DIFFERENT SEMANTIC EDITS CONVERGE TO GROUNDED TRUTH IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) must determine how a semantic change alters the truth of a proposition given visual evidence, rather than respond to surface cues such as negation or object identity. Yet how this grounded truth computation emerges inside a VLM remains unclear. We study this question using answer-balanced natural-image quartets in which negation and object substitution each reverse the correct answer in both directions. Across three VLMs and two natural-image benchmarks, we identify a common *edit-to-truth transition*: early representations preserve edit-specific structure, whereas intermediate layers organize distinct edits according to the truth change required by the image. This shared representation is both causally answer-relevant and visually grounded. At the transition, a one-dimensional truth-axis projection retains over 80% of the full attention-swap effect for both edits, while the orthogonal residual has little effect; the same geometric progression also appears when only the visual evidence changes. Localized removal of visual evidence suppresses the truth representation and its causal effect but largely preserves linguistic polarity. The resulting representation follows proposition truth rather than the literal output symbol under response-code reversal, and frozen RePOPE→AMBER probes transfer without target fitting, reaching 91–93% accuracy over all target images. Finally, source-derived truth directions selectively repair naturally occurring errors: on LLaVA, they correct 103 AMBER errors without introducing new ones, raising accuracy from 79.0% to 87.6%. Together, these results reveal a shared grounded-truth stage through which distinct semantic edits are transformed into visually grounded decisions in VLMs.

1 INTRODUCTION

A semantic edit does not determine an answer by itself; its effect depends on the image. If a dog is present, adding *no* to "Is there a dog?" changes Yes to No. If a cat is absent, the same edit changes No to Yes. Replacing *dog* with *cat* can produce the same answer reversals through a different operation. A vision-language model (VLM) must therefore combine the edit with visual evidence rather than map a word such as *no* directly to an answer. We ask how distinct edits become visually grounded truth judgments inside a VLM.

VLMs remain brittle under negation and compositional reasoning (Dobreva and Keller, 2021; Alhamoud et al., 2025; Zhang et al., 2025). Mechanistic studies localize object information and trace its influence on outputs (Li et al., 2026; Schaumlöffel et al., 2026; Kim et al., 2025), while language-model analyses identify truth-related linear structure across datasets and logical transformations (Marks and Tegmark, 2024; Bao et al., 2025). These results leave two possibilities unresolved. Semantic operations may remain separate until the final answer, or they may acquire a shared representation after being evaluated against the image. Decodability alone cannot distinguish grounded proposition truth from linguistic polarity or a pending output symbol.

We isolate these possibilities with answer-balanced 2×2 natural-image quartets (Fig. 1). Each image supplies a present object X and an absent object Y, queried in positive and negative-existential forms. Negation and object substitution each require both Yes-to-No and No-to-Yes changes while the image remains fixed. The construction separates edit identity from answer direction; response-code reversal further separates proposition truth from the literal symbol used to report it.

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.

Across three VLMs and two benchmarks, we identify an *edit-to-truth transition*. Early change vectors retain edit-specific structure. At model-specific intermediate layers, answer-direction classifiers transfer across edits while preserving decodable edit identity. Component swaps at these layers move native margins toward the paired source for both edits and strongly outperform norm-matched random replacements. On held-out Qwen3-VL-2B images, projecting the source–target attention difference onto a training-derived truth axis retains 88.6% and 82.3% of the full mean effect for negation and object substitution, while the orthogonal residual is weak. Localized removal of the queried object then weakens the truth projection and the attention-swap effect while largely preserving early polarity. Under a reversed Yes/No codebook, the representation follows proposition truth rather than the response symbol. Frozen RePOPE→AMBER truth probes also reach 91–93% accuracy over all target images across the three checkpoints.

The same geometry supports a functional test. A truth direction and router calibrated on RePOPE repair naturally occurring AMBER errors without target-label selection. On LLaVA, internal displacement fixes 103 errors without introducing new ones, raising accuracy from 79.0% to 87.6%. Our contributions are:

- We identify a cross-model *edit-to-truth transition*: initially edit-specific representations become organized by the proposition-truth change required by visual evidence; the same geometric progression appears for a visual-only edit that holds the question fixed.
- We establish causal specificity through both-edit component interventions and a truth-axis decomposition in which a one-dimensional truth-aligned component retains most of the full intervention effect, together with localized evidence removal across three checkpoints.
- We separate truth from output-symbol choice, demonstrate frozen cross-dataset transfer, and use a source-derived truth direction to raise LLaVA accuracy from 79.0% to 87.6% with 103 fixes and no new errors.

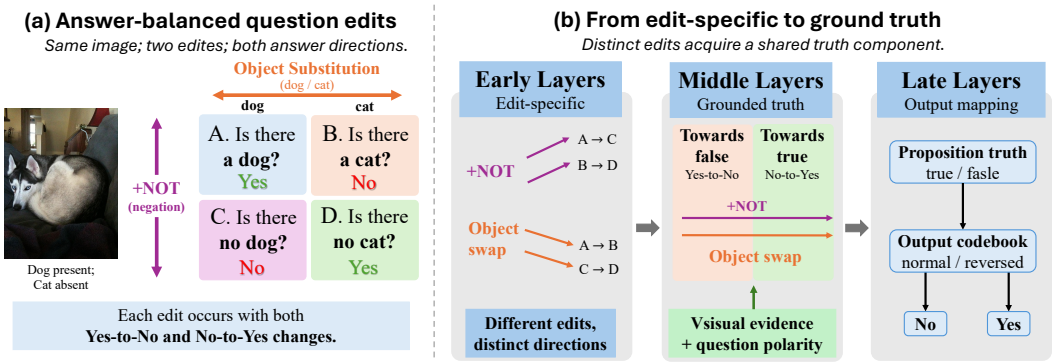


Figure 1: **From distinct semantic edits to a shared grounded-truth stage.** (a) A fixed image contains a dog but no cat. Negation (A→C, B→D) and object substitution (A→B, C→D) each produce both Yes-to-No and No-to-Yes changes, separating edit identity from answer direction. (b) The resulting mechanistic picture: early representations retain edit-specific structure; intermediate layers acquire a shared, visually grounded representation of proposition truth; later states map that truth to the required output code. The transition occurs at model-specific depths.

2 Related Work

Negation and compositional evaluation. Negation failures occur in multimodal encoders and generative VLMs (Dobreva and Keller, 2021; Alhamoud et al., 2025; Zhang et al., 2025); NAST uses causal layer signals for selective training (Abbasi et al., 2026). Winoground and ARO test broader compositional distinctions (Thrush et al., 2022; Yuksekgonul et al., 2023), while SugarCrepe addresses linguistic artifacts that permit image-blind solutions (Hsieh et al., 2023). Our answer-balanced quartets separate edit identity from the required Yes/No movement.

Truth representations. Language-model states can encode truth information accessible to contrast-consistent or supervised probes (Burns et al., 2023; Azaria and Mitchell, 2023). Marks and Tegmark (2024) study linear truth structure, cross-dataset transfer, and causal interventions; Bao et al. (2025) examine generalization across logical transformations, including negation, and question-answering tasks. Generalization tests further characterize the reliability of truth probes (Levinstein and Herrmann, 2025). We trace when negation and object substitution acquire a shared truth component, using cross-edit transfer, evidence removal, and codebook reversal to distinguish grounded truth from edit identity and output symbols.

Mechanistic analysis of VLMs. Mechanistic VLM work traces object representations (Li et al., 2026), localizes classification and localization heads (Schaumlöffel et al., 2026), and studies responses to absent concepts (Kim et al., 2025). Other work identifies spatial indexing for feature binding (Assouel et al., 2026) and framing-dependent use of visual input (Fan et al., 2026). Causal abstraction motivates testing intermediate variables through interventions (Geiger et al., 2021). We compare how negation and object substitution affect answer-position attention and MLP outputs.

