THEY KNOW WHERE IT POINTS, BUT NOT WHAT IT MEANS: THE CAUSAL GRAMMAR OF RELATIONAL VISUAL PROMPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graphical cues can identify targets and express relations. We test whether vision-language models use an arrow to attribute speech after locating its endpoint. Across 450 paired conditions from 150 scenes, five open VLMs locate endpoints 8.5–13.1 percentage points more accurately than they attribute speakers, even when given the arrow-to-speaker rule. Adding the cue changes candidate scores, and decoding these changes improves speaker attribution by 8.5–15.7 points. Complete three-target operator orbits recover relation-sensitive responses when raw answers remain unchanged, outperforming matched shams on a controlled benchmark and authored comics. A second directed relation reproduces the behavioral gap and cue-specific recovery. Activation replacement finds an early visual effect that fades as an answer-position effect emerges. Under a reversed semantic codebook, predictions remain more closely aligned with the physical operator than with the instructed mapping. These results distinguish reference grounding, sensitivity to the operator, and use of the relation to select an answer.

1 INTRODUCTION

Arrows, boxes, points, and scribbles can direct a vision-language model's attention (Shtedritski et al., 2023; Yang et al., 2023; Cai et al., 2024). A connector or speech-bubble tail also expresses a relation between the elements it joins. We call identifying the marked target reference grounding and using the connection's meaning relation use. A model may succeed at the first while failing at the second. A wrong relational answer leaves open whether the model missed the relation or represented it without using it to answer.

We study this distinction through speaker attribution. On the same prompted images, five open VLMs locate arrow endpoints 8.5–13.1 points more accurately than they attribute speakers; all five scene-bootstrap intervals exclude zero. Stating the arrow-to-speaker rule does not close the gap. As Figure 1 illustrates, adding a cue can increase the intended speaker's score most while another candidate remains the final answer. We call the candidate-specific change the *graphical prompt effect*. Decoding it improves speaker attribution by 8.5–15.7 points.

To test whether scores track the relation, we vary the operator across all three valid targets while holding the scene fixed. These *complete orbits* reveal cases of relational aliasing: categorical answers stay the same as candidate scores follow the operator. Matched shams and endpoint and label interventions test simpler visual and positional explanations. A second directed relation reproduces the behavioral gap and prompt-effect recovery; natural comics provide a separate domain test.

Activation replacement tests whether operator information affects the answer. On a separate mechanism set, an early effect at edited visual tokens fades as an answer-position effect emerges. When we reverse the instructed endpoint-to-answer mapping, predictions remain more closely aligned with the physical operator.

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.

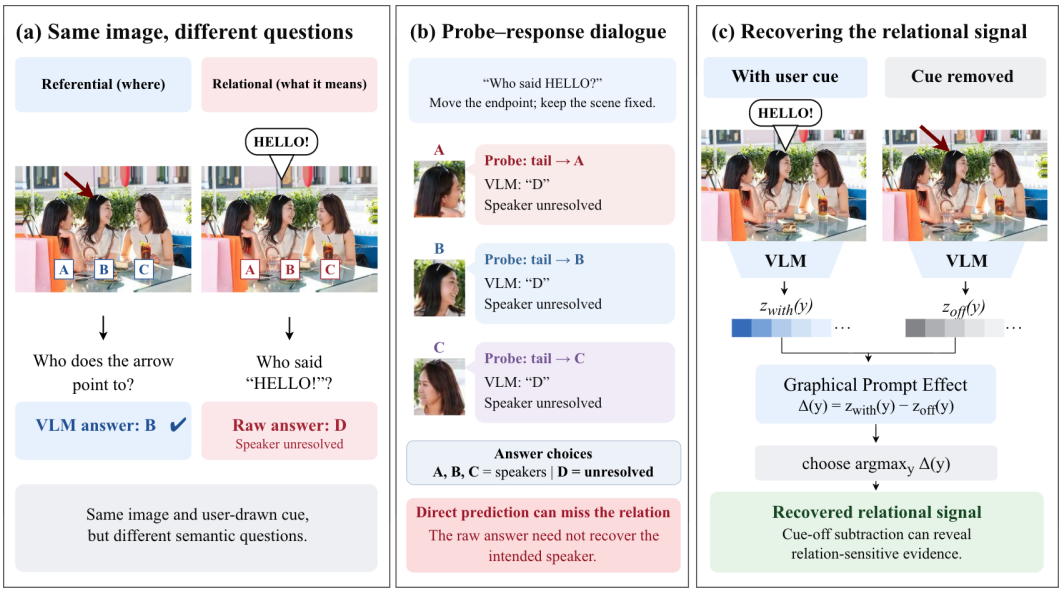

Figure 1: A graphical cue can change scores without changing the final answer. (a) The same user-drawn cue is easier to resolve as a referent than to use for speaker attribution. (b) Adding the cue can increase the intended target's score the most while another candidate remains the model's absolute prediction. (c) Contrasting the prompted and unprompted views isolates this cue-induced relational signal.

2 Related Work

Visual prompting and grounded multimodal models. "Visual prompting" includes both learned pixel prompts (Bahng et al., 2022) and human-readable marks drawn on an image. Red circles can steer CLIP toward a region (Shtedritski et al., 2023); Set-of-Mark assigns speakable IDs to regions (Yang et al., 2023); ViP-LLaVA trains an MLLM to consume points, boxes, scribbles, and arrows (Cai et al., 2024); and Kosmos-2, Shikra, Ferret, and Molmo strengthen region grounding or pointing (Peng et al., 2023; Chen et al., 2023; You et al., 2024; Deitke et al., 2025). VP-Bench separates visual-prompt perception from downstream use (Xu et al., 2026). These systems show that marks can help a model locate a referent. We test whether the model also uses the relation expressed by the mark.

