LOOK ONCE: A MID-DEPTH COMMITMENT WINDOW EXPLAINS SPATIAL INSTABILITY IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) perform inconsistently in spatial reasoning where some lateral relations are near ceiling, while real-world height and orientation remain at or below chance. Their predictions also change under geometric transformations that preserve the correct response. What remains unclear is where the decision forms and whether richer visual inputs, spatial post-training, and explicit reasoning alter this process or improve its reliability. We combine residual-state patching across token regions and depth with an answer-preserving scene reversal that jointly reverses the spatial relation and option order, changing the scene while keeping the correct answer slot fixed. This separates scene-dependent information transfer from an already resolved verdict. Across three architectures, we identify a narrow commitment window where option-token states become verdict-like as causal influence from the vision block declines. Restoring early visual states produces substantial scene-specific transfer before this boundary but little afterward, supporting reduced downstream use rather than content loss alone. Moreover, 88–98% of the most influential attention heads and MLPs lie downstream of commitment. Spatial post-training preserves the broad relay organization, while explicit reasoning improves accuracy by gaining 4.1%, yet prediction flips under answer-preserving reflection increase from 12.3% to 15.0%. These findings shift the design target from supplying more visual evidence to controlling how that becomes a verdict, the commitment window marks where spatial decisions take shape, while most influential downstream computation operates on the verdict already formed.

1 INTRODUCTION

Spatial reasoning is a core requirement of perceptual visual question answering, yet it remains unevenly developed in vision–language models (VLMs). Models achieve near-perfect scores on some lateral relations while remaining weak on real-world height and viewpoint-dependent orientation, a pattern that persists across benchmarks Kamath et al. (2023); Liu et al. (2023); Kim et al. (2026); Ma et al. (2025); Rajabi & Kosecka (2024); Zhang et al. (2026). Consider a model asked about the lateral relation between a book and a can in a reference image and its translated version. Although translation changes neither the objects nor their relative arrangement, 41.7% of the items answered correctly in the reference condition become incorrect under at least one translation, demonstrating that even correct spatial predictions are unstable.

Ordinary image replacement can change both the spatial arrangement and its correct answer, making scene-dependent information difficult to distinguish from the verdict derived from it. We introduce an *answer-preserving reversal* where changing to the opposite spatial arrangement while reversing the option order preserves the correct answer slot. Comparing this intervention with scene-only and order-only changes distinguishes scene-dependent from verdict-like causal transfer.

Existing approaches primarily address this weakness by improving the spatial evidence supplied to the language decoder. Spatially specialized models add depth estimates, region-level features, scene graphs, spatial positional encodings, or progressively constructed spatial-reasoning data Li et al. (2026); Chen et al. (2026); Cheng et al. (2024). General-purpose VLMs similarly preserve higher-resolution visual tokens or re-inject multi-level vision-encoder features into several decoder layers Li et al. (2024); Bai et al. (2025). These developments improve how visual information enters

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.

the model, but leave open when that information is converted into a decision and whether additional access remains useful after that conversion.

We therefore ask *RQ1. When and where is image-dependent spatial information converted into an answer-level verdict, and what do different token regions carry before and after this conversion?* Using residual-state patching across token regions and decoder depth, we identify a *commitment window* defined as a narrow transition where option-token states become verdict-like as causal influence from the vision block declines. Restoring early visual states has substantial scene-specific influence before this boundary but little afterward.

A complementary line of mechanistic work asks where spatial information is represented internally. Recent studies identify entangled spatial axes, linearly accessible spatial identifiers, distributed ordering signals, and attention heads associated with particular spatial operations Min et al. (2026); Kang et al. (2026); Cui et al. (2026); Ma et al. (2026) or tracing cross-modal information flow using attention knockout, finding a staged transfer from image tokens into question representations and subsequently to the final prediction position Zhang et al. (2025). These findings establish that spatial information can be localized or decoded, but do not establish where it becomes a particular answer. Component knockout and activation patching provide complementary evidence that 88–98% of the most influential attention heads and MLPs lie downstream of commitment. The median direct contribution accounts for only 2–3% of the total knockout effect. High component influence therefore need not identify the site where visual evidence first becomes a verdict.

Next research question is defined as *RQ2. How do architectural enrichment, spatial post-training, and explicit reasoning affect the commitment-and-relay mechanism and the reliability of spatial decisions?* Across three architectures with different vision encoders, projectors, decoder families, positional schemes, and visual feature injection, the depth and tempo vary, but the same broad organization recurs showing scene-dependent information becomes an answer-level verdict around the visual handoff and is subsequently relayed through text-side computation. Qwen3-VL tests whether visual features injected at several early decoder layers change this route. For the tested spatial decisions, replacing its base visual embeddings has a much larger effect than replacing its DeepStack features, even though both pathways supply visual input before commitment. Spatial post-training likewise preserves the relay peaks and verdict-like option states. During generation, Qwen3-VL-Thinking can let a renewed visual read determine its answer, yet its predictions remain sensitive to answer-preserving transformations. Together, these comparisons separate the availability of visual evidence from the reliability of the verdict formed from it.

Our contributions are:

- **A commitment window: locating the conversion from scene to verdict.** By patching residual states across token regions and decoder depth, we locate a narrow transition where option-token states become verdict-like as causal influence from the vision block declines. Controlled restoration of early visual states produces substantial scene-specific transfer before this boundary but little afterward, supporting reduced downstream use rather than content loss alone. The broad commitment-and-relay organization recurs across three distinct VLM architectures.

