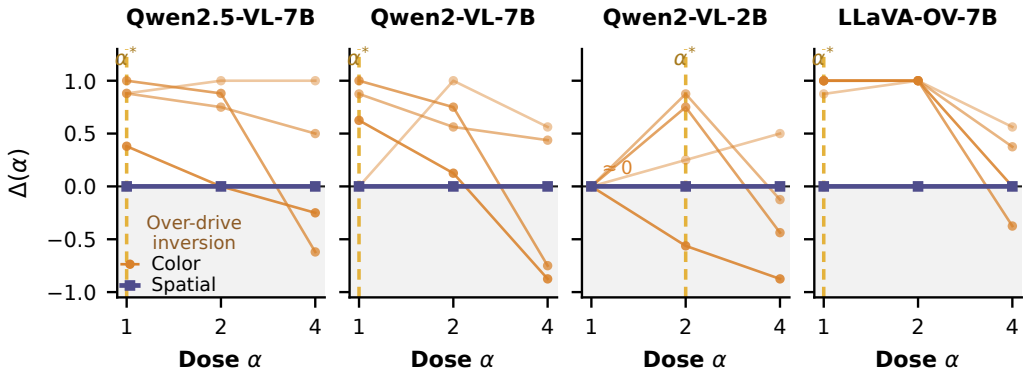


Figure 6: **Per-dose color calibration grids across four models** (Qwen2.5-VL-7B (Bai et al., 2025), Qwen2-VL-7B/2B (Wang et al., 2024), LLaVA-OneVision-7B (Li et al., 2025); per-layer thin curves). The dose α has an effective window: color peaks at a selected setting (gold), with negative responses at some higher-dose layers. The plotted spatial summaries (indigo) are near zero; Table 12 separately reports nonzero LLaVA measurements at selected layers. The 2B needs α^*=2; the 3B grid gives α^*=1 with color +1.00 at both calibration layers and inversion to −0.88 at $\alpha \geq 3$.

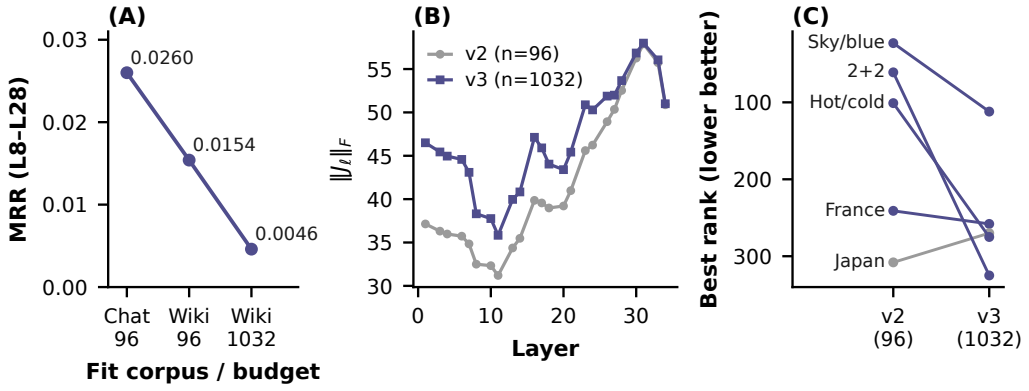


Figure 7: **The evaluated J-lens configurations fail the fact-readout check.** (A) Mean reciprocal rank (MRR) across L8–L28 is .0260, .0154, and .0046 for the three fits. The first two fits both use 96 prompts and differ in corpus and skip setting; only the last comparison increases the WikiText fit to 1032 prompts. (B) Jacobian norms retain a similar layer profile. (C) Four of five best fact ranks worsen from the second to the third fit. The comparison holds the target decoder and five-fact battery fixed.

color settings at $\alpha = 1$ or 2 and adverse responses at larger tested doses. Spatial performance is evaluated at the selected setting and inspected across the sweep. A spatial excursion at $\alpha = 6$ occurred where controls were already inverted. This setting provides a different perturbation condition from the one supporting the main comparison. The control establishes a responsive pathway at the selected dose, while the sweep tests how each payload behaves away from that setting.

7 Validating J-lens readouts before interpretation

The evaluated J-lens fits do not recover known fact tokens in the chosen sanity test. This is an instrument result on Qwen2.5-VL-3B, separate from the interchange, probe, and injection results above. The Jacobian lens maps a layer's activations through a prompt-averaged linearization before applying the model's output readout (Gurnee et al., 2026).

