WHAT HAPPENS BEFORE RECOGNITION? BACKGROUND EFFECTS ON TARGET STRUCTURE IN HIGH-RESOLUTION VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained target recognition remains challenging for vision language models (VLMs) in high resolution images with extensive background. Existing work has mainly treated background interference as a target-search problem, while recent work points to an overlooked effect on changing the target itself. However, without separating and measuring these two background effects, the source of recognition failure remains unclear, obscuring the true bottleneck and limiting existing remedies to partial recovery. In this paper, we introduce a target-fixed causal framework that separates background interaction in the vision transformer (ViT) from background competition in the LLM and traces their consequences for recognition. Results show that changes in target representations induced by background contribute substantially more to recognition failure than token competition. We identify an *attention fragility* mechanism in early ViT blocks, where limited background interaction with target is sufficient to disrupt the target structure relevant for recognition. We further uncover LLM's compensatory localization response: as background grows, target selectivity increases despite declining recognition. Finally, we use layerwise repair and reveal two distinct propagation pathways, completing a causal link from upstream representation damage to downstream answer formation. These results establish target structure preservation as a missing dimension of VLM robustness, providing a mechanistic basis for both architecture design and evaluation under large visual contexts.

1 INTRODUCTION

Large language models (LLMs) have developed strong capabilities in language understanding and reasoning in recent years (OpenAI, 2023; Yang et al., 2025). Vision language models (VLMs) extend these capabilities to visual inputs (Li et al., 2024a), typically by connecting a pretrained vision transformer (ViT) to an LLM through a lightweight projector (Bai et al., 2025a; Wang et al., 2025b). In real-world applications, visual inputs are moving from coarse, low-resolution imagery toward detailed high-resolution scenes, particularly in remote sensing (Wang et al., 2025a), autonomous driving (Sun et al., 2020), and human-centric video analysis (Wang et al., 2020). To support better visual understanding, VLMs are increasingly designed to preserve more details by processing high-resolution images as larger sets of visual tokens. A key challenge in such settings is recognizing small targets, where target-relevant information is surrounded by a much larger amount of background interference (Zhang et al., 2025b; Jiang et al., 2025a).

To address this challenge, existing works primarily focus on enabling the model to locate the target among increasing background and preserve its information as it is passed into the language backbone. Region-based methods improve access to small targets through active search (Wu & Xie, 2024), guided reinspection (Zhang et al., 2025a), task-aware cropping (Yi et al., 2026), or other region selection strategies (Shi et al., 2026; Jiang et al., 2025a; Wang et al., 2025c). Other studies show that even when the relevant region is found, recognition can still fail, and trace these failures to how visual information is attended to or propagated within the LLM (Chen et al., 2025; Kang et al., 2025b; Liu et al., 2026). These methods mainly frame additional background as a downstream problem of locating and using target information in the LLM. But this framing overlooks another effect of background growth: even when the target itself is fixed, progressively increasing the surrounding

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

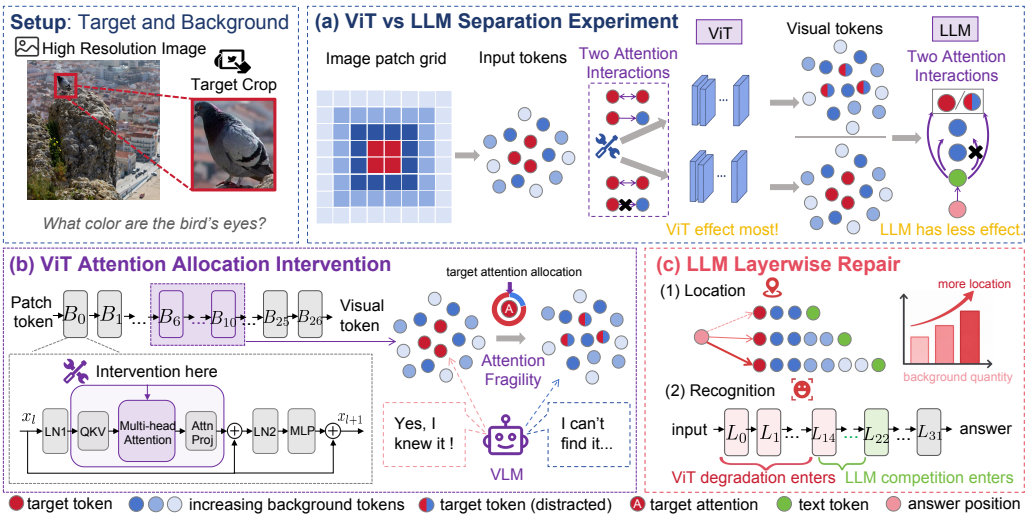

Figure 1: Causal analysis of background interference from target encoding to LLM decision. We separate background effects in the ViT and LLM (a), intervene on target computation inside the ViT (b), and trace how the resulting target information affects the LLM decision (c).

background can alter its representation and make it harder to recognize. It also remains unclear how additional background changes the representation of a fixed target inside high-resolution VLMs and whether these changes causally drive recognition failure.

In this work, we develop a target-fixed causal framework to trace how background interference propagates through the VLM pipeline to the final answer. (i) We first separate background interference during visual encoding from direct background exposure in the LLM. By separately controlling the two effects (Figure 1(a)), we find that a substantial part of the recognition failure originates in the vision encoder by changing the target representations. (ii) We then examine the underlying mechanism inside the vision encoder and identify *attention fragility* across the models. Although most attention from target tokens remains within the target, background tokens capture a growing share across the early and middle ViT blocks, disrupting the representation structure required for recognition (Figure 1(b)). (iii) At the representation level, we observe persistent changes in how target information is organized through the visual encoding. Crucially, repairing the token-relative component yields most of the correct-answer margin gain achieved by complete repair, identifying these structural changes as a causal contributor to the model's correct-answer margin. (iv) Finally, tracing these changes into the LLM reveals a compensatory increase in target selectivity and shows that ViT-induced degradation and direct LLM competition affect distinct stages of answer formation (Figure 1(c)). These analyses connect target structure damage to the internal decision pathway of the LLM, forming a complete causal chain from visual encoding to final answer formation.

Overall, our main contributions are as follows:

- We introduce a target-fixed causal framework that independently controls and quantifies background interaction in the ViT and direct background competition in the LLM, which form a unified basis for causal tracing across the VLM pipeline.
- We establish visual encoding as the dominant source of background-induced recognition failure: even when the target is fixed, background interaction alters its representation primarily by changing its internal relative shape inside the vision encoder.
- We uncover *attention fragility* in early blocks as the causal mechanism behind this degradation: even a small attention share captured by background has a disproportionate impact on target internal evolution, offering a concrete principle for future visual encoders.
- We revisit and explore the LLM decision process through the two background effects, revealing compensatory target localization and two distinct pathways to final recognition, providing a new causal approach for dissecting multimodal decision formation within the language backbone of VLMs under increasingly complex visual contexts.

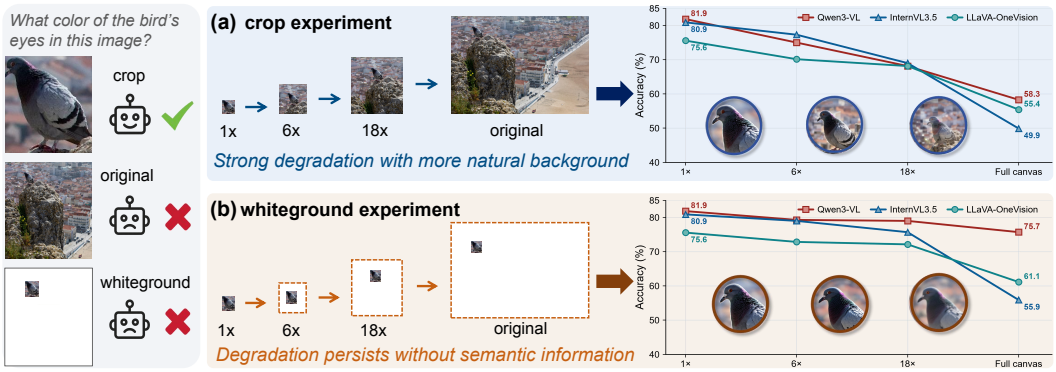

Figure 2: Recognition worsens as the surrounding context expands. Left: the tight crop succeeds, whereas the original and whiteground inputs fail. **(a)** Accuracy falls with added natural background. **(b)** The degradation persists without the semantic background.

2 What Happens as More Background Surrounds a Fixed Target?

VLMs often fail to recognize a target in the full high-resolution image (regardless of the background content) but succeed once the same target is cropped out (Wu & Xie, 2024; Zhang et al., 2025a), as illustrated in Figure 2(left). This failure in target recognition can lead to a more accurate question: when the target input is kept unchanged, how does increasing background affect target recognition in VLMs? To answer this question, we first design an experiment in which the same target is surrounded by progressively more background, allowing us to measure how target recognition changes as visual distraction increases (Section 2.1). Second, we adjust the proportion of target tokens and background tokens inside the model. This enables tighter causal control and lets us distinguish and compare how the two forms of background intervention affect target recognition (Section 2.2).

2.1 Recognition Degrades as More Context Is Retained

More context degrades target recognition. To test how retaining more background around the same target changes recognition accuracy, we establish a benchmark using 700 images with annotated target bounding boxes from high-resolution datasets (Appendix A.1). We start the *Crop* experiment by cropping the target region and progressively expand it to include more background. As shown in Figure 2(a), the scales $1\times$, $6\times$, and $18\times$ denote the crop area relative to the target region (for example, $6\times$ means six times the target area). For each scale, we evaluate nine open-source VLMs on all the image-question pairs. The result is that all nine VLMs become less accurate in target recognition as more surrounding context is included (Appendix A.3). For example, Qwen3-VL decreases from 81.9% accuracy at the tight crop to 58.3% on the original image. These results show that recognition accuracy generally declines as more background around the same target is retained.

Recognition degradation persists without semantic information. One possible explanation for this decline in recognition is that larger crops introduce more distracting content. We therefore design a *Whiteground* experiment that progressively expands the crop at the corresponding *Crop* scales but fills each newly added region with white pixels instead of natural image content. Expanding the input thus increases the surrounding canvas without introducing natural background content. Figure 2(b) shows that recognition still decreases substantially, with complete results in Appendix A.3. The magnitude varies considerably across models: InternVL3.5 loses about 25.0 percentage points, whereas Qwen3-VL loses about 6.1 points. Across all nine models, mean accuracy decreases from 74.5% to 55.3%, a 19.2-point gap versus 27.4 points under *Crop*. Thus, a blank but increasingly large background accounts for approximately 70% of the average *Crop* gap, suggesting that background extent alone is a major source of recognition loss in most cases.

2.2 Target-Fixed Interventions Localize the Dominant Background Effect

The results above show that recognition worsens as background increases, but do not reveal where this interference arises. We therefore construct increasing background conditions inside the model.

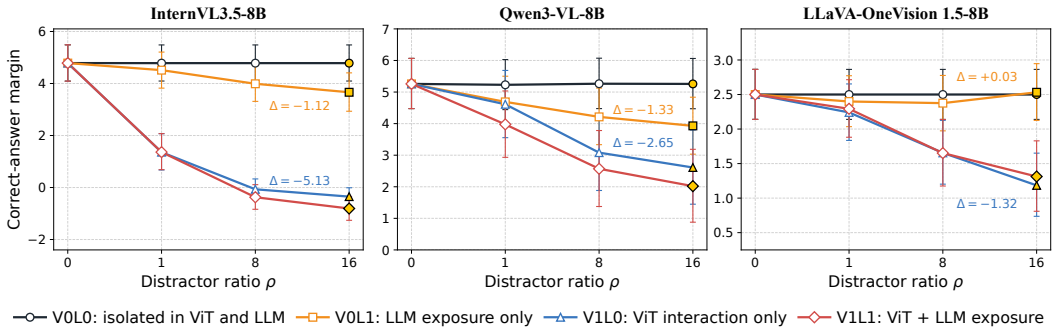

Figure 3: Target-fixed localization of background interference. $V1L0 - V0L0$ isolates background interaction in the ViT, while $V0L1 - V0L0$ isolates LLM exposure to background tokens.

By independently controlling background interaction in the vision encoder and LLM, we test whether recognition failure persists without direct LLM exposure to background tokens.

Setup. Figure 1 (a) illustrates the target-fixed intervention. Using the annotated target region, we separate target tokens T from distracting background tokens D. We keep the target-token embeddings fixed and increase the background-to-target ratio to $\rho = N_D/N_T \in \{1, 8, 16\}$, where N_T and N_D are the corresponding token counts. To distinguish where background exerts the stronger effect inside the model, we control whether target tokens interact with background in the vision encoder and whether text tokens read background visual tokens in the LLM. The two independent interventions yield four conditions: (i) $V0L0$ blocks background influence at both stages and serves as the reference. (ii) $V1L0$ enables background interaction only in the ViT, while (iii) $V0L1$ enables the LLM to read background tokens. (iv) $V1L1$ enables both pathways and is closest to the normal condition. We measure recognition using the correct-answer margin $M = z_{y^\star} - \max_{c \neq y^\star} z_c$, where $y^\star$ is the correct answer, c indexes the candidate answers, and z_c is the aggregated option score defined in Equation 8 of Appendix A.1. Comparing the correct-answer margin in each condition with the $V0L0$ reference at the same ρ measures the recognition effect of the enabled background pathway.

Background interference originates largely in the vision encoder. We use Qwen3-VL as our primary case study and InternVL3.5 and LLaVA-OneVision-1.5 for auxiliary validation. We first compare V1L0 with V0L0, with direct LLM background attention disabled in both. Figure 3 uses V0L0 as the baseline, which remains stable at $\rho = 1, 8, 16$. Compared with the matched V0L0 reference, the V1L0 margin decreases by 0.62, 2.18, and 2.65 points at $\rho = 1, 8, 16$, respectively. Since only V1L0 permits background interaction within the ViT, this growing gap isolates the recognition loss caused by background altering target representations and shows that this effect grows with background. We next compare this effect with direct LLM background exposure under V0L1. Although the two effects are similar at $\rho = 1$, V0L1 declines more slowly as background increases. At $\rho = 16$, the ViT effect reaches 2.65 points, nearly twice the 1.33-point LLM effect. We further test this phenomenon across a broader set of models and find that the substantial recognition loss caused by background interaction in the ViT remains consistent across architectures (Appendix A.15). Overall, these results separate background interference effects and show that background-induced changes to target representations have the stronger effect on recognition.

3 How Background Alter Target Structure in Vision Encoder?

Section 2.2 shows that background harms recognition primarily by altering target representations during visual encoding. We therefore examine how background changes target representations in Section 3.1, and which of these changes matter for recognition in Section 3.2.

3.1 Attention Fragility under Background Interaction

We study ViT encoders, where target and background tokens interact through stacked softmax self-attention blocks. To identify how background alters target representations, we trace target-to-target attention across depth and intervene by removing background values, restoring target attention mass, and varying the restoration layers. Together, these interventions separate background value writing

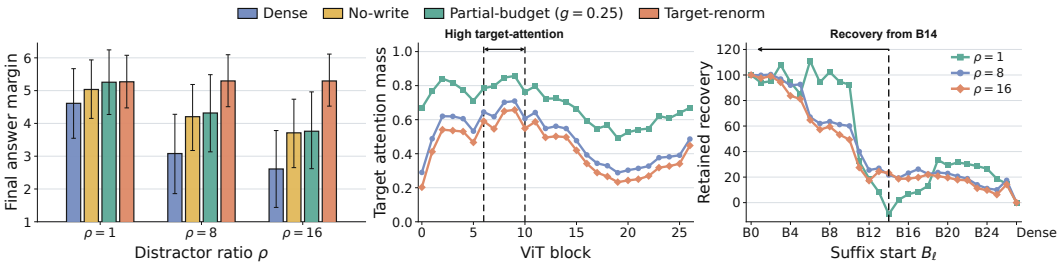

Figure 4: Left: final answer margins under four attention interventions across ρ. Middle: target-to-target attention mass at each ViT block. Right: recovery retained when *Target-renorm* starts at B_ℓ and continues through the final block, relative to its all-block recovery.

from attention-budget dilution and localize the blocks in which their effects become consequential for recognition, linking changes in visual attention allocation to their downstream behavioral effects.

Recognition recovery tracks restoration of target attention mass. To determine how background interaction disrupts target computation, we ask whether recognition loss is more closely associated with direct background value contributions or with reduced attention mass on target keys. We design controlled interventions to compare these two mechanisms. For target query i and head h, let a_{ij}^h be the original attention weight and v_j^h the value vector at key position j. We define:

$$t_i^h = \sum_{j \in T} a_{ij}^h v_j^h, \quad d_i^h = \sum_{j \in D} a_{ij}^h v_j^h, \quad m_{i,T}^h = \sum_{j \in T} a_{ij}^h, \quad m_{i,D}^h = \sum_{j \in D} a_{ij}^h, \tag{1}$$

For the original *Dense* attention, since $m_{i,T}^h + m_{i,D}^h = 1$, the unmodified *Dense* output is $t_i^h + d_i^h$. We compare it with three interventions that vary how much of the background contribution is retained and how the remaining target attention mass is normalized:

$$\begin{aligned} &\textit{No-write}: \quad \tilde{o}_{i,\mathrm{NW}}^h = \alpha_{\mathrm{NM},i}^h t_i^h, \qquad \alpha_{\mathrm{NM},i}^h = \frac{\|t_i^h + d_i^h\|_2}{\|t_i^h\|_2}, \\ &\textit{Partial-budget}(g): \quad \tilde{o}_i^h = \frac{t_i^h + g\, d_i^h}{m_{i,T}^h + g\, m_{i,D}^h}, \qquad \textit{Target-renorm}: \quad \tilde{o}_i^h = \frac{t_i^h}{m_{i,T}^h}. \end{aligned} \tag{2}$$

No-write removes the background values while preserving the original head-output norm but does not renormalize attention over target keys. *Partial-budget*(g) retains a fraction g of the background contribution and renormalizes the remaining attention mass, interpolating between *Dense* and *Target-renorm*. Lower g shifts more of the normalized allocation to target keys. *Target-renorm* removes background from the attention pool and normalizes the target weights to sum to one, assigning the full attention allocation to target keys. More details are in Appendix B.2.

Figure 4 (left) shows the same ordering across distractor ratios: *Dense* has the lowest margin, followed by *No-write* and *Partial-budget*, while *Target-renorm* performs best. Its advantage grows with background and reaches 1.56 points over the norm-matched *No-write* condition at $\rho = 16$. The continuous sweep in Appendix B.3 further shows that recognition improves smoothly as g decreases and more attention mass is restored to target keys. Together, these results show that recognition recovers as attention mass returns to target-internal computation through normalization aligned with the original queries and heads. This motivates us to localize the ViT blocks in which background competition over attention becomes most consequential for recognition.

Small losses of early target attention cause large recognition disruption. We next ask whether background harms recognition by capturing most of the attention from target queries, and at which ViT blocks its critical effect begins to emerge. As shown in the middle panel of Figure 4, attention from target queries to target keys rises sharply in the initial blocks, remains high through the early and middle blocks, and declines in later blocks as more attention shifts to background. If recognition damage were determined solely by the amount of attention diverted at each depth, restoring later blocks should produce the greatest recovery. However, the suffix intervention in the right panel shows limited recovery when *Target-renorm* begins in the middle or later blocks. Recovery begins to increase sharply when the intervention is extended back to around B14. Returning to the middle panel, target attention remains high from the early blocks and begins a sustained decline after B14. At $\rho = 16$, for example, target tokens constitute only $1/17$ of the visual sequence, yet receive an

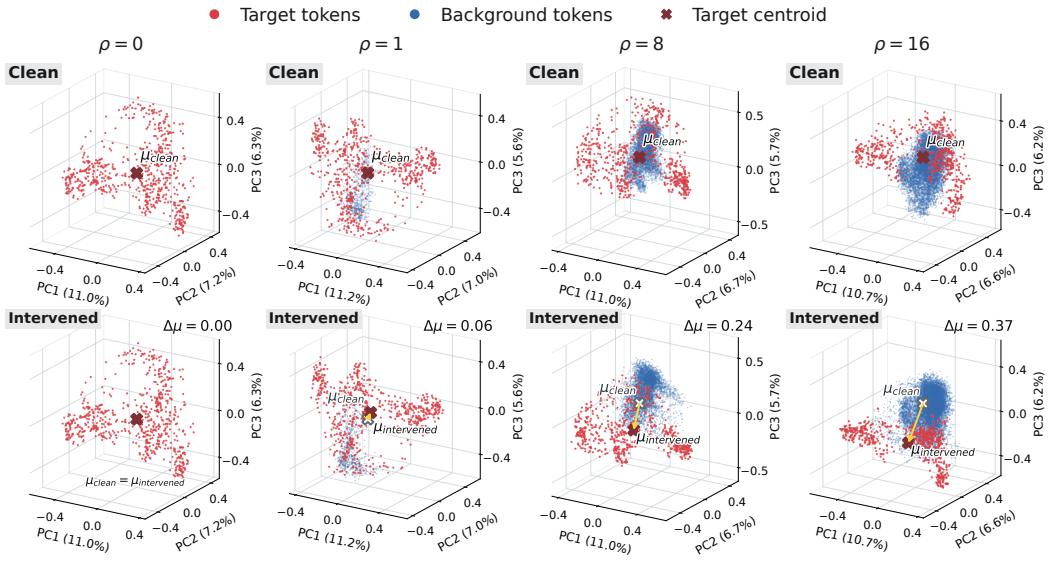

Figure 5: **Background reorganizes target representations.** Rows compare clean and intervened states; points mark target tokens • and background tokens •, while × marks the target centroid.

average attention mass of 56.4% across B6–B14. This high-attention interval overlaps the region where extending suffix restoration produces the largest increase in recognition recovery. Target tokens therefore retain most of the attention in this critical interval, while the remaining background share is sufficient to erode target-internal computation as recognizable information is formed. We call this mechanism *early-layer attention fragility* during target representation formation.

3.2 Recognition-Relevant Components of Target Representation

We next trace how the geometry of target representations changes across ViT blocks. We decompose this change into a shared centroid shift and changes in the relative arrangement of target tokens, then repair the components separately to determine how their contributions vary with depth.

Background produces distinct drifts in target geometry. For the $n_T^{(\ell)}$ matched target states $h_{i,c}^{(\ell)}$ at layer ℓ, we compare V0L0 ($c = 0$) and V1L0 ($c = 1$) by decomposing each representation into two complementary components, a shared mean and token-specific deviations:

$$\mu_c^{(\ell)} = \frac{1}{n_T^{(\ell)}} \sum_{i=1}^{n_T^{(\ell)}} h_{i,c}^{(\ell)}, \qquad \delta_{i,c}^{(\ell)} = h_{i,c}^{(\ell)} - \mu_c^{(\ell)}. \tag{3}$$

Centering removes shifts shared by all target tokens, and the remaining deviations describe how individual token states differ from their common mean. We quantify centroid drift using cosine distance between the two target means, and centered-token drift using the mean cosine distance between corresponding centered token states (Appendix B.11). Both metrics measure directional deviation from the background-isolated reference.

Figure 5 contrasts clean target representations (top row) with those affected by background interaction (bottom row). As ρ increases, the target tokens are progressively reorganized relative to the clean condition. Specifically, larger ρ shifts the target centroid farther from its clean position and makes the relative arrangement among target tokens increasingly different from the matched clean condition. Figure 6 quantifies both effects across ViT depth. Both increase with ρ, but follow distinct trajectories. Centroid drift peaks around B9–B14 and then decreases, indicating partial recovery of the average target state. Centered-token drift instead continues to increase and remains large near the output. Thus, the target mean can partially recover while the token-

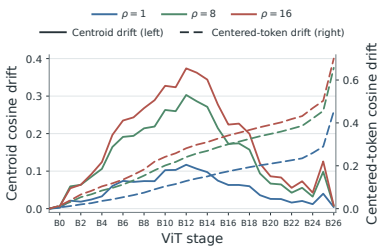

Figure 6: Centroid and centered-token drift across ViT layers.

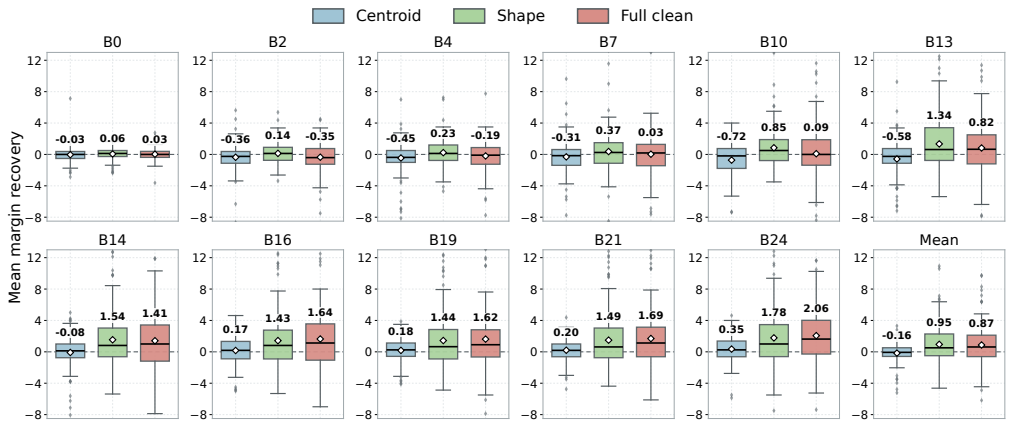

Figure 7: Correct-answer margin gain from centroid, centered-state (Shape), and full-state repair at selected ViT blocks, with the mean across the full ViT depth during target recognition.

specific organization remains persistently altered. We test this distinction through component-wise repair at matched ViT blocks, directly comparing their causal contributions to the answer margin.

Repairing centered-token drift produces most of the recovery gain. Starting from V1L0, we restore the target centroid, centered target states, or full target state from the matched V0L0 run. Figure 7 shows how the correct-answer margin gains from centered-state and full-state repair vary across ViT depth. Each repair here replaces the target state at one block's output. The main repairs produce little margin change in early blocks, begin to increase the margin around B13–B14, and yield substantially larger gains in later blocks. The emergence of positive margin gains in middle blocks suggests that attention erosion is no longer a dominant source of damage in later blocks, allowing a mid-layer repair to persist downstream. When margin gains are averaged across ViT blocks, centered-state repair captures most of the full-state gain and yields a larger gain than centroid repair. This result identifies the centered component as a major contributor to the correct-answer margin gain produced by full-state repair under strong background competition.

4 How Does the LLM React to Background Interference?

Prior work has primarily examined whether the LLM can locate and retrieve target information among background visual tokens (Zhang et al., 2025a; Liu et al., 2026). Building on our preceding analysis, we revisit this inquiry from a different perspective on background interference. In Section 4.1, we ask whether the LLM can still identify target information as background grows. We then use layerwise repair in Section 4.2 to establish a causal link between upstream damage to target representations and its downstream effects on answer formation, completing the causal chain.

4.1 LLM Compensation under Growing Background

Following evidence-ranking (Liu et al., 2026) and attention-based localization analyses (Zhang et al., 2025a), we measure target localization under growing background. For each attention head, AUROC of the answer query's pre-softmax QK scores measures target–background ranking (0.5 indicates chance). We also define the average per-token QK advantage of target set T over background set D:

$$\Delta_{\mathrm{QK}} = \log \frac{\frac{1}{N_T}\sum_{i\in T}\exp(q^\top k_i/\sqrt{d_h})}{\frac{1}{N_D}\sum_{j\in D}\exp(q^\top k_j/\sqrt{d_h})}, \tag{4}$$

where N_T and N_D count target and background tokens. Positive Δ_{QK} indicates greater per-token preference for targets than background. Let $\rho = N_D/N_T$, and let m_T and m_V denote the attention mass on target and all visual tokens, respectively:

$$\frac{m_T}{m_V} = \frac{1}{1+\rho\exp(-\Delta_{\mathrm{QK}})} = \sigma(\Delta_{\mathrm{QK}} - \log\rho)\,. \tag{5}$$

Equation 5 gives the target share within the answer position's visual attention: $\log\rho$ is the penalty from background token count, so Δ_{QK} must increase by the same amount to keep this share un-

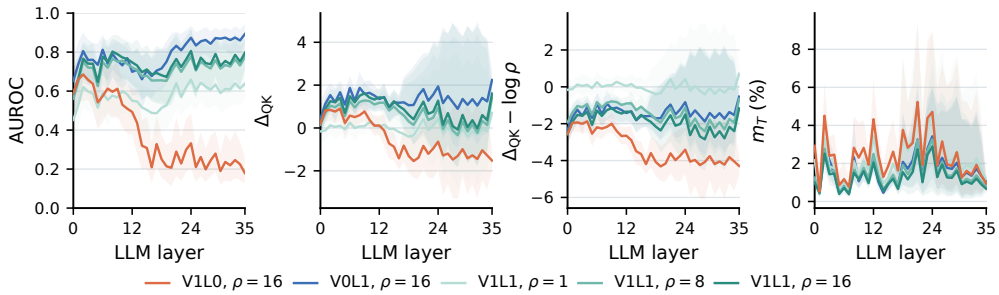

Figure 8: Target ranking (AUROC), per-token preference (Δ_{QK}), attention log-odds ($\Delta_{\mathrm{QK}} - \log \rho$), and target attention (m_T), left to right; red V1L0 uses pre-mask QK scores in panels 1–3.

changed. Because the answer also attends to text tokens, m_T is the absolute attention mass assigned to targets, obtained by multiplying this visual share by the total visual mass m_V.

The LLM compensates yet ViT degradation weakens this response. Figure 8 separates four aspects of target selection at the answer position. Under V1L1, layer-averaged AUROC increases from 0.5774 to 0.7128 and 0.7364 as ρ grows from 1 to 8 and 16, and Δ_{QK} increases accordingly (first two panels, green lines). As background grows, the LLM therefore increases its selectivity toward target tokens, exhibiting a compensatory localization response. Appendix C.10 further reveals adaptive query computation along the downstream decision pathway. This stronger preference nevertheless cannot overcome the growing background population. Because $\log \rho$ increases faster than Δ_{QK}, the target-attention log odds $\Delta_{\mathrm{QK}} - \log \rho$ and absolute target attention m_T both decrease (right two panels). Compensation thus improves target selection without preserving the attention ultimately received by the target. To isolate direct LLM competition, we compare V1L0 with V1L1 at $\rho = 16$. They share the same ViT representation, while only V1L1 exposes the answer-generation process to background visual tokens. Removing this competition increases m_T, showing that background tokens directly dilute target attention. Finally, we compare V0L1 with V1L1 at $\rho = 16$, holding background token count and LLM access fixed. Enabling ViT background interaction reduces AUROC from 0.7871 to 0.7364 and also lowers Δ_{QK}, the attention log-odds, and m_T. Upstream degradation therefore weakens the LLM's compensatory localization.

4.2 Recognition Recovery Through Visual and Text States

ViT-induced damage remains causal in early visual states. Having examined the localization problem above, we now turn to how the LLM forms its decision. We trace the decision pathway to identify where ViT-induced degradation and direct background competition within the LLM affect recognition. At each layer, we repair target visual states in V1L1 with matched V0L1 states. The conditions share the same background-token count and LLM access, differing only in ViT target–background interaction. Correct-answer margin recovery therefore isolates the effect of ViT-induced degradation while retaining direct LLM background competition.

The repair produces substantial gains in early layers, particularly L0–L14, but its effect declines sharply around L15–L20 (Figure 9, first panel), identifying early visual states as the primary causal intervention window. Yet the largest representational differences occur at L27–L35, where repair yields almost no recovery (second panel), showing that state distance is not a reliable proxy for causal importance. Since answer formation also involves text states, we next test whether they carry the two background effects and causally contribute to recognition.