- **Answer-preserving reversal separates the scene from its verdict.** A conventional counterfactual image changes both the spatial arrangement and the answer derived from it. By reversing the scene and option order together, we construct a donor whose scene changes while its correct answer slot remains fixed. Comparing this donor with image-only and order-only controls reveals a marked reduction in text-side transfer when the verdict is preserved, distinguishing scene-dependent causal content from an already resolved answer.

- **Architectural enrichment and higher accuracy do not ensure spatial stability.** We establish the scope of the commitment-and-relay mechanism across architectural and training differences, and show that improvements in spatial accuracy need not yield more stable decisions. These findings motivate evaluating how VLMs use visual evidence alongside whether they answer correctly.

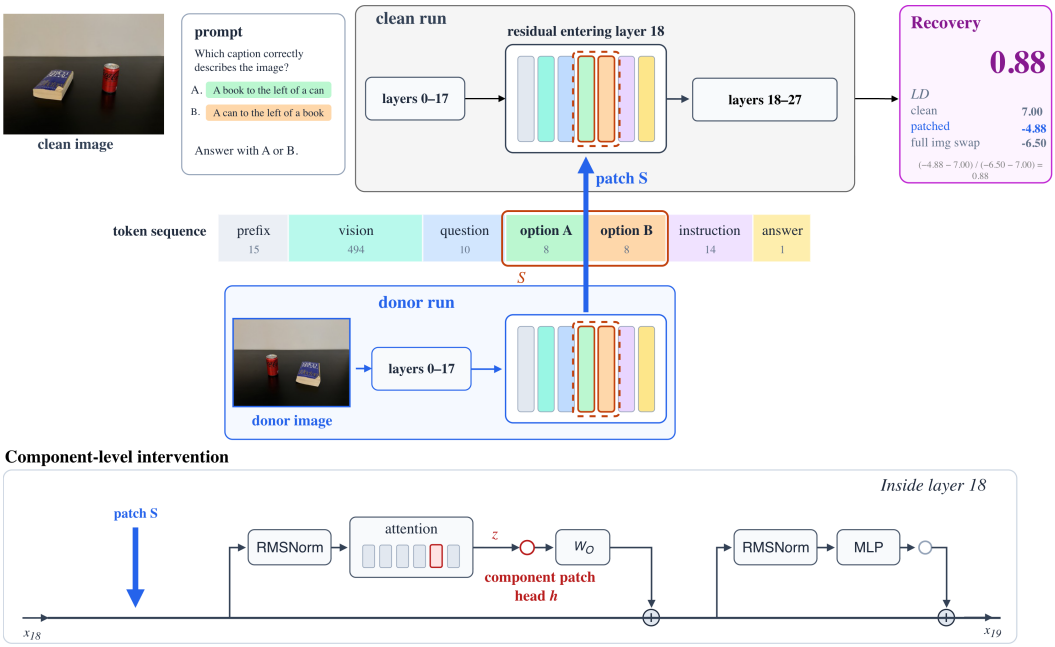

Figure 1: **Residual path patching traces the image-to-decision relay.** The clean run uses the shown image; the donor run uses its paired scene with the objects' positions reversed. At layer l(=18), donor states at the option tokens replace the clean states, and the remaining layers are recomputed. $LD = \ell_A - \ell_B$ measures preference for the clean image's correct option A. Recovery of 0.88 means this patch reproduces 88% of the answer change caused by replacing the entire image. The lower panel shows where an individual attention-head output can be patched within a layer.

2 From Scene to Verdict: Where VLMs Commit

Answer-preserving transformations expose unstable spatial decisions but not where they form. We separate scene changes from answer changes, trace their transfer across decoder depth, and test whether later visual access can revise the resulting verdict.

2.1 Counterfactual Design and Answer-Preserving Reversal

For item i, let y_i be the target answer token and $\bar{y}_i$ the alternative. We measure preference for the target with

$$LD_i = \ell_{i,y_i} - \ell_{i,\bar{y}_i}. \tag{1}$$

We partition the decoder input into the pre-image prefix $\mathcal{P}$, vision block $\mathcal{V}$, question $\mathcal{Q}$, option captions $\mathcal{O}$, instruction $\mathcal{I}$, and answer position $\mathcal{A}$. We also use $\mathcal{V}_n$, a sample of vision positions matched in size to the patched text region.

Donors. A *donor* is a second input whose internal states are copied into the original, or reference, run during patching. The table shows which parts of that input change and whether its correct answer slot changes. Reversing both the scene and option order changes the spatial relation while preserving the correct slot.

Donor	Scene	Option order	Correct slot
Image only	reversed	unchanged	changes
Order only	unchanged	reversed	changes
Question only(antonym)	unchanged	unchanged	changes
Answer-preserving reversal	reversed	reversed	preserved

2.2 RESIDUAL-STATE PATCHING: TRACING THE COMMITMENT WINDOW

We first report *attention enrichment* defined as attention mass assigned to a region divided by that region's share of accessible tokens. It serves as a preliminary test to describe attention allocation, while causal transfer is established by patching Serrano & Smith (2019); Jain & Wallace (2019).

Following activation-patching interventions Meng et al. (2022); Zhang & Nanda (2024); Wang et al. (2022), we record the residual stream for a reference and donor input. At layer L, we replace reference states at positions $\mathcal{S}$ with donor states and recompute the remaining layers, as shown in Figure 1. We measure recovery as

$$R_i(\mathcal{S}, L; d) = \frac{LD_i^{\mathrm{patch}(\mathcal{S}, L \leftarrow d)} - LD_i^{\mathrm{ref}}}{LD_i^{\mathrm{image}} - LD_i^{\mathrm{ref}}}, \tag{2}$$

where $d \in \{\mathrm{img}, \mathrm{ord}, \mathrm{both}\}$ is one of the three main donors and LD_i^{image} is the contrast after replacing the complete image. All three use this image-only denominator, $R_i = 0$ indicates no change in the answer contrast, and $R_i = 1$ reproduces the whole-image replacement effect. We report median recovery across items. The pre-image prefix is a structural null because it cannot receive image information under causal attention.