Across three fits, none of five model-known fact targets appears in the top eight over L8–L28. Mean reciprocal rank decreases from .0260 for a 96-prompt chat-templated fit to .0154 for a 96-prompt WikiText-based fit, then .0046 for a 1032-prompt fit using the latter corpus family (Merity et al., 2017). The largest fit costs 7.05 GPU-hours. The first comparison changes corpus and skip setting at fixed prompt count; the second increases the prompt count within the WikiText configuration. Table 13 lists the settings for each fit.

Figure 7 separates score, norm profile, and per-fact rank. A stable U-shaped Jacobian norm profile shows similar aggregate layer structure across budgets, while four of five best fact ranks worsen from the second to third fit. The WikiText fits also contain non-finite layer-0 rows, and L0 is excluded from the descriptive readouts (Appendix J). Norm profiles and fact recovery therefore supply complementary checks on the fitted readout. The usable conclusion is procedural: verify a lens on known content in the target decoder before interpreting failed token recovery as absence of a verbalizable representation.

8 Related work and implications

Decoding and causal use. Linear probes identify recoverable information, with controls needed to distinguish task structure from probe capacity (Alain & Bengio, 2017; Hewitt & Liang, 2019; Belinkov, 2022). Amnesic probing links representation removal to behavior (Elazar et al., 2021). Causal mediation, abstraction, and activation patching provide tools for testing internal contributions (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022; Wang et al., 2023; Heimersheim & Nanda, 2024). We characterize the Umbra effect using a mirrored interchange contrast, a separately trained probe, and a separately specified recipient-context injection, so each result answers a distinct question.

Verbal report and latent knowledge. Workspace theories emphasize access and broadcasting (Baars, 1988; Dehaene & Naccache, 2001; VanRullen & Kanai, 2021; Goyal et al., 2022; Butlin et al., 2023). The J-lens studies verbalizable representations in language models (Gurnee et al., 2026); logit and tuned lenses provide related token readouts (nostalgebraist, 2020; Belrose et al., 2023). Latent-knowledge and introspection studies examine information present in internal states or self-predictions (Kadavath et al., 2022; Azaria & Mitchell, 2023; Orgad et al., 2025; Binder et al., 2025). Our study tests a complementary operational question: how visual-relation payloads respond to calibrated interventions across causal use, decodability, and elicitation.

Visual information flow and steering. Prior work studies multimodal neurons (Goh et al., 2021; Schwettmann et al., 2023), visual information flow (Neo et al., 2025; Basu et al., 2024; Palit et al., 2023; Kaduri et al., 2025; Zhang et al., 2025c), and representation edits that reduce VLM hallucinations (Jiang et al., 2025). Activation addition and representation engineering study how vector interventions steer generation (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023; Li et al., 2023). Our handoff connects position-level interchange sensitivity to the layerwise probe profile, and the LLaVA comparison shows why payload form belongs in an elicitation audit. Spatial and perceptual benchmarks motivate the task domain (Tong et al., 2024; Rahmanzadehgervi et al., 2024; Zhang et al., 2025b); the present evaluation uses controlled relation stimuli and a small COCO subset.

9 Conclusion

The Umbra effect is a dissociation between the causal use of visual information and its elicitation under a specified intervention. Four Qwen checkpoints exhibit use–elicitation dissociation: weak spatial pooled-direction elicitation in tested settings where interchange affects answers; Qwen2.5-VL-3B additionally provides an Umbra band. Its vision-to-answer handoff occurs at L23. LLaVA's payload-dependent elicitation establishes a concrete boundary on generalization. For representation analysis, the resulting practice is to report causal contribution, probe accuracy, and controlled elicitation together, keeping model, payload, recipient context, and dose explicit.

Scope and limitations. The main band is selected on initially correct mirrored pairs and depends on fixed probe and injection thresholds. Cross-model evidence covers five checkpoints; image-domain transfer is confined to the 3B case study and small samples. The oracle edit supplies the true label and measures a margin.

Reproducibility Statement

Section 2 defines the interventions and measurements, and Appendices A–J contain the reported layerwise results. Appendix L records stimulus seeds, hashes, and compute settings. The accompanying manuscript source includes the original figure assets and plotting scripts.