Both background effects reach text states. We repair text states layerwise using two matched donors, each isolating one pathway while retaining the V1L1 recipient (protocols in Appendix C.5). The V0L1 $\rightarrow$ V1L1 repair holds the background-token count and LLM background access fixed, but replaces text states with those formed from visual representations computed without target–background interaction in the ViT. If text states did not incorporate the differing visual information, repairing them should produce little change. Instead, recovery at L18–L23 nearly matches the early visual-state recovery (Figure 9, third panel), showing that intermediate text states carry ViT-derived information that causally supports recognition. The complementary V1L0 $\rightarrow$ V1L1 repair holds the ViT output and token layout fixed, but supplies text states formed without direct LLM access to background tokens. Its gains concentrate at L17–L22 (fourth panel), showing that direct background

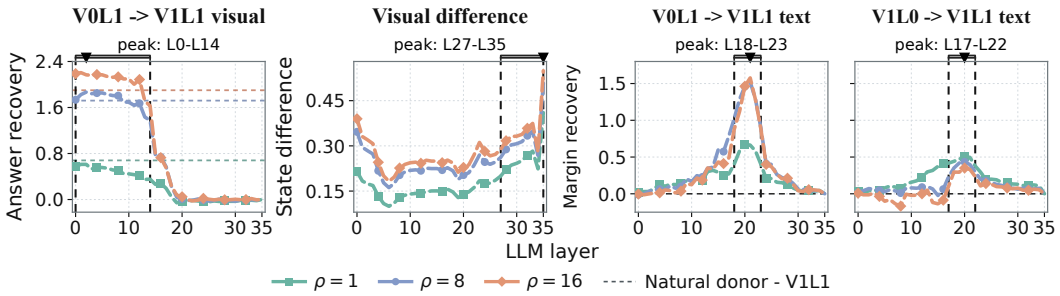

Figure 9: Early visual-state and intermediate text-state repair improve recognition most. Peak intervals indicate layers with recovery at least 50% of the maximum, or state difference in panel 2.

exposure also perturbs text states that causally support recognition. These overlapping recovery windows localize where the two background effects converge: although they arise at different stages, both affect intermediate text representations during answer formation. The larger gains from the V0L1 donor further indicate that, in this setting, degradation inherited from the ViT remains the stronger constraint after visual information has entered the text stream.

5 Related Work

Fine-grained recognition in high-resolution VLMs. High-resolution VLMs preserve fine information through patch-based detailed views (Li et al., 2024b; Guo et al., 2024; Yao et al., 2024). High-resolution pretraining strengthens these representations (Shi et al., 2025). Yet small targets remain difficult to recognize in cluttered scenes (Zhang et al., 2025b; Wang et al., 2025a). This shifts the problem from pixel retention to target location. Search and cropping revisit likely target regions (Wu & Xie, 2024; Zhang et al., 2025a). Patch aggregation and dynamic proposals consolidate regional evidence (Wang et al., 2025c; Jiang et al., 2025a). Hierarchical perception–reasoning methods combine local and global views to refine visual evidence (Yi et al., 2026). Complementary approaches reduce repeated inference through self-distilled RoI prediction (Shi et al., 2026) or region-to-image distillation (Wei et al., 2026). These studies mainly frame background interference as hindering the LLM from locating and retrieving the target. We ask whether it can instead change a fixed target representation during visual encoding, before the LLM retrieves it.

Mechanistic interpretability of VLMs. Recent mechanistic studies increasingly suggest that target recognition in VLMs cannot be explained by localization alone. VLMs may attend to the correct region while still producing incorrect answers (Zhang et al., 2025a; Liu et al., 2026), showing that successful localization does not guarantee reliable recognition. Analyses of specialized attention heads further reveal structured mechanisms for visual grounding, perception, and object-level processing (Kang et al., 2025a; Bi et al., 2025; Golovanevsky et al., 2024; Schaumlöffel et al., 2026). Beyond this initial selection, visual information is progressively integrated, transformed, and routed through the language backbone before contributing to the final prediction (Kaduri et al., 2025; Zhang et al., 2025c; Yin et al., 2025; Jiang et al., 2025b; Xi et al., 2026). Together, these studies explain how visual information is selected and subsequently used for recognition. Our work differs by treating the target representation itself as context-dependent: we test how increasing background alters a fixed target during visual encoding, and how this altered representation together with direct background exposure in the LLM shapes the final decision.

6 Conclusion

Our interventions separate two effects of growing background: harder target retrieval in the LLM and altered target representations in the ViT, with the latter causing greater recognition failure. Early attention fragility reorganizes target structure, and this disruption reaches the answer despite compensatory localization. This makes target preservation a central challenge for high-resolution VLMs: models must scale visual context without sacrificing the integrity of target representations. Future work should develop encoders with stronger target preservation and adaptive attention, and test degradation before retrieval in long videos, multiple images, and embodied multimodal systems.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and polishing of the manuscript. The authors remain responsible for all scientific content, including the claims, mathematical descriptions, experimental results, and citations reported in this paper.

ETHICS STATEMENT

This work investigates background-induced recognition failures in vision-language models through benchmark evaluation and controlled model interventions. Its intended contribution is a better understanding of visual reliability, not a guarantee of safe deployment. Improved fine-grained perception could benefit visual analysis but also enable misuse in surveillance or other privacy-sensitive applications. The reported findings do not establish robustness or fairness beyond the evaluated settings; applications involving sensitive imagery require assessment of privacy, bias, and failure risks.

REPRODUCIBILITY STATEMENT

The main text defines the target-fixed ViT–LLM conditions, attention interventions, representation decomposition, and layerwise state-repair procedures in Sections 2.2–4.2. Appendix A.1 documents the benchmark composition and evaluation protocol, Appendix A.2 lists the evaluated model suite, and Appendices A.3–A.5 specify the Crop, Whiteground, Downsample, and Fixed-Canvas image-space controls. Appendix A.8 provides a roadmap for the target-fixed factorial, and Appendix A.9 specifies the cohort, target and background units, fixed carrier, and V/L switches. Appendices A.10-A.13 document the cross-architecture contract and the model-specific construction and validation details for the Qwen3-VL, LLaVA-OneVision-1.5, and InternVL3.5 implementations; Appendix A.14 gives the additional-checkpoint adaptations. Geometry metrics and repair cohorts are in Appendices B.11 and B.13; attention-intervention scope, the background-gate sweep, and matched controls are provided in Appendix B.2-Appendix B.6. Behavioral interventions use paired image–question comparisons, with the outcome definition and experiment-specific bootstrap procedures reported in Appendix A.1. Cross-architecture stage-localization results are reported in Appendix A.15, with background-source controls in Appendix A.16.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. doi: 10.48550/arXiv.2404.14219.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. doi: 10.48550/arXiv.2509.23661.

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*, 2025a. doi: 10.48550/arXiv.2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. doi: 10.48550/arXiv.2502.13923.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4135–4144. IEEE, 2025.

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. *arXiv preprint arXiv:2503.01773*, 2025.

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*, 2024. doi: 10.48550/arXiv.2412.05271.

Michal Golovanevsky, William Rudman, Vedant Palit, Ritambhara Singh, and Carsten Eickhoff. What do vlms notice? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. *arXiv preprint arXiv:2406.16320*, 2024.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. LLaVA-UHD: An LMM perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pp. 390–406. Springer, 2024.

Yitong Jiang, Jinwei Gu, Tianfan Xue, Ka Chun Cheung, Pavlo Molchanov, Hongxu Yin, and Sifei Liu. Token-efficient VLM: High-resolution image understanding via dynamic region proposal. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24147–24158. IEEE, 2025a.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014. IEEE, 2025b.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image$f$ a deep-dive into the vision of vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558. IEEE, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9339–9350. IEEE, 2025a.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025b.

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*, 2024a. doi: 10.48550/arXiv.2408.03326.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26763–26773. IEEE, 2024b.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5334–5342, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

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.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774.

Timothy Schaumlöffel, Martina G Vilas, and Gemma Roig. Mechanisms of object localization in vision-language models. *arXiv preprint arXiv:2605.19792*, 2026.

Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, et al. Scaling vision pre-training to 4K resolution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9631–9640. IEEE, 2025.

Yuheng Shi, Xiaohuan Pei, Minjing Dong, and Chang Xu. Catching the details: Self-distilled RoI predictors for fine-grained MLLM perception. In *International Conference on Learning Representations*, volume 2026, pp. 136133–136155, 2026.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336. IEEE, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. doi: 10.48550/arXiv.2508.18265.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025c.

Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. Panda: A gigapixel-level human-centric video dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3265–3275. IEEE, 2020.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

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.

Gongli Xi, Ye Tian, Mengyu Yang, Huahui Yi, Liang Lin, Xiaoshuai Hao, Kun Wang, and Wendong Wang. Large vision-language models get lost in attention. *arXiv preprint arXiv:2605.05668*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024. doi: 10.48550/arXiv.2408.01800.

Xiaoyang Yi, Jing Chen, Li Peng, Yuru Bao, and Jian Zhang. Beyond the panorama: Training-free hierarchical perception-reasoning for fine-grained vision in MLLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30534–30549, 2026.

Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in MLLMs: Unlocking pathways to faster inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9382–9391, 2025.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025a.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025b.

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, 2025c.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX CONTENTS

A Background Effects and Stage Localization (Section 2) 15
A.1 Dataset, Evaluation Protocol, and Notation 15
A.2 Complete Cross-Model and Image-Space Results 16
A.3 Complete Cross-Model Crop and Whiteground Results 16
A.4 Absolute Target Pixel Sensitivity Under Downsampling 17
A.5 Fixed-Canvas Context Restoration 18
A.6 Visual Density Has a Stronger Effect Than Semantic Density 19
A.7 Native Image Preprocessing Across Models 20
A.8 Target-Fixed ViT and LLM Factorial Design 20
A.9 Shared Factorial Protocol Across Architectures 21
A.10 Architecture-Specific Designs 22
A.11 Qwen3-VL Reference Design 22
A.12 LLaVA-OneVision-1.5 Replication 23
A.13 InternVL3.5 Architecture-Adapted Replication 23
A.14 Extended Architecture-Adapted Replications 24
A.15 Stage Localization Results Across Architectures 24
A.16 Background Source Factorial Controls 29
A.17 Supplementary Full-Image Repair Check 31

B Target Representation in the ViT (Section 3) 31
B.1 Direction of Visual Interaction 31
B.2 Background Value Control and Target Attention Restoration 32
B.3 Continuous Target Attention Reallocation 33
B.4 Norm-Matched No-write 33
B.5 Target-Renormalization Structure Controls 35
B.6 Prefix and Suffix Target Attention Restoration 35
B.7 Model Generalization 37
B.8 Distributed ViT Mechanisms Across Heads and Frequencies 46
B.9 Distributed Head Effects Across ViT Depth 46
B.10 Structured Effects Across RoPE Frequencies 47
B.11 Target Geometry: Drift, Relational Structure, and Subspace Trajectories 48
B.12 A Multi-View Decomposition of Target-Token Shape Changes 48
B.13 Clean Target-State Repair 51
B.14 Exploratory Local Attention Repair 52

C Target Localization and Recognition in the LLM (Section 4) 52
C.1 Target Localization and Attention Allocation 52
C.2 Answer-Position Target Localization AUROC 53
C.3 Per-Token Selectivity and Absolute Target Attention 54
C.4 Answer-Stage Retrieval and Downstream Readout 55
C.5 Recognition Recovery Through Visual and Text States 57
C.6 Isolating Direct LLM Background Effects in Text States 58
C.7 Do High Background Text States Adapt or Degrade? 58
C.8 Two-Hop Relay from Target Tokens to the Answer 60
C.9 Full Q/K/V Factorial Analysis 63
C.10 Evidence of Damage-Specific Query Compensation 64
C.11 Spectral Decomposition of LLM Target States 66

A Background Effects and Stage Localization (Section 2)

A.1 Dataset, Evaluation Protocol, and Notation

Dataset composition. The benchmark in Section 2.1 contains 700 high-resolution images with annotated target bounding boxes, drawn from ZoomBench (Wei et al., 2026), V*Bench (Wu & Xie, 2024), and HRBench (Wang et al., 2025c). The HRBench portion is drawn primarily from its FSP subset. Table 1 summarizes the source composition of our evaluation set. For the subset of ZoomBench images lacking target bounding-box annotations, the missing annotations were provided by the authors of this paper. For each of the 700 visual question answering examples, we construct 26 controlled presentations: the original image, four Crop conditions, four Whiteground conditions, eleven Downsample conditions, and six Fixed-Canvas Context Restoration conditions. The resulting evaluation contains 18,200 image-question instances. The images have average dimensions of 2946.1×2228.7 pixels, and the mean target bounding-box area is 1.9716% of the image area. These statistics characterize a setting in which a single small target occupies a limited portion of a high-resolution image, motivating our analysis of recognition under extensive surrounding context and the mechanisms by which background interferes with fine-grained target evidence during high-resolution visual processing and downstream answer formation.

Table 1: Source composition.

Source	Images	Share (%)
ZoomBench	485	69.29
V*Bench	115	16.43
HRBench	100	14.29
Total	700	100.00

Image-space option-score evaluation. All nine models in the 700-example image-space evaluation use the same option-score rule. Under each model's tokenizer, let $\mathcal{T}_c$ contain all vocabulary tokens whose decoded text, after stripping surrounding whitespace, is exactly the candidate letter $c \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$. Free-generation strings are retained only for diagnostics and are not parsed for accuracy; in particular, no invalid output is mapped to A. We aggregate the variants at the answer-prediction position as

$$z_c = \log \sum_{t \in \mathcal{T}_c} \exp(\ell_t), \qquad \widehat{y} = \arg\max_c z_c, \tag{6}$$

where ℓ_t is the pre-softmax score of token variant t. Thus, z_c is an aggregated option score and is not necessarily a single vocabulary-token logit. The equivalent implementation sums the corresponding token probabilities in FP32 before taking the argmax.

For example s in condition c, accuracy is

$$a_{s,c} = \mathbf{1}[\widehat{y}_{s,c} = y_s^\star], \qquad \widehat{\mathrm{Acc}}_c = \frac{100}{n} \sum_{s=1}^{n} a_{s,c}. \tag{7}$$

Here $n = 700$, and all image-space comparisons are paired over the same example IDs. The same option-score evaluation is applied across all nine models and all image-space conditions.

Answer-letter scoring for target-fixed experiments. For each candidate letter $c \in \{A, B, C, D\}$, let $\mathcal{T}_c$ contain its accepted single-token spellings and let ℓ_t be the vocabulary logit of token t at the answer-prediction position. We define the option score and prediction by

$$z_c = \log \sum_{t \in \mathcal{T}_c} \exp(\ell_t), \qquad \widehat{y} = \arg\max_{c \in \{A,B,C,D\}} z_c, \qquad M = z_{y^\star} - \max_{c \neq y^\star} z_c. \tag{8}$$

The score z_c aggregates all accepted spellings of option c. Accuracy, cohort selection, and margin calculation use this same option-level prediction.

Paired comparisons and bootstrap inference. Outcomes from the same image and question remain paired across intervention conditions and competition levels. Unless stated otherwise, confidence intervals resample examples as paired clusters. The image-space and nine-model analyses use 2,000 bootstrap replicates, and the stage-localization and mechanistic analyses use 10,000.

A.2 Complete Cross-Model and Image-Space Results

We extend the image-space comparisons in Section 2.1 across models, then separate changes in natural background content from changes in canvas size and target pixel support. The evaluated open-source models include Qwen3-VL (Bai et al., 2025a), Qwen2.5-VL (Bai et al., 2025b), InternVL3.5 (Wang et al., 2025b), InternVL2.5 (Chen et al., 2024), LLaVA-OneVision-1.5 (An et al., 2025), LLaVA-NeXT (LLaVA-v1.6) at 7B and 13B (Liu et al., 2024), MiniCPM-V (Yao et al., 2024), and Phi-3.5-Vision (Abdin et al., 2024).

A.3 Complete Cross-Model Crop and Whiteground Results

Tables 2 and 3 report the per-model Crop and Whiteground accuracies at $1\times$, $3\times$, $6\times$, $18\times$, and the original-image extent used as the shared full-context endpoint. At each matched scale, the two conditions use the same canvas dimensions and identical target pixels, so a given model applies the same geometric preprocessing to the target region. Their difference therefore provides a useful within-model comparison of the interference associated with natural background content.

Crop expansion degrades recognition across all architectures. The unweighted nine-model mean decreases from 74.5% at the tight $1\times$ crop to 47.0% on the original image, corresponding to a 27.4-percentage-point loss. Every model declines between these endpoints, with losses ranging from 13.7 points for Qwen2.5-VL to 37.4 points for Phi-3.5-Vision. InternVL3.5 and InternVL2.5 lose 31.0 and 29.7 points, while LLaVA-v1.6-13B and LLaVA-v1.6-7B lose 32.7 and 29.3 points. Thus, retaining more natural image context consistently reduces recognition across diverse VLMs.

Whiteground losses vary substantially across architectures. The unweighted nine-model mean decreases from 74.5% at the tight $1\times$ condition to 55.3% on the full canvas, corresponding to a 19.2-percentage-point loss. Endpoint losses range from 4.1 points for Qwen2.5-VL and 6.1 points for Qwen3-VL to 34.1 points for Phi-3.5-Vision. InternVL3.5 and InternVL2.5 also show large losses of 25.0 and 23.3 points, while LLaVA-OneVision-1.5-8B loses 14.4 points. Figure 10 reports these endpoint losses. The trajectories also differ in when the loss emerges. Qwen3-VL and Qwen2.5-VL retain nearly all of their tight-crop accuracy through $18\times$, whereas InternVL, LLaVA-v1.6, MiniCPM-V, and Phi-3.5-Vision show larger declines as the blank canvas expands. The shared endpoint loss despite the absence of natural background content identifies canvas extent as a broadly relevant source of interference.

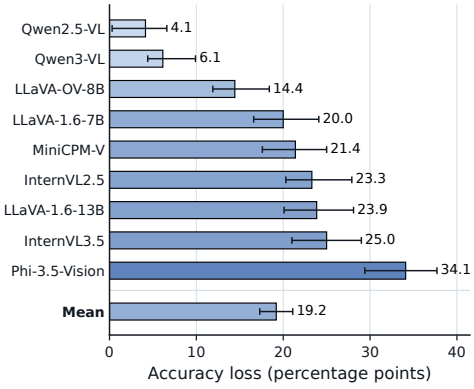

Figure 10: Whiteground accuracy loss.

Crop

Model	$1\times$	$3\times$	$6\times$	$18\times$	Original
Qwen3-VL	81.9	76.7	75.0	68.1	58.3
Qwen2.5-VL	72.4	74.0	73.6	68.1	58.7
InternVL3.5	80.9	78.0	77.3	69.0	49.9
InternVL2.5	81.7	78.0	74.6	67.3	52.0
LLaVA-OneVision-1.5-8B	75.6	73.7	70.1	68.1	55.4
LLaVA-v1.6-13B	69.1	66.0	64.4	55.9	36.4
LLaVA-v1.6-7B	65.0	63.3	59.0	53.6	35.7
MiniCPM-V	75.0	73.1	71.0	63.9	45.6
Phi-3.5-Vision	68.7	63.9	59.7	55.1	31.3
Mean	**74.5**	**71.9**	**69.4**	**63.2**	**47.0**

Table 2: Cross-model Crop accuracy (%) across expansion scales. Darker cells indicate higher accuracy at each scale.

Whiteground

Model	$1\times$	$3\times$	$6\times$	$18\times$	Original
Qwen3-VL	81.9	80.1	79.3	79.0	75.7
Qwen2.5-VL	72.4	73.1	74.9	75.0	68.3
InternVL3.5	80.9	79.4	79.0	75.7	55.9
InternVL2.5	81.7	78.9	76.4	73.9	58.4
LLaVA-OneVision-1.5-8B	75.6	75.1	72.9	72.1	61.1
LLaVA-v1.6-13B	69.1	67.7	67.4	62.1	45.3
LLaVA-v1.6-7B	65.0	65.4	63.4	57.6	45.0
MiniCPM-V	75.0	74.1	72.4	68.9	53.6
Phi-3.5-Vision	68.7	64.6	60.1	53.6	34.6
Mean	**74.5**	**73.2**	**71.8**	**68.7**	**55.3**

Table 3: Cross-model Whiteground accuracy (%) across expansion scales. Darker cells indicate higher accuracy.

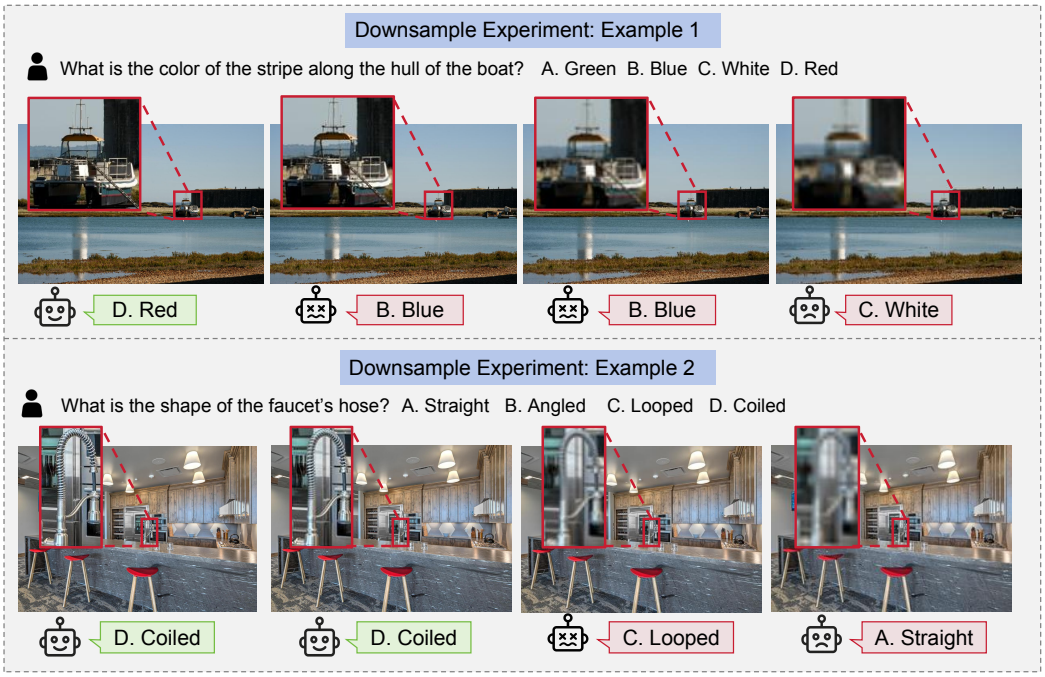

Figure 11: Whole-image downsampling examples. Lower resolution preserves scene composition and relative target size while removing fine target detail; insets show the target regions.

A.4 Absolute Target Pixel Sensitivity Under Downsampling

We further isolate absolute visual evidence by reducing target-pixel support while keeping scene extent, composition, and relative target area fixed. At the controlled image-construction level, this design is complementary to the Crop experiment: Crop changes the target's relative share of the displayed image while preserving its absolute pixels, whereas whole-image Downsample preserves the target's relative share while reducing its absolute pixels. Together, the two experiments separate relative visual density from absolute target-pixel evidence.

On the 700 examples, we downsample the full image at scales from $1\times$ to $0.0625\times$. This preserves the full scene and the target's relative area, while reducing the target bounding box's absolute area in the input pixel grid. Let $P_{s,q}$ be the target pixel area for sample s at scale q, and define the target-pixel reduction measure as $D^{\mathrm{abs}}_{s,q} = -\log_2 P_{s,q}$. We fit $y_{s,q} = \alpha_s + \beta_{\mathrm{abs}} D^{\mathrm{abs}}_{s,q} + \epsilon_{s,q}$, where $y_{s,q} \in \{0,1\}$ is correctness and α_s is a sample fixed effect. We report $\gamma_{\mathrm{abs}} = -100\beta_{\mathrm{abs}}$, the average percentage-point accuracy loss associated with each halving of the target's absolute pixel area. Table 4 reports the complete downsampling trajectories for all nine models together with the corresponding mean target pixel counts.

Downsampling provides controlled variation in absolute target-pixel area. The mean trajectory suggests a moderate-resolution plateau followed by a low-evidence regime. Reducing the linear scale to $1/2$ quarters mean target area from 89,094 to 22,273 pixels, yet mean accuracy falls by only 3.0 points. The next three area quarterings cost 3.9, 4.3, and 2.8 points. Across the full $256\times$ reduction, the observed 14.0-point loss is close to the 14.7 implied by $\gamma_{\mathrm{abs}} = 1.84$, so the fitted slope captures the overall trend but hides where the loss concentrates.

Across models, Qwen3-VL and Qwen2.5-VL have the largest slopes, 3.72 and 3.69 points per halving, lose 26.2 and 26.6 points, and both end at 32.1%. InternVL3.5 retains 38.9%, reversing the original ordering. The smaller LLaVA-v1.6 and Phi-3.5-Vision slopes partly reflect their low 31.3–36.4% baselines and limited room to decline, so γ_{abs} measures within-model pixel reliance rather than standalone robustness. The cross-model range contracts from 27.4 to 10.3 points, showing that severe pixel scarcity compresses architectural differences.

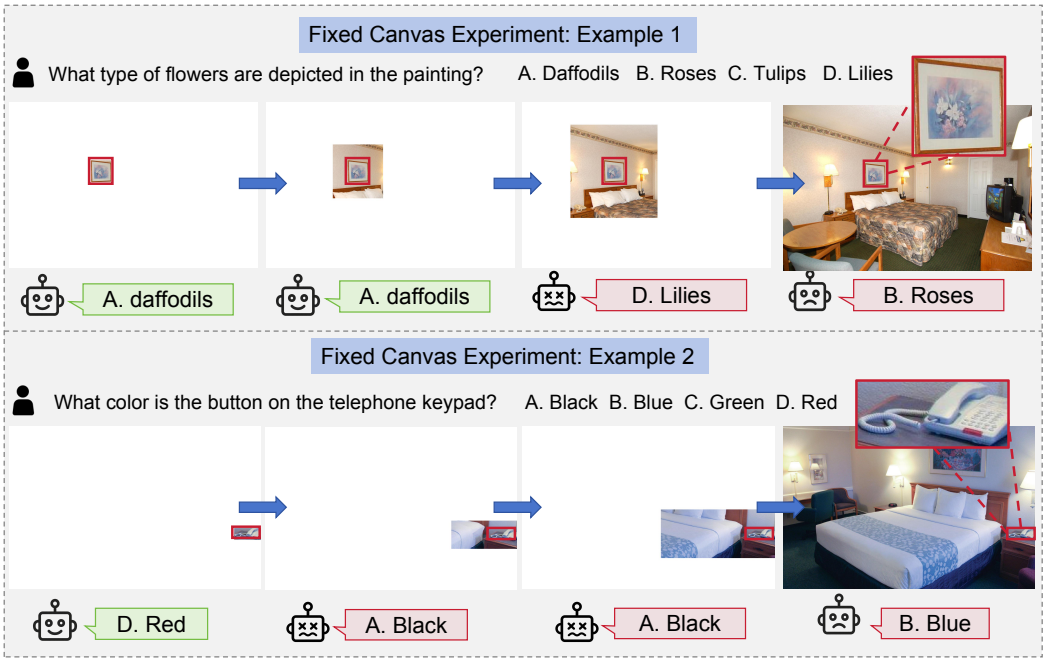

Figure 12: Fixed-Canvas Context Restoration. Natural image content is progressively restored around the target while the original canvas dimensions and target pixels remain fixed; the area outside the restored region is white.

Measure/model	**Orig.**	$\frac{15}{16}$	$\frac{7}{8}$	$\frac{13}{16}$	$\frac{3}{4}$	$\frac{11}{16}$	$\frac{5}{8}$	$\frac{9}{16}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	γ_{abs}
Mean target pixels	89,094	78,295	68,199	58,822	50,119	42,106	34,812	28,196	22,273	5,567	1,394	349	N/A
Qwen3-VL	58.3	58.6	58.3	56.6	56.4	55.6	52.3	51.9	50.7	41.0	33.6	32.1	3.72
Qwen2.5-VL	58.7	59.1	57.0	57.1	52.7	54.9	53.3	54.3	47.7	40.4	33.4	32.1	3.69
InternVL3.5	49.9	49.4	50.4	50.6	49.6	49.9	49.1	48.6	49.0	45.6	43.4	38.9	1.37
InternVL2.5	52.0	52.0	51.6	51.3	50.9	50.9	50.1	49.7	48.9	45.4	40.9	34.3	2.13
LLaVA-OneVision-1.5-8B	55.4	54.6	54.1	53.4	53.9	53.7	51.4	51.9	51.3	45.9	39.3	36.6	2.45
LLaVA-v1.6-13B	36.4	36.7	36.7	36.0	36.6	36.6	36.4	36.1	36.6	36.1	34.1	31.7	0.53
LLaVA-v1.6-7B	35.7	35.7	35.6	35.7	35.9	35.7	35.7	35.9	36.0	36.0	32.6	30.7	0.58
MiniCPM-V	45.6	45.9	46.6	45.4	45.7	45.3	44.3	43.9	44.9	39.4	35.4	32.0	1.84
Phi-3.5-Vision	31.3	31.6	30.7	31.6	30.6	30.0	31.4	31.1	31.1	30.7	29.9	28.6	0.28
Mean	**47.0**	**47.1**	**46.8**	**46.4**	**45.8**	**45.8**	**44.9**	**44.8**	**44.0**	**40.1**	**35.8**	**33.0**	**1.84**

Table 4: Cross-model accuracy and absolute target-pixel sensitivity under whole-image downsampling on the 700-example evaluation set. The Orig. column matches the original-image column of Table 2. Darker accuracy cells indicate higher accuracy using the same color scale as that table and Table 5; mean target pixels and γ_{abs} are unshaded.

A.5 Fixed-Canvas Context Restoration

Figure 12 illustrates the Fixed-Canvas Context Restoration construction.

The purpose of this experiment is to isolate the effect of semantic background content while holding the image geometry processed by each model fixed. Every condition uses the original canvas dimensions and identical target pixels; only progressively larger surrounding regions change from white to natural image content. Because the input dimensions are unchanged, each model's native resizing or downsampling applies the same transformation to the target across conditions, controlling changes in target resolution. Table 5 shows that the nine-model mean generally decreases as natural content is restored, from 55.3% at $1\times$ to 47.0% on the original image. Based on the unrounded accuracies, this 8.3-percentage-point decline therefore measures interference from background semantics with canvas extent and target-pixel support held fixed.

Model rankings depend on visual presentation. Tables 2 and 5 show that Crop and Fixed-Canvas Context Restoration produce different model orderings. Under the tight $1\times$ Crop, Qwen3-VL reaches 81.9%, followed by InternVL2.5 at 81.7%, InternVL3.5 at 80.9%, and LLaVA-OneVision-1.5-8B at 75.6%. When the original canvas is retained and natural content is restricted to the $1\times$ region, Qwen3-VL remains highest at 75.7%, followed by Qwen2.5-VL at 68.3%, LLaVA-OneVision-1.5-8B at 61.1%, InternVL2.5 at 58.4%, and InternVL3.5 at 55.9%. The expanded Crop also separates model generations: at $18\times$, Qwen3-VL, Qwen2.5-VL, and LLaVA-OneVision-1.5-8B each reach 68.1%, while LLaVA-v1.6-13B and 7B reach 55.9% and 53.6%. These comparisons show that architecture rankings depend strongly on how target extent and surrounding context are controlled. We next quantify that dependence through density sensitivity.

Model	1×	1.5×	3×	6×	12×	18×	Orig.
Qwen3-VL	75.7	75.0	73.0	70.3	67.6	66.6	58.3
Qwen2.5-VL	68.3	69.6	69.1	68.3	66.4	65.9	58.7
InternVL3.5	55.9	58.0	54.7	53.9	53.4	51.6	49.9
InternVL2.5	58.4	56.9	57.0	57.1	54.7	55.3	52.0
LLaVA-OneVision-1.5-8B	61.1	60.0	61.9	59.1	59.0	58.7	55.4
LLaVA-v1.6-13B	45.3	44.6	43.9	39.9	40.7	39.6	36.4
LLaVA-v1.6-7B	45.0	44.3	39.9	39.1	39.3	39.3	35.7
MiniCPM-V	53.6	54.7	54.4	53.0	53.3	51.4	45.6
Phi-3.5-Vision	34.6	35.1	34.3	32.3	33.4	31.1	31.3
Mean	**55.3**	**55.3**	**54.2**	**52.6**	**52.0**	**51.0**	**47.0**

Table 5: Cross-model accuracy under Fixed-Canvas Context Restoration.