Compositionality, equivariance, and shortcut-sensitive evaluation. Aggregate accuracy can conceal failures under controlled semantic changes. Winoground holds the words fixed while varying compositional structure (Thrush et al., 2022). SugarCrepe shows how artifacts can compromise a compositional benchmark when its hard negatives are not carefully designed (Hsieh et al., 2023). EqBen measures whether similarity responds faithfully to minimal semantic edits (Wang et al., 2023b), while MMVP tests basic visual distinctions that strong MLLMs miss (Tong et al., 2024). Work on object hallucination also documents strong answer priors in LVLMs (Li et al., 2023). Our within-scene interventions change the graphical operator while matching compute and foreground structure. A complete $K \geq 3$ orbit tests more than two valid targets, allowing us to distinguish relational aliasing across all views from a binary complement rule. ViP-LLaVA previously reported a 3,520-item bidirectional COCO arrow construction (Cai et al., 2024). We evaluate a separately frozen reconstruction with its own split and count convention.

Contrastive decoding and visual-prior correction. Visual Contrastive Decoding subtracts a distorted visual view (Leng et al., 2024); Contrastive Region Guidance compares prompted and region-ablated views (Wan et al., 2024); and CPCF contrasts target and misleading referring prompts (Zhu et al., 2025). These methods use logit subtraction and counterfactual views. Our two-view comparison changes only the user cue while fixing the scene and query; Equation 1 measures the resulting candidate-specific change. The K-way orbit addresses a different question: varying every

valid operator value removes candidate main effects and isolates the within-operator interaction without fitting a calibration model.

Hidden information and causal localization. Stable output can coexist with substantial changes in internal computation. Wani et al. show that answer invariance can hide representation drift under meaning-preserving image perturbations (Wani et al., 2026); VI-Probe studies persistence under inverted visual-illusion factors with matched controls (Sun et al., 2026); MMVU examines the broader gap between visual recognition and answering (Liu et al., 2025); and VLM-Fix changes standard and inverse textual rules, then partly repairs semantic fixation through activation steering (Alam, 2026). We use interchange-style causal interventions to test whether replacing an internal state changes the answer. The approach follows causal mediation and activation-patching studies (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022; Wang et al., 2023a; Zhang & Nanda, 2024) and, more broadly, interventionist causal analysis (Pearl, 2009). When replacing a state changes the counterfactual effect, that state carries a causal effect of the relational intervention.

3 Identifying Relational Sensitivity

A wrong relational answer may reflect either no response to the graphical relation or a response masked by a stronger candidate prior. We distinguish these cases with two interventions. First, we compare the scene before and after adding a cue. Then we vary the relational operator across all valid targets while holding the rest of the scene fixed.

3.1 Reference, relation, and the graphical prompt effect

Let I_0 denote a prompt-off image and I_g the same scene after the user adds a graphical cue g. A *referential* query asks for the cue target; a *relational* query asks for the semantic consequence of the connection. Absolute predictions cannot show whether a cue moves the intended candidate's score without overturning an existing answer prior. We measure the candidate-specific change caused by adding the cue.

For candidate y and fixed query q, define

$$\Delta_g(y) = z_y(I_g, q) - z_y(I_0, q), \qquad \hat{y}_{\mathrm{PED}} = \arg\max_y \Delta_g(y). \tag{1}$$

We call $\Delta_g(y)$ the graphical prompt effect and the corresponding training-free rule prompt-effect decoding (PED). The prompt-off image provides a relation-free baseline for the user's cue: in the speaker task, a tail-less speech bubble leaves the speaker unspecified. Subtracting its score removes preferences the model already had for the underlying scene.

3.2 A causal grammar and relational operator orbit

The paired contrast measures the response to one added cue. To test whether that response systematically tracks the *relational operator*, we vary the operator while holding the scene, participants, and rendering attributes fixed.

Complete operator intervention. Let a prompted image be $x = g(s, P, r, n)$, where s is the scene, P the participant set, r a relational-operator value, and n nuisance rendering. For a three-person speech balloon, $P = \{A, B, C\}$ and r_i selects person i. Here P denotes the valid participant set; a raw multiple-choice benchmark may also include a nonparticipant distractor, which is excluded when we report participant-restricted topology.

An aligned benchmark observes only $g(s, P, r_i, n)$ with target i. In this single view, operator following cannot be distinguished from proximity, salience, or a stable answer prior. A complete orbit instead evaluates every valid operator value $r_1, \ldots, r_K$ for the same underlying scene.

Orbit readout. For each factual operator value r_i, define

$$d_i(y) = z_y(r_i) - \frac{1}{K-1} \sum_{j \neq i} z_y(r_j), \qquad \hat{y}_i = \arg\max_{y \in P} d_i(y). \tag{2}$$

The contrast measures each candidate's response to operator r_i relative to the other valid relations in the same scene. It removes candidate preferences shared across the orbit and retains operator-dependent changes.

Matched structural control. For the evaluated $K = 3$ orbit, the primary control replaces the other two valid operator values with two structure-matched sham relations $s_{i,m}$ that point to no participant:

$$d_i^{\text{sham}}(y) = z_y(r_i) - \frac{1}{K-1}\sum_{m=1}^{K-1} z_y(s_{i,m}).$$

Orbit and sham decoding match the scene, candidate set, model-call count, and one visible graphical mark per view. Their difference tests whether recovery depends on varying a valid relational operator after accounting for generic foreground ink and multi-view centering. Sham centering is a structural control contrast, not the exact interaction decomposition in Equation 3.

3.3 Why every raw answer may be wrong

The orbit contrast removes candidate-specific main effects and retains the component that varies with the relational operator. Categorical predictions can therefore remain unchanged while candidate scores track the relation.

For response matrix Z, let $\bar{z}_y = K^{-1}\sum_j z_y(r_j)$ and $\epsilon_i(y) = z_y(r_i) - \bar{z}_y$. Then

$$d_i(y) = \frac{K}{K-1}\epsilon_i(y). \tag{3}$$

Orbit decoding is an exact main-effect/interaction decomposition. A shared candidate main effect can dominate every raw categorical answer, while a smaller operator-dependent interaction still ranks the relationally correct candidate first.