Activation steering and hallucination mitigation. Inference-time intervention targets truthfulness-related directions (Li et al., 2023a); activation addition and contrastive activation addition derive steering vectors from contrasting prompts or examples (Turner et al., 2023; Rimsky et al., 2024). Representation engineering develops population-level monitoring and control (Zou et al., 2023). Object-hallucination evaluation dates back to image captioning (Rohrbach et al., 2018). Mitigation includes visual contrastive decoding (Leng et al., 2024), attention-based penalties and retrospection (Huang et al., 2024), visual/textual latent steering in VTI (Liu et al., 2025), and sparse interventions restoring visual information in REVIS (Wu et al., 2026). We use source-derived truth geometry to repair routed polarity–evidence composition errors. External complementation tests repair selectivity, and norm-matched random directions test truth-axis specificity.

3 Experimental Design

3.1 Answer-balanced natural-image quartets

Each image contributes one object verified as present (X) and one verified as absent (Y). We ask A: "Is there an X in the image?", B: "Is there a Y in the image?", C: "Is there no X in the image?", and D: "Is there no Y in the image?" The respective truth values are Yes, No, No, and Yes. Polarity and visual match are therefore balanced within truth, and each semantic edit occurs in both answer directions. The same image is used in all four conditions. We use $+\mathrm{NOT}$ for the negative-existential edit implemented by adding *no*. This construction separates the semantic operation applied to the question from the truth change that the operation induces.

The images come from RePOPE, a corrected version of the MSCOCO-based POPE object-existence benchmark (Lin et al., 2014; Li et al., 2023b; Neuhaus and Hein, 2025), and the existence split of AMBER (Wang et al., 2023). We evaluate Qwen3-VL-2B-Instruct, Qwen3-VL-4B-Instruct (Bai et al., 2025), and LLaVA-1.5-7B (Liu et al., 2024a). For each prompt, we record

$$s = \log P(\text{Yes}) - \log P(\text{No}) \tag{1}$$

at the final prompt position, whose state predicts the first response token, and cache the residual stream after every decoder block. Layer 0 precedes the first block; layer $l > 0$ follows block l. Behavioral and population-level analyses include all images. Representation analyses center on all-correct quartets to characterize the internal geometry of successful computation, with all-image replications reported alongside them in Section 4.2.

3.2 Representation tests

Let $\mathbf{h}_l^A, \ldots, \mathbf{h}_l^D$ denote the answer-position states at layer l. For each edit, we form two vectors whose required answer movements oppose each other:

$$\Delta_l^{\mathrm{NOT},\downarrow} = \mathbf{h}_l^C - \mathbf{h}_l^A, \qquad \Delta_l^{\mathrm{NOT},\uparrow} = \mathbf{h}_l^D - \mathbf{h}_l^B, \tag{2}$$

$$\Delta_l^{\mathrm{OBJ},\downarrow} = \mathbf{h}_l^B - \mathbf{h}_l^A, \qquad \Delta_l^{\mathrm{OBJ},\uparrow} = \mathbf{h}_l^D - \mathbf{h}_l^C. \tag{3}$$

Table 1: **Behavior over all images.** A–D are percentages; "All" is the number of images on which all four conditions are correct. The lowest-accuracy quadrant varies across models and datasets.

Dataset	Model	A	B	C	D	All
RePOPE ($n = 491$)	Qwen3-VL-2B	98.2	95.9	87.4	93.7	387
RePOPE ($n = 300$)	Qwen3-VL-4B	96.3	98.3	98.7	94.7	274
RePOPE ($n = 100$)	LLaVA-1.5-7B	99.0	95.0	60.0	90.0	50
AMBER ($n = 300$)	Qwen3-VL-2B	99.3	76.0	96.7	85.3	215
AMBER ($n = 300$)	Qwen3-VL-4B	98.0	95.0	99.3	66.7	194
AMBER ($n = 300$)	LLaVA-1.5-7B	99.0	53.7	64.3	99.0	100

Their cosine measures whether the same edit retains a common direction across opposite required answers. Positive cosine indicates an edit-related component shared across contexts, whereas increasing influence of the opposing answer movements can drive the two vectors apart. To test whether distinct edits acquire a common answer-relevant organization, we train a linear classifier to distinguish ↑ from ↓ on one edit and evaluate it on the other. Near-chance transfer indicates predominantly edit-specific geometry; high transfer indicates that answer-direction information generalizes across edits. The layer-wise change between these regimes provides our primary measure of the edit-to-truth transition. Section 5 then tests whether the shared representation depends on visual evidence and follows proposition truth rather than the output code. Every probe uses five-fold image-held-out cross-validation, preventing any condition from a test image from entering training. Feature standardization is fit within each training fold and applied unchanged to its test fold.

The quartet defines a common truth-interaction contrast for both edits:

$$\mathbf{r}_l = \Delta_l^{\text{NOT},\uparrow} - \Delta_l^{\text{NOT},\downarrow} = \Delta_l^{\text{OBJ},\uparrow} - \Delta_l^{\text{OBJ},\downarrow} = \mathbf{h}_l^A + \mathbf{h}_l^D - \mathbf{h}_l^B - \mathbf{h}_l^C. \tag{4}$$

This identity follows directly from the balanced design. Cross-edit decoding measures how consistently this shared contrast is expressed across images relative to edit-specific offsets and variation within each answer-direction class. For each image, an exact factorial decomposition writes the states as $\boldsymbol{\mu}_l + p\mathbf{a}_l + v\mathbf{b}_l + pv\mathbf{c}_l$, with polarity p, visual match v, and $\mathbf{r}_l = 4\mathbf{c}_l$. The same-edit dot products are $4(\|\mathbf{a}_l\|^2 - \|\mathbf{c}_l\|^2)$ for negation and $4(\|\mathbf{b}_l\|^2 - \|\mathbf{c}_l\|^2)$ for object substitution. Thus, the cosine tracks the balance between edit-specific and truth-interaction components, while held-out cross-edit decoding measures how consistently the shared component generalizes across images (Appendix B).

We also fit balanced probes for polarity, visual match, and truth. In cross-dataset tests, both preprocessing and probe weights are fit on RePOPE and frozen before evaluation on AMBER. Probes measure decodable representation content (Hewitt and Liang, 2019; Belinkov, 2022), while component interventions test whether the corresponding internal states causally influence the answer. Appendix A gives the templates, layer numbering, and fold protocol; Table 5 summarizes experimental coverage across checkpoints.

3.3 Component interventions

To test whether the representational transition is causally involved in answer formation, we replace the answer-position output of self-attention or the MLP in a target condition with the corresponding output from its paired source, following causal mediation and tracing (Vig et al., 2020; Meng et al., 2022). Other positions and modules remain native. We patch both directions in both visual contexts, giving four directed relations per edit. For each relation, the oriented margin effect is

$$e = (s_{\text{patch}} - s_{\text{target}})\,\text{sign}(s_{\text{source}} - s_{\text{target}}). \tag{5}$$

A relation is directional when $e > 0$. This statistic measures whether the intervention moves the answer margin toward the paired source; signed effect size measures the magnitude of that movement, and an answer flip records a decision-boundary crossing. We report the directional fraction per image and the stricter indicator that all four relations are directional. Random controls replace the same component output with a centered Gaussian vector rescaled to the source-component norm, matching the replacement magnitude while removing source-specific information. Exact paired McNemar tests and bootstrap confidence intervals use images as the sampling unit, preserving dependence among the four relations. Following activation-patching guidance (Zhang and Nanda, 2024; Heimersheim and Nanda, 2024), we report directional margin movement and literal answer flips separately.

The intervention suite spans all three models at their model-specific transition intervals. The edit-by-component protocol uses 100, 95, and 50 held-out RePOPE images for Qwen3-VL-2B, Qwen3-VL-4B, and LLaVA, respectively, and localized visual-evidence removal uses model-specific A/C-correct subsets of 100, 100, and 60 images. The two Qwen checkpoints provide replicated response-code reversal tests; frozen cross-dataset transfer and source-calibrated repair cover all three models. Appendix A gives the complete coverage matrix.

