BACK TO THE QUESTION: TRACING VISUAL DISTRACTION THROUGH TEXT TOKENS

Anonymous authors
Paper under double-blind review

ABSTRACT

The visual world is cluttered: robust vision-language reasoning requires ignoring irrelevant information. Yet vision-language models (VLMs) can suffer substantial accuracy losses when answering questions about a target image with distractors present. Existing mitigations focus on selecting relevant images or limiting information exchange between images. Through a systematic analysis of three open 7–8B VLMs on five perception benchmarks adapted to include distractor images, we show that distractors reduce accuracy even when models correctly identify the relevant image, suggesting that selection alone is not enough. To understand the mechanism behind this interference, we combine attention analysis with targeted knockouts in Qwen2.5-VL and identify middle-layer distractor-to-text connections as a pathway through which irrelevant visual information affects reasoning. We introduce answer consistency regularization, which encourages consistent answer distributions with and without distractors. Without directly constraining attention, this objective reduces middle-layer attention from question tokens to distractor images, consistent with the pathway our analysis implicates. With five distractors, our approach improves circular accuracy by an average of 16 percentage points over the Qwen2.5-VL baseline across the five benchmarks while preserving single-image performance.

1 INTRODUCTION

Distraction tests true understanding: exams hide key variables among irrelevant data; interviews frame critical questions within long stories. The ability to reason despite distractions is essential for deploying vision-language models (VLMs) in real-world environments, where visual inputs are cluttered and noisy. Although VLMs excel on isolated single-image benchmarks, performance degrades sharply when irrelevant images are introduced (Sharma et al., 2024; Das et al., 2026; Wu et al., 2025; Liu et al., 2025a), exposing a crucial gap between evaluation and deployment.

Existing mitigations filter irrelevant images before interference (Wu et al., 2025), restrict information exchange between images (Das et al., 2026), or discourage answers grounded in the wrong image (Liu et al., 2025b). However, these interventions reduce but do not eliminate distractor interference. A possible source of residual interference is a text-mediated pathway through the model's internal question representations. In single-image description settings, Kaduri et al. (2025) show that input query-text representations encode global image context and can support response generation even when generated tokens cannot directly attend to image tokens. Masks that restrict attention between images (Das et al., 2026) leave this potential pathway open, since question tokens can still attend to distractor images (Figure 1). Whether distractors exploit this text-mediated pathway to interfere with reasoning remains an open question.

Answering a question in the presence of distractor images requires both identifying the relevant image and reasoning about it. To examine how distractors affect these two demands, we adapt five perception benchmarks (CLEVR (Johnson et al., 2017), CV-Bench (Tong et al., 2024), MIMIC (Das et al., 2026), What'sUp (Kamath et al., 2023), and BLINK (Fu et al., 2024)) to include distractor images, and evaluate three open 7–8B VLMs (Qwen2.5-VL-7B (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025), and LLaVA-OneVision-7B (Li et al., 2024)). We compare three inference regimes: standard prompting; two-stage prompting, in which the model first selects an image and then answers with all images retained; and prompting that explicitly specifies the target image. These com-

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.

parisons probe two failure modes: *selection failure*, where the model selects the wrong image, and *in-context corruption*, where distractors reduce accuracy even when the target is explicitly specified. Notably, with five distractors, explicitly naming the target image recovers on average only 12% of the drop in circular accuracy across the three models and five benchmarks, showing that distractor interference persists well beyond selection errors.

Our knockout experiments (Figure 1) show that blocking distractor information from entering text representations repairs significantly more distraction-induced failures than restricting inter-image attention. In Qwen2.5-VL with five distractors, severing distractor-to-text connections in layers 9–17 recovers 47% of failures while disrupting only 5% of correctly answered examples. In contrast, inter-image masking in the same layers recovers just 10% of failures while disrupting 56% of correct outputs. Attention analysis confirms that question tokens attend heavily to distractor images in middle layers—a pathway left entirely unblocked by inter-image masking. Together, these findings support text representations as an important pathway for distractor interference that inter-image masking leaves open.

To reduce distractors' effect on model predictions, we introduce answer consistency regularization. During fine-tuning, we minimize the Kullback–Leibler (KL) divergence between the model's answer distributions for the same question with and without distractor images, using its single-image prediction as the reference. This objective encourages consistent answers across the two contexts without explicitly constraining attention. Our attention analysis further shows that fine-tuning reduces question-token attention to distractor images in the middle layers, where knockouts identified a pathway for distractor interference.

Our contributions are threefold:

1. We distinguish selection failure from in-context corruption through comparisons that vary target guidance and which images remain in context, showing that selecting the right image does not eliminate distractor interference.
2. We identify middle-layer distractor-to-text attention as a key pathway for in-context corruption. Blocking it recovers 47% of distraction-induced failures, versus 10% for inter-image masking.
3. To reduce in-context corruption, we introduce answer consistency regularization, which aligns answers under distraction with single-image predictions. On Qwen2.5-VL with five distractors, it improves circular accuracy by 16 points, with similar gains when the target is named, and reduces middle-layer attention to distractors along the identified pathway.

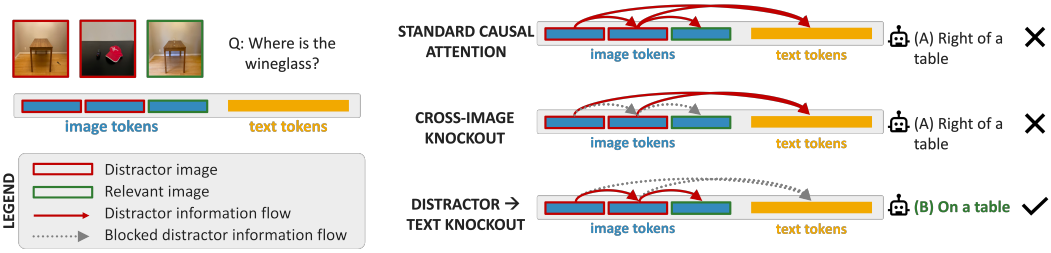

Figure 1: **How does distractor information reach the answer?** Cross-image knockout leaves distractor-to-text connections open. In this example (*Where is the wineglass?*), standard attention (top) and cross-image knockout (middle) yield incorrect answers, while distractor-to-text knockout (bottom) restores the correct answer. Knockouts target middle layers. Outlines identify distractor (red) and relevant (green) images. Arrows show allowed (solid red) and blocked (dotted gray) information flow from distractors; other connections are omitted.

2 Related Work

Multi-Image and Multimodal Reasoning. Multi-image benchmarks focus on evaluating whether VLMs can integrate information across separate visual inputs (Suhr & Artzi, 2019; Ying et al., 2024; Meng et al., 2025). Despite progress, frontier models frequently rely on superficial shortcuts rather than actual multimodal integration (Zhao et al., 2024; Fu et al., 2024; Wang et al., 2025). Success on these benchmarks does not necessarily establish robustness to irrelevant visual context.

Related work documents *cross-image information leakage* as a primary failure mode, where visual representations across images become entangled during processing (Park et al., 2025). Building on these observations, we study robustness to irrelevant visual context by augmenting perception benchmarks with distractor images, retaining a single query-relevant target.

Robustness to Visual Distractions. Prior studies show that VLMs lose accuracy as the number of distractors increases, while remaining sensitive to context length and image positioning (Sharma et al., 2024; Wu et al., 2025). Using MIMIC, Das et al. (2026) examine how relevant information is distributed across images and how visual sequence length affects performance, showing that reducing the number of visual tokens can improve counting accuracy. Our analysis extends this investigation across benchmarks, revealing that sensitivity to distractor content varies widely: while MIMIC is largely resilient, other benchmarks show far greater performance drops. To examine what underlies these performance losses, we also compare inference regimes that vary whether the target image is explicitly identified and whether distractors are included in context. This enables us to distinguish *selection failure* from *in-context corruption*.

Mitigation Methods. Prior approaches address visual distraction through *context reduction* or by improving how models process multiple images in context. Context-reduction methods use query-aware retrieval and token compression to limit the visual context supplied to the language model (Wu et al., 2025). Other approaches intervene within the model. Delimiter Token Scaling (DTS) strengthens the hidden states of image-delimiter tokens at selected layers, reinforcing image-wise attention boundaries and reducing cross-image information leakage without additional training (Lee et al., 2026). Inter-image attention masking instead restricts information exchange between image tokens during LoRA fine-tuning (Das et al., 2026). Training-based approaches also include MIA-DPO, which constructs distractor-augmented preference pairs for direct preference optimization (Liu et al., 2025b). Our complementary approach aligns prediction distributions between distracted and clean inputs. This enforces distractor robustness without restricting internal information flow, preserving single-image performance.

Mechanistic Analysis of Visual Information Flow. Mechanistic analyses of VLMs examine how visual information enters and influences textual representations. In single-image settings, Kaduri et al. (2025) use attention knockouts to show that query-token representations store global image information, while fine-grained visual details are accessed directly from image tokens. They further show that middle layers are central to this cross-modal information transfer. These findings motivate a complementary question: what internal process fails when distractors are present? Through attention analysis and targeted knockouts in Qwen2.5-VL, we find that blocking middle-layer distractor-to-text connections reduces distraction-induced errors, supporting their contribution to in-context corruption.

3 Experimental Setup

Our experiments examine why irrelevant images impair visual reasoning by considering both the model's ability to identify and use relevant evidence and the internal pathways through which distractors interfere. We construct controlled multi-image inputs from questions answerable with a single image, keeping the question and required evidence fixed. This setup supports complementary behavioral and mechanistic analyses. Inference comparisons test whether target guidance and distractor removal resolve errors. Attention analysis and targeted knockouts examine which internal connections contribute to interference.

Task Formulation. Each example contains a multiple-choice question q, a ground-truth answer a, and an image set $\mathcal{I}$. Exactly one image $I^* \in \mathcal{I}$ provides the required evidence; the others are distractors. We compare this multi-image input with its single-image counterpart containing only I^*, isolating the effect of irrelevant visual context.

Datasets. We evaluate 9,155 questions from five VQA datasets spanning spatial reasoning, counting, and perception (Table 1). We adapt each dataset to contain one relevant image and $k \in \{0, 1, 2, 3, 5\}$ distractors. The multiple-choice question, relevant image, and ground-truth answer

remain fixed across conditions, allowing us to measure how performance changes as irrelevant visual context increases.

Table 1: **Composition of the distraction test sets.** Each benchmark is evaluated at 0, 1, 2, 3 and 5 distractors; n is the number of questions, identical at every distraction level.

Benchmark	Competency	Subsets	n
CLEVR (Johnson et al., 2017)	Size reasoning	size	3554
CV-Bench (Tong et al., 2024)	Vision-centric spatial	count, relation, depth, distance	2638
What'sUp (Kamath et al., 2023)	Spatial relations	controlled_a, controlled_b	820
BLINK (Fu et al., 2024)	Low-level perception	spatial relation, relative reflectance, relative depth, localization, counting	643
MIMIC (Das et al., 2026)	Counting	count	1500
Total			**9,155**

Multi-Image Distractor Construction. To test robustness to visually plausible but irrelevant context, we sample distractors from the same dataset when available (CLEVR, What'sUp, MIMIC, and CV-Bench) and use domain-matched images for BLINK. We screen candidates using ground-truth annotations when available or VLM-based answerability checks, rejecting images judged capable of independently supporting an answer to the question. This screening aims to preserve the original answer and prevent distractors from introducing conflicting evidence. We combine accepted distractors with the target image while keeping the question and answer choices unchanged, allowing us to measure the effect of additional visual context without changing the required evidence. Figure 2 illustrates the construction process; dataset-specific details are provided in Appendix A.1.1.