This regime occurs empirically. Across Synthetic-51 and fixed-red COMICORDA, LLaVA answers D on all 297 views, yet the interaction ranks the correct speaker first among A/B/C on 197/297 (66.3%) views, compared with 105/297 for participant-restricted raw logits and 146/297 for the structure-matched sham. Removing the distractor or centering against additional views does not account for this recovery.

3.4 What response structure alone cannot establish

Operator-sensitive response structure shows that model scores track changes in the graphical relation. Without labels, it cannot determine the *semantic orientation* of that structure. Replacing the intended mapping f with $\pi \circ f$ for a nonidentity candidate permutation leaves the unlabeled response matrix Z unchanged, although correctness can change from perfect to zero.

At least one semantic anchor is required to distinguish the two mappings. The orbit diagnoses relational sensitivity, while the anchor determines whether that sensitivity matches the intended semantics. Appendix Section F.2 gives a constructive counterexample.

4 Experimental Design

Primary controlled evaluations. The primary behavioral evaluation uses two controlled populations. The paired reference–relation bank contains 150 COCO scenes screened by multiple human reviewers (Lin et al., 2014). Each scene has a tail-less prompt-off image and three prompted views connecting the speech bubble to A, B, or C, yielding 450 prompted conditions. We ask endpoint (Q1), semantic-relation (Q2), and explicit-rule (Q3) questions with five fixed paraphrases and forced A/B/C scoring; majority vote is computed per image.

A separate frozen 51-scene benchmark provides the complete $K = 3$ operator intervention used to test whether model scores systematically track all valid relational targets. It contains 153 factual directions. For this controlled speaker benchmark, raw scoring includes D ("speaker unresolved"), whereas participant-restricted topology, sham, and orbit decoding use the same first-token logits over A/B/C.

Table 1: Paired reference–relation controls and recovery. Five paraphrases on 450 user-arrow conditions from 150 human-screened scenes. Gap is Q1−Q2; PED is prompt-on−prompt-off Q2 decoding.

Model	Q1 endpoint	Q2 relation	Gap	Q3 rule	Q2+PED
Qwen3-VL-4B	93.5	85.0	+8.5	77.8	95.4
Qwen3-VL-8B	93.5	80.4	+13.1	75.2	96.1
InternVL3.5-4B	85.6	76.5	+9.2	72.5	88.9
Phi-3.5-Vision-4B	87.6	76.5	+11.1	58.2	85.0
LLaVA-1.6-7B	56.2	47.1	+9.2	47.1	62.1

Generalization and scope populations. We test generalization across image domains and relational semantics. COMICORDA provides 48 authored comic panels with native balloon–speaker links (Martinek et al., 2024), testing whether operator-sensitive response structure extends beyond the controlled renderer. A separately frozen 80-scene CoMix/DCM population (Vivoli et al., 2024) provides a harder domain-shift stress test.

A separately frozen 100-scene directed-recipient benchmark changes the relational semantics themselves, with three factual sender→receiver directions per scene and participant-restricted A/B/C scoring. Additional scope checks use an independent 500-scene COCO sample containing 1,000 matched box/arrow targets and a separately frozen 150-scene ViP-style arrow bank for Raw/VCD/CRG/PED comparisons and evaluation of the released ViP-LLaVA-7B checkpoint. A distinct 1,760-image ViP-style reconstruction supports the headless-line audit in the appendix.

Models, inference, and statistics. Unless otherwise stated, the primary model suite comprises five open general-purpose checkpoints: Qwen3-VL-4B/8B (Bai et al., 2025), InternVL3.5-4B (Wang et al., 2025), LLaVA-NeXT (v1.6-7B, denoted LLaVA-1.6-7B below) (Liu et al., 2023; 2024a;b), and Phi-3.5-Vision-4B (Abdin et al., 2024). We use first-token forced-choice logits. The source scene is the unit of uncertainty: paired differences use exact McNemar or sign-flip tests as appropriate, and 95% intervals use scene-cluster bootstrap resampling. Confirmatory multi-model families use Holm correction (Holm, 1979). For the paired bank, the prompt-off argmax serves as the model-prior baseline; speaker ground truth is supplied by the rendered relation.

Matched controls. The $K = 3$ experiments compare raw prediction, participant-restricted topology, structure-matched sham centering, and orbit decoding. The sham contrast controls for generic foreground marks and multi-view centering. Six renderings vary color, width, text, JPEG quality, endpoint offset, and shaft length to test dependence on appearance. A frozen A/B/C label derangement breaks fixed position-to-option mappings, while a headless copy of each arrow shaft isolates the endpoint operator from the much larger shared line. Multiple human reviewers screened the controlled renderings for readable participants and endpoints.

5 Results

5.1 Reference grounding and relation use diverge

Table 1 compares endpoint localization (Q1) with speaker attribution (Q2) on the same images. Q1 accuracy exceeds Q2 by 8.5–13.1 points in every model, and all five 95% scene-bootstrap intervals exclude zero. Locating the cue does not guarantee correct speaker attribution. This behavioral gap alone cannot show whether the model fails to represent the relation or fails to use it when answering.

Explicitly stating the arrow-to-speaker convention (Q3) does not close this gap: accuracy is lower than Q2 in four of five models and unchanged in LLaVA. PED raises Q2 accuracy by 8.5–15.7 points in every model, with positive 95% intervals. The cue thus induces a target-specific score change even when the explicit textual rule does not improve the categorical answer. The semantic-codebook intervention below further tests whether the model binds the visual relation to the requested answer.

Across the 150-scene bank, the 95% CIs for the referential–relational gap are [5.6,11.4], [9.8,16.4], [6.7,11.8], [7.7,14.5], and [6.5,11.8] points, respectively, for Qwen3-VL-4B, Qwen3-VL-8B, InternVL3.5-4B, Phi-3.5-Vision-4B, and LLaVA-1.6-7B. PED-minus-Q2 intervals, in the same

Table 2: Complete operator intervention. Frozen 51-scene, 153-direction $K = 3$ speaker result (%).