4 From Edit-Specific Geometry to a Shared Causal Stage

4.1 Layer-wise cross-edit convergence

All three models exhibit an edit-to-answer-direction crossover (Fig. 2). In Qwen3-VL-2B, cross-edit answer-direction accuracy remains near chance at layers 1 and 10 (49.7% and 54.3%), then rises sharply to 100% at layer 16. The two +NOT vectors show the complementary change: they are strongly aligned at layer 10 (cosine 0.841), nearly orthogonal at layer 16 (-0.022), and anti-aligned by layer 24 (-0.432). Object substitution follows the same pattern, although its cosine remains more positive at the boundary (0.870, 0.245, and -0.376 at layers 10, 16, and 24). The change vectors thus shift from a shared edit component toward the answer movement required in each context.

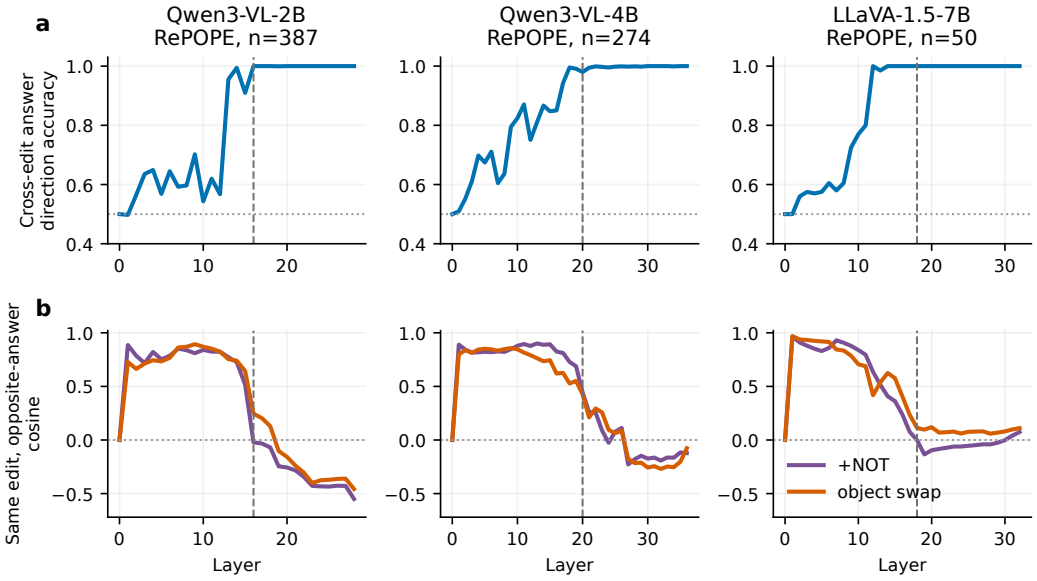


Figure 2: **Edit-specific trajectories become answer-specific.** Top: an answer-direction probe trained on +NOT and tested on object substitution, or vice versa, is averaged across both train/test directions and image-held-out folds. Bottom: same-edit cosine across opposite answer directions. Dashed lines mark the model-specific transition interval for visual reference.

The measurements resolve different aspects of the transition. Qwen3-VL-4B keeps its +NOT vectors aligned through layer 15 (0.893), followed by a separation to 0.445 at layer 20 and -0.026 at layer 24; cross-edit transfer reaches 99.5% at layer 18. In LLaVA, mean cross-edit accuracy is 100% at layer 12, 98.5% at layer 13, and 100% from layer 14 onward. Its +NOT cosine changes more gradually, from 0.839 at layer 10 to 0.239 at layer 16 and approximately zero at layer 18. Decoding and cosine therefore resolve complementary changes in cross-image predictability and within-image geometry. Section 4.2 tests both across benchmarks and sample populations.

Accurate cross-edit transfer coexists with preserved operation identity. On LLaVA, cross-answer-direction operation accuracy at layer 32 is 90% in one direction and 100% in the other. The representation thus carries both a shared answer-direction component and edit-specific structure. We next test its causal influence and use visual-evidence and response-code controls to characterize its grounding and relation to output symbols.

Visual-only edit. As a complementary test outside the same-image quartet algebra, we hold the question fixed and pair images with and without the same queried object, with object categories held out across folds. Its opposite-answer cosine follows the same progression as the linguistic edits, falling from 0.936 at L10 to 0.325 at L16 and -0.251 at L24 (Appendix I).

4.2 Cross-benchmark and population-level replication

We repeat the cross-edit analysis on AMBER and evaluate complete image populations alongside all-correct quartets (Table 2). On RePOPE, all-image cross-edit accuracy at the selected transition layers is 90.59% for Qwen3-VL-2B, 83.25% for Qwen3-VL-4B, and 97.50% for LLaVA. The corresponding all-correct accuracies are 100%, 98%, and 100%. Shared answer-direction information is therefore measurable in the complete populations as well as in successful computations.

Table 2: **Cross-edit generalization across benchmarks and sample populations.** Each probe is fit and evaluated within the named benchmark using image-held-out folds. Accuracy averages both edit-transfer directions and five folds. "Images" gives all/all-correct counts.

Dataset	Model	Layer	Images	All (%)	Correct (%)
RePOPE	Qwen3-VL-2B	16	491/387	90.59	100.00
RePOPE	Qwen3-VL-4B	20	300/274	83.25	98.00
RePOPE	LLaVA-1.5-7B	18	100/50	97.50	100.00
AMBER	Qwen3-VL-2B	16	300/215	95.67	100.00
AMBER	Qwen3-VL-4B	20	300/194	96.67	99.74

On AMBER, Qwen3-VL-2B again reaches perfect all-correct cross-edit transfer at layer 16, while Qwen3-VL-4B approaches it at layer 20. Both also exceed 95% on all target images. Same-edit alignment weakens over this interval: the all-correct +NOT cosine changes from .851 at layer 10 to $-.364$ at layer 20 in 2B, and from .861 to .422 in 4B, reaching $-.075$ at layer 24 (Appendix D). These within-benchmark replications test the recurring layer-wise pattern; Section 5 separately tests a frozen source-to-target representation.

4.3 Component swaps affect both edits

We test whether the transition is causally involved by patching component outputs on the reserved Qwen3-VL-2B images. At layer-16 self-attention, both edits move toward their paired source on all four relations for every image (Table 3). Norm-matched random vectors satisfy this strict all-four criterion on 5–6% of images. The mean source-minus-random gain in per-image directional fraction is 0.523 for +NOT and 0.513 for object substitution. The image-bootstrap intervals exclude zero, with image-level exact paired $p < 1.1 \times 10^{-28}$. Literal answer flips occur in 18.5% of negation relations and 17.0% of object-substitution relations.

Table 3: **Held-out component patching at Qwen3-VL-2B layer-16 self-attention.** "All-four" requires every directed relation for an image to move toward its source. Gain is the source-minus-random per-image directional fraction; confidence intervals bootstrap images.

Edit	Source all-four	Random all-four	Fraction gain (95% CI)	Image p
+NOT	100.0	6.0	.523 [.473, .573]	1.0×10^{-28}
Object substitution	100.0	5.0	.513 [.468, .558]	5.0×10^{-29}

Truth-axis decomposition. We further decompose each held-out source–target L16 attention difference into its projection onto a truth axis estimated from a disjoint training set and the orthogonal residual. The one-dimensional truth projection is directional on all four relations for all 50 test images for both edits and retains 88.6% (+NOT) and 82.3% (object substitution) of the full mean effect. The orthogonal residual satisfies the all-four criterion on only 2/50 images for each edit, with mean effects 0.158 and 0.284; paired-bootstrap truth-minus-orthogonal 95% CIs are [3.057, 3.334] and [2.849, 3.133] (Appendix J).