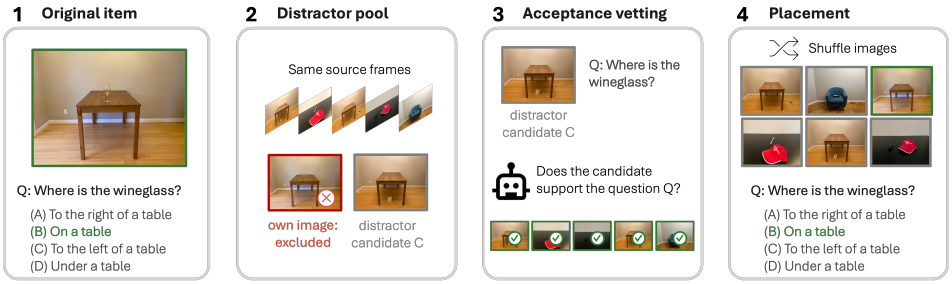

Figure 2: **Distractor augmentation**, illustrated on a What'sUp example. (1) A target image (green) provides the evidence for the question. (2) Candidate distractors are sampled from the same source, excluding the target image. (3) Candidates judged capable of independently supporting an answer are discarded. (4) Accepted distractors are shuffled with the target, keeping the question and answer choices unchanged.

Metrics To assess whether answers remain correct across option positions, we follow CircularEval (Liu et al., 2024). For example i with K_i answer choices, we evaluate all K_i circular rotations of the options, placing the correct answer in each position once. Let $c_{i,r} \in \{0, 1\}$ indicate whether the model answers example i correctly under rotation r. For an evaluation set of N examples, we report mean accuracy across rotations and circular accuracy:

$$\mathrm{Acc}_{\mathrm{mean}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_i}\sum_{r=1}^{K_i} c_{i,r}, \qquad \mathrm{Acc}_{\mathrm{circ}} = \frac{1}{N}\sum_{i=1}^{N}\prod_{r=1}^{K_i} c_{i,r}. \tag{1}$$

4 Visual Distractor Interference Persists After Correct Selection

Does visual distraction primarily prevent models from identifying the relevant image, or does irrelevant context also impair answering after successful selection? We address this question by quantifying the degradation, separating target selection from answering, and using black-image controls to test whether sequence length alone explains the effect.

Mean-rotation accuracy drops by 4 points with one distractor and 13 points with five. Across all three models, performance declines as the number of distractors k increases (Figure 3). Macro-averaged across models and benchmarks, mean-rotation accuracy drops by 4.0 percentage points with one distractor and 13.2 points with five, although the magnitude varies considerably across benchmarks. The stricter circular metric reveals even larger losses in some settings. Qwen2.5-VL's accuracy on CLEVR falls from 70.7% without distractors to 36.4% with five.

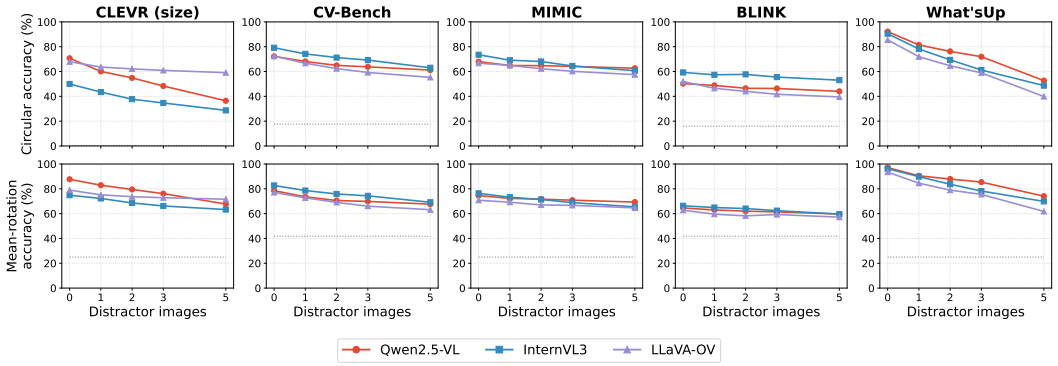

Figure 3: **Accuracy vs. number of distractor images.** All three 7-8B VLMs show degradation, with benchmark-dependent severity.

We compare three inference regimes to test whether identifying the relevant image is sufficient to overcome distraction:

- *Regular:* The model answers directly from all images without target guidance.
- *Ground-Truth:* The model answers from all images after the prompt identifies the relevant image.
- *Two-Step:* The model predicts the relevant image in phase 1, then answers from all images in phase 2 with its predicted selection identified in the prompt.

Selection failures depend strongly on the model. Selection accuracy decreases as distractors accumulate but remains above the $1/(k+1)$ chance rate in all evaluated conditions (Table 2). On CLEVR, accuracy from one to five distractors remains nearly perfect for Qwen2.5-VL ($99.9\% \rightarrow 99.4\%$), but falls from 99.9% to 84.1% for InternVL3 and from 93.4% to 29.8% for LLaVA-OV. At five distractors, Qwen2.5-VL still exceeds 81% on every benchmark, whereas LLaVA-OV reaches only 26-34%. Thus, selection failure is a major source of degradation for some models, but cannot explain Qwen2.5-VL's substantial performance loss.

Table 2: **Image-selection (phase-1) accuracy under the Two-Step protocol.** k is the number of random distractors. Chance is $1/(k+1)$.

Benchmark	Qwen2.5-VL				InternVL3				LLaVA-OV			
	$k=1$	$k=2$	$k=3$	$k=5$	$k=1$	$k=2$	$k=3$	$k=5$	$k=1$	$k=2$	$k=3$	$k=5$
CLEVR (size)	99.9	99.7	99.7	99.4	99.9	98.6	96.4	84.1	93.4	66.0	49.5	29.8
CV-Bench	95.2	93.0	91.2	85.2	93.3	90.1	81.7	61.8	90.9	56.5	46.7	26.0
MIMIC	99.7	99.5	99.2	99.1	99.9	99.1	96.3	80.5	99.0	76.3	67.1	34.3
BLINK	89.7	86.6	81.6	81.2	88.3	85.7	76.2	54.9	71.9	52.7	43.1	26.3
What'sUp	94.8	92.2	90.9	85.4	94.8	91.7	76.7	57.7	93.9	73.7	64.9	34.3

Correct target selection recovers at most a quarter of the loss. Whether the target image is named in the prompt or selected by the model itself, the performance deficit relative to clean single-image inputs largely persists (Figure 4). For Qwen2.5-VL with five distractors, against a 12.8-point mean-rotation drop under *Regular* prompting, *Ground-Truth* target naming recovers only 2.2 points (17%), while *Two-Step* selection recovers just 1.0 point (7%). Even conditioning *Two-Step* evaluation on correct phase-1 selection recovers only 24% of the degradation (Figure 4). Thus, neither specifying the target location nor verifying correct selection restores clean performance. By contrast, removing unselected images after correct selection fully restores accuracy, showing that distractors continue to interfere as long as they remain in context. Full per-benchmark results for

Qwen2.5-VL under all prompts and metrics are in Appendix A.3.2. The same analysis for InternVL3 and LLaVA-OV appears in Appendix A.3.3.

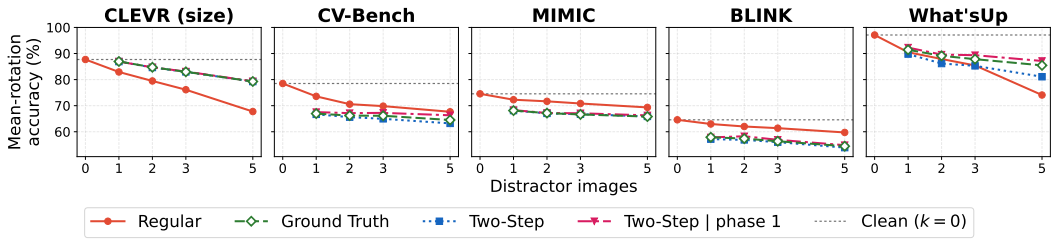

Figure 4: **Target guidance does not eliminate distraction in Qwen2.5-VL.** Mean-rotation accuracy across distractor counts under *Regular*, *Ground-Truth*, and *Two-Step* prompting. We report *Two-Step* both end to end, including phase-1 selection errors, and conditioned on correct phase-1 selection. Dashed lines indicate clean-accuracy.

Sequence length. Distractors add both visual content and sequence length. To separate these effects, we replace real distractors with solid black images of identical dimensions, preserving the number and positions of visual tokens. Under this matched-length control, mean-rotation accuracy drops by only 0.8 points with one distractor and 3.2 points with five, compared with 4.0 and 13.2 points for real distractors (Figure 17). Thus, at five distractors, competing visual content accounts for an additional 10.0-point loss—approximately 76% of the total degradation. The contribution of sequence length varies across benchmarks: it accounts for a larger share of the distraction penalty on MIMIC, while content drives most of the degradation on the other benchmarks (Table 11).

Together, these results separate selection failure from in-context corruption: distractors impair answering even when the relevant image is correctly identified. Improving selection alone therefore cannot close the gap.

5 Distractor Content Contaminates Text Representations

If selection is not the bottleneck, where does the interference arise? We analyze attention in Qwen2.5-VL-7B and use targeted knockouts to determine which connections contribute to distraction-induced errors. We focus on text-token representations as a potential pathway.

Distractors redirect question-token attention away from the relevant image. We examine how the answer readout and question tokens distribute attention across textual and visual inputs. Token-group details are provided in Appendix A.5.1. Across all benchmarks the readout attends predominantly to text, placing more than 90% of its attention there and never more than 13% on image tokens at any layer, with its attention to the question peaking sharply at layer 14 (46%; Figure 5). Introducing five distractor images barely alters this profile: despite a sixfold increase in image tokens, total image attention rises only from 3.6% to 5.1% averaged over layers, and the increase is confined to the first four layers (3.8% to 8.7%).

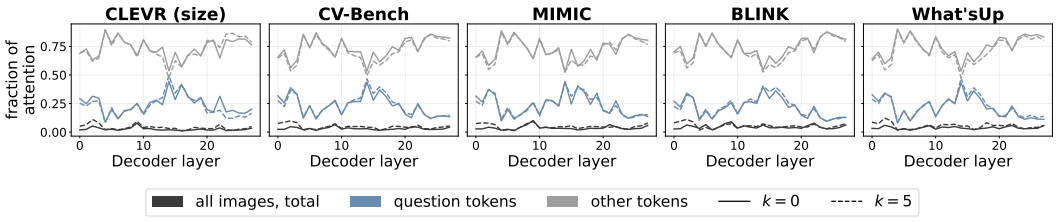

Figure 5: **Readout-token attention across decoder layers in Qwen2.5-VL-7B-Instruct.** Curves show the fraction of attention allocated to image tokens (black), question tokens (blue), and remaining tokens (gray; including template tokens, instructions, and image separators). Solid lines denote clean inputs (k=0), and dashed lines denote inputs with five distractors (k=5). Results use n=150 examples per benchmark, except What'sUp (n=300).

Question tokens, in turn, allocate about 10-15% of their total attention to images. Although this proportion changes little with distractor count, its distribution across images shifts substantially (Figure 6). Averaged across layers, the target's share of image attention falls from 0.57-0.66 with one distractor to 0.23-0.39 with five. The target remains above the uniform reference $1/(k+1)$ beyond the earlier layers, but distractors collectively receive most of the question tokens' image attention at $k = 5$. These observations motivate testing whether the same image-to-text pathway contributes to interference when irrelevant images are present.

