MANY PATHS, ONE COORDINATE: SEMANTIC MEDIATION AND READOUT DISSOCIATION IN VISION–LANGUAGE HALLUCINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models can hallucinate through diverse upstream failures, from missing visual evidence to relying on misleading contextual cues. We ask whether these distinct failure paths nevertheless converge on a common internal computation. Across multiple VLMs, we find that object omissions and fabrications become opposite movements along a shared semantic coordinate representing object presence. In Qwen3-VL and InternVL2, causal interventions on unseen clean–error transitions show that independently estimated versions of this coordinate mediate a large fraction of the late-layer behavioral effect, while matched controls have little influence. The semantic mediator is nevertheless distinct from the model's direct output readout: it remains stable across changes in answer surface even when the corresponding readout changes substantially, and removing its sample-specific coupling to the readout nearly eliminates direct behavioral control. Additional composition and trajectory experiments show that heterogeneous perturbations can follow nonlinear upstream paths while converging onto a compact downstream computation. Together, these results suggest a layered mechanism for vision–language hallucination in which many upstream paths converge on a shared semantic state that is subsequently translated into behavior by a distinct low-dimensional readout. In Qwen3-VL, the same mediation–readout organization also appears in binary counting, attribute, and relation judgments.

1 INTRODUCTION

Vision–language models (VLMs) can make the same wrong object-presence judgment for very different reasons. An object may be omitted because its visual evidence is weak, blurred, cropped, or occluded; an absent object may be fabricated because of contextual priors or learned co-occurrence (Li et al., 2023b; Si et al., 2025; Yang et al., 2025). These failure modes appear heterogeneous at their source. Yet interpretability studies often uncover surprisingly simple latent structure associated with model behavior. This raises a mechanistic question: *do different hallucination paths remain distinct throughout the network, or do they converge on a common internal computation before producing an incorrect answer?*

Answering this question requires distinguishing several notions that are easily conflated. A latent direction may first describe a *semantic organization*: different errors may become organized around a common variable in representation space. That variable may additionally be a *causal mediator*, in the sense that movement along it carries the behavioral effect of a clean–error transition. Neither property, however, implies that the same direction is the model's *readout*: the local rule that converts a hidden state into the final answer. Our hypothesis is therefore layered: heterogeneous upstream failures may converge onto a shared semantic state, which is subsequently translated into behavior by a distinct downstream readout.

We study this hypothesis through matched correct–error transitions in binary object-presence judgments (Fig. 1). Across multiple VLM families, omission and fabrication errors develop a common signed geometry: as representations progress through the network, the corresponding clean–error displacements become increasingly opposed along a shared object-presence coordinate. Controlled changes to visual evidence, contextual prior, and occlusion organize around the same coordinate. The

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.

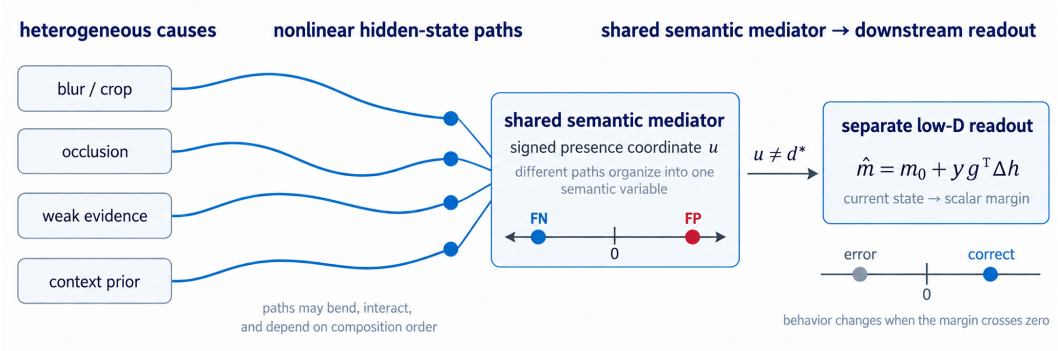


Figure 1: **Mechanistic hypothesis.** Heterogeneous clean–error paths become organized around a shared signed object-presence coordinate. An independently estimated version of this coordinate mediates most of the behavioral effect of held-out clean–error transitions, while a distinct low-dimensional readout maps the resulting semantic state to the answer.

effect is stable across resampling, reconstruction, and pairing controls, suggesting that it reflects a reproducible semantic organization rather than a peculiarity of a particular set of examples.

Geometry alone, however, does not establish a mechanism. We therefore estimate the presence coordinate independently and freeze it before evaluating unseen clean–error transitions. Across Qwen3-VL and InternVL2, decomposing each transition into its component along this frozen coordinate and its orthogonal residual reveals a strong asymmetry: the axis-aligned component recovers a large fraction of the late-layer behavioral effect, whereas matched orthogonal and wrong-layer controls contribute little. Thus the shared coordinate is not merely predictive of hallucination; it mediates a substantial part of the change in model behavior.

The semantic mediator is nevertheless distinct from the model's direct output readout. Changing the answer surface leaves the presence coordinate largely invariant while substantially changing the corresponding readout gradients. Conversely, directions that exert the strongest direct control over the answer lie in a compact readout field that is only weakly aligned with the semantic coordinate. Removing the sample-specific coupling between the presence direction and this readout nearly eliminates its direct behavioral effect. The model therefore need not represent semantic state and read it out along the same direction.

This separation also clarifies where the complexity of hallucination lies. Composition and trajectory experiments show that different visual perturbations can follow nonlinear and order-dependent paths through hidden space even when they ultimately converge on the same semantic variable. By contrast, the downstream mapping from the realized semantic state to the answer remains comparatively low-dimensional. Together, these results support a layered account of VLM hallucination: *many heterogeneous upstream paths converge on a shared object-presence coordinate, and a distinct compact readout translates that coordinate into behavior.*

Our contributions are threefold:

- We identify a cross-model signed geometry in which omission and fabrication errors become opposite movements along a shared object-presence coordinate.
- Using independently estimated axes and held-out interventions in Qwen3-VL and InternVL2, we show that this coordinate is a causal mediator of clean–error transitions rather than only a correlational probe.
- We dissociate semantic mediation from direct output readout, show that this layered organization extends to binary counting, attribute, and relation judgments, and show that nonlinear upstream paths can converge onto a comparatively compact downstream state-to-answer computation.

2 MECHANISTIC FRAMEWORK

We separate three questions that are often conflated when interpreting a "hallucination direction." First, do different errors become organized around a common semantic variable? Second, does movement along that variable carry the behavioral effect of a clean–error transition? Third, is that variable itself the direction that directly controls the model's answer? We formalize these questions as *semantic geometry*, *causal mediation*, and *readout*, respectively.

Throughout, we consider matched correct–wrong states at decoder layer ℓ. For example i, let

$$\Delta \mathbf{h}_i^{(\ell)} = \mathbf{h}_{i,\text{correct}}^{(\ell)} - \mathbf{h}_{i,\text{wrong}}^{(\ell)}. \tag{1}$$

For binary object-presence judgments, we define the answer margin

$$r_i(\mathbf{h}) = z_i(\text{Yes} \mid \mathbf{h}) - z_i(\text{No} \mid \mathbf{h}), \qquad m_i(\mathbf{h}) = y_i r_i(\mathbf{h}), \tag{2}$$

where $y_i \in \{+1, -1\}$ orients the margin toward the ground-truth answer. Thus larger m_i always corresponds to a more correct object-presence judgment.

