

Target-Binding Hallucination: A Position-Controlled Diagnostic for Multimodal LLMs in Multi-Image Contexts

Qichao Ma^{1*}

¹Liaoning University

Shenyang, China

20231491112@smail.lnu.edu.cn

Yongtai Liu^{2*†}

²Hanyang University

Seoul, Republic of Korea

yt2024259263@hanyang.ac.kr

Xuan Zhu^{3*}

³Central South University

Changsha, China

zhuxuan1@csu.edu.cn

Abstract

Multimodal LLMs are increasingly used with multi-image, interleaved image–text inputs, yet hallucination evaluation remains largely single-image. We study target-binding hallucination: a model answers as if an object appears in the queried target image when it appears only elsewhere. We introduce a position-controlled COCO haystack with POPE-style object-existence queries and split negatives into clean cases (object absent everywhere) and distractor-present cases (absent from the target but present elsewhere). On InternVL2-8B at $N = 10$, clean false positives remain near floor (1.15%), whereas distractor-present cases reach 10.55% and show a position-sensitive profile; the same regime-level separation appears across three additional architectures and in an object-balanced PASCAL VOC2012 check. Channel and paired-removal controls are consistent with a visual-channel-dominant source-misattribution interpretation, and a related MIHBench Existence re-audit finds that over 98% of false positives concentrate in partial-present negatives. For object-existence diagnostics and other source-sensitive multi-image evaluations with auditable negative evidence, we recommend reporting clean and evidence-supported negatives separately whenever non-target images may contain supporting evidence.

1 Introduction

Large vision-language models (LVLMs) have made rapid progress on multimodal understanding (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023a; Dai et al., 2023; Liu et al., 2023; Chen et al., 2024c; Wang et al., 2024a), yet they still hallucinate objects (Rohrbach et al., 2018; Li et al., 2023b; Wang et al., 2023; Liu et al., 2024b). This problem is usually evaluated in single-image

settings, whereas practical LVLM use increasingly involves multi-image, interleaved image–text contexts such as multi-page reports, visual histories, and retrieval-augmented multimodal inputs. In these settings, the model must not only recognize objects, but also bind evidence to the correct source image.

In such contexts, a false positive can mean either that the queried object is absent from the entire sequence, or that it appears in a non-target image and is incorrectly bound to the queried target image. For example, if a traffic light appears only in Image 3 but the user asks about Image 6, a Yes answer reflects source confusion rather than pure object invention. Aggregate false-positive rates conflate these qualitatively different failures, so the same metric can count errors with different causes and different implications for evaluation.

We call the second failure mode target-binding hallucination; source misattribution denotes the source-confusion interpretation supported by our controls. Related phenomena have been named as Sequence Confusion, Context Conflation, and cross-image information leakage in recent mitigation work (Liu et al., 2025b; Li et al., 2025c; Park et al., 2025). We do not claim to be first to observe cross-image misattribution. Instead, our contribution is to make one target-image regime measurable: we separate it from globally unsupported invention, hold position and distractor evidence under control, and test whether aggregate false positives are driven by evidence that is real but incorrectly scoped.

We construct a COCO val2017 haystack with controlled target-image depth and image–caption interleaving (Figure 1), and evaluate POPE-style object-existence queries (Li et al., 2023b). Negatives are split into clean cases, where the queried object appears nowhere, and distractor-present cases, where it is absent from the target but present in at least one distractor. We treat this haystack as

*Equal contribution.

†Corresponding author.

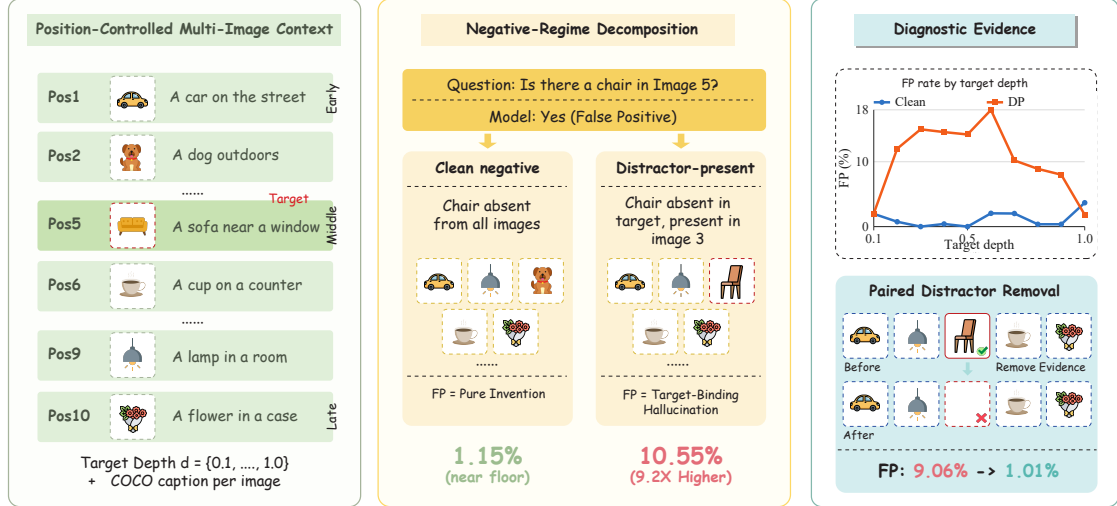


Figure 1: Target-binding hallucination diagnostic. We build position-controlled COCO haystacks with interleaved image–caption contexts, ask POPE-style target-image existence questions, and split negatives into *clean* and *distractor-present* regimes. The schematic highlights that false positives have different interpretations across regimes and summarizes the depth-wise false-positive pattern together with a paired distractor-removal counterfactual. Image strips and icons are schematic.

a controlled diagnostic rather than a comprehensive benchmark, and report Target Attention Score (TAS) only as an exploratory InternVL2-specific correlate.

Scope of the diagnostic. We deliberately begin with binary object existence as a foundational source-binding test, rather than as a comprehensive account of multimodal hallucination. This setting makes target-image absence and non-target supporting evidence directly auditable, and enables paired input-level interventions such as replacing the sole supporting distractor. Whether the same clean/evidence-supported decomposition transfers to attributes, relations, counting, identity consistency, fine-grained grounding, or free-form generation remains open; these settings require task-specific definitions of evidence scope and corresponding counterfactual controls.

The main decomposition shows that, on InternVL2-8B at $N = 10$, clean negatives remain near floor (1.15%) while distractor-present negatives reach 10.55% and account for the apparent position-dependent aggregate profile. We then test alternative explanations with input-channel, prompt-format, paired-removal, fixed-distance, cross-architecture, and MIHBench re-audit controls. The resulting claim is intentionally diagnostic rather than mechanistic: multi-image false positives should be interpreted by their evidence regime, and not only by their aggregate rate.

We separate three levels of claim. Empirically, we show a clean/distractor-present separation in a controlled object-existence diagnostic. Interpretively, our controls indicate that non-target visual evidence is a major driver of these false positives. Methodologically, our recommendation is limited to source-sensitive multi-image evaluations where negative queries may contain auditable supporting evidence outside the target.

Our contributions are: (i) a controlled diagnostic protocol for target-image source attribution; (ii) a clean/distractor-present decomposition showing that aggregate false positives conflate unsupported invention with distractor-supported misattribution; (iii) paired, channel, position, and benchmark-audit controls consistent with a visual-channel-dominant source-misattribution interpretation; and (iv) cross-architecture checks with an exploratory TAS analysis. The intended output is not a new leaderboard, but a reporting practice: for object-existence diagnostics and other source-sensitive multi-image evaluations with auditable negative evidence, negative errors should be separated into clean and evidence-supported regimes when non-target images may contain supporting evidence.

2 Related Work

Hallucination evaluation. Object hallucination has been widely studied in single-image caption-

ing and VQA, including CHAIR, POPE, AMBER, and broader LVLM suites (Rohrbach et al., 2018; Li et al., 2023b; Wang et al., 2023; Fu et al., 2025; Liu et al., 2024e; Guan et al., 2024). POPE-oriented work such as H-POPE and RePOPE is especially relevant because our diagnostic also uses object-existence queries and COCO-derived labels (Pham and Schott, 2024; Neuhaus and Hein, 2025). Rather than relabeling single-image benchmarks, we ask whether multi-image false positives are globally unsupported or supported by incorrectly scoped non-target evidence. This distinction is especially important because annotation noise and object-presence ambiguity can change the meaning of a negative query. We therefore include a manual audit and conservative sensitivity check for the clean slice, rather than relying only on COCO annotations. We diagnose rather than mitigate hallucination, complementing instruction tuning, preference alignment, decoding, and correction methods (Liu et al., 2024a; Yu et al., 2024; Leng et al., 2024; Chen et al., 2024a; Yin et al., 2024).