Ethics Statement

The study analyzes open-weight models using synthetic geometric stimuli and a small COCO image subset. It introduces no human-subject intervention. Original image rights and model license conditions apply to any redistribution. The answer-window intervention is a diagnostic using supplied relation labels. Implications for practical auditing are limited to the measured models and intervention settings.

AI Use Statement

Generative AI was used to assist with manuscript drafting and editing, finding and checking related literature, and writing code for the experiments.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. URL https://aclanthology.org/2023.findings-emnlp.68/.

Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

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.

Yonatan Belinkov. Probing classifiers: promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: language models can learn about themselves by introspection. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0a6059857ae5c82ea9726ee9282a7145-Abstract-Conference.html.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, et al. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023. URL https://arxiv.org/abs/2308.08708.

Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2):1–37, 2001.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics (TACL)*, 9:160–175, 2021.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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. URL https://distill.pub/2021/multimodal-neurons/.

Anirudh Goyal, Aniket Didolkar, Alex Lamb, Kartikeya Badola, Nan Rosemary Ke, Nasim Rahaman, Jonathan Binas, Charles Blundell, Michael Mozer, and Yoshua Bengio. Coordination among neural modules through a shared global workspace. In *International Conference on Learning Representations (ICLR)*, 2022.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://aclanthology.org/D19-1275/.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2411.17491.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: easy visual task transfer. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=zKv8qULV6n.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics (TACL)*, 11:635–651, 2023.

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.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

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.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`, 2020.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: on the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: a causal tracing tool for BLIP. In *ICCV Workshop on Closing the Loop Between Vision and Language*, 2023.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Asian Conference on Computer Vision (ACCV)*, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Association for Computational Linguistics (ACL)*, 2024.

Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. Multimodal neurons in pretrained text-only transformers. In *ICCV Workshop on Closing the Loop Between Vision and Language*, 2023.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL `https://arxiv.org/abs/2308.10248`.

Rufin VanRullen and Ryota Kanai. Deep learning and the global workspace theory. *Trends in Neurosciences*, 44(9):692–704, 2021.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. URL `https://arxiv.org/abs/2409.12191`.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: training-free perception of small visual details with multimodal LLMs. In *International Conference on Learning Representations (ICLR)*, 2025a.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. SPHERE: unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11591–11609, 2025b. doi: 10.18653/v1/2025.acl-long.568. URL https://aclanthology.org/2025.acl-long.568/.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c. URL https://arxiv.org/abs/2411.18620.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: a top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

APPENDIX

A BAND MASTER TABLE

Table 3: Full 25-layer table for Qwen2.5-VL-3B (Bai et al., 2025) spatial mirrored pairs (23/32 baseline-valid pairs; probe over 46 items). cw in nats; p from the paired permutation test; inj/pat are pooled and pattern injection rates; col/tok are the color and token-ceiling controls; probe is 4-way pooled accuracy (chance .25). Bold rows: the band. n.i.: not injected (non-causal).

L	cw	p	inj	pat	col	tok	probe
0	3.182	.0001	0.00	0.00	1.00	0.00	.490
1	3.649	.0001	0.00	0.00	1.00	1.00	.490
3	3.321	.0002	0.00	0.00	1.00	1.00	.490
4	3.421	.0001	0.00	0.00	1.00	1.00	.450
6	3.601	.0001	0.00	0.00	1.00	1.00	.450
7	3.938	.0001	0.00	0.00	1.00	0.00	.450
8	3.815	.0001	0.00	0.00	1.00	1.00	.445
10	3.066	.0001	0.00	0.33	1.00	1.00	.470
11	3.887	.0001	0.00	0.17	1.00	1.00	.445
13	3.517	.0001	0.00	0.00	1.00	1.00	.490
14	3.271	.0001	0.00	0.50	1.00	1.00	.510
16	3.512	.0001	0.00	0.33	1.00	1.00	.470
17	4.349	.0001	0.00	0.00	1.00	0.00	.530
18	3.660	.0001	0.00	0.00	1.00	0.00	.525
20	2.515	.0001	0.00	0.17	1.00	1.00	.550
21	1.029	.0001	0.00	0.00	1.00	0.00	.510
23	0.436	.0003	0.00	0.17	1.00	0.00	.575
24	0.387	.0004	0.00	0.00	1.00	0.00	.530
26	0.188	.0041	n.i.	–	–	–	.595
27	0.003	.4739	n.i.	–	–	–	.595
28	−0.005	.7171	n.i.	–	–	–	.640
30	−0.008	.8296	n.i.	–	–	–	.530
31	−0.009	.9572	n.i.	–	–	–	.465
33	0.005	.0609	n.i.	–	–	–	.485
34	0.002	.2141	n.i.	–	–	–	.465

Injection sample sizes: pooled 16 pairs $\times$ $\alpha\{1, 2, 4\}$; pattern 6 pairs $\times$ $\alpha\{1, 2, 4, 8\}$; color and token controls 8 pairs $\times$ $\alpha\{1, 2, 4, 8\}$. Causal weights and p values are identical across two independent instrumentation runs.