2.1 A SIGNED SEMANTIC COORDINATE

An omission and a fabrication reverse the semantic meaning of a correct–wrong transition. We therefore orient the two error types before averaging. Let $s_i = +1$ for false-negative (FN) transitions and $s_i = -1$ for false-positive (FP) transitions. Given a source set $\mathcal{S}$, we define the signed object-presence axis at layer ℓ as

$$\hat{\mathbf{u}}^{(\ell)} = \text{normalize}\left(\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} s_i \Delta \mathbf{h}_i^{(\ell)}\right). \tag{3}$$

With this convention, the positive orientation of $\hat{\mathbf{u}}$ corresponds to increasing object presence. If FN and FP transitions become antiparallel, signing them in this way maps both onto the same semantic coordinate.

Equation 3 describes the geometry of *transitions*; it does not require FN and FP hidden states themselves to coincide. Whether this geometry is stable across samples, perturbations, and model families is tested empirically in Section 3.

2.2 FROZEN-AXIS MEDIATION

A semantic coordinate need not carry the behavioral effect of the transition that revealed it. To test this, we estimate $\hat{\mathbf{u}}^{(\ell)}$ on a source set $\mathcal{S}$ and freeze it before evaluating a disjoint target set $\mathcal{T}$.

For target transition $i \in \mathcal{T}$, decompose

$$\Delta \mathbf{h}_{i,u}^{(\ell)} = \left(\Delta \mathbf{h}_i^{(\ell)\top} \hat{\mathbf{u}}^{(\ell)}\right) \hat{\mathbf{u}}^{(\ell)}, \tag{4}$$

$$\Delta \mathbf{h}_{i,\perp}^{(\ell)} = \Delta \mathbf{h}_i^{(\ell)} - \Delta \mathbf{h}_{i,u}^{(\ell)}. \tag{5}$$

Starting from the wrong state, we compare three interventions:

$$\mathbf{h}_{i,\text{full}} = \mathbf{h}_{i,\text{wrong}} + \Delta \mathbf{h}_i, \tag{6}$$

$$\mathbf{h}_{i,u} = \mathbf{h}_{i,\text{wrong}} + \Delta \mathbf{h}_{i,u}, \tag{7}$$

$$\mathbf{h}_{i,\perp} = \mathbf{h}_{i,\text{wrong}} + \widetilde{\Delta \mathbf{h}}_{i,\perp}, \tag{8}$$

where the orthogonal control is rescaled to match the mediator component exactly,

$$\widetilde{\Delta \mathbf{h}}_{i,\perp} = \frac{\Delta \mathbf{h}_{i,\perp}}{\|\Delta \mathbf{h}_{i,\perp}\|}\|\Delta \mathbf{h}_{i,u}\|. \tag{9}$$

We measure how much of the full behavioral effect is recovered by the axis-aligned intervention using

$$R_u = \frac{m(\mathbf{h}_u) - m(\mathbf{h}_{\text{wrong}})}{m(\mathbf{h}_{\text{full}}) - m(\mathbf{h}_{\text{wrong}})}. \tag{10}$$

We report the corresponding raw margin changes together with R_u and compare against equal-norm orthogonal, random-orthogonal, and wrong-layer controls. Here, *mediation* refers specifically to intervention-based recovery of an observed clean–error effect along an independently estimated component; it does not assume that this component is a unique causal variable.

2.3 Separating semantic mediation from readout

The semantic axis describes variation in hidden state, whereas direct output control is sample dependent. We characterize the latter through the local gradient of the answer margin,

$$\mathbf{g}_i^{(\ell)} = \nabla_{\mathbf{h}^{(\ell)}} r_i(\mathbf{h}). \tag{11}$$

A shared direction in this gradient field is summarized by the normalized mean

$$\mathbf{d}_\ell^* = \text{normalize}\left(\frac{1}{N}\sum_i \frac{\mathbf{g}_i^{(\ell)}}{\|\mathbf{g}_i^{(\ell)}\|}\right). \tag{12}$$

We test whether semantic mediation and readout coincide in three complementary ways. First, we change the answer surface while preserving the underlying object-presence judgment and compare the resulting semantic axes and gradient fields. Second, we measure the dimensionality and behavioral control strength of the gradient field relative to $\hat{\mathbf{u}}$. Third, for each example we remove from the semantic direction the component parallel to its local readout,

$$\mathbf{u}_i^{\perp g} = \mathbf{u} - \frac{\mathbf{g}_i^\top \mathbf{u}}{\|\mathbf{g}_i\|^2}\mathbf{g}_i. \tag{13}$$

If $\mathbf{u}$ remains semantically stable across answer surfaces while $\mathbf{u}_i^{\perp g}$ loses direct control of the answer, then semantic state and behavioral readout are experimentally dissociable.

For trajectory and composition experiments, we additionally use the same clean-state gradient as a first-order diagnostic. For a realized displacement $\Delta\mathbf{h}_i$ from clean state $\mathbf{h}_{i,0}$,

$$\widehat{m}_i = m_i(\mathbf{h}_{i,0}) + y_i \mathbf{g}_i^\top \Delta\mathbf{h}_i. \tag{14}$$

This approximation is used only to test how well a fixed local readout predicts behavior as upstream trajectories become nonlinear.

3 Experiments

3.1 Experimental setup

Our primary mechanistic analysis uses Qwen3-VL-2B-Instruct (Qwen Team, 2025) on binary object-presence queries from POPE (Li et al., 2023b), with hidden states taken at the final prompt position. Qwen3-VL-2B and InternVL2-8B (Chen et al., 2024a;b) independently support the complete geometry–mediation–readout analysis, with InternVL2 additionally replicating the composition stress test. Qwen2.5-VL-7B (Bai et al., 2025) provides a third-model test of the signed geometry; the full evidence scope is summarized in Appendix Table 5. Semantic-breadth tests additionally use binary counting judgments on CountBench (**?**) and attribute/relation judgments from AMBER (Wang et al., 2023). Unless stated otherwise, axes are estimated from data disjoint from the examples on which they are evaluated. Dataset construction, prompting, de-overlap rules, and complete robustness analyses are given in Appendix A.

3.2 Omission and fabrication converge onto a signed presence coordinate

We first ask whether omission and fabrication remain geometrically distinct throughout the network. In Qwen3, their mean correct–wrong displacements are only weakly related in early layers but become sharply antiparallel from mid-depth onward (Fig. 2a), reaching cosine -0.993 at late layers. Importantly, the full hidden states do not collapse: the convergence is specific to the signed transition component. The effect is stable under bootstrap resampling, class balancing, split-half reconstruction, leave-one-out analysis, and a pairing/sign-flip null (Appendix A).

The same organization appears in independent architectures. Qwen2.5-VL-7B and InternVL2-8B both reach late FN/FP displacement cosine near -0.93 (Fig. 2b). The layer at which the organization emerges differs across architectures, suggesting a shared depth-dependent computation rather than a universal layer index.

