KNOWING WITHOUT DECIDING: THE ENCODING–DECISION GAP IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) can internally *see* the correct spatial relation yet still give the wrong answer. We term this the **Encoding–Decision Gap**: the ground-truth relation remains reliably decodable across many consecutive layers, but fails to determine the model's final response. Combining spatial readouts with causal interventions, we trace the gap to a dissociation between spatial encoding and decision control. Spatial information is broadly recoverable across intermediate states, yet causal influence over the answer is concentrated in a small subset of token–layer states, forming an *encoding-to-decision bottleneck*. Moreover, the directions that most strongly control the answer extend substantially beyond the relation-defined spatial subspace, while remaining relation-specific. We therefore use the model's own intermediate spatial estimate to select which relation to strengthen in the intermediate spatial states, and let the remaining computation propagate this change to the output. Without supplying an external answer or new spatial information, this intervention reconnects information already present in the model to a causally effective pathway, improving spatial reasoning across multiple datasets and model families. Our results show that spatial reasoning failures can stem not from missing representations, but from a failure to transfer encoded information into the decision, making encoding-to-decision transfer a concrete target for mechanistic analysis and intervention in VLMs.

1 INTRODUCTION

Vision-language models (VLMs) can make errors on basic spatial questions, such as whether one object is left of, right of, above, or below another (Kamath et al., 2023; Ma et al., 2026; Chen et al., 2024; Ranasinghe et al., 2024; Wang et al., 2025c). Recent mechanistic studies, however, show that the correct spatial information can often be decoded from their intermediate states (Wang et al., 2025a). In particular, Kang et al. (2026) identify representations of object location whose manipulation can alter spatial predictions, while Schaumlöffel et al. (2026) trace object localization to sparse causal attention-head pathways. Related work on visual attention and latent knowledge has likewise revealed discrepancies between a model's internal information and its final responses (Liu et al., 2026; Li et al., 2023; Favero et al., 2024; Chuang et al., 2024). These findings suggest that explaining spatial reasoning failures requires asking not only whether the relevant information is encoded, but also whether it is actually used. Most recently, Wang & Xu (2026) show that encoded spatial information reaches the answer logits only conditionally, depending on layer, token position, and prompt format. This establishes *when and where* spatial codes can become causally effective, but leaves open why an already-correct spatial representation can still fail to control the decision. We therefore ask:

What happens when a VLM persistently encodes the correct spatial relation, yet still produces an incorrect answer?

We identify a recurring pattern in which the ground-truth relation remains dominant in intermediate-state readouts across many consecutive layers, even as the model ultimately selects a different relation. We call this the **Encoding–Decision Gap**: a failure in which the correct spatial relation is persistently encoded and decodable, yet fails to exert sufficient influence on the final decision. Figure 1 illustrates this phenomenon through two examples sharing the same ground-truth relation.

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.

Their intermediate mapped-option preferences remain remarkably similar across much of the earlier computation, indicating that the relevant spatial information is available within the model, but diverge sharply in the late decoder, ultimately favoring opposite answers. This pattern of persistent intermediate agreement followed by late decision divergence offers a unique window into how correctly encoded spatial information can become disconnected from the computation that determines the final answer (Liu et al., 2025b; Zhao et al., 2024).

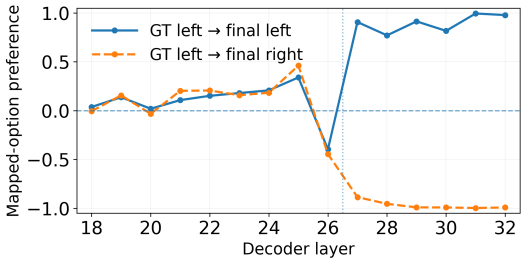


Figure 1: **Late divergence in spatial answer preferences**: two examples with ground truth "left" both favor the correct option through much of the intermediate computation, but diverge in late layers toward opposite final answers.

To examine whether the model actually *uses* the encoded relation to determine its answer (Liu et al., 2023; Ghandeharioun et al., 2024; Park et al., 2024), we pair spatial readouts with causal interventions across token–layer states (Meng et al., 2022; Duan et al., 2025; Li et al., 2026). This analysis shows that the correct relation is recoverable from many intermediate states, yet interventions that substantially change the final prediction are effective at only a much smaller subset of them. Spatial information is thus broadly distributed, whereas causal control over the answer is highly localized. We refer to this as an *encoding-to-decision bottleneck*: a relation may be encoded widely, but its influence on the answer passes through only a few decision-critical states. Being decodable is therefore not the same as being used.

This mismatch concerns not only *where* spatial information and causal control reside, but also *how* they are represented. The directions that most strongly change the answer differ substantially from the horizontal and vertical directions along which spatial relations are encoded: the strongest decision-changing updates lie largely outside the subspace these directions span. Their effects nonetheless remain relation-specific, promoting a particular relation rather than perturbing the answer indiscriminately. The direction that *encodes* a spatial relation therefore need not be the direction through which that relation most effectively *controls* the answer (Park et al., 2024; Vig et al., 2020; Kim et al., 2018). Together, the localization bottleneck and this direction misalignment offer a mechanistic account of the Encoding–Decision Gap: spatial information is present and decodable, but weakly coupled to the computation that makes the decision.