Multi-image and long-context LVLMs. Recent work evaluates multi-image, interleaved, or long-context LVLM behavior, including MIBench, MMIU, MuirBench, MileBench, MM-NIAH, Visual Haystacks, LLaVA-NeXT-Interleave, and LongHalQA (Liu et al., 2024c; Meng et al., 2025; Wang et al., 2025; Song et al., 2024; Wang et al., 2024b; Wu et al., 2025; Li et al., 2025a; Qiu et al., 2024). MIHBench is closest among existing benchmarks because it studies multi-image hallucination under object-centric tasks (Li et al., 2025b). It broadens coverage across models and task types, but its aggregate hallucination numbers do not by themselves distinguish globally unsupported invention from evidence that is present but incorrectly scoped. Our setting is narrower but more diagnostic: it varies target depth under image–text interleaving and decomposes negative queries into globally absent versus distractor-supported regimes, enabling counterfactual edits that are difficult to define in broad free-form benchmarks. This allows paired counterfactuals such as replacing the sole supporting distractor, which broader benchmark suites typically cannot isolate. We also re-audit MIHBench Existence under a related clean/partial-present split to show that regime mixing is not unique to our constructed haystack, while keeping clear that MIHBench asks all-image rather than target-image questions.

Misbinding, position, and attention. Recent mitigation papers name related failures as Sequence Confusion, Context Conflation, or cross-image information leakage (Liu et al., 2025b; Li et al., 2025c; Park et al., 2025). Our contribution is not a new mitigation method, but a controlled audit with paired interventions and caption-channel controls. In particular, we ask whether the same false-positive label should be interpreted differently depending on whether supporting evidence exists in another image. Work on long-context position bias and multimodal attention motivates our position-sensitive and TAS analyses (Liu et al., 2024d; Bai et al., 2024; Hsieh et al., 2024; Press et al., 2022; Tian et al., 2025; Liu et al., 2025a; Xiao et al., 2024; Kang et al., 2025; Huang et al., 2024; Clark et al., 2019; Michel et al., 2019). Because attention is difficult to interpret causally, we use TAS only as a within-model observational correlate.

3 Method

3.1 Problem and Negative Regimes

We study an interleaved sequence of N images and captions,

$$\mathcal{S} = (I_1, C_1, I_2, C_2, \dots, I_N, C_N), \quad (1)$$

where C_i denotes the caption paired with image I_i . Let j be the target-image position and $d = j/N$ its relative depth. The model is asked whether object o appears in I_j ; a false positive occurs when $o \notin I_j$ but the model answers Yes.

We split negatives into clean cases, where o is absent from every image, and distractor-present cases, where $o \notin I_j$ but appears in at least one non-target image:

$$r(I_j, o) = \begin{cases} \text{CLEAN}, & \sum_{i \neq j} \mathbf{1}[o \in I_i] = 0, \\ \text{DP}, & \sum_{i \neq j} \mathbf{1}[o \in I_i] > 0. \end{cases} \quad (2)$$

Thus the same Yes answer can indicate either globally unsupported invention or source misattribution, where real contextual evidence is treated as if it belonged to the queried image.

3.2 Haystack Benchmark and POPE Queries

We construct a COCO val2017 haystack benchmark (Lin et al., 2014). A fixed seed partitions annotated images into a 500-image target pool and a distractor pool. For each sample, we select a target image and sample $N - 1$ distractors after removing images that share target-image object cate-

gories, so positive queries are not trivially answerable from non-target images. For negative queries, the queried object is absent from the target by construction but may appear in distractors; this design creates both clean and distractor-present negatives from the same haystack. Each negative is labeled by Eq. 2 using COCO annotations. A 240-sample manual spot-check confirms target absence in 93.75% of cases and regime agreement in 96.10% of determinate cases; conservatively treating all uncertain or target-present clean false positives as non-errors leaves a confirmed clean-FP rate of 0.89% (23/2,598), far below the distractor-present rate of 10.55% (Appendix A.7). Each image is paired with its COCO caption; prompt details are in Appendix A.2.

Construction protocol. For each context, we (1) sample a target image and choose the target position; (2) sample $N - 1$ distractors after filtering target-category overlaps; (3) form an interleaved image-caption sequence; (4) sample positive objects from the target and negative objects absent from the target; and (5) label each negative as clean or distractor-present using non-target COCO annotations. This compact protocol is intended to make the diagnostic reproducible without relying on aggregate benchmark scores alone.

For $N = 10$, depths $d \in \{0.1, \dots, 1.0\}$ map one-to-one to positions $j = \text{round}(dN)$; $N = 5$ is used mainly as a context-length comparison (Appendix A.1). We pair each context with POPE-style positive and negative queries using Random, Popular, and Adversarial strategies (Li et al., 2023b). For each $N \in \{5, 10, 20\}$, we construct 2,000 contexts and three negative queries per context, yielding 6,000 negative model-query instances per N ; $N = 20$ is a limited low-token diagnostic (Appendix F.1).

3.3 Target Attention Score

TAS is an exploratory, InternVL2-specific observational summary of attention from question tokens to target-image visual tokens. Let $A^{(\ell^*)} \in \mathbb{R}^{H \times M \times M}$ denote the attention tensor at the final transformer layer ℓ^* , where H is the number of attention heads and M is the total multimodal sequence length. Let $\mathcal{V}_j$ be the visual-token indices of target image I_j , and let $\mathcal{Q}$ be the question-token

indices. We define

$$\begin{aligned} \bar{A}_{qv}^{(\ell^*)} &= \frac{1}{H} \sum_{h=1}^H A_{h,q,v}^{(\ell^*)}, \\ \text{TAS}(j) &= \frac{1}{|\mathcal{Q}| |\mathcal{V}_j|} \sum_{q \in \mathcal{Q}} \sum_{v \in \mathcal{V}_j} \bar{A}_{qv}^{(\ell^*)}. \end{aligned} \quad (3)$$

InternVL2 encodes each image as 256 IMG_CONTEXT tokens; we locate $\mathcal{V}_j$ from the target-image span and anchor a fixed question-token window to the final user-message delimiter. TAS is used only as a within-model correlate; the stronger evidence for source misattribution comes from paired input-level interventions. Details are in Appendix E.4–E.5.

3.4 Models and Setup

We evaluate InternVL2-8B (Chen et al., 2024c,b), MiniCPM-V-2.6 (Yao et al., 2024), Idefics3-8B-Llama3 (Laurençon et al., 2024), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025) with greedy decoding and max_new_tokens=5. Main runs use a single RTX 4090 24GB GPU; prompt-format robustness and target-only instruction runs use the same model, data, and decoding settings, with TAS disabled when attention is not analyzed. InternVL2-8B is the main diagnostic model; MiniCPM and Idefics3 run the full $N = 10$ sweep; MiniCPM also runs the paired distractor-removal control. Qwen2.5-VL and the InternVL2 low-token setting are constrained auxiliary diagnostics rather than rate-comparable estimates. Runtime and cross-architecture details are in Appendices A.5 and F.1.

We report 95% query-level bootstrap confidence intervals and verify the main decomposition with a context-clustered bootstrap, since each context contributes three negative queries; the clean and distractor-present intervals remain well separated (Appendix A.6). TAS statistics are exploratory and use depth-level correlations plus median-split tests.

4 Experiments

4.1 Position-Sensitive Target-Binding False Positives

At $N = 10$ on InternVL2-8B, all-negative false positives are 2.00% at $d = 0.1$, 11.50% at $d = 0.6$, and 2.67% at $d = 1.0$. The decomposition changes the interpretation: clean negatives remain near floor, whereas distractor-present negatives are

Regime	n	Mean FP	Peak FP	Peak d	End FP
All negatives	6000	6.48	11.50	0.6	2.67
Clean	2598	1.15	3.68	1.0	3.68
DP	3402	10.55	17.98	0.6	1.83

Table 1: Summary of false-positive rates on InternVL2-8B at $N = 10$ (percent). End FP denotes $d = 1.0$; depth-wise rates and confidence intervals are in Appendix B.1 and Figure 2. A context-clustered bootstrap gives well-separated overall CIs: clean 1.15 [0.74, 1.59] vs. distractor-present 10.55 [9.28, 12.02].

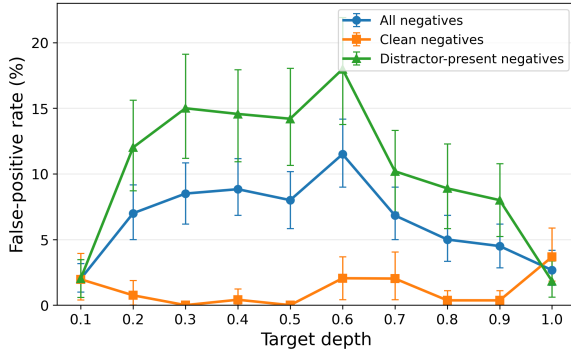


Figure 2: False-positive rate by target depth on InternVL2-8B at $N = 10$. The high-FP region in the aggregate curve is driven primarily by distractor-present negatives, while clean negatives remain near floor. Error bars show 95% bootstrap percentile confidence intervals; a context-clustered check is reported in Appendix A.6.

much higher and reach 17.98% at $d = 0.6$ (Table 1, Figure 2). The distractor-present rate is stable across depths (54–59%), so the aggregate profile is not caused by central depths containing more DP examples. Instead, the decomposition shows what the aggregate curve is counting: clean errors are rare, while errors with real non-target evidence are substantially more frequent. We interpret the pattern as position-sensitive, not a strictly middle-only or symmetric inverse-U effect; strategy-level results and a Yes-bias check are in Appendices B.2–B.3.

