BELIEVING IS SEEING: HOW VLMS ASSEMBLE TRUST IN AN IMAGE, AND WHERE IT CAN BE READ

Anonymous authors
Paper under double-blind review

ABSTRACT

When an image contradicts what a vision-language model expects, the model has to decide whether to believe it. We study how CLIP-based LLaVA models make that decision, using image–prior conflicts in which we can edit, shrink, or segment the contradicting object and intervene on the model's attention and residual stream. We find that the decision is read from a single representation at the answer position, and that three very different interventions move it in the same way: degrading the contradicting object, shrinking it, and redirecting a small set of attention heads toward it. A 256-dimensional subspace of the residual stream is enough to carry most of the effect of an image edit, and it generalizes to conflicts it was not fit on, although the model does not depend on it exclusively: removing it changes only a fraction of decisions. Layer-wise patching locates the image's contribution in the image tokens through the first third of the network and at the answer position from the middle onward, where the attention heads that mediate the decision act. The same map appears on a second encoder family (Qwen2.5-VL), and the readout and the edit sensitivity are already present in a LLaVA-1.5 that was never instruction-tuned on images. We then test the readout on real photographs from NaturalBench and find that the decision is again formed by the middle of the network, that an untrained mid-depth readout predicts it reliably, and that most of what is being overridden there is the model's default yes/no answer rather than knowledge about the object. The readout is accurate enough to let the model stop early on the items it is confident about, but the interventions that move trust on curated conflicts do not correct it on photographs. Together, these results describe trust in the image as one representation with a known content, location, and depth, and show that the curated-conflict results do not predict behavior on natural images beyond the readout.

1 INTRODUCTION

Vision-language models (VLMs) can describe an image and still answer from what they expected to see rather than from the image itself. When the image and the language prior disagree, something inside the model decides which to follow. Prior work located attention heads that mediate such conflicts (Ortu et al., 2026; Lietzow et al., 2026), showed that the visual evidence is decodable even when the prior wins (Nooralahzadeh et al., 2026), and cautioned that decodability is not equivalent to grounding (Liao et al., 2026). Together, these studies show that the model makes such a decision and that particular heads take part in it. They leave open what the decision is made of, where in the network it is assembled, and whether any of it holds outside curated conflicts.

We answer those three questions with a single measurement, the margin between the image-supported and prior-supported answers at the answer position, and a series of interventions on the same conflicts. We edit, shrink, and segment the contradicting object to find what the model responds to. We reallocate attention, inject directions, and patch subspaces to find what carries the decision inside the network. We patch activations one layer band at a time to find where it is assembled. Finally, we take the resulting readout to real photographs and to a model without visual instruction tuning, to test where these results hold and where they do not. The arbitration heads and the layer-20 readout were introduced by Vankadaru & DuTell (2026) on one model and one dataset; we build on both here.

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.

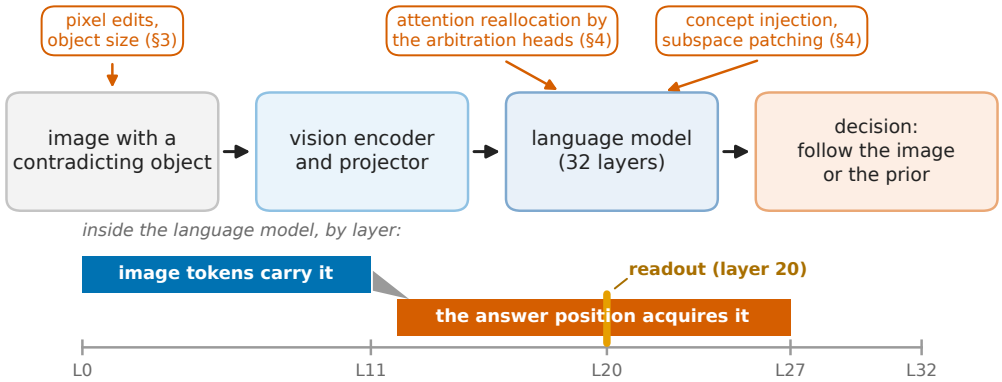

Figure 1: The contradicting object is encoded and projected into image tokens; through layer 11 those tokens carry the image's contribution to the decision, across layers 12–27 the answer position acquires it, and an untrained margin at layer 20 reads it before the answer. Three kinds of intervention act at the marked sites: pixel edits and object size on the input, attention reallocation by the arbitration heads, and concept injection or subspace patching in the residual stream. Same-image patching by layer band produces the map; the readout is then tested on real photographs.

Contributions. (1) We show that the model's trust in a conflicting image is set by how discernible the contradicting object is. Editing or shrinking only that object moves the decision several times more than editing the rest of the image or a random region of the same size, and a control that applies every pixel operation except the size change has no effect (§3). (2) We show that pixel edits, object size, and attention reallocation all move one representation at the answer position, and that a 256-dimensional subspace of the residual stream carries most of an image edit's effect on conflicts it was not fit on; the representation is sufficient rather than necessary, since removing it changes only a fraction of decisions (§4). (3) We map where the image's contribution is assembled by patching one layer band at a time, find the same order on two encoder families, and show that the readout and the edit sensitivity are present in a model that never received visual instruction tuning (§5). (4) On real photographs, the decision is again formed by the middle of the network and can be read there without training. Most of what the image overrides is the model's default yes/no answer. The readout is reliable enough to support early exit, but the interventions do not correct the override (§6).

2 Experimental Setup

Conflicts. We assemble 2,320 image–prior conflicts from four sources: 500 from WHOOPS-AHA (Ortu et al., 2026), 600 from ViLP (Luo et al., 2025), and 493 color and 727 size items from Visual CounterFact (Golovanevsky et al., 2025) (the color release has 493 rows; Lietzow et al. (2026) report 469 after their own validity filter, whereas we filter per model). In each conflict the image supports one answer and the model's prior supports another. The color items and the WHOOPS and ViLP items are *attribute* conflicts, where the evidence is a single object with an unexpected property; the size items are *relational* conflicts, where the evidence is a pair of objects whose relative size is unexpected. Our outcome measure throughout is the difference in logits between the image-supported and prior-supported answers at the answer position, which we call the *gap*; its sign is the decision. A conflict counts for a model only if the model's prior wins when no image is given. Two checks validate the outcome measure: the next-token decision agrees with free generation on 87–95% of conflicts, and 90–99% of image-wins disappear when the image is replaced by gray (Liao et al., 2026). Answer conventions per source are in App. A.

Conventions and terms. Layer L is the output of decoder block L (hidden-state index $L{+}1$; the NaturalBench readout uses hidden-state index 20, i.e. layer 19, inside the same 18–24 plateau). Edits: grayscale, saturation scaling, Gaussian blur, additive noise. *Head signal*: the summed logit contribution of the arbitration heads to the gap. *Gap*: the logit-difference margin between the image-

Table 1: Object-level effects on the gap (logits). Row 1: ViLP plus color counterfactuals, gray edit. Rows 2–3: the mask-matched WHOOPS subset (n = 364, 361, two models); Fig. 2a and App. G use the full set (n = 424, 416, 419). Rows 4–6: WHOOPS. Bracketed values are bootstrap 95% CIs.

	LLaVA-NeXT	LLaVA-1.5	BakLLaVA
Gray object / rest, ViLP+color (n≈970)	-2.9 / -0.2	-2.9 / -0.2	-3.3 / -0.2
Gray object [CI], WHOOPS (n=364/361)	-0.62 [-0.77, -0.46]	-0.65 [-0.82, -0.46]	–
Gray rest / matched random region	-0.30 / -0.24	-0.33 / -0.23	–
Object $0.5\times$ [CI], WHOOPS	-1.08 [-1.26, -0.91]	-0.60 [-0.78, -0.43]	-0.86 [-1.11, -0.62]
Recomposite $1.0\times$ (artifact control)	-0.01	-0.01	-0.00
Object at $0.75\times$	-0.55	-0.17	-0.25
Head signal r with the gap	$+0.71$	$+0.60$	$+0.65$

supported and prior-supported answers at the answer position; the *decision* is its sign, an *image-win* or *prior-win*; *vision-win* is the image-win fraction. *Discernibility*: how legible, large, and attended the conflict evidence is. *Arbitration heads*: heads whose attention reallocation moves the gap; ρ, α: reallocated attention fraction and injection strength (relative to the local residual norm); *recomposite*: the segmented object pasted back at a new scale over a blurred fill.

Models and heads. LLaVA-NeXT-Mistral-7B, LLaVA-1.5-7B, BakLLaVA-7B (CLIP ViT-L); Qwen2.5-VL-7B; LLaMA-3.2-11B-Vision (cross-attention; measurement only); and a projector-only LLaVA-1.5 assembled from Vicuna-7B-v1.5, the CLIP tower, and the public stage-1 projector. Arbitration heads follow the logit-inspection recipe of Ortu et al. (2026); our pipeline recovers 17/20 and 15/20 of the heads reported for LLaVA-NeXT by Vankadaru & DuTell (2026) and their numbers within noise (App. B). Conflict objects are segmented with SAM3 from the answer phrase (coverage 89%). Every headline result is on the three LLaVA models; four of the final controls (the cross-source subspace comparison, head necessity, the matched-area region, and natural-norm injection) are on LLaVA-NeXT and LLaVA-1.5, because BakLLaVA failed a patching self-test in that sweep (App. P); Qwen2.5-VL replicates the edits, the size dose, and the band map, and LLaMA-3.2-Vision is measured only. Bootstrap intervals are per row and do not cluster on repeated objects or questions, except in the mediation analysis. Sample sizes: the valid WHOOPS conflicts per model are 424, 416, 419 (written $n \approx 420$); rows that also require a SAM3 mask use 364 and 361; the image-right subset used for the loud-versus-quiet comparison has $n \approx 60$ against $n \approx 360$ image-wrong rows. Every effect is measured on the gap; patching uses the same image, edited, as the counterfactual (Zhang & Nanda, 2024) and acts on the residual stream (Makelov et al., 2024), with hooks verified against the unpatched forward pass (App. C).