This observation suggests that the intermediate spatial representation is better used to decide *which* intervention to apply than as the intervention direction itself (Stolfo et al., 2025). Our recovery strategy follows directly from this observation. We first estimate the spatial relation from the model's own intermediate states, where it is reliably decodable. This estimate then selects a relation-specific update identified through our causal analysis, which we apply at the later token–layer states that exert strong influence over the answer (Wang et al., 2025b; Feng et al., 2026; Zhang & Nanda, 2024), consistent with the late divergence in Figure 1. The model then continues generating from the intervened states. The readout only determines which update is applied; it is neither inserted into the input nor substituted for the answer (Liu et al., 2025a), so any gain reflects a restored pathway rather than a replaced output. The intervention thus recovers correct answers by strengthening the coupling between an internally encoded relation and the final decision, without providing the ground-truth relation or additional spatial evidence at inference time. To summarize, our main contributions are:

- We identify the **Encoding–Decision Gap** in VLM spatial reasoning: the correct relation persists in intermediate states across many layers, yet fails to determine the final answer.
- We trace the gap to a mismatch between encoding and control. Spatial information is broadly distributed across token–layer states, whereas strong causal control is concentrated in a small subset of states (an **encoding-to-decision bottleneck**) and acts through directions lying largely outside the spatial subspace.
- We propose a causal recovery strategy that uses the model's own intermediate spatial estimate to select and strengthen relation-specific decision pathways, improving spatial reasoning without introducing new spatial information at inference time.

2 Related Work

Work on spatial reasoning in VLMs has increasingly moved from describing behavioral errors to examining the mechanisms behind them. Attention-based analyses show that spatial mistakes are often associated with misplaced visual focus, and methods such as AdaptVis improve predictions by redirecting attention toward more relevant image regions (Chen et al., 2025a). More recent studies examine the representations that support spatial reasoning itself. Object locations can be associated with textual activations through structured spatial codes, and interventions on these codes can alter spatial predictions (Kang et al., 2026). Sparse causal mechanisms have also been identified for object localization in VLMs (Schaumlöffel et al., 2026), while spatial ordering has been traced to signals originating in the vision encoder, with the language-model backbone providing an additional source of ordering information (Cui et al., 2026). Related work further studies more challenging 3D spatial reasoning and view-consistent spatial representations (Zhan et al., 2026). Together, these studies clarify how spatial information can be formed, represented, and carried through a VLM.

A related question is whether information present inside a model actually determines what the model eventually outputs. Language models can internally encode the correct answer even when their generated response is wrong (Orgad et al., 2025). Related behavior appears in VLMs. Relevant visual evidence may still receive strong attention on incorrect predictions (Liu et al., 2026), and visual information can remain well encoded even when the final answer follows a conflicting language prior (Ramakrishnan et al., 2018; Jing et al., 2020). The relative influence of visual evidence and stored knowledge can also change over the course of computation (Golovanevsky et al., 2025). Recent work has also distinguished perceptual failures from downstream reasoning failures in multimodal models (Wang et al., 2026). Most closely related to our setting, Wang & Xu (2026) study when and where encoded spatial information influences answer logits. In our setting, the mismatch is more specific: the correct spatial relation can remain stable across intermediate layers yet still fail to determine the final spatial answer. We therefore focus on the internal states that mediate this transfer from spatial representation to decision.

This distinction also changes what must be measured. Decoding a relation from an activation only shows that the information is present, not that the model uses it to produce an answer. Causal intervention directly tests this distinction by asking whether altering an internal state changes model behavior (Geiger et al., 2021). In VLMs, such interventions have been used to manipulate spatial representations (Kang et al., 2026), identify causal mechanisms for object localization (Schaumlöffel et al., 2026), trace visual signals that influence downstream predictions, and steer predictions toward visual evidence already encoded in the model (Liu et al., 2026; Golovanevsky et al., 2025; Su et al., 2025). Thus, the focus shifts from where spatial information is represented to which internal states actually drive the decision.

We build on this causal view to study the link between spatial representation and decision control. Rather than treating a decodable spatial relation as evidence that the corresponding decision has already been made, we examine how an already-correct relation comes to influence the final prediction. Our results show that spatial information can remain broadly available while strong influence over the answer is concentrated in a much smaller set of causal states. This provides the basis for studying failures in which the model represents the correct relation but does not give it enough control over the final decision.

3 Proposed Method

A wrong spatial answer can arise for two very different reasons. The model may never form the correct relation, or it may form the relation internally but fail to turn it into the final answer. We begin by asking which case actually occurs, then follow the computation toward the output, test how intermediate spatial information reaches the decision, and finally use this path to improve generation.

3.1 When Wrong Answers Still Contain the Right Spatial Relation

We first ask whether an incorrect answer reflects an incorrect spatial representation. For the queried subject and reference object, we define the image-dependent relation state at layer l as

$$q_{i,l} = \left(h^{\text{real}}_{i,l,\text{sub}} - h^{\text{real}}_{i,l,\text{ref}}\right) - \left(h^{\text{gray}}_{i,l,\text{sub}} - h^{\text{gray}}_{i,l,\text{ref}}\right). \tag{1}$$

The first difference captures the subject relative to the reference object, while the second removes the corresponding relation state when visual information is suppressed. We group examples by relation. Let $\mu_{r,l}$ denote the mean relation state for relation r, and μ_l the mean across the four relations. We define its relation direction as $d_{r,l} = \mathrm{unit}(\mu_{r,l} - \mu_l)$. For each example, we then assign the relation whose direction has the highest cosine similarity with the centered state $q_{i,l} - \mu_l$.