4.2 Threat Model for the Source-Misattribution Interpretation

Table 2 maps alternative explanations to diagnostics; Table 5 summarizes the two paired evidence controls. The controls are organized to separate three possible explanations for the clean/DP gap: a prompt or Yes-bias artifact, leakage through captions rather than images, and true dependence on non-target visual evidence. They also test

Threat	Diagnostic	Main result
Yes bias / audit	Yes-rate check; manual spot-check	Yes-rate peaks at $d = 1.0$, not the FP peak; 93.75% target-absence and 96.10% regime agreement.
Prompt artifact	Prompt-format ablations	Clean/DP gap persists without the literal [TARGET] token and also under a target-marker-only question.
Caption leakage	Channel controls	Caption-only collapses DP FP to 1.26%; image-only and shuffling retain a gap; caption masking reduces exactly-one FPs.
Visual evidence	Sole-distractor replacement	Removing the only supporting distractor sharply reduces FPs (Table 5).
Position / distance	Fixed-distance intervention; logit model	Position contributes beyond distance, but not as a symmetric middle-only effect; DP remains predictive after controls (OR=7.78).
Task format	Source localization	Source accuracy 66.62%; target-bind error 5.58%.
Benchmark artifact	MIHBench re-audit	Partial-present cases account for >98% of FPs.

Table 2: Threat model: alternative explanations, the diagnostic addressing each, and the main result.

whether the position pattern survives when target–distractor distance is controlled.

Evidence strength and sampling. False positives increase with stronger and closer non-target evidence: cases with multiple object distractors, supporting images on both sides of the target, or an adjacent supporting image are more error-prone (Appendix D.1). However, the effect is not reducible to many repeated distractors: it remains in the exactly-one-distractor subset ($n = 1,291$; peak 11.11% at $d = 0.6$), where only one non-target image contains the queried object. A logistic model controlling for depth, sampling strategy, and queried-object frequency finds that the DP regime remains strongly predictive (OR=7.78, 95% CI [4.99, 12.14], $p = 1.51 \times 10^{-19}$); within DP, more object distractors and smaller target–distractor distance increase FP odds (Appendix D.2).

Prompt-format artifact. A no-marker ablation removes the [TARGET] prefix while keeping the Image- j reference. The gap persists (1.35% clean FP vs. 14.46% DP FP), ruling out the literal marker as the main artifact. We further run a target-marker-only variant that keeps the [TARGET] label in the context but removes the Image- j reference from the question, replacing it with “the image labeled [TARGET]” and “the target image.” This variant also preserves the separation: clean FP remains near floor at 0.81%, while distractor-present FP rises to 38.80% (48.00 $\times$ the clean rate). Thus the clean/DP gap is not induced by the Image- j

Prompt	Pos. Acc.	All FP	Clean FP	DP FP	DP/Clean
Original	70.05	6.48	1.15	10.55	9.17×
Target-only instr.	67.63	5.63	0.85	9.29	10.97×

Table 3: Lightweight target-only instruction intervention on InternVL2-8B at $N = 10$. The instruction modestly reduces false positives with a small positive-accuracy trade-off, but the clean/DP separation remains. Rates are percentages.

wording, although indexed or marker-based target specification remains a controlled diagnostic setting rather than a natural referring-expression interface (Appendix A.3). We use this variant as a robustness check for the clean/DP separation rather than as a rate-comparable replacement for the main prompt, since removing the Image- j reference weakens target specification.

Lightweight diagnostic intervention. We also add an explicit target-only instruction: the model is told to use only the image labeled [TARGET] and ignore objects in all other images. Table 3 shows that this intervention modestly reduces all-negative FP from 6.48% to 5.63% and DP FP from 10.55% to 9.29%, with a small positive-accuracy drop from 70.05% to 67.63%. The clean/DP gap remains large (0.85% vs. 9.29%), indicating that simple prompting does not solve the source-misattribution problem, but that the clean/DP split is useful as an intervention-sensitive diagnostic metric. Prompt details are in Appendix A.4.

Caption and visual-channel decomposition.

We isolate textual and visual channels with four controls (Table 4; Appendices C.1–C.3, D.3). Removing all images while preserving captions collapses DP FP to 1.26% and clean FP to 0.27%; within DP, whether non-target captions mention the object barely changes the rate (1.28% vs. 1.26%). Removing captions attenuates but preserves the gap (5.20% DP vs. 1.50% clean), and shuffling non-target captions leaves it essentially unchanged (10.96% DP). On the exactly-one subset, masking object mentions in non-target captions reduces FPs from 9.06% to 4.88%, so textual mentions can amplify the error when they align with a visual distractor. Overall, however, the controls are consistent with a visual-channel-dominant source-misattribution interpretation: the effect nearly disappears only when images are removed, while removing or disrupting captions leaves a substantial distractor-present gap.

Setting	Clean FP	DP FP	DP/Clean
Image + caption (original)	1.15	10.55	9.17×
Image only (no captions)	1.50	5.20	3.47×
Caption-shuffled	1.27	10.96	8.63×
Caption only (no images)	0.27	1.26	4.69×

Table 4: Caption-channel controls on InternVL2-8B at $N = 10$ (full 6,000 negatives, percent). Removing images nearly eliminates the effect; removing or shuffling captions does not.

Paired distractor removal. On the exactly-one-distractor subset, replacing the sole query-object distractor with a COCO image lacking the object reduces FPs from 9.06% to 1.01% on InternVL2-8B and from 9.22% to 1.55% on MiniCPM-V-2.6 (Table 5; all McNemar $p < .001$). This is the most direct input-level evidence for the source-misattribution interpretation in our diagnostic: when the only supporting non-target visual evidence is removed, most original false positives become correct rejections (Appendix D.3).

Position beyond distance. For $N = 10$, p_i denotes the condition in which the queried target occupies the i th image position (1-indexed); “ p_a vs. p_b ” denotes a matched contrast between two target placements. We move the target between matched middle and outer positions while holding the target image, query object, sole distractor, fillers, and target–distractor distance fixed. The aggregate middle-vs.-outer contrast is higher for middle placements (32.46% vs. 10.73%), but the late-tail contrast p_5 vs. p_9 reverses this pattern (16.19% vs. 23.08%). Thus position contributes beyond distance, but the effect is sequence-structure-dependent rather than uniformly middle-only (Table 5; Appendix D.6). We use “position-sensitive” operationally: errors vary with the relative placement of the queried target and supporting distractor, including adjacency and whether supporting evidence occurs before, after, or around the target, rather than following a universal middle-position curve.

Task-format specificity. In a forced-choice source-localization probe, the model is told that exactly one image contains the queried object and must answer with an image number. InternVL2-8B identifies the correct source in 66.62% of exactly-one cases but still assigns the object to the queried target in 5.58%. The depth profile differs from the binary Yes/No setting, so we do not use this as evidence for the same position curve; in-

Paired distractor replacement (exactly-one subset)		
Control	Orig. FP	Ctrl. FP
InternVL2 image-replaced	9.06	1.01
MiniCPM image-replaced	9.22	1.55
InternVL2 image-only replaced	6.74	2.48
Fixed-distance position intervention (matched contrasts)		
Contrast	Mid. FP	Outer FP
p_5 vs. p_1	16.19	2.56
p_6 vs. p_2	48.72	5.96
p_6 vs. p_{10}	48.72	11.31
p_5 vs. p_9	16.19	23.08
Aggregate	32.46	10.73

Table 5: Paired evidence controls. Replacing the sole query-object distractor sharply reduces FPs (top). Fixed-distance contrasts show position sensitivity, but not a uniformly middle-only pattern (bottom). Details in Appendices D.3 and D.6.

Model	Clean FP	Partial FP	All FP	FP from partial
InternVL2-8B	0.42	35.03	14.23	98.21
MiniCPM-V-2.6	0.14	16.77	6.77	98.75
Qwen2.5-VL (aux.)	0.00	15.92	6.35	100.00

Table 6: MIHBench Existence re-audit. Clean: object absent from all images; partial-present: object in at least one but not all. Rates are percentages; Qwen2.5-VL is a constrained auxiliary diagnostic.

stead, it shows that target-binding-like errors are not limited to Yes/No answering (Appendix D.4).

External benchmark re-audit. We re-audit MIHBench Existence negatives under a related clean/partial-present split. This is not a direct target-image binding test because MIHBench asks all-image existence questions. Among 1,200 negatives, 471 (39.25%) are partial-present; after excluding 19 conflicts, we evaluate 1,181 examples. Across InternVL2-8B, MiniCPM-V-2.6, and constrained Qwen2.5-VL, clean FPs are near zero (0.00–0.42%) while partial-present FPs are much higher (15.92–35.03%); over 98% of all FPs come from partial-present cases (Table 6). Absolute rates vary across models, with InternVL2 showing the highest partial-present rate, but the concentration of errors in the partial-present bucket is stable. Thus aggregate multi-image FP rates can mix globally unsupported negatives with cases where object evidence is present under a different scope. We use this result as external evidence for the usefulness of regime-aware reporting, not as direct evidence for the target-image binding mechanism in our haystack.