2.1 Related and concurrent work

Previous work on head-level levers is the closest: Ortu et al. (2026) identify counterfactual heads on WHOOPS-AHA and steer by reweighting their attention; Lietzow et al. (2026) give a prior-promoting head circuit whose ablation flips most prior-grounded predictions and few visually grounded ones; V-SEAM (Wang et al., 2025) and Gaze Heads (Gandikota & Bau, 2026) modulate head subsets to redirect what the model describes. Readouts and crossovers: Nooralahzadeh et al. (2026) report a logit-lens crossover on Visual CounterFact and show that image tokens carry the causal impact; Golovanevsky et al. (2025) place the prior-to-visual shift in the middle layers and steer it with learned vectors. On methodology, Liao et al. (2026), Makelov et al. (2024), and Zhang & Nanda (2024) supply the arbiter, subspace, and patching controls we adopt. This paper connects those pieces. We show that the heads, the crossover, and the causal dominance of image tokens are three views of one representation, that its content is the discernibility of the contradicting object, and that its location can be mapped by position and depth on the same edited image; we then test its readout on real photographs and on a model without visual instruction tuning (per-paper deltas: App. M).

Table 2: Sufficiency and necessity at the answer position and image tokens (WHOOPS, $n \approx 420$ per model). Recovery = fraction of the actual grayscale effect reproduced by patching; the last two rows: share of image-wins reversed by removing the subspace from the accumulated residual (from the band's own updates in parentheses), and the concept-injection gap shift. BakLLaVA values come from the same-image and band-map notebooks, whose identity checks passed (App. P); the dashes mark controls run only on the other two models.

	LLaVA-NeXT	LLaVA-1.5	BakLLaVA
Same-image patch at image tokens (L16–24): full / 256-D / random	69 / 67 / 5%	77 / 75 / 5%	73 / 68 / 1%
Cross-source 256-D [95% CI] / next-256 / random in top-1,024 span	68 [55, 81] / 8 / 26%	66 [55, 78] / 9 / 26%	–
Top-20 heads ablated: image-wins flipped (gap)	17% (−1.12)	7% (−0.87)	–
Interchange at the answer position, prior-wins flipped: full / 256-D / random	44 / 28 / 11%	36 / 21 / 5%	40 / 24 / 8%
Removing the 256-D subspace: image-wins flipped	12 (3)%	12 (2)%	9 (3)%
Concept injection α=1, gap shift, WHOOPS	+4.0	+2.4	+1.4

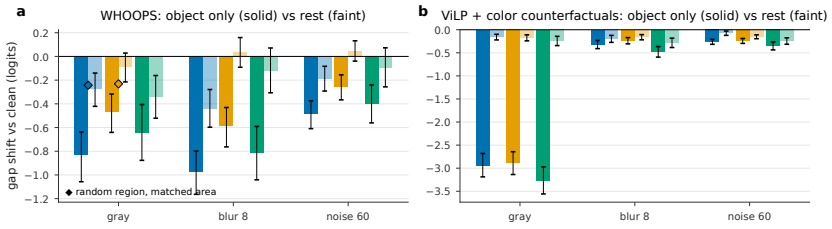

Figure 2: **Degrading only the object vs only the rest.** Gap shift vs the clean image for gray, blur, and noise edits at their strongest level (bootstrap 95% intervals), on the full WHOOPS set (n = 424/416/419) (a) and on ViLP plus color counterfactuals (b); diamonds in (a) are a random region of the object's own area, measured on the matched-mask subset (n = 364/361) that Table 1 rows 2–3 report. The size dose with its recomposite control is Fig. 4; steering onto object vs surround is Fig. 16.

3 THE OBJECT DECIDES

If a model believes an image because of what it can see of the contradicting object, then editing that object alone should move the decision, editing everything else should not, and making the object harder to see should move the decision in proportion. We tested each of these on all three models, with two controls that distinguish the object's role from the image's: a random region of the same size, and a recomposite that applies every pixel operation except the size change.

Object-only versus rest-only edits. To isolate the object, we segmented it with SAM3 and then applied one of three edits: graying the object, graying everything except it, or graying a random region of the same area (Fig. 2, Table 1). On WHOOPS, graying the object moved the margin about twice as much as graying the rest of the image and 2.6 times as much as the matched random region. On the color conflicts, where graying the object removes the conflicting color itself, the object edit was 12–18 times larger than the rest. The object-versus-rest contrasts survive Holm correction for blur and noise on all three models and for grayscale on two (App. G).

Smaller objects are less trusted. To test size directly, we cut the object out, scaled it, and pasted it back over a blurred fill at 0.5, 0.75, 1.0, 1.25, and 1.5 times its original size. At half size the margin fell by 1.08, 0.60, and 0.86 logits on the three models, at three-quarter size by about half that, while the recomposite at the original size, which applies every pixel operation except the size change, moved it by 0.01 (Fig. 4). The model responds to the size of the object rather than to the compositing: how much of the object it can see sets how much it believes.

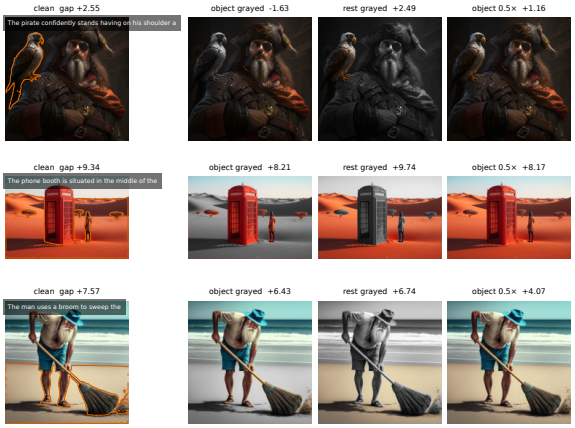

Figure 3: **The edits on real conflicts (LLaVA-NeXT).** For three WHOOPS conflicts: the image with the SAM3 object outlined and the statement, then the object-only gray edit, the rest-only gray edit, and the $0.5\times$ recomposite, with the gap above each. Graying the object alone reverses the margin in the first example and lowers it in the others, graying everything else leaves it, and shrinking the object lowers it.

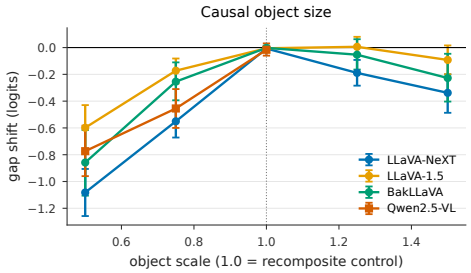

Figure 4: **Causal object size.** Gap shift when the segmented object is recomposited at $0.5\times$ to $1.5\times$ its size, four models, bootstrap intervals; the $1.0\times$ recomposite applies every pixel operation except the size change and is flat.

Relational conflicts. We also ran the attention lever of §4 on the relational conflicts, where the evidence is a pair of objects. There, steering attention onto the resized object lowered the margin and steering onto its surroundings, which contain the comparison object, raised it on two of three models: what the model needs to see is the evidence for the conflict, which is one object for attributes and the pair for relations. Discernibility is also a lever rather than a predictor: the object's area adds nothing to the model's own uncertainty in predicting which conflicts it will follow (App. F), and a mediation analysis attributes only a fifth to a half of each edit's effect to the object's concept overlap and attention mass. The remainder is not yet explained.

4 A SHARED REPRESENTATION AND A SUFFICIENT SUBSPACE

The edits of §3 act on the input, and the arbitration heads act in the middle of the network. We asked whether these levers move the same internal quantity, and if so, which part of the residual stream carries it.

Steerable and decodable directions. Two directions can be read at the answer position: the answer-concept direction (the difference of the two candidates' unembedding rows) and the direction a classifier fits to decode the decision. Injected at image tokens, the first moves the decision ($+2.61$, $+1.98$, $+2.30$ logits at $\alpha = 0.5$) and the second does nothing (-0.01, -0.13, -0.07), while an unrelated concept direction also does nothing (App. G). The direction that best decodes

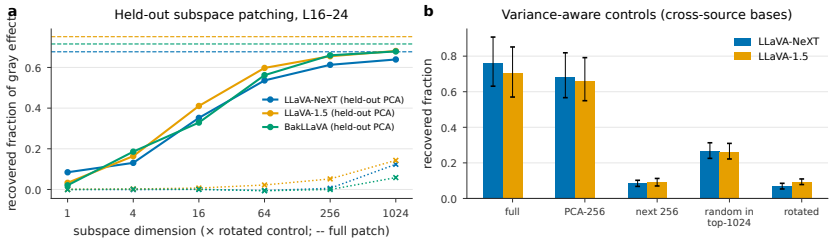

Figure 5: **The subspace is sufficient and specific.** (a) Recovered fraction of the gray edit by subspace dimension with bases fit on half the conflicts and evaluated on the other half; crosses are rotated subspaces of the same dimension, dashed lines full patching. (b) With bases fit on the other sources and evaluated on WHOOPS, the top-256 subspace against the next 256 principal components, a random subspace within the top-1,024 span, and a rotated subspace.

the decision is therefore not the one that steers it (both are tested by injection at image tokens). The answer-concept direction is candidate-dependent, a family of directions with one construction rule at one site, which is the sense in which we call the representation shared. The levers of §3 move the steerable direction: under pixel edits, object shrinking, and attention reallocation, the shift in the layer-20 answer-concept projection correlates with the shift in the decision (r = 0.56–0.84; Table 2, Fig. 11). Part of that correlation is structural, since the projection is a component of the margin, so the injection contrast is the stronger evidence. An additive surface fit on half the examples explains held-out variance at R^2 = 0.40, 0.27, 0.19 with one interaction (attention steering and injection compound) and no transfer across sources (App. H): the levers share a representation; they do not share a law.

4.1 A SUBSPACE THAT IS SUFFICIENT AND SPECIFIC

We then asked how much of the residual stream is needed to carry the decision, using a 256-dimensional subspace fit as the top principal components of image-token residuals on the conflict sets (construction and the two application sites in App. E).

Sufficiency on held-out conflicts. We patched the edited image's residuals into the clean run through only this subspace, at image tokens in layers 16–24, and compared the recovered effect with patching the full residual. The subspace recovered 67–75% of the gray edit's effect against 69–77% for the full residual; with the basis fit on the other sources and evaluated on WHOOPS, 0.68 and 0.66 against 0.76 and 0.71 (Table 2). At the answer position the same subspace flipped about 60% of the prior-wins that full cross-example patching flips. A 256-dimensional subspace, fit on other conflicts, therefore carries most of an image edit's effect.

Comparison with high-variance alternatives. A top-256 principal subspace also holds most of the residual's variance, so we compared it with other subspaces that carry substantial variance: the next 256 principal components recover 8–9% and a random 256-dimensional subspace inside the top-1,024 span recovers 26%, against 66–68% for the decision subspace and 7–9% for a rotated one (Fig. 5). Recovery tracks captured variance in the random control (a random quarter of the top-1,024 span recovers about a quarter), so the evidence that the subspace is more than variance comes from the next principal block, which carries the next-largest variance and recovers an eighth as much; the decisive control, an equal-variance subspace outside the top block, was not run.