A.6 Visual Density Has a Stronger Effect Than Semantic Density

The Crop and Fixed-Canvas Context Restoration trajectories measure two related sensitivities. Crop changes the target's share of the full visual input, while context restoration changes its share of natural image content without changing the canvas. We express both effects as accuracy loss per halving of target density, which places them on a common geometric scale and allows a direct comparison across models under the shared density parameterization.

Let A_s^{target} denote the annotated target area for sample s. At scale c, $A_{s,c}^{\text{crop}}$ is the full area shown under Crop, and $A_{s,c}^{\text{natural}}$ is the restored natural image area in the fixed-canvas condition. Visual density is $r_{s,c}^{\text{visual}} = A_s^{\text{target}}/A_{s,c}^{\text{crop}}$, which decreases as the crop expands. Semantic density is $r_{s,c}^{\text{semantic}} = A_s^{\text{target}}/A_{s,c}^{\text{natural}}$, which measures the target relative to the restored natural content. We convert either density to $C_{s,c} = -\log_2(r_{s,c})$ and fit $y_{s,c} = \alpha_s + \beta C_{s,c} + \epsilon_{s,c}$, where $y_{s,c} \in \{0, 1\}$ is correctness and α_s captures example difficulty. We report $\gamma = -100\beta$, the average percentage-point accuracy loss for each halving of target density.

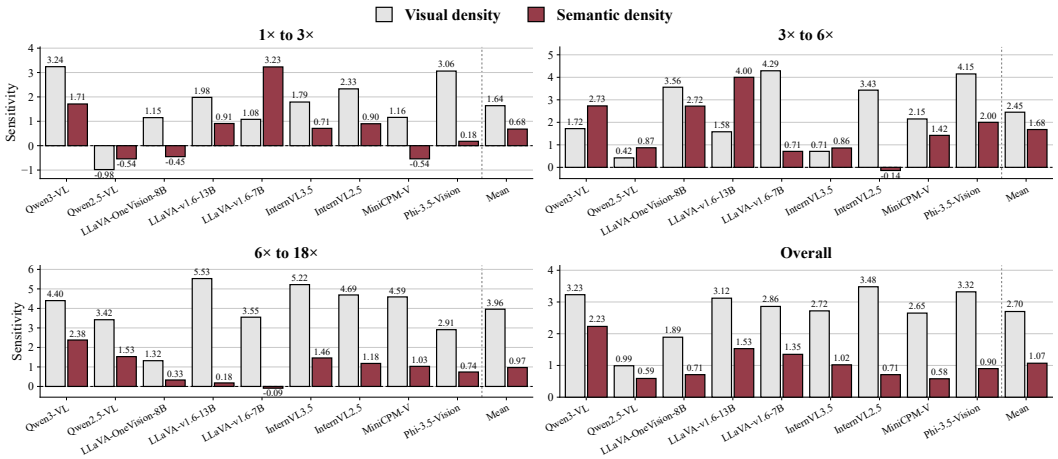

Figure 13: Visual-density and semantic-density sensitivity across the nine evaluated models.

Visual extent dominates overall, with the clearest separation at large context. The Overall panel of Figure 13 shows higher visual-density sensitivity than semantic-density sensitivity for all nine models. The model means are 2.70 and 1.07 percentage points per halving, respectively, so

the mean semantic-density sensitivity is about 40% of the visual-density sensitivity. The separation is largest for InternVL2.5, at approximately 3.48 versus 0.71 points, whereas Qwen2.5-VL retains the smallest gap. Qwen3-VL remains the most responsive to semantic density at approximately 2.23 points, although its visual-density sensitivity is still larger.

The interval panels show that this advantage emerges with context scale rather than holding uniformly over every local range. From $1\times$ to $3\times$, the mean sensitivities are 1.64 points for visual density and 0.68 for semantic density, with local non-monotonicity for Qwen2.5-VL and a larger semantic-density sensitivity than visual-density sensitivity for LLaVA-v1.6-7B (approximately 3.24 versus 1.08 points). From $3\times$ to $6\times$, the means increase to 2.45 and 1.68 points; semantic density is stronger for Qwen3-VL, Qwen2.5-VL, InternVL3.5, and LLaVA-v1.6-13B, showing that added natural content can be consequential at intermediate expansion. The ordering becomes uniform from $6\times$ to $18\times$: visual-density sensitivity exceeds semantic-density sensitivity for every model, and the mean gap widens to 3.96 versus 0.97 points. Recognition is therefore most consistently constrained by overall visual extent once the target occupies a small share of the input, rather than by the amount of natural background content alone. This scale-dependent separation motivates the target-fixed intervention in Appendix A.8, where target inputs are held constant while background competition is varied directly.

A.7 Native Image Preprocessing Across Models

The nine image-space models do not receive a common raster tensor. Table 6 records the native resizing, tiling, pixel or view budgets, and interpolation rules used in the evaluation. Consequently, the same external Crop, Whiteground, Downsample, or Fixed-Canvas presentation can produce different raster sizes, tile counts, and internal visual-token sequences across architectures.

Model	Native preprocessing	Input budget	Interpolation
Qwen3-VL-8B	Aspect ratio preserved; sides aligned to multiples of 32	65,536–16,777,216 pixels	Bicubic
Qwen2.5-VL-7B	Aspect ratio preserved; sides aligned to multiples of 28	3,136–12,845,056 pixels	Bicubic
InternVL3.5-8B	Dynamic 448×448 tiles	Model-native tile budget	Bicubic
InternVL2.5-8B	Dynamic 448×448 tiles	Model-native tile budget	Bicubic
LLaVA-OneVision-1.5-8B	Qwen2-VL resize; sides aligned to multiples of 28	3,136–3,240,000 pixels	Bicubic
LLaVA-v1.6-13B	AnyRes with 336×336 tiles	Up to 4 tiles + global view	Bicubic
LLaVA-v1.6-7B	AnyRes with 336×336 tiles	Up to 4 tiles + global view	Bicubic
MiniCPM-V-2.6	Approx. 448-pixel source + dynamic slices	Source view + up to 9 slices	Bicubic
Phi-3.5-Vision	336×336 grid crops + global view	Up to 4 crops + global view	Bilinear (grid); bicubic (global)

Table 6: Native preprocessing for the nine image-space models. Pixel limits apply to resize-based pipelines; view limits apply to tiled or sliced pipelines.

These heterogeneous native pipelines motivate the two-level evaluation design. The preceding cross-model image-space experiments treat each model as a black box and provide an intuitive validation that controlled image-level changes induce a consistent recognition trend, but they do not make the corresponding internal visual inputs equivalent across architectures. We therefore move inside selected architectures and adapt the target-fixed intervention to each model's native visual units, allowing background competition to be localized without erasing architectural differences.

A.8 Target-Fixed ViT and LLM Factorial Design

Our framework and mechanistic analyses center on Qwen3-VL-8B-Instruct, while the cross-architecture experiments below provide complementary validation of the identified mechanisms under distinct native designs. The target-fixed factorial is documented across the subsections

that follow. Appendix A.9 defines the shared cohort, target and background units, fixed carrier, and V/L switches. Appendix A.10 states the cross-architecture intervention contract, while Appendices A.11–A.13 give the model-specific construction and validation details for Qwen3-VL, LLaVA-OneVision-1.5, and InternVL3.5. Appendix A.14 then records the adaptations used for the additional checkpoints.

A.9 Shared Factorial Protocol Across Architectures

The image-space interventions above establish the recognition effect under controlled changes to the external canvas, but external image size is not itself the quantity processed by the model. Architecture-specific resizing, tiling, patch merging, and visual-token allocation can map the same geometric intervention to different internal visual sequences. We therefore reconstruct target–background competition directly within each model, where background units can be varied while the target input and downstream carrier remain fixed. In this controlled internal setting, we localize the recognition effect to background interaction during visual encoding versus direct background access during LLM processing. The model designs are described after the shared protocol below.

Mechanistic analysis cohort. For the mechanistic analysis, we select 120 examples that the model answers correctly from the target crop. A target crop from a high-resolution image is not necessarily recognizable in isolation, so correctness on the crop establishes that the target itself provides sufficient evidence for the model. The incorrect response on the full canvas then identifies cases in which that recognizable target ceases to support the correct answer when presented with the surrounding background. This selection directly matches the target-fixed question studied below: how recognition changes as background competition increases while the target input remains fixed. Paired resampling measures uncertainty across examples.

Target and background units. An experimental visual unit is the group of native ViT patches that produces one LLM visual token after the model-specific merger or projector. T contains all valid units from the standalone target crop. Background candidates are sampled from the original image outside the target. Across conditions, target patch embeddings, source coordinates, raster order, and native grouping are held fixed. We control background competition through the slot ratio $\rho = \frac{N_D}{N_T}$, where N_T and N_D denote the numbers of target and background visual-token slots presented downstream after the model-specific merger or projector. They are distinct from the stage-specific native token counts $n_T^{(\ell)}$ and $n_D^{(\ell)}$. In the Qwen reference design, for example, each downstream slot merges a 2×2 group of native patches, so the ViT contains $n_T^{\text{patch}} = 4N_T$ target patch tokens and $n_D^{\text{patch}} = 4N_D$ active background patch tokens. We evaluate $\rho \in \{0, 1, 8, 16\}$. At $\rho = 0$, no background slot is active; for each positive ratio, we construct $N_D = \rho N_T$ background slots from an ordered background pool $\mathcal{P}$. Thus, ρ controls competition in the visual token sequence. Appendix A.16 reports how the effect changes across controlled background source variants.

Fixed maximum carrier. For each example s, we construct the maximum downstream carrier at $\rho = 16$ and reuse it across all background ratios and all four V/L conditions. Let $\mathcal{C}_s^{\max}$ contain the input token IDs $\mathbf{x}$, total visual-placeholder count N_{vis}, sequence length L, position IDs $\mathbf{P}$, text positions $\mathcal{I}_{\text{text}}$, and answer position i_{ans}. For every $\rho \in \{0, 1, 8, 16\}$ and $(v, l) \in \{0, 1\}^2$, the following carrier identity holds across the shared layout and inactive capacity:

$$\begin{aligned}(\mathbf{x}, N_{\text{vis}}, L, \mathbf{P}, \mathcal{I}_{\text{text}}, i_{\text{ans}})_{s,\rho,v,l} &= \mathcal{C}_s^{\max},\\ N_D^{\text{active}}(\rho) &= \rho N_T,\\ (\mathbf{e}_u, m_u) &= (\mathbf{0}, 0), \qquad u \in U_\rho.\end{aligned} \tag{9}$$

Here, U_ρ denotes unused carrier slots, $\mathbf{e}_u$ their embeddings, and m_u their attention-mask entries. Only the number of active background slots and the specified ViT or LLM attention connections change. Unused slots remain in the carrier with zero embeddings and inactive attention masks.

The reference two-dimensional carrier preserves the target rectangle, including its raster order and row boundaries, at the same slot coordinates across all factorial conditions and background ratios. Row-end padding remains present but inactive, active background slots are filled consecutively below the target block, and all remaining capacity is zeroed and attention-masked. Thus, varying ρ changes background occupancy while keeping the target and overall sequence layout fixed.

ViT and LLM factorial switches. The term "target fixed" refers to the ViT input boundary: target patch embeddings, source positions, raster order, and native grouping are identical across conditions. Target hidden states remain free to change with context, allowing the experiment to measure how those changes propagate downstream. We independently control two background pathways. In the ViT, $v = 0$ blocks attention between target and background and $v = 1$ permits bidirectional attention. In the LLM, $l = 0$ blocks text access to background visual slots and $l = 1$ permits it; target slots remain accessible. The four conditions are V0L0 (isolated reference), V0L1 (LLM exposure), V1L0 (ViT interaction), and V1L1 (both). Under V1L0, background reaches the language model through target states contextualized inside the ViT.

Let $M_{vl}^{\mathrm{VL}}(\rho)$ denote the correct-answer margin under switches v, l, using Equation 8. For each matched example and background ratio, we define

$$\Delta_V = M_{10}^{\mathrm{VL}} - M_{00}^{\mathrm{VL}}, \quad \Delta_L = M_{01}^{\mathrm{VL}} - M_{00}^{\mathrm{VL}}, \quad \Delta_{\mathrm{joint}} = M_{11}^{\mathrm{VL}} - M_{00}^{\mathrm{VL}} \tag{10}$$

Negative contrasts indicate recognition loss relative to the isolated reference.

A.10 Architecture-Specific Designs

The protocol defines the intervention variables while preserving each model's native visual units, positional encodings, CLS structure, merger or projector, and decoder coordinates. The fixed maximum carrier remains shared across architectures.

A.11 Qwen3-VL Reference Design

Native visual path. Each experimental unit contains a native 2×2 patch group that becomes one LLM visual token through the spatial merger. Target groups retain crop raster order and background groups retain their original-image source order. Each patch keeps its native learned additive two-dimensional position embedding and native two-dimensional rotary encoding. Target patches retain crop-source encodings, while background patches retain original-image-source encodings. No patch is remapped to an artificial strip coordinate system.

V0 and V1 call the same visual execution function. For active patches $i, j \in T \cup D$, their only connectivity difference is specified by the following source-connectivity rule:

$$A_{ij}^{\mathrm{ViT}}(v) = \begin{cases} 1, & i, j \in T \text{ or } i, j \in D, \\ v, & i \in T,\ j \in D \text{ or } i \in D,\ j \in T. \end{cases} \tag{11}$$

V0 processes target and background as separate attention segments, whereas V1 places them in one segment. The main spatial merger and every DeepStack merger process T and D separately, so no merger group crosses the source boundary.

Decoder coordinates. The downstream carrier preserves the crop's two-dimensional target layout. Each target row and column maps to the same row and column in the carrier, unused row-end slots remain masked, and background slots begin below the target rectangle. Target M-RoPE coordinates are checked token by token against the standalone crop. Background tokens receive the M-RoPE coordinates associated with their carrier slots. These decoder coordinates are distinct from the source coordinates retained inside the ViT.

Paired input and position verification. The verification covers all selected examples at $\rho \in \{1, 8, 16\}$, giving matched V0L1 and V1L1 pairs. Let $\mathcal{F}_{vl}^{s,\rho}$ collect the input tokens, LLM attention mask, position IDs, and target, background, and answer indices. Let $\mathcal{Z}_T$ contain the target patch inputs, ViT source positions, ViT rotary encodings, and decoder M-RoPE coordinates. The following equality audit is applied to every matched (s, ρ) pair:

$$\mathcal{F}_{01}^{s,\rho} = \mathcal{F}_{11}^{s,\rho}, \qquad \max_{t \in T, \mathbf{z} \in \mathcal{Z}_T} \left\| \mathbf{z}_{t,01}^{s,\rho} - \mathbf{z}_{t,11}^{s,\rho} \right\|_\infty = 0. \tag{12}$$

The first equality confirms identical carrier and decoder inputs within each matched pair, and the second confirms identical target positional quantities. Q/K states and attention scores can still differ because the same positional encodings are applied to context-dependent hidden states. The comparison is performed within each background ratio.

Switch validation. The isolated reference remains approximately stable as ρ changes and is insensitive to distractor identity, permutation, and cross-sample replacement. The $\mathrm{V0L1}$ and $\mathrm{V0L0}$ conditions have identical target states at the ViT output. Under $\mathrm{V1L0}$, background visual positions are inaccessible to the LLM, so any background effect must reach the language model through target-side computation. Together, these checks confirm that the V and L switches isolate the intended background pathways at their designated stages.

Zero background anchor. The stage-localization trajectories use a shared $\rho = 0$ measurement with the same fixed carrier as the positive-ratio conditions and no active background slots. On the selected cohort, this anchor has mean margin 5.2583 (95% CI $[4.4771, 6.0646]$) and accuracy 0.9417 (95% CI $[0.9000, 0.9833]$). The four V/L labels coincide algebraically at $\rho = 0$, so the plot repeats one measurement to anchor all four trajectories. Architecture-adapted estimates below provide the cross-model replications.

A.12 LLaVA-OneVision-1.5 Replication

We reproduce the V/L factorial in LLaVA-OneVision-1.5 (An et al., 2025) while preserving RICE-ViT's visual operations and the decoder position system.

Native visual units and source positions. Each experimental unit is a native 2×2 RICE-ViT patch group that produces one LLM visual token through the merger. RICE-ViT uses native two-dimensional rotary encodings but has no additive ViT position embedding and no Qwen DeepStack. Target patches retain crop-source rotary encodings, while background patches retain original-image-source encodings. No additive position term is applied to the hidden states.

Shared V0/V1 visual path and CLS handling. Both conditions initialize the sequence as $[\mathrm{CLS}_T, T, \mathrm{CLS}_D, D]$ and use the same explicit RICE-ViT QKV/SDPA execution path. V0 isolates the two source segments. V1 adds only bidirectional target-patch/background-patch edges. Neither CLS token attends across sources. Target and background groups pass separately through the native merger. V0 target outputs match the target-only path numerically.

Carrier construction and decoder positions. The carrier retains the Qwen-style two-dimensional slot layout consisting of a target rectangle, masked row-end padding, background slots below the target, and masked remaining capacity. Slots are never compacted. Carrier construction supports up to 16,384 visual placeholders while leaving preprocessing of the crop and original image unchanged.

The LLaVA decoder uses one-dimensional RoPE. The primary carrier uses the natural scalar positions of the full sequence, including padding slots. Target and text coordinates remain fixed across factorial conditions and may differ from the standalone crop prompt.

Scoring and cohort. The LLaVA replication evaluates the selected cohort. Candidate screening and factorial evaluation use the option score in Equation 8 for both prediction and margin.

A.13 InternVL3.5 Architecture-Adapted Replication

We reproduce the same target-fixed factorial with InternVL3.5 while preserving the checkpoint weights, dynamic tiling, per-tile CLS structure, learned positions, pixel shuffle, projector, and decoder position system. Background ratio is defined by the number of occupied downstream slots.

Adapted visual construction. The target contains the tiles produced by native preprocessing of the standalone crop, while background units are selected from far-background locations in the original image's native tiles. Tile order, within-tile order, and learned tile-local positions are retained. We execute this shared construction under two controlled attention layouts: V0 isolates the target and background segments, whereas V1 permits cross-segment patch interaction. The two conditions otherwise share the same visual operations and parameters, and the V0 target output matches the native target extraction path.

Carrier and decoder positions. InternVL uses the one-dimensional placeholder sequence

$$[T][D_{16}][\text{text}],$$

with inactive background capacity retained under an inactive mask. Its decoder uses single-plane one-dimensional RoPE. Under the primary carrier, target tokens occupy the initial visual positions and retain their target-only scalar coordinates. The text follows the full carrier with monotonically increasing positions. The fixed-carrier and V/L invariants are the same as in the Qwen reference implementation, while decoder coordinates follow InternVL's native position system.

A.14 Extended Architecture-Adapted Replications

We extend the same V/L factorial to 12 additional checkpoints: Gemma3-4B-it (Team et al., 2025); InternVL2.5 at 1B, 2B, 4B, and 8B; InternVL3-8B (Zhu et al., 2025); Qwen2-VL-7B-Instruct (Wang et al., 2024); Qwen2.5-VL-3B- and 7B-Instruct; Qwen3-VL-2B- and 4B-Instruct; and Qwen3-VL-8B-Thinking. Each checkpoint is evaluated on the selected cohort with 10,000 paired bootstrap resamples.

Shared intervention contract. For every checkpoint, T comes from native preprocessing of the target crop, and D follows raster order through far background groups in the original image. Ratios $\rho = 1, 8, 16$ are nested prefixes of this sequence. Thus, ρ counts visual units after the merger or projector. At each ratio, all four V/L conditions share the maximum carrier, target inputs, target and text positions, and inactive slots. The displayed $\rho = 0$ point uses this same carrier with no active background.

Qwen2, Qwen2.5, and Qwen3 adaptations. Qwen2-VL retains its 32-block global-attention ViT, two-dimensional visual rotary encoding, 2×2 merger, and decoder M-RoPE. Qwen2.5-VL preserves its mixture of local windows and global blocks at depths 7, 15, 23, and 31. Within a local block, V1 adds cross-segment access while retaining the native window, and suffix restoration returns target queries to that window. The Qwen3 checkpoints preserve learned visual positions, two-dimensional rotary encoding, the merger, and all three DeepStack branches. Across these models, V0 and V1 share all model operations and differ only in permitted cross-segment attention.

InternVL2.5 and InternVL3 adaptations. The InternVL checkpoints retain dynamic 448×448 tiling, a local CLS token for each tile, learned positions, LayerScale, pixel shuffle, and the native projector. Target patches continue to read their own tile and CLS; V1 exposes background patches while target CLS queries remain local. After CLS removal, each tile produces 256 LLM visual units for the fixed one-dimensional carrier.

Gemma3 adaptation. Gemma3 uses a 27-block SigLIP encoder at 896×896, learned visual positions, 4×4 average pooling, RMS normalization, and its multimodal projector. The fixed carrier contains one target image block and 16 background image blocks with their native boundaries. The language model retains its mixture of global and sliding-window layers, and L1 exposes background keys according to that native visibility pattern.

A.15 Stage Localization Results Across Architectures

Qwen3-VL reference result. Table 7 reports the four factorial conditions. V0L0 remains nearly constant in aggregate across distractor ratios. Because each ratio is evaluated in a separate BF16 forward with a different packed visual-sequence length. Averaging the positive ratios $\rho \in \{1, 8, 16\}$ gives pooled effects of $\overline{\Delta}_V = -1.815$ and $\overline{\Delta}_L = -0.973$; the shared $\rho = 0$ anchor contributes no factorial contrast. Thus, when both context pathways are blocked, changing the number of active background slots has almost no effect on the baseline. The losses appear when either background pathway is enabled, with a larger pooled effect from ViT interaction than from direct LLM exposure. Recognition also declines under V1L0, where background reaches the LLM only through the contextualized target states.

Extended 12-model result. Figures 14 through 19 report the architecture-adapted factorials from Appendix A.14. Figure 20 additionally shows each model's accuracy drop from V0L0 across distractor ratios, together with the Qwen3-VL-8B-Instruct reference. In every checkpoint, the isolated

ρ	V0L0	V0L1	V1L0	V1L1
0	.942 / 5.258	.942 / 5.258	.942 / 5.258	.942 / 5.258
1	.942 / 5.229	.883 / 4.689	.767 / 4.613	.725 / 3.976
8	.925 / 5.264	.817 / 4.214	.658 / 3.081	.608 / 2.570
16	.933 / 5.257	.775 / 3.929	.608 / 2.610	.567 / 2.020

Table 7: Target-fixed Qwen3-VL factorial results on the fixed cohort and fixed maximum carrier used in Figure 3. At $\rho = 0$, the shared carrier contains no active background slots. Each entry reports accuracy / mean correct-answer margin. Contrasts are computed from unrounded sample means.

V0L0 curve is nearly constant within the fixed carrier, while V1L0 decreases as background slots are activated. At $\rho = 16$, the ViT-interaction change is more negative than the LLM-exposure change in all 12 models. The gap is small in InternVL2.5-2B (-4.67 versus -3.52 points) and substantially larger in Qwen2-VL and Qwen2.5-VL, where LLM-exposure changes are -0.34, -0.13, and -0.11 points and ViT-interaction changes reach -1.76, -1.33, and -1.70 points.

The family-level trajectories clarify how this common localization appears under different native designs. Qwen2 and Qwen2.5 keep the LLM-only curves close to the isolated reference while their ViT-interacting curves separate steadily with background, producing a particularly clean visual-encoder effect. Qwen3-VL-2B- and 4B-Instruct show larger ViT-only losses of -4.40 and -3.35 points, whereas Qwen3-VL-8B-Thinking has a smaller loss of -0.94 points and a direct LLM-exposure gain of 0.32 points. Across the InternVL checkpoints, both pathways can contribute, but the growing separation of V1L0 from V0L0 remains visible despite changes in model scale, tiling, CLS handling, and projector width. Gemma3 provides the strongest example of downstream compensation: direct LLM exposure improves the margin by 2.06 points at $\rho = 16$, alongside a much larger ViT-only loss of -10.09 points.

These differences explain why the natural V1L1 trajectory alone cannot identify where background interference originates. The joint condition combines an upstream loss with a downstream response that may be harmful, weak, or compensatory depending on the architecture. The factorial separates those contributions: the approximately stable V0L0 anchor verifies that slot count alone does not move the outcome, the V1L0 curve measures the consequence of visual interaction carried entirely through target states, and the contrast between V1L0 and V1L1 reveals the model-specific effect of direct background access in the LLM. Across all 12 checkpoints, the first of these effects is larger at $\rho = 16$, making upstream target-representation damage the shared component beneath the diverse joint trajectories.

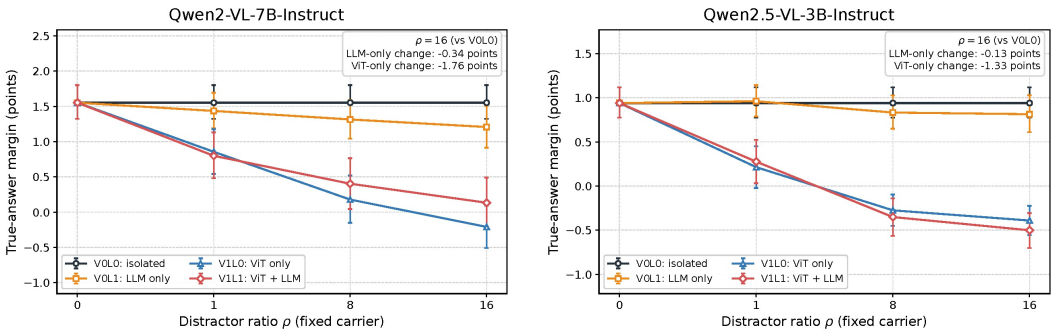

Figure 14: V/L factorials for Qwen2-VL-7B-Instruct (left) and Qwen2.5-VL-3B-Instruct (right). Each model is evaluated on the selected cohort; error bars are pointwise 95% paired bootstrap intervals, and $\rho = 0$ uses the fixed carrier with no active background.

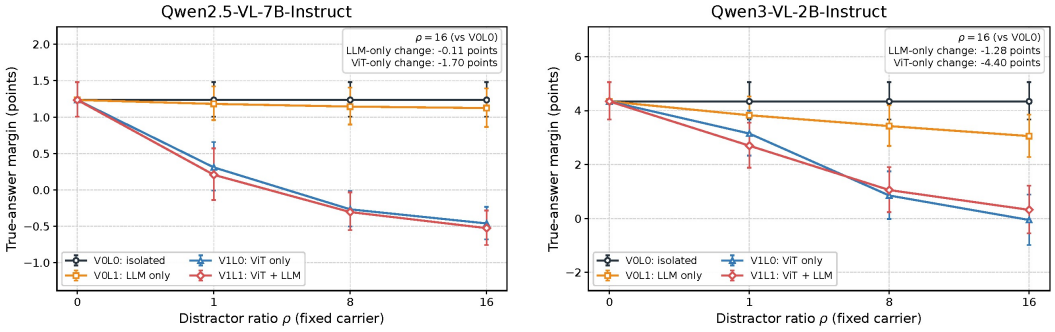

Figure 15: V/L factorials for Qwen2.5-VL-7B-Instruct (left) and Qwen3-VL-2B-Instruct (right). Plotting, cohort, and uncertainty conventions follow Figure 14.

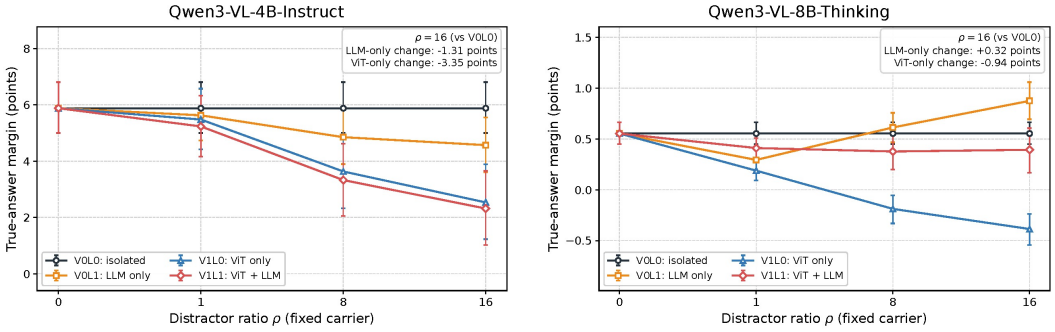

Figure 16: V/L factorials for Qwen3-VL-4B-Instruct (left) and Qwen3-VL-8B-Thinking (right). The Thinking checkpoint is scored from immediate option logits.

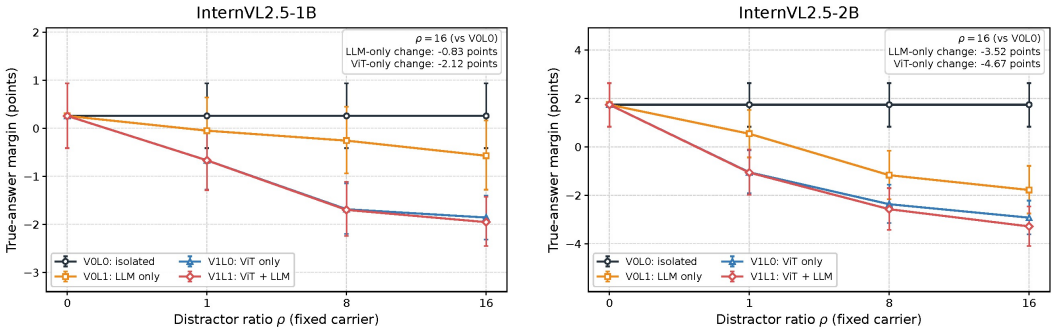

Figure 17: V/L factorials for InternVL2.5-1B (left) and InternVL2.5-2B (right). Here ρ counts background slots after the projector.

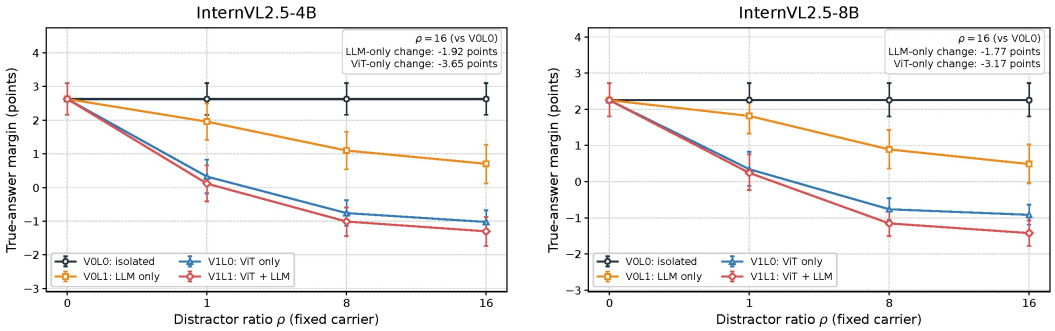

Figure 18: V/L factorials for InternVL2.5-4B (left) and InternVL2.5-8B (right), preserving native tiling, local CLS structure, pixel shuffle, and the projector.