4.3 Cross-Architecture and Context-Length Evidence

Table 7 summarizes context-length and cross-architecture diagnostics. Moving from $N = 5$ to $N = 10$ increases all-negative FP from 3.92% to 6.48% and DP FP from 5.87% to 10.55%, while clean negatives remain low; we interpret this as a crowded-context diagnostic, not a length-only causal estimate. The limited $N = 20$ low-token run preserves regime separation but should not be read as monotonic scaling, because lowering the visual-token budget changes both perception quality and context length. We therefore use it only as a within-model stress check.

Full $N = 10$ sweeps on MiniCPM-V-2.6 and Idefics3-8B-Llama3 preserve the clean/DP gap (0.46%/8.20% and 1.27%/10.23%), though depth profiles are model-specific. A constrained Qwen2.5-VL run shows the same within-model separation (0.38% vs. 3.20%), with a smaller but visible gap after removing person queries (0.39% vs. 1.42%). We therefore use these results as within-model regime comparisons rather than absolute model rankings; exact depth profiles and absolute rates remain model-specific. To test whether the separation is specific to COCO, we additionally use an object-balanced PASCAL VOC2012 val check (Everingham et al., 2010) with the same $N = 10$ multi-image format and neutral captions. The clean/DP FP pairs are 0.71%/3.00% for InternVL2-8B, 0.59%/13.65% for MiniCPM-V-2.6, 0.82%/12.00% for Idefics3-8B, and 0.82%/2.59% for Qwen2.5-VL-7B. Appendix F.3 reports sample sizes, gap ratios, significance tests, and model-specific input adaptations. These external results likewise support only within-model regime comparisons and dataset-level transfer of the diagnostic, not cross-model rankings or transfer beyond binary object existence.

4.4 TAS as an Observational Target-Binding Correlate

We report TAS as a secondary, InternVL2-specific correlate rather than a mechanism test. Depth-level TAS is negatively correlated with FP rate, especially in DP cases, but this uses only ten bins and is exploratory. We therefore emphasize the sample-level median split and controlled regressions rather than the depth correlation alone. At the sample level, low-TAS DP examples produce

Setting	All	Clean	DP	DP prop.
InternVL2, $N = 5$	3.92	2.49	5.87	42.33
InternVL2, $N = 10$	6.48	1.15	10.55	56.70
MiniCPM-V-2.6, $N = 10$	4.85	0.46	8.20	56.70
Idefics3-8B, $N = 10$	6.35	1.27	10.23	56.70
Qwen2.5-VL, $N = 10$	1.98	0.38	3.20	56.70
InternVL2, $N = 10$ LT	3.03	1.46	4.23	56.70
InternVL2, $N = 20$ LT	3.33	1.39	4.12	71.27

Table 7: Context-length, low-token, and cross-architecture summary (percent). The All, Clean, and DP columns are false-positive rates; DP prop. denotes the proportion of negative queries that are distractor-present, not an error rate. MiniCPM and Idefics3 are full $N = 10$ sweeps; LT denotes low-token InternVL2; Qwen2.5-VL is a constrained auxiliary diagnostic.

Group	n	Low FP	High FP	Ratio
All negatives	6000	9.52	3.44	$2.77\times$
Clean	2598	0.92	1.39	$0.66\times$
Distractor-present	3402	15.57	5.53	$2.82\times$

Table 8: TAS median-split analysis on negative samples. FP rates are percentages.

more FPs than high-TAS examples (15.57% vs. 5.53%, $p = 1.56 \times 10^{-21}$), while clean negatives show no significant split (0.92% vs. 1.39%, $p = 0.267$). Controlled regressions preserve the regime-specific association: in DP-only data, a one-standard-deviation TAS increase reduces FP odds (OR=0.424, 95% CI [0.297, 0.604], $p = 1.97 \times 10^{-6}$), and an all-negative interaction model finds a significant TAS $\times$ DP term (OR=0.279, $p = 1.81 \times 10^{-10}$). This pattern is consistent with the source-misattribution interpretation because the TAS split matters in the distractor-present regime, where non-target evidence can be misassigned, but not in the clean regime, where no contextual visual evidence supports the queried object. We nevertheless treat TAS only as an auxiliary correlate, not causal evidence; details are in Appendices E.2–E.4.

5 Conclusion

We presented a position-controlled diagnostic for target-binding hallucination in interleaved multi-image LVLM contexts. Aggregate false positives conflate globally unsupported clean errors with distractor-supported source-misattribution errors: on InternVL2-8B, clean FP remains near floor while DP FP is much higher, the same regime-level direction appears across the other tested architectures, and an auxiliary PASCAL VOC2012 check preserves it beyond the COCO construc-

tion. Channel and paired-removal controls are consistent with a visual-channel-dominant source-misattribution interpretation, while TAS remains only an exploratory InternVL2-specific correlate.

The broader implication is methodological. When non-target images may contain evidence for a negative query, aggregate hallucination rates alone do not reveal whether a model is inventing unsupported objects or mis-scoping real evidence from another image. The related MIHBench re-audit shows that this distinction can also affect the interpretation of existing multi-image hallucination benchmarks. For object-existence diagnostics and other source-sensitive multi-image evaluations where negative evidence can be audited, we therefore recommend accompanying aggregate false-positive rates with clean versus evidence-supported negative breakdowns whenever non-target images may contain supporting evidence.

Implications for practitioners. In document-review, retrieval-augmented, and agentic multi-image pipelines, an aggregate false-positive rate should not automatically be read as pure object invention when non-target sources may contain the queried evidence. When evidence scope is auditable, evaluations should record the queried source and report aggregate FP, clean-negative FP, distractor-present FP, and the prevalence of each regime. Low clean FP with high distractor-present FP points primarily to source-routing or evidence-binding failures and motivates testing per-source verification or explicit provenance checks; high clean FP indicates a larger globally unsupported-generation component. Paired edits that remove or replace the sole supporting distractor can further distinguish these cases from generic answer bias. This decomposition is diagnostic rather than a deployment-safety certificate, and any intervention should be evaluated separately in both regimes.

6 Limitations

Our diagnostic deliberately focuses on binary object-existence queries with controlled target references. The primary COCO benchmark and auxiliary PASCAL VOC2012 check support dataset-level transfer of the clean/DP separation, but not transfer to attributes, relations, counting, identity consistency, natural referring expressions, or free-form generation. The MIHBench re-audit is likewise a related regime-mixing check rather than a

direct target-image binding test.

The clean/DP split relies on dataset annotations, so missed or ambiguous objects may introduce label noise, although our manual audit suggests this is unlikely to explain the main gap. Absolute rates are not directly comparable across architectures because visual preprocessing and token budgets differ. TAS is an InternVL2-specific observational correlate rather than causal evidence, and position profiles may be model-specific. Longer full-token contexts, larger or proprietary models, and more natural tasks remain future work.

7 Ethical Considerations

This work is a diagnostic evaluation study of multimodal hallucination in LVLMs using public datasets and publicly available model checkpoints. It does not collect private user data, involve human subjects, train or release a new model, or deploy a user-facing system. The supplementary material submitted during review did not redistribute COCO or PASCAL VOC images or model weights; it provided metadata, prompts, audit labels, model outputs, and analysis scripts for reproducibility.

Target-binding hallucinations may be harmful in downstream multi-image applications where attributing visual evidence to the correct image matters, such as document review, visual retrieval, and medical or legal image triage. Our results should not be interpreted as certifying any model as safe or unsafe for deployment. Rather, they show that aggregate hallucination rates can obscure source-misattribution failures and should be complemented with clean versus evidence-supported negative analyses; reducing aggregate hallucination alone may not ensure correct source attribution. Because our experiments remain limited to object-category existence and controlled target references, despite the additional VOC2012 dataset check, the results should not be generalized to richer multimodal reasoning or deployment settings without further validation.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian

Borgeaud, and 8 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024a. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

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

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 49250–49267.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010.