Visual handoff. We patch the changed scene into the vision-token positions at successive decoder depths. Recovery $R(\mathcal{V}, L; \mathrm{img})$ is initially high, then falls as those positions lose influence on the answer. We call the depth where it falls below 0.5 the **visual handoff**, this threshold is estimated by interpolating between sampled depths.

A low recovery value alone cannot tell us whether scene information has disappeared from the vision states or whether later layers have stopped using them. We therefore restore donor vision states captured at an early, high-recovery layer to positions before and after the handoff, this difference measures scene-specific transfer. We also replace the vision states with their mean to test whether perturbing these positions still affects the answer. Both effects are substantial before the handoff and small afterward, supporting an end to downstream use of the vision positions.

Commitment window. At option-caption positions, the donor pattern changes from

$$R(\mathcal{O}, L; \mathrm{ord}) \approx R(\mathcal{O}, L; \mathrm{both}) > R(\mathcal{O}, L; \mathrm{img})$$

at early depths to

$$R(\mathcal{O}, L; \mathrm{img}) \approx R(\mathcal{O}, L; \mathrm{ord}) > R(\mathcal{O}, L; \mathrm{both})$$

at later depths. Initially these states primarily track which caption occupies each slot; later they transfer the answer verdict. The **commitment window** brackets this transition, with the descending half-crossing of $R(\mathcal{O}, L; \mathrm{both})$ as its estimated depth. Order-only recovery remains high at this transition, showing that the option positions still influence the answer while the content they transfer changes. One-layer scans in Qwen2.5-VL and InternVL2.5 resolve its timing further, their full curves and crossing estimates are reported in the supplementary material.

2.3 COMPONENT INFLUENCE AFTER COMMITMENT WINDOW

Patching locates the transfer between token regions. Component interventions independently test where the computation most influential to the final answer occurs. At the answer position, we zero each attention head's contribution before its output projection or each MLP's output before residual addition, then recompute the remaining network. The total knockout effect is

$$a_i(c) = LD_i^{\mathrm{ref}} - LD_i^{\mathrm{zero}(c)}. \tag{3}$$

We rank components by aggregated knockout influence and measure what fraction of the top 5% lies after the commitment depth. Matched- and unrelated-image patches test their visual dependence; their definitions and controls are in the appendix.

A component's direct logit attribution (DLA) measures its own write toward the answer contrast, whereas knockout includes the response of subsequent layers. Holding the final RMSNorm denominator at its reference value gives the comparable pinned knockout effect $a_i^{\mathrm{pin}}(c)$. For selected

components, we compute the direct-to-total ratio

$$r(c) = \frac{|\mathbb{E}_i[\mathrm{DLA}_i(c)]|}{|\mathbb{E}_i[a_i^{\mathrm{pin}}(c)]|}. \tag{4}$$

We summarize its median across eligible components in the two models with both measurements.

2.4 Testing Changes to Visual Evidence and Reasoning

Spatial post-training. Spatial specialization could change how visual evidence reaches the option states or when those states become verdict-like. We compare SenseNova-SI-1.1 and SenseNova-SI-1.3 with their respective Qwen2.5-VL and Qwen3-VL backbones on behavioral stability, token-region recovery, and commitment timing.

Explicit reasoning. Additional reasoning over the available vision tokens and RL training to enhance visual attention during reasoning offer two ways to revisit a spatial decision. We compare Qwen3-VL-Thinking and Reflection-V Jian et al. (2025) with their respective Instruct backbones, reporting accuracy and answer-preserving flip rates on matched valid examples. Direct answers are read from candidate logits; reasoning answers are extracted after generation. For Reflection-V, residual patching measures the initial answer preference before the reasoning trace, which is reported separately from the final generated answer.

KV Cache swap to check visual access during generation. To test whether a later visual read controls the final answer, we prefill the reference input and replace cached keys and values at vision positions in every layer with those from a donor image before generation. The reference text cache remains in place. We compare the swapped answer with the full-donor answer on examples where the reference and full-donor answers differ. Residual patching uses the fixed target answer in Eq. 1; this donor-following comparison uses the models' generated answers, including for relations on which the model may be wrong.

Additional visual input through DeepStack. Qwen3-VL places base visual embeddings at the decoder input and projects features from three intermediate vision-encoder layers through separate mergers, adding them at image positions after decoder layers 0, 1, and 2. DeepStack is intended to expose the decoder to visual features from multiple encoder depths. We ask whether this additional input pathway carries the spatial difference that changes an answer. For a reference–donor image pair, we run both images to obtain their base embeddings and three DeepStack tensors. We then replay the decoder with donor base embeddings and reference DeepStack features, or reference base embeddings and all three donor DeepStack features. Recovery in each condition isolates which visual input pathway transmits the tested scene change.

3 Experimental Setup

Models. We evaluate Qwen2.5-VL-7B-Instruct, InternVL2.5-8B, and Qwen3-VL-8B-Instruct Bai et al. (2025); Chen et al. (2024c), replicating the central commitment, reversal, and component analyses across these architectures. SenseNova-SI-1.1 Cai et al. (2026) and SenseNova-SI-1.3 extend it to show effects on spatially post-trained models. Qwen3-VL-8B-Thinking and Reflection-V-7B Jian et al. (2025) provide reasoning and KV-cache comparisons with their respective backbones. Model-specific prompts, decoding settings, and answer-extraction rules are documented in the appendix.