B INJECTION-SITE ATLAS

Table 4: Qwen2.5-VL-3B raw elicitation rates: spatial pooled contrast (sp), probe direction (pd; band only), and color (col). Color is weakest at the final position and reaches 1.00 at post-image text positions in L17–L18.

	vision			post-image text			final		
L	sp	pd	col	sp	pd	col	sp	pd	col
10	0.00	–	1.00	0.00	–	0.17	0.00	–	0.00
14	0.00	–	1.00	0.00	–	0.67	0.00	–	0.00
17	0.00	0.00	1.00	0.00	0.00	**1.00**	0.00	0.00	0.00
18	0.00	0.00	1.00	0.00	0.00	**1.00**	0.00	0.00	0.00
20	0.00	0.00	1.00	0.00	0.00	0.67	0.00	0.00	0.00
21	0.00	0.00	1.00	0.00	0.00	0.67	0.00	0.00	0.00
23	0.00	0.00	1.00	0.00	0.00	0.67	0.00	0.00	0.17
24	0.00	0.00	1.00	0.00	0.00	0.50	0.00	0.00	0.00
best	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.17

C POSITIONAL-CODE ANATOMY

Table 5: Injection-arm anatomy (6 pairs, $\alpha\{1, 2, 4, 8\}$): intact vs spatially shuffled pattern, object-only vs background-only vs random-k positions, and the color-pattern control. The intact-vs-shuffled advantage at L14 was smaller in the expanded repeat (12 pairs $\times$ 3 seeds: L14 intact 0.25 vs shuffled 0.14; L16 0.33 vs 0.25) and is reported as a descriptive comparison.

L	intact	shuffled	object	background	random	color-pattern
10	0.33	0.50	0.00	0.00	0.00	1.00
11	0.17	0.17	0.00	0.00	0.00	1.00
14	0.50	0.17	0.33	0.00	0.17	1.00
16	0.33	0.00	0.17	0.00	0.17	1.00
17	0.00	0.00	0.17	0.00	0.00	1.00
20	0.17	0.00	0.00	0.00	0.00	1.00
23	0.17	0.00	0.00	0.00	0.17	1.00

Probe arm (per-position PCA-40 vs pooled, 46 items, GroupKFold): the mean per-position minus pooled gap over L0–L13 is 0.006, providing little evidence that pooling explains the early probe deficit; in the late stack pooled exceeds per-position (L23 .575 vs .535; L26–L28 .595–.640 vs .51–.53), consistent with consolidation into pooled form.

D LEXICAL-CONFOUND ARM (DIAGONAL RELATIONS)

Table 6: Diagonal-relation diagnostic (13/16 baseline-valid pairs). The causal band ends earlier (L21) with roughly half the causal weight of the cardinal arm; pooled injection is 0.00 at every causal layer, matching the cardinal arm exactly. UI=0.165, band [18,20,21]; diagonal relations are pooled-decodable only at L18–L21 (peak .567).

L	cw	p	inj	L	cw	p	inj
0	1.370	.0001	0.00	16	2.020	.0001	0.00
1	1.647	.0007	0.00	17	1.255	.0004	0.00
3	2.085	.0001	0.00	18	1.734	.0005	0.00
4	1.391	.0007	0.00	20	1.715	.0001	0.00
6	2.075	.0001	0.00	21	1.011	.0001	0.00
7	1.711	.0001	0.00	23	−0.057	.6996	n.i.
8	1.763	.0001	0.00	24	0.088	.1487	n.i.
10	1.729	.0003	0.00	26	0.139	.0084	n.i.