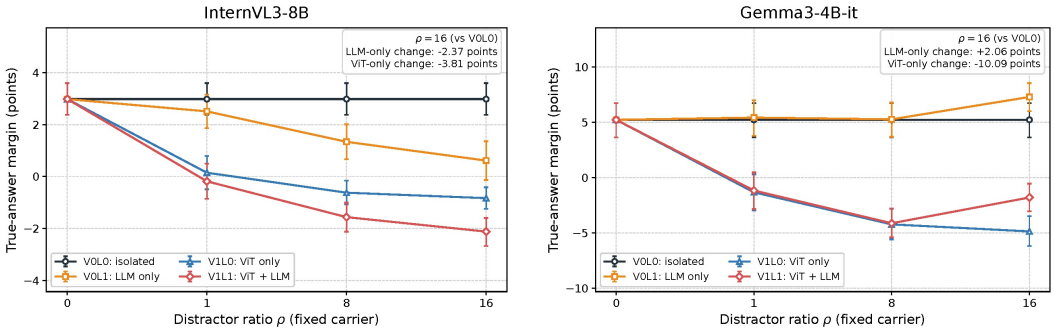

Figure 19: V/L factorials for InternVL3-8B (left) and Gemma3-4B-it (right). Gemma3 preserves its multi-image boundaries and mixed sliding and global language attention; its positive LLM exposure change accompanies a substantially larger ViT interaction loss.

Table 8: Mean margins from the LLaVA-OneVision-1.5 factorial on the selected cohort.

ρ	V0L0	V0L1	V1L0	V1L1
0	2.500	2.500	2.500	2.500
1	2.500	2.399	2.241	2.293
8	2.500	2.375	1.655	1.652
16	2.500	2.532	1.184	1.312

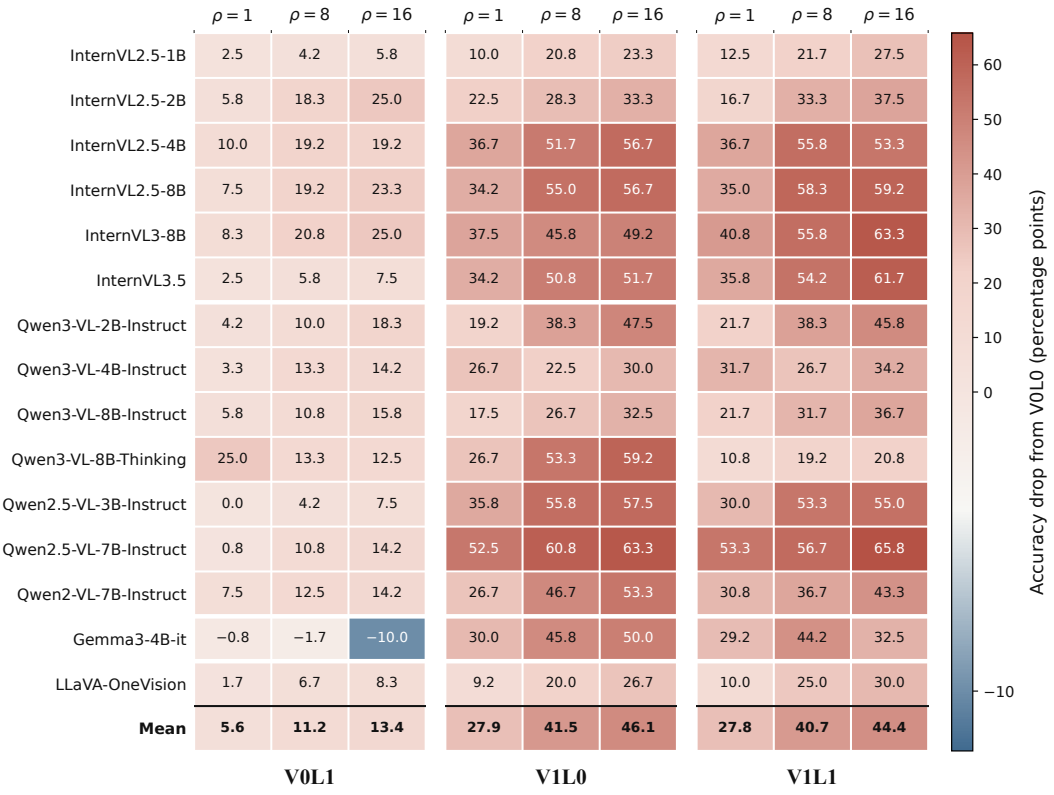

Figure 20: Cross-model target-fixed accuracy drops relative to V0L0 across distractor ratios. Each checkpoint is evaluated on the selected cohort. For $vl \in \{01, 10, 11\}$, the plotted value is $D_{vl}(\rho) = 100[\mathrm{Acc}_{00}(\rho) - \mathrm{Acc}_{vl}(\rho)]$, where accuracy is expressed on a 0–1 scale. Values are in percentage points; positive values indicate an accuracy drop relative to V0L0, whereas negative values indicate an improvement under this sign convention.

LLaVA-OneVision-1.5 replication result. Table 8 reports the LLaVA-OneVision-1.5 factorial on the selected cohort. The isolated margin remains exactly 2.500 across ratios, whereas V1L0 decreases to 1.184 at $\rho = 16$. Direct LLM exposure is small in isolation and mildly compensatory after ViT interaction, raising the joint margin from 1.184 to 1.312 at the largest ratio. The widening gap between the isolated and ViT-interacting conditions therefore reproduces the stage-localization pattern under RICE-ViT and a different decoder position system rather than relying on Qwen-specific visual processing.

InternVL3.5 replication result. The isolated trajectory again remains stable, while the two conditions that permit ViT interaction decline as more background slots become active. Because InternVL retains dynamic tiles, local CLS tokens, pixel shuffle, and its native projector, this replication tests a substantially different route from image patches to LLM visual units. The InternVL3.5 diagnostics in Appendix B.7.3 show that target attention decreases with background

and suffix protection becomes effective in early-to-middle visual depth; Appendix B.7.4 extends this evidence to InternVL2.5 and InternVL3 checkpoints.

Cross-architecture interpretation. Across architectures, the factorial variables remain fixed while each checkpoint retains its visual tokenization, positional encoding, merger or projector, CLS structure, attention schedule, and decoder coordinates. The resulting 15-model suite shows that stage localization recurs across distinct internal designs. The ViT analyses in Appendices B.7.1, B.7.2, B.7.3, and B.7.4 connect these trajectories to target attention and early computation.

A.16 Background Source Factorial Controls

We evaluate stage localization under four background constructions that vary sampling, semantic relation, and low-level appearance: a raster prefix from the same image, the mean over five random samples from the same far-background pool, matched far-background tokens from an unrelated image, and a mean-color control at the same indices and positions. Figures 21 and 22 report the InternVL3.5 and LLaVA-OneVision-1.5 results. Figure 23 reports the corresponding Qwen background-source factorial controls.

ViT-Side Degradation Persists across Background Sources. Across all three models, the isolated V0L0 trajectory remains nearly flat and V0L1 changes comparatively little, whereas conditions that allow target tokens to interact with context inside the ViT separate increasingly as competition grows. The same stage ordering appears for a raster prefix from the source image, randomized samples from the same far-background pool, tokens from an unrelated image, and a mean-color construction. The natural-background conditions generally produce the strongest losses, indicating that structured visual features amplify the interaction, but the persistence of the effect under unrelated and mean-color inputs shows that it does not require semantic agreement with the target scene.

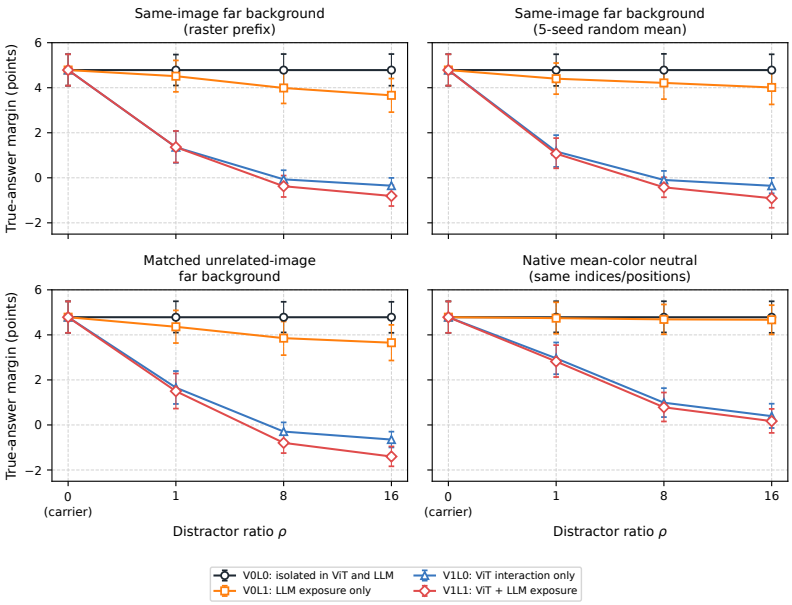

Figure 21: **InternVL3.5 background-source factorial.** Across raster-prefix, randomized same-image, and unrelated-image backgrounds, the isolated V0L0 trajectory remains nearly flat, while both conditions with ViT interaction (V1L0 and V1L1) deteriorate sharply as ρ increases. The mean-color control weakens this degradation, indicating that natural background features amplify the effect while the same stage-localization pattern also appears with semantically unrelated or neutral content.

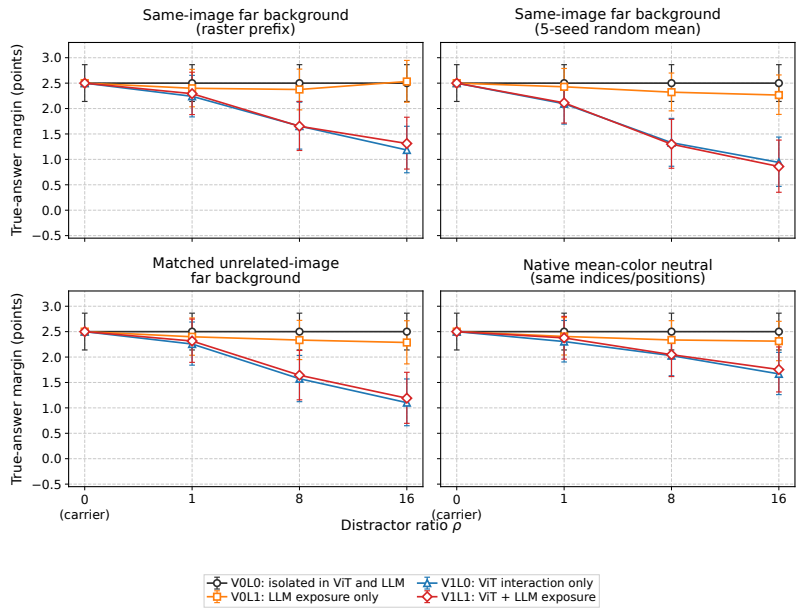

Figure 22: **LLaVA-OneVision-1.5 background-source factorial.** Across all sources, V0L0 and V0L1 remain stable, while ViT interaction drives the decline; V1L1 closely tracks V1L0 even for unrelated-image and mean-color backgrounds.

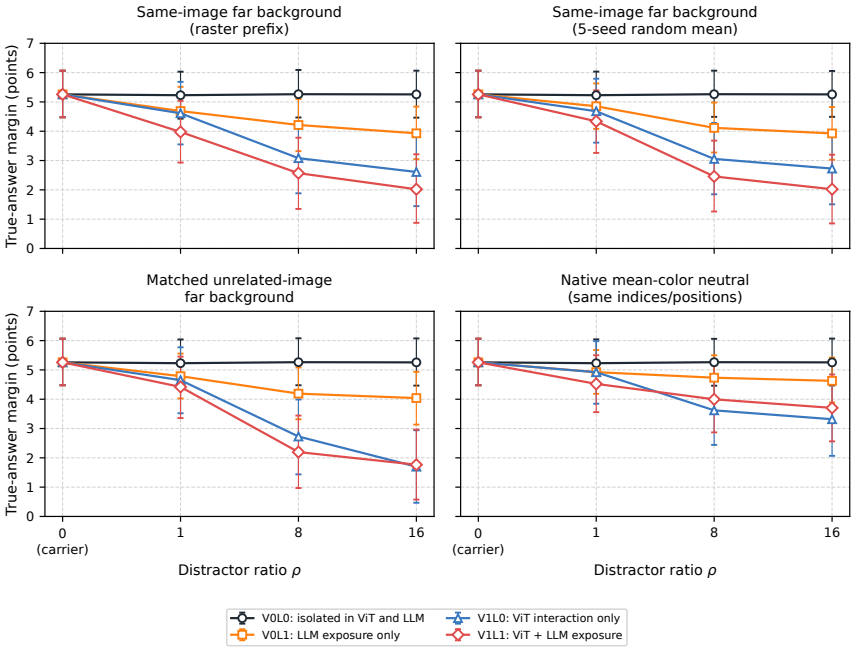

Figure 23: **Qwen3-VL background-source factorial.** ViT interaction consistently degrades performance across sources; additional LLM exposure compounds the loss for natural same-image backgrounds but partially offsets it for mean color.

The downstream response is more architecture-dependent. In LLaVA-OneVision-1.5, V1L1 remains close to V1L0 across the controls, with direct background access mildly compounding or offsetting the ViT-side loss depending on the source. Qwen3-VL likewise changes this or-

dering with background content: natural same-image inputs can compound the upstream loss, while the mean-color condition produces a compensatory joint trajectory. InternVL shows a strong ViT-driven decline under all natural-source variants and a weaker decline for mean color. These source controls therefore reinforce the cross-architecture conclusion: visual contextualization supplies the stable harmful component, while the LLM contribution depends on what background evidence remains available downstream.

A.17 Supplementary Full-Image Repair Check

The target-fixed factorial intentionally uses a controlled carrier so that increasing background does not alter the target input itself. We additionally test whether the same qualitative stage ordering appears on a matched set of real full-image examples with Qwen3-VL-8B-Instruct. Here the model retains its native full-image preprocessing, and the interventions begin only after the resulting visual units enter the ViT. Starting from the natural $\mathrm{V1L1}$ trajectory, $\mathrm{V}_{1\to0}$ blocks target–background interaction inside the ViT, while $\mathrm{L}_{1\to0}$ blocks direct background access during subsequent LLM computation.

This check is not used as the primary causal test because it cannot hold the target fixed before visual encoding. Native preprocessing recomputes resizing, tiling, patchification, and token grouping from the full image, so the same raw target region can acquire a different effective resolution, token count, positional encoding, and merger grouping before the first ViT block. LLM isolation is also less exact because target visual units remain at their native positions rather than preceding all background units; causal ordering therefore does not block preceding background tokens. These interventions are post-preprocessing repairs, not estimates of the strictly target-fixed effect of adding background.

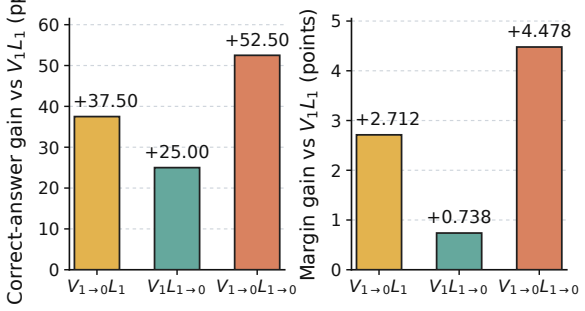

Figure 24: Full-image repair check with Qwen3-VL-8B-Instruct. Panels show accuracy and mean recovery in the correct-answer margin from ViT repair, LLM isolation, and their combination.

Figure 24 nevertheless shows the same directional pattern as the controlled factorial. The natural $\mathrm{V1L1}$ condition answers none of the examples correctly and has a mean correct-answer margin of -4.048. Repairing ViT interaction alone raises accuracy to 37.5% and the margin to -1.336, whereas LLM isolation alone reaches 25.0% and -3.310. Applying both repairs gives the largest recovery, reaching 52.5% accuracy and a positive margin of 0.430. Thus, this supplementary real-image check reproduces the trend that intervention at visual encoding provides the larger single-stage recovery and that additional LLM isolation can further improve recognition, without replacing the controlled target-fixed evidence.

B Target Representation in the ViT (Section 3)

This section develops the ViT mechanism in three steps. We first identify the harmful direction of visual interaction and test how recognition changes as attention returns to target tokens. We then localize this effect across depth, heads, frequencies, and model families. Finally, we connect the attention changes to target geometry and repair the representation components that support recognition.

B.1 Direction of Visual Interaction

On the same examples and fixed-max-carrier protocol as Table 7, we evaluate Dense attention, bidirectional visual isolation, and the two directional masks with $L = 0$ at $\rho \in \{1, 8, 16\}$. For the equivalence audit, bidirectional isolation and T-query$\to$ D-key protection use the same explicit Boolean-mask SDPA implementation at every ViT block. All other computations, including normalization, Q/K/V projection, visual RoPE, softmax, output projection, MLP, and residual updates, are identical;

the two conditions differ only in whether D-queries may access T-keys. Let T and D denote the sets of target and distractor token positions, respectively. For layer ℓ and head h, the attention weights are

$$A_{ij}^{(\ell,h)} = \mathrm{softmax}_j \left(\frac{q_i^{(\ell,h)} {k_j^{(\ell,h)}}^\top}{\sqrt{d_h}} + \mathcal{M}_{ij} \right). \tag{13}$$

For the target-to-distractor intervention, we use

$$\mathcal{M}_{ij}^{T \not\to D} = \begin{cases} -\infty, & i \in T,\ j \in D, \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

whereas the reverse intervention uses

$$\mathcal{M}_{ij}^{D \not\to T} = \begin{cases} -\infty, & i \in D,\ j \in T, \\ 0, & \text{otherwise}. \end{cases} \tag{15}$$

The arrow follows the query-to-key convention. Thus, blocking T-query$\to$ D-key access prevents background value vectors from entering target-token updates. The two interventions remove the same number of cross-region attention edges but in opposite directions, while preserving all within-target and within-background edges. This is an equal-edge directional control; it does not assume that the removed attention mass or attention-output norm is equal across directions. Those quantities are examined separately by the value and norm controls below. Their behavioral difference therefore isolates the direction of the harmful visual interaction.

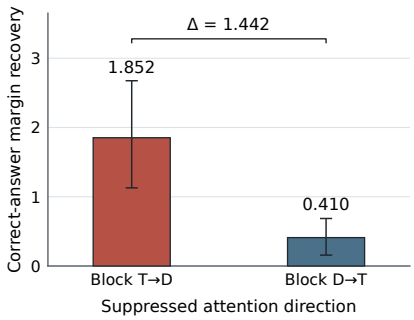

Figure 25: Directional ViT interventions. Bars show pooled margin gains over Dense attention.

Blocking target-query access to background keys in all 27 ViT blocks improves the correct-answer margin by 1.852 points (95% CI [1.128, 2.675]), compared with 0.410 points [0.157, 0.686] for the reverse direction. The resulting directional gap is 1.442 points [0.726, 2.268].

In the unified-kernel audit, bidirectional isolation and T-query$\to$ D-key protection produce bitwise-identical target patch states at the input and after every ViT block, across 10,080 comparisons including the input, with a global maximum absolute difference of 0. Their three DeepStack outputs, final merged target states, option scores, margins, and predictions also match exactly for all 360 sample-ratio pairs. Consequently, the unified full-isolation gain and the $T \to D$ gain are both 1.852 points, corresponding to exactly 100% recovery.

Paired gains are first averaged over ρ within each example and then across examples. Recovery is the ratio of these mean paired gains; per-example ratios are not averaged.

B.2 Background Value Control and Target Attention Restoration

We next trace a sequence of interventions from removing background value contributions to progressively restoring normalized attention toward target tokens. Magnitude and structure controls then characterize the properties of the restored update that support recognition.

The interventions in Equation 2 modify the per-head target-query output within the vision encoder's attention sublayer, before head concatenation and the original output projection (Figure 1). For target query i, let $\widetilde{\mathbf{o}}_i^h$ denote the intervened output of head h. The resulting multi-head attention update is

$$\widetilde{\mathbf{z}}_i = W_O \ \mathrm{Concat}_{h=1}^{H} \left(\widetilde{\mathbf{o}}_i^h \right). \tag{16}$$

The interventions modify target and background contributions within each head while preserving W_O, patch embeddings, and LLM states. The main comparison applies them throughout the ViT.

Throughout the main text, *No-write* removes the background value contribution and preserves the Dense head output norm at the same layer input. We also evaluate *Unscaled No-write* (UNW), which

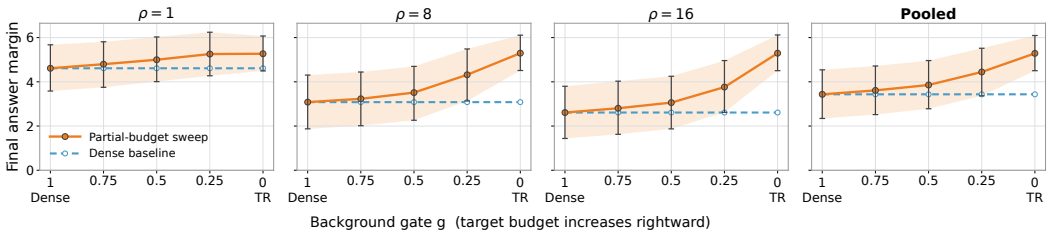

Figure 26: Continuous target-budget reallocation at $\rho = 1, 8, 16$ and in the pooled results. Reducing the background gate g from 1 (Dense) to 0 (Target-renorm) returns attention budget to target-internal interaction. Error bars show 95% sample-cluster bootstrap confidence intervals over the selected examples; dashed lines mark the Dense baseline in each panel.

removes background values and leaves the smaller target contribution unchanged. Comparing the two conditions separates value removal from update magnitude. The gated conditions in Equation 2 test progressive target-attention restoration along a Dense-to-Target-renorm path. For any paired behavioral outcome Y, the endpoint difference can be written as the following sequence of changes:

$$Y_{\mathrm{TR}} - Y_{\mathrm{Dense}} = (Y_{\mathrm{NM}} - Y_{\mathrm{Dense}}) + (Y_{\mathrm{TR}} - Y_{\mathrm{NM}}) = \Delta_{\mathrm{NM-Dense}} + \Delta_{\mathrm{TR-NM}}. \tag{17}$$

The first contrast records the change from Dense to norm-matched background-value removal. The second records the subsequent change from *No-write* to *Target-renorm*, which further reallocates normalized attention toward target tokens through query- and head-specific rescaling. These contrasts describe successive steps along the intervention path. Appendix B.4 provides the direct NM-versus-UNW comparison used to separate value removal from update magnitude.

B.3 Continuous Target Attention Reallocation

To move beyond the Dense and Target-renorm endpoints, we vary the background gate g in Equation 2 continuously from 1 to 0 and intervene on target-query attention outputs throughout the ViT. Figure 26 shows a graded response on the same examples: as g decreases, less normalized attention is spent on background keys, a larger share returns to target-internal interaction, and the correct-answer margin rises. The curves are ordered by competition strength. At $\rho = 1$, the available background consumes less of the target budget and the behavioral gain is modest; at $\rho = 8$ and 16, the margin grows more strongly, particularly near the Target-renorm endpoint. The pooled trajectory follows the same smooth pattern rather than changing only at $g = 0$, which connects recognition continuously to the amount of attention restored to the target. This dose-response pattern supports an attention-budget mechanism, while the norm and structure controls below determine whether the gain can be explained by update magnitude alone.

B.4 Norm-Matched No-write

In Section 3.1, *No-write* refers to Norm-Matched No-write (NM). To test what happens without norm restoration, we additionally compare against Unscaled No-write (UNW), which removes the background value term and leaves the resulting target-query update at its reduced norm. NM preserves the same target-only direction but rescales each target-query and attention-head output to its Dense norm before the shared output projection W_O.

We evaluate Dense, UNW, and NM at $\rho \in \{1, 8, 16\}$ across all 27 ViT blocks. The target, background selection, carrier, positions, masks, weights, and scoring remain shared across conditions. Pooled estimates average ratios within each example.

Intervention definition. At each layer, t_i^h and d_i^h are recomputed from that layer's Q/K/V states, as in Equation 1. The Dense output is $t_i^h + d_i^h$, and NM is the No-write branch defined in Equation 2. UNW uses the following unscaled target-only update:

$$o_{i,\mathrm{UNW}}^h = t_i^h. \tag{18}$$

UNW preserves the remaining target direction, and NM rescales it by $\alpha^h_{\mathrm{NM},i}$ from Equation 2 to match the Dense head norm. Rescaling occurs before head concatenation and W_O.

Norm matching separates background removal from update shrinkage. We recompute Dense, UNW, and NM from identical layer inputs on the NM trajectory and compare head output norms before concatenation, averaged over target queries, heads, layers, and examples. Table 9 shows that NM reproduces the Dense norm at every ratio, while UNW becomes progressively smaller as competition increases. The behavioral ordering follows this distinction: UNW loses margin as the remaining target update shrinks, whereas NM consistently improves over UNW and exceeds Dense at $\rho = 8$ and 16. Removing the background value contribution is therefore useful when the local update magnitude is preserved, and the growing NM to UNW gap at larger ratios shows why simple deletion is an inadequate control. We consequently use NM as the No-write condition in the main comparison.

Branch	$\rho = 1$	$\rho = 8$	$\rho = 16$
Dense	2.60	2.53	2.51
UNW	1.96	1.24	1.05
NM	2.60	2.53	2.51

Table 9: Matched head norms.

Scale assignment must remain query aligned. For target query i and attention head h, Target-renorm uses the scale $\alpha^h_{\mathrm{TR},i}$, while *Scale shuffled* preserves the same scale factors within each sample, layer, and head but permutes their assignment across target queries:

$$\alpha^h_{\mathrm{TR},i} = \frac{1}{m^h_{i,T}}, \qquad o^h_{i,\mathrm{shuffle}} = \alpha^h_{\mathrm{TR},\pi(i)} t^h_i, \tag{19}$$

where $m^h_{i,T}$ is the attention mass assigned to target keys. The target contribution t^h_i is unchanged, and only the scale assignment is permuted before concatenation and W_O. We evaluate this control on the same examples and background ratios, intervening in all ViT blocks.

Random reassignment sharply reduces the margin relative to TR, especially at $\rho = 8$ and 16 (Figure 27). The scale factors are therefore useful because they remain aligned with the queries from which the missing target mass was measured, not simply because their distribution produces larger updates. Since permutation can also alter the norms of individual scaled vectors, the next controls match the resulting residual magnitude token by token and test whether the spatial and sample-specific direction remains necessary.

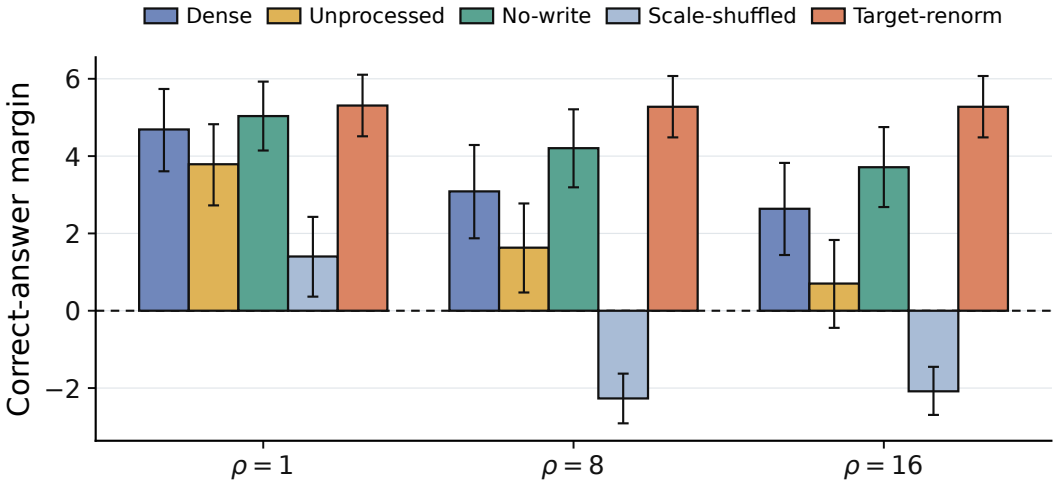

Figure 27: No-write and Target-renorm controls. In the legend, Unprocessed denotes UNW and No-write denotes NM. Scale shuffled preserves the TR scale factors but randomly reassigns them across target queries. TR remains strongest, while random reassignment removes most recovery, especially at larger background ratios. Error bars show 95% paired bootstrap confidence intervals.

B.5 Target-Renormalization Structure Controls

We next test how the target-aligned structure of the restored update contributes beyond its magnitude. We reuse the cohort at $\rho \in \{1, 8, 16\}$ and apply each intervention throughout the ViT. The intervention unit is the post-W_O attention residual written to each target patch token.

Let $R^{\mathrm{TR}}_{s,\rho,\ell,i}$ denote the correct Target-renorm residual for sample s, ratio ρ, ViT layer ℓ, and target token i. For every control candidate $\widehat{R}^{\mathrm{ctrl}}_{s,\rho,\ell,i}$, we match its norm tokenwise:

$$R^{\mathrm{ctrl}}_{s,\rho,\ell,i} = \frac{\left\| R^{\mathrm{TR}}_{s,\rho,\ell,i} \right\|_2}{\left\| \widehat{R}^{\mathrm{ctrl}}_{s,\rho,\ell,i} \right\|_2} \widehat{R}^{\mathrm{ctrl}}_{s,\rho,\ell,i}. \tag{20}$$

All controls therefore share the residual norm of the correct Target-renorm update.

The spatial-permutation control applies a deterministic derangement over the target's 2×2 merger cells, preserving the target-value multiset while breaking its spatial correspondence. The cross-sample control retains the receiver's Q/K states but replaces its target values with same-layer values from a non-self donor having exactly the same target patch and merged-token counts. The orthogonal control uses a deterministic direction that is tokenwise orthogonal to the correct Target-renorm residual before norm matching. Finally, the Dense-gain control preserves the native Dense residual direction while scaling it to the Target-renorm norm. All conditions evolve along their own intervened trajectories across layers.

Recovery requires target-aligned structure. Correct Target-renorm exceeds the spatial permutation, cross sample, and orthogonal controls by 7.051, 7.633, and 7.398 margin points, respectively. It also exceeds the Dense gain control by 2.634 points, so matching the magnitude of the restored update preserves little recovery when its target-aligned direction is lost. The representation results agree with the behavioral contrasts. Correct Target-renorm is nearly V0-like, with centered cosine drift 0.0021 and relative centered Frobenius drift 0.0524, whereas the four norm-matched controls remain substantially farther from V0 (Figure 28). Recovery therefore follows the sample-specific, spatially aligned target-update structure.

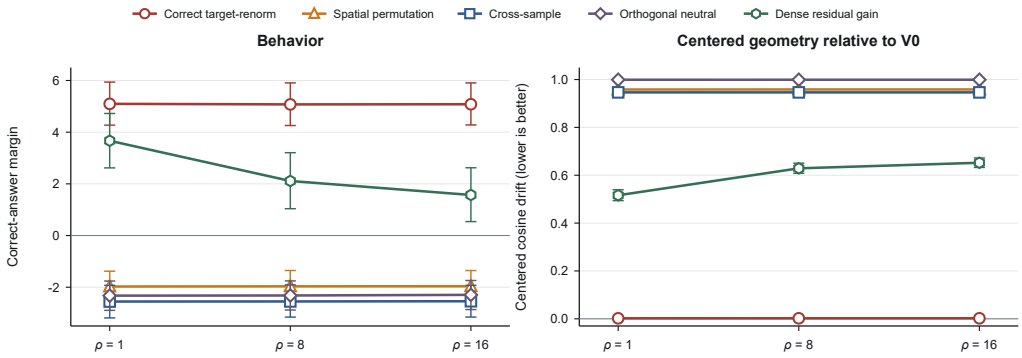

Figure 28: Structure controls for target renormalization. Spatial permutation, donor states from another example, and an equal-norm orthogonal update remove recovery, showing that the restored update must preserve target-aligned structure.

B.6 Prefix and Suffix Target Attention Restoration

We next scan every possible prefix endpoint and suffix starting block to map when target-attention protection accumulates and retains recognition recovery. A prefix restores target-only attention in blocks B_0 through B_ℓ; a suffix restores it in blocks B_ℓ through B_{26}. Blocks outside the intervened range retain Dense attention. The prefix and suffix scans each use paired Dense and full-depth Target-renorm endpoints.