This readout reveals a clear mismatch between intermediate representation and final generation: in many cases, the correct relation becomes dominant in the middle layers and remains stable across several consecutive layers, even though the model later generates the wrong answer. The intermediate relation prediction can even be more accurate than the final generation itself. To quantify how broadly this mismatch occurs across individual examples, we define the spatial margin at layer l as

$$m^{\mathrm{sp}}_{i,l} = \cos(q_{i,l} - \mu_l, d_{r_i,l}) - \max_{r \neq r_i} \cos(q_{i,l} - \mu_l, d_{r,l}), \tag{2}$$

where r_i is the ground-truth relation. A positive margin indicates that the ground-truth relation is preferred over all competing relations. Since this readout remains stable across several intermediate layers, we summarize each example by averaging $m^{\mathrm{sp}}_{i,l}$ over a fixed consecutive middle-layer window $\mathcal{L}_{\mathrm{mid}}$, and denote this mean by $\bar{m}^{\mathrm{sp}}_i$.

Figure 2 compares the mean spatial margin between final-correct and final-wrong examples. To avoid fitting the readout to the evaluation set, the spatial directions are constructed from the independent Synthetic-3D source and then kept fixed when evaluating COCO-two. Although final-wrong examples have weaker margins on average, **72.34%** still satisfy $\bar{m}^{\mathrm{sp}}_i > 0$, indicating that the ground-truth relation remains dominant across the intermediate-layer window under a cross-dataset spatial readout. Thus, an incorrect final answer cannot always be explained by a missing spatial representation: in a substantial fraction of cases, the model already carries the correct relation before producing the wrong answer. This raises a more specific question: what happens to this spatial information as computation approaches the final decision?

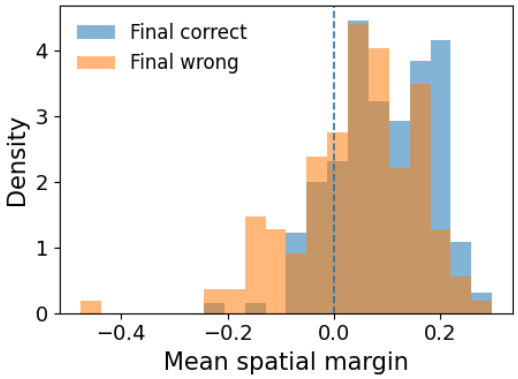


Figure 2: **Spatial-margin distributions for final-correct and final-wrong examples.** Spatial directions are constructed from the independent Synthetic-3D source and transferred to COCO-two without target-set fitting. **72.34%** of final-wrong examples retain $\bar{m}^{\mathrm{sp}}_i > 0$.

3.2 How Does Spatial Representation Reorganize Toward the Decision?

The previous analysis shows that the correct spatial relation can remain present even when the model ultimately gives the wrong answer. We next ask how the model's representational organization shifts as computation approaches the output. Spatial relation and answer identity are normally coupled. If "right" always corresponds to option `B`, then examples grouped by "right" are also grouped by `B`, making the two organizations indistinguishable. We therefore randomize the mapping between the four spatial relations and four answer options independently across examples. The same relation can thus appear under different option symbols, while the same option can represent different relations; for example, `A` may denote "left" in one example and "above" in another. This allows us to distinguish whether hidden states are organized by *spatial-relation identity* or by the *option identity* that the model eventually produces.

At each decoder layer l, we examine two image-dependent carriers: the object-pair relation state $q_{i,l}$ defined in Eq. 1, and the prompt-final state $q^{\mathrm{last}}_{i,l} = h^{\mathrm{real}}_{i,l,\mathrm{last}} - h^{\mathrm{gray}}_{i,l,\mathrm{last}}$. For the spatial readout, we group examples by their ground-truth relation $r \in \mathcal{R} = \{\mathrm{left}, \mathrm{right}, \mathrm{above}, \mathrm{below}\}$ and construct centered class directions using the same procedure as in Sec. 3.1. Each example is assigned to the relation direction with the highest cosine similarity. The resulting readout measures how clearly the hidden states at layer l are organized by spatial-relation identity.