- The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. [MME: A comprehensive evaluation benchmark for multimodal large language models](#). In *Advances in Neural Information Processing Systems*, volume 38. Datasets and Benchmarks Track.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoub, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Krizan, Shantanu Acharya, Dima Rekish, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What’s the real context size of your long-context language models? In *Proceedings of the First Conference on Language Modeling (COLM)*.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2025a. LLaVA-Interleave: Tackling multi-image, video, and 3d in large multimodal models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Jiale Li, Mingrui Wu, Zixiang Jin, Hao Chen, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. 2025b. [MIHBench: Benchmarking and mitigating multi-image hallucinations in multimodal large language models](#). In *Proceedings of the 33rd ACM International Conference on Multimedia (MM ’25)*, pages 3143–3152.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 19730–19742.
- Xudong Li, Mengdan Zhang, Peixian Chen, Xiawu Zheng, Yan Zhang, Jingyuan Zheng, Yunhang Shen, Ke Li, Chaoyou Fu, Xing Sun, and Rongrong Ji. 2025c. [Zooming from context to cue: Hierarchical preference optimization for multi-image MLLMs](#). In *Advances in Neural Information Processing Systems*, volume 38.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoub, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and Weiming Hu. 2024c. MIBench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024d. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

- Shi Liu, Weijie Su, Xizhou Zhu, Wenhai Wang, and Jifeng Dai. 2025a. CoMemo: LVLMS need image context with image memory. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024e. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 216–233.
- Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2025b. MIA-DPO: Multi-image augmented direct preference optimization for large vision-language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2025. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.
- Yannic Neuhaus and Matthias Hein. 2025. RePOPE: Impact of annotation errors on the POPE benchmark. *arXiv preprint arXiv:2504.15707*.
- Yeji Park, Minyoung Lee, Sanghyuk Chun, and Jun-suk Choe. 2025. Mitigating cross-image information leakage in multi-image understanding with large vision-language models. *arXiv preprint arXiv:2508.13744*.
- Nhi Pham and Michael Schott. 2024. H-POPE: Hierarchical polling-based probing evaluation of hallucinations in large vision-language models. *arXiv preprint arXiv:2411.04077*.
- Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Han Qiu, Jiaying Huang, Peng Gao, Qin Qi, Xiaoqin Zhang, Ling Shao, and Shijian Lu. 2024. Long-HalQA: Long-context hallucination evaluation for multimodal large language models. *arXiv preprint arXiv:2410.09962*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4035–4045.
- Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. MileBench: Benchmarking MLLMs in long context. In *Proceedings of the First Conference on Language Modeling (COLM)*.
- Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. 2025. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10599–10609.
- Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2025. MuirBench: A comprehensive benchmark for robust multi-image understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2023. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. 2024b. Needle in a multimodal haystack. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37.
- Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. 2025. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Haitao Zheng, Maosong Sun, and Tat-Seng Chua. 2024. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13807–13816.

A Implementation and Dataset Details

A.1 Depth-to-position mapping

The main text defines relative depth using the 1-indexed target position $j \in \{1, \dots, N\}$, with $d = j/N$, and uses the 0-indexed implementation index $k = j - 1$ when constructing prompts and image sequences. For $N = 10$, the ten nominal depths $d \in \{0.1, 0.2, \dots, 1.0\}$ map one-to-one to the ten image positions. For $N = 5$, multiple nominal depths map to the same discrete image position; we therefore use the $N = 5$ slice primarily for context-length comparison.

Nominal depth	$N = 5$ target position j	0-indexed k
0.1/0.2	1	0
0.3/0.4	2	1
0.5/0.6	3	2
0.7/0.8	4	3
0.9/1.0	5	4

Table 9: Depth-to-position mapping for the $N = 5$ setting. Because several nominal depths map to the same discrete position, we interpret the $N = 5$ results mainly as a context-length baseline.

A.2 Prompt template

For InternVL2-8B, each prompt has the following interleaved form:

```
Image 1: <image>
[caption_1]
Image 2: <image>
[caption_2]
...
Image N: <image>
[caption_N]
Look at Image  $j$  (labeled [TARGET]).
Is there a/an {object} in Image  $j$ ? Answer only
Yes or No.
```

The target image line is prefixed with [TARGET] to make the queried image unambiguous. For Qwen-family models, the same semantic prompt is wrapped with the model’s apply_chat_template interface.

A.3 Prompt-format robustness ablations

To test whether the main clean/distractor-present separation is an artifact of target-reference wording, we run two full InternVL2-8B $N = 10$ prompt ablations over the same 6,000 negative queries. The no-[TARGET] ablation removes the [TARGET] prefix and replaces “Look at Image j (labeled [TARGET])” with “Look at Image j .” The target-marker-only ablation keeps the [TARGET]

label in the context but removes the Image- j reference from the question, replacing it with “Look at the image labeled [TARGET]” and “the target image.” Image order, captions, target positions, object queries, and audit labels are unchanged. The target-marker-only run uses greedy decoding with TAS disabled because attention is not analyzed in this robustness check.

Table 10 shows that neither prompt change removes the regime separation. Removing the explicit marker preserves a 1.35% clean FP versus 14.46% DP FP gap. Removing Image- j wording from the question also preserves the gap: clean FP remains 0.81%, while DP FP rises to 38.80%. Thus the clean/DP separation is not induced by the literal [TARGET] token or by Image- j wording, although both settings remain controlled target-specification prompts rather than fully natural referring expressions.

Prompt	All FP	Clean FP	DP FP	DP/Clean
Original Image- j	6.48	1.15	10.55	9.17×
No-[TARGET] marker	8.78	1.35	14.46	10.73×
Target-marker-only	22.35	0.81	38.80	48.00×

Table 10: Prompt-format robustness ablations on InternVL2-8B at $N = 10$ over 6,000 negative queries. The no-[TARGET] ablation removes the explicit target marker while preserving Image- j wording. The target-marker-only variant removes the Image- j reference from the question and refers only to the image labeled [TARGET]. Rates are false-positive percentages.

A.4 Target-only instruction intervention

To test whether the clean/distractor-present split can serve as an intervention-sensitive diagnostic metric, we add a lightweight target-only instruction to the original Image- j prompt:

```
Look at Image  $j$  (labeled [TARGET]).
For this question, use only the image labeled
[TARGET] and ignore objects in all other images.
Is there a/an {object} in Image  $j$ ? Answer only
Yes or No.
```

We run InternVL2-8B at $N = 10$ over the full 12,000-query set (6,000 positive and 6,000 negative queries), using the same images, captions, target positions, object queries, and greedy decoding as the main run. TAS is disabled because attention is not analyzed in this intervention.

The main text reports the results in Table 3. All-negative FP decreases from 6.48% to 5.63%, and DP FP decreases from 10.55% to 9.29%, while positive accuracy decreases from 70.05% to 67.63%. The clean/DP separation remains

large, so this simple instruction does not solve target-binding errors; rather, it illustrates how the clean/DP split can evaluate whether a proposed intervention specifically reduces distractor-supported false positives.

A.5 Runtime and reproducibility details

All experiments use greedy decoding (`do_sample=False`, `max_new_tokens=5`). Main runs use a single RTX 4090 24GB GPU; prompt-format robustness and target-only instruction runs use the same model, data, and decoding settings, with TAS disabled when attention is not analyzed. InternVL2 uses FP16 and deterministic decoding; Yes/No outputs are parsed from the first generated token, with no unknown outputs in the main $N = 10$ run. Pool partition and bootstrap resampling use seed 42. Main InternVL2 runs with attention extraction take roughly 2–3 hours and peak at about 21GB VRAM; most other diagnostics finish within two hours. Qwen-family runs are constrained dynamic-resolution diagnostics because strict visual-token matching is infeasible under this memory budget.

Reproducibility materials. The supplementary material submitted during review contained haystack metadata, POPE-style questions, audit labels, model outputs, TAS extraction code, and scripts for the main controls, cross-model sweeps, MIHBench re-audits, and statistical analyses. It reproduced the reported diagnostics from stored outputs without requiring readers to rerun all models.

A.6 Context-clustered bootstrap

Each haystack context contributes three negative queries, so query-level resampling may underestimate uncertainty. We repeat the main bootstrap by resampling haystack contexts within each target depth. Table 11 shows that the clean CI remains far below the distractor-present CI.

Regime	Overall FP [95% CI]	Peak FP [95% CI]
All	6.48 [5.75, 7.33]	11.50 [8.50, 14.83] at $d = 0.6$
Clean	1.15 [0.74, 1.59]	3.68 [1.53, 6.48] at $d = 1.0$
Distr.-present	10.55 [9.28, 12.02]	17.98 [12.98, 23.01] at $d = 0.6$

Table 11: Context-clustered bootstrap check for the main InternVL2-8B $N = 10$ negative-query results (percent).

A.7 Manual spot-check of audit labels

We manually spot-check 240 sampled negative instances: all 30 clean false positives, plus 70 examples from each of clean correct rejections, distractor-present false positives, and distractor-present correct rejections. Annotators verify target absence and non-target object presence. Table 12 shows that target absence is confirmed in 225/240 cases (93.75%), or 225/229 determinate cases (98.25%), and regime labels agree in 222/231 determinate cases (96.10%). Residual ambiguity is concentrated in the already rare clean-FP slice and is therefore unlikely to explain the much larger distractor-present false-positive rate.

Conservative clean-FP sensitivity. Because all 30 clean false positives are audited, we can bound the main clean-FP estimate directly. Treating all uncertain and target-present clean false positives as non-errors gives a confirmed clean-FP rate of 23/2,598=0.89%; keeping uncertain cases but removing clear conflicts gives 27/2,598=1.04%. Both remain far below the distractor-present rate of 10.55% (359/3,402), so the clean/DP gap is not driven by ambiguous clean false positives.

Group	n	Target absent	Regime agree	Uncertain
Clean CR	70	98.6	100.0	5
Clean FP	30	76.7	89.7	4
DP CR	70	95.7	95.5	7
DP FP	70	94.3	95.7	4
Overall	240	93.8	96.1	20

Table 12: Manual audit of COCO-based negative-query labels. Target absent is the percentage judged absent from the target image; Regime agree excludes uncertain regime labels. Rates are percentages.