For an intervened block set $\mathcal{S}$, retained margin gain is the ratio of paired mean gains:

$$R(\mathcal{S}) = 100\,\frac{\mathbb{E}_s[M_{\mathrm{TR},s}(\mathcal{S}) - M_{\mathrm{Dense},s}]}{\mathbb{E}_s[M_{\mathrm{TR},s}(\mathcal{B}) - M_{\mathrm{Dense},s}]}, \qquad \mathcal{B} = \{B_0, \ldots, B_{26}\}. \tag{21}$$

For pooled results, $\rho = 1, 8, 16$ are first averaged within each example. Thus, $R = 0$ denotes the Dense endpoint and $R = 100$ the mean gain of full-depth Target-renorm. The ratio is formed from paired mean gains, with intervals from 10,000 paired-cluster bootstrap replicates.

The two cumulative scans identify a broad early and middle transition. Figure 29 shows how retained recovery changes as restoration extends farther from the input. Across all ratios, the prefix curves begin a sustained rise once the endpoint reaches roughly B8, with the strongest accumulation continuing through B14 and a more gradual increase thereafter. The prefix turning point is therefore around B8.

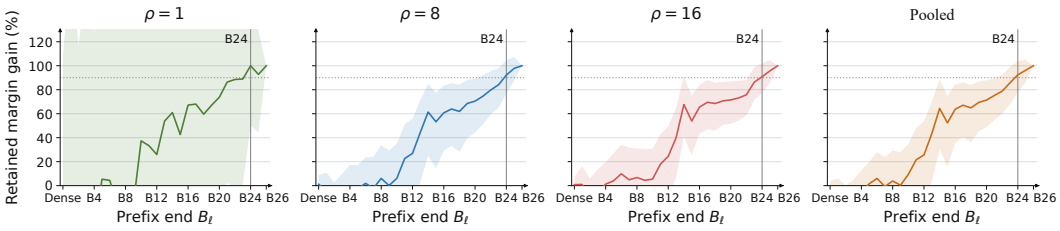

Figure 29: Prefix Target-renorm retention. Columns show $\rho = 1, 8, 16$ and the within-sample pooled result. The horizontal coordinate is the final restored block; all preceding blocks also use Target-renorm, while Dense is the no-intervention endpoint. Recovery begins its sustained rise around B8 and accumulates most strongly through B14. Bands are 95% sample-cluster bootstrap intervals.

Figure 30 provides the complementary view by restoring attention from each starting block through B26. Reading suffix starts from late to early blocks, recovery first begins a sustained rise when the protected range reaches the region around B14, and then grows as the starting point moves into earlier blocks. Together, the prefix and suffix turning points localize the main recovery transition to an extended early-to-middle stage rather than a single isolated block; later computation largely propagates and refines the altered target state.

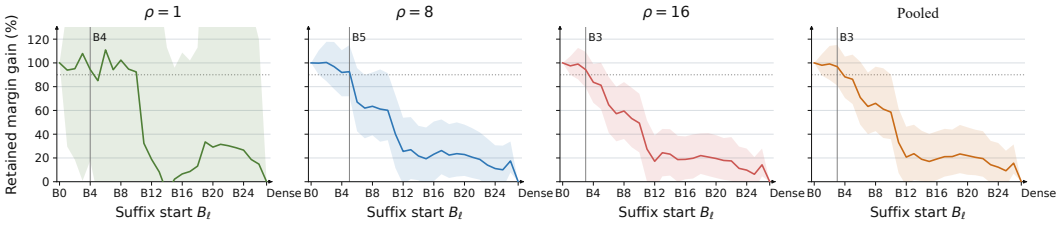

Figure 30: Suffix Target-renorm restores target attention from the indicated block through B26; B27 is the Dense endpoint. Recovery is measured relative to the corresponding Dense and full-depth Target-renorm endpoints. Pooled results average background ratios within each example. When read from late to early blocks, recovery begins its sustained rise around B14; this is the turning point rather than the layer of maximal recovery.

The target-internal attention peak marks an attention-fragility regime. Together with the natural Qwen3-VL profile in Figure 31, the depth scans connect two quantities: where target queries place their largest attention mass back onto target keys, and when recovery begins to rise as suffix protection is extended backward from late to earlier blocks. Their relative locations define the model's fragile interval. High target-to-target attention therefore does not imply robustness: target structure is actively forming, and background access during this window introduces changes that later restoration cannot fully undo. To test whether this peak-and-upturn structure is cross-architectural, we next repeat both diagnostics in LLaVA-OneVision-1.5 and InternVL3.5, then extend them to 12 additional models, including five InternVL checkpoints.

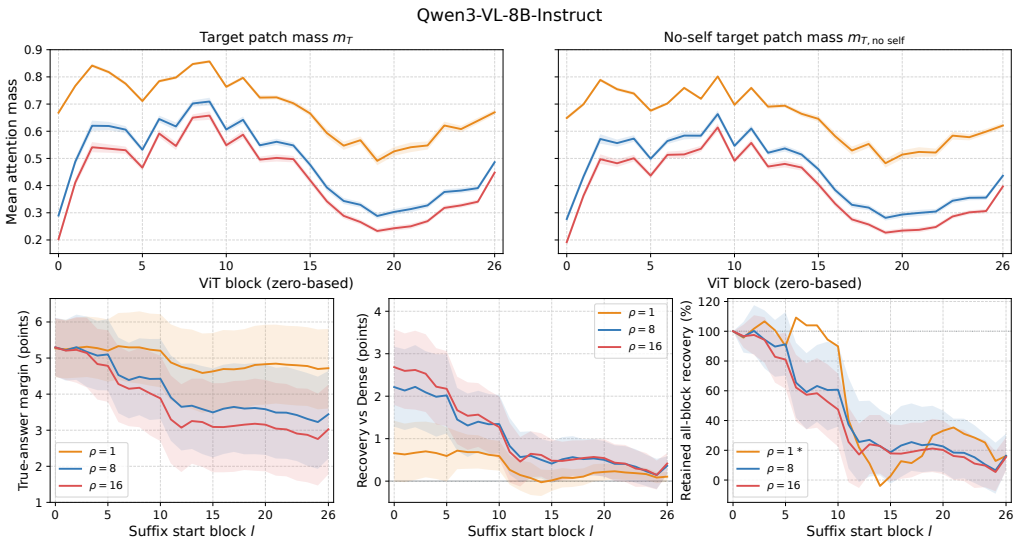

Figure 31: Qwen3-VL-8B-Instruct attention-fragility diagnostic. Top: target-query attention mass assigned to target patch keys, with and without the diagonal self edge; the target-internal peak spans approximately B6–B14 and is reduced as ρ grows. Bottom: true-answer margin, recovery over Dense, and retained all-block recovery across suffix Target-renorm starts. The suffix-onset block is about B14. Bands are pointwise 95% paired-bootstrap intervals.

B.7 Model Generalization

Throughout this subsection, we define the *suffix-onset block* as the point where recovery begins a sustained rise as the suffix start moves from later to earlier blocks; when the onset spans adjacent blocks, we report a region.

Attention-mass convention for native CLS tokens. For encoders with native visual CLS tokens, we report the raw attention mass assigned to target patch keys under the full native softmax:

$$m^h_{i,T} = \sum_{j \in \mathcal{T}^{\text{patch}}} a^h_{ij}.$$

Any native CLS keys remain in the attention denominator, and their aggregate mass $m^h_{i,C}$ is recorded separately, so that $m^h_{i,T} + m^h_{i,C} + m^h_{i,D} = 1$. The self-excluded diagnostic subtracts the query patch's diagonal attention weight and keeps the original normalization, i.e., $m^h_{i,T,\text{excl}} = m^h_{i,T} - a^h_{ii}$.

B.7.1 Primary Qwen3-VL Attention and Suffix Restoration

Figure 31 reports the natural attention profiles and their suffix-start causal diagnostics. Their alignment identifies the attention-fragility signature: background access becomes behaviorally consequential while target structure is actively forming. The normalized effect at $\rho = 1$ is noisier because its absolute all-block recovery is smaller.

B.7.2 LLaVA-OneVision-1.5 Attention and Suffix Restoration

Figure 32 reports the corresponding LLaVA-OneVision-1.5-8B diagnostic on the selected examples. The signed retained-effect ratio is noisy at $\rho = 1$ because the corresponding all-block gain is small; the absolute margin and recovery panels provide the stable comparison. Replication under RICE-ViT recovers the same attention-fragility relationship without relying on the Qwen merger or attention schedule.

B.7.3 InternVL3.5 Attention and Suffix Restoration

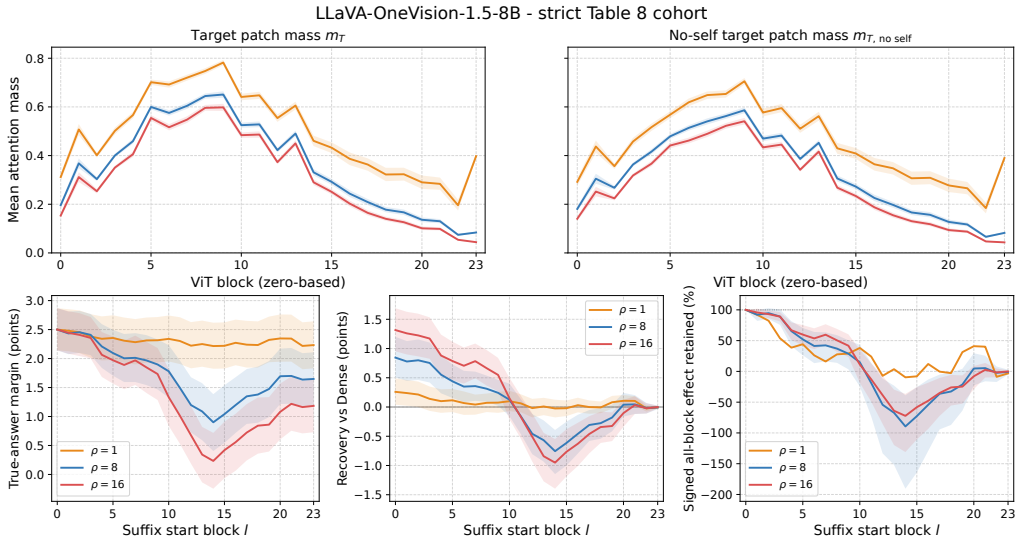

Figure 32: Attention-fragility diagnostic for LLaVA-OneVision-1.5-8B at $\rho = 1, 8, 16$. Top panels report target-to-target attention mass with and without the diagonal self edge; the broad target-internal maximum spans approximately B5–B10. Bottom panels report margin, recovery over Dense, and retained all-block effect across suffix starts. The suffix-onset block is around B11. Bands are pointwise 95% paired-bootstrap intervals.

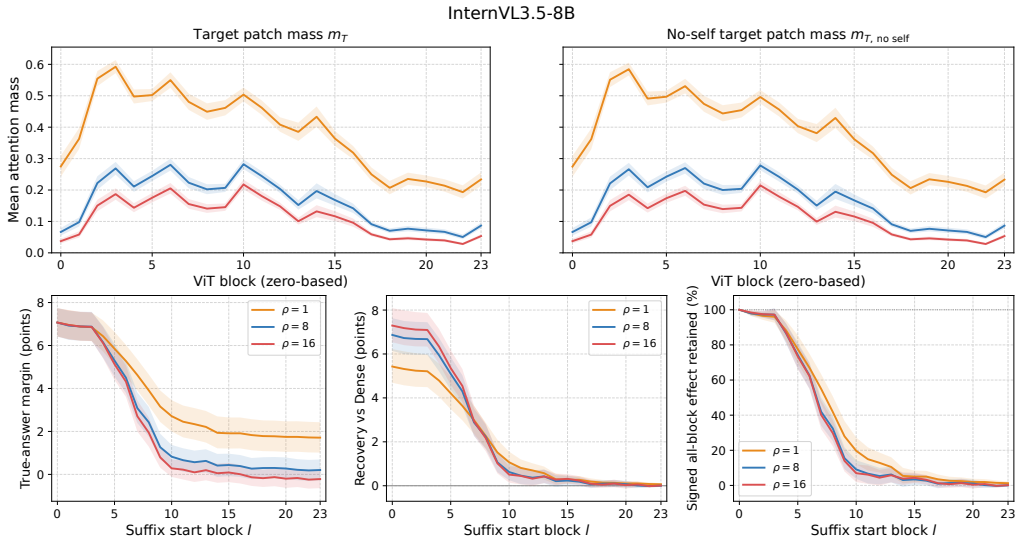

Figure 33: Attention-fragility diagnostic for InternVL3.5-8B with native tiling, local CLS attention, pixel shuffle, and projector preserved. Target-to-target attention exhibits repeated peaks across approximately B2–B10 and remains ordered by ρ after self-edge removal. The suffix-onset region is around B10–B11. Bands are pointwise 95% paired-bootstrap intervals.

Figure 33 reports the corresponding InternVL3.5-8B diagnostic under its native tiled visual pipeline. Target-patch attention exhibits repeated early-to-middle peaks and remains ordered by background ratio after removing diagonal self-attention. The suffix scan places recovery onset near the late edge of this peak sequence, extending the attention-fragility relationship to dynamic tiling, local CLS tokens, pixel shuffle, and the native projector.

B.7.4 Extended 12-Model Attention and Suffix-Restoration Suite

We apply the same paired diagnostic to the 12 additional checkpoints introduced in Appendix A.14 as a cross-model test of attention fragility. Figures 34 through 45 report one checkpoint per figure, with two vertically stacked figures on each page. Under V1, the top row locates where target queries assign their largest mass to target patch keys, with and without the diagonal self edge, and tests whether background competition suppresses that target-internal attention. The bottom row is read from late to early suffix starts and identifies the layer at which extending Target-renorm backward first produces a clear rise in margin and recovery over the paired Dense V1L0 baseline. Attention fragility predicts a model-specific early or middle recovery boundary associated with target-representation formation rather than a purely late effect. Each checkpoint is evaluated on the selected cohort at $\rho = 1, 8, 16$, with pointwise 95% paired bootstrap intervals shared across conditions and ratios.

Target-internal peaks and recovery upturns generalize across all five architecture families. Across Gemma3, the five InternVL checkpoints, Qwen2-VL, both Qwen2.5-VL checkpoints, and the three additional Qwen3-VL variants, the upper panels locate an architecture-specific target-internal attention peak or peak sequence. Target attention is also ordered by background ratio through most of visual depth: $\rho = 1$ receives the most target mass and $\rho = 16$ the least. The ordering remains after subtracting each query's diagonal self-attention, so the effect cannot be reduced to changes in a single dominant self-edge. Qwen2.5 places its peaks at global-attention blocks, InternVL preserves the repeated geometry induced by tiled processing and local CLS tokens, Qwen3 follows its DeepStack visual path, and Gemma3 concentrates its largest target-internal mass in the first few blocks.

The suffix scans provide the causal localization. When read from late to early starting blocks, margin and recovery remain near the Dense endpoint until protection reaches a model-specific early or middle stage, where they turn upward. For Qwen and InternVL, this boundary lies near or shortly after the late edge of the main target-internal peak sequence; Gemma3 instead exhibits an earlier raw-attention maximum followed by a later behavioral upturn. The exact relationship therefore shifts with visual design, but no model moves the first clear recovery rise to a purely late stage. Taken together with the factorial suite, the background-dependent attention reduction and the early-to-middle recovery boundaries support a shared attention-fragility mechanism.

The individual pages also distinguish the shared mechanism from architecture-specific expression. In Qwen2.5, periodic global blocks create visible target-internal peaks, while the Qwen3 variants have smoother peak bands. InternVL produces repeated tile-associated peaks, whereas Gemma3 separates an early attention maximum from a later behavioral upturn. These differences rule out a mechanism tied to one visual position system, one global attention schedule, or one merger design while preserving the common result that target-internal attention is background sensitive and protection becomes behaviorally consequential well before purely late computation.

The total and self-excluded curves provide a second consistency check. If the ratio ordering were driven mainly by diagonal self-attention, subtracting the self-edge would remove or reverse it. Instead, the separation remains across the suite, showing that background changes the broader target-internal attention pattern. The suffix curves need not mirror raw attention at a single block because each point protects a cumulative range. Together, the background-dependent target-mass separation, the location of the target-internal peak, and the early-to-middle recovery upturn provide the cross-model evidence for attention fragility.

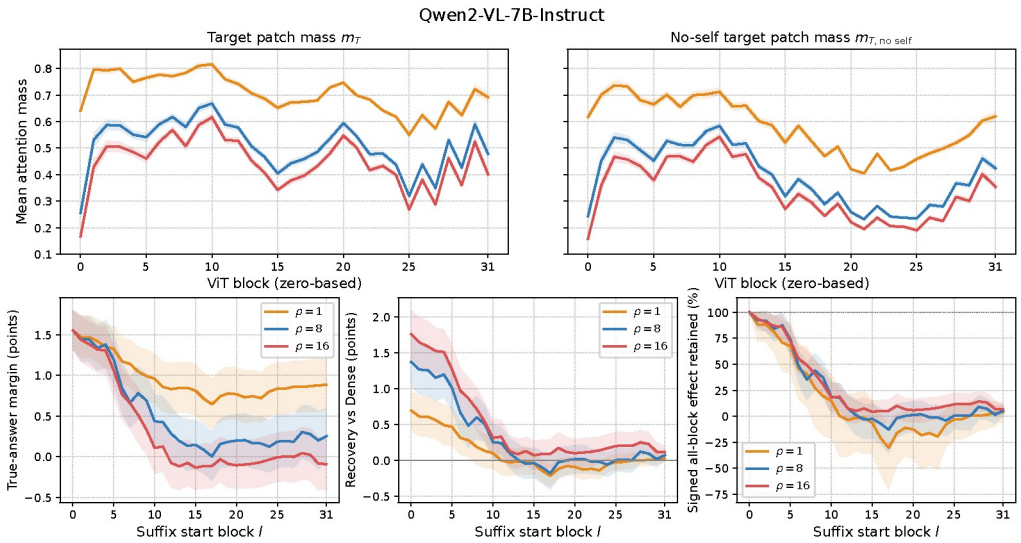

Figure 34: Attention-fragility diagnostic for Qwen2-VL-7B-Instruct. Target-to-target attention has a dominant early plateau across approximately B1–B12 and remains ordered by ρ. The suffix-onset block is around B12. Bands are pointwise 95% paired-bootstrap intervals.

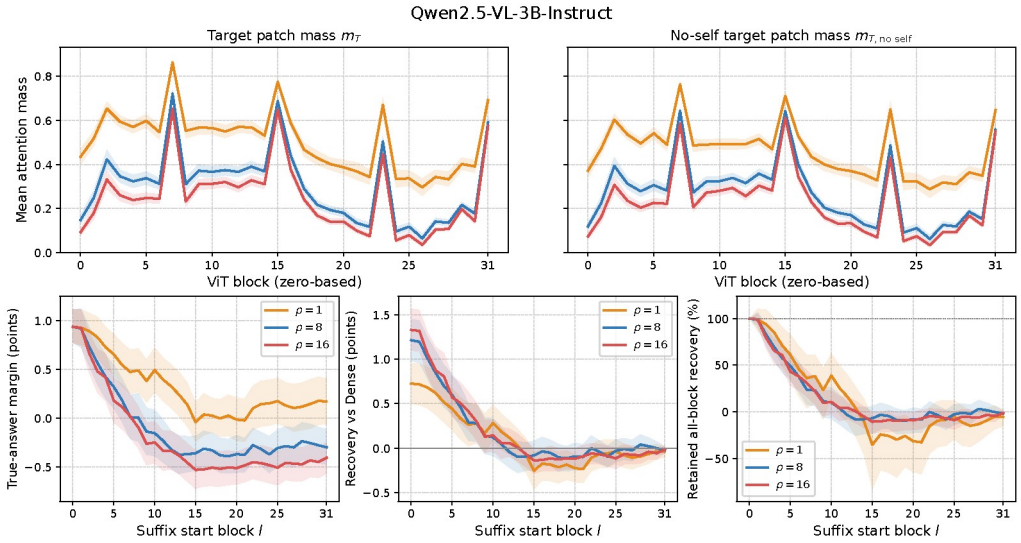

Figure 35: Attention-fragility diagnostic for Qwen2.5-VL-3B-Instruct. The mean target-to-target attention profile is highest across approximately B6–B16, with sharp periodic peaks at B7, B15, B23, and B31. These four layers are the model's native full-attention blocks; all other layers use window attention, so the periodic spikes are an architectural signature rather than a data artifact. The ρ ordering persists between peaks and after self-edge removal. The suffix-onset region is B12–B15. Bands are pointwise 95% paired-bootstrap intervals.

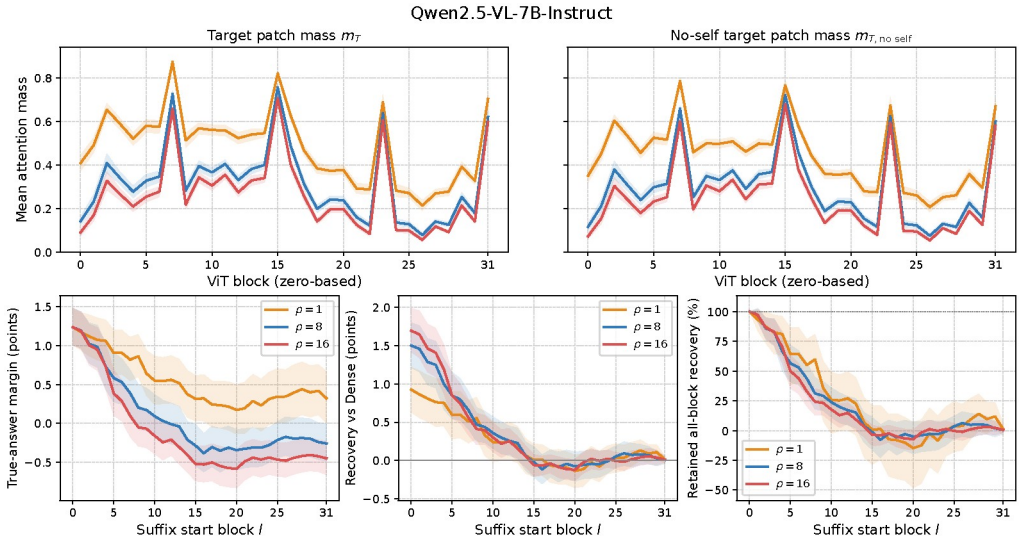

Figure 36: Attention-fragility diagnostic for Qwen2.5-VL-7B-Instruct. The mean target-to-target attention profile is highest across approximately B6–B16, with sharp periodic peaks at B7, B15, B23, and B31. These four layers are the model's native full-attention blocks; all other layers use window attention, so the periodic spikes are an architectural signature rather than a data artifact. Target mass decreases as ρ grows. The suffix-onset block is around B15. Bands are pointwise 95% paired-bootstrap intervals.

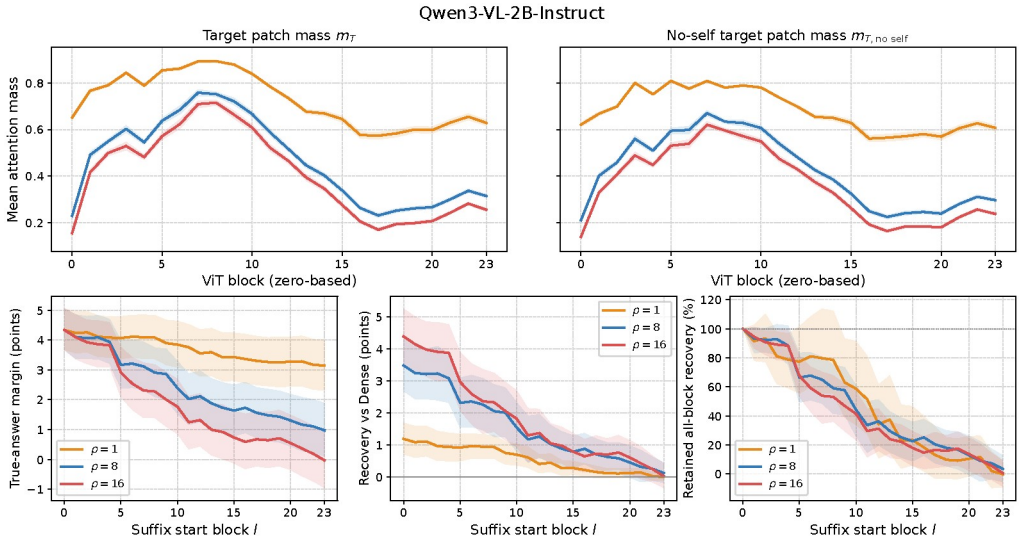

Figure 37: Attention-fragility diagnostic for Qwen3-VL-2B-Instruct under its DeepStack visual encoder. Target-to-target attention is highest across approximately B5–B10 and remains separated by ρ. The suffix-onset region is B11–B12. Bands show pointwise 95% paired-bootstrap intervals.

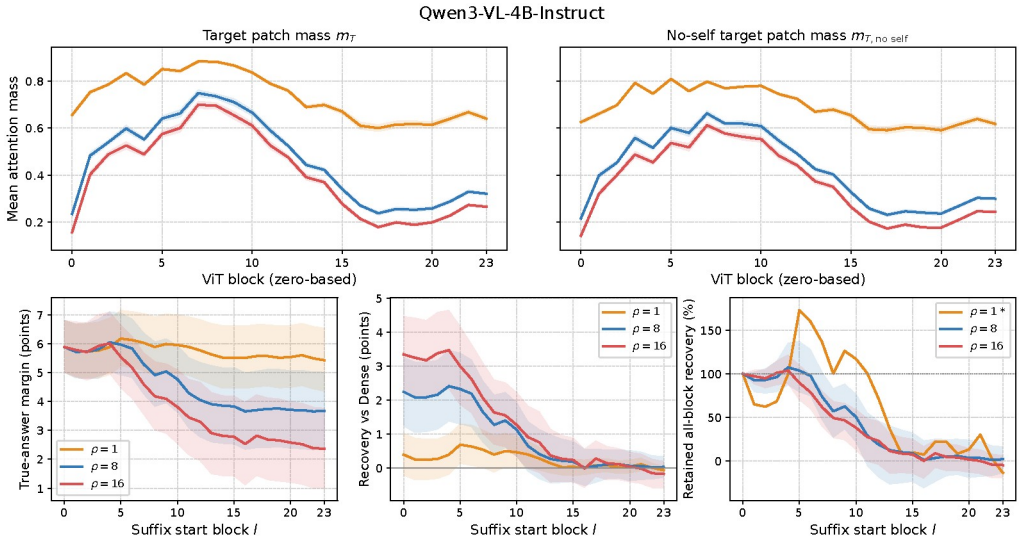

Figure 38: Attention-fragility diagnostic for Qwen3-VL-4B-Instruct. Target-to-target attention has a smooth maximum across approximately B5–B10, with higher ρ receiving less mass after self-edge removal. The suffix-onset region is B12–B13. Bands are pointwise 95% paired-bootstrap intervals.

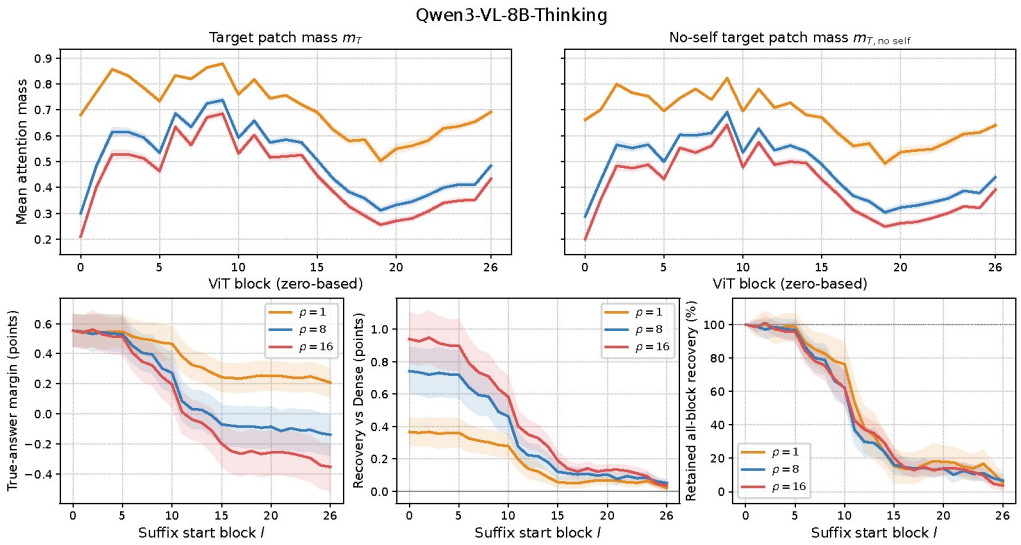

Figure 39: Attention-fragility diagnostic for Qwen3-VL-8B-Thinking, evaluated from immediate option logits. Target-to-target attention has a dominant peak band across approximately B6–B14 and remains ordered by ρ. The suffix-onset block is around B14. Bands show pointwise 95% paired-bootstrap intervals.

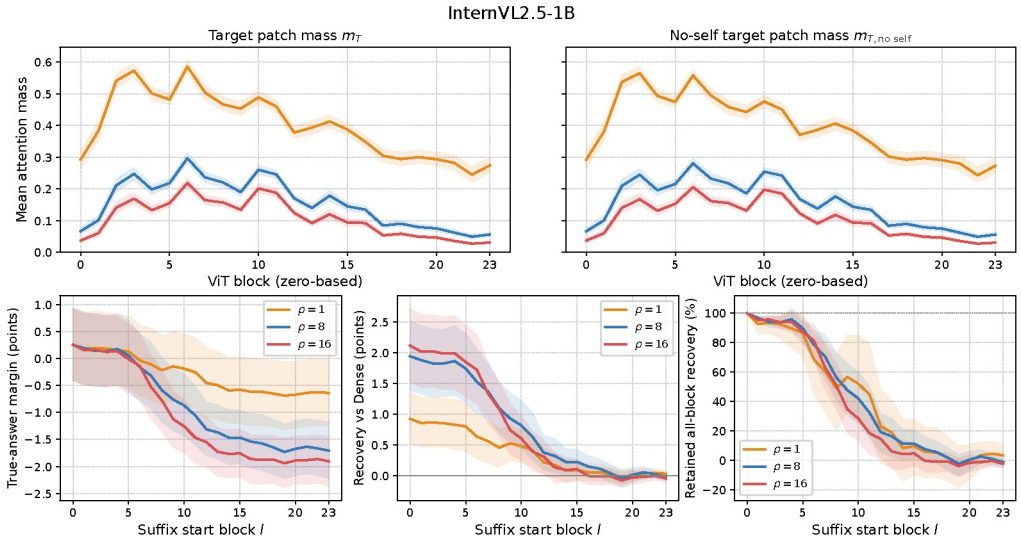

Figure 40: Attention-fragility diagnostic for InternVL2.5-1B with local tile CLS attention, pixel shuffle, and projector preserved. Repeated target-to-target attention peaks span approximately B3–B11 and remain ordered by ρ after self-edge removal. The suffix-onset region is B12–B13. Bands are pointwise 95% paired-bootstrap intervals.