Redundancy outside the subspace. We also injected the answer-concept direction split into its component inside the subspace and its component outside, each at its natural norm. At natural norms the part of the injected concept direction inside it carries 27–36% of the effect from 26–29% of the norm, about four times a random subspace per unit norm, while the complement still carries the majority; the injection passes the in-range check of Makelov et al. (2024) and an unrelated concept direction does nothing (App. G). Across models the layer-20 subspaces align up to a rotation (16–18° against 24–27° shuffled and 78° random), though per-example directions port only partially (App. I).

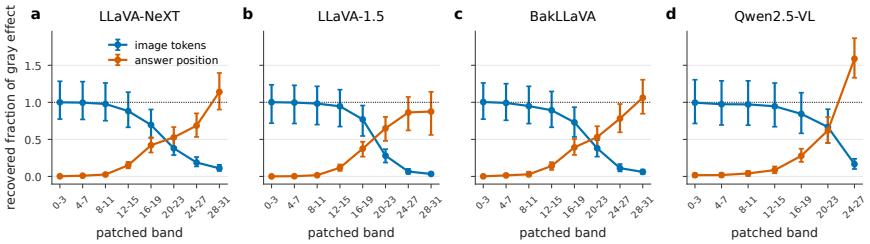

Figure 6: **Transport map.** Residuals of the grayscale copy of the same image patched into the clean run per 4-layer band, at image tokens (blue) and at the answer position (orange); y = fraction of the actual grayscale effect recovered, with bootstrap intervals. Through layer 11 only the image tokens suffice; across layers 12–27 the answer position comes to suffice; the crossover at 20–23 overlaps the arbitration heads' layers (mean 23–25). Qwen2.5-VL (d, 28 layers) shows the same ordering with the crossover in its last band. Values above 1 are overshoot on the paired difference. $n \approx 420$ WHOOPS conflicts per LLaVA model, 300 for Qwen.

5 Where the image's contribution is assembled

Next we locate the representation in the network. We patched the activations of a grayscale copy of each conflict image into the clean run, one 4-layer band at a time and at either the image tokens or the answer position, and measured how much of the gray edit's effect each patch reproduced (Fig. 6).

Location by layer band. We found that through layer 11 the image's contribution is present only in the image tokens: patching them reproduces the whole effect and patching the answer position reproduces none. Across layers 12–27 the two curves cross, and by layer 28 the answer position carries everything. The crossover, at layers 20–23, overlaps the band where the arbitration heads sit (mean layer 23–25) and is where the readout of §6 becomes predictive. The same order holds on Qwen2.5-VL, a different encoder family, with the crossover later in relative depth. Two qualifications apply. The map measures sufficiency, not necessity, and the layer-0 recovery is expected by construction because only the image differs between the paired inputs. It also measures where the image's contribution to this decision sits, which need not be specific to trust, since any causal transformer moves image information toward the answer position with depth (Salazar et al., 2026); the control that would separate the two, the same band map on images where the prior agrees, was not run. The map also explains a discrepancy in prior work: Nooralahzadeh et al. (2026) found that last-token patching flips almost no decisions, but they patched at the crossover layer, where the answer position carries little; at layers 16–24 the same patch flips 36–44% of prior-wins (Table 2).

Necessity of the heads and the subspace. We then asked whether the heads and the subspace are necessary, by ablating them at the answer position. The heads are necessary in a graded but partial way. Mean-ablating the top 5, 10, and 20 arbitration heads at the answer position flips 2.5, 12.7, and 16.9% of image-wins on LLaVA-NeXT and 1.8, 2.3, and 7.0% on LLaVA-1.5, against 0.8–2.5% for layer-matched random heads, and prior-wins are unaffected. These rates are well below the 68–96% that Lietzow et al. (2026) report for prior-promoting heads under a prompt contrast; the two experiments measure different things. The subspace is less necessary than it is sufficient: removing it from a band's own updates flips 2–3% of image-wins, and from the accumulated residual 9–12%, costing about a logit of margin (App. E). The decision is read at the answer position from a signal written across many upstream updates, consistent with a redundant encoding.

Presence before visual instruction tuning. To ask when the mechanism appears in training, we assembled a LLaVA-1.5 from Vicuna-7B-v1.5, the CLIP tower, and only the stage-1 projector, with no visual instruction tuning. It followed the image on 0.755 of WHOOPS conflicts against 0.733 for the tuned model (a paired difference indistinguishable from zero), read its own decisions at 0.82 AUROC against 0.83, and lost 58–77% as much margin to grayscale and blur (Fig. 7). Visual instruction tuning sharpens the response; it does not create the mechanism, and since Vicuna is already chat-tuned, the comparison isolates the visual stage alone.

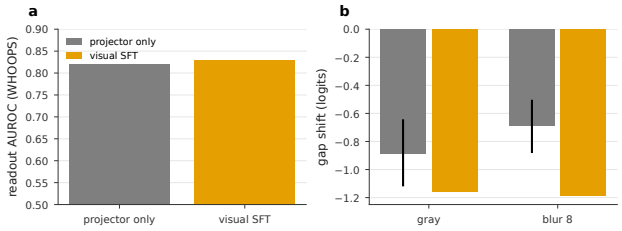

Figure 7: **Present before visual instruction tuning.** The projector-only LLaVA-1.5 against the tuned model: readout AUROC on WHOOPS (a) and the gap shift under grayscale and blur with bootstrap intervals (b).

Table 3: Readout. AUROC of the untrained layer-20 margin for the decision (hidden-state index 20 on NaturalBench; see Conventions); NaturalBench values after subtracting the median layer-20 margin over all evaluation items; n = 1,200 natural conflicts per model.

	LLaVA-NeXT	LLaVA-1.5	BakLLaVA
WHOOPS, within source (n=424/416/419)	0.80	0.83	0.80
NaturalBench, centered on all items (raw)	0.88 (0.55)	0.98 (0.93)	0.92 (0.86)
NaturalBench, centered on a disjoint half	0.89	0.98	0.95
Text-only control (predicting the LM's own text-only answers)	0.85	0.91	0.94
NaturalBench, image-dependent items (gray arbiter; n=398/584/621)	0.96	0.97	0.81
NaturalBench, within yes / within no items	0.95 / 0.92	0.99 / 0.97	0.95 / 0.96
5-fold linear probe on the L20 residual (L2, C=0.01; App. N)	0.64	0.66	0.75
Prior override on NaturalBench (accuracy points)	29	22	37

6 Reading the decision before the answer

If the decision is assembled at the answer position by layer 20, it should be readable there without training. We tested this by projecting the layer-20 residual onto the two candidate answers' unembedding rows and using the difference as a score. On WHOOPS it predicts the decision at 0.80–0.83 AUROC (Table 3; layer 20 was fixed in the prior study and not re-tuned, and the centered curve plateaus across 18–24 on every model, Fig. 19). We then tested the same readout on real photographs.

The prior override on photographs. On NaturalBench each yes/no question comes with two photographs whose answers differ, so the model's text-only answer is right for one and wrong for the other. The models lose 22–37 accuracy points when the image contradicts that answer (Fig. 8a). The text-only answer, though, is mostly a default yes or no rather than knowledge about the object (yes-rates 0.33, 0.52, 0.66), so what is overridden here is a polarity default.

Readout on photographs. The same layer-20 margin, applied unchanged, reads the decision on photographs: raw, at 0.55, 0.93, 0.86 AUROC, and after centering at 0.88, 0.98, 0.92. The score is $s_i = p_i\,(m_i - c)$: m_i the yes-minus-no margin at layer 20, c its median over the evaluation items, and $p_i = \pm 1$ the orientation toward the image-supported answer; centering changes the ranking because p_i differs across items (App. L). The early exit uses $m_i - c$ alone. The raw figure on LLaVA-NeXT is low because its margin carries a large polarity offset on this benchmark (yes-rate 0.33), which centering removes. By layer 20 the answer is decided. A text-only control shows how much of this is the language model: with no image, the same margin predicts the model's own text-only answer at 0.85–0.94. The readout is therefore a general property of the language model, and we used the gray-image arbiter to measure what the image adds (Fig. 8b): on the third to half of items where replacing the photograph with gray changes the answer, the readout is 0.96 and 0.97 on LLaVA-NeXT and LLaVA-1.5, and on LLaVA-NeXT it is higher than on default items (0.82); on BakLLaVA it is 0.81 where the image decides against 0.98 where it does not. On two models the

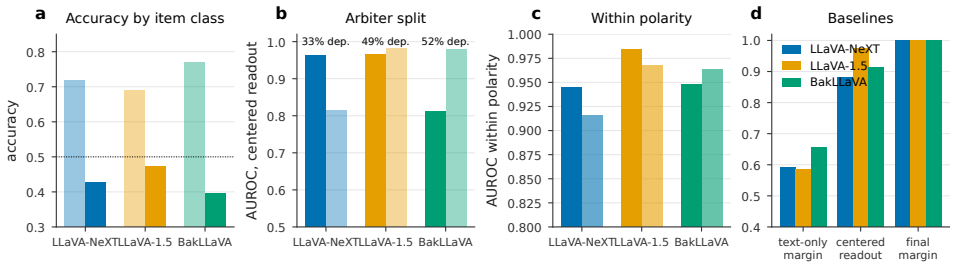

Figure 8: **Real photographs.** NaturalBench yes/no pairs, 1,200 natural conflicts per model. (a) Accuracy when the prior agrees (faint) vs when the image contradicts it (solid). (b) The centered readout on the items where a gray image changes the answer (solid) vs the default items (faint), with the image-dependent share above each pair. (c) The readout within each answer polarity. (d) The readout against the text-only margin and the final-layer margin on the same items. Depth curves are in Fig. 19.

readout reflects the image's contribution where it matters; on the third it mostly reflects the default answer. Within-polarity readouts, the probe baselines, and the cross-model matrix are in App. L.

Early exit. Exiting on the items where the centered margin is largest matches full-depth accuracy from 30% through 90% coverage on all three models (at 50%: 0.610 vs 0.620, 0.635 vs 0.635, 0.642 vs 0.642) and saves 12 of 32 decoder layers on the exited items (about 19% of decoder compute at 50% coverage, not an end-to-end latency); the exited answer agrees with the full-depth answer on 100% of items at 30% coverage and 82–95% at 90% (App. J). Layer 20 is not special within the plateau: an exit at layers 18–26 does as well (0.619–0.654 at matched coverage), while an exit at layers 8–16 falls to 0.50–0.59 (Fig. 18, App. J). The decision is complete by the plateau rather than at one layer, and the exit saves computation rather than error, since it reproduces the model's own answer, right or wrong.