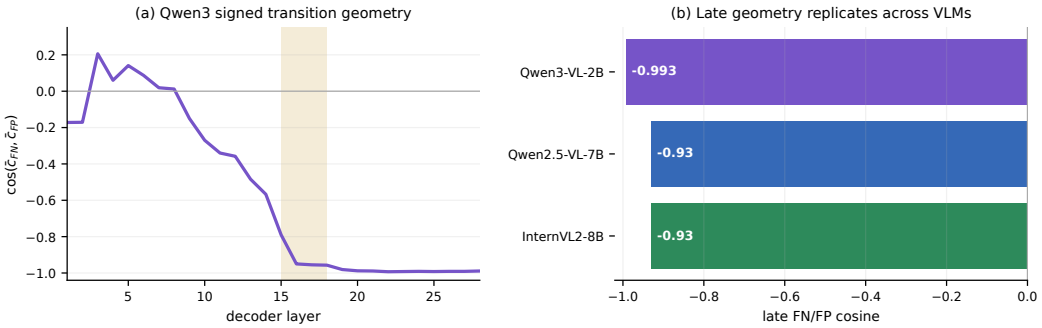


Figure 2: **Omission and fabrication converge onto a shared signed presence geometry.** (a) In Qwen3, FN and FP correct–wrong displacements become strongly antiparallel from mid-depth onward. (b) The late signed geometry replicates in Qwen2.5-VL-7B and InternVL2-8B.

Table 1: **Heterogeneous upstream factors converge onto a shared late presence organization.** Alignment is measured against the Qwen3 L21 presence axis. Repair reports the targeted intervention at an early pre-convergence layer (L9) and a later post-convergence layer (L18). Dashes denote conditions not used for the targeted repair comparison.

Factor	Alignment at L21	Early signature	Targeted repair L9 → L18
Reduced object evidence	0.82	L11	–
Contextual prior	0.51	L5	0% → 59%
Occlusion	−0.89	L9	0% → 38%
Scale control	0.28	–	–

Different upstream factors converge onto the same organization. Controlled perturbations provide an independent test of what feeds this coordinate (Table 1). Different factors show their strongest signatures at different early depths, yet align with the same late presence organization. Consistently, prior- and occlusion-derived interventions fail to repair targeted errors at L9 but become effective by L18. Because the late directions also affect untargeted controls, we interpret this as convergence onto a shared semantic variable rather than recovery of factor-specific inverse mechanisms or an exact causal onset layer.

Semantic specificity. A separate donor-transport control asks whether the late organization is merely a generic binary-decision geometry (Appendix Table 8). Same-semantic presence donors transfer across distinct visual perturbation paths, whereas color-attribute and shuffled donors do not. A stricter cross-task test reaches the same conclusion: presence and attribute directions can be highly aligned geometrically, yet cross-semantic donors fail where matched donors succeed. Thus high cosine does not imply causal interchangeability. These controls support interpreting the converged direction as a task-semantic mediator rather than as an arbitrary binary decision axis (Appendix A).

3.3 A FROZEN PRESENCE AXIS MEDIATES HELD-OUT TRANSITIONS

Geometry does not establish that a direction carries behavioral effect. We therefore estimate the presence axis on an independent source set and freeze it before intervention. The final source contains 44 natural transitions spanning 36 images. We evaluate on 163 clean–error transitions spanning 130 different images, with zero source–target image or exact-question overlap. No target transition is used to construct or update the axis.

Figure 3 shows a sharp transition in causal relevance, and Table 2 reports the corresponding intervention effects. The frozen-axis component has little effect at early layers but recovers most of the clean–error margin change from mid-to-late layers onward. In contrast, an exactly norm-matched orthogonal component and a pre-specified wrong-layer axis remain weak at the same late layers. Thus the late effect cannot be explained simply by intervention magnitude or by applying an arbitrary presence-related direction.

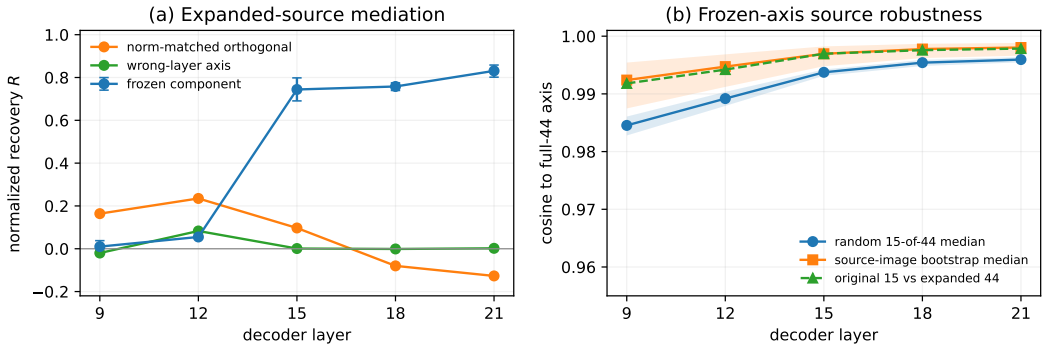


Figure 3: **Independent frozen-axis mediation.** Left: across 163 held-out clean–error pairs, the independently estimated presence component becomes dominant after mid-depth, while norm-matched and wrong-layer controls remain weak. Right: source-image bootstrap, subsampling, and an independently constructed earlier axis recover nearly the same late direction. Source and target sets have zero image and exact-question overlap.

Table 2: **Frozen-axis mediation in Qwen3.** Orthogonal vectors are exactly norm-matched to the frozen-axis component. Raw margin change is reported together with the normalized mediated fraction.

Layer	R_u	$R_{\perp,\text{match}}$	R_{wrong}	Δm_u	Δm_{full}
L9	0.010	0.164	–	0.03	0.20
L12	0.055	0.235	0.083	0.01	0.32
L15	0.744	0.097	0.001	1.98	2.72
L18	0.758	−0.080	−0.001	5.85	7.66
L21	0.831	−0.127	0.002	7.35	8.80

The result is robust to how the source axis is constructed. Source-image bootstrap, random source subsampling, polarity-balanced reconstruction, and the original independent source axis all produce nearly collinear late directions. Target-image-cluster bootstrap intervals remain tight across the late layers, and the recovered effect remains substantial when FN and FP targets are analyzed separately (Appendix A).

A reverse intervention provides a complementary check. Removing only the frozen-axis component from a clean state produces a late-layer margin change comparable to the full clean-to-error transition, whereas its norm-matched orthogonal control causes no flips in the archived test. InternVL2 independently reproduces the late-layer asymmetry: $R_u = 0.568/0.632/0.681$ at L20/L26/L31, while exactly norm-matched orthogonal recovery is $0.074/-0.021/-0.058$ at the same layers. This replication uses 38 source transitions spanning 31 images and 124 held-out targets spanning 101 image clusters, with zero source–target image or exact-question overlap. At L31, recovery remains similar for FN and FP targets (0.706 and 0.645), with a target-image-cluster bootstrap 95% interval of $[0.625, 0.729]$ (Appendix A.14).