The layer profile further separates the roles of attention and the MLP (Fig. 3a–b). Attention source swaps produce large directional effects around layers 15–16. By layer 18, source swaps in both self-attention and the MLP are nearly or fully directional, with MLP signed-margin gains of 4.22 for +NOT and 4.45 for object substitution. The interventions identify a shared component-level causal locus for the two edits.

The same causal ordering replicates at model-specific depths (Fig. 3b). In Qwen3-VL-4B, layer-23 attention source swaps satisfy the strict all-four criterion on 92.6% of images for +NOT and 93.7% for object substitution, compared with 5.3% and 4.2% under norm-matched random replacement. A later layer-24 MLP effect remains source-specific (26.3% and 30.5% all-four, versus 2.1% random). LLaVA reaches 100% strict all-four attention movement for both edits at layer 17, compared with 2% and 8% random; at layer 18 the MLP reaches 64% for +NOT and 84% for object substitution, versus 2% and 0% random. Together with the Qwen3-VL-2B layer-16 attention and later layer-18 MLP profile, the transition depth varies by architecture while the attention-to-MLP ordering recurs across checkpoints.

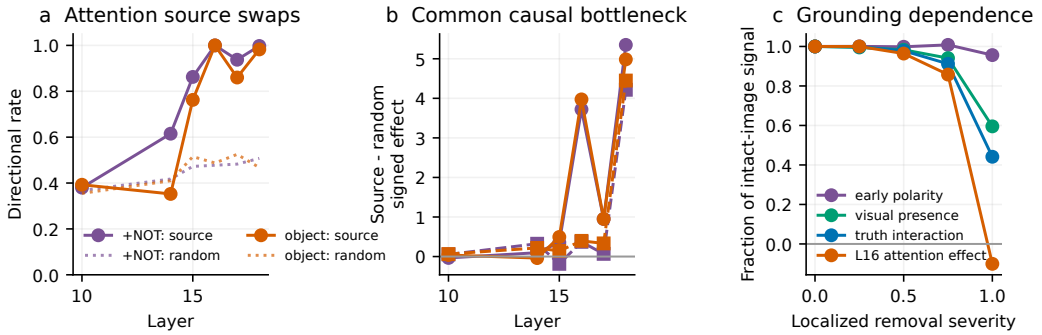


Figure 3: **The causal transition and its visual dependence replicate across models.** (a) Qwen3-VL-2B attention source swaps versus norm-matched random controls across layers. (b) Strict image-level all-four movement for both edits at each model's strongest transition-stage attention layer. (c) Under full localized visual removal, early polarity remains near its intact value across all three models, while visual presence and truth fall and the transition-stage attention effect collapses.

5 Testing the Grounded-Truth Interpretation

The preceding tests identify shared geometry and a causal intervention locus. We now ask whether the inferred component depends on visual evidence, remains stable when output symbols change, and transfers beyond the source benchmark.

5.1 Removing visual evidence weakens truth but preserves polarity

We test visual grounding by weakening detector-localized evidence for the queried present object while holding the question fixed. Grounding-DINO (Liu et al., 2024b) defines the region; intermediate severities blur that region and full removal replaces it with the surrounding local mean color. Polarity, visual-presence, and truth directions are estimated from intact images and then frozen. We apply the same protocol to Qwen3-VL-2B, Qwen3-VL-4B, and LLaVA.

The cross-model dissociation is strikingly consistent (Fig. 3c). At full removal, the early polarity projection retains 95.6%, 95.0%, and 95.9% of its intact value in 2B, 4B, and LLaVA. The visual-presence projection retains 59.5%, 59.9%, and 61.3%, while the transition-stage truth projection retains 44.1%, 52.0%, and 45.9%. The corresponding attention-patch effects change from 5.08 to -0.51, 4.57 to -0.38, and 1.71 to 0.06. Thus the fixed linguistic polarity remains represented while the visually grounded truth signal and its causal attention effect weaken together.

5.2 Reversing output symbols preserves truth

A decodable Yes/No state may reflect either proposition truth or the symbol used to report it. We separate the two with normal and reversed response codebooks, under which the same proposition requires the opposite output symbol. In Qwen3-VL-4B, among 191 images answered correctly in all four conditions under both codes, truth transfer rises from 93.8% at layer 15 to 100% at layer 24, while literal-symbol transfer falls from 6.2% to 0% (Fig. 4a). Qwen3-VL-2B independently reproduces the separation on 30 jointly correct images: truth transfer is 97.5% at layer 15 and 100% at layers 16–20, whereas literal-symbol transfer is 2.5% and then 0%. The middle representation therefore follows proposition truth rather than the chosen reporting symbol.

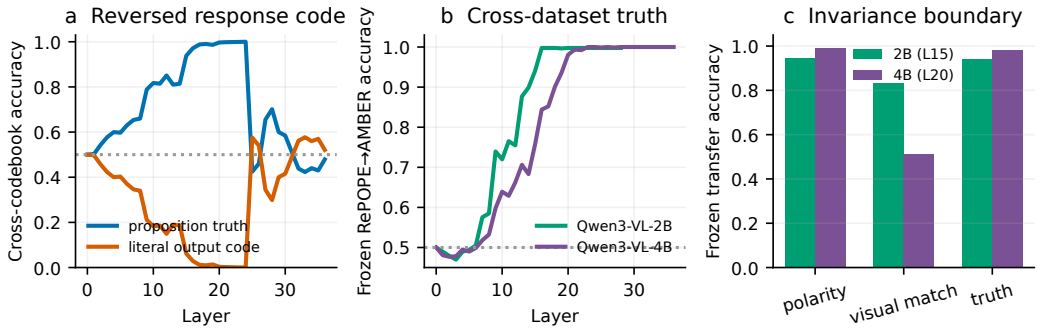


Figure 4: **Truth is stable across output codes and natural-image benchmarks.** (a) Qwen3-VL-4B cross-codebook transfer through the transition interval: proposition truth remains aligned while the literal Yes/No symbol reverses. (b) Frozen RePOPE truth probes evaluated on AMBER all-correct quartets for all three models. (c) At representative model-specific layers, frozen truth accuracy is 90.8–92.5% over all target images and 99.8–100% on all-correct target quartets.

5.3 Truth transfers across natural-image benchmarks

The truth geometry also transfers across datasets without target fitting. At representative transition layers, a probe fit only on RePOPE reaches 90.8% on all 300 AMBER images for Qwen3-VL-2B (L16), 92.5% for Qwen3-VL-4B (L24), and 91.1% for LLaVA (L18). On the corresponding all-correct target quartets, the same frozen probes reach 99.77%, 100%, and 100%, respectively (Fig. 4b–c). The shared truth geometry is therefore portable across the two natural-image benchmarks in all three checkpoints.

6 Target-Label-Free Repair with Source Truth Geometry

We use repair as a downstream test: can the inferred truth direction correct composition errors without fitting to target labels? The intervention operates on the inferred direction and retains the VLM's own readout.

6.1 Source-calibrated intervention

We connect truth-direction analysis (Marks and Tegmark, 2024) with inference-time activation steering (Li et al., 2023a; Rimsky et al., 2024). For each negated query q^-, the router also evaluates its positive counterpart q^+. Let s^- and s^+ be their native Yes-minus-No margins, and $\hat{y}^{\pm} = \mathbf{1}[s^{\pm} \geq 0]$. A query is eligible when

$$\hat{y}^+ = \hat{y}^-, \qquad |s^+| \geq \tau, \qquad |s^+| - |s^-| \geq \gamma. \tag{6}$$

The route side, thresholds, transition layer, and dose are selected on RePOPE, prioritizing source accuracy, fewer new errors, more repairs, and lower coverage. At the selected layer, the proposition-truth direction is

$$\mathbf{t}_l = \mathbb{E}[\mathbf{h}_l \mid A, D] - \mathbb{E}[\mathbf{h}_l \mid B, C] = \tfrac{1}{2}\mathbb{E}_{\text{source}}[\mathbf{r}_l]. \tag{7}$$

On target queries, native predictions and source-calibrated thresholds determine routing. C-like (same-Yes) inconsistencies move along $-\mathbf{t}_l$ and D-like (same-No) inconsistencies along $+\mathbf{t}_l$. Source selection gives layer 16/same-No/projection-q75 for Qwen3-VL-2B, layer 20/same-No/add-3 for Qwen3-VL-4B, and layer 18/both/add-1 for LLaVA. Conditions A and B remain unchanged, and the VLM produces the downstream answer. Hard complementation and sample-wise norm-matched Gaussian directions serve as controls. No AMBER label selects the route, layer, direction, threshold, or dose; Appendix M reports the search grids and additional statistics.