Model	Raw	Topology	SHAM	ORBIT
Qwen3-VL-4B	73.9	77.1	89.5	96.7
Qwen3-VL-8B	43.1	57.5	81.0	94.1
InternVL3.5-4B	24.8	55.6	73.2	85.0
LLaVA-1.6-7B	0.0	37.3	59.5	82.4
Phi-3.5-Vision-4B	38.6	40.5	57.5	66.7

Table 3: Natural-comic replication. Fixed-carrier COMICORDA results on 48 authored panels; gain is orbit minus matched sham.

Model	SHAM	ORBIT	Gain [95% CI]	Holm p
Qwen3-VL-4B	68.8	80.6	+11.8 [6.3,17.4]	0.00170
Qwen3-VL-8B	77.1	84.0	+6.9 [1.4,12.5]	0.03139
InternVL3.5-4B	66.7	75.7	+9.0 [3.5,15.3]	0.02300
LLaVA-1.6-7B	38.2	49.3	+11.1 [4.2,18.1]	0.02300
Phi-3.5-Vision-4B	52.8	62.5	+9.7 [2.8,16.7]	.02436

model order, are also positive: [7.4,13.5], [12.4,19.0], [9.4,15.4], [4.4,12.6], and [10.7,19.1] points. These results are consistent with prior–prompt competition: the cue can raise the intended target's score while another candidate remains on top.

5.2 WRONG IN EVERY VIEW, RIGHT IN THE ORBIT

PED identifies the candidate whose score changes most when a cue is added. The complete $K = 3$ intervention tests whether scores track each valid operator target within the same scene. Table 2 shows that they do. Relative to matched sham, ORBIT improves accuracy by 7.2, 13.1, 11.8, 22.9, and 9.2 points across the five models; all five scene-clustered tests survive Holm correction. Appendix Table A1 reports the prespecified weak-model boundary.

LLaVA's raw categorical accuracy is 0% on this benchmark, yet orbit decoding recovers 82.4%. Its score changes track the relational operator despite every raw decision being wrong, as described by the main-effect/interaction decomposition in Section 3.3.

Appearance invariance and relation sensitivity separate. Across six renderings, Qwen3-VL-4B's orbit accuracy is 94.8–97.4% and LLaVA's is 73.9–84.3%. LLaVA's raw answers agree 100% because every answer is D and incorrect. Orbit predictions remain correct under all six appearances for 62.1% of directions, compared with 41.2% under sham centering. The scores retain relation-sensitive variation even though the categorical answers do not change.

5.3 DOMAIN AND RELATION TRANSFER

We next test the score pattern on authored images and a directed-recipient task.

Natural-domain replication. A fixed-carrier COMICORDA replication extends the complete-operator result to 48 authored comic panels. Orbit-minus-sham gains are +6.9 to +11.8 points; all five scene-bootstrap intervals exclude zero and all Holm-adjusted p values are below 0.05 (Table 3). The same red query carrier is used for all 48 panels and all models, with no model-specific carrier selected after scoring. The orbit advantage therefore extends beyond the controlled renderer.

A separately frozen 80-scene CoMix/DCM stress test provides a harder domain shift. Its pooled scene-macro orbit-minus-sham gain is +7.1 points with a 95% CI of [4.4,9.8], although per-model gains range from -3.3 points for LLaVA to +16.7 for InternVL (Appendix Table A2).

A second directed relation. A separately frozen 100-scene task uses a directed sender→receiver convention with three factual receiver candidates per scene. The referential–relational gap is positive in both evaluated architectures: +8.4 points [6.0,10.9] for LLaVA and +3.7 [0.7,6.7] for Qwen. PED also improves receiver attribution in both models, by +18.6 [14.9,22.3] and +3.6 [0.8,6.3] points,

Table 4: Directed-recipient replication. 100 frozen scenes, 300 factual directions. Gap and PED-gain CIs use the source scene as the paired uncertainty unit.

Model	Q1 endpoint	Q2 receiver	Gap [95% CI]	Q2+PED	PED gain [95% CI]
LLaVA-1.6-7B	46.8	38.3	+8.4 [6.0,10.9]	56.9	+18.6 [14.9,22.3]
Qwen3-VL-4B	48.9	45.2	+3.7 [0.7,6.7]	48.8	+3.6 [0.8,6.3]

respectively (Table 4). The endpoint–relation gap and cue-specific recovery also occur outside speaker attribution.

5.4 Operator sensitivity survives matched shortcut controls

Visual features or answer positions correlated with the operator could also produce an orbit advantage. We test these alternatives with matched interventions.

Structure-matched shams control for generic foreground ink and multi-view centering. A separate headless-line audit compares the intact arrow with its shaft on the 1,760-image ViP-style reconstruction (Appendix Section E). A frozen A/B/C label derangement breaks the original left/middle/right-to-answer mapping. Before rendering or scoring, every scene is assigned a nonidentity three-cycle so that each label moves to a different spatial position. Under this intervention, Qwen3-VL-4B retains a +5.9-point orbit–sham gain (96.1% vs. 90.2%, 95% CI [2.6,9.8], Holm $p = 0.00734$). The complete two-model control is reported in Appendix Table A6.

For Qwen, the orbit advantage persists after the label permutation; the LLaVA comparison is inconclusive (Appendix Table A6). Together with the sham result, this rules out a fixed spatial-label mapping as the sole explanation for Qwen's gain.

Decoder and model scope. On the frozen 150-scene ViP-style bank, CRG exceeds PED in aggregate for LLaVA-1.6-7B (73.1% vs. 57.1%), while PED exceeds VCD by +3.0 points [1.1,5.0] on the same model. The released ViP-LLaVA-7B checkpoint has nearly identical raw endpoint and relation accuracies (50.2% and 50.0%); PED raises relation readout to 62.1% (+12.1 [9.4,14.9]).

Across the tested images and relations, graphical cues change model scores even when categorical answers do not follow them. We next examine where the effect appears inside the model.