L	cw	p	inj	L	cw	p	inj
11	1.929	.0001	0.00	27–31	$\leq$0.049	$>$.08	n.i.
13	2.109	.0001	0.00	33/34	.001/−.003	.50/.94	n.i.
14	1.492	.0001	0.00				

E STATISTICS BATTERY

Causal-weight CIs (2000 pair-resample bootstrap, 23 pairs). L17 $+3.744$ $[+2.446, +5.177]$; L18 $+3.683$ $[+2.424, +5.047]$; L20 $+2.868$ $[+1.731, +3.976]$; L21 $+1.167$ $[+0.893, +1.457]$; L23 $+0.392$ $[+0.248, +0.548]$; L24 $+0.329$ $[+0.202, +0.466]$. All exclude zero. Band permutation p-values (Table 3) are at most 4×10^{-4}, so all six remain below .05 after Bonferroni correction over the 25-layer grid. The sampled Δ scores are constant, yielding zero-width empirical bootstrap intervals.

Δ key cells (degenerate CIs, zero variance). Color L17 $+1.000$ $[+1.000, +1.000]$ (8/8 pairs); spatial L17 and L23 both 0.000 $[0.000, 0.000]$ (12/12 pairs each).

UI. Primary reported value $0.2268 = 12.376/54.556$ nats (the index numerator selects exactly the contiguous Umbra band, so the band-restricted index coincides); Separate control-seed repeat per-pair point 0.2412, CI $[0.2066, 0.2809]$ (classifications fixed, so the CI covers causal-weight sampling noise only). Probe-threshold sensitivity, pre-declared: $\text{probe_min} = 0.45 \rightarrow$ UI 0.7989; $0.50 \rightarrow 0.2268$ (operative); $0.55 \rightarrow 0.0541$; the spread is structural because early-layer accuracies cluster at .445–.490, and the per-position analysis (Table 5) gives a similarly weak early probe.

Multi-seed controls. Band causal weights across three stimulus seeds (order L17/18/20/21/23/24): 3.26/3.24/2.30/0.96/0.36/0.31; 4.09/4.03/3.25/1.14/0.34/0.28; 3.86/3.82/3.12/1.16/0.38/0.30, all inside the CIs above.

Probe-direction injection. Injecting the probe's raw-space class directions at all six band layers, $\alpha \in \{1, 2, 4, 8\}$, four relations each: rate 0.00 at every layer; per-relation hit counts {left:0, right:0, above:0, below:0} at each of L17/18/20/21/23/24.

Signed responses at the calibrated dose. $\Delta(\alpha^*{=}1)$, 12 pairs: pooled direction 0.00 at seven layers with a single $+0.08$ cell at L21; probe direction exactly 0.00 at all six band layers.

F HANDOFF FULL TABLE

Table 7: Component measurements for Figure 4: vision-position interchange weight, answer-window (last-four-position) interchange weight with its permutation p, vision mean-ablation drop, and pooled probe accuracy. Vision weights are reproduced from the 23-pair master table; the answer-window comparison is reported for 12 pairs. The first layer where the answer-window weight exceeds the vision weight is L23.

L	cw^V	cw^A	p^A	abl	probe	L	cw^V	cw^A	p^A	abl	probe
0	3.182	+0.001	.342	4.812	.490	17	4.349	+0.085	.231	4.419	.530
1	3.649	+0.001	.500	4.820	.490	18	3.660	−0.023	.565	4.495	.525
3	3.321	−0.007	.740	4.813	.490	20	2.515	−0.012	.528	3.957	.550
4	3.421	−0.009	.815	4.822	.450	21	1.029	+0.172	.206	1.156	.510
6	3.601	0.000	.479	4.820	.450	**23**	0.436	**+0.971**	**.0057**	0.543	.575
7	3.938	−0.001	.510	4.807	.450	24	0.387	+0.757	.0164	0.509	.530
8	3.815	+0.007	.326	4.794	.445	26	0.188	+1.346	.122	0.503	.595
10	3.066	−0.000	.504	4.740	.470	27	0.003	+2.801	.0122	0.049	.595
11	3.887	+0.005	.420	4.747	.445	28	−0.005	+1.772	.0487	−0.025	.640
13	3.517	−0.039	.815	4.655	.490	30	−0.008	+1.270	.180	−0.014	.530
14	3.271	−0.059	.936	4.634	.510	31	−0.009	+3.496	.0038	−0.005	.465
16	3.512	−0.052	.676	4.604	.470	33	0.005	**+5.750**	**.0002**	0.001	.485
						34	0.002	+4.411	.0007	−0.001	.465