6.2 Target results and controls

Table 4 evaluates the frozen intervention on AMBER. Qwen3-VL-2B and Qwen3-VL-4B select same-No routing on RePOPE, so the intervention moves D-like states toward truth. LLaVA selects both sides and applies the same axis with opposite signs to C-like and D-like inconsistencies. The internal intervention improves all three models while leaving A and B unchanged. LLaVA shows the largest effect: overall accuracy increases from 79.0% to 87.58%, with C improving from 64.33% to 98.33% and D from 99.0% to 99.33%. The intervention fixes 103 errors and introduces none. With at most two editable queries per image, these repairs improve at least 52 images, giving a conservative image-level paired sign-test bound of $p \leq 4.44 \times 10^{-16}$ (Appendix M.2).

Table 4: **Target-label-free RePOPE→AMBER repair.** "External" hard-complements routed predictions; "Internal" intervenes on the VLM state. Fix/break counts compare internal repair with the original model. Random max is the largest net repair among norm-matched random directions.

Model	Source route	Original	External	Internal	Fix/break	Random max
Qwen3-VL-2B	same-No	89.33	89.75	89.75	5/0	2
Qwen3-VL-4B	same-No	89.75	94.58	92.92	39/1	22
LLaVA-1.5-7B	both	79.00	81.67	87.58	103/0	29

On LLaVA, hard complementation fixes the same 103 errors but also changes 71 previously correct routed queries; internal displacement fixes 103 and breaks none. Across 20 sampled norm-matched random axes, net repair averages 4.45 (SD 8.49) with a maximum of 29. Qwen3-VL-4B yields 38 net internal repairs against a random maximum of 22. These controls connect the source-derived truth direction to selective correction of naturally occurring target errors while retaining the model's native readout.

7 Discussion

Across architectures, the measurements reveal a common ordering: early states retain the semantic operation, intermediate states acquire answer-direction structure grounded in the image, and later states map proposition truth to the response code. Transition depth differs across Qwen3-VL-2B, Qwen3-VL-4B, and LLaVA, while edit identity remains decodable throughout. The shared grounded-truth component therefore coexists with operation-specific information.

Four convergent controls support the grounded-truth interpretation. Component swaps establish causal answer relevance. Localized evidence removal dissociates stable early polarity from a weakened truth projection and attention effect. Codebook reversal tracks proposition truth across output symbols, and frozen RePOPE→AMBER probes carry the geometry across benchmarks without target fitting.

8 Conclusion

Distinct semantic edits become organized by a shared, visually grounded truth component inside VLMs. Causal decomposition, visual removal, codebook reversal, frozen cross-dataset transfer, and source-derived repair jointly connect this representation to answer formation, while visual-only, paraphrase, and color-attribute controls extend the same geometry beyond the original quartet.

AI USE STATEMENT

Generative AI tools were used for literature search and language refinement during manuscript preparation.

REPRODUCIBILITY STATEMENT

Section 3 defines the representation and intervention measurements, and Section 6 specifies the repair rule. Appendix A gives the quartet templates, checkpoint coverage, and image-held-out protocol; Appendix B derives the factorial contrasts. Appendices E–H describe the component, grounding, codebook, and frozen-transfer controls; Appendices I–L give the focused robustness addenda; Appendix M documents repair settings. Appendix M.2 derives aggregate-count bounds for image-level repair tests, and Appendix N describes the executable demo and its included analyses.

REFERENCES

Ali Abbasi, Mehdi Taghipour, and Rahmatollah Beheshti. NAST: Improving Negation Handling in Medical Vision-Language Models through Negation-Aware Selective Training. In *Proceedings of the 11th Machine Learning for Healthcare Conference*, volume 340 of *Proceedings of Machine Learning Research*, pp. 1–29, 2026. https://proceedings.mlr.press/v340/abbasi26a.html.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-Language Models Do Not Understand Negation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29612–29622, 2025. CVF Open Access.

Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual Symbolic Mechanisms: Emergent Symbol Processing in Vision Language Models. In *International Conference on Learning Representations (ICLR)*, 2026. https://arxiv.org/abs/2506.15871.

Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 967–976, 2023. https://aclanthology.org/2023.findings-emnlp.68/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. https://arxiv.org/abs/2511.21631.

Yuntai Bao, Xuhong Zhang, Tianyu Du, Xinkui Zhao, Zhengwen Feng, Hao Peng, and Jianwei Yin. Probing the Geometry of Truth: Consistency and Generalization of Truth Directions in LLMs Across Logical Transformations and Question Answering Tasks. In *Findings of the Association for Computational Linguistics: ACL*, pp. 682–700, 2025. https://aclanthology.org/2025.findings-acl.38/.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1), pp. 207–219, 2022. https://aclanthology.org/2022.cl-1.7/.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations (ICLR)*, 2023. https://arxiv.org/abs/2212.03827.

Radina Dobreva and Frank Keller. Investigating Negation in Pre-trained Vision-and-Language Models. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 350–362, 2021. https://aclanthology.org/2021.blackboxnlp-1.27/.

Wan-Cyuan Fan, Jiayun Luo, Declan Kutscher, Leonid Sigal, and Ritwik Gupta. Tinted Frames: Question Framing Blinds Vision-Language Models. *arXiv preprint arXiv:2603.19203*, 2026. https://arxiv.org/abs/2603.19203.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. https://arxiv.org/abs/2106.02997.

Stefan Heimersheim and Neel Nanda. How to Use and Interpret Activation Patching. *arXiv preprint arXiv:2404.15255*, 2024. https://arxiv.org/abs/2404.15255.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. https://aclanthology.org/D19-1275/.

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 (NeurIPS)*, 36, 2023. https://arxiv.org/abs/2306.14610.

Qidong Huang, Xiaoyi Dong, Pan Zhang, et al. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. https://arxiv.org/abs/2311.17911.

Sohee Kim, Soohyun Ryu, Joonhyung Park, and Eunho Yang. Unveiling the Response of Large Vision-Language Models to Visually Absent Tokens. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 21535–21557, 2025. https://aclanthology.org/2025.emnlp-main.1092/.

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 (CVPR)*, 2024. https://arxiv.org/abs/2311.16922.

Benjamin A. Levinstein and Daniel A. Herrmann. Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks. *Philosophical Studies*, 182, pp. 1539–1565, 2025. https://doi.org/10.1007/s11098-023-02094-3.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023a. https://arxiv.org/abs/2306.03341.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023b. https://aclanthology.org/2023.emnlp-main.20/.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal Tracing of Object Representations in Large Vision Language Models: Mechanistic Interpretability and Hallucination Mitigation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37), pp. 31645–31653, 2026. https://ojs.aaai.org/index.php/AAAI/article/view/40431.

Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014. https://arxiv.org/abs/1405.0312.

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 (CVPR)*, 2024a. https://arxiv.org/abs/2310.03744.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, et al. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *European Conference on Computer Vision (ECCV)*, 2024b. https://arxiv.org/abs/2303.05499.