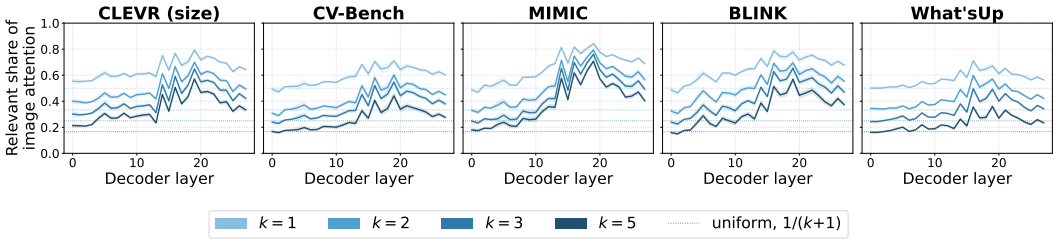

Figure 6: **Share of the question tokens' image attention on the relevant image**, $\mathrm{rel}/(\mathrm{rel} + \sum \mathrm{distractors})$, by decoder layer in Qwen2.5-VL-7B-Instruct, at $k = 1, 2, 3, 5$ distractors; dotted lines are uniform allocation $1/(k+1)$. ($n = 150$ per benchmark; 300 for What'sUp)

Attention knockouts. We test the distractor-to-text pathway by setting selected entries of the additive attention mask to $-\infty$ during prefill, thereby blocking specific connections within designated layer bands. Arrows below denote information flow from source key/value tokens to receiving query tokens. We compare four interventions:

P distractor $\rightarrow$ everything (full-deletion control).
D distractor $\rightarrow$ text (primary hypothesis).
MM image $\rightarrow$ all preceding (MIMIC cross-image mask (Das et al., 2026)).
R distractor $\rightarrow$ readout token.

We evaluate each intervention on two subsets. *Broken* examples are answered correctly without distractors but incorrectly under circular evaluation with five distractors; recovery is the fraction made correct by the intervention. *Preserved* examples are correct both with and without distractors under circular accuracy; damage is the fraction made incorrect by the intervention.

In layers 9–17, blocking distractor-to-text attention (**D**) recovers 47% of broken examples (95% CI $[45, 49]$), approaching the 56% recovered by blocking all distractor connections (**P**; $[54, 58]$). The two interventions largely repair the same failures: 92% of the examples recovered by **D** are also recovered by **P** (Table 15). Both damage only 5% of preserved examples. By contrast, the MIMIC-style cross-image mask (**MM**) (Das et al., 2026) recovers only 10% of broken examples ($[9, 11]$) while damaging 56% of preserved examples ($[55, 58]$). Blocking direct distractor-to-readout attention (**R**) is less destructive but recovers only 3% ($[2, 4]$). Together, these results identify distractor-to-text attention as an effective and selective pathway for repairing in-context corruption (Table 3).

Table 3: **Knocking out distractor→question attention repairs items that distraction broke.** Qwen2.5-VL, middle layers 9-17, pooled over 5 benchmarks at k=5 distractors. Recovery is the fraction of broken items knockout fixes; damage is the fraction of preserved items it damages. **range** and **macro** are the min–max and unweighted mean over individual benchmarks. Intervals are 95% bootstrap CIs over items (10,000 resamples).

cond.	route	recovery	range	macro	damage
P	$\mathrm{IMG}_D \rightarrow$ everything	0.56 [0.54, 0.58]	0.46–0.77	0.59	0.05 [0.04, 0.05]
MM	IMG $\rightarrow$ all preceding	0.10 [0.09, 0.11]	0.04–0.47	0.19	0.56 [0.55, 0.58]
D	$\mathrm{IMG}_D \rightarrow$ all text	0.47 [0.45, 0.49]	0.43–0.65	0.50	0.05 [0.05, 0.06]
R	$\mathrm{IMG}_D \rightarrow$ readout	0.03 [0.02, 0.04]	0.01–0.08	0.04	0.02 [0.01, 0.02]

$n_{\text{broken}} = 2353$, $n_{\text{right}} = 4162$.

Layer dependence. The effect of blocking distractor-to-text attention depends on where the intervention is applied. Masking layers 9-17 recovers 47% of broken examples, compared with 58% ($[56, 60]$) when masking all 28 layers. These nine middle layers therefore achieve about 80% of the recovery obtained by the full-depth intervention. Per-benchmark recovery ranges from 7-32%

for early-layer masking (0-8) and 5-22% for late-layer masking (18-27), with detailed results in Appendix A.5.2. The strong effect of middle-layer masking coincides with the peak in readout attention to question tokens around layer 14 (Figure 5). This identifies middle-layer distractor-to-text attention as a promising target for distractor invariance.

6 MITIGATING DISTRACTION THROUGH DISTRACTOR INVARIANCE

As established in Section 5, middle-layer distractor-to-text attention serves as a key pathway for in-context corruption. While blocking this pathway mitigates errors, hard intervention relies on distractor labels at test time. To bypass this constraint, we train the model to output consistent answer distributions regardless of distractor presence, minimizing output divergence between clean and distractor-augmented contexts during fine-tuning.

Distractor Invariance Objective. For an aligned group of inputs sharing the same multiple-choice question and target image, let p_N denote the model's probability distribution over the answer options at distractor count N. We augment the standard cross-entropy loss with a Kullback-Leibler (KL) divergence penalty relative to the clean baseline ($N = 0$ distractors):

$$\mathcal{L} \;=\; \mathcal{L}_{\text{CE}} \;+\; \frac{\lambda}{|\mathcal{N}|} \sum_{N \in \mathcal{N}} \mathrm{KL}\big(\mathrm{sg}\left[p_0\right] \,\big\|\, p_N\big)\,, \qquad \mathcal{N} = \{1, 2, 4, 8\}. \tag{2}$$

The divergence is computed over the complete option distribution rather than $P(\text{correct})$ alone, constraining how probability mass is reallocated among alternative options under distraction. The clean distribution p_0 is produced by the model itself within the same training step and is detached via a stop-gradient operator. We evaluate regularization strengths $\lambda \in \{0.1, 0.5, 1.0, 2.0\}$.

Experimental Setup. We fine-tune *Qwen2.5-VL-7B-Instruct* using LoRA (rank $r = 16$, $\alpha = 32$) on a subset of the FineVision (Wiedmann et al., 2025) dataset. Each question is expanded to include distractor counts $N \in \{0, 1, 2, 4, 8\}$, yielding 12,708 training and 1,000 validation aligned groups.

To isolate the impact of the consistency loss, we establish three primary baselines:

1. **CE-Only Control:** Fine-tuned on identical data, configurations, but with $\lambda = 0$, isolating the effect of distribution alignment beyond data augmentation.
2. **Inter-Image Attention Masking:** Re-implementation of Das et al. (2026), restricting image tokens in deep decoder layers (14–27) to their local image blocks during training and inference.
3. **Delimiter Token Scaling (DTS):** Training-free method by Lee et al. (2026), which rescales image-delimiter hidden states by $s = 8$ across layers 0–3 on the base model.

Full hyperparameter details are available in Appendix A.6.1. We evaluate models across our five held-out benchmarks at $N \in \{0, 1, 2, 3, 5\}$ distractors, reporting circular accuracy.

Results. Table 4 reports circular accuracy across all benchmarks and distractor levels. The KL-regularized model ($\lambda = 1.0$) achieves the highest pooled mean performance under distracted conditions at 69.5%, outperforming both the CE-only control (66.4%) and the base model (59.1%). At five distractors, $\lambda = 1.0$ maintains 67.4% accuracy, compared to 63.6% for CE-only and 51.4% for the base model. Performance improvements over the CE control are more pronounced on CLEVR-size, MIMIC, and What'sUp (Table 17). Increasing the regularization weight beyond optimal levels ($\lambda = 2.0$) yields a slight decrease in pooled performance (67.6%). For comparison, the reference baselines achieve pooled means of 66.5% (Inter-Image Attention Masking) and 60.5% (DTS).

These gains persist when the target is named: under *Ground-Truth* and *Two-Step* prompting, KL-clean $\lambda = 1.0$ improves circular accuracy by 15.3-16.8 points at every distractor count (Table 18), so training reduces in-context corruption, not just selection errors.

Attention after fine-tuning. Although the objective does not constrain attention, fine-tuning reshapes the pathway identified in Section 5 (Figure 7). Total image attention remains stable, but question tokens shift their image attention from distractors toward the target, concentrated in middle layers 9–17, where distractor-to-text knockouts were most effective. The shift is largest on CLEVR-size and What'sUp, which also show the largest accuracy gains. It appears in both CE-only and

Table 4: **Circular accuracy under distraction.** Each benchmark column reports mean accuracy over $k \in \{1, 2, 3, 5\}$ distractor images. Pooled columns average across the five benchmarks and report accuracy at $k = 0$, at $k = 5$, and the mean over $k > 0$. Δ_{base} is the pooled mean over $k > 0$ minus the base model's corresponding mean. Per-k results are shown in Table 17; boldface marks the best model in each column.

	Per benchmark (k>0 mean)					Pooled			
Model	**CLEVR**	**CV-B**	**MIMIC**	**WhatsUp**	**BLINK**	**k=0**	**k=5**	**mean**	**Δ_{base}**
Qwen2.5-VL-7B *(base)*	49.9	64.5	64.1	70.6	46.4	70.7	51.4	59.1	0.0
Ours									
CE only	77.9	68.1	56.4	77.0	52.6	71.9	63.6	66.4	+7.3
CE + KL λ=0.1	76.1	67.4	60.6	77.5	52.6	73.6	63.7	66.8	+7.7
CE + KL λ=0.5	75.0	**68.9**	64.8	**79.8**	**54.5**	74.2	66.0	68.6	+9.5
CE + KL λ=1.0	**82.1**	66.9	**65.8**	79.0	53.8	**74.7**	**67.4**	**69.5**	**+10.4**
CE + KL λ=2.0	78.4	67.4	60.0	78.1	53.8	73.4	64.9	67.6	+8.4
Reference									
CE + MIMIC mask (Das et al., 2026)	79.4	68.4	57.7	74.9	51.9	72.7	62.7	66.5	+7.4
DTS *(no training)* (Lee et al., 2026)	59.3	64.0	60.3	72.6	46.4	70.6	55.6	60.5	+1.4

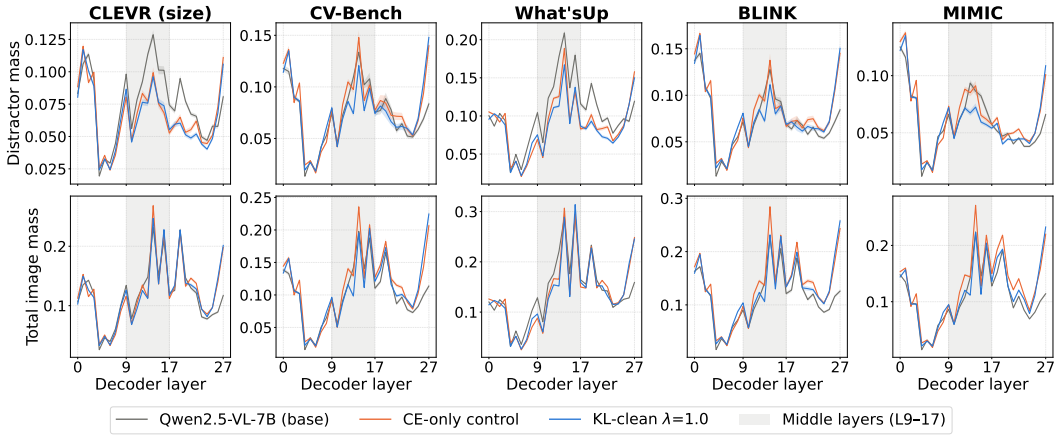

Figure 7: **Question-token attention to images by decoder layer.** Attention is averaged over heads and question-token query positions, then summed over each image's visual-token span. Rows show distractor attention mass, and total image attention mass. Columns show the five benchmarks at $k = 5$ ($n = 150$ per benchmark; $n = 300$ for What'sUp). Curves compare the frozen base model, the CE-only control, and KL-clean model $\lambda = 1$.