G TASK-TYPE PER-LAYER TABLES

Table 8: Semantic and counting arms under the identical criterion (per-layer cw in nats, pooled injection rate, pooled probe accuracy). Semantic (24/24 valid): large causal weights with injection succeeding broadly, UI=0.000. Counting (20/24 valid): probe at .78–.975, pooled injection decaying to 0.00 from L23, band [14,16,17], UI=0.3345. The text arm (24/24 valid) has probe $\equiv 1.000$ at all 25 layers because the label appears verbatim in the input (§6).

	semantic			counting		
L	cw	inj	probe	cw	inj	probe
0	5.701	.50	.645	4.017	.42	.825
1	5.590	.50	.645	3.134	.33	.825
3	5.132	.58	.645	3.771	.33	.800
4	5.090	.67	.645	3.641	.58	.775
6	4.663	.67	.625	2.680	.33	.800
7	4.316	.67	.625	3.159	.42	.800
8	4.738	.75	.625	2.969	.42	.800
10	4.757	.58	.625	3.570	.42	.800
11	5.818	.67	.605	4.605	.25	.825
13	5.578	.58	.605	2.741	.42	.850
14	6.399	.67	.520	2.874	.33	.800
16	6.253	.50	.540	3.309	.33	.800
17	6.778	.50	.520	3.232	.25	.800
18	6.797	.50	.520	3.259	.42	.825
20	5.358	.50	.560	3.681	.17	.800
21	5.106	.58	.560	3.717	.08	.825
23	6.110	.42	.585	3.347	.00	.850
24	5.239	.42	.625	2.063	.00	.825
26	5.425	.42	.645	2.079	.00	.950
27	4.427	.42	.620	0.778	.00	.975
28	4.208	.50	.620	0.903	.00	.975
30	3.507	.58	.620	0.711	.00	.975
31	1.710	.50	.660	0.559	.00	.975
33	0.008	1.00	.680	0.450	.00	.975
34	0.012	1.00	.705	0.314	.17	.975

Gradient stimuli were frozen separately (seed and hashes in §L); baseline validity: text 24/24, semantic 24/24, counting 20/24, with each arm evaluated on its own valid subset.

H NATURAL-IMAGE TABLES

Table 9: Beyond-clean-canvas arms (cw/p/probe/pooled inj). COCO (Lin et al., 2014): 9/16 valid mirrored photo pairs via horizontal-flip construction; Δ spot checks at L17/L20/L23 all 0.00. Cluttered synthetic: 16/16 valid; Δ spots 0.00/0.00/$-$0.10. The photo probe (.40–.55, 18 items, 4-way) is underpowered; the decodability leg is not claimed on photos.

	COCO photos				cluttered synthetic			
L	cw	p	probe	inj	cw	p	probe	inj
10	4.491	.0013	.550	0.00	4.729	.0002	.867	.08
14	4.233	.0013	.400	0.00	3.651	.0003	.900	0.00
17	3.441	.0013	.550	0.00	3.225	.0002	.967	0.00
18	3.304	.0055	.450	0.00	3.909	.0001	.967	0.00
20	0.941	.0678	.500	1.00	0.056	.44	1.000	1.00
21	0.503	.0242	.450	1.00	0.379	.0046	1.000	0.00
23	-0.019	.56	.500	–	-0.076	.88	–	–
24	0.056	.32	.550	–	-0.039	.72	–	–

The COCO analysis uses the nine baseline-valid pairs listed in this table.

I CROSS-FAMILY PER-LAYER TABLES

Table 10: Qwen2.5-VL-7B (Bai et al., 2025) (32 baseline-valid pairs; cw/p from 12-pair interchange; probe over all 64 items; injections at causal layers only). Probe $\geq .533$ at every causal layer, pooled injection 0.00 and pattern $\leq .25$ everywhere causal: the Umbra band covers the entire causal region, UI=1.0. Color grid: α=1 [L0 +0.88, L3 +0.88, L6 +1.00, L10 +0.38]; α=2 [+1.00, +0.75, +0.88, 0.00]; α=4 [+1.00, +0.50, −0.62, −0.25]; α^*=1. Spatial Δ at α^*: 0.00 at L0/L1/L2 and L13/L14/L15.