Correction. We then tried to correct the override with the levers that move trust on curated conflicts, and they do not. Object-attention steering with a mask of the question noun changes conflicting-item accuracy by at most $+0.7$ points; the object's area does not predict image-following (r within ± 0.07 with parsed nouns and 63% mask coverage). Injecting the model's own text-only answer direction does move accuracy, but only from one item class to the other: with sign and strength selected on 400 conflicting items, held-out conflicting accuracy rises by $+2.9$, $+13.6$, $+13.5$ points while consistent-item accuracy falls by 2.9, 11.8, 10.6, a random direction of equal norm changes nothing, and gating on the readout reduces both sides without a net gain (Fig. 21, App. L). The direction is a polarity-bias vector with a fitted sign, not a correction of the decision. Steering leaves POPE, HallusionBench, and MME within two points (App. K).

7 Conclusion

On the LLaVA models studied here, whether the model believes an image depends on how much of the contradicting object it can see. We found that this weighing has a representation that three levers move, a subspace that is sufficient for it, and a depth by which it is assembled at the answer position and readable without training; that the representation is redundant rather than necessary; that the map locates the image's contribution rather than trust as such; and that on real photographs the prior being overridden is mostly a default yes/no answer. These results connect the heads of prior work, the logit-lens crossover, and the causal dominance of image tokens as three views of the same representation, and they provide a readout that works on photographs well enough to support early exit but not correction.

Limitations. Two controls that would sharpen the claims were not run: a band map on images where the prior agrees, and an equal-variance subspace outside the top principal block. Four final controls are on two models; Qwen2.5-VL's outcome statistics vary with the library version; intervals do not cluster on repeated items; the mechanism models share one frozen encoder at 7B (App. P).

Reproducibility statement

Every number in the main text maps to a notebook and an output file listed in App. B; intervention hooks and the steering path are verified in App. C; data construction, answer conventions, and prompts are in App. A. Per-model heads, subspace bases, the readout, and the annotated conflicts will be released as supplementary material.

AI use statement

Generative AI tools were used for writing assistance (polishing and drafting sections), for literature retrieval and discovery, and for execution support (drafting analysis code); not for research ideation, for generating synthetic datasets, or for proving mathematical claims. All AI-assisted text, code, and claims were reviewed and verified by the authors, who take responsibility for the final content.

References

Chan et al. System-mediated attention imbalances make vision-language models say yes. *arXiv preprint arXiv:2601.12430*, 2026.

Rohit Gandikota and David Bau. Gaze heads: How vlms look at what they describe. *arXiv preprint arXiv:2606.14703*, 2026.

Michal Golovanevsky, William Rudman, Michael Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *EMNLP 2025 (Main)*, 2025. arXiv:2505.17127.

Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. Halp: Detecting hallucinations in vision-language models without generating a single token. In *EACL*, 2026.

Chih-Ting Liao, Fei Shen, Xin Cao, and Tat-Seng Chua. Decodable is not grounded: A vision-ablation arbiter for vlm spatial reasoning. *arXiv preprint arXiv:2606.31257*, 2026.

Niclas Lietzow, Danielle Bitterman, Carsten Eickhoff, William Rudman, and Michal Golovanevsky. Vision-default, prior-override: Causal mechanisms of perception-knowledge conflict in vision-language models. *arXiv preprint arXiv:2606.28273*, 2026.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in vlms. In *ICML*, 2025.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *ICLR*, 2024. arXiv:2311.17030.

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.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. In *Proceedings of ACL 2026 (Long)*, 2026. arXiv:2507.13868.

Palacios et al. Visual instruction tuning aligns modalities through abstraction. *arXiv preprint arXiv:2606.03871*, 2026.

Salazar et al. Pathways of visual information flow in vision-language models. *arXiv preprint arXiv:2607.03358*, 2026.

Vijay Vankadaru and Vasha DuTell. Vision-language models trust what they can discern: Image and representational properties governing modality arbitration. In *Explainable Computer Vision Workshop at ECCV*, 2026.

Qidong Wang, Junjie Hu, and Ming Jiang. V-seam: Visual semantic editing and attention modulating for causal interpretability of vision-language models. In *EMNLP*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *ICLR*, 2024.

A Data and prompts

Sources. WHOOPS-AHA: 500 conflicts (435 in the prior study of Vankadaru & DuTell (2026) plus 65 further rows from the same release); ViLP: 600 (200 questions $\times$ 3 images); Visual CounterFact color: 493 rows, 493 unique objects and images; Visual CounterFact size: 727 rows over 410 unique objects ("dog" 25$\times$, "car" 8$\times$), one image per row. Rows are (object, counterfactual image) pairs as released. Per-model prior filter (text-only answer must be the prior answer, three outcome rules agreeing on $\geq 92\%$): valid conflicts across all four sources: LLaVA-NeXT 2,121, LLaVA-1.5 1,950, BakLLaVA 2,024; Qwen2.5-VL 1,133.

Answer conventions. On the size split the released `correct_answer` is the world-knowledge (prior) answer and `incorrect_answer` the image-supported one, the reverse of the color split; confirmed by the prior filter's acceptance rates (0.85, 0.64, 0.75).

Prompts. Statement continuation for WHOOPS and Visual CounterFact; question form for ViLP; model-family templates (Mistral `[INST]`, Vicuna `USER/ASSISTANT`, Qwen chat template). Yes/no benchmarks use the in-context continuation scorer with the attention-mask fix (§C); first-token and in-context scoring agree in sign on 100% of LLaVA-1.5 items.

Outcome validation. Free-generation agreement with the next-token outcome: 0.87–0.95 across models; gray-image arbiter: 90–99% of image-wins do not survive image removal (per-model values in the outcomes files listed in App. B). Quiet-truth ratios (prior wrong, image right, $n \approx 60$): gray 0.56/0.29/0.12, blur 0.44/0.84/0.69, size 0.50/0.85/0.89; saturation boost turns 4.5%, 7.1%, 6.5% of prior-wins into wins for the false answer.

Masks. SAM3 on the image-supported answer phrase; coverage 89% (WHOOPS), 702/727 (size split); median object area 8–10% of pixels. NaturalBench: yes/no items selected by their answer strings (1,450 of 1,900 rows are yes/no), first 600 pairs; the question noun for masks is the last noun phrase after a determiner (printed examples: "hat", "couch").

B Reproduction of prior results on LLaVA-NeXT

On LLaVA-NeXT the rebuilt pipeline recovers 17/20 counterfactual and 15/20 factual heads; attended-region size partial r +0.150 vs +0.174; attention concentration -0.110 vs -0.111; decodability 0.72 vs 0.71; DINOv3 gap +0.157 vs +0.156; saturation dose 0.718$\rightarrow$0.805 vs 0.72$\rightarrow$0.81; blur and noise curves within noise. Every number in this paper maps to a notebook and an output file: 01 outcomes, 02 heads and geometry, 03/03b captures and detector, 04 dose, 05 mediation, 06 masks, 07 steering, 08 localized edits, 09 detector, 10–10d benchmarks, 11 representation steering, 12 Qwen, 13/13b interchange, 14 CHAIR, A transfer, B same-image patching, E cross-attention, G size dose and encoder law, G2 error detection, H unification/necessity/universality/stage-1/loud-vs-quiet, I transport map and portable directions, J NaturalBench.

C Verification of interventions

The steering attention path is registered as a custom attention implementation and verified equal to eager attention with no steering set (max logit difference $< 10^{-3}$, and bit-identical outcomes); sdpa equals eager on all three CLIP models (max difference 0.016 in logits, identical outcomes). Interchange and ablation hooks return the unpatched forward exactly when their plans are empty. Qwen2.5-VL requires bfloat16 (fp16 overflows to NaN; a finite-logit guard aborts the run otherwise). The KV-cached decode path handles single-row steering; multi-row steering uses the query-row interface.

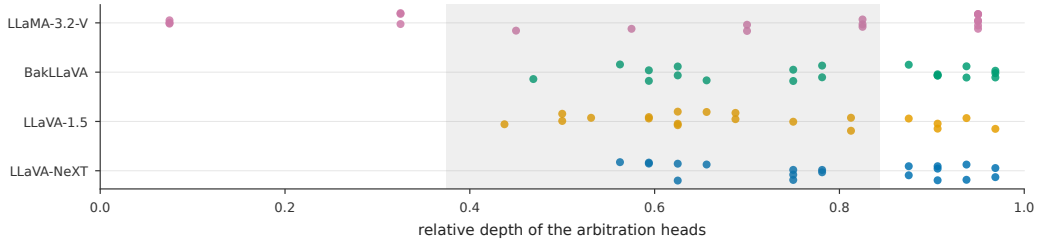

Figure 9: Relative depth of the arbitration heads on five models, including the cross-attention model; shaded band is layers 12–27 of 32.

D TRANSPORT MAP, EXTENDED

Band patches replace the residual stream entering each of the four decoder blocks in the band (forward pre-hooks on the blocks, applied simultaneously) at the chosen positions; the final block's output and the final normalization are never replaced, so even the last band does not reduce to copying the donor's logits, and recoveries near but not equal to one in that band are expected. Per-band recovered fractions (grayscale copy patched into the clean run), image tokens / answer position: NeXT 1.00/0.00, 1.00/0.01, 0.98/0.03, 0.88/0.15, 0.69/0.42, 0.38/0.53, 0.19/0.69, 0.11/1.14 for bands 0–3 through 28–31; LLaVA-1.5 1.00/0.00, 1.00/0.00, 0.98/0.02, 0.95/0.12, 0.77/0.37, 0.28/0.65, 0.07/0.86, 0.03/0.88; BakLLaVA 1.00/0.00, 0.99/0.01, 0.95/0.03, 0.89/0.14, 0.73/0.39, 0.38/0.53, 0.11/0.78, 0.06/1.06. Qwen2.5-VL (28 layers, $n = 300$): image tokens 1.00, 0.97, 0.97, 0.95, 0.84, 0.67, 0.17; answer position 0.02, 0.02, 0.04, 0.09, 0.28, 0.62, 1.59 for bands 0–3 through 24–27. Recoveries above 1.0 in the last bands (1.14, 1.06, 1.59) are overshoot on the paired difference: a late-band patch carries the gray copy's entire downstream state at that site, including components unrelated to the edit that happen to push the same way, and the denominator is the mean actual effect. Vision-wins lost per band follow the same profile (0.18 at the interface, 0.02 at the answer position, crossing at 20–23). The same-image notebook's L16–24 recovery (67–77% gray, 54–68% blur at image tokens; 256-D subspace 64–75%/44–64%; random 2–7%; 1-D probe 0%) is consistent with the band profile.