3.4 The semantic mediator is distinct from the output readout

We next test whether the shared semantic mediator is simply the model's native decision direction. Changing the answer surface from Yes/No to Present/Absent preserves semantic behavior on 400 examples and leaves the signed presence coordinate nearly unchanged ($\cos = 0.973$), while the corresponding answer-margin gradients align only at 0.347. As a control, swapping answer-option order while retaining the same answer surface leaves the gradients much more strongly aligned ($\cos = 0.942$). The large gradient change therefore reflects a change in readout rather than prompt formatting alone. A complementary control projects the empirical decision-gradient span out of the FN/FP transition displacements before recomputing their geometry. The late antiparallel relation remains essentially unchanged ($\cos = -0.98$ to -0.99), and the archived signed transport effect

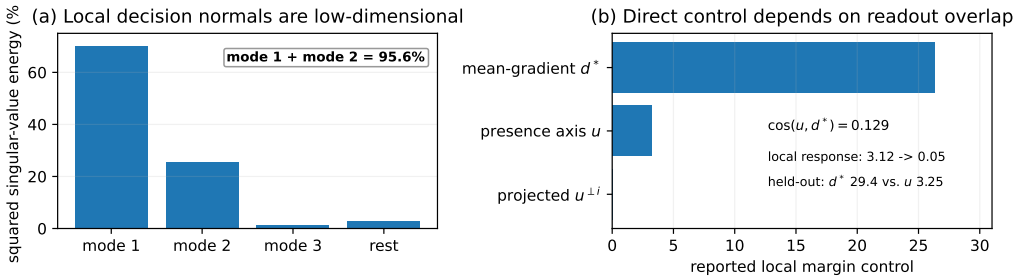


Figure 4: **Semantic mediation and direct readout are distinct.** (a) L21 answer gradients are dominated by two modes. (b) The dominant readout direction controls the margin much more strongly than the semantic axis; removing sample-local readout overlap from the semantic direction nearly eliminates its direct effect.

persists. Thus the semantic convergence is not created by the measured answer-gradient component (Appendix A).

The readout field is itself simple, but it is geometrically different from the semantic axis. Across a separate balanced set of 600 examples, its first two singular modes dominate the gradient field. The dominant mean-gradient direction d^* is only weakly aligned with the presence axis, yet produces far stronger direct margin control in held-out forward evaluation (Fig. 4).

A direct projection test sharpens this dissociation (Fig. 4b). For each sample we remove from the presence direction its component parallel to the local answer gradient; the resulting direction has almost no direct local effect. Thus the semantic coordinate affects the answer through its sample-dependent coupling to the readout, rather than because the semantic and readout directions are identical. InternVL2 independently reproduces this dissociation: answer-surface remapping preserves the semantic axis ($\cos = 0.958$) while changing the gradient field ($\cos = 0.412$; option-order control 0.906). Its first two normalized-gradient modes explain 92.9% of squared singular-value energy, and per-example projection reduces the measured direct response from 2.68 to 0.08 (Appendix A.15).

3.5 The layered mechanism generalizes across semantic variables

We next test whether the organization above is specific to object presence. Using the same held-out, frozen-axis construction, we form same-image, same-question clean–error pairs for binary counting, attribute, and relation judgments. Table 3 shows that the semantic component recovers 0.77–0.93 of the late-layer transition effect, while norm-matched orthogonal controls remain near zero. Across answer-surface remappings, the semantic direction remains highly stable (0.95–0.99) while the corresponding readout gradient changes substantially (0.18–0.35). The sharpness of FN/FP geometry varies by domain, so we claim generality of the *layered mediation–readout organization*, not a universal hallucination axis. Full layerwise results and pairing controls are in Appendix A.12.

Table 3: **Semantic generality in Qwen3-VL-2B (L21).** Presence uses the matched Exp43B protocol for comparability (primary estimate: $R_u = 0.831$ in Table 2); surface columns compare u and g across equivalent answer forms.

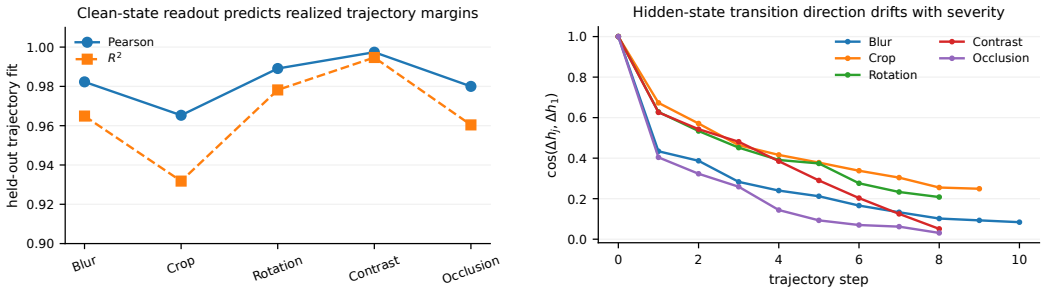
Variable	FN/FP cos.	R_u	$R_\perp$	surface $\cos(u)$ / $\cos(g)$
Presence	-0.993	0.824	-0.130	0.973 / 0.347
Counting	-0.733	0.920	0.043	0.960 / 0.221
Attribute	-0.571	0.773	0.006	0.966 / 0.230
Relation	-0.398	0.931	-0.043	0.964 / 0.183

3.6 Nonlinear upstream paths coexist with a compact downstream readout

Finally, we ask whether a compact downstream readout implies that the upstream path is itself simple. It does not. Subcritical perturbations compose strongly non-additively (Table 4): combining blur,

Table 4: **Non-additive perturbation composition.** Each single perturbation is chosen to remain subcritical. Joint flips report the fraction of eligible examples whose answer changes under the combination. "Realized-state" applies the frozen clean-state readout to the actual joint hidden state.

Model / combination	Joint flips	Additive r	Realized-state r
Qwen3: blur+crop	79.9%	0.230	**0.818**
Qwen3: blur+rotation	55.1%	0.134	**0.901**
Qwen3: crop+rotation	57.8%	0.243	**0.904**
Qwen3: triple	77.6%	0.224	**0.829**
InternVL2: blur+crop	61.8%	0.173	**0.691**



(a) Frozen clean-state readout vs. realized margin.

(b) Hidden trajectories rotate strongly.

Figure 5: **Nonlinear upstream path, stable downstream map.** Across five perturbation families, hidden trajectories change direction substantially while a readout measured at the clean state remains predictive of the realized margin.

crop, and rotation frequently changes the answer even though each perturbation alone is selected to preserve it. Additive combinations of the single-perturbation effects predict the joint outcome poorly, whereas applying the fixed clean-state readout to the *realized* joint hidden state remains strongly predictive. InternVL2 shows the same qualitative separation.

A trajectory analysis shows the same phenomenon continuously (Fig. 5). Across 14,976 held-out states generated by blur, crop, rotation, contrast, and occlusion, the answer gradient measured once at the clean state remains highly predictive of the realized correctness margin (pooled Pearson 0.972). Meanwhile the hidden trajectory bends sharply: by the end of the trajectory, displacement directions have cosine only 0.09–0.21 with the initial direction. Upstream image-to-state dynamics can therefore be strongly nonlinear even when the downstream state-to-answer map remains comparatively compact.

4 Discussion

Our results support a layered account of object-presence hallucination, with the same mediation–readout organization extending in Qwen3-VL to binary counting, attribute, and relation judgments. Different failures can arise from different upstream perturbations and follow different trajectories through hidden space, yet become organized around a shared signed semantic variable at later stages of computation. In both Qwen3-VL and InternVL2, an independently estimated version of this variable carries a large fraction of the behavioral effect of held-out clean–error transitions while remaining distinct from the model's direct output readout. The readout is itself low-dimensional, but changes substantially under answer-surface remapping. Together, these findings support a sequence in which heterogeneous upstream computation converges onto a task-relevant semantic state that is subsequently translated into behavior by a separate readout.