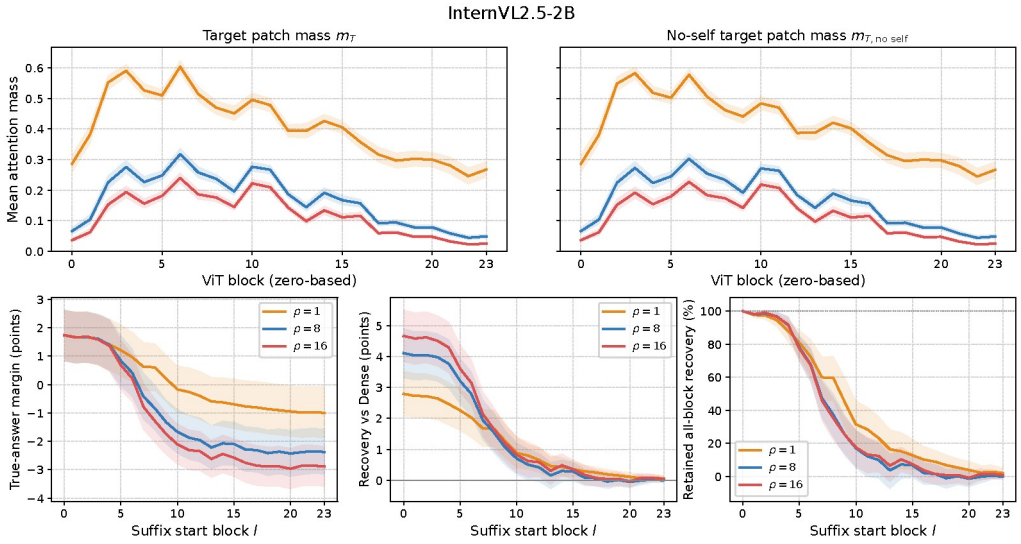

Figure 41: Attention-fragility diagnostic for InternVL2.5-2B with native tiling, local CLS attention, pixel shuffle, and projector preserved. Tile-induced target-to-target attention peaks span approximately B3–B11 and weaken as ρ grows, including after self-edge removal. The suffix-onset block is around B13. Bands show pointwise 95% paired-bootstrap intervals.

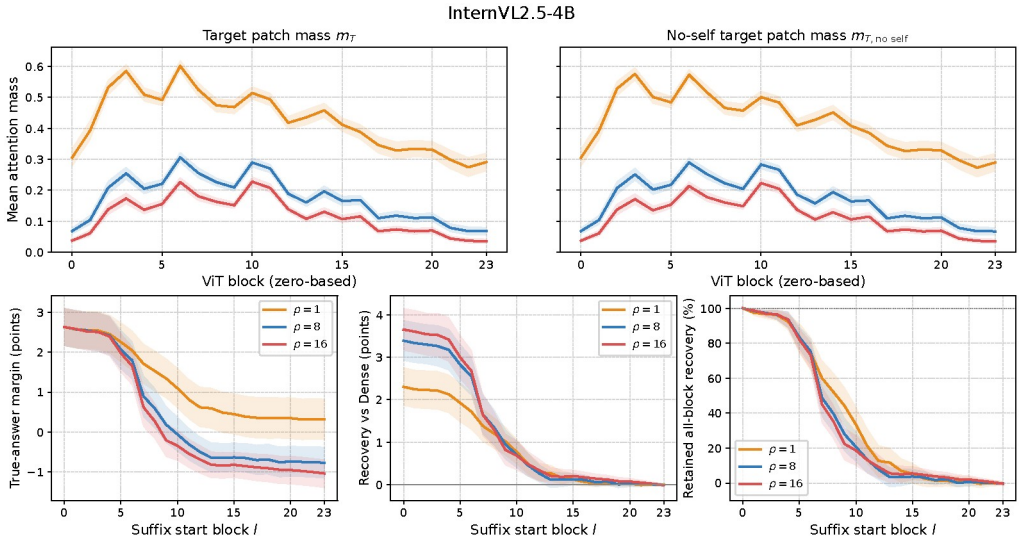

Figure 42: Attention-fragility diagnostic for InternVL2.5-4B. Repeated tile-level target-to-target attention peaks span approximately B3–B11, with persistent separation by ρ. The suffix-onset region is B12–B13. Bands are pointwise 95% paired-bootstrap intervals.

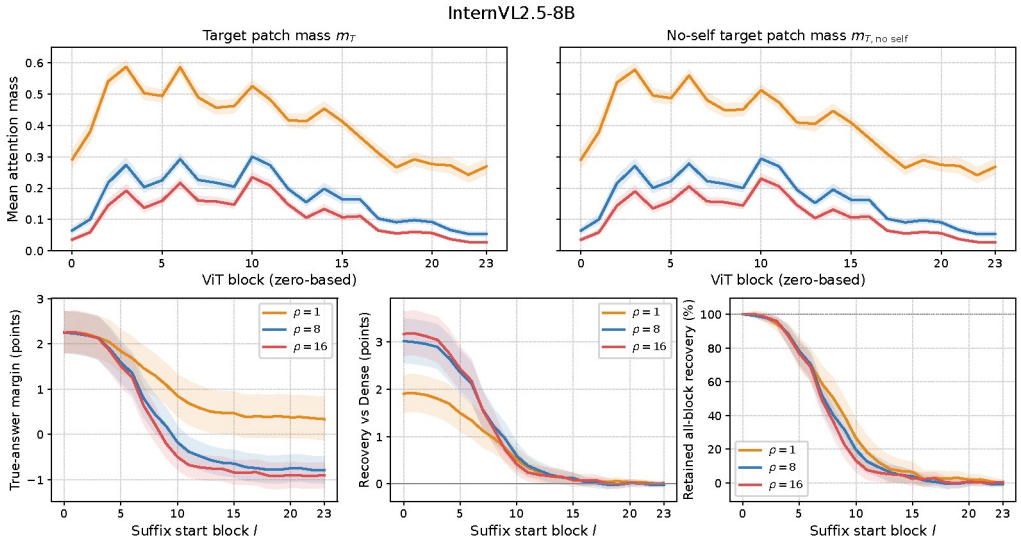

Figure 43: Attention-fragility diagnostic for InternVL2.5-8B. Tile-structured target-to-target attention peaks span approximately B3–B11 and retain monotonic ρ ordering. The suffix-onset block is B13. Scaling changes the magnitude, not the early-to-middle onset localization. Bands show pointwise 95% paired-bootstrap intervals.

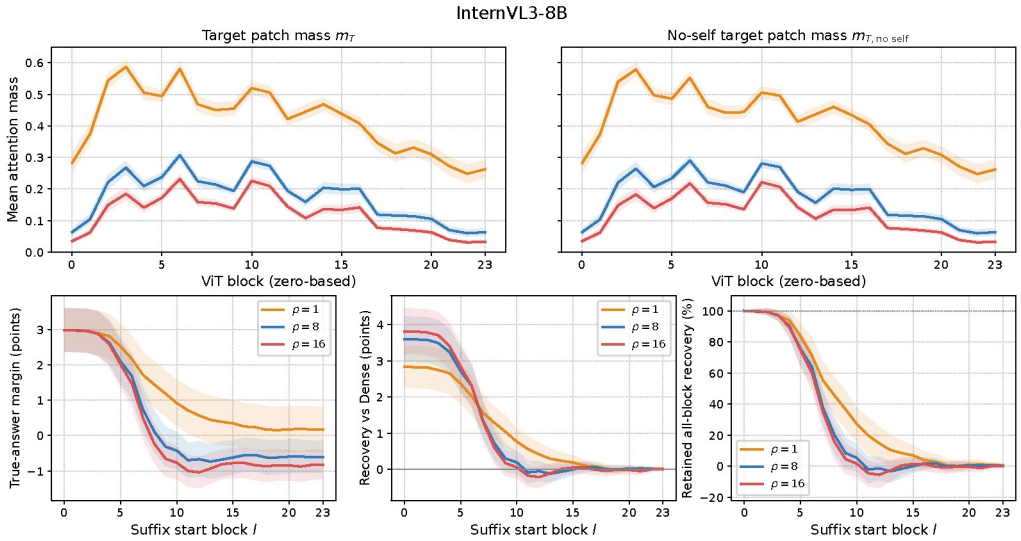

Figure 44: Attention-fragility diagnostic for InternVL3-8B under its tiled InternViT encoder. Repeated target-to-target attention peaks span approximately B3–B11 and retain ρ separation after self-edge removal. The suffix-onset region is B10–B11. The newer generation preserves the same early-to-middle onset localization. Bands show pointwise 95% paired-bootstrap intervals.

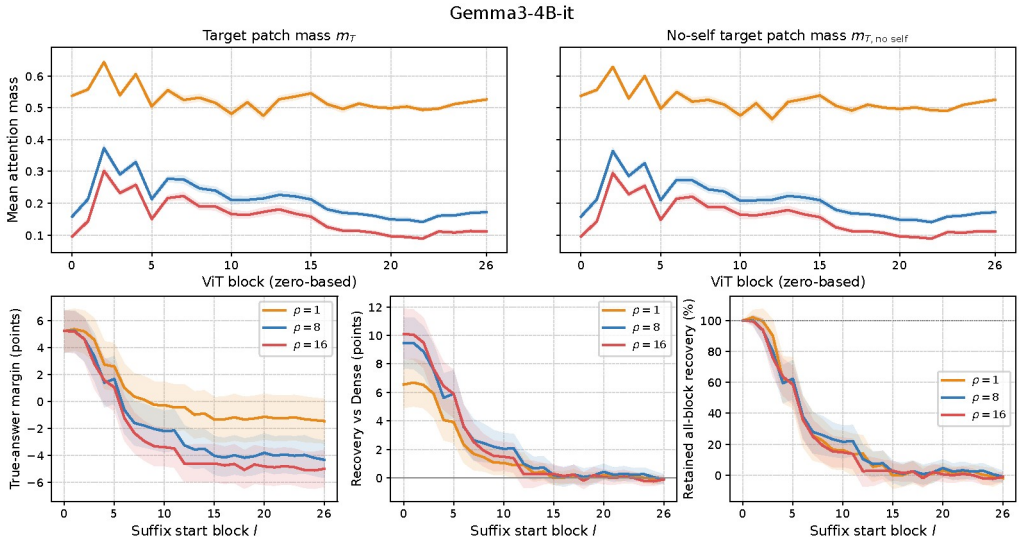

Figure 45: Attention-fragility diagnostic for Gemma3-4B-it under its pooled SigLIP encoder. Target-to-target attention has its main early maximum around B2–B7 and remains ordered by ρ. The suffix-onset block is B10. Thus, the behavioral onset can lag the raw attention maximum in a smoother encoder. Bands show pointwise 95% paired-bootstrap intervals.

B.8 Distributed ViT Mechanisms Across Heads and Frequencies

We broaden the mechanism analysis from aggregate attention to all 432 layer and head units and multiple RoPE frequency subsets. These experiments show how background interference is organized across depth, heads, and positional frequencies.

B.9 Distributed Head Effects Across ViT Depth

This experiment asks whether background interaction is concentrated in a small reusable set of attention heads or distributed across a broader stage of ViT computation. A sparse mechanism would preserve head identity across background ratios, while a stage mechanism would distribute effects across heads within a consistent depth range.

Head selection used five-fold cross-fitting on the frozen 30-example discovery set. Each fold contained six held-out examples, with two examples from each historical selection stratum. For every fold, we ranked all $27 \times 16 = 432$ zero-based ViT block–head units by their mean single-head margin gain over the 24 training examples and both ρ values, then jointly masked the top 87 units only on that fold's held-out examples. Recovery is the ratio of the mean paired 87-unit margin gain to the mean paired all-unit target-to-background masking gain on the same 60 sample–ρ outcomes. Confidence intervals use 10,000 paired bootstrap resamples over sample IDs, keeping the two ρ outcomes from each example together. For sample s, the stratification score and all-unit gain are

$$\begin{aligned} D_s^{\mathrm{sel}} &= \frac{1}{2} \sum_{\rho \in \{8,16\}} \left(M_{00,\rho}^{\mathrm{VL},(s)} - M_{10,\rho}^{\mathrm{VL},(s)} \right), \\ D_s^{\mathrm{head}} &= \frac{1}{2} \sum_{\rho \in \{8,16\}} \left(M_{\mathrm{allhead},\rho}^{(s)} - M_{\mathrm{dense},\rho}^{(s)} \right). \end{aligned} \tag{22}$$

The score D_s^{sel} forms the high, middle, and low historical selection strata. The recovery denominator D_s^{head} is the paired gain from masking all target-query-to-background-key/value edges. Figure 46 reports the complete single-unit effect maps.

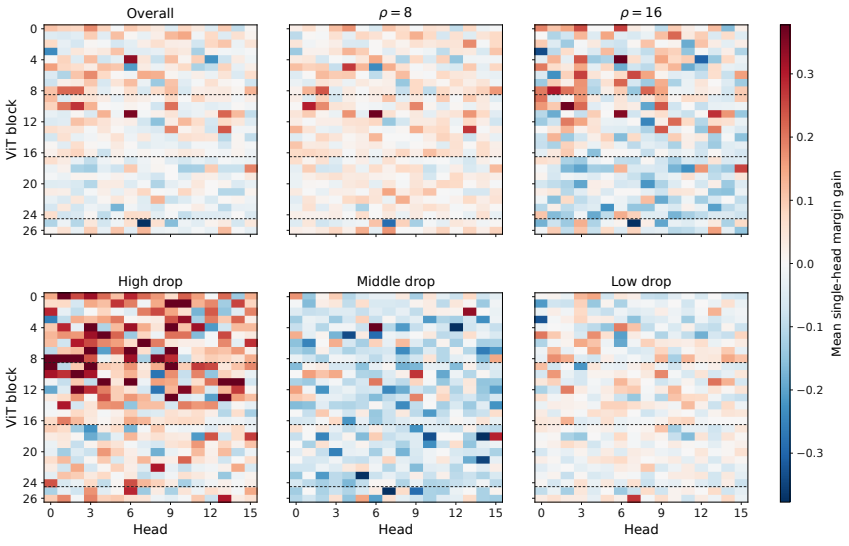

Figure 46: Single-head intervention effects on the frozen discovery cohort, shown overall, by ratio, and across baseline-drop strata. The cross-fitted top 87 units recover 41.19% of the all-unit masking gain.

Head effects remain distributed across the frozen discovery cohort. We reran the head-localization analysis under the same fixed-max-carrier protocol as Table 7 while freezing the original 30-example discovery cohort and five stratified folds. For each fold, all 432 layer–head units were

ranked using only the other 24 examples at $\rho \in \{8, 16\}$, and the top 87 units were jointly evaluated on the six held-out examples. Across the 60 held-out sample–ρ outcomes, the selected units improved the correct-answer margin by 1.938 points (95% sample-cluster bootstrap CI $[0.981, 2.977]$), compared with 4.704 points $[2.435, 7.246]$ for masking all target-query-to-background-key/value edges on the same pairs. The corresponding recovery was 41.19% $[20.15\%, 78.97\%]$, below the prespecified 50% point-estimate gate. Two individual units, B4H3 and B11H13 in zero-based indexing, had positive effects after BH correction ($q = 0.043$ for each), but the overall positive effect remained distributed (effective head count 243.0; top-10% share 36.0%). These estimates characterize the historically selected discovery cohort and are not an independent full-cohort or held-out result.

B.10 Structured Effects Across RoPE Frequencies

We next map the background effect across RoPE frequency components and test how much behavioral recovery can be obtained from cross-fitted frequency subsets.

We estimate effects for each layer and frequency, select candidate subsets through cross-fitting, and compare their recovery with random subsets of the same size. Figures 47 and 48 report the resulting structure and behavioral effects.

RoPE frequency effects are structured and strongly depth dependent. The layer and frequency maps contain coherent bands rather than uniform noise across the spectrum, but their magnitude and sign change with visual depth and with the score used to rank components. Several local bands contribute across multiple blocks, while no single frequency remains dominant under every ratio and selection rule. A cross-fitted four-frequency subset outperforms matched random selection in one setting, whereas the advantages for $K = 8$ and $K = 11$ remain unresolved and the selected identities change under alternative scores. Positional frequency therefore shapes the distributed visual mechanism without defining a fixed architecture-independent bottleneck. The stable pattern again lies at the level of a structured stage: particular bands become useful within certain depth ranges, and their effect depends on the computation in which they participate.

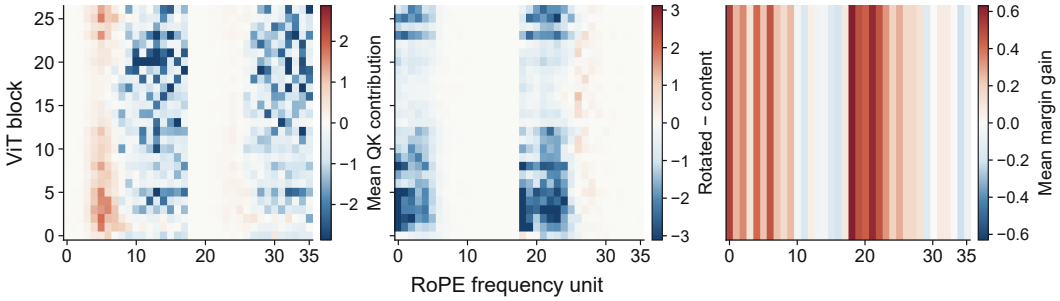

Figure 47: Layer-frequency effects in the RoPE-frequency analysis.

Small frequency subsets recover substantial but incomplete behavioral effects. Figure 48 compares cross-fitted top-K subsets with the all-frequency intervention, direct target-to-background masking, and random subsets of the same size. Recovery rises rapidly as frequencies are added, and the $K = 18$ subset slightly exceeds the mean gain of the full-frequency intervention. Even this result remains much smaller than the recovery from directly controlling target access to background, so frequency manipulation captures one component of the mechanism rather than replacing the attention explanation. The $K = 4$ subset has a clear advantage over matched random subsets, while the intervals for $K = 8$ and $K = 11$ overlap zero. Together with the changing selected identities, the curve shows how a few well-chosen components can carry a meaningful local effect even though the full causal process remains distributed across frequencies and depth.

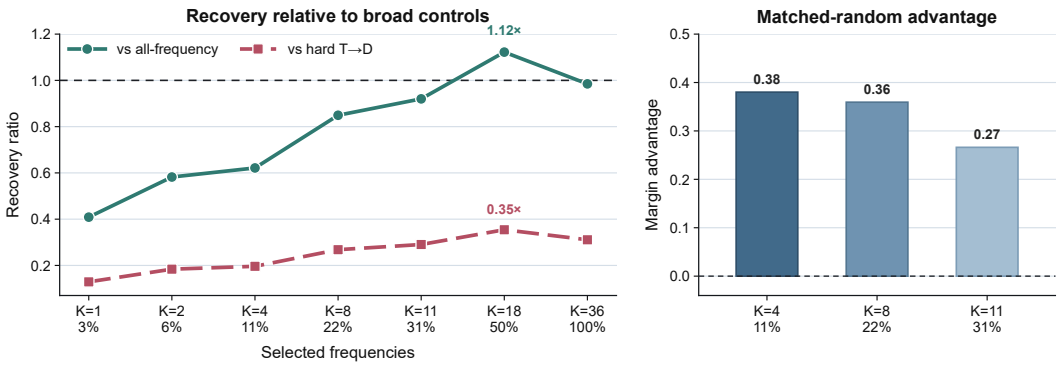

Figure 48: Cross-fitted RoPE-frequency repair. The left panel reports recovery relative to the all-frequency and hard target-to-distractor controls across subset sizes. The right panel reports the mean correct-answer margin advantage over matched-random subsets for $K = 4, 8, 11$. Small selected subsets recover substantial effects, but the stronger target-to-distractor damage remains only partially explained by frequency-selective repair alone.

B.11 Target Geometry: Drift, Relational Structure, and Subspace Trajectories

This section supplements the analysis in Section 3.2 with a finer-grained account of how the shape of the target-token configuration changes. We examine the residual change after optimal alignment, token–token cosine relations, and pairwise token distances, and then trace how the leading target subspace evolves from the ViT into the LLM.

Definitions supplementing the main-text decomposition. Let $X_c^{(\ell)} \in \mathbb{R}^{n^{(\ell)} \times d}$ denote the complete state matrix at stage ℓ, where $n^{(\ell)}$ is the total number of represented positions and $c = 0$ denotes the background-isolated V0L0 run while $c = 1$ denotes the context-interacting V1L0 run. Let $\mathcal{I}_T^{(\ell)}$ index the target rows and let $H_c^{(\ell)} = X_c^{(\ell)}[\mathcal{I}_T^{(\ell)}, :] \in \mathbb{R}^{n_T^{(\ell)} \times d}$ denote the target-only submatrix. Within the ViT, $n_T^{(\ell)} = n_T^{\text{patch}}$; after the native merger, pooling, or projector, it equals the downstream target-slot count N_T defined in Appendix A.9. Thus, the stage-specific count in Equation 3 keeps native patch tokens distinct from post-merger slots throughout the paper: $\mu_c^{(\ell)} = (n_T^{(\ell)})^{-1} \sum_{r=1}^{n_T^{(\ell)}} h_{r,c}^{(\ell)}$ and $\delta_{i,c}^{(\ell)} = h_{i,c}^{(\ell)} - \mu_c^{(\ell)}$. Writing the centered states as the rows of $\bar{H}_c^{(\ell)}$ gives

$$\begin{aligned} H_c^{(\ell)} &= \mathbf{1}(\mu_c^{(\ell)})^\top + \bar{H}_c^{(\ell)}, \qquad \mathbf{1}^\top \bar{H}_c^{(\ell)} = 0, \\ \left\| H_1^{(\ell)} - H_0^{(\ell)} \right\|_F^2 &= n_T^{(\ell)} \left\| \mu_1^{(\ell)} - \mu_0^{(\ell)} \right\|_2^2 + \left\| \bar{H}_1^{(\ell)} - \bar{H}_0^{(\ell)} \right\|_F^2 . \end{aligned} \tag{23}$$

The identity separates the shared displacement of all target tokens from changes to their centered, token-relative organization. The main-text trajectories measure these components using

$$D_\mu^{(\ell)} = 1 - \cos\left(\mu_1^{(\ell)}, \mu_0^{(\ell)}\right), \qquad D_{\cos,\text{ctr}}^{(\ell)} = 1 - \frac{1}{n_T^{(\ell)}} \sum_{i=1}^{n_T^{(\ell)}} \cos\left(\delta_{i,1}^{(\ell)}, \delta_{i,0}^{(\ell)}\right). \tag{24}$$

Larger values indicate greater deviation from the matched isolated representation.

B.12 A Multi-View Decomposition of Target-Token Shape Changes

Beyond the centroid and centered-token drifts summarized in the main text, we further characterize target-token shape changes from three complementary views. The Procrustes residual measures how much of the centered configuration differs after its best orthogonal alignment, the cosine Gram matrix captures changes in angular relations among target tokens, and pairwise Euclidean distances capture changes in their relative spacing. Together, these measures provide a more detailed account of how background interaction reorganizes the target configuration across depth and background ratio.

For sample s at ViT block ℓ, let $H_0^{(s,\ell)}$ and $H_1^{(s,\ell)}$ denote the isolated and background-interacting target states. Target inputs, source positions, raster order, and native grouping are identical across the two conditions. At patch-level stages, we concatenate the four raw patches belonging to each native merger unit, so each row is aligned to the same eventual LLM visual-token slot. We then center each matrix over its N_T target units,

$$\bar{H}_c^{(s,\ell)} = H_c^{(s,\ell)} - \mathbf{1}\big(\mu_c^{(s,\ell)}\big)^\top, \qquad \mu_c^{(s,\ell)} = \frac{1}{N_T}\sum_{i=1}^{N_T} H_{c,i}^{(s,\ell)}. \tag{25}$$

Here $c \in \{0,1\}$ indexes the isolated and background-interacting conditions. Each $H_c^{(s,\ell)} \in \mathbb{R}^{N_T \times d}$, where d is the feature dimension; $H_{c,i}^{(s,\ell)}$ is its ith target row, $\mu_c^{(s,\ell)}$ is the target centroid, and $\mathbf{1}$ broadcasts that centroid across rows. Thus, $\bar{H}_c$ removes shared translation while retaining token-relative structure. Below, $\|\cdot\|_F$ and $\|\cdot\|_2$ are the Frobenius and Euclidean norms, Q is orthogonal, and $\widehat{H}_c$ is the row-normalized centered matrix. The matrices G_c and E_c store pairwise cosine similarities and Euclidean distances, with i, j indexing aligned target units. Each discrepancy D is dimensionless and normalized by the isolated reference $c = 0$. We suppress the sample and block superscripts below. Orthogonal Procrustes alignment removes the best shared feature-space rotation,

$$Q^* = \arg\min_{Q^\top Q = I} \big\|\bar{H}_1 Q - \bar{H}_0\big\|_F, \qquad D_{\mathrm{Proc}} = \frac{\big\|\bar{H}_1 Q^* - \bar{H}_0\big\|_F}{\big\|\bar{H}_0\big\|_F}. \tag{26}$$

We additionally normalize every centered row and compare the resulting token–token cosine relations through the corresponding Gram matrices,

$$G_c = \widehat{H}_c \widehat{H}_c^\top, \qquad D_{\mathrm{Gram}} = \frac{\|G_1 - G_0\|_F}{\|G_0\|_F}, \tag{27}$$

and compare all centered pairwise Euclidean distances,

$$E_c(i,j) = \big\|\bar{H}_{c,i} - \bar{H}_{c,j}\big\|_2, \qquad D_{\mathrm{pair}} = \frac{\|E_1 - E_0\|_F}{\|E_0\|_F}. \tag{28}$$

The three quantities emphasize different aspects of the same shape change. D_{Proc} reports the residual discrepancy after the best shared orthogonal alignment, D_{Gram} tracks angular relations while also removing uniform rescaling, and D_{pair} tracks the distance structure among target tokens. Their agreement indicates that the observed reorganization is expressed consistently across aligned states, angular relations, and relative spacing.

Figure 49 shows a consistent dose-ordered reorganization across the fixed Qwen3-VL-8B cohort of matched examples. The effect is already visible in B0–B8, increases sharply at the transition to B9, and remains elevated through the rest of the ViT. In the middle region B9–B16, mean D_{Proc} increases from 0.374 $[0.345, 0.403]$ at $\rho = 1$ to 0.674 $[0.640, 0.710]$ at $\rho = 8$ and 0.895 $[0.840, 0.952]$ at $\rho = 16$. The same ordering appears in token–token relations: D_{Gram} rises from 0.343 to 0.560 and 0.739, while D_{pair} rises from 0.314 to 0.604 and 0.815. The synchronized increase across all three views shows that the early-to-middle transition changes both the aligned target states and their internal angular and distance structure.

The regional summaries in Table 10 further show that both adjacent dose contrasts are positive for every metric in the early, middle, and late regions. At block resolution, $\rho = 8$ exceeds $\rho = 1$ at all 27 blocks; $\rho = 16$ exceeds $\rho = 8$ at 26 of 27 blocks, with only B0 statistically unresolved. The depth profile itself is not monotone: the geometric change is sharply amplified around B9 and remains large thereafter, while some measures recede from their middle-region maxima. Background interaction therefore changes when and how strongly the target configuration is reorganized, rather than accumulating as a uniform perturbation at every block.

Together, these analyses add three layers of detail to the main-text drift decomposition: the target configuration changes after optimal alignment, its token–token angular relations are reorganized, and its pairwise spacing shifts in the same dose-ordered pattern. They further localize the strongest amplification to the transition around B9 and show that the subsequent trajectory varies by geometric property rather than following a uniform accumulation across blocks. The clean-state repairs in Appendix B.13 subsequently test which components of the altered target structure support recognition.

Metric	ρ	Early B0–B8	Middle B9–B16	Late B17–B26
D_{Proc}	1	.096 [.093, .100]	.374 [.345, .403]	.403 [.391, .414]
	8	.151 [.146, .156]	.674 [.640, .710]	.576 [.559, .593]
	16	.175 [.168, .182]	.895 [.840, .952]	.682 [.656, .709]
D_{Gram}	1	.137 [.132, .142]	.343 [.330, .356]	.523 [.511, .536]
	8	.229 [.221, .237]	.560 [.532, .591]	.688 [.672, .703]
	16	.266 [.257, .276]	.739 [.688, .792]	.743 [.724, .761]
D_{pair}	1	.041 [.039, .043]	.314 [.287, .342]	.207 [.196, .218]
	8	.067 [.064, .070]	.604 [.570, .638]	.355 [.338, .372]
	16	.084 [.080, .088]	.815 [.761, .869]	.458 [.432, .484]

Table 10: Regional target-geometry changes in Qwen3-VL-8B. Entries are means with 95% paired sample-cluster bootstrap intervals over matched examples. Layers are averaged within each example before bootstrap aggregation.

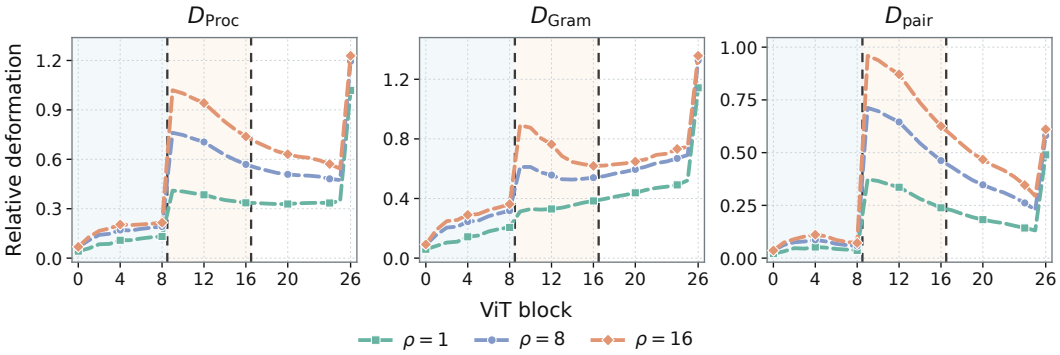

Figure 49: **Background interaction reorganizes target-token geometry in Qwen3-VL-8B.** From left to right, the panels show the Orthogonal-Procrustes residual D_{Proc}, normalized centered-token cosine-Gram difference D_{Gram}, and centered pairwise-distance difference D_{pair}. Lower values indicate greater similarity between the isolated V0L0 and background-interacting V1L0 target states. Lines show means over the same matched examples; bands show 95% paired sample-cluster bootstrap intervals from 10,000 draws. Blue and peach backgrounds mark the prespecified early (B0–B8) and middle (B9–B16) regions; the late region is B17–B26.

The leading target subspace adds a fourth view of this shape change. Extending the relational measures above, we measure alignment between the dominant directions of the centered target representation. We record $\bar{H}_c^{(\ell)}$ after patch embedding, every ViT block, the visual merger, and every LLM layer for the same examples at $\rho \in \{1, 8, 16\}$. Let $Q_{c,k}^{(\ell)} \in \mathbb{R}^{d\times k}$ contain the top k right singular vectors of $\bar{H}_c^{(\ell)}$. Their normalized squared overlap is

$$S_k^{(\ell)} = \frac{1}{k}\left\|\left(Q_{0,k}^{(\ell)}\right)^\top Q_{1,k}^{(\ell)}\right\|_F^2 = \frac{1}{k}\sum_{j=1}^{k}\cos^2\theta_j^{(\ell)}, \qquad k \in \{1, 5, 10\}, \tag{29}$$

where $\theta_j^{(\ell)}$ are the principal angles between the isolated and context-interacting target subspaces. Thus, $S_k^{(\ell)} = 1$ denotes identical leading subspaces, while smaller values indicate greater rotation. To localize where this rotation changes, Figure 50 plots the sample-averaged stagewise increment

$$\Delta S_k^{(\ell)} = \text{mean}_s\left[S_{k,s}^{(\ell)} - S_{k,s}^{(\ell-1)}\right]. \tag{30}$$

Positive values indicate local realignment, and negative values indicate further rotation apart. This quantity describes geometry and is interpreted alongside the causal repair results.