E NECESSITY

Subspace construction. The 256-dimensional subspace is one basis per layer: the top principal components of the attended image-token residuals captured on the conflict sets, fit once per layer. It is applied at the image tokens for same-image patching and at the answer position for interchange, necessity, and injection; the half-split and cross-source held-out designs and the cross-model alignment use the layer-20 basis. Ablation at the answer position over the band $\{16, 20, 24\}$. Increment

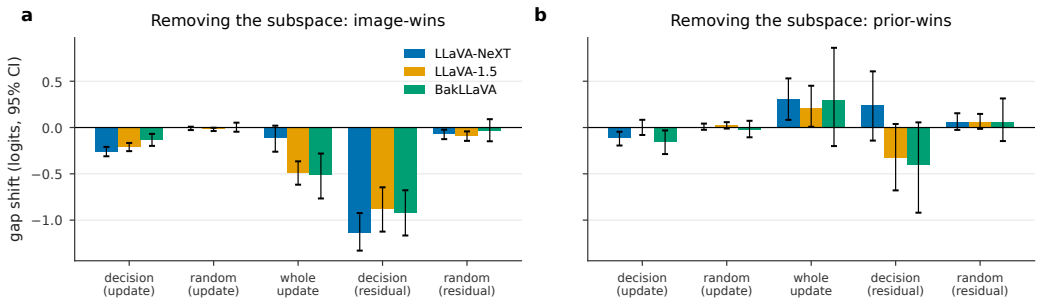

Figure 10: Removing the decision subspace from the layer updates and from the accumulated residual, on image-wins and on prior-wins, with random and whole-update arms and bootstrap intervals.

ablation (subspace removed from each layer's own update): decision subspace flips 2.6, 1.6, 2.6% of

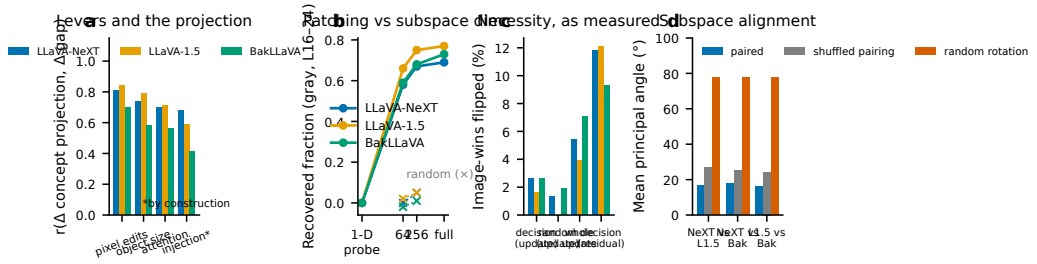

Figure 11: **Three levers, one projection.** (a) Shift of the answer-concept projection at the answer position vs shift of the gap under pixel edits, object size, attention steering, and concept injection (three models pooled). (b) Held-out subspace patching by dimension. (c) Necessity: decision subspace vs random vs whole update, removed from the layer updates. (d) Principal angles between models' subspaces after Procrustes vs shuffled and random baselines.

vision-wins (gap -0.22, -0.15, -0.14); random subspace 1.3, 0.0, 1.9% ($+0.01$, -0.01, $+0.02$); whole update 5.4, 3.9, 7.1% (-0.006, -0.31, -0.31; on LLaVA-NeXT the whole-update flips are symmetric, 5.4% of image-wins and 6.3% of prior-wins, so the mean gap shift is near zero while the decision is perturbed). Residual ablation (subspace removed from the accumulated stream): 11.8, 12.1, 9.3% of image-wins flipped (gap over all items -0.78, -0.73, -0.79; over image-wins only -1.14, -0.88, -0.93, the figure quoted in §5); random 2.2, 1.0, 2.2%. Interpretation in §5.

F Mediation and encoder analysis

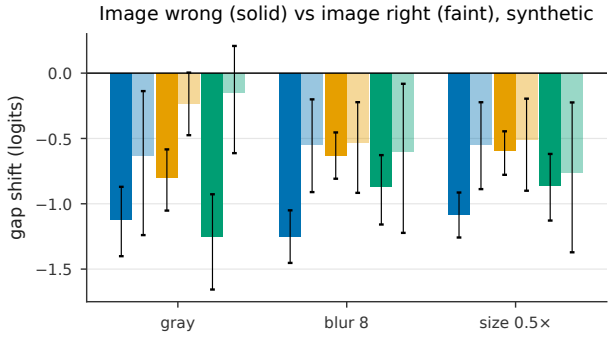

Figure 12: The same manipulations on synthetic rows where the image is wrong (solid) and right (faint), $n \approx 360$ vs $n \approx 60$, with bootstrap intervals.

Product-of-coefficients mediation with a cluster bootstrap over prompts: saturation acts $\approx 20\%$ through concept overlap on all three models (and through probe decodability), blur 27–45% through attention mass on the object; direct effects remain. The CLIP encoder's generic representational change of the object (cosine and L2 between clean and edited pooled object features) does not predict the trust shift across seven edits (grayscale changes the encoder less than blur 8 yet moves trust more); concept-specific change correlates positively but not significantly (Spearman 0.72, 0.36, 0.46, $n = 7$).

G Attention steering

Full localized-edit set on WHOOPS (gray, n = 424/416/419): object -0.83, -0.47, -0.64; rest -0.16, -0.09, -0.24. Holm-corrected paired inside-vs-outside contrasts (localized edits, all levels): significant at corrected $p < 0.05$ for blur 5/8 and noise 40/60 on all three models, for grayscale on LLaVA-NeXT ($p < 0.001$) and LLaVA-1.5 ($p = 0.001$), not for grayscale on BakLLaVA

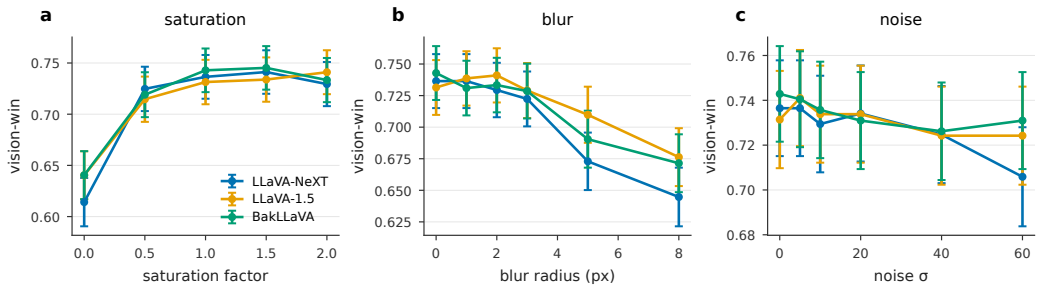

Figure 13: Dose curves for saturation, blur, and noise on the three CLIP models with binomial intervals; the Qwen2.5-VL dose files were not retained (its saturation flatness is reported from the run log).

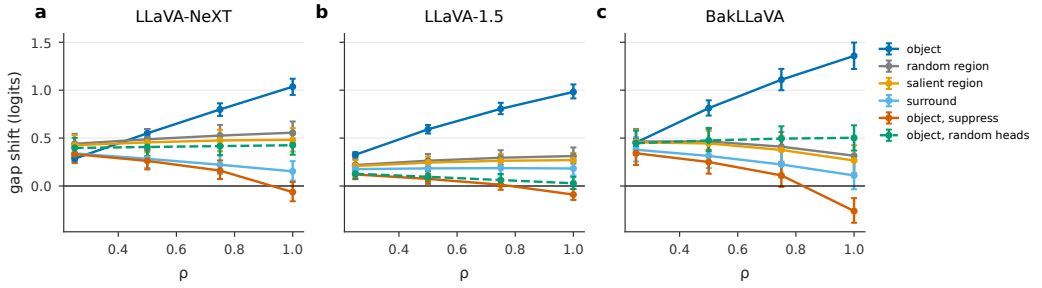

Figure 14: Every steering arm by dose on four models with bootstrap intervals.

($p = 0.14$), and not for saturation 1.5/2.0 or noise 20 on any model. Arms (Fig. 15): object (SAM3 mask), random region (matched token count), salient region (spectral-residual saliency), surround (annulus), object with random heads, object suppression, all heads, head-output scaling. WHOOPS at $\rho = 1$ (H, same masks): object $+1.07$, $+0.82$, $+1.45$; 10 heads $+0.64$, $+0.45$, $+0.85$; 5 heads $+0.26$, $+0.20$, $+0.40$; random region $+0.15$, -0.08, -0.07; mask confidence split $+1.00$ vs $+1.13$, $+0.70$ vs $+0.93$, $+1.46$ vs $+1.44$. Specificity contrasts (object vs each control) $p < 10^{-4}$ on all decoder-only backbones in the 07 runs; suppression monotone negative; all-heads steering breaks the model. Head-level necessity (mean ablation at the answer position; LLaVA-NeXT / LLaVA-1.5): top-5 flips 2.5/1.8% of image-wins (gap $-0.27/-0.18$), top-10 12.7/2.3% ($-0.79/-0.32$), top-20 16.9/7.0% ($-1.12/-0.87$); layer-matched random heads 0.8–2.5% (-0.02 to -0.06); prior-wins flip 0–0.4% and shift -0.09 to -0.47. Heads' write projection: 0.315/0.303 of the answer-position output norm inside PCA-256 vs 0.248/0.249 for a random 256-D subspace. Matched-area random region (gray, WHOOPS, n = 364/361): object $-0.62/-0.65$, rest $-0.30/-0.33$, random region of the object's area $-0.24/-0.23$. Injection controls (gap shift at $\alpha = 0.5$; LLaVA-NeXT, LLaVA-1.5, BakLLaVA): a discriminant direction fit to classify the decision on held-out conflicts -0.01, -0.13, -0.07 (inert); an unrelated concept direction -0.05, -0.11, $+0.04$; the answer-concept direction $+2.61$, $+1.98$, $+2.30$. In-range check: injected concept projections sit 0.5–0.6 standard deviations above the natural mean, with 4–8% beyond the natural 99th percentile. Unnormalized injection components at $\alpha = 0.5$ (full/in/out): $+2.82/+1.00/+2.02$ and $+2.25/+0.60/+1.78$; random-subspace split $+0.20/+2.67$ and $+0.17/+2.12$. BakLLaVA did not pass the notebook's self-test in this pass and is not reported for these four items. Transfer to ViLP and Visual CounterFact color: $+1.24$, $+1.30$, $+2.32$, with random regions also positive ($+0.70$, $+0.75$, $+1.40$) because the object fills the frame; grayscale baselines lower the gain (NeXT $+0.66$ at saturation 0). Size split: object -0.93, -0.28, -0.93; surround positive on LLaVA-1.5 and BakLLaVA; random region positive on all. Qwen2.5-VL: object $+1.61$ at full dose; object steering with random heads $+0.30$ (20% of the effect); random region $+0.18$; no discernibility interaction. LLaMA-3.2-Vision: measurement only (heads at cross-attention layers 23–38, $r = 0.45$; attention mass to the image predicts trust on ViLP, +0.29).