KL-clean models, indicating that it stems from training on distractor-augmented data, while the consistency term adds further accuracy gains (Table 4).

7 Conclusion

Irrelevant images can degrade VLM performance even when target visual context is correctly identified. To systematically study this failure mode, we adapt five visual perception benchmarks to incorporate controlled distractor images. Attention knockout experiments on Qwen2.5-VL reveal that a substantial share of visual interference flows through text-token representations; specifically, blocking distractor-to-text attention in layers 9–17 restores 47% of failure cases caused by distractors. To mitigate this vulnerability, we enforce answer-level distractor invariance by aligning model output distributions on clean and distracted inputs. Across five held-out benchmarks, optimizing KL divergence to clean distributions ($\lambda = 1.0$) improves pooled circular accuracy by 16 percentage points over the base model under five-distractor conditions. Attention analysis further indicates that fine-tuning shifts question-token attention toward target visual features, particularly within intermediate layers. Finally, our evaluations rely on randomly sampled distractors, leaving open how specific distractor properties impact robustness. Investigating structured spatial distractors will be essential to decouple perceptual confusion from spatial reasoning failures, while extending this framework to video domains will test robustness under dynamic, time-varying distractor contexts.

AI USE STATEMENT

We used generative AI tools to assist with language editing, code development, debugging, and constructing the distractor-augmented evaluation examples detailed in Section A.1.1.

Generative AI was not used to do a literature review, generate experimental results, determine the scientific conclusions of the paper, or replace the authors' evaluation of the work. The authors reviewed all AI-assisted text and code, and tested AI-assisted code before using it in the experimental pipeline. Reported numerical results were produced by the authors' experimental pipelines or transcribed from the cited original sources and independently checked against those sources. All interpretations and scientific claims were independently verified by the authors, who take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce our dataset construction, evaluation, mechanistic analyses, and fine-tuning experiments. Section 3 defines the task, datasets, models, and evaluation metrics. Appendix A.1.1 documents the benchmark-specific distractor sources, exclusion criteria, answerability-filter prompts, parsing rules, image transformations, and validation procedures. Appendix A.2 reports the prompting protocols, visual-token budgets, decoding settings, and other inference details. Sections 5, and A.5.2 specify the attention measurements, knockout interventions, and evaluated layer ranges, while Appendix A.6.1 and Appendix A.7 detail the targeted finetuning setup and the construction of the training set respectively. We use fixed evaluation examples across experimental conditions and report results across all evaluated benchmarks and distractor levels. Upon acceptance, we will publicly release the code, dataset-construction, evaluation scripts, and prompts needed to reproduce the reported experiments.

REFERENCES

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, 2025. URL https://arxiv.org/abs/2502.13923.

Anurag Das, Adrian Bulat, Alberto Baldrati, Ioannis Maniadis Metaxas, Bernt Schiele, Georgios Tzimiropoulos, and Brais Martinez. More images, more problems? a controlled analysis of vlm failure modes. *arXiv preprint arXiv:2601.07812*, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14549–14558, 2025.

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

Minyoung Lee, Yeji Park, Dongjun Hwang, Yejin Kim, Seong Joon Oh, and Junsuk Choe. Enhancing multi-image understanding through delimiter token scaling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7QFf05KrOm.

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

Ming Liu, Hao Chen, Jindong Wang, and Wensheng Zhang. On the robustness of multimodal language model towards distractions. *arXiv preprint arXiv:2502.09818*, 2025a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 34583–34610, 2025b.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 38405–38453, 2025.

Yeji Park, Minyoung Lee, Sanghyuk Chun, and Junsuk Choe. Mitigating cross-image information leakage in multi-image understanding with large vision-language models. *arXiv preprint arXiv:2508.13744*, 2025.

Aditya Sharma, Michael Saxon, and William Yang Wang. Losing visual needles in image haystacks: Vision language models are easily distracted in short and long contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5429–5451, 2024.

Alane Suhr and Yoav Artzi. Nlvr2 visual bias analysis. *arXiv preprint arXiv:1909.10411*, 2019.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pp. 62624–62650, 2025.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, David Chan, et al. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 64592–64615, 2025.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024.

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.

A APPENDIX

A.1 EXPERIMENTAL SET-UP

A.1.1 DATASET AUGMENTATION

We will release all distractor-augmented benchmarks upon acceptance to support reproducibility.

Every benchmark we use was written for a single image input. To study distraction, we therefore construct multi-image versions of each, in which the original relevant image is presented alongside k distractor images that do not support the answer to the question. The underlying task and its ground-truth answer are preserved, but the surface form is adapted where a single-image phrasing would no longer be well posed once several images are shown. Questions lose their references to a specific image, so that *How many organs are in the image?* becomes *How many organs are there?*; the rewritten question is used at every level, including k=0, so wording never covaries with the number of images. Answer options are adapted in the same spirit where the original task was framed around one image: WhatsUp, for instance, is natively a caption-selection task whose options are complete sentences, which is ambiguous when a caption may also hold of a distractor. We reframe it as a question over spatial relations, with the relation alone as the option. We build $k \in \{1, 2, 3, 5\}$ for every benchmark, together with the unmodified k=0 case, and hold the row set fixed across levels so that a degradation curve compares the same questions at every point.

All five benchmarks share one skeleton: (1) build a distractor candidate pool, (2) exclude images that could legitimately answer the question, (3) vet what remains, and (4) place eligible distractors alongside the relevant image in the shuffled order. Table 5 gives the per-benchmark steps.

Table 5: **Pipeline steps by benchmark.** All benchmarks share the core structure: pool construction, exclusion, vetting, and placement. Key differences are the vetting method and any special handling required.

Benchmark	Pipeline Steps
CLEVR	1. Load questions filtered by category (size/color/count/material) 2. Build attribute index from scene graphs 3. Per-question vetting: exclude renders containing queried attribute 4. Sample k distractors, shuffle with relevant image
What's Up	1. Parse caption to extract subject + spatial relation 2. Reframe as MCQ: "Where is/are the X?" 3. Domain-stratified pool selection (tabletop vs. natural) 4. Structural exclusion (own image, other subsets' images) 5. VLM answerability gate (per-candidate check) 6. Sample k distractors, shuffle with relevant image
BLINK	1. Load COCO val2017 + Visual Genome as candidate pool for distractors 2. Remove duplicate images 3. VLM answerability gate 4. Sample k distractors, shuffle with relevant image
CV-Bench	1. Parse questions to extract queried COCO categories 2. Route: COCO pool (if annotated) or CV-Bench pool (if not) 3. COCO route: exclude images annotated with queried class 4. Own pool route: VLM answerability gate 5. Draw colored marker boxes on distractors (1,653 annotated rows) 6. Sample k distractors, shuffle with relevant image
MIMIC	1. Use MIMIC's own images as in-domain pool 2. Structural exclusion (remove own image) 3. Sample k distractors, shuffle with relevant image

Distractor pools. Table 6 details the source and size of each distractor pool. Where available, distractors are drawn directly from the benchmark's own image distribution to ensure they are visually confusable with target images and cannot be dismissed based on rendering style alone. BLINK is an exception: because its images contain drawn annotations essential to the questions and span mul-

tiple source corpora, it uses a "matched-domain" pool of real photographs matched to image kind rather than exact distribution. CV-Bench uses a hybrid strategy: queries matching COCO categories are vetted directly against COCO annotations, while remaining queries fall back to the benchmark's unannotated images. MIMIC requires no external pool, as it natively provides six distractor images per item; lower pool sizes are generated by subsampling these native distractors.

Table 6: Distractor pool sources by benchmark. Pool size indicates the number of candidate images available for distractor selection. What's Up uses domain-stratified pools to avoid cross-domain artifacts.

Benchmark	Pool Source	Pool Size	Domain Type
CLEVR	CLEVR val renders	15,000	In-domain
What's Up (controlled_a/b)	controlled_images + controlled_clevr	829	In-domain
What's Up (coco_qa)	COCO val2017	5,000	In-domain
What's Up (vg_qa)	Visual Genome	4,797	In-domain
BLINK	COCO val2017 + Visual Genome	∼9,800	Matched-domain
CV-Bench	COCO train2017 (annotated rows) CV-Bench box-free (unannotated)	118,287 985	Hybrid
MIMIC	MIMIC dataset images	∼1,500	In-domain

Vetting. A candidate image is only a valid distractor if it cannot answer the target question. Because metadata formats vary, we apply the strongest guarantee supported by each benchmark rather than a single uniform rule (Table 7). Three benchmarks permit exact verification: CLEVR leverages scene graphs for ground-truth attribute matching; CV-Bench excludes images labeled with the target COCO category for annotated rows; and MIMIC's native distractors are verified against COCO instance counts to confirm irrelevance. Where no such guarantee exists, we ask the row's own question of each candidate *in isolation* and reject the whole set if any image could support an answer; verdicts are single-token and strictly parsed, and a rejected set is redrawn, up to 15 attempts.

```
Question: Where is the beer bottle?
Answer choices: Under an armchair, On a armchair, To the left of a armchair, To the right of
     a armchair

Does this image contain any of the objects the question asks about, or could this image on
     its own support answering the question?

Answer with exactly one word: YES if it contains any of those objects or could support an
     answer, NO if it is completely unrelated. No explanation.
```

Figure 8: **Answerability gate prompt.**

Transformations. Accepted distractors are resized to match the reference image's exact dimensions (Table 8). Beyond resizing, only CV-Bench requires active visual editing. In CV-Bench, 1,653 of 2,638 questions target an object highlighted by a colored box drawn on the image; unedited distractors lacking these box strokes would allow models to trivially identify the target image via a "pick the image with markers" heuristic without reading the question. We address this by drawing synthetic boxes on distractors using the exact colors named in the question. Marker pixel area is matched to that of the target image with added per-distractor jitter, placing the target image's marker density within the overall distribution rather than at an extreme.

A.1.2 Examples

Figure 9 presents sample benchmark items at the highest distraction level ($k = 5$). Each item consists of one relevant target image and five distractors sampled from the same domain, presented in randomized order alongside the question and options. All prompts use neutral phrasing stripped of single-image references so that question text does not vary with k. For CV-Bench's annotated

Table 7: **VLM gate configurations.** Each benchmark uses the strongest vetting guarantee its metadata supports. CLEVR uses exact scene-graph vetting; other benchmarks use VLM-based answerability checks with identical prompts and model.

Benchmark	Gate Type	Description
CLEVR	Ground truth	Scene graphs provide exact attribute vetting; no VLM needed
What's Up	Answerability	Per-candidate check: "Does this image contain any of the objects the question asks about, or could this image on its own support answering the question?"
BLINK	Answerability	Same prompt as What's Up; applied per candidate image
CV-Bench	Hybrid: COCO annotations or Answerability	Annotated rows: COCO category exclusion (ground truth) Unannotated rows: same answerability prompt
MIMIC	Answerability	Same prompt as What's Up; applied per candidate image

Table 8: **Special transformations applied during distractor construction.** Most benchmarks only require size normalization; CV-Bench additionally requires marker drawing to match visual cues in annotated questions.

Benchmark	Transformation	Details
CV-Bench	Marker drawing	Extract marker colors from question text Draw matching colored bounding boxes on distractors Match marker pixel mass to relevant image
BLINK	Duplicate filtering	Duplicate detection Remove duplicate images
All	Resize to match	Distractors resized to relevant image dimensions
All	Alternative variants	Random noise images (rnd) Black images (black) Used as control conditions

questions (row 5), distractors feature matched synthetic boxes to ensure marker presence cannot be exploited as a shortcut.