Sheng Liu, Haotian Ye, and James Y. Zou. Reducing Hallucinations in Large Vision-Language Models via Latent Space Steering. In *International Conference on Learning Representations (ICLR)*, 2025. https://openreview.net/forum?id=LBl7Hez0fF.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *Conference on Language Modeling (COLM)*, 2024. https://arxiv.org/abs/2310.06824.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 2022. https://arxiv.org/abs/2202.05262.

Yannic Neuhaus and Matthias Hein. RePOPE: Impact of Annotation Errors on the POPE Benchmark. *arXiv preprint arXiv:2504.15707*, 2025. https://arxiv.org/abs/2504.15707.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. https://aclanthology.org/2024.acl-long.828/.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object Hallucination in Image Captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. https://aclanthology.org/D18-1437/.

Timothy Schaumlöffel, Martina G. Vilas, and Gemma Roig. Mechanisms of Object Localization in Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31356–31365, 2026. CVF Open Access.

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 (CVPR)*, 2022. https://arxiv.org/abs/2204.03162.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering. *arXiv preprint arXiv:2308.10248*, 2023. https://arxiv.org/abs/2308.10248.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, et al. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. NeurIPS proceedings.

Junyang Wang, Yuhang Wang, Guohai Xu, et al. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. *arXiv preprint arXiv:2311.07397*, 2023. https://arxiv.org/abs/2311.07397.

Jialin Wu, Wei Shi, Han Shen, et al. REVIS: Sparse Latent Steering to Mitigate Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2602.11824*, 2026. https://arxiv.org/abs/2602.11824.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and Why Vision-Language Models Behave Like Bags-of-Words, and What to Do About It? In *International Conference on Learning Representations (ICLR)*, 2023. https://arxiv.org/abs/2210.01936.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations (ICLR)*, 2024. https://arxiv.org/abs/2309.16042.

Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. NegVQA: Can Vision Language Models Understand Negation?. In *Findings of the Association for Computational Linguistics: ACL*, pp. 3707–3716, 2025. https://aclanthology.org/2025.findings-acl.191/.

Andy Zou, Long Phan, Sarah Chen, et al. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023. https://arxiv.org/abs/2310.01405.

A Experimental Protocol

A.1 Data construction

The quartet requires both a present and an absent queried object for the same image. We use the corrected RePOPE object-existence split built from MSCOCO images, selecting X and Y only when the benchmark records support their presence and absence, respectively. For AMBER, we use the official discriminative existence questions and retain object-existence cases with benchmark-supported labels.

The four templates are fixed. Throughout, $+\mathrm{NOT}$ denotes the negative-existential edit implemented by adding *no*:

Case	Question	Polarity	Match	Truth
A	Is there an X in the image?	positive	present	Yes
B	Is there a Y in the image?	positive	absent	No
C	Is there no X in the image?	negative	present	No
D	Is there no Y in the image?	negative	absent	Yes

We select articles grammatically and constrain responses to Yes or No. The decision score is the difference between the corresponding first-token logits; the reported answer is Yes iff $s \geq 0$. All forward passes are deterministic. Holding the image fixed across A–D makes the two edits comparable while reversing the correct answer in each edit direction.

A.2 Models and layer numbering

We use locally stored, unmodified open checkpoints: Qwen3-VL-2B-Instruct (28 decoder blocks), Qwen3-VL-4B-Instruct (36 blocks), and LLaVA-1.5-7B (32 blocks). Layer 0 denotes the embedding/output state before decoder block 1; layer $l > 0$ is the residual stream after block l. The answer position is the final prompt position whose state predicts the first response token. Images, processors, chat templates, and model weights are held fixed within each checkpoint.

A.3 Image-held-out evaluation

Every learned probe uses five image-grouped folds: no condition from a test image enters training. We fit feature standardization on the training fold and apply it unchanged to the test fold. Curves average folds and both transfer directions unless a direction is named. Bootstrap intervals resample whole images to retain dependence among A–D; behavioral comparisons use exact paired tests over images.

A.4 Checkpoint coverage

Table 5: **Experimental coverage by checkpoint.** A checkmark denotes a completed control; addenda marked 2B are focused tests on Qwen3-VL-2B.

Evidence	Qwen3-VL-2B	Qwen3-VL-4B	LLaVA-1.5-7B
Semantic crossover, RePOPE	✓	✓	✓
Semantic crossover, AMBER	✓	✓	✓
Component source/random controls	✓	✓	✓
Localized visual removal	✓	✓	✓
Reversed response codebook	✓	✓	–
Frozen RePOPE→AMBER transfer	✓	✓	✓
Target-label-free truth-axis repair	✓	✓	✓
Visual-only edit / truth-axis decomposition	✓	–	–
Paraphrase / color-attribute robustness	✓	–	–

B Factorial Geometry of the Quartet

Assign $p = +1$ to positive questions and $p = -1$ to negative questions, and $v = +1$ to a present queried object and $v = -1$ to an absent one. The truth label is pv. At every layer and for each image, the four states have the exact expansion

$$\mathbf{h}_l^{p,v} = \boldsymbol{\mu}_l + p\mathbf{a}_l + v\mathbf{b}_l + pv\mathbf{c}_l, \tag{8}$$

$$\boldsymbol{\mu}_l = \tfrac{1}{4}(\mathbf{h}_l^A + \mathbf{h}_l^B + \mathbf{h}_l^C + \mathbf{h}_l^D), \qquad \mathbf{a}_l = \tfrac{1}{4}(\mathbf{h}_l^A + \mathbf{h}_l^B - \mathbf{h}_l^C - \mathbf{h}_l^D), \tag{9}$$

$$\mathbf{b}_l = \tfrac{1}{4}(\mathbf{h}_l^A - \mathbf{h}_l^B + \mathbf{h}_l^C - \mathbf{h}_l^D), \qquad \mathbf{c}_l = \tfrac{1}{4}(\mathbf{h}_l^A - \mathbf{h}_l^B - \mathbf{h}_l^C + \mathbf{h}_l^D). \tag{10}$$

This expansion is an identity for any four states; it does not show that the model has learned truth. The resulting edit vectors are

$$\Delta_l^{\text{NOT},\downarrow} = -2\mathbf{a}_l - 2\mathbf{c}_l, \qquad \Delta_l^{\text{NOT},\uparrow} = -2\mathbf{a}_l + 2\mathbf{c}_l, \tag{11}$$

$$\Delta_l^{\text{OBJ},\downarrow} = -2\mathbf{b}_l - 2\mathbf{c}_l, \qquad \Delta_l^{\text{OBJ},\uparrow} = -2\mathbf{b}_l + 2\mathbf{c}_l. \tag{12}$$

Subtracting opposite answer directions gives $4\mathbf{c}_l$ for either edit, by construction. Their within-edit dot products are $4(\|\mathbf{a}_l\|^2 - \|\mathbf{c}_l\|^2)$ and $4(\|\mathbf{b}_l\|^2 - \|\mathbf{c}_l\|^2)$. For nonzero edit vectors, each cosine has the sign of the corresponding norm difference. The reported mean cosine averages image-level cosines; it is not a ratio of averaged norms.

Equal contrasts need not transfer between edits. The edit-specific class midpoints, $-2\mathbf{a}_l$ and $-2\mathbf{b}_l$, and their variation across images remain unconstrained; they affect the learned weights, intercept, and feature standardization. In one dimension, $a = 0$, $b = 10$, and $c = 1$ give negation classes $-2, +2$ and object-substitution classes $-22, -18$. Both contrasts equal four, yet a zero-threshold classifier trained on the first edit has 50% accuracy on the balanced second edit. Cross-edit performance is therefore an empirical test, not a consequence of the quartet identity.

On an equally weighted set of complete quartets, the source truth axis in Eq. 7 satisfies $\mathbf{t}_l = 2\mathbb{E}[\mathbf{c}_l] = \frac{1}{2}\mathbb{E}[\mathbf{r}_l]$. The remaining questions concern whether this contrast is stable across images, depends on visual evidence, and affects the model's answer. The following controls address these questions at the representation, component, and output levels.