Datasets and tasks. We use binary-choice tasks from What'sUp and 3DSRBench Kamath et al. (2023); Ma et al. (2025). The core What'sUp analysis covers two lateral, front/behind and on/under subsets. Its paired photographs provide matched causal donors with opposite spatial arrangements. Reflection evaluation covers ten 3DSRBench categories, while translation uses validated crops from real-world height, camera depth, and viewpoint-dependent orientation.

Geometric transformations. We shift the image crop by $k \in -4, -2, -1, 0, 1, 2, 4$ units of 28 pixels, retaining the relevant objects without artificial borders. Reflection reverses lateral and subject-centered left/right relations, while preserving What'sUp front/behind and on/under relations

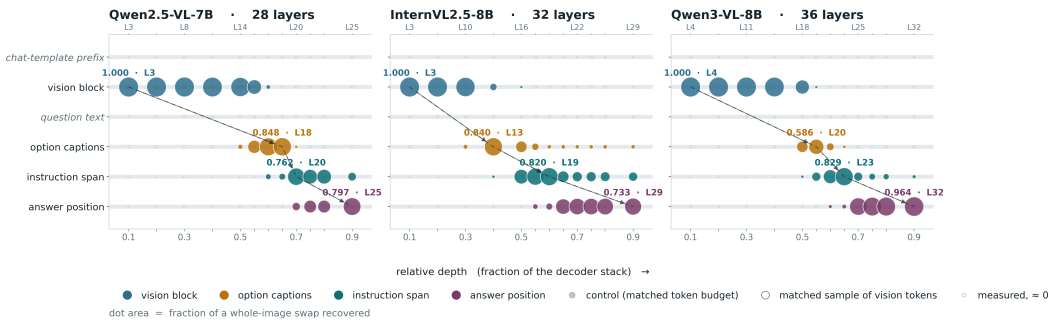

Figure 2: **Image-to-decision relay across relative decoder depth.** Circle area denotes recovery from a whole-image swap. Blue, orange, teal, and purple mark the vision block, option captions, instruction span, and answer position, respectively.

Table 1: **High spatial accuracy can coexist with fragile correct answers.** This shows the flips caused by mirror and displacement transformations to clean image on What'sUp and 3DSRBench. Correct-to-fragile is the fraction of initially correct answers that change under at least one shift.

Model	Accuracy	112 px flip	Any-shift flip	Correct-to-fragile	Mirror error
What'sUp: three relations					
Qwen2.5-VL	95.9%	10.2%	21.2%	20.2%	3.4%
InternVL2.5	98.5%	9.0%	19.5%	19.0%	1.0%
Qwen3-VL	99.3%	7.9%	16.0%	15.9%	0.0%
SenseNova-1.1	95.9%	12.7%	25.6%	24.3%	8.8%
SenseNova-1.3	99.7%	9.4%	18.9%	19.0%	0.0%
3DSRBench					
Qwen2.5-VL	62.2%	22.4%	47.9%	41.7%	13.7%
InternVL2.5	63.6%	22.0%	42.4%	36.9%	13.3%
Qwen3-VL	68.6%	20.1%	41.7%	35.9%	12.4%
SenseNova-1.1	63.2%	20.5%	44.4%	38.3%	13.1%
SenseNova-1.3	67.4%	19.7%	38.6%	32.2%	11.6%

and the other nine 3DSRBench categories. We report accuracy, prediction flips under each answer-preserving transformation, flips under any tested shift, and correct-to-fragile rates among initially correct examples. Answer-reversing reflection cases are evaluated separately.

Residual-state patching. We sample 12 relative decoder depths mapped to the nearest valid layer, and patch the seven position sets defined in Section 2.2. Donors share the question and instruction, with scene and option order varied as specified. Qwen3-VL patches begin after its DeepStack additions, and option spans are restricted to their option lines.

4 Results

4.1 Stable Accuracy Hides Unstable Spatial Decisions

Table 1 shows substantial prediction changes across the three base architectures. Spatial post-training reduces some flip rates but does not remove the instability. Aggregate accuracy also conceals opposing changes as at 112 pixels, Qwen2.5-VL changes 11.7% of predictions from correct to incorrect and repairs 10.7% in the other direction, leaving only a 1.1-point net accuracy change despite a 22.4% flip rate.

The contrast across relations is equally sharp: Qwen3-VL is accurate but shift-sensitive on What'sUp *a_lateral*, while its on/under answers are both accurate and stable under the tested transformations. The following interventions examine how those decisions are formed.

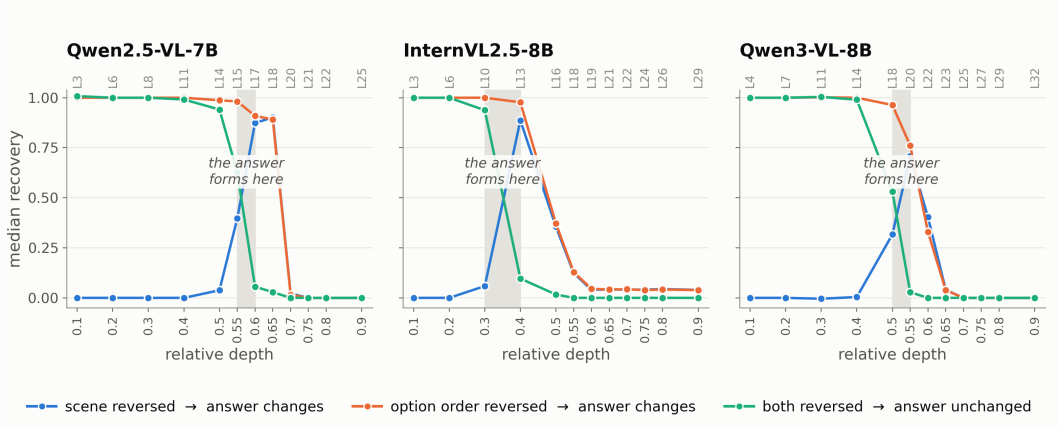

Figure 3: **The commitment window converts the changed scene into an answer-level verdict.** Shaded regions mark the transition where recovery from the answer-preserving donor collapses while both verdict-changing donors remain effective.