L	cw	p	probe	inj	pat	L	cw	p	probe	inj	pat
0	3.732	.0002	.550	0.00	0.00	13	2.932	.0002	.533	0.00	0.17
1	3.136	.0002	.550	0.00	0.00	14	4.533	.0002	.533	0.00	0.17
2	3.475	.0005	.550	0.00	0.08	15	1.717	.0088	.579	0.00	0.17
3	3.903	.0006	.564	0.00	0.17	16	0.260	.0604	.579	n.c.	n.c.
4	3.092	.0008	.567	0.00	0.17	17	0.187	.0566	.581	n.c.	n.c.
5	3.135	.0015	.612	0.00	0.00	18	0.104	.1197	.600	n.c.	n.c.
6	3.365	.0008	.579	0.00	0.08	20	0.032	.0856	.552	n.c.	n.c.
8	3.795	.0005	.581	0.00	0.25	21	0.007	.3625	.550	n.c.	n.c.
9	4.019	.0016	.579	0.00	0.17	22	−0.017	.9750	.519	n.c.	n.c.
10	3.563	.0002	.564	0.00	0.17	23	0.003	.3577	.498	n.c.	n.c.
11	4.459	.0009	.581	0.00	0.25	24–26	$\leq$0.006	$>$.15	.498–.529	n.c.	n.c.
12	4.506	.0002	.550	0.00	0.00						

Table 11: Qwen2-VL cross-generation tables (Wang et al., 2024) (cw/p/probe/pooled inj/pattern). 7B: UI=0.9506, band [0–13] (pattern .50 at L14 breaks contiguity; L16/L18 also meet the raw elicitation thresholds); color grid α=1 [L0 0, L4 .875, L9 1.0, L13 .625], α=2 [1.0, .5625, .75, .125], α=4 [.5625, .4375, −.75, −.875]; spatial Δ 0.00 at L0/1/2. 2B: probe sub-threshold at every causal layer, so band [ ] and UI=0.0 by rule; weak pooled elicitation recurs (pooled 0.00 everywhere causal; spatial Δ 0.00 at L11/12/13); α^*=2 with color grid α=1 $\approx$ 0 across four layers, α=2 [.25, .875, .75, −.5625], α=4 [.5, −.125, −.4375, −.875].

	Qwen2-VL-7B					Qwen2-VL-2B				
L	cw	p	probe	inj	pat	cw	p	probe	inj	pat
0	4.070	.0002	.617	0	0	2.542	.0025	.335	0	.17
1	3.747	.0002	.600	0	0	3.260	.0002	.335	0	0
2	3.340	.0005	.600	0	0	2.940	.0002	.335	0	0
3	3.713	.0016	.586	0	.17	3.214	.0002	.385	0	0
4	2.832	.0002	.631	0	.08	3.108	.0004	.425	0	.17
5	3.453	.0023	.614	0	.08	3.833	.0002	.405	0	.17
6	3.122	.0020	.583	0	.17	3.531	.0002	.385	0	.17
8	3.277	.0002	.612	0	0	3.949	.0002	.380	0	.17
9	3.894	.0005	.519	0	.33	3.382	.0002	.335	0	0
10	3.854	.0003	.514	0	.17	2.849	.0005	.420	0	0
11	3.917	.0021	.514	0	.17	4.194	.0002	.425	0	0
12	3.882	.0002	.562	0	.25	2.517	.0002	.395	0	.08
13	3.775	.0006	.645	0	.33	2.749	.0005	.425	0	.08
14	2.482	.0033	.645	0	**.50**	L14+ non-causal (cw $\leq$0.18)				
15	0.296	.24	n.c.	–	–	probe .425–.545, crosses .50				
16	0.553	.0044	.595	0	.17	only at non-causal L21/L22				
17	0.103	.36	n.c.	–	–					
18	0.367	.0031	.645	0	0					
20–26	$\leq$0.03	–	.579–.662	–	–					

Table 12: LLaVA-OneVision-7B (Li et al., 2025) (12 interchange pairs; probe over all valid items; pooled and pattern injection rates). Pooled injection is 0.00 except for 0.08 at L0; pattern rates in the table range from 0.50 to 0.83 over L0–L14. The focused comparison below reports coherent low-dose generations. Color grid α=1 [L0 .875, L3 1.0, L6 1.0, L10 1.0], α=2 [1.0 ×4], α=4 [.5625, .375, −.375, 0], so α^*=1; spatial Δ [L13 +.125, L14 0.00, L17 +.375].