Treatment	Count	Rate
Original clean-FP count	30/2,598	1.15
Confirmed target-absent only	23/2,598	0.89
Confirmed + uncertain labels	27/2,598	1.04
Original DP-FP count	359/3,402	10.55

Table 13: Conservative sensitivity check for the clean false-positive estimate. Treating uncertain and target-present clean false positives as non-errors lowers the already small clean-FP rate, while the distractor-present rate remains much higher. Rates are percentages.

B Main Result Diagnostics

This section provides depth-wise and strategy-wise diagnostics for the main InternVL2-8B $N = 10$ setting, together with a positive-query Yes-bias check.

B.1 InternVL2 $N = 10$ depth-wise decomposition

Table 14 gives the full clean/distractor-present decomposition for the main InternVL2-8B $N = 10$ setting.

<i>Depth-wise decomposition</i>				
Depth	All FP	Clean FP	Distr. FP	Distr. rate
0.1	2.00	1.97	2.02	57.67
0.2	7.00	0.75	12.01	55.50
0.3	8.50	0.00	15.00	56.67
0.4	8.83	0.41	14.57	59.50
0.5	8.00	0.00	14.20	56.33
0.6	11.50	2.05	17.98	59.33
0.7	6.83	2.02	10.20	58.83
0.8	5.00	0.36	8.90	54.33
0.9	4.50	0.36	8.00	54.17
1.0	2.67	3.68	1.83	54.67
Mean	6.48	1.15	10.55	56.70
<i>Strategy-level decomposition</i>				
Strategy	All FP	Clean FP	Distr. FP	Distr. rate
Random	1.80	0.46	6.21	23.35
Popular	8.70	0.71	10.84	78.90
Adversarial	8.95	3.11	11.72	67.85

Table 14: InternVL2-8B $N = 10$ false-positive decomposition (percent), including depth-wise and POPE-strategy splits.

B.2 Strategy-level clean/distractor-present decomposition

The strategy-level panel of Table 14 shows that Popular and Adversarial queries are harder partly because they more often create distractor-present conflicts.

B.3 Positive-query accuracy and Yes-bias check

Table 15 reports positive-query accuracy, overall Yes rate, and negative false-positive rate by target depth on InternVL2-8B at $N = 10$. Each depth contains 600 positive and 600 negative queries. The highest overall Yes rate occurs at $d = 1.0$, where negative false positives are low, whereas the negative false-positive peak occurs at $d = 0.6$. The negative-query pattern is therefore not explained by a position-dependent tendency to answer Yes.

Depth	Yes rate	Pos. Acc.	Neg. FP
0.1	36.08	70.17	2.00
0.2	38.50	70.00	7.00
0.3	37.17	65.83	8.50
0.4	37.00	65.17	8.83
0.5	35.75	63.50	8.00
0.6	38.58	65.67	11.50
0.7	36.50	66.17	6.83
0.8	38.50	72.00	5.00
0.9	41.25	78.00	4.50
1.0	43.33	84.00	2.67
Mean	38.27	70.05	6.48

Table 15: Positive-query accuracy, overall Yes rate, and negative false-positive rate by target depth on InternVL2-8B at $N = 10$ (percent). The negative FP peak at $d = 0.6$ does not coincide with the highest overall Yes rate at $d = 1.0$.

C Caption-Channel Controls

This section provides full results for the four caption-channel controls summarized in Table 4 of the main paper: caption-only, image-only, caption-shuffled, and caption-masked. Together they separate the contributions of textual and visual channels to target-binding errors.

C.1 Caption-only ablation

We remove all images while preserving image-position structure, target marker, captions, and queries, and run the full 6,000 $N = 10$ negative-query set on InternVL2-8B. Table 16 shows that distractor-present false positives collapse to 1.26% and clean false positives to 0.27%, eliminating the multimodal position-sensitive FP pattern. Although the DP/Clean ratio is $4.69\times$, both regimes are near floor and the absolute gap is only about one percentage point.

Depth	All FP	Clean FP	Distr. FP	Distr. rate
0.1	0.67	0.00	1.16	57.67
0.2	1.00	0.00	1.80	55.50
0.3	1.00	0.38	1.47	56.67
0.4	1.00	0.41	1.40	59.50
0.5	1.17	0.38	1.78	56.33
0.6	1.17	0.41	1.69	59.33
0.7	0.50	0.81	0.28	58.83
0.8	0.67	0.00	1.23	54.33
0.9	0.50	0.00	0.92	54.17
1.0	0.67	0.37	0.91	54.67
Mean	0.83	0.27	1.26	56.70

Table 16: Caption-only ablation on InternVL2-8B at $N = 10$ over all 6,000 negative queries. Removing all images collapses both regimes to near floor and eliminates the multimodal position-sensitive FP pattern. The DP rate column is determined by haystack construction and unchanged.

Non-target caption mention split. Within the 3,402 distractor-present negatives, we further split

on whether any non-target caption explicitly mentions the queried object name. Table 17 shows that false-positive rates are nearly identical across the two groups (1.28% vs. 1.26%). The residual textual effect is therefore not explained by direct lexical caption matching.

Non-target caption mentions object?	<i>n</i>	FP	FP rate
No	2,307	29	1.26
Yes	1,095	14	1.28

Table 17: Caption-only ablation, distractor-present negatives only, split by whether any non-target caption explicitly mentions the queried object. Rates are nearly identical, indicating that the residual caption-only errors are not explained by direct lexical caption matching.

Combined with the image-only, caption-shuffled, and caption-masked controls below, this supports a visual-channel-dominant interpretation: removing images nearly eliminates the gap, while removing or disrupting captions mainly attenuates it.

C.2 Image-only caption ablation

The image-only control keeps the same images, target positions, target marker, and negative queries, but removes all captions. Table 18 shows that removing captions attenuates but does not eliminate the effect: distractor-present FP drops from 10.55% to 5.20%, while remaining above clean FP (1.50%).

Depth	All FP	Clean FP	Distr. FP	Distr. rate
0.1	1.67	1.97	1.45	57.67
0.2	3.17	2.25	3.90	55.50
0.3	3.83	0.38	6.47	56.67
0.4	6.17	2.06	8.96	59.50
0.5	4.17	0.00	7.40	56.33
0.6	6.00	2.05	8.71	59.33
0.7	3.33	1.62	4.53	58.83
0.8	2.50	1.09	3.68	54.33
0.9	2.67	0.36	4.62	54.17
1.0	2.50	3.31	1.83	54.67
Mean	3.60	1.50	5.20	56.70

Table 18: Image-only caption ablation on InternVL2-8B at $N = 10$ over all 6,000 negative queries. Removing captions reduces false positives, but distractor-present negatives remain more error-prone than clean negatives.

C.3 Non-target caption-shuffled control

This control keeps images, target caption, target marker, and queries fixed, but permutes captions among non-target images. Table 19 shows that disrupting non-target image-caption alignment leaves the regime gap essentially unchanged

(1.27% clean vs. 10.96% DP).

Depth	All FP	Clean FP	Distr. FP	Distr. rate
0.1	3.67	1.97	4.91	57.67
0.2	6.17	1.12	10.21	55.50
0.3	9.33	0.77	15.88	56.67
0.4	8.67	0.82	14.01	59.50
0.5	7.33	0.38	12.72	56.33
0.6	11.00	1.64	17.42	59.33
0.7	6.50	1.21	10.20	58.83
0.8	5.50	1.09	9.20	54.33
0.9	6.00	0.36	10.77	54.17
1.0	3.50	3.31	3.66	54.67
Mean	6.77	1.27	10.96	56.70

Table 19: Caption-shuffled InternVL2-8B $N = 10$ false-positive decomposition (percent). Non-target captions are shuffled across non-target images while all images, target caption, target marker, and queries remain fixed.

D Distractor Evidence Controls

This section provides full results behind the threat-model controls summarized in Table 2 of the main paper: descriptive evidence-strength trends, a multivariate logistic robustness check, paired input-level controls on the exactly-one-distractor subset, a forced-choice source-localization probe, the MI-HBench Existence re-audit, and the paired fixed-distance target-position intervention.

D.1 Distractor evidence strength and placement

Using audit metadata, we count how many non-target images contain the queried object, where they occur relative to the target, and the nearest target-distractor distance. Table 20 shows that false positives are higher with stronger, two-sided, or adjacent non-target evidence. These are descriptive trends because distractor placement is audited after construction.

Evidence facet	Group	<i>n</i>	FP rate
# object distractors	1	1291	6.74
	2	446	8.30
	≥ 3	1665	14.11
Relative location	Before only	1051	5.61
	After only	1091	6.78
	Both sides	1260	17.94
Nearest distance	1	1711	16.13
	2-3	964	5.91
	≥ 4	727	3.58

Table 20: False-positive rate by distractor-evidence strength and placement on distractor-present negatives for InternVL2-8B at $N = 10$ (percent). Stronger and closer non-target evidence is associated with higher target-binding false-positive rates.

D.2 Multivariate logistic robustness analysis

To check whether the clean/distractor-present gap is explained by simple covariates, we fit logistic regressions on InternVL2-8B $N = 10$ negative queries with standard errors clustered by haystack context. The all-negative model predicts false positives from the distractor-present indicator while controlling for target depth, POPE sampling strategy, and queried-object frequency. The distractor-present-only model further examines evidence-structure factors: the number of object distractors, nearest target–distractor distance, and whether supporting distractors occur before, after, or on both sides of the target.