Table 2: **Visual transfer reaches a verdict-like option state.** Recovery is normalized to the whole-image effect. Instruction and answer entries show their own peak recoveries and layers. Handoff and verdict formation are interpolated relative depths on the sampled grid.

Model	n	Option peak	Recovery at option peak			Later recovery peaks		Handoff → verdict
			Image	Order	Both	Instruction	Answer	
Qwen2.5-VL	446	L18	0.848	0.810	**0.030**	0.762 (L20)	0.797 (L25)	0.554 → 0.555
InternVL2.5	320	L13	0.840	0.969	**0.141**	0.820 (L19)	0.733 (L29)	0.355 → 0.355
Qwen3-VL	446	L20	0.586	0.529	**0.024**	0.829 (L23)	0.964 (L32)	0.500 → 0.494
SenseNova-1.1	444	L18	0.850	0.858	**0.039**	0.757 (L20)	0.889 (L25)	0.556 → 0.557
SenseNova-1.3	446	L20	0.744	0.814	**0.049**	0.852 (L23)	0.964 (L32)	0.480 → 0.475
Reflection-V	446	L18	0.957	0.977	−0.002	0.893 (L20)	0.861 (L25)	0.565 → 0.560

4.2 The Image-to-Decision Relay Replicates across Architectures

Attention offers a preliminary view of routing. Vision-attention enrichment is low on average at the option captions, but attention allocation alone cannot establish which visual information affects the answer. We therefore measure causal transfer with residual-state patching.

Figure 2 and Table 2 show the same progression across architectures: influence from the vision block appears at the option captions, then at the instruction and answer positions. On the four-relation What'sUp pool, the sampled question positions recover at most 0.002 and the pre-image prefix remains at zero. This localizes the transfer more specifically than attention or token proximity alone. The three option-position interventions in Table 2 allow us to determine what is transferred at this handoff.

4.3 The Commitment Window Converts a Scene into a Verdict.

Image-only replacement changes both the scene and its correct answer, so its recovery alone cannot identify what the option states carry. Reversing both the scene and option order preserves the correct answer slot. In Figure 3, recovery from this preserved-answer donor declines at the option positions while the order-only patch remains effective at the same depth. The option positions can still change the decision, but they transfer little of a scene change that leaves the answer slot fixed. Changing only the relation word, without moving the captions, gives the same separation.

We call the sampled interval in which option states change from following the captions in their slots to following the correct slot the *commitment window*. On What'sUp relations, the vision handoff and verdict formation occur close together in the sampled depth curves; orientation relation is an exception discussed below.

Table 3: **Orientation verdicts form after the visual handoff.** Option-peak recovery and relative-depth estimates on 3DSR orientation.

Model	n	Image	Both	Handoff $\rightarrow$ verdict
Qwen2.5-VL	68	0.525	0.243	0.556 $\rightarrow$ 0.594
InternVL2.5	78	0.290	0.089	0.380 $\rightarrow$ 0.501
Qwen3-VL	72	0.474	0.135	0.502 $\rightarrow$ 0.536
SenseNova-1.1	56	0.483	0.300	0.558 $\rightarrow$ 0.609
SenseNova-1.3	42	0.320	0.214	0.511 $\rightarrow$ 0.555

Table 4: **A later visual read can determine the answer without improving stability.** Behavioral changes are relative to each model's direct baseline on matched valid examples. KV donor-following is includes cases where the full-donor answer differs from the reference answer.

Model	Accuracy Δ	Mirror flip Δ	Any-shift flip Δ	KV follows donor(%)
Qwen3-VL-Thinking	$+4.1$	$+2.7$	$+1.0$	90.9%
Reflection-V	$+1.0$	$+4.3$	$+3.9$	4.1%

The loss of influence from vision positions reflects when they are used, not simply what their late states contain. Before the handoff, inserting early donor-image states transfers substantial scene information. Afterward, inserting those states has no appreciable scene-specific effect beyond inserting the reference image's own early states. The raw effect is explained by moving early states to a late depth, because it also occurs without donor-scene information. Therefore, tested vision positions contribute little to the remaining prefill decision after the handoff.

If the scene has already become a verdict at this boundary, the components with the largest effects on the final answer should lie mostly beyond it. In the component sweep, 88%, 98%, and 90% of the top-5% most influential heads and MLPs do so across the three base architectures. These components remain responsive to the paired image, but their median direct writes explain only 1.6–3.1% of their total pinned knockout effects in the two models where both were measured. Knockout therefore detects substantial influence that a direct answer-direction measurement misses. This indirect influence occurs across depth rather than beginning abruptly at commitment.

The timing alignment is relation-dependent. In the 3DSR orientation runs, option commitment follows the vision handoff in all five evaluated models (Table 3), most markedly in InternVL2.5. The commitment window is therefore a measurable boundary, but the handoff and answer formation are not locked together for every relation.