Table 4: Attention steering on WHOOPS (same masks, same conflicts as the factorial): gap shift toward the image by arm. Salient, surround, random-head, and suppression arms in App. G.

Model	Object, 20 heads	Object, 10 heads	Object, 5 heads	Random region
NeXT-Mistral	+1.07	+0.64	+0.26	+0.15
LLaVA-1.5	+0.82	+0.45	+0.20	−0.08
BakLLaVA	+1.45	+0.85	+0.40	−0.07
Qwen2.5-VL	+1.61	–	–	–

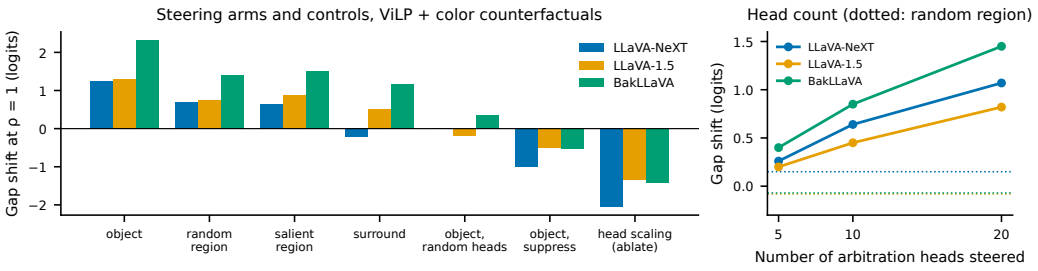

Figure 15: Steering arms and controls at full dose on ViLP and color counterfactuals (left), and the head-count ablation on WHOOPS (right); dotted lines are the random-region arm.

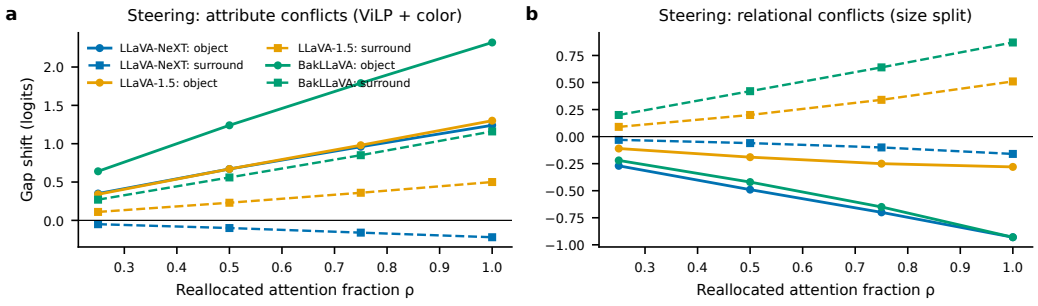

Figure 16: Attention steering onto the object (solid) vs the surround (dashed) by dose, on attribute conflicts (a) and on the relational size split (b), where the object arm inverts.

H TRUST EQUATION AND PREDICTION ACROSS SOURCES

Additive fit on half the WHOOPS examples, R^2 on the other half: NeXT $\Delta\text{gap} = -0.64\,\text{gray} - 0.86\,\text{blur} - 0.66\,\text{small} + 1.20\,\rho_{\text{obj}} + 5.21\,\alpha - 0.41$ ($R^2 = 0.40$); LLaVA-1.5 -0.36, -0.38, -0.23, $+0.94$, $+3.67$ (0.27); BakLLaVA -0.02, -0.32, -0.21, $+2.17$, $+4.86$ (0.19). Interactions ($p < 0.01$, all models): attention $\times$ injection $+2.7$, $+2.9$, $+7.7$; edit $\times$ injection positive. The equation does not predict effect magnitudes on other sources (Spearman 0.12): grayscale on color counterfactuals is -2.9 against a predicted -0.6, and the object term reverses on the size split.

I UNIVERSALITY AND PORTABLE DIRECTIONS

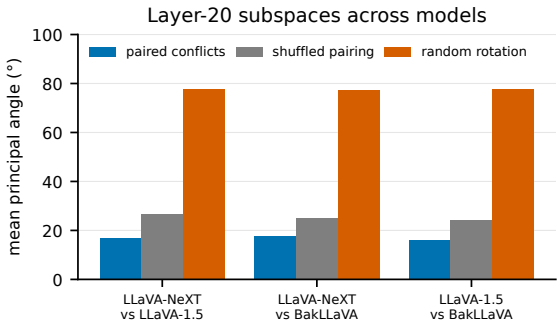

Figure 17: Principal angles between models' layer-20 subspaces after Procrustes, against shuffled pairing and random rotation.

Principal angles between models' layer-20 PCA-256 subspaces after orthogonal Procrustes on paired conflicts (NeXT–1.5, NeXT–BakLLaVA, 1.5–BakLLaVA): 16.7°, 17.6°, 16.2°; shuffled pairing 26.8°, 25.1°, 24.3°; random rotation 77.6°. Per-example unembedding-difference directions ported through the same map have cosine 0.01–0.04 with the target's own direction (random 0.016) and move the target by $+0.52$ (into LLaVA-NeXT), $+0.43$ (into LLaVA-1.5), $+0.21$ (into BakLLaVA) at $\alpha = 0.5$ (the target's own direction in that run: $+2.63$, $+1.97$, $+2.30$; random ≈ 0). Principal angles are for the pairs NeXT–1.5, NeXT–BakLLaVA, 1.5–BakLLaVA. The three $\alpha = 0.5$ full-injection figures in this paper ($+2.82$/$+2.25$ natural norms, $+2.61$/$+1.98$ renormalized, $+2.63$/$+1.97$ here) come from three runs on overlapping item subsets; each comparison is within its run.

J DETECTOR AND CALIBRATION

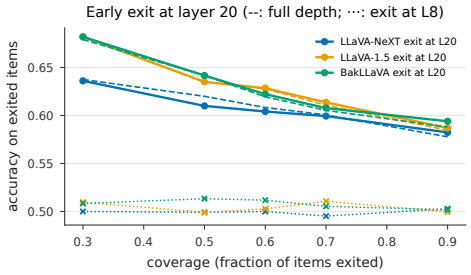

Figure 18: **Early exit at layer 20.** Accuracy of the exited answer on the items the centered margin is most confident about (solid) against full depth on the same items (dashed) and an exit at layer 8 (dotted), by coverage.

Untrained margin AUROC on WHOOPS by layer (LLaVA-NeXT): 0.51 at layer 12, 0.67 at 16, 0.80 at 20, 0.79 at 22, 0.72 at 28. Early exit on NaturalBench uses the centered layer-20 margin and exits

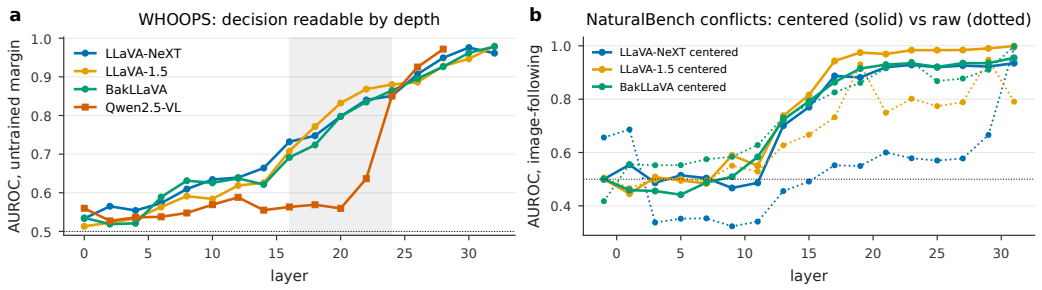

Figure 19: Untrained margin AUROC by layer on WHOOPS (four models; shaded band 16–24) and on NaturalBench conflicts, centered (solid) and raw (dotted).

when its magnitude exceeds the coverage quantile; Table 5 gives the exited accuracy against full depth on the same items, with a random-layer control at matched coverage (one draw per row, the layer in parentheses). The correctness-AURC (lower is better) of the centered layer-20 margin is 0.368, 0.348, 0.350 against 0.357, 0.335, 0.341 for the final margin and 0.436, 0.418, 0.434 for random ordering. Trained last-position probes: ViLP 0.86, Visual CounterFact 0.85–0.88, WHOOPS

Table 5: Early exit at layer 20 on NaturalBench: exited accuracy / full-depth accuracy on the same items, and a random-layer exit at matched coverage.

Coverage	LLaVA-NeXT	LLaVA-1.5	BakLLaVA
30%	0.636 / 0.637 (L26: 0.654)	0.682 / 0.682 (L26: 0.675)	0.682 / 0.679 (L26: 0.686)
50%	0.610 / 0.620 (L18: 0.619)	0.635 / 0.635 (L18: 0.636)	0.642 / 0.642 (L18: 0.631)
60%	0.604 / 0.608 (L16: 0.569)	0.628 / 0.628 (L16: 0.582)	0.622 / 0.619 (L16: 0.588)
70%	0.599 / 0.601 (L8: 0.495)	0.614 / 0.611 (L8: 0.511)	0.608 / 0.605 (L8: 0.505)
90%	0.582 / 0.578 (L10: 0.504)	0.587 / 0.585 (L10: 0.499)	0.594 / 0.588 (L10: 0.509)

0.73, 0.60, 0.70. Selective prediction (NeXT): the trained probe calibrates better (AURC 0.09 vs 0.23 on ViLP; 0.14 vs 0.19 on WHOOPS). Benchmark error prediction with the mid-depth yes/no margin: POPE 0.62, 0.57, 0.74; MME 0.5–0.6; HallusionBench ≈ 0.5; post-hoc final margin 0.75–0.84.

K BENCHMARKS

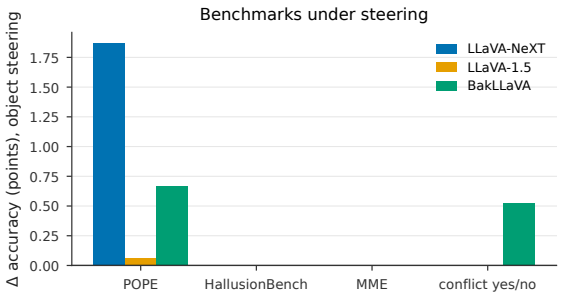

Figure 20: Accuracy change under object steering per benchmark and model.