Table 21 shows that the distractor-present indicator remains a strong predictor of false positives after controls (OR=7.78, 95% CI [4.99, 12.14], $p = 1.51 \times 10^{-19}$). Within distractor-present negatives, the number of object distractors increases false-positive odds (OR=1.91, 95% CI [1.18, 3.07], $p = 0.0083$), while greater nearest distance decreases false-positive odds (OR=0.75, 95% CI [0.64, 0.89], $p = 0.0011$). Relative side indicators are not significant after these controls.

Model	Predictor	OR [95% CI]	p
All negatives	Distractor-present	7.78 [4.99, 12.14]	1.51×10^{-19}
All negatives	Log object frequency	0.91 [0.79, 1.05]	.211
All negatives	Popular strategy	0.87 [0.78, 0.97]	.014
All negatives	Random strategy	0.29 [0.16, 0.51]	2.49×10^{-5}
DP only	Log # distractors	1.91 [1.18, 3.07]	.008
DP only	Nearest distance	0.75 [0.64, 0.89]	.001
DP only	Distractor before only	1.00 [0.60, 1.66]	.988
DP only	Distractors on both sides	1.15 [0.69, 1.92]	.581
DP only	Popular strategy	0.89 [0.79, 0.99]	.036
DP only	Random strategy	0.31 [0.15, 0.63]	.001

Table 21: Multivariate logistic robustness analysis on InternVL2-8B $N = 10$ negatives. Standard errors are clustered by haystack context. Odds ratios above one indicate higher false-positive odds.

D.3 Exactly-one-distractor subset and removal counterfactual

The exactly-one-distractor subset contains 1,291 negatives and remains non-flat, with 6.74% mean FP and an 11.11% peak at $d = 0.6$. On this subset, caption masking replaces explicit query-object mentions in non-target captions with a neutral token, while image replacement swaps the sole query-object distractor for a COCO image without the object. Table 22 shows that caption masking reduces FPs but image replacement nearly eliminates them, confirming that captions amplify the effect while non-target visual evidence is the pri-

mary factor. The “Orig. FP” baseline is the paired-rerun baseline for the control scripts.

Control	Orig. FP	Ctrl. FP	Yes→No	No→Yes
InternVL2 caption-masked	9.06	4.88	47.01	0.09
InternVL2 image-replaced	9.06	1.01	90.60	0.17
MiniCPM image-replaced	9.22	1.55	94.96	1.19
InternVL2 img-only replaced	6.74	2.48	70.11	0.50

Table 22: Exactly-one-distractor caption-masked and removal controls ($n = 1291$ per row). Yes→No is the percentage of original false positives that become correct rejections; No→Yes is the reverse flip rate among original correct rejections. Rates are percentages.

D.4 Forced-choice source-localization probe

To test whether target-binding is tied to the binary Yes/No format, we add a source-localization probe on the exactly-one-distractor subset. Each query is converted from target-existence verification into forced-choice image-index localization: the model is told that exactly one image contains the queried object and must answer with an image number from 1 to 10. The correct answer is the sole non-target source image containing the object. We count a *target-binding error* when the model answers with the target-image number, even though the queried object is absent from the target.

Table 23 shows that InternVL2-8B correctly localizes the source in 66.62% of exactly-one cases. It still assigns the object to the queried target in 5.58% of cases, with no invalid or “none” responses. The largest target-binding localization errors occur at outer depths rather than reproducing the binary-existence position profile, so we use this result as a task-format check rather than as evidence for the same position dynamics.

Depth	n	Source acc.	Target-bind	Other wrong	None
0.1	133	66.17	12.78	21.05	0.00
0.2	135	69.63	2.96	27.41	0.00
0.3	113	64.60	2.65	32.74	0.00
0.4	133	59.40	2.26	38.35	0.00
0.5	129	72.87	2.33	24.81	0.00
0.6	153	75.82	2.61	21.57	0.00
0.7	138	64.49	5.07	30.43	0.00
0.8	122	65.57	4.10	30.33	0.00
0.9	112	56.25	8.93	34.82	0.00
1.0	123	68.29	13.01	18.70	0.00
Mean	1291	66.62	5.58	27.81	0.00

Table 23: Forced-choice source-localization probe on exactly-one-distractor cases. Target-bind denotes cases where the model incorrectly selects the queried target image. Rates are percentages.

D.5 External MIHBench Existence re-audit

MIHBench Existence asks all-image questions rather than target-image questions, so we report

a related regime-mixing audit rather than a direct target-binding test. We use a clean/partial-present split: clean negatives have the queried COCO category absent from all images, partial-present negatives contain it in at least one but not all images, and all-present annotation conflicts are excluded. Table 24 shows that 39.25% of 1,200 negatives are partial-present, including 92.50% of adversarial negatives.

Strategy	n	Clean	Partial-present	Conflict
Adversarial	400	3.50	92.50	4.00
Popular	800	87.00	12.62	0.38
All	1200	59.17	39.25	1.58

Table 24: Annotation-only re-audit of MIHBench Existence negative examples. Clean: object absent from all images; partial-present: object in at least one but not all images; conflict: COCO annotations indicate presence in all images despite a negative label. MIHBench Existence uses only Adversarial and Popular sampling strategies. Rates are percentages.

We then run InternVL2-8B, MiniCPM-V-2.6, and a constrained Qwen2.5-VL diagnostic on the 1,181 non-conflicting negatives. Table 25 shows near-zero clean FPs (0.00–0.42%) but much higher partial-present FPs (15.92–35.03%); partial-present cases account for over 98% of all FPs in all three runs. The cross-model invariant is this regime-level concentration, not the absolute rate.

Model	Strategy/Bucket	n	FP	FP rate	Unknown
InternVL2	Adv./Clean	14	0	0.00	0
	Adv./Partial	370	157	42.43	0
	Pop./Clean	696	3	0.43	0
	Pop./Partial	101	8	7.92	0
MiniCPM	Adv./Clean	14	1	7.14	0
	Adv./Partial	370	76	20.54	0
	Pop./Clean	696	0	0.00	0
	Pop./Partial	101	3	2.97	0
Qwen2.5-VL	Adv./Clean	14	0	0.00	0
	Adv./Partial	370	72	19.46	0
	Pop./Clean	696	0	0.00	0
	Pop./Partial	101	3	2.97	0

Table 25: Strategy-level MIHBench Existence re-audit. Conflicting annotation cases are excluded. “Partial” denotes partial-present negatives. The MiniCPM Adv./Clean row is a single false positive in $n = 14$ and should not be over-interpreted. Rates are percentages.

D.6 Paired fixed-distance target-position intervention

For $N = 10$, let p_i denote the condition in which the queried target occupies the i th image position (1-indexed). For each exactly-one-distractor example, we reconstruct matched sequences that keep

the target image, query object, sole distractor, filler images, and target–distractor distance fixed while moving the target to middle or outer positions. Table 26 shows that middle placements are substantially more error-prone than matched outer placements in aggregate (32.46% vs. 10.73%), with the strongest contrast at position 6. However, the p_5 vs. p_9 contrast goes in the opposite direction (16.19% vs. 23.08%). We therefore interpret this intervention as evidence that target position contributes beyond target–distractor distance, not as evidence for a uniformly middle-only or symmetric inverse-U pathology.

Contrast	Mid. FP	Outer FP	M-only	O-only	p
p_5 vs. p_1	16.19	2.56	183	7	< .001
p_5 vs. p_9	16.19	23.08	56	145	< .001
p_6 vs. p_2	48.72	5.96	558	6	< .001
p_6 vs. p_{10}	48.72	11.31	501	18	< .001
Aggregate	32.46	10.73	1298	176	< .001

Table 26: Paired fixed-distance target-position intervention on the exactly-one-distractor subset. M-only counts cases where the middle variant is a false positive but the matched outer variant is a correct rejection; O-only is the reverse. The aggregate contrast favors middle placements, but the p_5 vs. p_9 row shows that the effect is sequence-structure-dependent rather than uniformly middle-only. Rates are percentages; very small paired-test values are capped at $p < .001$.

E TAS Diagnostics and Ablations

TAS analyses in this appendix are exploratory and InternVL2-specific. They are intended to show whether target-image attention co-varies with distractor-present false positives, not to establish a causal or cross-architecture mechanism.

E.1 TAS distribution and per-depth visualization

Figure 3 pairs depth-aggregated TAS with the false-positive rate on distractor-present negatives, visualizing the negative correlation reported in the main text.