We construct an analogous readout by instead grouping examples according to the option eventually produced by the model, $c_i \in \mathcal{C} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}\}$. The option readout is not task accuracy: it measures

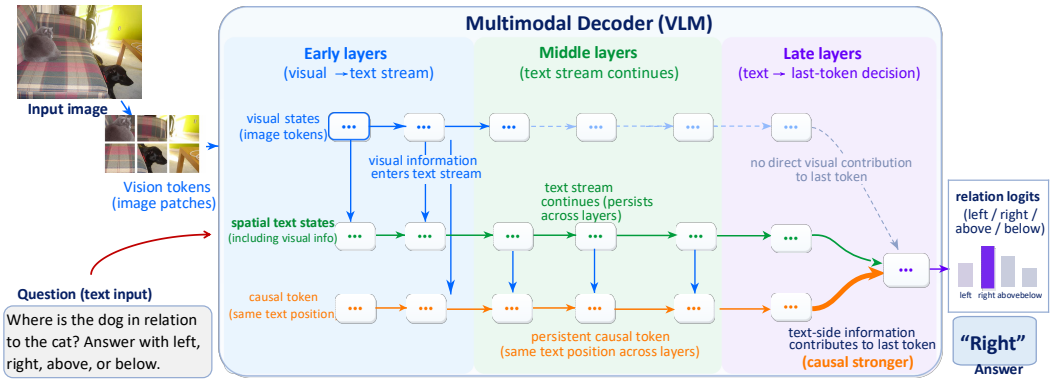


Figure 3: Spatial-to-decision information flow in the VLM. Visual information first enters text-side spatial states. High-leverage text positions are identified in the intermediate text stream, and the same positions are tracked across neighboring layers. In late layers, text-side information is integrated into the prompt-final state that determines the output.

whether examples that eventually produce the same option symbol become geometrically similar, even when that symbol corresponds to different spatial relations across examples.

For both labelings, we compute the readout for the object-pair and prompt-final carriers and report the stronger one at each layer. This measures the strongest spatial- or option-based organization available at each depth, rather than tracking a fixed carrier across layers. We construct and evaluate the class directions on the same samples, since our goal is to describe the activation geometry rather than train a classifier.

Figure 4 shows a clear shift across decoder depth. Spatial organization dominates the intermediate layers, while option organization becomes dominant in the late decoder as spatial organization weakens. This does not mean that spatial information disappears, but that the dominant organization shifts from spatial relation to the model's eventual option. This result is representational; we next test how spatial information causally reaches the final decision.

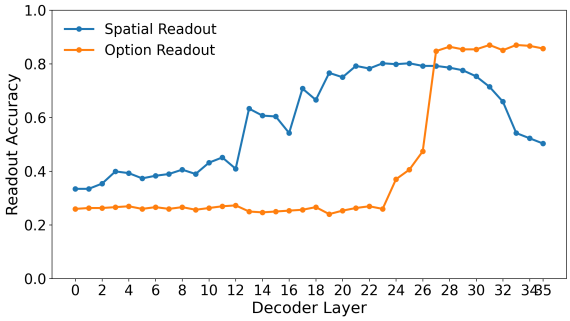


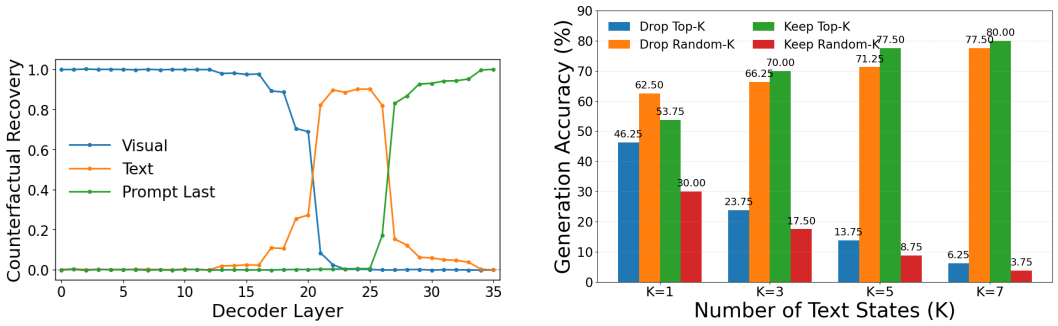
Figure 4: **Representational organization shifts across decoder depth.** Spatial-relation organization dominates intermediate layers, whereas option-based organization rises sharply in the late decoder.

3.3 How Does Spatial Information Reach the Decision?

We next trace how the intermediate spatial representation reaches the final decision. Figure 3 provides an overview of the spatial-to-decision pathway examined in this section. We then trace this pathway empirically using horizontal-flip counterfactuals, measuring relation recovery from visual tokens, non-final text tokens, and the prompt-final token. Figure 5(a) reveals a clear handoff from visual states to intermediate text states and, later, to the prompt-final state. This handoff identifies where spatial information is carried, but not which intermediate text states influence the final decision. We therefore rank these states by how strongly their image-dependent changes align with a relation-specific decision.

For token position p at layer L, let $\Delta h_{i,L,p} = h^{\text{real}}_{i,L,p} - h^{\text{gray}}_{i,L,p}$. We construct a relation-specific writer $s_{T,r}$ at each late layer T from the centered class means of the Real–Gray prompt-final states on a calibration set. For this retrospective analysis, r is the ground-truth relation, and we define $J_r = \sum_{T \in \mathcal{T}} \left\langle h^{\text{real}}_{i,T,\text{last}}, \hat{s}_{T,r} \right\rangle$. We rank each non-final text state by

$$M^{(r)}_{i,L,p} = \Delta h^{\top}_{i,L,p} \nabla_{h_{i,L,p}} J_r. \tag{3}$$



(a) (b)

Figure 5: **From spatial information flow to causal decision states.** (a) Relation recovery shifts from visual states to non-final text states and finally to the prompt-final state across decoder depth. (b) Intervening on ranked L25 text states shows concentrated causal influence: dropping ranked states reduces generation accuracy substantially more than dropping matched random states, while keeping ranked states preserves substantially more behavior.