POPE, HallusionBench, MME with object steering at $\rho \in \{0.25, 0.5\}$ and concept injection at $\alpha \in \{0.5, 1\}$; CHAIR on 500 COCO captions. Object steering at full dose changes POPE by $+1.9$, $+0.1$, $+0.6$ points and leaves HallusionBench and MME unchanged (Table 6); concept injection improves conflict yes/no accuracy by 6.3 points on BakLLaVA (prior-override splits +11–12), $+0.1$ on LLaVA-NeXT, and 0 on LLaVA-1.5; CHAIR_s 0.110→0.102 on NeXT, higher on LLaVA-1.5, mixed

Table 6: Benchmark accuracy, baseline → object steering at full dose (answer position), from the 10c/10d runs; conflict yes/no under concept injection at α=1 in the last column.

	POPE	HallusionBench	MME	Conflict yes/no	Conflict yes/no, injection
LLaVA-NeXT	0.697 → 0.716	0.485 → 0.485	0.617 → 0.617	0.744 → 0.744	0.744 → 0.745
LLaVA-1.5	0.741 → 0.742	0.484 → 0.484	0.652 → 0.652	0.786 → 0.786	0.787 → 0.787
BakLLaVA	0.743 → 0.749	0.463 → 0.463	0.677 → 0.677	0.727 → 0.733	0.728 → 0.791

on BakLLaVA. Concept rows on benchmark items without a candidate answer pair are undefined and excluded.

L NATURALBENCH INSTRUMENT TEST

Scoring. For each item we compute the layer-20 logit-lens margin $m_i = z_{\text{yes},i} - z_{\text{no},i}$ at the answer position. The raw readout is $s_i = p_i m_i$ with $p_i = +1$ if the image-supported (correct) answer is yes and -1 otherwise, so that a high score predicts that the model follows the image; the target is whether the model's final answer equals the image-supported answer. The centered readout is $s_i = p_i(m_i - c)$ with $c = \text{median}_i\, m_i$ over the evaluation items; because p_i is item-dependent, subtracting c before orienting changes the ranking across yes and no items, which is why centering raises AUROC on LLaVA-NeXT, whose m_i carries a large negative offset. The orientation p_i is used only to define the evaluation target; the early exit and the gate use $m_i - c$ alone and predict yes when it is positive. The trained probes use the same target and the same folds. Scorer validation: on 200 items per model,

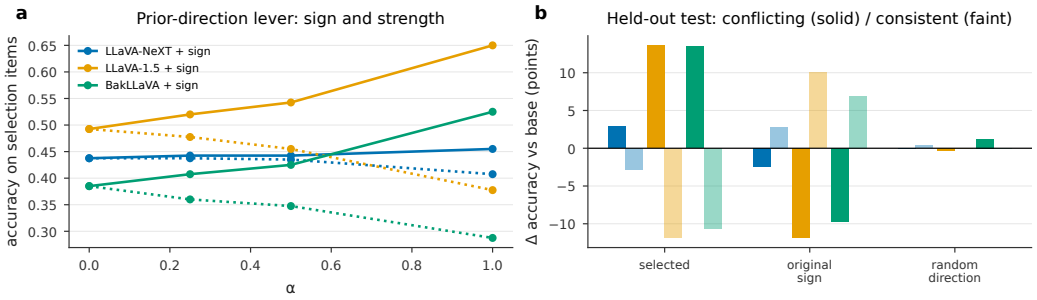

Figure 21: The polarity-bias lever by sign and strength on the selection items (a) and held out with controls, in accuracy points (b).

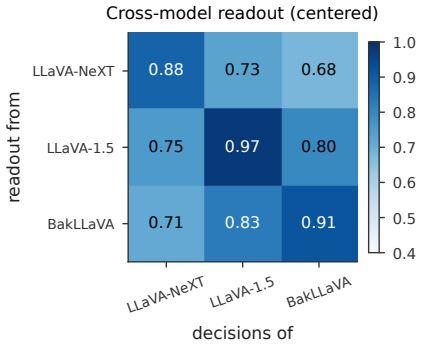

Figure 22: One model's centered readout predicting another model's decisions on the same photographs (rows: readout from; columns: decisions of).

the in-context yes/no scorer's sign agrees with the greedy generated answer on 99.5%, 100%, 98%; the statement-continuation scorer used for the synthetic sets agrees on 62%, 69%, 66.5% and is therefore not used for yes/no benchmarks (its earlier NaturalBench numbers, 0.78/0.65, 0.78/0.57, 0.82/0.55 and raw AUROC 0.88–0.92, are superseded). With the validated scorer: yes-rates 0.33, 0.52, 0.66; accuracy 0.72/0.43, 0.69/0.47, 0.77/0.40; raw layer-20 AUROC 0.55, 0.93, 0.86; centered 0.88, 0.98, 0.92; text-only-subtracted 0.82, 0.95, 0.93. Depth curves (hidden-state index): NeXT 0.66, 0.34, 0.35, 0.46, 0.55, 0.60, 0.58, 0.67, 0.99 at 0, 4, 8, 14, 20, 22, 26, 30, 32; LLaVA-1.5 0.51 to 0.93 at 20 and 0.95 at 30; BakLLaVA 0.42 to 0.86 at 20, 0.94 at 24, 1.00 at 32. Within-polarity AUROC 0.95/0.92, 0.99/0.97, 0.95/0.96. Trained 5-fold probe 0.64, 0.66, 0.75. Held-out centering (median from a disjoint half): 0.886, 0.976, 0.947 (n = 615, 587, 579). Non-default-prior subset: 0.916, 0.985, 0.948. Text-only control: 0.847, 0.907, 0.935. 1-D margin probe, same folds: 0.478, 0.930, 0.861; 2-D (margin, text-only margin): 0.850, 0.967, 0.953; final-layer margin 1.000. Gray-blank arbiter: answers change under a gray image on 0.320, 0.495, 0.530 of all items (0.332, 0.487, 0.517 of natural conflicts); centered readout AUROC on image-dependent conflicts 0.963, 0.967, 0.812

(n = 398, 584, 621; follows-image rates 0.588, 0.452, 0.343) and on image-independent conflicts 0.816, 0.983, 0.980 (n = 802, 616, 579). Cross-model, centered with each source model's own median (readout from → decisions of): NeXT→LLaVA-1.5 0.73, NeXT→BakLLaVA 0.68, LLaVA-1.5→NeXT 0.75, LLaVA-1.5→BakLLaVA 0.80, BakLLaVA→NeXT 0.71, BakLLaVA→LLaVA-1.5 0.83 (own: 0.88, 0.97, 0.91); an earlier uncentered computation gave 0.46–0.83 and is superseded. Object-area correlation $r = +0.04, -0.00, -0.05$ (n.s.). Size recomposites: conflicting -0.13, -0.07, -0.09 at $0.5\times$. Prior-direction sign test (400 conflicting items per model): base 0.438, 0.492, 0.385; original sign 0.435, 0.455, 0.347; opposite sign 0.443, 0.542, 0.425. Gated interventions ($\alpha = 0.5$): all changes within -4.7 to $+2.4$ points. Held-out lever ($+\alpha = 1$ selected on 400 conflicting items): conflicting base/always-on/gated 0.421/0.450/0.450, 0.464/0.600/0.551, 0.400/0.535/0.446 ($n = 800$); consistent 0.717/0.688/0.685, 0.690/0.572/0.612, 0.768/0.662/0.723 ($n = 1{,}200$); random direction 0.421/0.460/0.412 on conflicting (flat); all items base/always-on/gated 0.599/0.593/0.591, 0.600/0.583/0.588, 0.621/0.611/0.613. NaturalBench v2 (dependency-parsed nouns, 63% mask coverage): object area $r = +0.07, -0.01, -0.07$ (n.s., n = 797, 746, 706); gated object steering $+0.0$ to $+0.7$; $\alpha = 0.5$ lever conflicting $+0.3$, $+5.1$, $+4.0$ always-on and $+0.2$, $+2.4$, $+1.1$ gated, consistent $+0.1$, -4.4, -3.5. Early-exit values are in App. J; head-necessity, matched-region, and injection-component values in App. G.

M Related-work deltas

Table 7: Positioning, one row per paper: what it measures and what this paper adds.

Paper	Measures	This paper adds
Ortu et al. (2026)	Counterfactual heads by logit inspection on WHOOPS-AHA; attention reweighting; head attention localizes the region	The quantity those heads move; object-level edits with matched-area and recomposite controls; the band map; the projector-only checkpoint; the real-photograph readout
Lietzow et al. (2026)	Prior-promoting head circuit; ablation flips 68–96% of prior-grounded vs 0.8–7.5% of visually grounded predictions (prompt contrast); a critical window at 52–76% depth	Graded head necessity under the image contrast (17%, 7% at 20 heads) and subspace-level necessity, both weaker; the mirror asymmetry
Nooralahzadeh et al. (2026)	Logit-lens crossover; image tokens carry the causal impact; last-token patching 0–1%; final logit gap predicts grounding	Per-band, per-site sufficiency on the same image; the reconciliation of last-token flip rates by layer; the mid-depth readout on real photographs and its text-only control
Golovanevsky et al. (2025)	Prior-to-visual shift in middle-to-late layers; PvP steering vectors move 99% of color predictions	Single-direction injection as a sufficiency check rather than a steering method; the size-split inversion
Salazar et al. (2026)	Image-token to final-token pathway by layer on LLaVA-1.5-7B and two others	Recovered fraction per band at both sites on the same edited image; the crossover band; the Qwen ordering
Palacios et al. (2026)	Stage-1 vs tuned LLaVA-1.5; tuning's effect in layers 6–13	The readout and edit sensitivity on the stage-1 checkpoint
Chan et al. (2026)	LLaVA-1.5 yes-bias on NaturalBench from attention imbalances in layers 25–32	The same default read as a polarity offset at layer 20; the arbiter split; the lever that acts through it
Liao et al. (2026)	Gray-blank arbiter; decodable is not grounded	The arbiter applied to the outcome protocol and to the NaturalBench readout
Makelov et al. (2024); Zhang & Nanda (2024)	Subspace illusions; patching metrics and corruption	The controls adopted; the in-range check; the whole-update comparison
Kogilathota et al. (2026)	Hallucination probes at about 0.9 AUROC	An untrained arbitration readout that is not a hallucination detector (POPE 0.57–0.74)

Ortu et al.: heads by logit inspection, attention reweighting, SAM-3 attention validation on one CLIP model; here the quantity those heads move, its interface origin, its transport, three converging levers plus an injection check, and real-photo readout. V-SEAM: concept-level region edits with head modulation for VQA accuracy; here the same-image residual patching that links edits to a concentrated subspace, and the conflict-trust outcome. Gaze Heads: head-subset masking redirects description; here steering tied to the trust quantity and its controls, with the frozen-encoder contrast noted. Lietzow et al.: head-level necessity and the vision-default asymmetry; here coordinate-level necessity as measured, weaker. Liao et al.: the gray-image arbiter; applied here to validate the outcome protocol and stated as the reason decodability is not the necessary evidence. Nooralahzadeh et al.: encoding–grounding dissociation with the final logit gap as predictor; here the mid-depth margin as a pre-answer readout, its centering, and its transfer to real photographs. HALP: hallucination probes; here an arbitration signal that is deliberately not a hallucination detector. Makelov et al. and Zhang & Nanda: the controls adopted (residual-stream site, in-distribution counterfactuals, logit-difference metric, subspace decomposition, whole-update comparison).