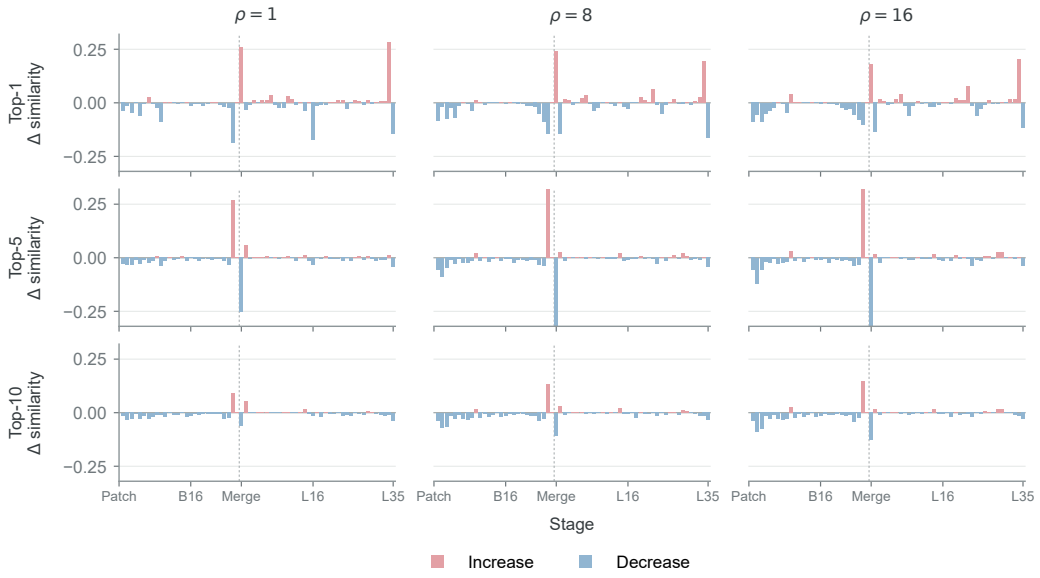

Figure 50: Stagewise change in target-subspace similarity between V0L0 and V1L0. Rows show $k = 1, 5, 10$, columns show background ratios, and the dashed line marks the visual merger. Positive bars indicate local realignment; negative bars indicate further rotation apart.

Subspace rotation accumulates in the ViT and changes abruptly near the merger. Figure 50 shows predominantly negative stagewise changes from patch embedding through the ViT, with larger and more sustained departures at $\rho = 8$ and 16. The leading centered directions therefore rotate progressively as target tokens interact with background rather than changing only at the downstream interface. The largest opposite-sign changes occur around the visual merger, particularly for the top-5 and top-10 subspaces, where compression reorganizes the directions delivered to the LLM. Most subsequent LLM changes are smaller and closer to zero, with isolated positive and negative spikes that do not undo the earlier separation. Because the same pattern appears for $k = 1, 5, 10$, it extends beyond a single dominant component. Together with the centroid and centered-token drifts in the main text, these trajectories show that background interaction changes both the shared target state and the orientation of token-relative structure. The centered subspace remains altered even where the centroid partially recovers, and the clean-state repairs below identify which part of that geometry carries recognition-relevant information.

B.13 Clean Target-State Repair

We specify the intervention used for the component-wise target-state repair in the main text.

The clean donor is the matched V0L0 run and the recipient is V1L0; both hide background tokens from the LLM. A repair applied at block output ℓ replaces target patch rows before the next block. Other rows remain on the recipient trajectory, after which the remaining ViT blocks, DeepStack computation, visual merger, and LLM continue normally.

Full target-state repair replaces only target positions:

$$\widetilde{X}^{(\ell)}_{\text{clean}} = P^{(\ell)}_T X^{(\ell)}_0 + \left(I_{n^{(\ell)}} - P^{(\ell)}_T\right) X^{(\ell)}_1, \tag{31}$$

where $P^{(\ell)}_T \in \{0,1\}^{n^{(\ell)} \times n^{(\ell)}}$ is the diagonal projector onto the target rows $\mathcal{I}^{(\ell)}_T$. Consequently, the target-row submatrix of $\widetilde{X}^{(\ell)}_{\text{clean}}$ is $H^{(\ell)}_0$, while every non-target row remains equal to the corresponding row of $X^{(\ell)}_1$. Centroid-only repair restores the shared target centroid while retaining contextualized centered states; shape-only repair restores centered target states while retaining the contextualized centroid. The remainder of the forward pass is unchanged.

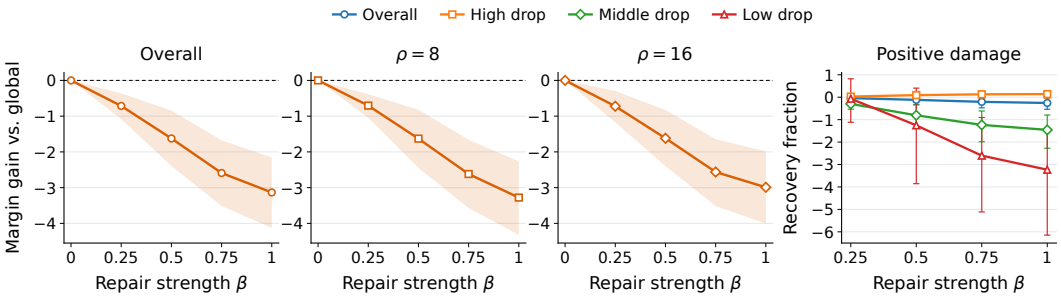

Figure 51: Post-B16 interpolation toward the same-token local-attention anchor. The first three panels show correct-answer margin change relative to the unmodified global state, pooled across ratios and separately at $\rho = 8$ and 16. The rightmost panel reports the ratio of mean local-anchor gain to mean full-$V0$ repair gain for positive-damage pairs, with strata defined by the ViT-side selection loss D_s^{sel} in Equation 22. Bands and error bars show 95% paired-bootstrap confidence intervals.

B.14 Exploratory Local Attention Repair

We additionally explore a simple repair strategy that suppresses long-range ViT interaction. Local attention offers a state-independent way to reduce broad visual mixing while preserving nearby computation within each source segment.

Let G_i denote target token i after B16 in the ordinary contextualized V1 run, L_i the corresponding state obtained under local attention, and $V0_i$ its matched isolated-target state. From B0 through B16, each patch attends only to positions within the following same-segment 5×5 neighborhood:

$$\mathcal{N}(i) = \{j : \max(|x_j - x_i|, |y_j - y_i|) \le 2,\ \text{seg}(j) = \text{seg}(i)\}\,. \tag{32}$$

This restriction blocks long-range interaction while preserving local computation within each target or background segment of the original visual layout.

Immediately after B16, we interpolate the ordinary contextualized target state toward the resulting local-attention state using the following convex path:

$$H_i^{(\beta)} = G_i + \beta(L_i - G_i), \qquad \beta \in \{0.25, 0.50, 0.75, 1.00\}. \tag{33}$$

Here, $\beta = 0$ recovers the original contextualized state and $\beta = 1$ uses the full local-attention state. The remaining ViT blocks and downstream LLM computation then proceed normally.

Local attention suppresses useful integration together with background interaction. Increasing β consistently reduces the correct-answer margin at both $\rho = 8$ and $\rho = 16$ in Figure 51; even the weakest intervention, $\beta = 0.25$, decreases the pooled margin by 0.714 points, and the recovery fraction remains near or below zero across repair strengths. The corresponding geometry audit explains why this seemingly protective mask does not produce a clean anchor. Under every geometry measure, the local-attention state lies farther from the matched isolated state $V0$ than the original contextualized state. The 5×5 restriction removes long-range background access, but it simultaneously prevents target tokens from integrating evidence across the object. Fine-grained recognition therefore requires selective separation: harmful target-to-background reads must be controlled while target-internal computation remains globally available.

C Target Localization and Recognition in the LLM (Section 4)

C.1 Target Localization and Attention Allocation

We first separate three quantities at the answer position: target and background ranking, relative attention preference, and absolute attention received by target tokens. Direct interventions then connect these quantities to recognition.

ρ	Condition	Δ_{QK}	m_T/m_V	m_T (%)	$M\uparrow$
1	V1L1	−0.019	0.496	2.166	3.976
	V0L1	**0.212**	**0.543**	**2.241**	4.689
	Centroid	0.072	0.513	2.216	4.005
	Shape	−0.009	0.501	2.154	**4.782**
	Full	0.026	0.508	2.175	4.610
8	V1L1	0.631	0.255	1.471	2.570
	V0L1	**1.073**	**0.321**	**1.758**	**4.214**
	Centroid	0.769	0.272	1.567	2.579
	Shape	0.687	0.263	1.520	4.092
	Full	0.913	0.295	1.660	4.051
16	V1L1	0.871	0.201	1.290	2.020
	V0L1	**1.306**	**0.251**	**1.527**	**3.929**
	Centroid	0.993	0.212	1.336	2.286
	Shape	0.954	0.211	1.348	3.828
	Full	1.097	0.224	1.394	3.868

Table 11: Answer-position arithmetic means. Both m_T/m_V and m_T are averaged over example, layer, and head observations; column maxima within each ρ block are shaded.

C.2 Answer-Position Target Localization AUROC

We measure how answer-position target and background discriminability changes across background ratios and ViT-interaction conditions, using the LLM layers, attention heads, and matched V0L1, V1L0, and V1L1 trajectories from Figure 8 for all examples. For each example, layer, and head, we take the original pre-softmax scores $s_i = q^\top k_i/\sqrt{d_h}$ from the answer-prediction position to all target and background carrier slots used for the score diagnostic. Target tokens have label 1 and background tokens label 0. Under V1L0, the background slots are masked from realized attention, while their scores support the target and background ranking comparison. AUROC is computed independently for each example, layer, and head, equivalently as

$$\mathrm{AUROC} = \frac{1}{N_T N_D} \sum_{i\in T} \sum_{j\in D} \left[\mathbf{1}(s_i > s_j) + \tfrac{1}{2}\mathbf{1}(s_i = s_j)\right]. \tag{34}$$

An AUROC of 0.5 indicates chance-level pairwise ranking; larger values mean that target tokens more often receive higher scores than background tokens. The metric captures score-based localization, while the next subsection measures attention allocation and recognition. For the layerwise curves, we first average AUROC across heads within each example and layer, then across examples. Figure 52 additionally averages across the evaluated layers within each example before reporting the sample mean and 95% paired bootstrap CI. Resampling preserves the pairing of conditions and background ratios within each example.

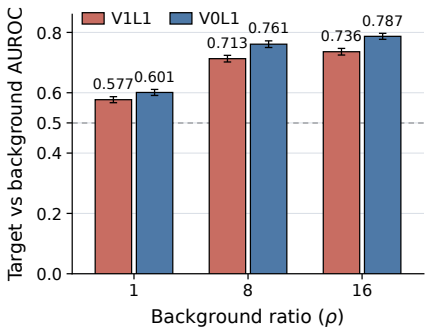

Figure 52: Layer-averaged answer-position target/background AUROC.

The LLM sharpens target ranking as background grows, while ViT interaction weakens that response. Under V1L1, layer-averaged AUROC increases from 0.5774 to 0.7128 and 0.7364 as ρ grows from 1 to 8 and 16. The corresponding V0L1 values are higher at 0.6010, 0.7615, and 0.7871. Thus, the answer query becomes better at ranking an individual target token above an individual background token when competition increases, but the ranking is consistently weaker when the target has already interacted with background in the ViT. The growing AUROC should not be interpreted as growing access to target evidence: it is a pairwise ordering statistic and does not account for the number of background tokens, the total visual attention budget, or the quality of information stored within the target states. The next decomposition makes these distinctions explicit and explains how localization can improve while recognition declines.

C.3 Per-Token Selectivity and Absolute Target Attention

We derive an exact attention-budget decomposition to explain how stronger per-token target preference can coexist with lower total target attention and correct-answer margin.

For a fixed answer query and attention head, let $s_i = q^\top k_i/\sqrt{d_h}$ and let X contain all remaining nonvisual keys accessible to that answer query. We define

$$Z_T = \sum_{i\in T} e^{s_i}, \qquad Z_D = \sum_{j\in D} e^{s_j}, \qquad Z_X = \sum_{x\in X} e^{s_x}. \tag{35}$$

Then $\Delta_{\mathrm{QK}} = \log[(Z_T/N_T)/(Z_D/N_D)]$ and $Z_D/Z_T = \rho\exp(-\Delta_{\mathrm{QK}})$. Consequently,

$$\begin{aligned} m_V &= \frac{Z_T + Z_D}{Z_T + Z_D + Z_X}, \\ m_T &= \frac{Z_T}{Z_T + Z_D + Z_X} = m_V \frac{1}{1+\rho\exp(-\Delta_{\mathrm{QK}})} \\ &= m_V\,\sigma(\Delta_{\mathrm{QK}} - \log\rho). \end{aligned} \tag{36}$$

Here, $\sigma(u) = 1/(1+e^{-u})$. The identity holds for each query and head before aggregation and separates per-token preference, the background token-count penalty, and the total attention budget allocated to the visual tokens relative to all accessible nonvisual tokens. For each example s, layer ℓ, and head h, we compute the realized target share

$$\left(\frac{m_T}{m_V}\right)_{s\ell h} = \sigma(\Delta_{\mathrm{QK},s\ell h} - \log\rho)\,. \tag{37}$$

Table 11 reports the arithmetic mean of these realized shares,

$$\overline{\frac{m_T}{m_V}} = \frac{1}{NLH}\sum_{s=1}^{N}\sum_{\ell=1}^{L}\sum_{h=1}^{H}\left(\frac{m_T}{m_V}\right)_{s\ell h}. \tag{38}$$

Because the sigmoid is nonlinear,

$$\mathbb{E}\left[\frac{m_T}{m_V}\right] = \mathbb{E}[\sigma(\Delta_{\mathrm{QK}} - \log\rho)] \neq \sigma(\mathbb{E}[\Delta_{\mathrm{QK}}] - \log\rho)\,. \tag{39}$$

The table therefore reports the mean realized share computed for each observation. Total target attention is averaged in the same way,

$$\overline{m_T} = \frac{1}{NLH}\sum_{s,\ell,h} m_{V,s\ell h}\left(\frac{m_T}{m_V}\right)_{s\ell h}, \tag{40}$$

and the percentage column in Table 11 reports $100\,\overline{m_T}$.

Experimental objective. We use this comparison to determine whether recognition loss under growing background is primarily a failure of answer-position target localization or a loss of recognition-relevant information within the target states. Across the natural, ViT-isolated, and repaired conditions, we jointly measure per-token target preference (Δ_{QK}), the target share of visual attention (m_T/m_V), absolute target attention (m_T), and the correct-answer margin. This separates changes in localization and token-count dilution from behavioral recovery, testing whether improved target selection alone explains restored recognition.

Token-count competition converts stronger selectivity into lower absolute target attention. Table 11 makes the token-count penalty explicit. In the natural V1L1 condition, Δ_{QK} rises from -0.019 at $\rho = 1$ to 0.631 at $\rho = 8$ and 0.871 at $\rho = 16$, indicating stronger target-over-background preference per visual token. Nevertheless, the realized target share decreases from 0.496 to 0.255 and 0.201, total target attention decreases from 2.166% to 1.471% and 1.290%, and the correct-answer margin decreases from 3.976 to 2.570 and 2.020. Background competition therefore still dilutes target allocation, although the decrease is less severe than suggested by applying the sigmoid to the mean QK contrast. The growing number of background tokens outweighs the compensatory increase in per-token selectivity. The first two panels of Figure 53

show the corresponding layerwise comparison at $\rho = 16$: relative to V0L0, activating direct LLM background access in V0L1 produces a larger QK evidence term through most middle and late layers, while realized target attention is generally lower.

The V0L1 comparison shows that upstream ViT interaction further weakens the target signal. At every ratio, isolating target and background in the ViT increases Δ_{QK}, the target share within visual attention, total target attention, and the correct-answer margin. At $\rho = 16$, V0L1 raises Δ_{QK} from 0.871 to 1.306 and total target attention from 1.290% to 1.527%, while the margin increases from 2.020 to 3.929. The LLM sharpens target ranking as competition grows, yet degraded visual evidence continues to constrain recognition.

The repair rows further separate recognition from these mean attention statistics. Centroid repair modestly increases selectivity and target attention but produces little behavioral recovery; at $\rho = 8$, its margin is 2.579 compared with 2.570 for V1L1. Shape repair, by contrast, raises the margin to 4.092 while leaving Δ_{QK} and total target attention close to their V1L1 values (0.687 versus 0.631, and 1.520% versus 1.471%). Full-state repair shifts the mean attention quantities more strongly, yet its margin is similar to Shape and becomes only slightly higher at $\rho = 16$.

Behavioral recovery therefore tracks restoration of centered target structure more closely than average target ranking or attention mass. Damage to this structure reaches the answer stage as degraded evidence, while stronger target selectivity partially compensates for the growing token count. Aggregate target and background ranking can improve even as organization within the target deteriorates: the LLM retains a coarse *where* signal, yet the fine-grained *what* structure needed for recognition remains degraded. This separation also explains why centroid repair can move the attention summaries without recovering the answer, whereas Shape repair restores the margin with only small changes in the same summaries. Localization and recognition are related computations, but they diagnose different properties of the target representation.

C.4 Answer-Stage Retrieval and Downstream Readout

We organize the analysis around three questions: whether late target retrieval can directly restore recognition, when the consequence of ViT background interaction becomes readable at the answer position, and which source states carry the late answer-aligned signal. We combine causal interventions with local readout diagnostics to distinguish information that remains directly recoverable from target visual tokens from information that has already been integrated into downstream nonvisual states.

Can late target retrieval directly restore recognition? We test this question with sparse retrieval controls and direct target-to-answer repair. In the 30-example pilot at $\rho = 8, 16$, oracle target-only attention changes the margin by -0.015 points and QK top-K selection by -0.004 points, with both intervals including zero. For direct target-to-answer repair, we scanned all 36 layers on a separate $n = 30$ selection set, used positive sufficiency and necessity effects to define L5–13 as the candidate region, and compared prespecified contiguous three- and five-layer windows. L7–9 achieved the largest combined positive score and was frozen before evaluation on the full evaluation sample. On this sample, the L7–9 sufficiency effect is $+0.022$ with a 95% confidence interval of $[-0.003, 0.046]$, and the necessity effect is $+0.007$ with an interval of $[-0.016, 0.029]$. Both remain close to zero. These results establish that stronger target ranking and concentrated selection do not by themselves provide a late direct recovery path from target visual tokens to the answer. Combined with the AUROC increase, they instead motivate testing where the recognition-relevant difference becomes readable and which downstream states carry it.

When does the ViT-induced difference become readable at the answer position? We use a natural-trajectory logit-lens diagnostic to locate when the consequence of ViT background interaction becomes linearly readable. The fourth panel of Figure 53 compares answer-position residual states from the natural V0L0 and V1L0 runs. For each sample s, background ratio ρ, and LLM layer ℓ, we extract the states at the end of the layer,

$$h^{\mathrm{V0L0}}_{s,\rho,\ell,\mathrm{ans}} \quad \text{and} \quad h^{\mathrm{V1L0}}_{s,\rho,\ell,\mathrm{ans}}. \tag{41}$$

We pass each intermediate state directly through the model's final normalization and unembedding to obtain the following proxy answer margin at every layer:

$$m^{z}_{s,\rho,\ell} = (W_{c_s} - W_{w_s})^{\top} \operatorname{Norm}\big(h^{z}_{s,\rho,\ell,\mathrm{ans}}\big), \qquad z \in \{\mathrm{V0L0}, \mathrm{V1L0}\}, \tag{42}$$

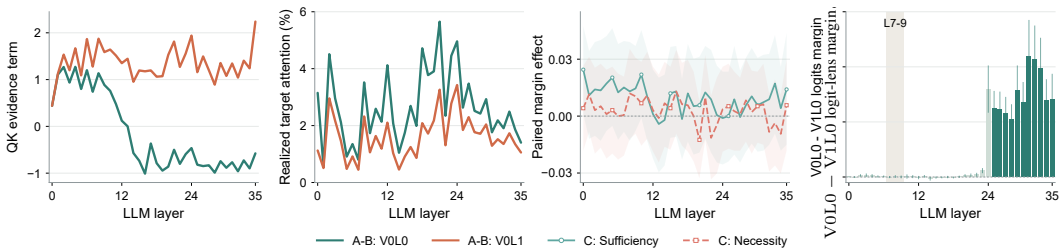

Figure 53: Additional retrieval and answer-readout diagnostics. The first two panels compare the QK evidence term and realized target attention across LLM layers at $\rho = 16$. Relative to V0L0, V0L1 exhibits a larger QK evidence term through most middle and late layers but generally lower realized target attention, showing that stronger per-token evidence need not increase absolute target allocation. The third panel reports the near-zero sufficiency and necessity effects of direct target-to-answer repair. The fourth panel shows the natural-trajectory V0L0 − V1L0 logit-lens margin; positive values indicate more linearly readable correct-versus-strongest-wrong information under ViT isolation. Its L7–9 shading marks the window frozen by the third-panel repair selection procedure, not a significance interval or a window selected from the fourth-panel curve.

Layer	V0L0 − V1L0 logit-lens margin	95% CI
L7	−0.002	[−0.041, 0.037]
L8	+0.015	[−0.036, 0.067]
L9	+0.002	[−0.048, 0.054]
L21	+0.053	[0.006, 0.101]
L24	+0.228	[0.138, 0.324]
L25	+3.171	[2.334, 4.051]
L32	+4.376	[3.126, 5.703]

Table 12: Selected layers from the natural-trajectory logit-lens diagnostic in Figure 53. Results use the full evaluation sample. We first average $\rho \in \{8, 16\}$ within each sample and then construct 95% confidence intervals using 10,000 paired-cluster bootstrap resamples over sample IDs.

where c_s is the correct answer token, w_s is the strongest incorrect answer under the final V1L0 output and is fixed across layers and conditions, and W_{c_s} and W_{w_s} are the corresponding LM-head readout directions. The panel plots

$$\Delta^{\mathrm{V0-V1}}_{s,\rho,\ell} = m^{\mathrm{V0L0}}_{s,\rho,\ell} - m^{\mathrm{V1L0}}_{s,\rho,\ell}. \tag{43}$$

A positive value means that the V0L0 answer state contains more information that the final head can linearly read along the correct-versus-strongest-wrong direction. Both conditions use the L0 attention pattern, so the answer position cannot directly read background visual tokens in either run; they differ only in whether target and background interact inside the ViT. The difference therefore diagnoses how the effect of ViT background interaction is expressed at the answer position across LLM depth. More precisely, the vertical axis is the V0L0 − V1L0 *logit-lens margin*, rather than a generic V0 − V1 prediction margin.

The logit lens projects intermediate states through the final normalization and output head, providing a diagnostic readout rather than an intermediate prediction. We use it to measure when the V0L0 − V1L0 difference becomes readable. The shaded L7–9 interval is the independently selected direct-repair window from the third panel.

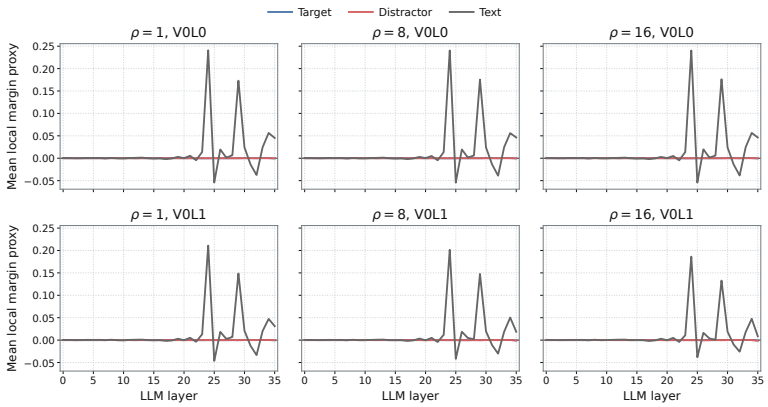

Figure 54: Value-output proxies at the answer-query position across LLM depth. Curves show projections from target visual tokens, distractor visual tokens, and visible nonvisual keys onto the condition-local correct-versus-strongest-wrong readout direction. The late signal is dominated by nonvisual sources, while the visual-source proxies remain near zero. Under V0L1, the late nonvisual proxy is lower than under V0L0, indicating weaker downstream readout under direct LLM background exposure. These quantities are local linear proxies.

The answer-state difference becomes readable only in later layers. The frozen L7–9 window shows essentially no logit-lens difference at the answer position. The difference becomes reliably positive around L21 and grows sharply after L25. This establishes that the recognition difference induced by ViT background interaction is not yet accessible to the final readout in the early direct-repair window; it is progressively expressed during later text and answer computation.

Which source states carry the late answer-aligned readout? We diagnose the source of the late signal by decomposing the attention output received by the answer query into value-output proxies from target visual tokens, distractor visual tokens, and visible nonvisual keys across LLM depth. Figure 54 uses matched V0L0 and V0L1 runs at each background ratio and projects each source contribution onto the condition-local correct-versus-strongest-wrong readout direction. Target and distractor visual proxies remain close to zero across depth, whereas the nonvisual proxy grows in the later LLM layers and supplies the visible answer-aligned signal. Under V0L1, this late nonvisual component is consistently weaker than under V0L0, with the largest reduction under stronger background competition. This establishes a downstream readout dominated by nonvisual states that have already integrated visual evidence, rather than a strong late contribution directly from target visual values. Because these quantities are local linear proxies, the following state interventions test the same pathway causally.

C.5 Recognition Recovery Through Visual and Text States

We extend the main matched-donor repair with two additional intervention families: a complete 36-layer scan using background-masked text donors and cross-ratio swaps that transfer text states between low- and high-background runs. Together, the three designs separate ViT-induced damage, direct LLM background exposure, and the quality of text states formed at different competition levels.

Target visual-state difference. For sample s, background ratio ρ, and zero-based decoder layer ℓ, let $H^{01}_{s,\rho,\ell,T}$ and $H^{11}_{s,\rho,\ell,T}$ denote the position-aligned target visual-token residual states under V0L1 and V1L1, respectively. States are recorded after the decoder block and its same-layer DeepStack contribution, before the next decoder block. We define the per-example state difference as

$$d_{s,\rho,\ell} = 1 - \frac{1}{N_{T,s}} \sum_{i \in T_s} \frac{\left\langle h^{11}_{s,\rho,\ell,i}, h^{01}_{s,\rho,\ell,i} \right\rangle}{\left\| h^{11}_{s,\rho,\ell,i} \right\|_2 \left\| h^{01}_{s,\rho,\ell,i} \right\|_2}. \tag{44}$$

The second panel of Figure 9 plots $D_{\rho,\ell} = N^{-1} \sum_s d_{s,\rho,\ell}$, where N is the number of examples. Thus, cosine similarity is computed separately for corresponding target tokens before av-

eraging over target tokens and examples; no centering or additional norm normalization is applied. Shaded intervals are pointwise 95% paired-sample bootstrap intervals based on 10,000 resamples of sample IDs shared across layers and ratios.

C.6 Isolating Direct LLM Background Effects in Text States

The main text shows that both ViT interaction and direct LLM background exposure affect the intermediate text states used for answering. Its V0L1 $\rightarrow$ V1L1 repair measures damage inherited from the ViT. The complementary V1L0 $\rightarrow$ V1L1 repair keeps ViT outputs and token layouts fixed while changing LLM access to background visual tokens. Improvement from the donor states therefore measures degradation created during LLM processing.

Writing m_{bg} for LLM access to background visual tokens, the matched contrast is

$$H_{\mathrm{ViT}}^{\mathrm{V1L0}} = H_{\mathrm{ViT}}^{\mathrm{V1L1}}, \qquad m_{\mathrm{bg}}^{\mathrm{V1L0}} = 0, \qquad m_{\mathrm{bg}}^{\mathrm{V1L1}} = 1. \tag{45}$$

At each decoder layer, we replace the recipient states of the instruction, question, and answer options with their position-aligned donor states. Visual tokens and the final answer-query state retain the recipient values, after which the remaining layers continue normally. The intervention measures whether text representations formed under background masking provide better evidence for the same downstream answer process.

Let $\mathcal{R}$ denote these pre-answer text positions. At the output of layer ℓ, the repair is

$$\widetilde{H}_\ell[i,:] = \begin{cases} H_\ell^{\mathrm{V1L0}}[i,:], & i \in \mathcal{R}, \\ H_\ell^{\mathrm{V1L1}}[i,:], & i \notin \mathcal{R}. \end{cases} \tag{46}$$

Scanning L0 to L35 produces the complete depth profile behind the fourth panel of Figure 9. The pooled gain peaks at L20, and layers L17 to L22 reach at least half of this peak. Direct LLM background exposure therefore becomes behaviorally consequential while visual evidence is integrated into text. We use this window and the fixed L20 comparison for the cross-ratio swaps in Appendix C.7, which tests representation adaptation across background ratios.

Formally, the pooled layerwise gain and the resulting half-maximum window are

$$\begin{aligned} G_\ell &= \mathbb{E}_{s,\,\rho\in\{1,8,16\}}\left[M_{s,\rho,\ell}^{\mathrm{repair}} - M_{s,\rho}^{\mathrm{V1L1}}\right], \\ \ell^* &= \arg\max_\ell G_\ell = 20, \qquad \mathcal{W}_{1/2} = \{\ell : G_\ell \geq \tfrac{1}{2} G_{\ell^*}\} = \{17,18,19,20,21,22\}. \end{aligned} \tag{47}$$

Together with the V0L1 $\rightarrow$ V1L1 repair in the main text, this contrast separates two routes to the same downstream bottleneck: ViT interaction degrades the evidence entering the LLM, whereas direct LLM exposure further degrades the text states that integrate that evidence. Their overlapping intermediate-layer recovery windows show where these otherwise distinct background effects converge during answer formation in the same middle-layer band.

C.7 Do High Background Text States Adapt or Degrade?

Appendix C.6 shows that direct LLM background exposure alters intermediate text states. To determine the value of these changes, we transfer states formed at a lower background ratio into a recipient with higher competition. Improvement from the lower ratio donor indicates that it retains more information useful for recognition.

We perform the transfers separately within V0L1 and V1L1, keeping the pathway fixed while changing the background ratio under which donor text states were formed. The three comparisons are $8 \leftarrow 1$, $16 \leftarrow 1$, and $16 \leftarrow 8$. At each layer from L17 to L22, the pre-answer text states are transplanted and the recipient resumes its downstream computation; L20 provides the fixed-depth summary. Comparing the two conditions also shows how lower-background text states interact with ViT-induced target degradation.

The plotted contrast is $C = M_{\mathrm{native}} - M_{\mathrm{swap}}$, where the native baseline is the unmodified recipient at its own background ratio. Positive C indicates an advantage of the native text states; negative

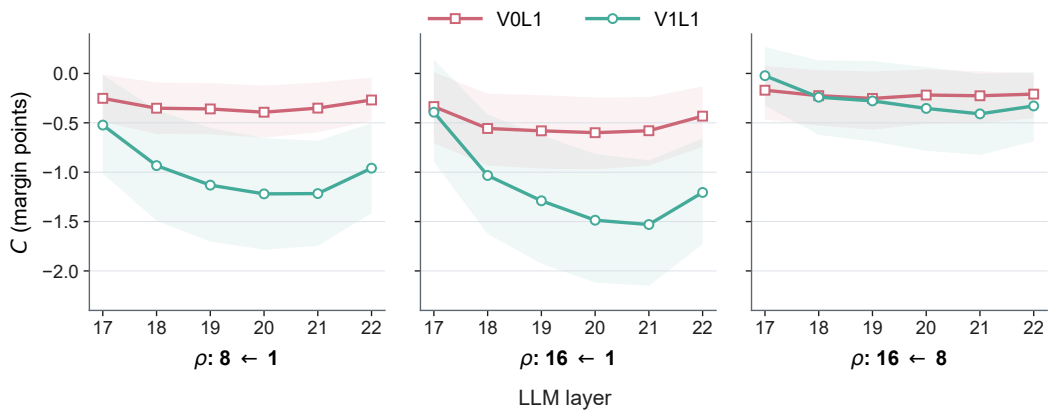

Figure 55: Cross ratio text state swaps across L17 to L22. Panels show recipient $\rho \leftarrow$ donor ρ for $8 \leftarrow 1$, $16 \leftarrow 1$, and $16 \leftarrow 8$ within V0L1 and V1L1. Each point represents a single layer swap; bands are pointwise 95% paired bootstrap confidence intervals. Negative $C = M_{\text{native}} - M_{\text{swap}}$ denotes improvement after swapping.