6 Mechanism and Repair

In conflict cases, the user cue raises the intended target's score while another answer remains on top, consistent with prior–prompt competition. The K-way decomposition separates a stable candidate component from the operator-dependent response. Activation replacement tests whether the counterfactual signal affects the answer.

Intervention logic. A readout can show that a hidden state contains information, but cannot show whether the model uses it to answer. For each factual/inverted image pair, we rerun the base forward pass with selected visual activations replaced by their counterfactual values. We measure the induced change in the output answer-margin contrast, fixing its direction before examining layers. This interchange intervention follows causal-abstraction and activation-patching work (Geiger et al., 2021; Meng et al., 2022; Zhang & Nanda, 2024). An all-vision effect concentrated in tokens whose image cells changed identifies a local carrier of the counterfactual effect.

6.1 From edited pixels to answer binding

On a separate 20-scene mechanism set, causal patching identifies an early visual carrier of the operator effect. Replacing all vision tokens at layers 2, 4, 8, and 12 strongly changes the answer margin. Restricting the replacement to tokens whose image cells contain pixels changed by tail inversion recovers 91.4%, 88.1%, 88.5%, and 90.4% of the all-vision effect. The edit mask covers 11–41 of roughly 192–380 vision tokens, with 92.5–95% of paired directions moving with the correct causal

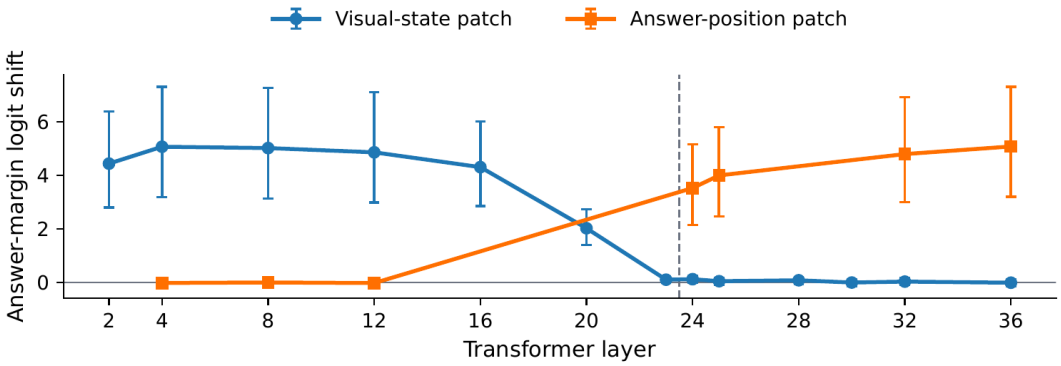

Figure 2: Layerwise patch effects in Qwen3-VL-4B. Counterfactual replacement of visual states has a large early effect on the answer margin and becomes negligible by layers 23–24. Counterfactual replacement at the answer position is null at early tested layers and becomes strongly effective from layer 24 onward. Error bars are source-scene bootstrap 95% CIs. Both curves use the same raw first-token answer-margin logit-shift metric.

Table 5: All-vision causal transfer on the 20-scene mechanism set. Mean shift is the raw first-token answer-margin logit shift. Peak normalizes within model to compare layerwise timing, not magnitude across architectures; correct sign is the fraction of paired directions moving with the counterfactual intervention.

Model	Layer	Mean shift [95% CI]	Peak	Correct sign
Qwen3-VL-4B	2	3.072 [1.903,4.322]	87.4%	95.0%
	4	3.503 [2.197,4.956]	99.6%	95.0%
	8	3.516 [2.191,4.975]	100.0%	95.0%
	12	3.425 [2.119,4.863]	97.4%	92.5%
LLaVA-1.6-7B	2	0.096 [0.073,0.117]	98.4%	90.0%
	4	0.098 [0.074,0.119]	100.0%	90.0%
	8	0.092 [0.070,0.113]	94.4%	92.5%
	12	0.084 [0.066,0.101]	86.0%	97.5%

sign. By layer 20, the local effect falls to 60%; by layer 23, its absolute magnitude is nearly zero. The answer-position patch effect appears after this decay (Figure 2).

Directionally consistent all-vision effects also appear in LLaVA on the same 20-scene mechanism set. At layers 2, 4, 8, and 12, 90–97.5% of directions have the correct causal sign, and every scene-clustered test survives within-model Holm correction ($p \leq 4.5 \times 10^{-5}$). Because raw logit scales are model-specific, absolute shifts are reported but not compared across architectures; peak normalization is used only to compare layerwise timing within each model. Qwen3-VL-4B supports exact edit-token localization, whereas LLaVA's multiresolution packing lacks an exact flat pixel–token map.

Matched causal controls and answer-position effects. A companion Qwen control run on the same 20 source scenes tests the specificity of this layerwise pattern. Size-matched random visual-token patches and self-patches have no measurable effect at any tested layer. Counterfactual answer-position patches are likewise null at layers 4, 8, and 12, but become strongly effective after the visual-token effect has faded: the mean answer-margin shift is 3.522 [2.144,5.160] at layer 24 and 5.075 [3.200,7.300] at layer 36 (scene-level sign-flip $p < 10^{-4}$ for both). The controls locate the measured effects at edited visual states in early layers and at the answer position in late layers. Representative results are in Appendix Table A4.

Semantics-reversing codebook. We next hold the image and physical endpoint fixed while explicitly permuting the endpoint-to-answer rule in the prompt with a frozen three-cycle. Under the normal codebook, physical- and semantic-orientation accuracy are both 65.0%. Under the reversed codebook, physical orientation remains 67.5% while semantic accuracy is 32.5%; predictions change on 20% of paired conditions, and only 5% switch from the normal answer to the reversed semantic

target. The responses remain more closely aligned with the physical relation than with the instructed remapping, showing that operator tracking does not ensure semantic answer binding.