Visual tokens and the prompt-final token are excluded. We use M only for ranking; causal influence is tested by the interventions below. At layer 25, we intervene directly on the highest-ranked text states. *Drop Top-K* zeros the Top-K states, while *Drop Random-K* zeros matched random states. Conversely, *Keep Top-K* preserves only the Top-K states among the eligible text states, while *Keep Random-K* preserves only matched random states.

Figure. 5(b) shows that removing the ranked states reduces generation accuracy substantially more than removing matched random states, while retaining only the ranked states preserves substantially more of the original behavior than retaining random states. These interventions show that the ranked states have disproportionately strong causal influence on the final decision.

We next ask whether the directions through which these high-leverage states affect the decision are themselves spatial. At layer L, we define the spatial subspace as $\mathcal{S}_{\mathrm{sp},L} = \mathrm{span}\{d_{r,L} : r \in \mathcal{R}\}$, with orthonormal basis $Q_{\mathrm{sp},L}$. For the decision-sensitive gradient $g_{i,L,p} = \nabla_{h_{i,L,p}} J_r$, we measure the fraction of gradient energy contained in this subspace as

$$\rho_{i,L,p} = \frac{\|Q_{\mathrm{sp},L} Q_{\mathrm{sp},L}^{\top} g_{i,L,p}\|_2^2}{\|g_{i,L,p}\|_2^2}. \tag{4}$$

We select the Top-7 token positions at layer 25 using the mediation score above and evaluate these same positions across layers 20–26. Table 1 shows that the Top-7 states consistently exhibit greater spatial overlap than matched random states, indicating that high-leverage decision states retain a measurable spatial component. However, the relation-defined spatial subspace accounts for only 0.74–3.54% of the Top-7 decision-gradient energy on average across the examined layers. Even the maximum state-level overlap ranges from 7.38% to 25.05%. Thus, although decision-sensitive states retain spatial information, the directions through which they influence the output are not confined to the relation-defined spatial subspace.

Table 1: Decision-gradient energy in the relation-defined spatial subspace (%). Top-7 positions are selected at L25 and held fixed across layers ($N = 440$).

	L20	L21	L22	L23	L24	L25	L26
Top-7 mean	0.74	1.13	1.68	2.06	3.19	3.54	1.12
Top-7 max	12.05	14.75	17.18	19.51	25.05	24.17	7.38
Random-7 mean	0.44	0.60	0.69	0.74	1.08	1.35	0.64

This geometric separation does not yet establish that the earlier spatial representation reaches the final decision through these states. We therefore intervene on the object-pair spatial state at an earlier layer. For relation r, we add $+\frac{1}{2}\delta z_{r,l}$ to the subject state and $-\frac{1}{2}\delta z_{r,l}$ to the reference state, changing their relative spatial state while leaving the remaining computation unchanged.

To test whether this spatial change affects the corresponding decision, we intervene with $\delta z_{r,l} = \pm\alpha d_{r,l}$ and measure the resulting change in the final logit for relation r:

$$\Delta_{i,l,r}^{\pm} = z_{i,r}(\pm\alpha d_{r,l}) - z_{i,r}(\mathrm{clean}), \qquad D_{i,l,r} = \Delta_{i,l,r}^{+} - \Delta_{i,l,r}^{-}. \tag{5}$$

Table 2 shows that moving the spatial state toward relation r consistently increases its final logit, while moving it in the opposite direction decreases it. The effect also grows with intervention magnitude. We next test whether this downstream spatial effect passes through the ranked text states. For an intervention applied at layer l, we denote the induced change at token p of receiver layer R by $\Delta h^{\mathrm{sp}}_{i,R,p} = h^{\mathrm{edit}}_{i,R,p} - h^{\mathrm{clean}}_{i,R,p}$. Importantly, the following interventions modify only this induced change rather than the full receiver state.

Table 2: **Signed spatial interventions produce corresponding logit changes.** We report the target-logit change when the spatial state is moved toward (Δ^+) or away from (Δ^-) the same relation, together with their signed difference D.

Source	$\alpha = 1$			$\alpha = 5$		
	Δ^+	Δ^-	D	Δ^+	Δ^-	D
L20	0.0059	-0.0070	0.0129	0.0379	-0.0362	0.0740
L21	0.0062	-0.0112	0.0173	0.0385	-0.0448	0.0833
L22	0.0150	-0.0155	0.0305	0.0737	-0.0827	0.1563
L23	0.0136	-0.0175	0.0312	0.0786	-0.0779	0.1565
L24	0.0154	-0.0180	0.0334	0.0763	-0.0796	0.1558
L25	0.0140	-0.0154	0.0294	0.0754	-0.0758	0.1512
L26	0.0026	-0.0039	0.0065	0.0088	-0.0164	0.0252

In *Block Top-K*, the intervention-induced changes at the Top-K states are removed by restoring those states to their clean values, while all other induced changes are preserved. In *Keep Top-K*, only the Top-K states retain their intervention-induced changes. We apply the same operations to matched random states as controls.