4.4 Targeted Training Preserves the Scene-to-Verdict Readout.

Spatial post-training does not substantially change the measured readout organization. Both SenseNova variants retain their respective backbones' option, instruction, and answer peak layers. On the same four-relation What'sUp evaluation, changing the image still transfers strongly at the option positions, while the preserved-answer donor has little effect. The option states thus remain visually dependent and verdict-like after spatial specialization. This conclusion comes from the same causal interventions on each backbone and trained variant, rather than from their similar accuracy alone. Reflection-V tests training that encourages visual attention during reasoning. Its initial option preference still peaks at L18, with strong image-only and order-only transfer and near-zero preserved-answer transfer (Table 2). The initial decision forms near the vision handoff even though the model is trained to continue looking during generation. These patching measurements concern the preference before the trace; we test the trace's influence on the final answer in §4.5.

Additional visual features through DeepStack. DeepStack supplies Qwen3-VL with additional visual features before the commitment window. Yet replacing its base embeddings alone recovers 0.972 of the tested whole-image effect, whereas replacing all three DeepStack feature sets alone recovers 0.028. The same pathway asymmetry appears in SenseNova-SI-1.3 and in the tested orientation subset. These replacements locate the route by which the counterfactual scene affects the trained model; they do not measure the benefit of training with DeepStack.

4.5 Renewed Visual Access Does Not Ensure Stable Answers

Does the image still control an answer after generation begins? We answer this by processing the reference image, replacing the vision positions in its KV cache with those from a paired scene, and then generating the answer. This leaves the initial computation based on the reference image while changing what the generated tokens can retrieve. Neither direct-answer model follows the substituted image on an eligible What'sUp item. Qwen3-VL-Thinking, by contrast, follows it on 90.9% of samples and a later visual read can determine the final answer, so the initial verdict is not irrevocable during reasoning.

Reflection-V exposes a different outcome. Its generated trace changes under every tested swap and can describe the substituted scene, yet its final answer follows that scene on only 4.1% of eligible samples. Renewed visual information can enter the trace without displacing the answer favored before generation. Despite these different responses to the cache swap, neither reasoning model becomes more stable under answer-preserving transformations. Both gain accuracy relative to their matched direct baselines, while mirror and shift flips increase (Table 4). In Qwen3-VL-Thinking, real-world height improves but viewpoint-dependent orientation remains below chance, consistent with the relation-dependent timing in Table 3. The cache intervention establishes the sharper point that even when a second visual read controls most final answers, that capability alone does not make spatial decisions stable.

5 Related Work

Spatial benchmarks reveal uneven performance across relations and viewpoints Kamath et al. (2023); Ma et al. (2025); Yang et al. (2025); Stogiannidis et al. (2025). Spatially specialized models address these weaknesses with spatial supervision, structured visual information, or depth inputs Cheng et al. (2024); Cai et al. (2026); Chen et al. (2024a); Cai et al. (2025). These works establish which abilities improve, but accuracy alone does not show whether an individual spatial decision is stable. Mechanistic studies locate spatial representations and trace visual information into text-side computation Min et al. (2026); Kang et al. (2026); Zhang et al. (2025); Palit et al. (2023); Neo et al. (2025). Other approaches reduce visual tokens in later layers Chen et al. (2024b); Endo et al. (2025), inject features at multiple decoder depths Meng et al. (2024); Bai et al. (2025), or encourage renewed visual access during reasoning Wu & Xie (2024); Jian et al. (2025). These studies leave open when visual evidence is converted into an answer, why influence from the vision block declines, and whether later visual access can revise the decision.

6 Conclusion

We locate a commitment window in which spatial evidence becomes an answer-level state at the option positions. Answer-preserving reversal shows what changes at this point. Substituting either the scene or the option order strongly affects the answer, while substituting both has little effect when the correct slot remains fixed. On the core relations, this conversion occurs near the end of causal transfer from vision positions. Early-state restoration and its controls indicate that later decoder layers make little further use of those positions. Component interventions support the downstream influence. The same organization persists after spatial post-training and visual-attention RL. Additional early visual features do not change the dominant pathway for the tested scene differences. A reasoning model can even let a later visual read determine its final answer without becoming stable under answer-preserving transformations. These findings suggest two concrete intervention targets: retain a causally usable spatial relation through commitment, or enable a later visual read to revise an earlier verdict. Our commitment-window measure can test whether either change actually occurs.

Limitations Our main localization analysis intervenes on decoder residual states and selected component outputs; it does not characterize the latent feature basis of the spatial relation. We compare architectures with different vision encoders and positional schemes but do not isolate the causal contribution of either design choice. The experiments also use forced-choice spatial answers, and we have not demonstrated a training or architectural intervention that improves reliability by changing the commitment window.

AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing the writing, particularly the results section where a variety of metrics and analysis is mentioned to support the hypothesis. We have also used AI for retrieval and discovery of related work and discovery as our work involves multiple subdomains across VLMs, spatial understanding, mechanistic interpretability and reasoning models. We have not used generative AI tools for drafting entire sections of the paper, and other tasks are not applicable to this work. Additionally, we used generative AI tools for execution of certain architecture specific debugging such as how vision tokens are handled in InternVL 2.5 versus Qwen family. We have reviewed all AI-assisted work through manual literature survey and cross-checking with published architecture details about the chosen models in our work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Code and evaluation scripts for reproducing the reported results are available in an anonymous repository at `https://anonymous.4open.science/r/spatial-verdict-7244`. We will make the repository public upon publication.

REFERENCES

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

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9490–9498. IEEE, 2025.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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*, pp. 14455–14465, 2024a.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024b.