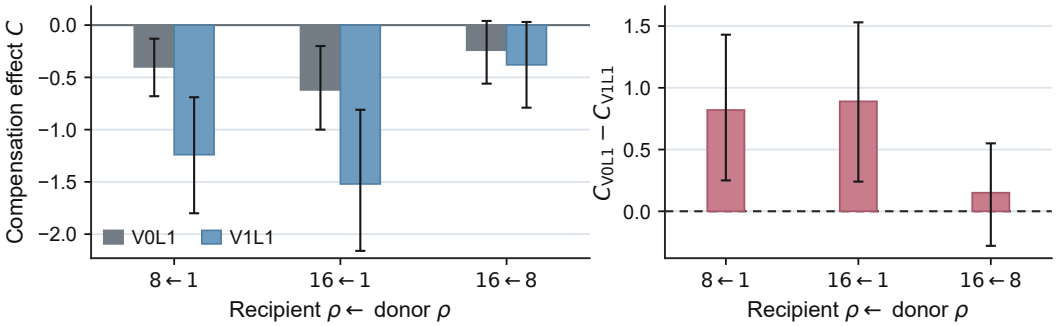

Figure 56: Cross-ratio text-state swaps at fixed L20 (the 21st decoder block). Panel A reports $C = M_{\text{native}} - M_{\text{swap}}$ within each V/L condition; negative values favor the swapped states. Panel B reports the paired difference $C_{\text{V0L1}} - C_{\text{V1L1}}$; a positive value means a larger swap gain under V1L1. Error bars show 95% paired-bootstrap confidence intervals.

C indicates that the cross-ratio swap improves the correct-answer margin. Figure 55 shows the six independently tested layers, and Figure 56 summarizes the fixed L20 comparison.

Lower-Background Text States Improve High-Background Recipients. Across L17 to L22, the layerwise curves are predominantly negative for all three transfers, so text states formed with less background usually improve the high-competition recipient rather than disrupting an adaptation needed at the larger ratio. At the fixed L20 comparison, the mean contrasts are negative in both V0L1 and V1L1 for $8 \leftarrow 1$, $16 \leftarrow 1$, and $16 \leftarrow 8$. The benefit is strongest for donors from $\rho = 1$, which provide the largest reduction in competition during state formation. For these transfers, the paired difference $C_{\text{V0L1}} - C_{\text{V1L1}}$ has a 95% interval above zero, showing that the low-background donor helps more when the recipient also carries ViT-induced target degradation; the interval for $16 \leftarrow 8$ includes zero. Intermediate text computation therefore does not simply adapt to the denser visual sequence. It progressively loses information useful for recognition, and this downstream loss compounds the degraded representation inherited from the ViT.

C.8 Two-Hop Relay from Target Tokens to the Answer

Section 4.2 localizes causally relevant target-visual and intermediate-text states by showing that repairing either can recover answer margin. State localization alone, however, does not establish whether these states form one transmission pathway or merely provide separate intervention sites. We therefore move beyond localizing causally relevant states to test how repaired target information is transmitted between them on its way to the answer. Following the two-stage organization used in prior cross-modal information-flow analyses (Zhang et al., 2025c; Yin et al., 2025; Lin et al., 2025), we denote the target visual tokens by T, the pre-answer text tokens by R, and the answer position by A, and test the relay $T \rightarrow R \rightarrow A$.

Tracing an upstream target repair to the answer. The primary experiment compares two matched trajectories. In the V1L0 baseline, target tokens interact with background inside the ViT and the LLM reads only the resulting target states. The repaired S1 trajectory replaces the target representation with its matched V0L0 state while preserving the remaining visual inputs and downstream computation. Their margin difference measures the value of repaired target information under background-masked LLM processing.

We then trace how the S1 gain travels. Each path intervention changes the contribution along one designated connection. Layerwise intervention localizes the first hop, from target visual tokens to pre-answer text, to L17–L21. The second hop, from those relay states to the answer position, follows at L20–L24. This ordered but overlapping depth profile is consistent with late target-to-text writing and early text-to-answer transmission acting on the same evolving relay states.

Each hop is tested in both directions. The sufficiency intervention inserts the repaired contribution into the baseline, and the necessity intervention inserts the contextualized contribution into S1. Agreement between these directions shows that the same connection carries and preserves the upstream repair benefit along the relay.

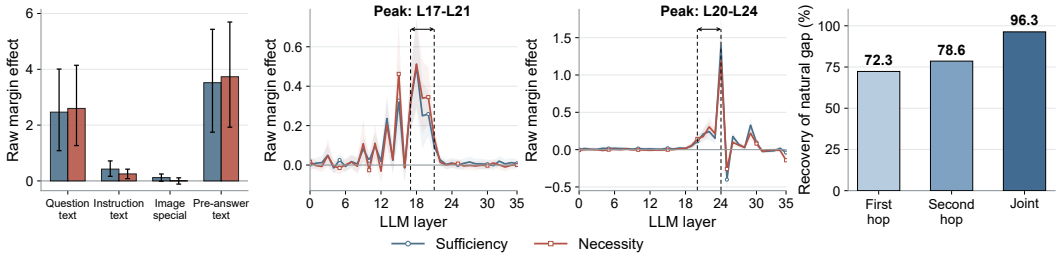

Figure 57: Primary two-hop causal relay result. From left to right, the panels identify the receiver class that carries repaired target information, localize the target-to-text and text-to-answer transfers across LLM depth using sufficiency and necessity, and report recovery of the upstream target-repair gain. Each hop carries a substantial but overlapping part of the same repaired signal; restoring both transfers jointly recovers 96.3%.

The first hop transfers repaired target information into text. The $T \rightarrow R$ effect concentrates at L17–L21, before the peak of the second hop. Its sufficiency and necessity effects are closely matched at 1.619 and 1.677 margin points, respectively. Averaging these two directional estimates recovers 72.3% of the mean S1 gain. The agreement between inserting the repaired contribution into the baseline and removing it from S1 shows that this earlier connection is both sufficient to transmit and necessary to preserve a substantial portion of the repaired target signal.

The second hop carries the relay signal to the answer. The $R \rightarrow A$ effect occurs later, at L20–L24, and again agrees across causal directions: the sufficiency and necessity effects are 1.864 and 1.718 margin points. Their directional mean recovers 78.6% of the S1 gain. The later depth window and the consistent bidirectional result identify pre-answer text as an active relay rather than a correlated readout of target information.

Jointly restoring the relay recovers 2.195 margin points, or 96.3% of the upstream repair gain. Let M_{00}^{relay} denote the polluted path, M_{11}^{relay} the condition with both transfers restored, and M_{S1} the naturally propagated repaired trajectory. The 96.3% recovery is the following ratio of the factorial joint-repair effect to the mean S1 repair gain:

$$R_{\text{relay}} = \frac{\mathbb{E}[M_{11}^{\text{relay}} - M_{00}^{\text{relay}}]}{\mathbb{E}[M_{\text{S1}} - M_{00}^{\text{relay}}]} = 0.963, \tag{48}$$

with a 95% interval of 92.0%–102.7%. Thus, the complete $T \to R \to A$ route accounts for nearly all of the behavioral effect produced by repairing the upstream target state.

The two single-hop effects are not additive because they transmit the same repaired signal along successive steps. Letting $a, b \in \{0, 1\}$ indicate restoration of the first and second hops,

$$M_{ab}^{\text{relay}} = M_{00}^{\text{relay}} + aE_1 + bE_2 + abI_{12}, \tag{49}$$

where $E_1 = M_{10}^{\text{relay}} - M_{00}^{\text{relay}}$, $E_2 = M_{01}^{\text{relay}} - M_{00}^{\text{relay}}$, and $I_{12} = M_{11}^{\text{relay}} - M_{10}^{\text{relay}} - M_{01}^{\text{relay}} + M_{00}^{\text{relay}}$.

Table 13 summarizes the corresponding margin effects. The factorial main effects are $E_1 = 1.649$ and $E_2 = 1.883$, while their joint effect is 2.195 margin points. The interaction $I_{12} = -1.337$ quantifies the expected overlap between two interventions on successive stages of the same signal path rather than two independent sources of gain.

Analysis	Estimand	Mean
Directional	Hop 1, suff.	1.619
	Hop 1, nec.	1.677
	Hop 2, suff.	1.864
	Hop 2, nec.	1.718
Factorial	E_1 (hop 1)	1.649
	E_2 (hop 2)	1.883
	$M_{11}^{\text{relay}} - M_{00}^{\text{relay}}$ (both)	2.195
	I_{12} (interaction)	-1.337
Reference	S1 repair gain	2.279

Table 13: Two-hop relay margin effects.

The effect depends on the relay positions and transferred content. Matched non-relay positions change the margin by only -0.001 points, and bypassing the relay with direct $T \to A$ repair changes it by only 0.011 points. An equal-norm orthogonal-content intervention also fails to recover the signal and instead decreases the margin by 0.806 points. Conversely, reversing the intervention direction reproduces 2.109 points of recovery. The effect therefore depends on both the relay positions and the transferred content, rather than arising from a generic perturbation or direct target-to-answer shortcut.

Together, these results turn the state-repair localization in Section 4.2 into a causal transmission account. Early target-state repair improves the answer because its information is first written into pre-answer text at L17–L21 and then passed from those relay states to the answer at L20–L24. The near-complete joint recovery connects the causally relevant states identified in the main text into one ordered pathway, while the near-zero alternatives rule out spatially nonspecific transfer, arbitrary replacement content, and a direct visual-to-answer explanation.

Measuring the two transfers. We quantify the two relay transfers using a common set of causally valid receiver queries. For the first hop, the source is the target visual-token set $S_1 = T$, the receivers are the pre-answer text positions $Q_1 = R$, and the evaluation window is $W_1 = \{17, 18, 19, 20, 21\}$. For the second hop, the source is the relay-text set $S_2 = R$, the receiver is the answer position $Q_2 = \{A\}$, and the evaluation window is $W_2 = \{20, 21, 22, 23, 24\}$.

For both hops, we enforce causal access through the following valid receiver set:

$$Q_j^{\text{valid}} = \{q \in Q_j : \max(S_j) \le q\}.$$

The valid receiver set contains positions that occur after their source. This excludes two pre-answer text positions that precede the target image; the answer query in the second hop always satisfies this causal-order condition by construction.

B24 target-state repair. The Shape repair and Full repair conditions in Table 14 intervene immediately after the output of zero-based ViT block B24 (the 25th block). The intervention acts only on the target-token rows: Full repair replaces those rows with their matched clean-donor states, whereas Shape repair replaces only their centered component while retaining the recipient centroid. All non-target rows remain on the recipient trajectory. The resulting hybrid state is then

propagated normally through the remaining ViT blocks, DeepStack computation, visual merger, and LLM, with no further replacement. We fix B24 across all background ratios because it provides a strong and stable repair site rather than selecting a ratio-specific optimum.

For layer ℓ, head h, receiver query q, and source key k, let

$$s_{\ell hqk} = \gamma_\ell \langle \mathrm{RoPE}(q_{\ell hq}), \mathrm{RoPE}(k_{\ell hk}) \rangle + m_{qk},$$

where γ_ℓ is the model's native attention scaling factor and m_{qk} is the causal attention mask. The corresponding attention weight follows from the source-key softmax:

$$\alpha_{\ell hqk} = \mathrm{softmax}_k \left(s_{\ell hqk} \right).$$

Routing. Routing measures the absolute attention mass assigned to the source set:

$$\mathrm{Routing}_j = \frac{1}{|W_j|} \sum_{\ell \in W_j} \frac{1}{H |Q_j^{\mathrm{valid}}|} \sum_{h=1}^{H} \sum_{q \in Q_j^{\mathrm{valid}}} \sum_{k \in S_j} \alpha_{\ell hqk}.$$

QK compatibility. QK compatibility is the following window-averaged log-mean-exp source compatibility over valid receiver queries:

$$\mathrm{QK}_j = \frac{1}{|W_j|} \sum_{\ell \in W_j} \frac{1}{H |Q_j^{\mathrm{valid}}|} \sum_{h=1}^{H} \sum_{q \in Q_j^{\mathrm{valid}}} \log \left[\frac{1}{|S_j|} \sum_{k \in S_j} \exp \left(s_{\ell hqk} \right) \right].$$

This source-set log-mean-exp score complements the answer-position Δ_{QK} used in Section 4.1.

Content fidelity. For condition C, we define content fidelity from the following source-edge attention-output contribution received by query q on hop j:

$$\mathbf{c}_{\ell q}^{C,j} = W_O^\ell \left[\bigoplus_{h=1}^{H} \sum_{k \in S_j} \alpha_{\ell hqk}^C \mathbf{v}_{\ell hk}^C \right].$$

We compare this contribution with the matched V0L1 reference from the same sample, background ratio, layer, and hop using an ϵ-stabilized cosine similarity:

$$F_j^C = \frac{1}{|W_j|} \sum_{\ell \in W_j} \frac{1}{|Q_j^{\mathrm{valid}}|} \sum_{q \in Q_j^{\mathrm{valid}}} \frac{\mathbf{c}_{\ell q}^{C,j} \cdot \mathbf{c}_{\ell q}^{\mathrm{V0L1},j}}{\max \left(\|\mathbf{c}_{\ell q}^{C,j}\|_2, \epsilon \right) \max \left(\|\mathbf{c}_{\ell q}^{\mathrm{V0L1},j}\|_2, \epsilon \right)}, \qquad \epsilon = 10^{-8}.$$

The table reports arithmetic means over the matched examples.

The first hop carries the main degradation. The relay experiment above uses L0 to trace the propagation of repaired target information, while Table 14 measures the same transfers under natural LLM background access. All three measures use the same causally valid receiver queries, so Routing, QK compatibility, and content fidelity describe a common transfer rather than different token subsets. The first hop shows a competition-dependent loss in both access and transmitted content. At $\rho = 16$, target-to-relay QK compatibility decreases from 4.514 under V0L1 to 4.081 under V1L1, Routing decreases from 0.0515 to 0.0362, and fidelity falls from approximately 1.000 to 0.581. Shape repair raises fidelity to 0.713, while full-state repair raises it to 0.783 and restores compatibility to 4.313. The same ordering is already present at $\rho = 1$ and strengthens at $\rho = 8$ and 16, showing that background interaction progressively changes what target visual tokens can write into the text stream. The repair does more than increase attention mass: it restores the alignment and content of the contribution that reaches the relay.

The second hop is comparatively stable in how it reads the relay. At $\rho = 16$, Routing and QK compatibility are nearly identical between V0L1 and V1L1 (0.4827 versus 0.4854, and 7.916 versus 7.925). Its remaining difference lies mainly in content fidelity, which falls to 0.936 under V1L1 and returns to 0.974 and 0.983 after Shape and full-state repair. The behavioral margins follow this restoration: at $\rho = 8$, Shape and full repair raise the V1L1 margin from 2.570 to

		First hop: $T \to R$			Second hop: $R \to A$		
ρ	Condition	Routing	QK	F_1	Routing	QK	F_2
1	V0L1	0.0837	4.405	1.000	0.4782	7.225	1.000
1	V1L1	0.0790	4.329	0.795	0.4792	7.241	0.975
1	Shape repair	0.0807	4.329	0.888	0.4792	7.240	0.991
1	Full repair	0.0808	4.339	0.899	0.4790	7.238	0.993
8	V0L1	0.0593	4.467	1.000	0.4809	7.664	1.000
8	V1L1	0.0441	4.109	0.663	0.4827	7.658	0.948
8	Shape repair	0.0469	4.153	0.782	0.4825	7.659	0.980
8	Full repair	0.0499	4.313	0.821	0.4814	7.640	0.985
16	V0L1	0.0515	4.514	1.000	0.4827	7.916	1.000
16	V1L1	0.0362	4.081	0.581	0.4854	7.925	0.936
16	Shape repair	0.0393	4.151	0.713	0.4855	7.914	0.974
16	Full repair	0.0417	4.313	0.783	0.4841	7.890	0.983

Table 14: B24 target-row repair diagnostics over the matched examples, where B24 denotes an intervention after the output of zero-based ViT block 24. The first hop connects target visual tokens to relay text over L17 to L21, and the second connects relay text to the answer over L20 to L24. Routing is source attention mass, QK is source compatibility, and F_1 and F_2 measure contribution fidelity relative to V0L1.

4.092 and 4.051, approaching the V0L1 margin of 4.214; at $\rho = 16$, they raise it from 2.020 to 3.828 and 3.868, compared with 3.929 for the reference. A complementary decomposition beginning from the fully contextualized trajectory gives the same separation and negligible effects from mismatched donors. Together, Figure 57 and Table 14 close the serial account: the main loss occurs when target information is written into relay text, the answer position continues to read that relay with similar routing and compatibility, and restoring the complete $T \to R \to A$ route reproduces nearly all of the upstream repair benefit.

C.9 Full Q/K/V Factorial Analysis

Appendix C.8 establishes a serial transfer from target visual tokens to relay text and then to the answer. We next identify which attention components support each hop, separating receiver queries, source keys, and source values within the reconstructed source contribution.

The factorial uses a V1L1 **recipient and matched** V0L1 **donor.** The analysis uses the same examples, relay positions, and layer windows as Appendix C.8. The two hops connect target visual tokens to pre-answer relay text over L17 to L21 and relay text to the answer over L20 to L24 under natural LLM background access. For each hop, the intervention reconstructs the designated source contribution before residual addition. The live non-source output remainder is preserved, and subsequent computation proceeds normally from the intervened hidden state.

For each hop, Q is the receiving position's query, whereas K and V are the source tokens' keys and values. We use P for a component from the V1L1 recipient and C for the corresponding component from the matched V0L1 donor. Triplets are ordered as (Q, K, V). CPP, PCP, and PPC replace only Q, K, and V, respectively, within the reconstructed source contribution. CCP, CPC, and PCC replace QK, QV, and KV; and CCC replaces all three components. PPP is the unrepaired V1L1 baseline. Figure 58 reports final answer-margin changes relative to PPP as bars and QK compatibility at the intervened transfer as dashed curves.

Factorial contrasts keep the non-intervened component fixed. In the K/V subfactorial, the receiver query remains fixed to its polluted value $Q = P$. Let $k, v \in \{0, 1\}$ indicate replacement of K and V by their V0L1 counterparts. The four conditions admit

$$M_{kv}^{Q=P} = M_{\mathrm{PPP}} + kE_K^{Q=P} + vE_V^{Q=P} + kvI_{KV}^{Q=P}, \tag{50}$$

The corresponding main effects and interaction are defined by

$$\begin{aligned} E_K^{Q=P} &= M_{\mathrm{PCP}} - M_{\mathrm{PPP}}, \\ E_V^{Q=P} &= M_{\mathrm{PPC}} - M_{\mathrm{PPP}}, \\ E_{KV}^{Q=P} &= M_{\mathrm{PCC}} - M_{\mathrm{PPP}}, \\ I_{KV}^{Q=P} &= M_{\mathrm{PCC}} - M_{\mathrm{PCP}} - M_{\mathrm{PPC}} + M_{\mathrm{PPP}}. \end{aligned} \tag{51}$$

Analogously, the Q and K interaction with $V = P$ fixed is

$$I_{QK}^{V=P} = M_{\mathrm{CCP}} - M_{\mathrm{CPP}} - M_{\mathrm{PCP}} + M_{\mathrm{PPP}}. \tag{52}$$

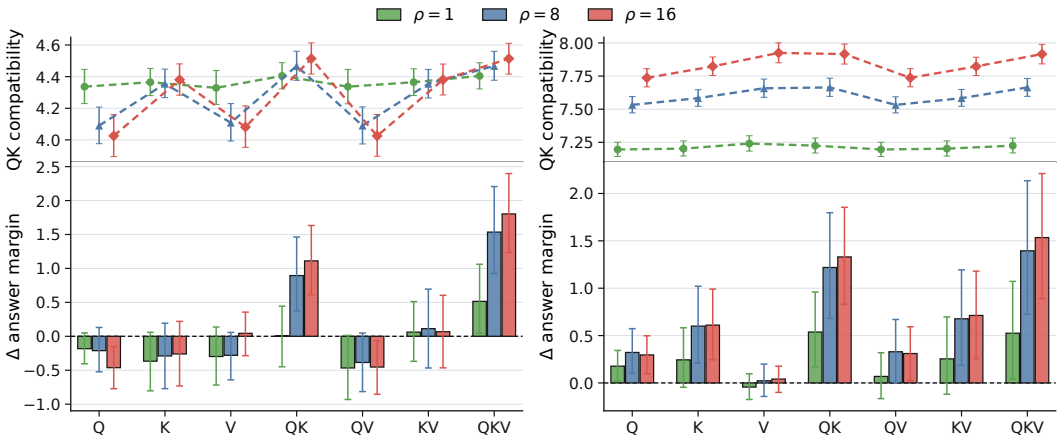

Figure 58: Q/K/V decomposition of the two relay transfers. Panel A connects target visual tokens to relay text over L17 to L21, and Panel B connects relay text to the answer over L20 to L24. P components come from the V1L1 recipient, and C components come from the matched V0L1 donor. Bars show answer margin changes relative to PPP; dashed curves show QK compatibility. The limited recovery from K and KV under $Q = P$, compared with QK and QKV, identifies joint query and key matching as the dominant first hop interaction.

The first hop depends primarily on joint Q and K matching, and the second hop shows the same preference more weakly. Figure 58A shows that repairing K can raise local compatibility without recovering the answer when the receiving query remains on the degraded trajectory. At $\rho = 16$, the margin effects of Q, K, V, QK, QV, KV, and QKV repair are -0.463, -0.260, 0.044, 1.111, -0.453, 0.068, and 1.803 points. The K effect and the KV effect both include zero, and the K and V interaction is small and unresolved at 0.284 (95% CI $[-0.080, 0.659]$). By contrast, the Q and K interaction reaches 1.834 (95% CI $[1.311, 2.391]$). The first transfer therefore requires a receiver query that is compatible with the repaired source keys; changing the source representation alone does not reconstruct the computation that reads it.

The second hop is more individually sensitive to K at high background ratios, but its largest effects still occur for QK and QKV. Across both panels, the dashed compatibility curves and behavioral bars serve complementary roles: QK repair restores access to the source set, while the final margin also depends on the values and downstream state produced by the full configuration. All non-source keys used in the hybrid source-term softmax retain their V1L1 references. These effects therefore localize the interaction to the designated transfer rather than a global donor replacement. The shared pattern motivates the next experiment, which asks whether the natural recipient query has become specifically matched to the degraded evidence it encounters.

C.10 Evidence of Damage-Specific Query Compensation

Appendix C.9 shows that the receiver query participates in the Q/K/V interactions that transmit target evidence. We next ask whether the query computed on the contextualized trajectory merely inherits

upstream damage or becomes specifically adapted to the degraded K/V evidence it encounters. We use the same examples at $\rho = 1, 8, 16$ and scan L0 to L35. At each layer, the selected source contribution is recomputed with a specified Q/K/V combination before residual addition, while all other rows and subsequent computation remain on the V1L1 trajectory. Figure 59 reports both paths: the upper panel follows target visual tokens into relay text, and the lower panel follows relay text into the answer. Within each panel, the left column measures the natural query benefit under degraded evidence, the middle column measures the same query under reference evidence, and the right column isolates the resulting damage-specific interaction.

The four conditions form a 2×2 crossing of query source and K/V evidence source. With degraded K/V from V1L1 fixed, we compare the natural V1L1 query, $M(\mathrm{PPP})$, with the reference V0L1 query, $M(\mathrm{CPP})$. With reference K/V from V0L1 fixed, we analogously compare $M(\mathrm{PCC})$ with $M(\mathrm{CCC})$. Thus, within the reconstructed source contribution, each contrast changes only the query reference while holding the source K/V evidence fixed. Let $M(Q, K, V)$ denote the resulting correct-answer margin. For each sample, ratio, path, and layer, we evaluate

$$\begin{aligned} M_{00}^{\mathrm{QE}} &= M(\mathrm{CCC}), \quad M_{01}^{\mathrm{QE}} = M(\mathrm{PCC}), \\ M_{10}^{\mathrm{QE}} &= M(\mathrm{CPP}), \quad M_{11}^{\mathrm{QE}} = M(\mathrm{PPP}) = M_{\text{natural V1L1}}, \end{aligned} \tag{53}$$

For this QE family, the first and second indices identify the K/V-evidence and query sources, respectively, with $0 = C$ and $1 = P$; hence M_{11}^{QE} is the unrepaired natural V1L1 condition. The query effects under degraded and reference K/V evidence, and their interaction, are

$$\begin{aligned} C_{\mathrm{damaged}} &= M_{11}^{\mathrm{QE}} - M_{10}^{\mathrm{QE}}, \\ C_{\mathrm{clean}} &= M_{01}^{\mathrm{QE}} - M_{00}^{\mathrm{QE}}, \\ I_{\mathrm{comp}} &= C_{\mathrm{damaged}} - C_{\mathrm{clean}}. \end{aligned} \tag{54}$$

A positive C_{damaged} means that the natural query uses degraded evidence better than the reference query; a positive I_{comp} means that this advantage is larger than under reference evidence. Their joint positive sign therefore provides evidence of damage-specific query compensation. Inference uses 20,000 paired sample-level bootstrap resamples with shared sample indices across layers, conditions, and ratios; Figure 59 reports pointwise 95% intervals.

The compensation profile develops in two stages. On the target-to-relay path, C_{damaged} peaks at L18, L15, and L18 for $\rho = 1, 8, 16$, with values of 0.0677, 0.0646, and 0.1313. The exact peak shifts, but all three occur in the same middle-layer band where target information is entering text states. In this region, C_{damaged} becomes positive while C_{clean} becomes negative, producing a positive I_{comp} that is largest at $\rho = 16$. The sign reversal is important: the natural query is not uniformly stronger than the reference query; it is better matched to degraded K/V and worse matched to reference K/V. Although the pointwise intervals overlap zero on this first hop, the aligned middle-layer profiles provide a descriptive localization of query and evidence coadaptation during the target-to-relay handoff.

Path	ρ	Layer	Mean
First	1	L18	0.0677
	8	L15	0.0646
	16	L18	0.1313
Second	1	L25	0.1604
	8	L25	0.3052
	16	L25	0.3094

Table 15: Peak C_{damaged} margin differences by layer.

The relay-to-answer path converts that conditional structure into a larger and statistically resolved margin contribution. Its peak occurs at L25 for every ratio, with C_{damaged} values of 0.1604, 0.3052, and 0.3094, and all three pointwise 95% intervals above zero. The benefit nearly doubles between $\rho = 1$ and $\rho = 8$ and then remains stable at $\rho = 16$, indicating that compensation strengthens once degradation is substantial and then saturates. At the same layer, negative C_{clean} and positive I_{comp} again show specialization to the degraded evidence rather than a generally improved query. Downstream computation therefore does more than passively inherit the upstream loss: it develops queries adapted to the source K/V states during target-to-text transfer and turns that matching into measurable compensation near answer formation. This compensation is real but incomplete, which is consistent with the remaining recognition gap under high-background competition.

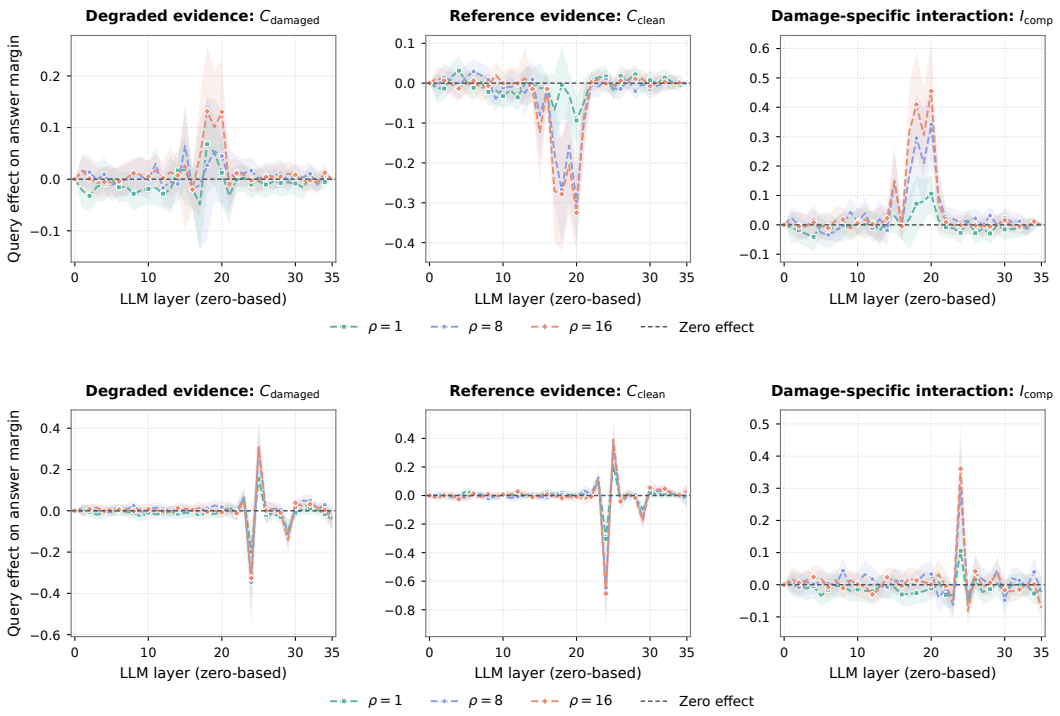

Figure 59: Damage-specific query compensation along the two relay paths. The upper panel shows target visual tokens to relay text ($T \rightarrow R$), and the lower panel shows relay text to the answer ($R \rightarrow A$). Columns report the natural query benefit under degraded evidence, its effect under reference evidence, and the resulting interaction.

C.11 Spectral Decomposition of LLM Target States

We close the downstream analysis by decomposing the conspicuous mid-layer spectral collapse around L16 in LLM target states. Across examples and ratios, we select the 190 states whose raw rank falls below two at that layer and measure their first principal-component variance, maximum token norm, maximum-to-median norm ratio, and largest absolute hidden activation. Figure 60 shows that the collapse is produced by one exceptional token rather than a collective compression of the target representation. Removing the largest-norm token and recentering raises the mean effective rank at L16 to 38.50, returning it to the earlier range.

The right panel identifies the common source of this behavior. The first principal component explains a mean 98.85% of centered variance, increasingly extreme token-norm ratios approach complete PC1 dominance, and hidden dimension 2276 contains the largest absolute activation in all 190 selected states. A repeated activation in one token and one shared dimension therefore creates the apparent low-dimensional geometry. Despite its visual prominence, the signature is weakly associated with answer margin (Pearson $r = -0.114$, Spearman $\rho = -0.061$). This mid-layer collapse should consequently be treated as a recurrent geometric artifact rather than evidence of the recognition mechanism; the behaviorally consequential states are those identified by the causal repair and relay analyses above.

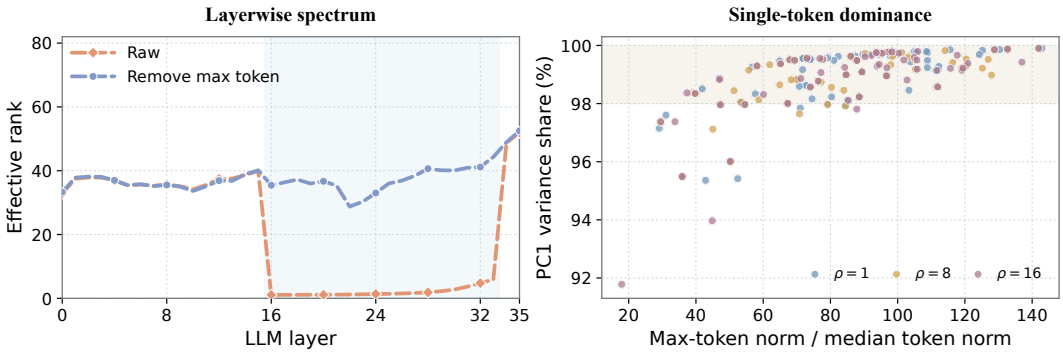

Figure 60: Decomposition of the mid-layer spectral collapse around L16. Left: raw centered effective rank drops sharply at L16, while removing the largest norm target token and recentering the remaining tokens restores the earlier rank range. Right: stronger maximum token dominance corresponds to near complete first PC concentration across the 190 selected states.