The cross-model replication sharpens this mechanistic picture. Qwen3-VL and InternVL2 independently exhibit the same geometry–mediation–readout separation, while Qwen2.5-VL provides a third architectural replication of the signed geometry. In InternVL2, the frozen coordinate also transfers to naturally occurring errors on unperturbed images, and matched alternative directions recover substantially less of the held-out transition effect (Appendix A.16–A.17). The frozen coordinate

therefore transfers beyond a single architecture and beyond the perturbation-induced transitions used for the controlled mediation test. Across semantics, the mediator remains task-specific: the strength of signed FN/FP geometry varies, even when frozen-axis mediation and answer-surface dissociation remain strong (Table 3).

The separation between semantic state and readout also explains why simple downstream structure can coexist with complex error formation. Object evidence, contextual prior, occlusion, and image perturbations become informative at different depths, and composed perturbations interact non-additively while their hidden trajectories rotate substantially. Yet once a semantic state is realized, a compact local readout remains strongly predictive of the answer. This distinction suggests that mechanistic diagnosis and direct behavioral control need not use the same direction: the coordinate that best summarizes and mediates a naturally occurring semantic transition can differ from the direction that most efficiently changes the output.

5 Related Work

Vision–language hallucination. Object hallucination has been studied from early image captioning systems to modern VLMs (Rohrbach et al., 2018; Li et al., 2023b). Existing work documents multiple sources of error, including weak visual evidence, language and co-occurrence priors, attention imbalance, and distinct mechanisms underlying omission and fabrication (Zhou et al., 2024; Si et al., 2025; Yang et al., 2025). A complementary literature focuses on mitigation through alignment and hallucination-aware training, decoding or post-hoc correction, and activation- or representation-level intervention (Sun et al., 2024; Yu et al., 2024; Leng et al., 2024; Huang et al., 2024; Jiang et al., 2025; Liu et al., 2025; Zhang et al., 2026). These studies establish that hallucination can arise through heterogeneous routes and can often be reduced by targeted intervention. We instead ask a mechanistic question: whether different upstream routes become organized around a common internal variable before producing the same object-presence decision.

Latent representations and behavioral steering. Language-model representations often contain low-dimensional directions associated with truth, knowledge, refusal, and other behavioral variables; intervening along such directions can predictably alter model outputs (Burns et al., 2023; Ferrando et al., 2025; Li et al., 2023a; Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023). Related VLM work likewise uses internal representations to diagnose or modify hallucination behavior (Jiang et al., 2025; Liu et al., 2025; Yang et al., 2025). However, decodability and steerability alone do not determine what role a direction plays in the model's native computation. In particular, a direction that organizes semantic states need not be the component that mediates a naturally occurring error transition, nor the direction that most directly controls the output.

Causal mechanisms and readout. Causal-interpretability work formalizes this distinction by studying model components through counterfactual intervention rather than representation alone (Geiger et al., 2021), while recent work further distinguishes conceptual representations from functional directions and documents cases in which intervention produces latent states different from those arising naturally (Opielka et al., 2026; Grant et al., 2026). Our work brings these distinctions together in VLM object hallucination. Rather than treating a "hallucination direction" as a single object, we separately test whether a signed presence coordinate (i) captures a stable semantic organization, (ii) carries the behavioral effect of held-out clean–error transitions, and (iii) coincides with the model's direct output readout. This separation lets us distinguish semantic representation, causal mediation, and behavioral control within the same mechanistic setting.

6 Conclusion

Across multiple VLMs, omission and fabrication become opposite movements along a shared signed presence coordinate, and independently estimated versions of this coordinate carry a large fraction of held-out clean–error effects. The semantic mediator is distinct from the model's direct output readout; in Qwen3-VL, this mediation–readout separation also extends to binary counting, attribute, and relation judgments. Together, these results support a layered account in which nonlinear upstream paths converge onto a semantic state before a compact readout maps it to behavior.

Reproducibility statement

Appendix A records the fixed data roles and de-overlap rule, signed-axis stability analysis, independent frozen-axis mediation protocol, norm-matched controls, answer-surface intervention, readout analysis, controlled factor experiments, semantic-generality and pairing controls, cross-model replication, natural-error transfer, direction baselines, and nonlinear downstream stress tests.

AI use statement

Generative AI tools were used to assist with manuscript drafting, language editing, and polishing. All scientific content, analyses, results, interpretations, and conclusions were developed and verified by the authors. The authors reviewed all AI-assisted text and take full responsibility for the final manuscript.

References

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

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024b.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations*, 2025.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586, 2021.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, 2026.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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*, 2024.

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

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*, 2024.

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*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b.

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*, 2025.

Gustaw Opielka, Hannes Rosenbusch, and Claire Stevenson. Causality neq invariance: Function and concept vectors in llms. In *International Conference on Learning Representations*, 2026.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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*, pages 15504–15522, 2024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

Guangzong Si, Hao Yin, Xianfei Li, Qing Ding, Wenlong Liao, Tao He, and Pai Peng. Two causes, not one: Rethinking omission and fabrication hallucinations in mllms. *arXiv preprint arXiv:2509.00371*, 2025.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

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

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *International Conference on Learning Representations*, 2025.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination begins where saliency drops. In *International Conference on Learning Representations*, 2026.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Supplementary Material

A.1 Scope of the evidence chain

Table 5: Scope of the evidence chain. The complete geometry–mediation–readout analysis is independently established in Qwen3-VL-2B and InternVL2-8B; Qwen2.5-VL-7B provides an additional test of the signed geometry.

Model	signed geometry	frozen mediation	readout	composition
Qwen3-VL-2B	✓	✓	✓	✓
Qwen2.5-VL-7B	✓	–	–	–
InternVL2-8B	✓	✓	✓	✓

A.2 Primary data and hidden-state extraction

The primary signed-geometry analysis uses a fixed balanced pool of 800 examples from the POPE Random setting: 400 positive and 400 negative object-presence queries, of which 728 are clean-correct. The paired-mediation target pool is constructed separately from a deduplicated union of POPE Random, Popular, and Adversarial candidates using the same clean-correct criterion and a fixed blur/crop perturbation ladder. Hidden states are read from the final prompt position at each decoder layer. Natural matched transitions are constructed from examples for which a controlled or naturally occurring change yields a correct/wrong pair with the same object-presence query. The signed-geometry discovery analysis uses 35 unique transitions (11 FN and 24 FP) from the Random pool. The mediation axis uses a separate expanded natural-transition candidate pool described below. No per-pair normalization is applied before averaging in Eq. 3.

For cross-model signed geometry, the same orientation convention is used in Qwen3-VL-2B, Qwen2.5-VL-7B, and InternVL2-8B. Late FN/FP mean transition displacements have cosine approximately -0.99, -0.93, and -0.93, respectively. The onset layer of strong antiparallel organization differs across architectures; we therefore report the layerwise curves rather than treating a normalized onset index as an invariant quantity.

A.3 Signed-axis stability and paired mediation protocol

The 35-pair Qwen3 source set contains 11 FN and 24 FP transitions. A 10,000-draw stratified bootstrap gives late FN/FP cosine mean -0.982 with percentile interval approximately $[-0.992, -0.950]$. In a 2,000-draw pairing/sign-flip null, no draw is at least as extreme as the observed late geometry, giving $p = 1/(2000+1) < 5 \times 10^{-4}$ with +1 correction. Balanced subsampling of FP down to the FN count and split-half/leave-one-out analyses give the same qualitative result.