L	cw	p	probe	inj	pat	L	cw	p	probe	inj	pat
0	8.391	.0002	.490	.08	.58	9	6.793	.0002	.557	0	.67
1	8.425	.0002	.490	0	.50	10	7.803	.0002	.610	0	**.83**
2	5.729	.0002	.490	0	.58	11	8.094	.0002	.610	0	.75
3	7.377	.0012	.507	0	.58	12	6.662	.0002	.590	0	.75
4	4.835	.0002	.507	0	.58	13	7.857	.0002	.573	0	.58

L	cw	p	probe	inj	pat	L	cw	p	probe	inj	pat
5	7.311	.0002	.507	0	.67	14	6.874	.0009	.593	0	.67
6	6.395	.0010	.543	0	.67	15/16	1.95/1.07	.056/.063	n.s.		
8	6.845	.0002	.577	0	.58	17	1.283	.0100	.627	0	.25
						18–26	non-causal (cw $\leq$0.61 or $p >$.01); probe .52–.68				

Focused payload comparison (6 pairs): pooled 0.00 against pattern 0.83 at L10 and 0.67 at L13; sample hits at α=1–2 are coherent full sentences correctly naming both objects and the injected relation.

J INSTRUMENT TABLES

Table 13: J-lens (Gurnee et al., 2026) fits on the Qwen2.5-VL-3B decoder and the text-sanity battery over five model-known facts. WikiText fits use the reference corpus family (Merity et al., 2017). The J_0 NaN rows scale with prompt count (1 at 96 to 47 at 1032), indicating non-finite values in the fitted Jacobian; L0 is excluded from all descriptive readouts.

lens	corpus	prompts	skip	fit GPU-h	battery	J_0 NaN rows	MRR(L8–28)
v1	chat-templated	96	4	0.36	0/5	0	.0260
v2	WikiText	96	16	0.60	0/5	1	.0154
v3	WikiText	1032	16	7.05	0/5	47	.0046

Per-fact best rank over L8–L28 (v1/v2/v3): France 375/241/258; Japan 1352/308/270; hot–cold 49/101/275; sky–blue 13/23/112; 2+2 34/61/325. The layerwise Jacobian norm profile is a stable U-shape across budgets, minimum at L11 (31.20 at n=96; 35.83 at n=1032) and maximum at L31 (57.90; 57.99), showing a similar aggregate norm profile despite poor fact recovery. The battery's model-knows flags are computed from the same encode-forward-unembed pipeline that feeds the lens.

K CALIBRATION AND READOUT CONVENTIONS

Layer classifications use raw pooled and position-preserving success rates. The signed-response analysis uses Δ and color-selected doses, so its results are reported alongside the raw-rate index. Dose sweeps are reported by dose because the color response can collapse or reverse at larger perturbations. Positional-pattern comparisons distinguish the smaller six-pair study from the 12-pair, three-seed repeat. J-lens fact recovery is evaluated separately from the interchange, probe, and injection measurements; layer 0 is excluded from descriptive lens readouts because the WikiText fits contain non-finite Jacobian rows.

L REPRODUCIBILITY DETAILS

Stimuli: 280 px square canvases; 32 cardinal-relation mirrored pairs (cardinal set: seed 20260710, spec hash `9e65d848eb557b51`, rendered-image hash `b5e09adaee631395`) and 16 diagonal-relation pairs; gradient set frozen separately (seed 20264710, spec `003855854ebb381b`, images `496246967f9a0121`). Layer grid: 25 selected decoder layers, listed in Table 3 for the primary model. Probes: 4-way logistic regression, standardized features, C=1, GroupKFold(5) by pair (Pedregosa et al., 2011). Injection dose grid $\alpha \in \{1, 2, 4\}$ with $\{8\}$ on control arms and $\{1, 2, 3, 4, 6\}$ in the calibration study; payloads scaled by the mean residual norm at the target layer. For the lens fits, corpora: 64/1000 WikiText-103 prompts (Merity et al., 2017) plus 32 multimodal scene prompts, selection hashes `274179995346c5d7` and `c85eb5a18586b4d5`. Models: Qwen2.5-VL-3B/7B fp16 (Bai et al., 2025), Qwen2-VL-2B/7B (Wang et al., 2024), LLaVA-OneVision-7B (Li et al., 2025). The experiments use one RTX 4090 (24 GB) for frozen-model forward passes and Jacobians; logistic probes are fitted separately.