7 Limitations and Discussion

The measurements in this paper answer different questions. Endpoint accuracy tests reference grounding. PED measures how the cue changes each candidate's score, including cases where the categorical answer stays fixed. Orbit decoding then asks whether scores track every valid operator value within a scene; matched nonparticipant marks separate this pattern from generic foreground ink. The three readouts locate different steps between seeing a mark and using it to answer. The codebook reversal adds a semantic test: responses remain closer to the physical operator than to the instructed target even when scores are operator-sensitive. The permutation argument explains why the mapping itself needs a labeled anchor (Appendix Section F.2).

PED and orbit remove different baselines. PED subtracts the unmarked scene, using the view before a user adds the cue. Orbit instead compares every valid operator variant within one scene, removing candidate preferences shared by those prompted views. This distinction matters if a prompt-off image is unavailable: orbit still provides a within-scene contrast when the operator can be edited. Conversely, PED needs only the paired before-and-after views. The two designs thus make different demands on the evaluation interface.

The evidence has specific limits. Sham views control foreground structure and model-call count, the headless-line audit probes the shared shaft on a separate population, and deranged labels test a spatial shortcut in Qwen. The edited images still differ locally, and the authored-comic replication does not cover every graphical style. Activation replacement establishes causal effects at tested visual and answer-position states, but does not trace a unique path between them. The mechanistic tests use two open VLMs on speaker attribution; the directed-recipient task broadens the behavioral evidence, not the mechanistic claim. PED requires candidate scores and a prompt-off image, while complete orbits also require editable operator values and a known participant set. These conditions fit controlled evaluations and some interactive interfaces. Finally, the frozen VCD/CRG comparisons show that PED is not the best decoder in every setting.

For a new graphical-prompt task, we recommend retaining a prompt-off view, varying the operator over valid targets, and including matched nonparticipant marks when editing is possible. Report raw answers alongside score contrasts; an instruction-remapping check can then test whether the response follows the requested semantics. Uncertainty should be clustered by source scene rather than treating edited views as independent. This separates reference grounding, operator sensitivity, and semantic answer binding.

8 Conclusion

Across five open general-purpose VLMs, endpoint localization exceeds speech attribution. A directed-recipient task shows a similar gap and PED recovery in Qwen and LLaVA. Cue-induced score changes and complete operator orbits reveal relation-sensitive responses masked by answer priors. The effect survives matched controls and extends to authored comics, while contrastive baselines and a visual-prompt-tuned checkpoint clarify its scope. Causal interventions find early visual and later answer-position effects. Under codebook reversal, predictions follow the physical operator more often than the instructed mapping. These results motivate separate evaluation of reference grounding, operator sensitivity, and semantic answer binding.

Reproducibility statement

Appendix Sections A–F report the frozen evaluation populations, matched controls, response decomposition, and causal-intervention details.

AI use statement

Generative AI tools assisted with manuscript drafting and language editing. The authors reviewed the AI-assisted text and take responsibility for the final manuscript.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. URL https://arxiv.org/abs/2404.14219.

Md Tanvirul Alam. Beyond perception errors: Semantic fixation in large vision-language models. *arXiv preprint arXiv:2604.12119*, 2026. URL https://arxiv.org/abs/2604.12119.

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. URL https://arxiv.org/abs/2203.17274.

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

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. ViP-LLaVA: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12914–12923, 2024. URL https://arxiv.org/abs/2312.00784.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. URL https://arxiv.org/abs/2306.15195.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 91–104, 2025. URL https://arxiv.org/abs/2409.17146.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *arXiv preprint arXiv:2106.02997*, 2021. URL https://arxiv.org/abs/2106.02997.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.14610.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. URL https://arxiv.org/abs/2311.16922.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. URL https://aclanthology.org/2023.emnlp-main.20/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. URL https://arxiv.org/abs/1405.0312.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.08485.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a. URL https://arxiv.org/abs/2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, and Bo Zhao. Unveiling the ignorance of MLLMs: Seeing clearly, answering incorrectly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9087–9097, 2025. URL https://arxiv.org/abs/2406.10638.

Jiri Martinek, Josef Baloun, Martin Prantl, Ladislav Lenc, and Pavel Kral. COMICORDA: Dialogue act recognition in comic books. In *Proceedings of LREC-COLING*, pp. 3566–3578, 2024. URL https://aclanthology.org/2024.lrec-main.316/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2202.05262.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. URL https://arxiv.org/abs/2306.14824.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? visual prompt engineering for VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11987–11997, 2023. URL https://arxiv.org/abs/2304.06712.

Xiaoxiao Sun, Mingyang Li, Kun Yuan, Min Woo Sun, Mark Endo, Shengguang Wu, Changlin Li, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. Do VLMs perceive or recall? probing visual perception vs. memory with classic visual illusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25861–25870, 2026. URL https://arxiv.org/abs/2601.22150.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022. URL https://arxiv.org/abs/2204.03162.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024. URL https://arxiv.org/abs/2401.06209.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020. URL https://arxiv.org/abs/2004.12265.

Emanuele Vivoli, Marco Bertini, and Dimosthenis Karatzas. CoMix: A comprehensive benchmark for multi-task comic understanding. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2407.03550.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*, pp. 198–215, 2024. URL https://arxiv.org/abs/2403.02325.

Kevin Ro 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*, 2023a. URL https://arxiv.org/abs/2211.00593.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11998–12008, 2023b. URL https://arxiv.org/abs/2303.14465.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Farooq Ahmad Wani, Alessandro Suglia, Rohit Saxena, Aryo Pradipta Gema, Wai-Chung Kwan, Fazl Barez, Maria Sofia Bucarelli, Fabrizio Silvestri, and Pasquale Minervini. Same answer, different representations: Hidden instability in VLMs. *arXiv preprint arXiv:2602.06652*, 2026. URL https://arxiv.org/abs/2602.06652.