The paired mediation experiment uses a stricter expanded frozen source axis. The natural-transition candidate pool contains 64 pairs. Removing all 20 candidates whose exact question occurs in a mediation target leaves 44 source pairs (32 FN, 12 FP) spanning 36 source-image clusters. The 163 targets are same-image, same-question clean–error pairs (101 FN, 62 FP) spanning 130 target-image clusters. Exact image-signature and question-string audits give zero source–target overlap. The clean image is answered correctly; a blur/crop ladder produces the paired error. The frozen axis is never re-estimated on target transitions. The archived validation contains exactly $163 \times 5 = 815$ target–layer patch rows and 100 norm-matched random orthogonal nulls per target–layer pair ($81{,}500$ null rows total). Reconstructing each saved axis from the signed mean of the 44 archived source deltas agrees to maximum elementwise error 5.1×10^{-9}.

For each target and layer, the full clean–error displacement is decomposed into Δh_u and $\Delta h_\perp$. The primary controls are (i) an orthogonal vector scaled to exactly $\|\Delta h_u\|$; (ii) a Gaussian random vector projected orthogonal to u and scaled to the same norm; and (iii) a pre-specified wrong-layer control that applies u_{L9} at later layers. We report both pair bootstrap and a 10,000-draw target-image-cluster bootstrap in which all questions/perturbations from the same image move together. Raw margin change is the primary outcome; the normalized recovery R_u in Eq. 10 is reported when the full-patch denominator is safely nonzero. At the archived threshold $|m_{\text{full}} - m_{\text{error}}| \geq 0.5$, no target is excluded.

Table 6: Independent frozen-axis mediation in Qwen3. Orthogonal vectors are exactly norm-matched to the frozen-axis component.

Layer	R_u	$R_{\perp,\mathrm{match}}$	R_{wrong}	Δm_u	Δm_{full}
L9	0.010	0.164	–	0.03	0.20
L12	0.055	0.235	0.083	0.01	0.32
L15	0.744	0.097	0.001	1.98	2.72
L18	0.758	-0.080	-0.001	5.85	7.66
L21	0.831	-0.127	0.002	7.35	8.80

The expanded axis is highly stable to source sampling. A 5,000-draw bootstrap over the 36 source-image clusters gives axis-cosine 95% intervals $[0.9948, 0.9982]$, $[0.9962, 0.9987]$, and $[0.9966, 0.9988]$ at L15/L18/L21. Across 3,000 random 15-of-44 subsets, median cosines to the full axis are $0.9937/0.9954/0.9959$; the original Random-only 15-source axis has cosine $0.9970/0.9976/0.9979$ with the expanded mixed-setting axis. Because the expanded source is polarity-imbalanced (32 FN, 12 FP), we also repeat 3,000 reconstructions with exactly 12 FN and 12 FP per draw; median cosines are $0.9974/0.9981/0.9982$ and every draw exceeds 0.95.

For the target effect, 10,000 target-image-cluster bootstrap intervals for R_u are $[0.691, 0.798]$, $[0.740, 0.775]$, and $[0.802, 0.858]$ at L15/L18/L21. A companion analysis with the original 15-source axis—nearly collinear with the expanded axis—provides polarity-stratified and reverse-intervention checks: at L18, FN and FP R_u are 0.779 and 0.736; at L21 they are 0.886 and 0.727. Subtracting only Δh_u from the clean state gives mean margin changes -8.84 at L18 and -9.30 at L21, versus -8.11 and -8.83 for the full clean-to-error transition; its norm-matched orthogonal control causes zero flips. We treat these companion numbers as robustness evidence rather than as the primary expanded-source estimates.

A.4 Controlled factor construction

The factor-alignment analysis uses deterministic paired constructions. *Object evidence* uses natural POPE pairs with the same question and one object-present versus one object-absent image. *Contextual prior* uses questions with multiple object-absent images and contrasts the highest- versus lowest-margin negative image under the identical question. *Occlusion* takes clean-correct positive examples and replaces a centered square of side length 0.35 times the image width and height with RGB $(128, 128, 128)$. The *scale control* downsizes the original image by 0.6 and pastes it at the center of an equal-sized gray canvas. The archived alignment sweep uses eight pairs per factor, with no stochastic editing.

Table 7: Controlled factor construction and alignment with the signed presence axis.

Factor	n	Contrast	L15 cosine	L21 cosine
object evidence	8	present vs. absent natural pair	0.76	0.82
context prior	8	high- vs. low-margin absent image	0.53	0.51
occlusion	8	original vs. centered gray box	-0.62	-0.89
scale control	8	original vs. $0.6\times$ resized image	-0.02	0.28

A.5 Native factor layers and natural-error discrimination

A separate archived layerwise analysis places the strongest early factor signatures at different depths: contextual prior around L5, occlusion around L9, object evidence around L11, and scale around L18. These layer labels are descriptive summaries of the archived sweep and are not used as inferential thresholds in the main mediation test. The existence of different native layers supports heterogeneous upstream computation, but native scores are weak error detectors. FP versus CP discrimination from the tested native scores reaches AUROC 0.594, while FN versus CP ranges from approximately 0.40 to 0.62. By contrast, after convergence, a simple L18 combination $s_{\mathrm{prior}} - s_{\mathrm{obj}}$ reaches AUROC 0.892 on the same error-discrimination task. These results are used only to distinguish early factor

Table 8: Semantic specificity of causal transport. Same-semantic donors transfer across distinct visual paths, whereas geometrically similar but semantically mismatched donors do not.

Donor	Receiver	Relation	Repair
Presence, different path	Presence	same semantic	52/58/58%[†]
Color attribute	Presence	different semantic	0%
Shuffled	Presence	negative control	0–3%
Presence	Attribute	$\cos = 0.88$–0.93	0%
Matched attribute	Attribute	same semantic	26–39%

[†]L18/L21/L23; other ranges summarize the archived late-layer tests.

representation from post-convergence decision organization; no claim is made that the L18 linear combination is an optimal hallucination detector.

A.6 Pre- versus post-convergence interventions

The archived intervention test contrasts an early pre-convergence layer (L9) with a post-convergence layer (L18). Directions and doses are calibrated independently of target outcomes. For the prior-derived direction, the early intervention repairs $0/34$ targeted FPs; at L18, a one-standard-deviation intervention repairs 59% of targeted FPs and damages 5% of CP controls. For the occlusion-derived direction, the early intervention repairs $0/91$ targeted FNs; at L18, the corresponding intervention repairs 38% of targeted FNs and damages no CP controls at that dose. However, the same L18 direction strongly affects the opposite correct-negative polarity, pushing all tested CN controls toward FP. This broad effect motivates the interpretation in the main text: after convergence, a factor-derived direction acts as generic presence manipulation rather than a factor-selective inverse.

The main paper deliberately does not claim an exact single-layer causal onset from this two-regime comparison. A dense matched-norm layer sweep is the appropriate follow-up for estimating an onset interval.