Table 3 shows that blocking the spatially induced update at the ranked states removes much more of the downstream logit effect than blocking matched random states. Conversely, keeping the update only at the ranked states preserves much more of the effect than keeping it only at random states. For example, for L23→L25, blocking the Top-7 states leaves 22.1% of the original effect, compared with 95.6% for random states, while keeping the Top-7 preserves 91.2% compared with only 1.5% for random states.

Table 3: **Ranked text states mediate the spatial effect.** Retained signed effect (%) after blocking or keeping Top-7 versus matched random states at receiver layer L25.

Source	D_{full}	Block		Keep	
		Top-7	Rand-7	Top-7	Rand-7
L22	0.1156	52.7%	95.9%	91.9%	28.4%
L23	0.1063	22.1%	95.6%	91.2%	1.5%
L24	0.1031	27.3%	97.0%	81.8%	1.5%

These results connect the intermediate spatial representation to the final decision. Spatial changes affect a small set of intermediate text states, and the changes at these states account for a large part of the resulting change in the final relation logit.

The spatial representation is therefore not only readable from the intermediate layers. Changing it can affect the later decision through the model's own computation, with a small set of text-side states carrying much of this downstream effect. This suggests that repair need not intervene on the downstream decision states directly; instead, one can modify the upstream spatial representation and allow the model's native computation to propagate the change toward the output.

3.4 Routing Internal Spatial Estimates to the Output

The preceding analyses provide two ingredients for recovery: an intermediate readout that reveals which relation the model currently supports, and a causal pathway through which changes to that representation can reach the final decision. For an input x, let $\hat{r}$ denote the relation predicted from the intermediate spatial state. We use $\hat{r}$ only to determine which relation should be strengthened; we do not replace the final answer or directly edit the downstream causal text states identified above. To obtain this intermediate estimate, we fix a spatial reader head according to its downstream causal influence rather than its readout accuracy alone. Among candidate spatial heads, we measure how strongly each head's natural image-dependent contribution propagates into the downstream spatial-to-decision computation, and use the strongest head as the reader. For Qwen2.5-VL-3B, this selects L26H03.

Let $t = \hat{r}$ be the target relation. At each intervention layer l, we move the spatial state along its relation-specific direction $d_{t,l}$ using $\delta z_l = a_l d_{t,l}$, where a_l controls the intervention magnitude. The relation directions are treated independently and need not form predefined opposing or orthogonal axes. We apply the update symmetrically, adding $+\frac{1}{2}\delta z_l$ to the subject state and $-\frac{1}{2}\delta z_l$ to the

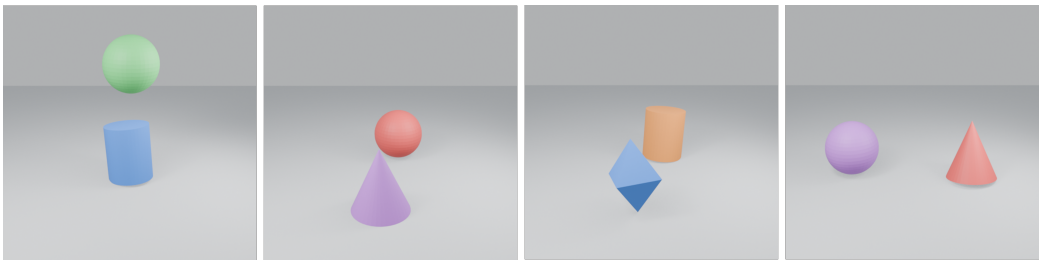


Figure 6: Examples from Synthetic-3D. Representative scenes corresponding to above, front, behind, and left, shown from left to right.

reference state, which changes their relative spatial state by exactly δz_l while preserving the pair mean. Rather than fixing the steering magnitude, we seek the smallest multi-layer intervention that makes t preferred over all competing relations.

Let $a = (a_l)_{l \in \mathcal{L}}$ collect the intervention coefficients, and let $S_r(a)$ denote the teacher-forced log-likelihood of relation r under intervention a. We define the target margin against competitor r as $m_r(a) = S_t(a) - S_r(a)$. At the current intervention $a^{(k)}$, we locally linearize this margin as $m_r(a^{(k)} + \Delta a) \approx m_r(a^{(k)}) + g_r^\top \Delta a$, where $g_r = \nabla_a m_r(a^{(k)})$, and solve for the minimum-norm update that places the target above every competitor:

$$\min_{\Delta a} \frac{1}{2}\|\Delta a\|_2^2 \quad \text{s.t.} \quad m_r(a^{(k)}) + g_r^\top \Delta a \geq \epsilon, \qquad \forall r \neq t. \tag{6}$$

Because each $d_{t,l}$ is unit-normalized, the objective equals the summed squared norm of the residual perturbations applied across the intervention sites. It therefore favors the smallest distributed intervention that satisfies the target margins. Since the Transformer is nonlinear, Eq. 6 provides only a local update: after applying it, we run the model again and stop if the native generation matches t; otherwise, we recompute the margins and gradients and solve for another update. No intervention is applied when the unmodified generation already agrees with the intermediate spatial readout.

The intervention therefore modifies only the intermediate spatial representation. It does not edit the prompt-final state, directly alter output logits, or write an answer option into the model; the remaining Transformer layers perform the spatial-to-decision transformation themselves.