Mingjie Xu, Jinpeng Chen, Yuzhi Zhao, Jason Chun Lok Li, Yue Qiu, Zekang Du, Mengyang Wu, Pingping Zhang, Kun Li, Hongzheng Yang, Wenao Ma, Jiaheng Wei, Qinbin Li, Kangcheng Liu, and Wenqiang Lei. VP-Bench: A comprehensive benchmark for visual prompting in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11332–11341, 2026. URL https://arxiv.org/abs/2511.11438.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023. URL https://arxiv.org/abs/2310.11441.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.07704.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16042.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Haiyang Wu, De Wen Soh, and Jun Liu. CPCF: A cross-prompt contrastive framework for referring multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79836–79849, 2025. URL https://proceedings.mlr.press/v267/zhu25h.html.

APPENDIX

A COMPLETE CONTROLLED RESULTS

A.1 WEAK-REPRESENTATION BOUNDARY

Table A1 gives the small-model result omitted from the primary table. The 1.3-point orbit–sham gain is statistically indistinguishable from zero, and orbit accuracy remains close to chance. This result does not support reliable recovery of a relational direction in SmolVLM2-2.2B.

Table A1: Frozen Synthetic-51 weak-model boundary (153 directions, %).

Model	Raw	Topology	SHAM	ORBIT	Gain
SmolVLM2-2.2B	34.0	35.3	47.7	49.0	+1.3

A.2 RENDERING NUISANCE ROBUSTNESS

Six nuisance renderings vary color, width, text, JPEG quality, endpoint offset, and shaft length while holding the scene, participants, and relational target fixed. The main paper reports the aggregate robustness ranges; this control serves to separate appearance invariance from relation-sensitive decoding.

A.3 A SECOND COMPLETE OUTPUT-INVARIANT ORBIT

Figure A1 shows the complete orbit for a frozen football scene from the controlled benchmark. The source, people, labels, text, and answer candidates are identical across its three views; only the tail endpoint moves. LLaVA predicts D in each view, while the orbit recovers A, B, and C.

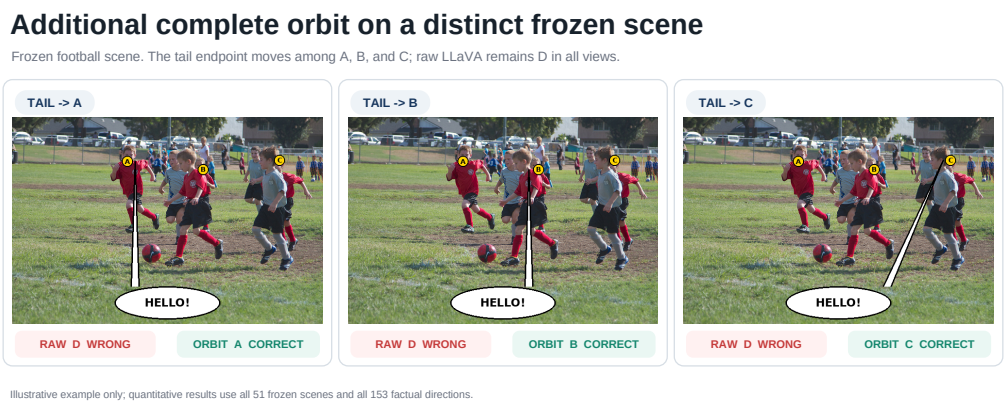

Figure A1: Additional complete orbit. A second scene exhibits the same all-view categorical failure and correctly oriented response geometry.

B NATURAL-DOMAIN RESULTS AND TRANSFER BOUNDARY

The fixed-carrier COMICORDA replication uses one red rectangular query outline for all 48 panels and all five models. No model-specific carrier was selected after scoring. Main-text Table 3 reports the per-model results.

Figure A2 illustrates the natural-domain task. Given a dialogue-text region, the model must identify its officially linked speaker among the yellow-labeled people. The displays preserve the original speech tails and comic artwork. For legibility, panels (a–c) outline the complete official text boxes; the quantitative inputs use predeclared equal-size carriers centered on the same official query endpoints.

The quantitative sham control places the same carrier away from all text, balloons, and annotated people. The displays are illustrative.

We separately evaluate a frozen 80-scene CoMix/DCM domain-shift stress test. All estimates use the complete population of 80 scenes, 240 factual queries, and 160 matched shams frozen before model scoring. Table A2 reports the full per-model sham and orbit accuracies together with the updated scene-bootstrap intervals; the pooled scene-macro gain is +7.1 points with a 95% CI of [4.4,9.8].

Table A2: CoMix/DCM domain-shift stress test frozen before model scoring.

Model	SHAM	ORBIT	Gain [95% CI]
Qwen3-VL-4B	60.0	63.3	+3.3 [-0.8,7.5]
Qwen3-VL-8B	58.3	70.0	+11.7 [7.5,15.8]
InternVL3.5-4B	48.3	65.0	+16.7 [9.6,23.8]
LLaVA-1.6-7B	35.0	31.7	-3.3 [-9.6,2.9]
Pooled macro	–	–	+7.1 [4.4,9.8]

C EXACT RESPONSE DECOMPOSITION

For candidate y and operator value r_i, let $\bar{z}_y = K^{-1}\sum_j z_y(r_j)$ and $\epsilon_i(y) = z_y(r_i) - \bar{z}_y$. The identity

$$z_y(r_i) - \frac{1}{K-1}\sum_{j\neq i} z_y(r_j) = \frac{K}{K-1}\epsilon_i(y) \tag{4}$$

holds for every response matrix. We verify the identity numerically to tolerance 10^{-10} for all 15 frozen model–domain matrices. This decomposition describes the response matrix; matched sham remains the causal control.

Table A3: LLaVA decomposition over Synthetic-51 plus fixed-red COMICORDA.

Quantity	Count / 297
Raw prediction D	297
Interaction correct over A/B/C/D	165
Raw topology correct over A/B/C	105
Interaction topology correct over A/B/C	197
Matched-sham topology correct	146
Interaction vs. sham: improves / harms	76 / 25

The targets are exactly balanced, with 99 cases each. The interaction-topology decoder is correct on 65 A targets, 59 B targets, and 73 C targets. Relative to raw topology, it recovers 126 failures and changes 34 successes to errors. Neither removing D nor generic candidate centering therefore explains the reported result.