E.2 Controlled TAS logistic regression

To check that the TAS association is not a depth, sampling-strategy, or object-frequency artifact, we fit logistic regressions on InternVL2-8B $N = 10$ negatives with TAS standardized and standard errors clustered by haystack context. The distractor-present-only model predicts false positives from TAS while controlling for target depth, POPE strategy, and queried-object frequency. The all-

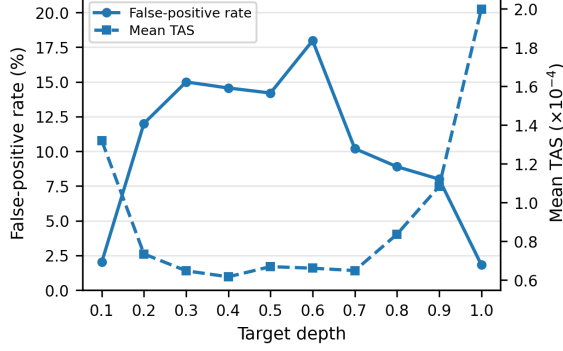


Figure 3: Target Attention Score (TAS) and false-positive rate across depths for distractor-present negatives on InternVL2-8B at $N = 10$.

negative interaction model adds a $\text{TAS} \times \text{DP}$ term under the same controls.

Model / term	OR	95% CI	p
DP-only TAS	0.424	[0.297, 0.604]	1.97×10^{-6}
All-neg. TAS	1.377	[0.922, 2.056]	.118
All-neg. $\text{TAS} \times \text{DP}$	0.279	[0.189, 0.413]	1.81×10^{-10}

Table 27: Controlled logistic regression for TAS on InternVL2-8B at $N = 10$. TAS is standardized; odds ratios are per one-standard-deviation increase. Models control for target depth, sampling strategy, and queried-object frequency, with cluster-robust standard errors by context.

E.3 Target–distractor attention contrast

TAS measures attention to the target image, but target-binding errors may also involve attention to query-object distractors. We compute a target–distractor attention ratio, $\text{TDAR}_{\max} = a_{\text{target}} / (a_{\text{dist}}^{\max} + \epsilon)$, where $a_{\text{dist}}^{\max}$ is the maximum attention over query-object distractor-image spans. Table 28 shows lower target-relative attention for false positives than correct rejections.

Metric	FP med.	CR med.	p
Target attention	6.9×10^{-5}	8.4×10^{-5}	$< 10^{-4}$
Distractor attn. mean	4.6×10^{-5}	4.1×10^{-5}	$< 10^{-4}$
Distractor attn. max	6.5×10^{-5}	5.3×10^{-5}	$< 10^{-4}$
TDAR mean	1.388	2.022	$< 10^{-4}$
TDAR max	0.953	1.578	$< 10^{-4}$
log TDAR mean	0.328	0.704	$< 10^{-4}$
log TDAR max	-0.049	0.456	$< 10^{-4}$

Table 28: Target–distractor attention contrast on all distractor-present negatives for InternVL2-8B at $N = 10$ ($n = 3402$). TDAR is the ratio of question-token attention to target-image visual tokens over attention to query-object distractor-image visual tokens.

E.4 Layer choice

We re-run the $N = 10$ InternVL2-8B experiment with hooks at layers $\{0, 8, 16, 24, 31\}$. Table 29 shows that mid-to-late layers carry the strongest TAS–FP signal, supporting the last-layer TAS choice used in the main text.

Layer	Pearson r	p	Median-split ratio
0	-0.566	.088	$1.42 \times$
8	-0.712	.021	$1.65 \times$
16	-0.842	.002	$1.52 \times$
24	-0.892	$< .001$	$1.68 \times$
31	-0.891	$< .001$	$2.32 \times$

Table 29: Per-layer all-negative TAS–false-positive relationship on InternVL2-8B at $N = 10$. Pearson r is computed across depth-aggregated points; the median-split ratio is computed at the individual-sample level on negative samples.

E.5 Head aggregation and question-token range

For each question token q and target-image visual token v , TAS averages $A_{h,q,v}^{(\ell^*)}$ uniformly over the H attention heads, as defined in Eq. 3. The question-token set $\mathcal{Q}$ is a fixed 15-token window anchored to the final $\langle \text{im_end} \rangle$ token; this avoids late-position contamination from a naive final-token window. Head-max aggregation, learned head weighting, and question-window sensitivity remain future work.

F Context Length, Cross-Architecture, and Token-Budget Results

This section provides the auxiliary tables behind the compact cross-model summary in Table 7 of the main paper.

F.1 Auxiliary cross-architecture and context-length settings

Table 30 consolidates the auxiliary context-length and cross-architecture settings. The $N = 5$ slice is interpreted as a coarse context-length baseline because several nominal depths map to the same discrete image position. The $N = 20$ run uses a low-token configuration and is used only as a limited diagnostic. Qwen2.5-VL is evaluated under constrained dynamic-resolution preprocessing and is not token-matched to InternVL2; we therefore use it only as a within-model diagnostic. Its errors are concentrated around $d = 0.7$ and many involve *person* queries; the person-excluded row is a sanity check.

Setting	n	All FP	Clean FP	Distr. FP
InternVL2, $N = 5$	6000	3.92	2.49	5.87
InternVL2, $N = 20$ LT	6000	3.33	1.39	4.12
Qwen2.5-VL	6000	1.98	0.38	3.20
Qwen2.5-VL (excl.)	4361	0.80	0.39	1.42
MiniCPM-V-2.6	6000	4.85	0.46	8.20
Idefics3-8B	6000	6.35	1.27	10.23

Table 30: Compact auxiliary context-length and cross-architecture diagnostics at $N = 10$ unless noted (percent). LT denotes the low-token configuration; “excl.” excludes *person* queries.

F.2 MiniCPM-V-2.6 and Idefics3-8B-Llama3 depth-wise decomposition

Table 31 gives per-depth profiles for the two full cross-architecture sweeps. Both preserve the clean/distractor-present separation, but their exact position profiles differ from InternVL2 and from each other: MiniCPM has elevated distractor-present rates at several positions, while Idefics3 peaks earlier at $d = 0.4$. We use these results as evidence for regime-level robustness across model families, not as evidence for a universal position curve.

Depth	MiniCPM-V-2.6				Idefics3-8B-Llama3			
	All	Clean	DP	DP r.	All	Clean	DP	DP r.
0.1	1.67	1.57	1.73	57.67	2.00	0.79	2.89	57.67
0.2	9.83	0.37	17.42	55.50	5.50	0.75	9.31	55.50
0.3	1.50	0.00	2.65	56.67	8.67	0.77	14.71	56.67
0.4	9.17	0.00	15.41	59.50	11.67	0.82	19.05	59.50
0.5	0.83	0.00	1.48	56.33	9.50	0.38	16.57	56.33
0.6	6.50	0.82	10.39	59.33	7.17	2.46	10.39	59.33
0.7	2.00	0.40	3.12	58.83	5.50	1.62	8.22	58.83
0.8	8.17	0.00	15.03	54.33	4.00	1.09	6.44	54.33
0.9	8.33	1.09	14.46	54.17	5.17	2.55	7.38	54.17
1.0	0.50	0.37	0.61	54.67	4.33	1.47	6.71	54.67
Mean	4.85	0.46	8.20	56.70	6.35	1.27	10.23	56.70

Table 31: Full $N = 10$ negative-query depth-wise decomposition for MiniCPM-V-2.6 and Idefics3-8B-Llama3 (percent). “DP r.” is the per-depth distractor-present rate.

Low-token and matched-budget interpretation.

Reducing InternVL2 to roughly 64 visual tokens/image lowers distractor-present FP to 4.23%, but the clean/distractor-present separation remains visible. The MiniCPM and Idefics3 full sweeps, together with the constrained Qwen2.5-VL diagnostic, further suggest the separation is not solely an InternVL2 token-budget artifact. Absolute cross-model rates should nevertheless be interpreted cautiously because visual preprocessing, model architectures, and token budgets are not matched.

F.3 External validation on PASCAL VOC2012

To examine external validity beyond the COCO-based setup, we use PASCAL VOC2012 val (Everingham et al., 2010) in an object-balanced setting with the same $N = 10$ multi-image format and neutral captions. On the common-support subset, each model is evaluated on 1,700 clean and 1,700 distractor-present negative queries across 17 object classes. Table 32 reports within-model regime comparisons; the significance values use Pearson χ^2 tests without continuity correction.

Model	Clean FP	DP FP	DP/Clean	p
InternVL2-8B	0.71	3.00	$4.25 \times$	7.06×10^{-7}
MiniCPM-V-2.6	0.59	13.65	$23.20 \times$	1.31×10^{-49}
Idefics3-8B	0.82	12.00	$14.57 \times$	2.26×10^{-40}
Qwen2.5-VL-7B	0.82	2.59	$3.14 \times$	7.09×10^{-5}

Table 32: PASCAL VOC2012 val object-balanced external check. Each model is evaluated on 1,700 clean and 1,700 distractor-present negatives from the common-support subset across 17 object classes, using the $N = 10$ format and neutral captions. False-positive rates are percentages. Results are intended as within-model regime comparisons rather than cross-model rankings.

Every model has a higher distractor-present false-positive rate than clean false-positive rate in this setting. For Idefics3-8B, we disable image splitting and resize images to fit the 10-image input. For Qwen2.5-VL-7B, we use the budget-relaxed H800 rerun with a higher dynamic-resolution visual budget. Because these adaptations are architecture-specific, we do not interpret the absolute rates as cross-model rankings. The VOC2012 experiment tests dataset-level external validity while retaining binary category-existence queries; it therefore does not establish task-level transfer to attributes, relations, counting, identity consistency, or free-form generation.