A.7 Cross-path transport and semantic specificity

These archived controls are separate from the frozen-axis mediation experiment and are used only to test causal specificity. The natural-image transport pool contains 60 local boundaries (33 FN and 27 FP). Using raw donor displacements, same-semantic cross-path repair reaches 0.52, 0.58, and 0.58 at L18/L21/L23, compared with 0.70, 0.81, and 0.83 for same-path donors. A different-semantic color donor produces zero repair at all three layers, and shuffled donors give 0.00–0.03. Donor normalization destroys the transport effect because the archived displacement norms differ by orders of magnitude, so these tests preserve the original displacement scale.

For the cross-task control, presence and attribute directions are estimated under the same binary answer interface. Their late-layer cosine is 0.88–0.93, but presence-to-attribute repair is 0, whereas an attribute-matched donor repairs 0.26–0.39 of attribute receivers. The result therefore separates geometric similarity from causal interchangeability; it is not used to claim that semantic axes are globally orthogonal. These archived donor-transport tests use a different contrast from the same-image frozen mediation protocol in Appendix A.12; the cross-image variant is retained explicitly as a negative control in Appendix A.13.

A further archived control computes exact answer-margin gradients by autograd and projects their empirical span out of the FN and FP boundary displacements. The late FN/FP cosine remains approximately -0.98 to -0.99, and the signed cross-polarity transport effect survives the projection. This test addresses a different question from the per-example projection below: here the goal is to ask whether the semantic boundary geometry itself is reducible to the measured decision-gradient subspace, not how much direct local output control remains after removing one sample's gradient component.

A.8 Answer-surface remapping and readout geometry

Normal Yes/No, Present/Absent remapping, and option-order swap each achieve 100% semantic compliance on 400 examples. The signed semantic coordinate remains highly aligned between Yes/No and Present/Absent ($\cos = 0.973$), while the corresponding readout gradients have cosine 0.347. Under option-order swap, the readout gradient remains much more aligned with the normal condition ($\cos = 0.942$), consistent with a smaller change in the semantic answer mapping.

A separate fixed set of 600 Qwen3 clean examples (300 positive, 300 negative) is used to analyze the readout field at L21. Each raw Yes–No gradient is L2-normalized before SVD. The first three singular modes account for 70.2%, 25.4%, and 1.4% of squared singular-value energy. The dominant mean-gradient mode has cosine 0.129 with the signed presence axis. Its analytic local derivative is 26.3 and held-out forward derivative 29.4, compared with 3.2 and 3.25 for the presence axis.

A.9 Per-example readout projection

For each evaluation sample, we construct

$$\mathbf{u}^{\perp i} = \mathbf{u} - \frac{\mathbf{g}_i^\top \mathbf{u}}{\|\mathbf{g}_i\|^2}\mathbf{g}_i. \tag{15}$$

The original presence direction has measured local response $+3.12$, close to the analytic $+3.22$. The projected direction has response $+0.05$, close to zero. In the archived steering evaluation, the global omit/fabricate tradeoff changes by roughly nine percentage points along the original direction and only about 0.3 points after the per-example projection. The projection calculation was rechecked after correcting an earlier mean-centering implementation bug; the final per-example residual orthogonality is numerical ($\sim 8.6 \times 10^{-10}$ in the archived audit).

A.10 Additional downstream stress test: non-additive perturbation composition

These experiments are retained as auxiliary evidence for downstream readout stability; they are not part of the paper's central novelty claim. For each single factor we search a fixed severity ladder and retain a strength only when it reduces the clean correctness margin while leaving the answer correct. The first 400 clean-correct examples in fixed pool order are used without margin selection; 390 are valid in the stored Qwen3 run.

Table 9: Qwen3 subcritical-combination results. "Joint projection" uses the realized combined hidden state rather than a sum of single-factor effects.

Combination	n	flips	additive r	sum-proj. r	joint-proj. r	AUROC
blur+crop	214	79.9%	0.230	0.110	0.818	0.998
blur+rot	136	55.1%	0.134	0.051	0.901	0.989
crop+rot	109	57.8%	0.243	0.139	0.904	0.997
triple	107	77.6%	0.224	0.136	0.829	0.997

The corresponding joint-state MCC values are 0.941, 0.870, 0.962, and 0.864. Blur→crop and crop→blur disagree on $36/214 = 16.8\%$ of eligible examples. In InternVL2-8B, subcritical blur+crop flips $105/170 = 61.8\%$ of eligible cases; additive margin change has Pearson 0.173, while projection of the realized joint hidden state reaches Pearson 0.691 and AUROC 0.898.

A.11 Additional downstream stress test: complete perturbation trajectories

The trajectory experiment splits the 390 Qwen3 base examples with seed 2027 into 78 calibration and 312 held-out images, with no image overlap. The held-out grid contains 11 blur states, 10 crop states, and 9 states each for rotation, contrast, and occlusion per image, for 14,976 perturbation states in total. A local answer-gradient direction is computed once from each clean input and frozen across its trajectory.

Table 10: Clean-state readout explanation of complete held-out trajectories.

Factor	points	Pearson	Spearman	R^2
Blur	3432	0.982	0.991	0.965
Crop	3120	0.965	0.989	0.932
Rotation	2808	0.989	0.995	0.978
Contrast	2808	0.997	0.998	0.995
Occlusion	2808	0.980	0.998	0.960
Pooled	14976	0.972	0.995	0.945

Direction drift is substantial: $\cos(\Delta\mathbf{h}_j, \Delta\mathbf{h}_1)$ falls to roughly 0.21–0.38 by mid-trajectory and 0.09–0.21 at the end, while the frozen clean-state readout remains predictive of the observed margin. This is consistent with nonlinear upstream state evolution coexisting with a simpler downstream readout; we do not treat the first-order readout law as a primary contribution.

A.12 Semantic generality across binary variables

We repeat the frozen-axis mediation and answer-surface tests on three additional binary semantic variables in Qwen3-VL-2B: counting on CountBench (**?**), and attribute and relation judgments from AMBER (Wang et al., 2023). The scope remains binary decisions with an identifiable answer margin and zero crossing; we do not use these experiments to claim open-ended generation generality. For mediation, every target is a same-image, same-question clean-correct $\rightarrow$ perturbed-wrong pair. Perturbations follow the fixed blur ladder $4/8/12/20$ and crop ladder $2/3/4$, taking the first flip. For each domain, the semantic axis is estimated on a disjoint half of the questions and frozen before evaluation on the held-out half. Orthogonal controls are norm-matched to within 3×10^{-5}; the wrong-layer control applies the L9 axis at L18/L21. AMBER relation images are capped at 1.5 MP before the model processor; attribute images are used at their native 500×500 resolution.

Table 11: Frozen-axis mediation across semantic variables in Qwen3-VL-2B. The late signed geometry varies in strength, while mediation remains strong.

Variable	FN/FP cosine			n	R_u	
	L15	L18	L21	targets	L18	L21
Presence	–	–	−0.993	43	0.764	0.824
Counting	0.309	−0.464	−0.733	76	0.855	0.920
Attribute	0.355	−0.302	−0.571	96	0.634	0.773
Relation	0.847	0.003	−0.398	119	0.927	0.931

For protocol comparability, the presence row uses the matched Exp43B subset; the primary expanded-source estimate in Table 2 is $R_u = 0.831$ on 163 targets. At L18/L21, the corresponding norm-matched orthogonal recoveries are $-0.085/-0.130$ (presence), $0.126/0.043$ (counting), $0.084/0.006$ (attribute), and $0.014/-0.043$ (relation). Wrong-layer recovery is approximately zero where measured: $0.002/0.007$ for attribute and $-0.000/0.010$ for relation. At L21 the semantic component alone restores the correct sign for 100%, 95%, 85%, and 95% of presence, counting, attribute, and relation targets, respectively. The available transition pools contain 177/43 attribute FN/FP pairs, 197/23 relation FN/FP pairs, and 126/25 counting FN/FP pairs; we therefore interpret the weaker relation FN/FP cosine cautiously and do not require equally sharp global antiparallel geometry across domains.