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. Sd-vlm: Spatial measuring and understanding with depth-encoded vision-language models. *Advances in Neural Information Processing Systems*, 38:33938–33968, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

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.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22826–22835. IEEE, 2025.

Sarthak Jain and Byron C. Wallace. Attention is not explanation, 2019. URL https://arxiv.org/abs/1902.10186.

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9262–9281, 2025.

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.

Yoonji Kim, Jieun Kim, Yujin Jeong, and Sung-Bae Cho. Diagnosing spatial consistency across perspectives and viewpoints in large vision-language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32803–32827, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1514. URL https://aclanthology.org/2026.acl-long.1514/.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 76566–76592, 2026.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso De Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934, 2025. doi: 10.1109/ICCV51701.2025.00651.

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. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Cheolhong Min, Jaeyun Jung, Daeun Lee, Hyeonseong Jeon, Yu Su, Jonathan Tremblay, Chan Hee Song, and Jaesik Park. Why far looks up: Probing spatial representation in vision-language models. In *European Conference on Computer Vision*, pp. 22–39. Springer, 2026.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2848–2853. IEEE, 2023.

Navid Rajabi and Jana Kosecka. Gsr-bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms. *arXiv preprint arXiv:2406.13246*, 2024.

Sofia Serrano and Noah A. Smith. Is attention interpretable?, 2019. URL https://arxiv.org/abs/1906.03731.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 70072–70141, 2026.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791. IEEE, 2025.

A Appendix

This appendix extends the main paper with concrete prompts, the full commitment and boundary tests, and the controls behind the other measurements. All recovery values use the whole-image scale in Eq. 2.

B Prompts and donor construction

Prompts (Sections 2.1 and 3). The following are example items. Each line in a prompt is separated by a semicolon here, and each prompt ends with Answer with A or B. The objects vary by item. What'sUp flips option order on odd-indexed pairs, balancing the correct letter. Table 5 gives example prompts for each tested relation. Generated-answer models receive the same item prompts. Qwen3-VL-Thinking is decoded greedily up to 2,048 tokens and scored after its last </think> tag. Reflection-V uses an R1-style system prompt and is scored inside <answer> tags. Unparseable answers are excluded. The supplied method notes do not include the literal Reflection-V system string or rendered model chat templates; these must be pasted from the released scripts before submission.

Donors. Table 6 gives each change from clean to donor. Paired image means the opposite arrangement from the same pair. For example, apple on armchair with A: on and B: under becomes apple under armchair with A: under and B: on. A is correct both times. For the word control, left changes to right in both captions without moving either caption. **Takeaway:** answer-preserving donors change the relation without changing the correct slot, allowing scene transfer to be separated from verdict transfer.

Table 5: **Example prompts.** Semicolons mark line breaks; every row ends with Answer with A or B.

Relation	Prompt before the final instruction
What'sUp lateral (B)	Which caption correctly describes the image?; A. A book to the left of a can; B. A can to the left of a book
What'sUp depth	Which caption correctly describes the image?; A. A book in front of a can; B. A can in front of a book
What'sUp lateral (A)	Which caption correctly describes the image?; A. An apple to the left of an armchair; B. An armchair to the left of an apple
What'sUp vertical	Which caption correctly describes the image?; A. An apple on an armchair; B. An apple under an armchair
3DSR orientation	Consider the real-world 3D locations and orientations of the objects. If I stand at the person in white sweater's position facing where it is facing, is the blue arrow sign on the left or right of me?; A. on the left; B. on the right
3DSR height	Consider the real-world 3D locations of the objects. Which object has a higher location?; A. step stool; B. notebook
3DSR camera depth	Consider the real-world 3D location of the objects. Which object is closer to the camera?; A. red bus; B. stop sign

Table 6: **Clean-to-donor changes.** Same means that the correct answer slot stays fixed.

Donor	Image	Prompt change	Gold slot
Image	Paired	None	Flips
Order	Same	Exchange the A and B captions	Flips
Both	Paired	Exchange the A and B captions	Same
Antonym	Same	Change left to right in both captions without moving them	Flips
Scene + antonym	Paired	Make the same word change	Same
Question	Same	Change higher to lower in the question	Flips
Question + order	Same	Change the question word and exchange the options	Same

Scoring and position sets (Section 2.2). We patch seven sets: pre-image prefix, full vision block, sampled vision positions, sampled question positions, complete option captions, sampled instruction positions, and answer position. Samples are fixed per item and contain at most as many tokens as the options. The twelve relative depths map to these layers:

Depth	.10	.20	.30	.40	.50	.55	.60	.65	.70	.75	.80	.90
Qwen2.5-VL	3	6	8	11	14	15	17	18	20	21	22	25
InternVL2.5	3	6	10	13	16	18	19	21	22	24	26	29
Qwen3-VL	4	7	11	14	18	20	22	23	25	27	29	32

All donors divide by the image-only answer change, including the donors that keep the correct slot fixed. Recovery is the median over eligible items. Values with denominator magnitude at most 10^{-6} are undefined; additional relation and boundary analyses require magnitude at least one. Reference and donor region indices must match. Direct answers use float32 gold-letter logit differences; KV donor-following compares generated answers. **Takeaway:** donor recovery values share one scale and correspond to the same token slots.

C TIMING OF THE COMMITMENT WINDOW

Extension of Section 4.3. We repeat the image and answer-preserving patches at every layer around the transition in two models. The pre-image prefix has recovery 0.000 at all nine layers in both scans. Table 7 shows recovery at each scanned layer.