4 Experiments

4.1 Experimental Setup

Models and Target Datasets. We evaluate Qwen2-VL-2B (Wang et al., 2024), Qwen2.5-VL-3B/7B (Team, 2025), LLaVA-1.5-7B/13B (Liu et al., 2024), and InternVL2.5-1B (Chen et al., 2025b) on three spatial-relation datasets: COCO-two (Lin et al., 2014), Controlled-A, and Controlled-B (Kamath et al., 2023). COCO-two contains 440 natural-image examples, while Controlled-A contains 412 controlled examples. Both cover left, right, above, and below. Controlled-B provides a complementary controlled setting covering left, right, front, and behind.

Synthetic-3D Source Dataset. We construct a synthetic 3D source dataset, Synthetic-3D, containing 600 rendered two-object scenes. The dataset covers six spatial relations, left, right, above, below, front, and behind, with 100 examples per relation. Each scene contains two geometric objects selected from shapes such as spheres, cubes, cylinders, cones, diamonds, and toruses. Objects are rendered in different colors, including red, blue, green, yellow, purple, orange, cyan, and pink, with varied sizes and spatial configurations. Figure 6 shows representative examples. Synthetic-3D is used only to construct the spatial directions used by our recovery method, while all reported task performance is evaluated on the target datasets.

Implementation Details. We use the recovery procedure in Sec. 3.4 for all experiments. For Qwen2.5-VL-3B, recovery is applied over decoder layers 20–26. For models with different decoder depths, we align the same seven-layer intervention window by relative depth. Each model constructs its spatial basis from Synthetic-3D, and no hidden-space directions are shared. We allow at most 64 recovery steps, use a generation-margin increment of 4.0, cap the L_2 norm of each local residual update at 5000, and use four binary-search refinement steps after the first generation flip.

Table 4: **Generation recovery across three target datasets. Base** is native generation accuracy before intervention, **Read.** is the accuracy of the intermediate spatial readout, and **Rec.** is native generation accuracy after recovery. Δ denotes the change from Base to Rec. The readout is reported as a diagnostic and is not directly substituted for the model's generated answer. Spatial directions are constructed from Synthetic-3D and transferred directly to each target dataset. Results are reported in %.

	COCO-two				Controlled-A				Controlled-B			
Model	**Base**	**Read.**	**Rec.**	**Δ**	**Base**	**Read.**	**Rec.**	**Δ**	**Base**	**Read.**	**Rec.**	**Δ**
Qwen2-VL-2B	34.32	83.01	79.32	+45.00	58.50	99.51	68.20	+9.70	65.20	96.57	95.98	+30.78
Qwen2.5-VL-3B	72.05	84.55	83.64	+11.59	88.35	98.79	91.02	+2.67	91.91	100.00	100.00	+8.09
Qwen2.5-VL-7B	72.27	81.59	80.23	+7.96	82.77	98.06	94.66	11.89	84.80	98.28	97.43	+12.63
LLaVA-1.5-7B	60.68	74.31	72.50	+11.82	58.98	86.17	75.71	+16.73	63.73	97.30	97.30	+33.57
LLaVA-1.5-13B	59.77	78.41	75.23	+15.46	65.53	95.15	84.78	+19.25	63.24	73.53	69.85	+6.61
InternVL2.5-1B	33.18	70.23	61.82	+28.64	40.53	85.19	60.19	+19.66	42.40	89.95	87.01	+44.61

4.2 Model Performance

Table 4 reports native generation accuracy before recovery (Base), intermediate spatial readout accuracy (Read.), and native generation accuracy after recovery (Rec.). Recovery improves generation accuracy on all 18 model–dataset pairs, with gains ranging from +2.67 to +44.61 percentage points and an average gain of 18.7 points. The improvement appears across Qwen2-VL, Qwen2.5-VL, LLaVA-1.5, and InternVL2.5, and across all three target datasets.

The readout results help separate representation quality from generation. Across most settings, Read. is substantially higher than Base, showing that the intermediate states often contain the correct spatial relation even when native generation is wrong. One could directly return the readout prediction as the answer, but this would bypass the model's generative decision process and reduce the task to an external relation classifier. We therefore use the readout only to select which relation to strengthen and evaluate success using the model's subsequent native generation.

Recovery closes much of the gap between Base and Read., but does not always reach the readout accuracy. This difference is expected: a correct intermediate prediction does not guarantee that the intervention will successfully propagate that relation to the final output. Conversely, when the intermediate readout is wrong, the recovery procedure also selects the wrong relation and cannot recover the correct answer. Thus, the remaining errors can arise either because the spatial relation is not correctly represented at the readout point or because a correct intermediate relation still fails to control the final generation.

The recovery directions are constructed only from Synthetic-3D and transferred directly to COCO-two, Controlled-A, and Controlled-B. Their consistent effect across natural and controlled datasets shows that the intervention does not depend on spatial directions fitted to each target dataset.