C Behavioral Statistics

The behavioral asymmetry is not fixed across models or datasets. On 491 RePOPE images, Qwen3-VL-2B has a C–D difference of -6.31 percentage points (image-bootstrap 95% CI $[-10.18, -2.44]$; exact McNemar $p = 0.0015$). A versus C differs by 10.79 points ($p \approx 2.2 \times 10^{-16}$). On AMBER, by contrast, Qwen3-VL-4B is strongest on C (99.3%) and weakest on D (66.7%). We therefore test the representation transition across settings without requiring a universal ordering of quadrant errors.

On the same 300-image AMBER sample, LLaVA obtains A/B/C/D accuracies of 99.0%, 53.7%, 64.3%, and 99.0%, respectively. It has 100 all-correct quartets. Errors concentrate in the two false-proposition quadrants, unlike the Qwen3-VL-4B D failure. This difference matters for repair: the bidirectional router evaluates both types of negative-query inconsistency using source-selected thresholds.

D AMBER Representation Replication

Changing the benchmark while retaining object-existence questions tests whether the layer pattern is specific to RePOPE. Table 6 reports the within-edit cosines on all-correct AMBER quartets. Qwen3-VL-2B reproduces the transition near layer 16; in Qwen3-VL-4B, the change spans a broader interval toward layer 24.

Held-out cross-edit decoding sharpens this comparison. It reaches 100% at layer 16 in 2B; in 4B it develops across layers 15–20 and approaches 100% by layer 20. LLaVA rises from 83.5% at L10 to 100% from L15 through L24, while its $+\text{NOT}$ cosine falls from .840 to $-.024$ at L20. Cross-condition polarity remains perfect through layer 15 in 2B and through roughly layer 20 in 4B, then drops as truth-specific structure becomes more prominent. The different timings of cosine and decoding caution against treating either as a universal transition marker.

Table 6: Selected same-edit/opposite-answer cosines on AMBER.

Model	Edit	L10	L15	L20	Late
Qwen3-VL-2B ($n = 215$)	+NOT	.851	.585	−.364	−.617 (L28)
	object swap	.835	.648	−.326	−.564 (L28)
Qwen3-VL-4B ($n = 194$)	+NOT	.861	.872	.422	−.075 (L24)
	object swap	.819	.733	.366	.050 (L24)
LLaVA-1.5-7B ($n = 100$)	+NOT	.840	.383	−.024	.141 (L32)
	object swap	.645	.492	.092	.123 (L32)

D.1 ALL-IMAGE ROBUSTNESS

All-correct selection is not required for a measurable crossover, although it sharpens the estimates. Table 2 compares within-benchmark cross-edit accuracy on all images and all-correct quartets. On RePOPE, the corresponding all-image/all-correct +NOT cosines are .072/−.022 for Qwen3-VL-2B at L16, .474/.445 for Qwen3-VL-4B at L20, and .140/−.0003 for LLaVA at L18. The checkpoint, layer, and population definitions match those in the table.

E COMPONENT PATCHING DETAILS

Cross-edit decoding establishes predictability, not influence on the answer. To test influence, a forward hook replaces the target's answer-position component output with the source output at the same layer for each directed source–target pair. Other positions and modules remain native. A relation is directional when the oriented effect in Eq. 5 is positive. The random control uses a Gaussian vector centered and rescaled to the source component norm before the same replacement. Each image contributes A↔C and B↔D for +NOT, and A↔B and C↔D for object substitution.

Table 7: Qwen3-VL-2B source-swap directional rates. Values aggregate four relations per image.

Edit	Component	L10	L14	L15	L16	L17	L18
+NOT	attention	.38	.615	.863	1.00	.938	.998
	MLP	.38	.838	.343	.673	.548	1.00
Object	attention	.39	.605	.758	1.00	.855	.983
	MLP	.39	.858	.520	.675	.555	1.00

An independently frozen Qwen3-VL-2B confirmatory run gives strict all-four directional rates at L16 of 100% for attention source swaps, 3% for attention random controls, 1% for MLP source swaps, and 7% for MLP random controls. LLaVA shows its strongest strict attention effect at L17 (100%) and a later MLP effect at L18 (64%). This replicates the component sequence, not the full edit-by-edit random-control grid, which was run on Qwen3-VL-2B.

The confirmatory image-level analysis collapses the four directed relations before testing. At L16 attention, source swaps produce strict all-four movement on 100/100 images for both edits; norm-matched random vectors do so on 6/100 images for +NOT and 5/100 for object substitution. The image-level source-only/random-only discordances are 94/0 and 95/0, giving exact paired $p = 1.01 \times 10^{-28}$ and 5.05×10^{-29}. Mean image-level directional-fraction gains over random replacements are .523 (95% image-bootstrap CI [.473,.573]) and .513 ([.468,.558]). Treating each image as one sampling unit retains dependence among its four relations.

F LOCALIZED VISUAL-EVIDENCE INTERVENTION

The component effect could reflect text-driven answer changes rather than visual grounding. To test its visual dependence, Grounding-DINO receives the present-object name and returns the highest-scoring box above fixed detection thresholds. We intervene at severities $\{0, .25, .5, .75, 1\}$. Intermediate levels apply Gaussian blur only inside the box with a feathered edge. At severity 1, the crop is replaced by the mean color of a narrow surrounding ring to preserve a local color baseline. Detector scores,

box coordinates, and box-area fractions are saved for every image. All 100 selected A/C-correct images were localized.

Directions for polarity, presence, and truth are estimated from severity-0 states and frozen. Full removal changes the mean A score from 13.29 to 3.16 and C from -5.86 to 2.58. The L10 polarity projection changes little, from 4.029 to 3.853, whereas L15 presence falls from 21.760 to 12.955 and L16 truth from 56.106 to 24.769. The L16 A-to-C attention patch delta falls from 5.075 to -0.511. This selective weakening supports visual dependence of the middle-layer signal.

G Response-Codebook Control

Visual dependence alone does not distinguish proposition truth from the symbol used to express it. We therefore prompt Qwen3-VL-4B under normal and reversed codebooks, swapping the response symbols assigned to true and false propositions. The analysis retains 191 images answered correctly on all four conditions under both mappings. Five-fold image-held-out probes train under one mapping and test under the other, in both directions. Proposition-truth accuracy is 93.8%, 97.1%, 99.0%, 99.7%, and 100% at layers 15, 16, 18, 20, and 24. Literal output-code accuracy is 6.2%, 2.9%, 1.0%, 0.3%, and 0%. Under binary label reversal, these accuracies are complementary views of one control, not two independent results.

H Frozen Cross-Dataset Transfer

A probe fitted on one benchmark provides a stricter portability test than a new probe fitted on AMBER. Source probes use 387 all-correct Qwen3-VL-2B RePOPE images or 274 all-correct Qwen3-VL-4B images. No AMBER label enters fitting, preprocessing, layer choice, or hyperparameter selection. On all 300 target images, Qwen3-VL-2B truth accuracy is 85.5% at L14, 86.6% at L15, and 90–91% at L16–20. Qwen3-VL-4B rises from 73.9% at L15 to 91.3% at L20 and 92.5% at L24. Restricting evaluation to target all-correct quartets gives the higher values in Fig. 4; both populations use the same frozen source probes.

Truth transfer does not imply that every factor transfers equally. For Qwen3-VL-2B at L15, visual-match transfer is 65.2% from A/B→C/D but 96.0% from C/D→A/B. With 274 all-correct source images, Qwen3-VL-4B frozen quadrant composition attains 69.83% overall accuracy. The visual-match basis therefore remains less portable than the truth signal in these model–dataset settings.

I Visual-Only Edit

To test the transition outside the same-image textual quartet, we construct 133 image pairs spanning 50 object categories. Each pair contains one image with object X and one without it; the question is held fixed while only the visual evidence changes. Object categories, rather than individual images alone, are held out across five folds. All 133 pairs enter the analysis regardless of native behavioral correctness.