The crossings in Table 8 are interpolated between adjacent layers; their difference is option crossing minus vision crossing. The gap intervals use 2,000 bootstrap samples. InternVL uses FlashAttention-2 for both curves in this scan and is not mixed with its eager-attention coarse run.

Table 7: **The visual handoff and option-state conversion in the one-layer scans.** Values are median recovery on the three core What'sUp relations. The pre-image control is 0.000 at all nine depths in both models.

Model	Patched region / donor	Decoder layer								
Qwen2.5-VL		12	13	14	15	16	17	18	19	20
	Vision / image	.984	.967	.959	.608	.491	.056	.009	.008	.000
	Options / both	.983	.949	.941	.627	.464	.056	.028	.008	.000
InternVL2.5		8	9	10	11	12	13	14	15	16
	Vision / image	1.000	1.000	.947	.891	.761	.076	.077	.053	.020
	Options / both	.993	.991	.944	.885	.789	.098	.088	.050	.022

Table 8: **Crossing estimates in decoder layers.** Both intervals exclude a gap of one full layer. These measurements concern the core What'sUp relations; orientation is separately reported in the main paper.

Model	Visual handoff	Option conversion	Gap [95% CI]
Qwen2.5-VL	15.92	15.78	-0.14 [-0.42, $+0.14$]
InternVL2.5	12.38	12.42	$+0.04$ [-0.03, $+0.08$]

Takeaway: on the core What'sUp relations, the two crossings occur together within measurement resolution. Orientation is an exception in the main paper.

D Restoring visual states at the handoff

Extension of Sections 2.2 and 4.3. Low late vision recovery could mean missing content or an ended read. We restore early donor states before and after the handoff. Restoring the reference item's own early states controls for writing an early-layer state at a late depth; the mean-state condition tests whether a perturbation of these positions still matters. There are 240 attempted What'sUp items per model, with 12 and 6 exclusions, respectively. A same-depth self patch has maximum $|\Delta LD| = 0.00000$ in both models. Table 9 compares the interventions before and after handoff.

At InternVL L16, donor-early yields .269 but reference-early yields .271. Thus the large raw effect comes from the cross-depth write, not the other scene. Mean-state changes the answer by only .016 there, against .492 before handoff. The net column is the reported difference of summary effects; a paired estimate would require calculation from the per-item records. **Takeaway:** early scene content does not regain scene-specific influence after the tested visual handoff.

E Supporting measurements

E.1 Attention enrichment

Extension of Section 4.2. Enrichment is attention to vision divided by vision's proportion of accessible tokens. It describes allocation, while patching measures causal transfer. Table 10 reports enrichment at the option and answer positions.

Options attend more to vision than the answer position on average. Maximum cells exceed one in both models, so some cells do attend above the proportional share. **Takeaway:** attention suggests a route, while patching shows what information affects the decision.

E.2 Relation-word control

Extension of Section 4.3. On 160 lateral What'sUp items, changing left to right in both captions changes the answer without moving their positions. Combining the word change with the paired scene preserves the answer. Table 11 compares the scene-only, word-only, and combined changes.

Table 9: **Scene transfer ends after the handoff.** Donor columns are recovery; reference-early and mean-state effects are changes in logit difference divided by the item's whole-image denominator. Brackets give 2,000-sample bootstrap intervals for the three source-state columns. The net column is the reported difference between donor-early and reference-early effects; see the note below before interpreting it as an item-paired estimate.

Model	Layer	Donor same depth	Donor early	Reference early	Net	Mean state
Qwen2.5-VL	L12 (before)	.989 [.983,.994]	.787 [.762,.810]	.197 [.166,.225]	+.590	.494
	L18 (after)	.009 [.006,.011]	.034 [.025,.042]	.032 [.024,.040]	+.002	.015
	L20 (after)	.000 [.000,.001]	.025 [.023,.028]	.025 [.022,.027]	.000	.006
InternVL2.5	L9 (before)	.998 [.996,1.001]	.839 [.831,.862]	.134 [.119,.152]	+.705	.492
	L14 (after)	.078 [.057,.098]	.110 [.103,.119]	.094 [.090,.104]	+.015	.075
	L16 (after)	.014 [.012,.016]	.269 [.260,.277]	.271 [.261,.281]	−.002	.016

Table 10: **Vision-attention enrichment.** Option and answer entries average across measured layers and heads; maximum is over measured layer/head/query cells.

Model	Options	Answer	Maximum cell
Qwen2.5-VL	.162	.040	1.158
InternVL2.5	.187	.088	1.022

Takeaway: answer-preserving recovery stays small without moving captions, supporting the verdict reading of the option states.

E.3 Component influence

Extension of Section 2.3. Knockout removes one head or MLP output at the answer position and recomputes later layers. A matched patch uses the paired image; an unrelated patch uses an image from another pair. The same-image patch is the identity check. Components are ranked by absolute mean knockout effect; the main paper locates the top 5%. Direct attribution measures the component's own answer-directed write. For comparison with total knockout, the final RMSNorm denominator is pinned to its reference value. Components below .02 logits of mean knockout effect and ratios above five are excluded. A separate depth test finds no additional mediation step at commitment after accounting for a smooth depth trend ($p = .99/1.00/.59$ across base models). This depth test uses an unpinned ratio rather than the main pinned estimate. **Takeaway:** influential components are mostly downstream, but indirect influence itself does not begin at the window.

Table 11: **A verdict-preserving control without moving captions.** Median option-position recovery on 160 lateral items.

Model	Scene only	Word only	Scene + word
Qwen2.5-VL	.907	.967	$-.048$
InternVL2.5	.852	.925	.083
Qwen3-VL	.674	.787	.039