C.1 CLOSED-BATCH ASSIGNMENT

PED uses the images before and after the user adds the cue and improves Q2 for all five models in main-text Table 1. Complete operator orbits test additional valid counterfactuals; the independent-query decoder produces the gains in main-text Table 2 without training. When all K relations form a closed batch with a known one-to-one mapping, maximizing the total score over assignments raises LLaVA-1.6-7B from 82.4% to 90.8% on Synthetic-51 (Holm $p = .0333$). For any one-to-one



(a) Queried dialogue text; answer A

(b) Queried dialogue text; answer B

(c) Queried dialogue text; answer C

Figure A2: Natural-comic speaker task. The yellow tags mark candidate people; the red outline identifies the queried dialogue text. Panels (a–c) cycle through three native utterances whose official CoMix links identify different speakers. Source: CoMix validation / public-domain DCM.

assignment, subtracting a candidate-specific column baseline changes every assignment by the same constant. Raw, sham, and orbit matrices thus have the same optimum under bijective matching. The batch gain comes from the topology constraint; the independent-query recovery is already present without it.

D CAUSAL INTERVENTION DETAILS

The causal intervention uses the separate 20-scene mechanism set described in the main paper. For each factual/counterfactual pair, causal patching measures the patched change in the counterfactual-minus-factual answer margin relative to the unpatched margin. The source scene is the independent unit, and the two directions are averaged within each scene. Early-layer p values are corrected with Holm's procedure within each model over layers 2, 4, 8, and 12.

The complete all-vision causal-transfer results are reported in main-text Table 5; here we give the intervention and localization details.

For Qwen, the exact pixel-difference mask intersects 11–41 of approximately 192–380 vision tokens. Patching only these tokens reproduces 91.4%, 88.1%, 88.5%, and 90.4% of the all-vision effect at layers 2, 4, 8, and 12. The local effect falls to 60% by layer 20 and becomes negligible in absolute size by layer 23. For LLaVA, packed multiresolution image tokens support the all-vision causal-transfer analysis but lack the exact flat pixel–token map required for edit-token localization.

Table A4: Qwen causal specificity and semantic reversal. Patch shifts use source-scene CIs; codebook rows report physical versus instructed-semantic orientation accuracy on the same 20-scene mechanism set.

Type	Condition	Primary metric	Secondary metric	p
Patch	random visual @ 4	0.099 shift	CI $[-0.029, 0.222]$	0.1405
Patch	answer pos. @ 4	-0.025 shift	CI $[-0.150, 0.094]$	0.7326
Patch	answer pos. @ 24	3.522 shift	CI $[2.144, 5.160]$	$< 10^{-4}$
Patch	answer pos. @ 36	5.075 shift	CI $[3.200, 7.300]$	$< 10^{-4}$
Codebook	normal	65.0% physical	65.0% semantic	–
Codebook	reversed	67.5% physical	32.5% semantic	–

E CROSS-SYMBOL REPAIR AND NEAREST-PRIOR PRESSURE TEST

The target-blind arrow decoder receives only paired model-score rows and condition names. Before predictions are frozen, we remove targets, category strings, boxes, endpoints, prompts, and correctness; ground truth is joined only for evaluation. This procedure reproduces both reported gains exactly. A condition-only rule scores 50.9%, so neither row order nor condition names explain the result.

Table A5: Separate 500-scene COCO cross-symbol sample (1,000 targets).

Model	Box	Raw arrow	Reversed contrast	Paired p
Qwen3-VL-4B	92.2	60.0	77.5	1.33×10^{-31}
LLaVA-1.6-7B	62.5	36.9	66.0	1.14×10^{-53}

The same-order ViP-style reconstruction contains 1,760 COCO-val images and 3,520 factual directions covering all 80 categories, 424 unordered category pairs, and 579 left-to-right arrangements. It is separate from the unreleased ViP-LLaVA split. An exhaustive audit covers all 5,280 rendered views. Every pixel hash is unique, each factual-minus-line change is confined to the intended arrowhead, and each line-minus-source change is confined to the shared shaft. Arrowhead changes contain 40–227 pixels (median 91), whereas the shared shaft contains 352–5,724 changed pixels (median 2,050). The headless-line control is therefore stricter than an absent-overlay baseline.

F SHORTCUT AND IDENTIFIABILITY BOUNDARIES

F.1 FROZEN LABEL PERMUTATION

Before scoring, every A/B/C label is moved from its original left/middle/right position by a derangement. Qwen retains a significant gain. LLaVA's gain is positive but statistically inconclusive after Holm correction across the two frozen model comparisons. Qwen's significant gain rules out the fixed spatial-label shortcut for that model; LLaVA remains unresolved.

The complete label-permutation results are reported in Table A6.

Table A6: Frozen label-permutation control. A/B/C labels are deranged before scoring; gains are orbit minus matched sham.

Model	Raw	SHAM	ORBIT	Gain [95% CI]	Holm p
Qwen3-VL-4B	73.2	90.2	96.1	+5.9 [2.6,9.8]	.00734
LLaVA-1.6-7B	0.0	63.4	71.2	+7.8 [0.0,15.7]	.08871

F.2 WHY TARGET-FREE SEPARABILITY CANNOT CERTIFY MEANING

Consider an unlabeled orbit response matrix Z whose row-wise maxima realize the intended mapping f. Replace the semantic mapping with $\pi \circ f$ for a derangement π of the candidates. The numerical matrix and every target-free statistic remain unchanged, while accuracy changes from one to zero. Hence no unlabeled separability statistic can certify semantic orientation in general.

Empirically, OASI correlates with absolute orbit accuracy over 15 frozen model–domain pairs ($\rho = .761$, permutation $p = .00147$), but not with orbit–sham gain ($\rho = .050$, $p = .864$); the correlation after model/domain demeaning is negative. We therefore use OASI as an exploratory descriptor of response sensitivity.