A.1.3 Answer Distribution and Relevant Image Distribution

Because our distractor construction inserts the relevant image at a random slot, its position carries no information about the answer. Figure 10 verifies this: across all benchmarks and distractor counts, the relevant-image index is close to uniform, so accuracy under distraction reflects genuine cross-image grounding rather than a positional shortcut.

The same holds for the answers themselves: because each question is evaluated under all four circular rotations of its options, the correct answer appears in every position exactly once, eliminating answer-position bias by construction.

A.2 Inference Details.

Before examining whether selection alone explains distractor degradation, we must ensure that observed accuracy drops reflect genuine interference rather than context-truncation artifacts.

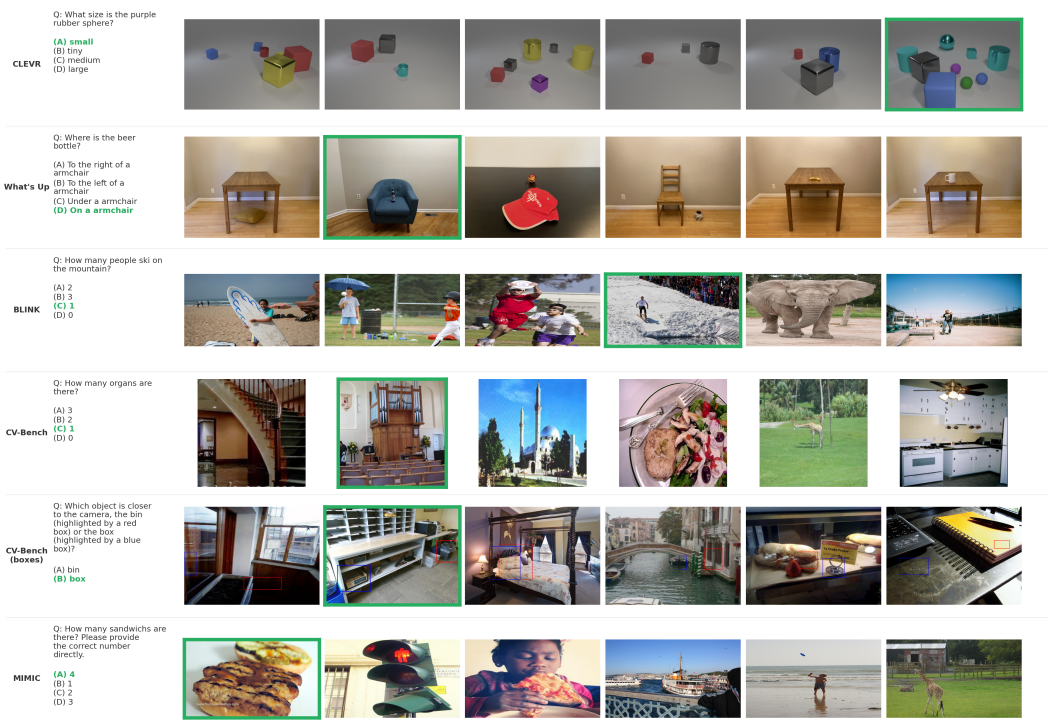

Figure 9: **Example items at the maximum distraction level ($k = 5$).** Each row shows a question, its choices (correct answer in green), and the six images shown to the model: one relevant image (green outline) among five distractors from the same image distribution, in shuffled order.

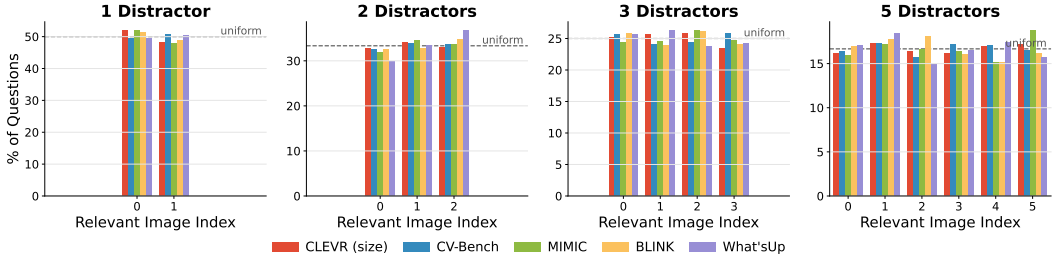

Figure 10: **The relevant image is placed uniformly across positions at every distractor level.** For each benchmark, we plot the fraction of questions in which the question-relevant image occupies each slot, split by the number of distractors (1, 2, 3, 5). Distributions track the uniform reference (dashed line) across all positions and benchmarks, confirming that a model cannot exploit a fixed-position prior to locate the relevant image.

A.2.1 Budget Selection

Multi-image evaluation with $k = 5$ distractors requires careful management of the context length. Each image consumes visual tokens proportional to its resolution, and with 6 images per sample (1 relevant + 5 distractors), high-resolution inputs risk exceeding the model's 32K context window, leading to truncation.

We evaluated four visual token budgets (256, 512, 1280, and native (uncapped)) on 200 samples per benchmark at $k = 0$ (single image) using Qwen2.5-VL-7B. Figure 11(a) shows that accuracy is preserved across all budgets, with no significant degradation at 1280 tokens compared to native resolution. Figure 11(b) estimates worst-case context length at k=5 by computing $6\times$ the maximum visual tokens observed plus text overhead. At the 1280-token budget, all benchmarks remain safely under the 32K limit, with BLINK as the tightest constraint at approximately 28K tokens.

Based on these results, we use 1280 visual tokens per image (corresponding to 1M pixels) for all multi-image experiments, ensuring no context truncation while maintaining accuracy parity with native resolution.

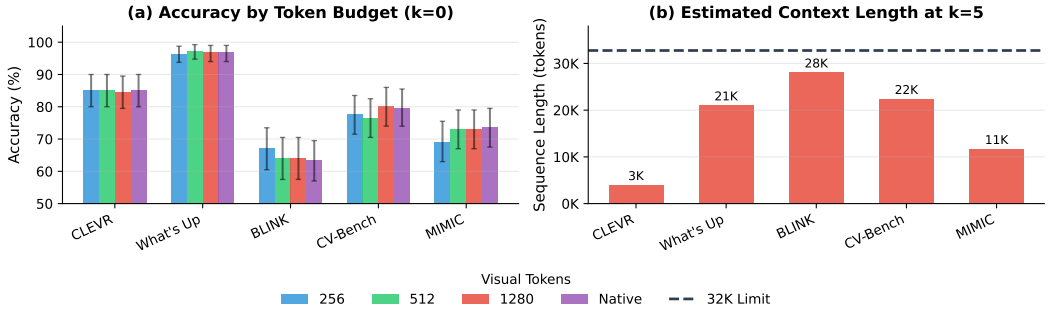

Figure 11: **Token budget selection for multi-image evaluation.** (a) Accuracy at k=0 (single image) across token budgets; 1280 tokens matches native resolution. (b) Worst-case context length at k=5, estimated as 6× max visual tokens plus text overhead. The 1280-token budget keeps all benchmarks under the 32K limit while preserving accuracy. Results based on 200 samples per dataset with 95% bootstrap confidence intervals.

A.2.2 PROMPTS

Regular. No hint about distractors; the model must implicitly identify the relevant image.

```
{question}
A. {choice_a}  B. {choice_b}
C. {choice_c}  D. {choice_d}
Answer with the option's letter from the given choices directly.
```

Ground Truth (Oracle). The relevant image index is revealed to the model (upper bound).

```
You are given {n} images (Image 1, ..., Image {n}).
Only Image {k} is relevant for answering the question; ignore the other images.
{question}
A. {choice_a}  B. {choice_b}
C. {choice_c}  D. {choice_d}
Answer with exactly one letter (A, B, C, D). No punctuation, no explanation.
```

Two-Step/Hard Isolation, Phase 1: Image Selection. First, ask only which image is relevant.

```
You are given {n} images. Only one is relevant for answering the question below.
{question}

Which image should be used to answer the question?

A. First image
B. Second image
...
Answer **only** with A, B, ... No additional text.
```

Two-Step, Phase 2: Answer. Condition the answer on the selected image from Phase 1.

```
You are given {n} images (Image 1, ..., Image {n}).
Only Image {selected} is relevant for answering the question; ignore the other images.
{question}
A. {choice_a}  B. {choice_b}
C. {choice_c}  D. {choice_d}
Answer with exactly one letter (A, B, C, D). No punctuation, no explanation.
```

Hard Isolation, Phase 2: Answer. After Phase 1, distractors are *physically removed* from context; only the selected image is shown with the Regular prompt.

```
{question}
```

```
A. {choice_a}  B. {choice_b}
C. {choice_c}  D. {choice_d}
Answer with the option's letter from the given choices directly.
```

Note: Two-Step Phase 2 and Ground Truth share the same phase 1 template; they differ in phase 2. Hard Isolation uses the Regular prompt but over a single image.

A.3 RESULTS

A.3.1 BASELINE.

Table 9 reports the per-benchmark accuracies underlying Figure 3, for distractor counts $k \in 0, 1, 2, 3, 5$. From zero to five distractors, mean-rotation accuracy declines on every benchmark for all three models. What'sUp shows the largest drops: 23.0, 26.4, and 31.7 percentage points for Qwen2.5-VL, InternVL3, and LLaVA-OV, respectively. BLINK shows the smallest drops, ranging from 4.9 to 6.7 percentage points.

Table 9: **Baseline accuracy under random-image distraction.** All three metrics are computed on the same runs: vanilla is the rotation-0 pass (actual benchmark question), circular requires every rotation to be correct, and mean-rotation is the per-item fraction of rotations passed, so circular $\leq$ mean-rotation holds cell by cell. k is the number of distractor images; k=0 is the clean single-image baseline.

Benchmark	Qwen2.5-VL					InternVL3					LLaVA-OV				
	$k=0$	$k=1$	$k=2$	$k=3$	$k=5$	$k=0$	$k=1$	$k=2$	$k=3$	$k=5$	$k=0$	$k=1$	$k=2$	$k=3$	$k=5$
Vanilla accuracy															
CLEVR (size)	88.0	83.3	79.1	75.1	67.6	74.6	71.7	69.4	65.4	63.3	79.0	74.9	73.5	72.9	70.6
CV-Bench	79.1	73.6	71.2	70.4	68.4	82.9	78.7	76.2	74.9	69.7	78.0	73.0	70.1	66.0	64.0
MIMIC	73.6	72.3	71.8	71.1	68.9	76.3	73.5	70.6	69.2	65.7	70.9	69.5	66.3	66.4	65.0
BLINK	66.1	64.7	63.8	62.1	60.7	66.1	65.9	65.2	62.2	58.9	65.3	60.0	59.1	60.3	56.3
What'sUp	97.2	90.6	88.9	85.9	74.1	95.7	91.0	84.3	78.7	71.3	93.2	84.0	78.2	77.2	62.0
Circular accuracy															
CLEVR (size)	70.7	60.1	54.9	48.3	36.4	50.0	43.4	37.7	34.7	28.7	67.9	63.6	62.1	60.9	59.1
CV-Bench	72.2	68.1	64.9	63.8	61.2	79.2	74.2	71.2	69.2	63.0	72.4	66.7	62.4	59.2	55.3
MIMIC	68.0	64.9	64.7	64.1	62.7	73.5	69.1	68.1	64.5	60.7	66.7	64.9	62.3	60.2	57.5
BLINK	50.2	48.8	46.5	46.3	44.0	59.3	57.4	57.7	55.5	53.0	51.9	46.5	44.0	41.7	39.5
What'sUp	92.2	81.5	76.2	72.0	52.7	90.5	78.2	69.4	61.2	48.7	85.5	72.0	64.6	58.9	39.9
Mean-rotation accuracy															
CLEVR (size)	87.7	82.9	79.5	76.1	67.7	74.9	72.2	68.6	66.2	63.3	79.0	75.1	73.7	72.8	71.6
CV-Bench	78.5	73.6	70.6	69.8	67.7	82.7	78.6	75.9	74.3	69.2	77.1	72.7	69.1	66.0	63.2
MIMIC	74.6	72.3	71.7	70.8	69.3	76.4	73.2	71.4	68.9	65.5	70.8	69.2	67.0	66.7	64.6
BLINK	64.6	63.0	62.0	61.4	59.7	66.3	64.9	64.1	62.4	59.6	62.8	59.6	58.2	59.3	57.2
What'sUp	97.1	90.5	87.9	85.4	74.1	96.2	89.8	83.7	78.2	69.8	93.5	84.5	79.0	75.5	61.8

A.3.2 PROMPTING STRATEGIES ON QWEN2.5-VL

Table 10 reports the full results with the different prompts on Qwen2.5-VL.