The within-edit opposite-answer cosine shows the same layerwise progression as the linguistic edits. At L10, the cosines are 0.879 (+NOT), 0.856 (object substitution), and 0.936 (visual edit); at L16 they are 0.101, 0.306, and 0.325; by L24 they are -0.323, -0.309, and -0.251. A parameter-free mean-difference direction also transfers strongly among all three edits at L16 (96.6–99.3% across directed pairs), providing a readout that does not rely on fitting a high-dimensional discriminative boundary. These results extend the edit-to-answer-direction geometry to a change induced solely by visual evidence.

J Truth-Axis Causal Decomposition

We reuse the disjoint 50-train/50-test Qwen3-VL-2B split for L16 answer-position self-attention. A unit truth direction is estimated only on the training images as true (A,D) minus false (B,C). For each held-out directed source–target relation, let δ be the source-minus-target attention-output difference. We compare the full source replacement with the truth-parallel component $\delta_{\parallel}$ and the orthogonal residual $\delta_{\perp} = \delta - \delta_{\parallel}$, using the same oriented margin effect as Eq. 5.

Table 8: L16 causal decomposition on 50 held-out images. "All-four" requires all four directed relations for an image to move toward the paired source.

Edit	Intervention	Directional	All-four	Mean effect
+NOT	full	100.0%	50/50	3.788
	truth-only	100.0%	50/50	3.356
	orthogonal-only	55.5%	2/50	0.158
Object substitution	full	100.0%	50/50	3.985
	truth-only	100.0%	50/50	3.280
	orthogonal-only	62.0%	2/50	0.284

Truth-only retains 88.6% and 82.3% of the full mean effect for negation and object substitution. The paired image-bootstrap truth-minus-orthogonal differences are 3.198 [3.057, 3.334] and 2.996 [2.849, 3.133], respectively; exact McNemar tests on the all-four criterion compare 50 versus 2 successful images for each edit ($p = 7.1 \times 10^{-15}$). The decomposition therefore localizes most of the measured transition-stage causal effect to the training-derived truth-aligned component.

K Paraphrase Robustness

We evaluate four prespecified question families on the same 100 Qwen3-VL-2B images: the original "there/no" form, "contain/not contain", "show/not show", and "visible/not visible". A frozen truth probe trained on the original wording transfers at L16 to the three alternatives with 97.0%, 93.25%, and 95.75% accuracy; reverse transfer back to the original wording gives 97.25%, 96.25%, and 97.75%. Cross-template, cross-edit transfer at L16 is likewise high: original-negation to alternative-object substitution gives 97.5%, 96.0%, and 86.0%, while original-object substitution to alternative-negation gives 97.5%, 97.0%, and 97.5%. Thus the middle-layer truth geometry transfers across substantial wording changes even when final answer accuracy varies by phrasing.

L Color-Attribute Extension

As a small non-existence extension, we construct 100 GQA scene-graph examples in which a uniquely named object has one explicit color X and a contrasting color Y. The quartet asks whether the object is X/Y and whether it is not X/Y, yielding negation and color-substitution edits with opposing answer directions. Five-fold image-held-out cross-edit transfer rises from 56%/54% at L10 (negation→color / color→negation) to 82%/74.5% at L14 and 89.5%/90% at L16. The corresponding opposite-answer cosines also move from positive early values (0.947 for negation and 0.740 for color at L10) to negative late values (-0.229 and -0.589 at L24). This preliminary predicate extension reproduces the central layerwise reorganization beyond object existence.

M Source-Calibrated Truth-Axis Repair

M.1 Routing and dose selection

The repair tests whether the source truth direction can alter naturally occurring target errors without target-label fitting. For each negated query q^- and its positive counterpart q^+, let s^- and s^+ be native Yes-minus-No margins. A route is eligible when the predictions agree, $|s^+| \geq \tau$, and $|s^+| - |s^-| \geq \gamma$. We select the route side from same-Yes, same-No, or both. The search over $\tau \in \{0, 1, 2, 4, 6, 8, 10\}$ and $\gamma \in \{-8, -4, -2, 0, 2, 4, 8\}$ uses RePOPE only, prioritizing source accuracy, fewer new errors, more repairs, and lower coverage in that order.

At the source transition layer, the truth direction is the difference between mean true- and false-proposition states. Additive doses are $\{.25, .5, 1, 1.5, 2, 3\}$ times this mean difference. Projection doses move a state only far enough to reach the 25th, 50th, 75th, or 90th percentile of the appropriate source truth class. A C-like same-Yes route uses the negative direction; a D-like same-No route uses the positive direction. We choose the dose on source routes only, breaking final ties by lower intervention norm. Section 6 reports the selected configurations and their target outcomes.

Table 9: Detailed RePOPE→AMBER repair outcomes. Internal case values are percentages.

Model	Orig.	Internal	A	B	C	D	Fixed	Broken
Qwen3-VL-2B	89.33	89.75	99.3	76.0	96.7	87.0	5	0
Qwen3-VL-4B	89.75	92.92	98.0	95.0	99.0	79.7	39	1
LLaVA-1.5-7B	79.00	87.58	99.0	53.7	98.3	99.3	103	0

M.2 Repair counts and image-level inference

The strongest selective repair is on LLaVA, which fixes 103 queries and breaks none. Its 20 sampled random directions have mean net repair 4.45 (SD 8.49) and maximum 29. Qwen3-VL-4B fixes 39 queries and breaks one; its 20 random directions have mean net repair 4.75 (SD 6.39) and maximum 22. None of these sampled directions reaches the corresponding truth-axis net repair. Qwen3-VL-2B fixes five D queries and breaks none; ten random directions fix at most two. These controls compare the selected direction with a finite random sample, not with all possible interventions.

Fix/break counts refer to queries, but the paired sampling unit is an image. Let z_i be the change in correct answers for image i. Because A and B are fixed, $z_i \in \{-2, -1, 0, 1, 2\}$. An exact two-sided paired sign test uses positive and negative z_i and omits zero changes. Conditional on m nonzero images under an independent-image, symmetric-sign null, the positive count is $\mathrm{Binomial}(m, \frac{1}{2})$. These bounds characterize the evaluated configurations under that null; Section 6 specifies the evaluation protocol.

We derive conservative sign-test bounds from the aggregate counts by maximizing the p-value over possible allocations of changed queries to images. For LLaVA, 103 fixes and no breaks imply at least $\lceil 103/2 \rceil = 52$ positive images and no negative images. Hence $p \leq 2^{1-52} = 4.44 \times 10^{-16}$. For Qwen3-VL-4B, a break on a separate image leaves at least 20 positive images and one negative image, giving $p \leq 2(\binom{21}{0} + \binom{21}{1})/2^{21} = 2.10 \times 10^{-5}$. A break sharing an image with a fix cancels that image and leaves at least 19 positive images, giving the smaller bound 2^{1-19}. These are worst-case bounds from aggregate counts, not measured image-level discordance counts. Qwen3-VL-2B changes five distinct D answers, giving $p = 2^{1-5} = .0625$.

M.3 Evaluation protocol

Section 6 specifies the source-calibrated evaluation and its timing. The initial same-No LLaVA configuration routes and corrects one of 1,200 target queries. The source-selected bidirectional router has wider coverage, but routing alone is not sufficient: on the same route, external complementation fixes 103 queries and breaks 71, whereas internal displacement fixes 103 and breaks none. The target labels do not select the route or dose.

N Executable Demo

We provide an executable demo using real Qwen3-VL forward passes to illustrate the main analysis procedures in the paper. The demo covers the transition from edit-specific to answer-direction geometry and the layer-16 self-attention source-swap analysis for negation and object substitution. It includes three representative RePOPE examples with Qwen3-VL-2B-Instruct, together with the scripts and instructions needed to reproduce the corresponding demo results. Full code will be made public.