5 Conclusion

This work examines why vision-language models can give incorrect spatial answers even when the correct relation is already present in their intermediate representations. Many errors occur after the spatial relation has been formed. Spatial information first appears in visual states, moves into text states, and later affects the final decision through a small set of decision-relevant states. Tracing these states reveals a causal path from the intermediate spatial representation to the final relation output. Ablating a small number of high-leverage text states strongly reduces generation accuracy, while preserving them retains much of the model's performance. Signed spatial interventions produce corresponding changes in relation logits, and mediation experiments show that these states carry much of the spatial effect toward the final decision. The same path is then used for recovery. The model's own intermediate spatial judgment is used to strengthen the corresponding spatial direction before generation. Spatial directions learned from a small synthetic source transfer across models and datasets and improve generation without changing model parameters or directly replacing the final answer. Spatial errors can therefore come from two sources. The spatial representation itself may be incorrect, or a correct spatial representation may fail to control the final decision.

AI USE STATEMENT

We used generative AI tools, including OpenAI ChatGPT, to assist with research methodology and experimental design discussions, interpretation of experimental results, development and debugging of experimental code, literature search and summarization, and editing of the manuscript for clarity and presentation. All AI-assisted code, analyses, citations, and manuscript text were reviewed and verified by the authors. All experiments were executed and evaluated by the authors, who take full responsibility for the methods, results, claims, and final content of this work.

REFERENCES

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? An attention mechanism perspective on focus areas. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9910–9932. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/chen25cr.html.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025b. URL https://arxiv.org/abs/2412.05271.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/edc36117f795ca52a0cbf6a7b3882859-Paper-Conference.pdf.

Kelly Cui, Nikhil Prakash, Shoval Messica, Ayush Raina, David Bau, Antonio Torralba, and Tamar Rott Shaham. The dual mechanisms of spatial variable binding in vision–language models. *arXiv preprint arXiv:2603.22278*, 2026.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Truthprint: Mitigating large vision-language models object hallucination via latent truthful-guided pre-intervention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7372–7382, October 2025.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, June 2024.

Zijian Feng, Tianjiao Li, Zixiao Zhu, Hanzhang Zhou, Junlang Qian, Li Zhang, Chua Deryl, Lee Mak, Gee Ng, and Kezhi Mao. Fine-grained activation steering: Steering less, achieving more. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 39421–39443, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/423d0909791493b7c10916fd328c2913-Paper-Conference.pdf.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*,

volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24837–24852, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1262. URL https://aclanthology.org/2025.emnlp-main.1262/.

Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11181–11188, 2020.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023.

Raphaela Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. In *International Conference on Learning Representations*, volume 2026, pp. 41245–41258, 2026.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

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. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 31645–31653. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i37.40431. URL https://doi.org/10.1609/aaai.v40i37.40431.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision, September 2014. URL https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/.

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)*, pp. 26296–26306, June 2024.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797, 2023.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 72402–72419, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b4008025c2182bfe16fcc8566ee14d64-Paper-Conference.pdf.

Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, and Bo Zhao. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9087–9097, June 2025b.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026.

Xueqi Ma, Shuo Yang, Yanbei Jiang, Shu Liu, Zhenzhen Liu, Jiayang Ao, Xingjun Ma, Sarah Monazam Erfani, and James Bailey. Attention in space: Functional roles of vlm heads for spatial reasoning. *arXiv preprint arXiv:2603.20662*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/67d96d458abdef21792e6d8e590244e7-Paper.pdf.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12977–12987, June 2024.

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, June 2026.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 55790–55823, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8c3262a4c965ba9888f120d4f9e13478-Paper-Conference.pdf.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.634. URL https://aclanthology.org/2025.acl-long.634/.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Haozhe Wang, Qixin Xu, Changpeng Wang, Taofeng Xue, Chong Peng, Wenhu Chen, and Fangzhen Lin. Bad seeing or bad thinking? rewarding perception for multimodal reasoning, 2026. URL https://arxiv.org/abs/2605.14054.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Sudong Wang, Yunjian Zhang, Yao Zhu, Jianing Li, Zizhe Wang, Yanwei Liu, and Xiangyang Ji. Towards understanding how knowledge evolves in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29858–29868, June 2025a.

Weixuan Wang, JINGYUAN YANG, and Wei Peng. Semantics-adaptive activation intervention for llms via dynamic steering vectors. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 79334–79351, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/c4d26a95fd83f8e590f81c54ae670b5d-Paper-Conference.pdf.

Zehan Wang, Sashuai Zhou, Shaoxuan He, Haifeng Huang, Lihe Yang, Ziang Zhang, Xize Cheng, Shengpeng Ji, Tao Jin, Hengshuang Zhao, and Zhou Zhao. Spatialclip: Learning 3d-aware image representations from spatially discriminative language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29656–29666, June 2025c.

Zeyu Wang and Xinming Xu. Knowing isn't always saying: When do spatial encodings reach answers in vision-language models?, 2026. URL https://arxiv.org/abs/2608.22916.

Shaoxiong Zhan, Yanlin Lai, Zheng Liu, Hai Lin, Shen Li, Xiaodong Cai, Zijian Lin, Wen Huang, and Hai-Tao Zheng. 3viewsense: Spatial and mental perspective reasoning from orthographic views in vision-language models, 2026. URL https://arxiv.org/abs/2603.07751.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/06a52a54c8ee03cd86771136bc91eb1f-Paper-Conference.pdf.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *ECCV*, 2024.