A.3.3 PROMPTING STRATEGIES ON INTERNVL3 AND LLAVA-OV.

Figures 12 and 13 replicate the prompting-strategy analysis of Fig. 4 on InternVL3-8B and LLaVA-OneVision. Two model-specific patterns emerge.

On InternVL3, the ground-truth, two-step, and hard-isolation curves are essentially indistinguishable from the regular baseline across all five benchmarks: revealing the relevant image, or even removing the distractors entirely, yields no consistent gain. Conditioning instead on correct phase-1 selection separates the two failures: hard isolation then recovers clean-image accuracy, whereas ground-truth and two-step do not. The residual end-to-end penalty is thus attributable to selection errors compounding with distractor interference, and the pattern mirrors the main-text finding: locating the relevant image is not sufficient—only removing the distractors from the input restores performance.

LLaVA-OV illustrates the other half of the decomposition: whereas InternVL3 selects reliably and suffers mainly from residual interference, LLaVA-OV's dominant failure is selection itself. Hard isolation (the best strategy for Qwen2.5-VL) is the worst here, falling well below the regular baseline as distractors are added. Because the two-step and hard-isolation curves are reported end-to-end, they inherit phase-1 selection errors, and LLaVA-OV's selection accuracy collapses toward chance with

Table 10: **Qwen2.5-VL under the relevance prompts, random-image distraction.** Ground Truth names the relevant image; Two-Step selects it then answers; Hard Isolation re-asks with only the selected image in context. k is the number of distractor images. Pooled is the mean of the five benchmark rows.

Benchmark	Ground Truth				Two-Step				Hard Isolation			
	$k=1$	$k=2$	$k=3$	$k=5$	$k=1$	$k=2$	$k=3$	$k=5$	$k=1$	$k=2$	$k=3$	$k=5$
Vanilla accuracy												
CLEVR (size)	86.4	84.6	83.3	78.6	86.4	84.6	83.2	78.3	87.4	87.0	87.3	87.3
CV-Bench	70.3	69.2	69.0	67.0	70.2	68.4	67.9	65.6	78.0	77.2	75.8	74.7
MIMIC	68.4	67.0	66.8	65.7	68.3	66.9	66.9	65.8	74.5	74.3	73.8	73.4
BLINK	60.7	60.5	58.0	56.6	59.9	59.9	57.4	56.5	64.9	64.1	63.3	62.8
What'sUp	91.8	90.1	88.5	84.5	90.4	87.8	86.3	78.7	93.2	92.1	91.2	85.9
Pooled	75.5	74.3	73.1	70.5	75.0	73.5	72.3	69.0	79.6	78.9	78.3	76.8
Circular accuracy												
CLEVR (size)	68.2	64.5	61.0	53.0	68.2	64.5	60.9	52.9	71.2	71.3	70.4	70.3
CV-Bench	42.2	42.3	41.4	38.8	41.9	41.8	40.5	37.7	71.5	71.0	69.2	68.4
MIMIC	45.9	47.0	47.2	45.7	45.9	47.0	47.2	45.6	67.9	67.3	67.1	67.2
BLINK	26.1	25.3	23.6	21.8	25.2	24.7	23.6	21.3	48.4	48.4	48.2	47.0
What'sUp	80.6	75.7	76.2	70.1	78.4	71.2	72.2	64.6	87.3	84.6	84.5	81.0
Pooled	52.6	51.0	49.9	45.9	51.9	49.8	48.9	44.4	69.2	68.5	67.9	66.8
Mean-rotation accuracy												
CLEVR (size)	87.0	84.7	83.0	79.3	86.9	84.6	83.0	79.2	87.9	87.8	87.6	87.1
CV-Bench	67.0	66.2	66.1	64.5	66.7	65.6	65.0	63.2	77.4	76.8	75.3	74.2
MIMIC	68.1	67.2	66.6	65.8	68.1	67.0	66.7	65.9	74.2	74.2	73.7	73.9
BLINK	57.9	57.4	56.4	54.5	57.2	56.9	56.0	53.9	62.8	62.4	62.6	61.5
What'sUp	91.4	89.1	87.8	85.5	89.8	86.2	85.3	81.1	93.1	91.7	90.8	86.7
Pooled	74.3	72.9	72.0	69.9	73.7	72.1	71.2	68.7	79.1	78.6	78.0	76.7

more distractors (Tab. 2). Hard isolation is maximally sensitive to this failure: once the wrong image is selected, the evidence-bearing image is discarded from the input altogether, making a correct answer impossible. Inspecting the phase-1 responses (Figure 16) reveals the source of the collapse: the model exhibits a strong positional bias toward option B). It selects the second image in 35-60% of cases across all five benchmarks, well above the 25% chance rate (one-proportion binomial test, $p < 10^{-10}$), even though the relevant image is near-uniformly distributed over positions.

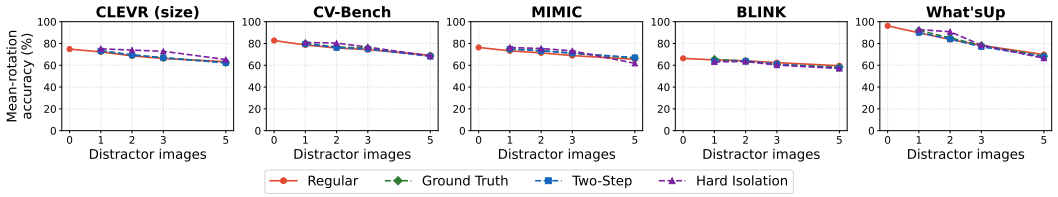

Figure 12: **Prompting strategies under distraction (InternVL3-8B)**. Mean-rotation accuracy as a function of the number of distractor images for four inference strategies (Regular, Ground-truth, Two-step, and Hard isolation) on each benchmark.

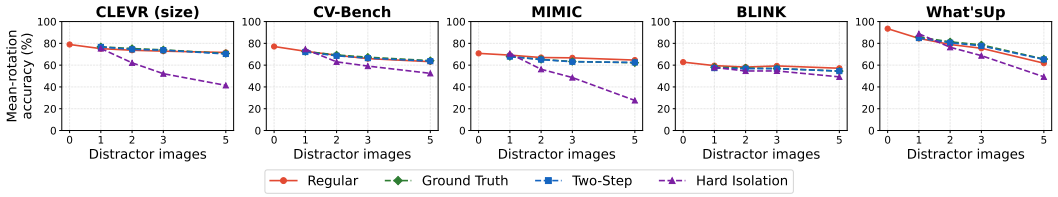

Figure 13: **Prompting strategies under distraction (LLaVA-OV)**. Mean-rotation accuracy as a function of the number of distractor images for three inference strategies (Regular, Ground-truth, Two-step) on each benchmark.

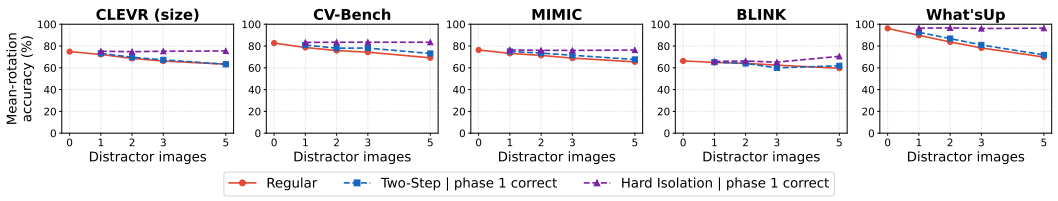

Figure 14: **Correctly locating the relevant image does not undo distractor-induced degradation. (InternVL3-8B)**

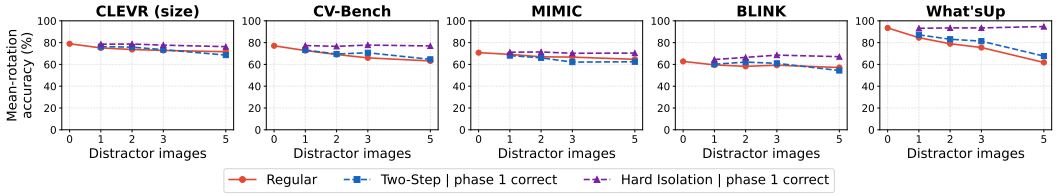

Figure 15: **Correctly locating the relevant image does not undo distractor-induced degradation. (LLaVA-OneVision-7B)**

A.4 Sequence Length

A natural hypothesis for distractor-induced degradation is that longer visual context simply dilutes attention or exceeds the model's effective context window. To test this, we construct a *black-image control* for every distractor variant: each distractor image is replaced by a solid black image of identical pixel dimensions, preserving the number and positions of visual tokens while removing distractor *content*.

Results. Figure 17 shows accuracy as a function of the number of distractors under the black-image control. In stark contrast to the random distractors (Figure 3), the curves are more robust: averaged over three models and five benchmarks, mean-rotation accuracy drops by only 0.8 points at k=1 and 3.2 points at k=5, compared to 4.0 and 13.2 points respectively under random distractors.

Table 11 quantifies the per-benchmark degradation. We report the fraction of the penalty attributable to visual content (as opposed to sequence length) :

$$\text{Content fraction} = \frac{\Delta_{\text{rnd}} - \Delta_{\text{black}}}{\Delta_{\text{rnd}}}$$

where Δ denotes the drop from the clean baseline k=0 to k=5. Across all model-benchmark pairs, 76-79% of the degradation is attributable to competing visual content rather than context dilution.

Table 11: **Content vs. sequence-length contribution to distractor penalty.** Δ_{rnd} and Δ_{black} are the drops in mean-rotation accuracy from k=0 to k=5 distractors. The content fraction measures how much of the penalty is attributable to distractor content rather than sequence length: $(\Delta_{\text{rnd}} - \Delta_{\text{black}})/\Delta_{\text{rnd}}$.