N DESIGN CHOICES AND CONTROLS, IN REVIEWERS' TERMS

Why edited-to-clean patching, and what it measures. The same image, edited, is the counterfactual; patching the edited copy's residuals into the clean run is a noising patch, so a recovered fraction says the edited activation at that site suffices to reproduce the edit, equivalently that the clean activation there is needed for the clean outcome. Noise or cross-example corruptions were avoided on the recommendation of Zhang & Nanda (2024); cross-example interchange is reported separately at the answer position.

Why the gap, and its scale. The gap is the logit difference between the two candidate answers at the answer position; its sign is the decision. Its scale is given by the vision-win companions in §2 (grayscale moves vision-win from 0.78 to 0.69 on LLaVA-NeXT). The gap is not a sufficient statistic for every ablation: the whole-update ablation on LLaVA-NeXT flips 5.4% of image-wins and 6.3% of prior-wins with a mean shift near zero, which is why flip rates are reported alongside shifts.

Why layer 20, and the hidden-state offset. Layer 20 was fixed on the LLaVA-NeXT WHOOPS depth curve in the prior study and not re-tuned; the NaturalBench notebook indexes hidden states from the embeddings, so its index 20 is decoder layer 19 in this paper's convention. The centered curve plateaus across 18–24 on all three models, so the choice is not delicate within that band.

Why median centering, and what the checks show. The margin carries a per-model answer-polarity offset on NaturalBench. Subtracting the median over the evaluation items is label-free; estimating the median on a disjoint half changes the AUROCs by at most 0.03; and the within-polarity AUROCs (0.92–0.99) show the ranking exists before centering.

Why a trained probe underperforms the untrained direction. The unembedding difference is one direction in the 4,096-dimensional residual, so a probe could in principle recover it. Two things stop the 5-fold probe: with 1,200 items and L2 regularization at $C = 0.01$ it is strongly shrunk, and the class-conditional distributions of the margin differ by answer polarity, so a single threshold fits neither polarity (a 1-D probe on the margin itself with the same folds reaches only 0.48, 0.93, 0.86). Supplying the polarity as a second feature (the text-only margin) lifts the 2-D probe to 0.85, 0.97, 0.95, matching the centered readout. The readout's advantage is that it fixes the direction and needs only an offset; the probe has to learn both from 1,200 items in 4,096 dimensions.

What the text-only control does and does not show. The same lens, run without an image, predicts the language model's own text-only answer at 0.85–0.94; those are AUROCs for a different target than the image-run decision. The comparison shows the lens is a general property of the language model; it does not by itself say what the image run adds. The gray-image arbiter split does: 0.96 and 0.97 on the image-dependent items of LLaVA-NeXT and LLaVA-1.5, 0.81 on BakLLaVA.

Why the lever's sign inverted, and why it is not selection-free. Suppressing the model's text-only answer direction at image tokens lowered conflicting-item accuracy; the opposite sign raised it. The direction reaches the answer through attention value projections whose sign need not match the unembedding; we have no further account. Sign and strength were selected on 400 items and tested on the remaining 800 with random-direction and text-site controls; the gain held out on two models and was offset by consistent-item losses. One hypothesis for the inversion: the unembedding difference points along the direction that *writes* the answer at the final position, while at image tokens the direction that the arbitration heads' value projections carry forward need not align with it, and the reported steering effects say that pulling attention onto the image raises the margin; adding the prior direction at image tokens may simply make those tokens look more like the prior to the heads that read them.

BakLLaVA. The final-sweep notebooks open with a three-example self-test that patches layer 0 alone; BakLLaVA failed it (recovery 0.24–0.80) while the band-map notebook's identity checks passed and its L0–3 band recovers 1.00 on average. The four final-sweep controls are therefore reported on two models; BakLLaVA's grayscale contrast also fails Holm correction ($p = 0.14$) and its injection dose is non-monotone.

Qwen2.5-VL version sensitivity. Three outcome regenerations under two transformers minor versions gave valid counts of 356/457/345 and 477/574/407 and vision-wins of 0.79/0.34/0.48 and 0.98/0.57/0.79 by source. The object-level battery, band map, and readout by layer are reported on the later run and replicate the LLaVA ordering; the outcome statistics of this model should be read as version-dependent.

Clustering and multiplicity. Confidence intervals in the main tables are per-row bootstraps; only the mediation analysis clusters on prompts. The size split has 727 rows over 410 objects and ViLP has 200 questions over 600 images, so the intervals there are narrower than a cluster bootstrap would give. Inside-versus-outside contrasts are Holm-corrected across the edit grid.

Two held-out designs. The half-split design fits the subspace on one half of WHOOPS and evaluates on the other; the cross-source design fits on the non-WHOOPS conflicts with pooled captures and evaluates on WHOOPS, and is the one used for the variance-aware comparison.

O QUALITATIVE EXAMPLES

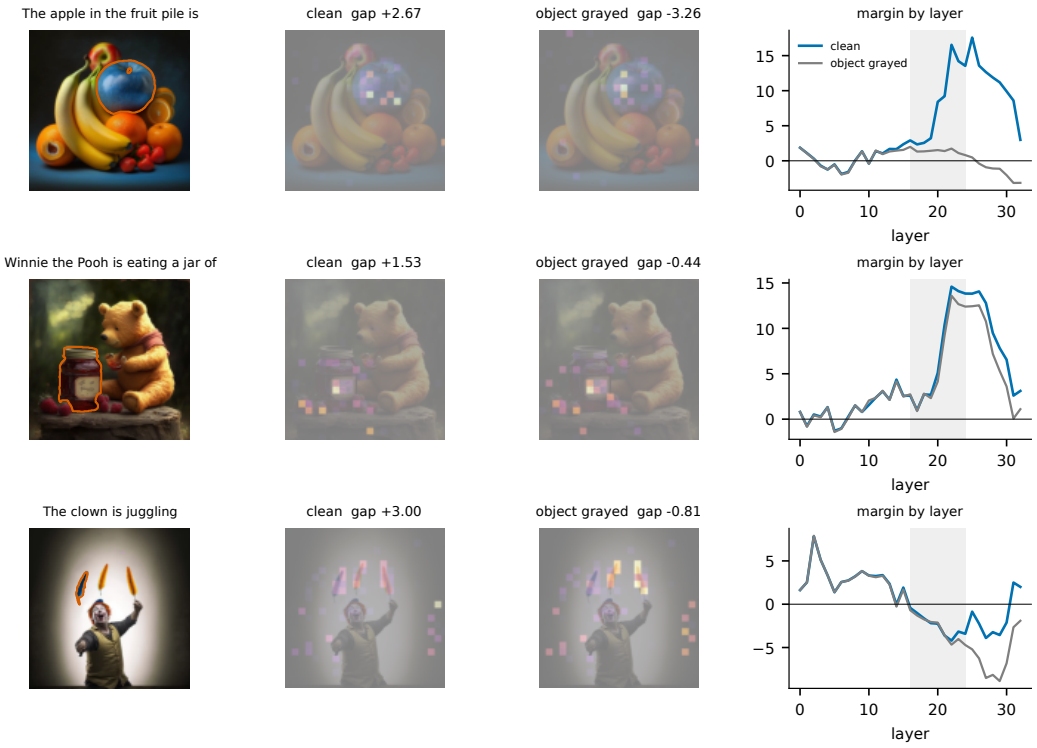

Figure 23: The arbitration heads on real WHOOPS examples (LLaVA-1.5): the clean image with the SAM3 object outlined and the statement; the heads' mean attention from the answer position over image tokens on the clean image and on the object-grayed image, with the gap above each; and the logit-lens margin by layer for both conditions (band 16–24 shaded). The heads attend to the contradicting object, graying it flips the decision, and the two margins separate inside the band where the heads act. Examples are chosen to illustrate the mechanism, not to estimate its size.



Figure 24: NaturalBench pairs (LLaVA-1.5): the same question on two photographs; each title gives the model's answer, the truth, its text-only prior, and the centered layer-20 readout (blue when correct, orange when the prior overrides the image).

P LIMITATIONS, EXTENDED

Confidence intervals in the main tables are per-row bootstraps that do not cluster on repeated objects in the size split or repeated questions in ViLP; only the mediation analysis uses a cluster bootstrap. Typical gap magnitudes per model and source are not tabulated in this version; vision-win companions in §2 give the scale for the headline shifts. Not completed before submission: steering, the band map, and subspace injection on the projector-only checkpoint; the union-mask test on the relational split; and, on BakLLaVA, the head-necessity, matched-region, natural-norm injection, and variance-matched subspace controls: the final-sweep notebooks open with a three-example self-test that patches layer 0 alone and requires near-complete recovery, and BakLLaVA failed it (0.24–0.80). The BakLLaVA numbers reported in Table 2, Fig. 6, and App. D come from the earlier same-image and band-map notebooks, whose hooks passed their identity checks and whose values are averages over about 420 conflicts (band L0–3 recovers 1.00 there); the two are not in conflict, since a single-layer, single-example check is stricter than a band average, but we flag it; a read-space versus nullspace decomposition of the injected direction was attempted and not reported because the next layers' input projections are full rank. The transport map is measured for grayscale and blur on WHOOPS; the relational (size) split's union-mask prediction is registered and unrun; enlargement is confounded with legibility; the stage-1 comparison uses one assembled checkpoint of one architecture; NaturalBench scorers are polarity-biased and the noun heuristic for masks is crude; models are 7–11B; cross-attention steering is not implemented.