Answer-surface invariance. For attribute, relation, and counting we evaluate 60 sample pairs under three semantically equivalent answer surfaces: Yes/No (YN), True/False (TF), and Correct/Incorrect (CI). The signed semantic transition direction is stable across surfaces at every measured late layer, whereas the answer-margin readout gradient changes substantially.

For counting, $\cos(g_{YN}, g_{CI}) = 0.180$. Presence uses the separately reported Yes/No versus Present/Absent remapping, with $\cos(u) = 0.973$ and $\cos(g) = 0.347$ (Sec. 3.4). Thus the semantic/readout dissociation is not tied to one token surface.

Table 12: Semantic direction versus readout under answer-surface remapping. Entries for u report L15/L18/L21 cosine; the final column is the L21 YN–TF gradient cosine.

Variable	$\cos(u_{YN}, u_{TF})$	$\cos(u_{YN}, u_{CI})$	$\cos(u_{TF}, u_{CI})$	$\cos(g_{YN}, g_{TF})$
Attribute	.986/.994/.966	.978/.987/.956	.987/.993/.978	.230
Relation	.991/.994/.964	.984/.990/.951	.993/.995/.979	.183
Counting	.971/.988/.960	.960/.979/.970	.983/.988/.965	.221

A.13 Why the mediation contrast must be image matched

The image-matched construction is a design requirement rather than a cosmetic pairing choice. If the same statistic is computed using the same question but a correct state from one image and a wrong state from a different image, the semantic component no longer separates from a norm-matched orthogonal control.

Table 13: Image matching is required to isolate semantic mediation.

Domain	construction	n	R_u L18/L21	$R_\perp$ L18/L21
Attribute	cross-image	50	.282/.349	.388/.493
Relation	cross-image	36	.581/.632	.348/.456
Attribute	image-matched	96	.634/.773	.084/.006
Relation	image-matched	119	.927/.931	.014/$-$.043

The cross-image variant therefore serves as a negative control: its loss of specificity is consistent with image-identity variation contaminating Δh, whereas same-image, same-question pairs more cleanly isolate the semantic transition targeted by the mediator. We treat image matching as part of the intervention protocol, not as a post-hoc filtering choice.

A.14 InternVL2-8B independent frozen-axis mediation replication

We repeat the frozen-axis protocol with an independently estimated InternVL2-8B signed object-presence axis. The source set contains 38 transitions (26 FN, 12 FP) spanning 31 source-image clusters; the disjoint target set contains 124 clean–error transitions (73 FN, 51 FP) spanning 101 target-image clusters. Source and target sets have zero image-signature and exact-question overlap. Five representative decoder layers give 620 target–layer patch rows, with 100 norm-matched random-orthogonal draws per target–layer pair (62,000 null rows).

Table 14: Independent frozen-axis mediation in InternVL2-8B. The frozen component becomes dominant at late layers while matched controls remain weak.

Layer	R_u	$R_{\perp,\mathrm{match}}$	R_{rand}	R_{wrong}	$\Delta m_u/\Delta m_{\mathrm{full}}$
L8	0.028	0.118	0.032	–	0.01/0.36
L14	0.105	0.146	0.018	0.061	0.06/0.57
L20	0.568	0.074	0.011	0.018	1.62/2.85
L26	0.632	-0.021	0.006	0.004	3.96/6.27
L31	0.681	-0.058	0.009	0.013	5.24/7.69

At L31, polarity-stratified recovery is 0.706 for FN targets and 0.645 for FP targets. Source-axis reconstruction is stable: source-image-cluster bootstrap median cosine to the full axis is 0.982/0.989/0.991 at L20/L26/L31, and random 15-of-38 source subsets give median cosine 0.976/0.985/0.989. Target-image-cluster bootstrap 95% intervals for R_u are $[0.508, 0.623]$, $[0.581, 0.681]$, and $[0.625, 0.729]$ at L20/L26/L31.

A.15 InternVL2-8B answer-surface remapping and readout projection

A held-out pool contains 360 examples, of which 354 (98.3%) preserve the semantic answer under Yes/No, Present/Absent, and option-order swap. The signed semantic axis remains strongly aligned

between Yes/No and Present/Absent ($\cos = 0.958$), while answer-margin gradients align only at 0.412. Under option-order swap, gradient alignment is much higher ($\cos = 0.906$).

On a separate balanced set of 480 clean examples, the first two singular modes of normalized L31 answer gradients explain 92.9% of squared singular-value energy. The dominant mean-gradient direction has cosine 0.176 with the semantic axis and exerts much stronger direct margin control: held-out forward derivative 24.6 versus 2.91 for the semantic direction. For each of 354 evaluation examples, removing the local-gradient component from u reduces measured direct response from $+2.68$ to $+0.08$; the corresponding global omit/fabricate steering tradeoff falls from 7.4 to 0.6 percentage points. The maximum residual $|g_i^\top u_i^\perp|$ in the numerical audit is 1.2×10^{-9}.

A.16 Natural-error transfer on unperturbed images

We evaluate the frozen InternVL2 presence axis on naturally occurring errors from original, unperturbed images. A unit step along $+u$ at L31 repairs 35/74 natural FNs (47.3%), compared with 4/74 (5.4%) under a norm-matched orthogonal control. Conversely, a unit step along $-u$ repairs 23/58 natural FPs (39.7%), compared with 2/58 (3.4%) for the control. On correct examples, the same global directions flip 11/120 correct negatives (9.2%) toward false positives and 8/120 correct positives (6.7%) toward false negatives.

Table 15: Transfer of the frozen InternVL2 presence axis to natural errors on unperturbed images.

Group	n	signed-axis flips/repairs	orthogonal control
Natural FN, $+u$	74	35 (47.3%)	4 (5.4%)
Natural FP, $-u$	58	23 (39.7%)	2 (3.4%)
Correct negative, $+u$	120	11 (9.2%)	2 (1.7%)
Correct positive, $-u$	120	8 (6.7%)	1 (0.8%)

A.17 Direction baseline comparison

All baseline directions are estimated on source/training data only, including orientation and scaling, and are frozen before target evaluation. The unsigned mean uses the un-signed average clean–error displacement; PCA-1 is the leading principal component of source transition deltas with orientation fixed on source data; the presence and correctness probe normals come from linear probes fit on the corresponding source labels; and the orthogonal control is exactly norm-matched to the signed-axis component. No baseline is selected or re-oriented using target recovery.

Table 16: Matched direction baselines. Normalized recovery of held-out clean–error effects under identical intervention norms.

Direction	Qwen3 L18	Qwen3 L21	InternVL2 L31
signed semantic u	**0.758**	**0.831**	**0.681**
unsigned mean Δh	0.421	0.462	0.389
PCA-1 transition direction	0.379	0.418	0.352
presence-probe normal	0.514	0.557	0.471
correctness-probe normal	0.286	0.304	0.259
norm-matched orthogonal	-0.080	-0.127	-0.058