	Qwen2.5-VL			InternVL3			LLaVA-OV		
Benchmark	Δ_{rnd}	Δ_{black}	Content	Δ_{rnd}	Δ_{black}	Content	Δ_{rnd}	Δ_{black}	Content
CLEVR (size)	20.0	1.7	91%	11.6	4.3	63%	7.3	−1.4	119%
CV-Bench	10.8	3.0	72%	13.5	2.2	84%	13.9	3.0	78%
MIMIC	5.2	4.0	24%	10.9	6.1	44%	6.2	2.2	63%
BLINK	4.8	0.7	86%	6.7	1.3	80%	5.6	1.9	66%
What'sUp	23.0	2.6	89%	26.3	6.8	74%	31.7	8.8	72%
Average	12.8	2.4	72%	13.8	4.2	69%	12.9	2.9	80%

Benchmarks differ in how much degradation persists under the matched-length black-image control. On CLEVR (size), What'sUp, and BLINK, replacing real distractors with black images substantially

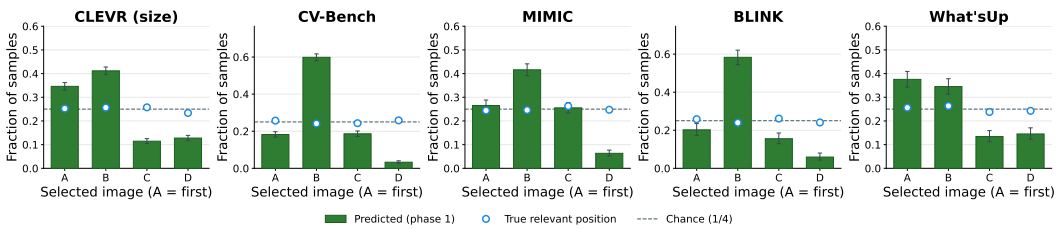

Figure 16: **Phase-1 image selection is positionally biased toward option B (second image) in LLaVA-OV.** We zoom on the case where there are 3 distractors. For each benchmark, green bars give the distribution over the image slot selected in phase 1 (A=first image, ...). Error bars are 95% Wilson score confidence intervals. Although the relevant image is near-uniformly distributed across positions, the model selects option B far above chance (25%) in every benchmark.

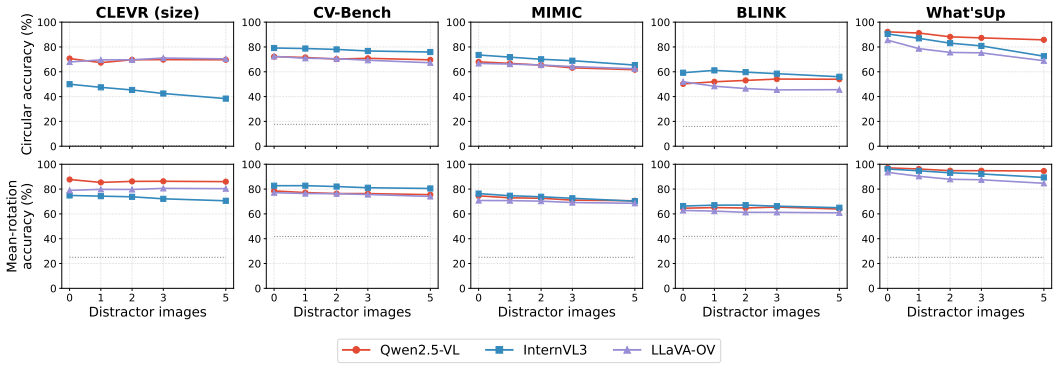

Figure 17: **Degradation under black-image distractors (sequence-length control).** Same layout as Figure 3, but each distractor is replaced by a solid black image of identical dimensions. The curves are more flat, demonstrating that sequence length alone does not explain the distractor penalty.

reduces the accuracy loss, suggesting that visual content contributes beyond sequence length alone. On BLINK, this reduction accounts for 66-86% of the original penalty. MIMIC retains a larger proportion of the degradation under the control: for Qwen2.5-VL, approximately three-quarters of the penalty persists with black images. This pattern is consistent with a greater contribution from sequence length on MIMIC.

A.5 Where does the interference occur? An internal representation analysis

A.5.1 Attention Analysis

All attention measurements use Qwen2.5-VL-7B-Instruct (28 decoder layers, 28 heads, indexed 0-27) at `max_pixels`$= 1,003,520$, the budget used for the accuracy evaluations. Values are head-averaged and normalized by the query row's total, so each row sums to 1 over the token groups and "group g receives x% of the attention" is well defined.

Token groups. Every sequence position is assigned to one of the groups illustrated in Table 12.

Where the readout's attention goes. Figure 18 decomposes the readout row across all groups.

A.5.2 Attention Knockouts

We use attention knockouts to test whether specific information-flow routes contribute to distraction-induced errors. Unlike the descriptive inference analysis in Section 4, these interventions modify the model's computation at inference time. For selected query-key pairs, we set the corresponding entry of the additive attention mask to $-\infty$ before the softmax. The masked query therefore cannot

Table 12: **Token groups**.

GROUP	CONTENTS
system / header	attention sink (position 0) + system prompt + user header
target image	visual tokens of the relevant image
distractors images	visual tokens of each distractor, in positional order
question	question stem + answer choices (content words only)
readout (answer)	the last position, where the answer is produced
separator	vision delimiters (between the image blocks), answer-format instruction and generation scaffold (after the last image)

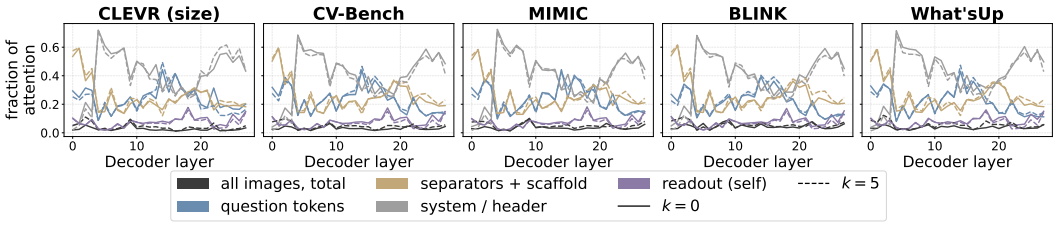

Figure 18: **Readout-token attention across decoder layers in Qwen2.5-VL-7B-Instruct.** Curves show the fraction of attention allocated to all image tokens (black), question tokens (blue), separators and prompt-scaffold tokens (gold), system and header tokens (gray), and the readout position itself (purple). Solid lines denote clean inputs (k=0), and dashed lines denote inputs with five distractor images (k=5). Results use n=150 examples per benchmark, except What'sUp (n=300).

attend to the selected key or receive information through its value vector. Model parameters and all other attention connections remain unchanged.

Table 13 summarizes the interventions evaluated in our experiments. **D** blocks the distractors' visual tokens from every position after the final image (question, answer-format instruction, generation scaffold and readout). This excludes the system header, positioned before the images, and the delimiters interleaved between earlier image blocks. **R** is the readout alone and **P** is every position that is not itself a distractor token, so $\mathbf{R} \subset \mathbf{D} \subset \mathbf{P}$. **MM** sits outside this chain: it prevents each *image* from reading what precedes it (ie. the system header and all earlier images) and it does so for every image, the relevant one included.

Table 13: **Knockout conditions: source and target spans.** Entries of the additive attention mask at (query = target, key = source) are set to $-\infty$ during prefill, within the designated layer band only. Because the text block follows every image, $\mathbf{R} \subset \mathbf{D} \subset \mathbf{P}$.

Cond.	Source	Target	Definition of the target span
P	IMG_D	everything	every position that is not itself a distractor token
MM	each image	all preceding tokens	per image, every position before its first token: the system prefix and all earlier images (MIMIC)
D	IMG_D	all text	every position after the final image: question stem, choices, answer-format instruction, generation scaffold, readout
R	IMG_D	readout	the last prompt position, $L-1$

Table 14 reports the middle-layer attention-knockout results separately for each benchmark. We apply each intervention in layers 9-17 of Qwen2.5-VL-7B-Instruct with five distractor images. A *broken* item is answered correctly in the corresponding clean condition but incorrectly when distractors are added; a *right* item remains correct despite the distractors. For each intervention, RECOV./BROKEN measures the fraction of broken items whose correct answer is restored, whereas BROKEN/RIGHT measures the fraction of previously correct items that the intervention causes the

Table 14: Attention knockouts with 95% bootstrap intervals (10,000 resamples over items, band `middle`). RECOV./BROKEN is the fraction of broken items the knockout fixes. BROKEN/RIGHT is the fraction of the already-correct population it damages.

BENCHMARK	COND.	n_{broken}	n_{right}	ROUTE	RECOV./BROKEN	BROKEN/RIGHT
CLEVR	P	1331	1181	$\text{IMG}_D \rightarrow$ everything	0.54 [0.51, 0.56]	0.07 [0.06, 0.09]
	MM	1331	1181	IMG $\rightarrow$ all preceding (MIMIC)	0.04 [0.03, 0.05]	0.86 [0.83, 0.87]
	D	1331	1181	$\text{IMG}_D \rightarrow$ all text	0.44 [0.41, 0.47]	0.08 [0.07, 0.10]
	R	1331	1181	$\text{IMG}_D \rightarrow$ readout	0.02 [0.02, 0.03]	0.04 [0.03, 0.05]
CV-Bench	P	439	1465	$\text{IMG}_D \rightarrow$ everything	0.62 [0.57, 0.66]	0.04 [0.03, 0.05]
	MM	439	1465	IMG $\rightarrow$ all preceding (MIMIC)	0.18 [0.14, 0.21]	0.32 [0.30, 0.35]
	D	439	1465	$\text{IMG}_D \rightarrow$ all text	0.54 [0.49, 0.59]	0.04 [0.03, 0.05]
	R	439	1465	$\text{IMG}_D \rightarrow$ readout	0.04 [0.02, 0.06]	0.01 [0.01, 0.02]
BLINK	P	90	233	$\text{IMG}_D \rightarrow$ everything	0.52 [0.42, 0.62]	0.04 [0.02, 0.07]
	MM	90	233	IMG $\rightarrow$ all preceding (MIMIC)	0.47 [0.37, 0.57]	0.44 [0.37, 0.50]
	D	90	233	$\text{IMG}_D \rightarrow$ all text	0.43 [0.33, 0.53]	0.04 [0.02, 0.06]
	R	90	233	$\text{IMG}_D \rightarrow$ readout	0.08 [0.02, 0.13]	0.00 [0.00, 0.01]
What'sUp (A+B)	P	339	417	$\text{IMG}_D \rightarrow$ everything	0.66 [0.61, 0.71]	0.04 [0.02, 0.06]
	MM	339	417	IMG $\rightarrow$ all preceding (MIMIC)	0.14 [0.10, 0.18]	0.50 [0.45, 0.55]
	D	339	417	$\text{IMG}_D \rightarrow$ all text	0.54 [0.48, 0.59]	0.04 [0.02, 0.06]
	R	339	417	$\text{IMG}_D \rightarrow$ readout	0.03 [0.01, 0.04]	0.01 [0.00, 0.02]
MIMIC	P	154	866	$\text{IMG}_D \rightarrow$ everything	0.46 [0.38, 0.54]	0.04 [0.03, 0.06]
	MM	154	866	IMG $\rightarrow$ all preceding (MIMIC)	0.10 [0.05, 0.15]	0.62 [0.59, 0.65]
	D	154	866	$\text{IMG}_D \rightarrow$ all text	0.45 [0.37, 0.53]	0.05 [0.03, 0.06]
	R	154	866	$\text{IMG}_D \rightarrow$ readout	0.05 [0.01, 0.08]	0.00 [0.00, 0.00]

Table 15: **Item-level overlap between knockout P and knockout D in the middle layers (9–17).** Percentages are of the broken population. **P** blocks distractor attention to every token group; **D** blocks it to text only. The final column is the fraction of D's repairs that P also repairs.

Benchmark	n	**P**	**D**	both	D-only	P∩D / D
CLEVR	1331	54%	44%	39%	4%	90%
CV-Bench	439	62%	54%	50%	4%	93%
BLINK	90	52%	43%	41%	2%	95%
What'sUp	339	66%	54%	52%	1%	97%
MIMIC	154	46%	45%	38%	6%	86%
Pooled	2353	56%	47%	43%	4%	92%

model to answer incorrectly. Confidence intervals are percentile 95% bootstrap intervals obtained from 10,000 item-level resamples within each benchmark.

Table 15 shows that **D** and **P** have strong overlap: 92% items repaired by **D** are also repaired by **P**. Table 16 reports recovery for every condition across the four layer bands.

A.6 Mitigating distraction through distractor invariance.

A.6.1 Experimental Setup

Model and adaptation. All fine-tuned models start from Qwen2.5-VL-7B-Instruct. We train LoRA adapters with rank 16, $\alpha = 32$, and dropout 0.05. The vision encoder and the vision-language merger are frozen, so every intervention we report acts on the language model's use of visual evidence rather than on the visual features themselves.

Optimization. We use AdamW at a learning rate of 2×10^{-4} with a cosine schedule. All arms share the same data order, and LoRA configuration, so differences between arms are attributable to the objective alone.

Table 16: **Knockout recovery by layer band.** Qwen2.5-VL at N=5 distractors, pooled over five benchmarks at the item level. RECOVERY (fraction of broken items repaired). Bands partition the 28 decoder layers; ALL masks every layer. The MIDDLE column reproduces Table 3. Intervals are 95% bootstrap CIs over items (10,000 resamples).

COND.	ROUTE	EARLY 0–8	MIDDLE 9–17	LATE 18–27	ALL 0–27
P	$\text{IMG}_D \rightarrow$ everything	0.34 [0.32, 0.36]	0.56 [0.54, 0.58]	0.07 [0.06, 0.08]	0.76 [0.75, 0.78]
MM	IMG $\rightarrow$ all preceding (MIMIC)	0.28 [0.26, 0.30]	0.10 [0.09, 0.11]	0.03 [0.02, 0.04]	0.11 [0.10, 0.12]
D	$\text{IMG}_D \rightarrow$ all text	0.19 [0.17, 0.21]	0.47 [0.45, 0.49]	0.07 [0.06, 0.08]	0.58 [0.56, 0.60]
R	$\text{IMG}_D \rightarrow$ readout	0.03 [0.03, 0.04]	0.03 [0.02, 0.04]	0.03 [0.03, 0.04]	0.03 [0.03, 0.04]

Validation, early stopping, and checkpoint selection. We hold out a fixed random subset of 1,000 validation questions (Appendix A.7) and evaluated at the end of each epoch on all five variants ($k \in 0, 1, 2, 4, 8$). Early stopping uses a patience of 2 epochs, and the exported checkpoint maximizes mean accuracy over the five variants.

A.7 TRAINING SET CONSTRUCTION

Source and scope. The mitigation experiments train on a distractor ladder built from FineVision (Wiedmann et al., 2025), deliberately drawn from a different domain than the perception benchmarks we evaluate on. Training rows come from thirteen FineVision sub-datasets grouped into five categories: general scene VQA (A-OKVQA), general VQA (Visual7W), science and diagram reasoning (AI2D, ScienceQA, ScienceQA without context), medical imaging (PMC-VQA), and math and geometry (GeoQA+, Geometry3K, Geo170K, InterGPS, UniGeo, MAVIS-metagen, MAVIS-rule). We keep only multiple-choice turns. The source pool includes 12,708 training questions and 1,000 validation questions, composed of A-OKVQA 33%, science and diagram reasoning 19%, PMC-VQA 19%, Visual7W 16%, and math and geometry 13%. None of the five evaluation benchmarks contributes to training. Hence, gains reported in Section 6 measure transfer of a distraction-robustness behavior rather than in-domain adaptation.

Distractor construction. Distractors are sampled from the same sub-dataset as the source row (e.g., geometry questions receive geometry distractors). Domain-matching prevents models from shortcutting the task using low-level visual differences, forcing them to resolve the question-image relation. Following the LoCoVQA pipeline (Sharma et al., 2024), we apply a two-stage VLM filter to each domain-matched candidate, which is more likely than a randomly sampled image to contain question-relevant content. The first stage extracts the entities present in the candidate image, and the second returns a binary verdict indicating whether any extracted entity is relevant to the target question. Failed candidates are resampled up to 15 times; rows that cannot be filled are dropped. Both filter prompts enforce a strict constraint on people: if the question involves a person, distractors must contain no people to avoid referential ambiguity.

A.8 CLEAN-ANCHORED ANSWER CONSISTENCY

Table 17 reports circular accuracy for every benchmark, distractor count, and training condition.

To examine whether clean-anchored answer consistency affects the attention pathways identified in our mechanistic analysis, we compare attention patterns under CE-only and CE+KL fine-tuning. We measure how question tokens allocate attention to target and distractor images across layers, focusing on the middle layers where distractor-to-text knockouts were most effective. This analysis tests whether regularizing answer distributions also changes attention allocation, even without imposing an explicit attention constraint.

Finally, we ask whether these gains persist when the relevant image is identified explicitly rather than inferred. We evaluate the KL-clean model (λ=1.0) under the ground-truth and two-step prompts against the untrained base model (Table 18). The training transfers: circular accuracy improves by 15.3-16.8 points at every distractor count under both prompts.

Table 17: **Circular accuracy by distractor count.** Results for Qwen2.5-VL-7B and its LoRA fine-tuning arms; k denotes the number of distractor images. Each block reports one benchmark, followed by an unweighted average across the five benchmarks. The final two columns give mean accuracy over distracted conditions ($k > 0$) and Δ_{base}, the difference between that mean and the base model's mean in the same block. Boldface marks the best value within each block and column. The main results are summarized in Table 4.

	Distractor images k					Summary	
Model	0	1	2	3	5	Mean k>0	Δ_{base}
CLEVR-size							
Qwen2.5-VL-7B *(base)*	70.7	60.1	54.9	48.3	36.4	49.9	0.0
CE only	82.5	79.9	78.5	77.9	75.4	77.9	+28.0
CE + KL λ=0.1	82.8	79.4	77.8	75.4	71.9	76.1	+26.2
CE + KL λ=0.5	79.8	77.9	75.9	74.2	72.0	75.0	+25.1
CE + KL λ=1.0	**86.7**	**84.7**	**82.8**	**81.4**	**79.5**	**82.1**	+32.2
CE + KL λ=2.0	83.4	80.3	80.0	77.5	75.7	78.4	+28.4
CE + MIMIC mask	83.6	81.4	79.9	80.0	76.1	79.4	+29.4
DTS *(no training)*	70.7	67.1	61.4	58.0	50.6	59.3	+9.3
CV-Bench							
Qwen2.5-VL-7B *(base)*	72.2	68.1	64.9	63.8	61.2	64.5	0.0
CE only	74.8	70.3	**69.8**	67.1	65.2	68.1	+3.6
CE + KL λ=0.1	74.0	70.8	68.3	66.3	64.1	67.4	+2.9
CE + KL λ=0.5	74.3	71.7	69.7	**67.7**	**66.4**	**68.9**	+4.4
CE + KL λ=1.0	74.1	70.0	67.6	65.2	64.9	66.9	+2.4
CE + KL λ=2.0	73.7	70.0	68.2	66.1	65.4	67.4	+2.9
CE + MIMIC mask	**75.4**	**71.9**	69.7	67.0	65.1	68.4	+3.9
DTS *(no training)*	72.3	67.7	64.1	63.2	61.1	64.0	-0.5
MIMIC							
Qwen2.5-VL-7B *(base)*	68.0	64.9	64.7	64.1	62.7	64.1	0.0
CE only	61.2	58.7	55.9	55.7	55.3	56.4	-7.7
CE + KL λ=0.1	66.5	62.9	61.3	60.9	57.5	60.6	-3.5
CE + KL λ=0.5	**68.3**	65.3	65.1	64.6	64.4	64.8	+0.7
CE + KL λ=1.0	67.9	**66.3**	**66.1**	**66.0**	**64.6**	**65.8**	+1.6
CE + KL λ=2.0	64.5	62.3	61.3	59.1	57.4	60.0	-4.1
CE + MIMIC mask	61.4	60.3	58.3	56.7	55.4	57.7	-6.4
DTS *(no training)*	67.3	63.2	60.9	59.7	57.5	60.3	-3.8
WhatsUp							
Qwen2.5-VL-7B *(base)*	**92.2**	81.5	76.2	72.0	52.7	70.6	0.0
CE only	85.9	80.5	79.3	76.8	71.6	77.0	+6.5
CE + KL λ=0.1	88.8	81.7	78.0	77.0	73.2	77.5	+6.9
CE + KL λ=0.5	90.9	**84.6**	**80.9**	**79.6**	**74.0**	**79.8**	+9.2
CE + KL λ=1.0	88.4	82.6	79.9	79.4	74.0	79.0	+8.4
CE + KL λ=2.0	86.1	81.3	79.4	78.8	73.0	78.1	+7.6
CE + MIMIC mask	88.3	82.8	76.8	73.4	66.7	74.9	+4.4
DTS *(no training)*	92.2	81.0	73.8	72.4	63.2	72.6	+2.0
BLINK							
Qwen2.5-VL-7B *(base)*	50.2	48.8	46.5	46.3	44.0	46.4	0.0
CE only	55.2	54.9	51.8	53.0	50.7	52.6	+6.2
CE + KL λ=0.1	55.7	53.7	53.8	51.0	52.1	52.6	+6.2
CE + KL λ=0.5	57.7	**55.7**	**55.4**	**54.0**	53.0	**54.5**	+8.1
CE + KL λ=1.0	56.6	54.1	54.1	53.3	**53.8**	53.8	+7.4
CE + KL λ=2.0	**59.3**	55.2	53.5	53.7	52.9	53.8	+7.4
CE + MIMIC mask	54.6	54.9	52.9	49.9	49.9	51.9	+5.5
DTS *(no training)*	50.4	48.2	46.2	45.6	45.7	46.4	+0.0
Pooled (5 benchmarks)							
Qwen2.5-VL-7B *(base)*	70.7	64.7	61.5	58.9	51.4	59.1	0.0
CE only	71.9	68.9	67.1	66.1	63.6	66.4	+7.3
CE + KL λ=0.1	73.6	69.7	67.9	66.1	63.7	66.8	+7.7
CE + KL λ=0.5	74.2	71.0	69.4	68.0	66.0	68.6	+9.5
CE + KL λ=1.0	**74.7**	**71.5**	**70.1**	**69.1**	**67.4**	**69.5**	+10.4
CE + KL λ=2.0	73.4	69.8	68.5	67.0	64.9	67.6	+8.4
CE + MIMIC mask	72.7	70.3	67.5	65.4	62.7	66.5	+7.4
DTS *(no training)*	70.6	65.4	61.3	59.8	55.6	60.5	+1.4

Table 18: **Circular accuracy (%) of the KL-clean model (λ=1.0) against the untrained base model**, under the regular, ground truth and two-step prompts. Accuracy is macro-averaged over five benchmark groups weighted equally, over 9,155 items per cell (9,154 at k=5). Random distractors, `max_pixels` = 1,003,520. The relevance prompts have no k=0 point: selecting among zero distractors is not a task.

Prompt	**Model**	k=0	k=1	k=2	k=3	k=5
Regular	base	70.7	64.7	61.5	58.9	51.4
	KL-clean	74.7	71.5	70.1	69.1	67.4
	Δ	+4.1	+6.9	+8.7	+10.2	+16.0
Ground Truth	base	—	52.6	51.0	49.9	45.9
	KL-clean	—	68.0	66.6	65.2	62.2
	Δ	—	+15.4	+15.6	+15.3	+16.3
Two-Step	base	—	51.9	49.8	48.9	44.4
	KL-clean	—	67.3	65.7	64.4	61.2
	Δ	—	+15.4	+15.8	+15.5	+16.8