

Lost at the End: Primacy Bias in Multimodal Retrieval-Augmented Question Answering

Jieyuan Liu¹ Jianyang Gu² Shijie Chen² Jefferson Chen¹ Zhen Wang^{1*}

¹University of California, San Diego ²The Ohio State University

{jil029, zhw085}@ucsd.edu

Abstract

Knowledge-based visual question answering (KB-VQA) lets vision-language systems answer questions that exceed their parametric knowledge by conditioning a reader on passages retrieved from a Wikipedia-derived knowledge base. In pure-text long-context LLMs, retrieved-context use follows the U-shaped *lost-in-the-middle* effect of Liu et al. (2024): information at the start and end of context is used, the middle is lost. Whether this transfers to deployed multimodal KB-VQA is open. To close this gap, we design the first controlled probe of reader-side position dependence in multimodal KB-VQA: a gold-position protocol in which only the gold passage’s prompt slot varies within question. We run it on three open-source 7B/8B VLM readers and two KB-VQA benchmarks with up to 20 retrieved passages. The shape flips from U to primacy: gold-at-first beats gold-at-last by 16 to 26 points on all six combinations of reader and benchmark, an effect we call *Lost at the End*; the gap holds at every scale we test, 3B to 32B, attenuating at 32B. Three targeted ablations narrow the cause. A text-only control that removes the image and changes nothing else shows the primacy is already present in text mode and does not depend on the image. Image-position and distractor-shuffle ablations trace the effect to prompt slot 0 of the instruction-tuned reader, where a second answer-bearing passage placed later is largely wasted. On a frozen reader, three retrieval-side fixes (MMR, oracle reranking, rank-based reordering) all fail to improve on the deployment default. Our findings indicate that $\text{recall}@k$ is the wrong metric for deployed KB-VQA and that the remaining headroom sits on the reader side; we release our protocol as a controlled instrument for evaluating reader-side interventions.

* Corresponding author.

1 Introduction

Knowledge-based visual question answering (KB-VQA) asks fine-grained questions about entities in an image that no vision-language model reliably memorizes (Lin et al., 2024; Yan and Xie, 2024; Caffagni et al., 2024; Cocchi et al., 2025). Systems therefore retrieve supporting passages and pass them to a frozen reader. When the reader is fed multiple competing passages, whether the system returns the answer depends not only on whether retrieval surfaces it but on where that passage sits in the reader’s context: a passage that helps at the start of the prompt can be invisible at the end. In a deployed pipeline routing dozens of candidates per query through a frozen reader, this position dependence drives system reliability, yet the choices that optimize retrieval — $\text{recall}@k$ targets, larger candidate pools — implicitly assume the reader can use whatever retrieval surfaces. That assumption has not been characterized for multimodal KB-VQA. Text-only long-context QA has a well-tested answer: a U-shaped curve in which the middle of the context is underused (Liu et al., 2024).

Existing multimodal-RAG position-bias work has not characterized this setting. Yao et al. (2025) report a U-shaped Position Sensitivity Index of 2 to 11 pp. Their benchmarks span text-only, image-only and mixed-modality comprehension; ours is KB-VQA, where the distractors are the wrong passages a retriever ranked most relevant. Hu et al. (2025) observe qualitatively that LVLMs prioritize early candidates on the benchmarks we use, and route around it with a top-1-only pipeline. Without controlled measurement at deployed scale, three basic questions are unanswered: is the U-shape the right model for multimodal RAG, or is the shape different? What drives the position effect, given that image proximity, retrieval similarity and prompt position are confounded in a standard pipeline? And do the obvious retrieval-side re-

sponses recover the loss?

To answer these questions, we design a gold-position protocol¹ (§3) that holds everything bit-identical within question except the gold passage’s prompt slot. Any accuracy difference must therefore come from position, and the prompts pair question-by-question for an exact bootstrap. We run it at the scale a deployed pipeline produces (top-50 PreFLMR ViT-G (Lin et al., 2024) distractors, k up to 20, frozen instruction-tuned readers with no test-time fine-tuning) on three open-source 7B/8B VLM readers (Qwen2.5-VL-7B (Bai et al., 2025b), InternVL3-8B (Zhu et al., 2025), Qwen3-VL-8B (Bai et al., 2025a)) and two KB-VQA benchmarks (INFOSEEK (Chen et al., 2023), E-VQA (Mensink et al., 2023)).

Running this protocol yields four findings. *The shape is primacy, not U* (§4.1). We use *primacy bias* for the reader’s systematic preference for evidence early in its prompt over otherwise-identical evidence at later positions: gold-at-first beats gold-at-last by 16 to 26 pp on all six reader-by-benchmark *cells* — three readers times two benchmarks — with no recency rebound on five of six, and the gap holds across an order of magnitude of reader scale, from 3B to 32B parameters. Three ablations then locate the cause. *The effect is a text-side property of the reader* (§4.2): removing the image and changing nothing else does not shrink the gap, so primacy is not introduced by multimodal conditioning. *Primacy is anchored at prompt slot 0 of the instruction-tuned reader* (§4.3): image-position and distractor-shuffle ablations rule out image-token proximity and retrieval similarity as primary drivers, and a multi-gold experiment shows that a second answer-bearing passage placed later contributes little once slot 0 already holds one. *The obvious retrieval-side fixes do not help* (§4.4): MMR diversification, oracle reranking, and rank-based distractor reordering all fail to improve on the deployment default. Together these answers indicate that recall@ k is the wrong metric for deployed KB-VQA and that the remaining headroom sits on the reader side; we discuss what this implies for the design space of multimodal RAG in §4.5.

2 Related Work

Position bias in text LLMs. The *lost-in-the-middle* U-shape of Liu et al. (2024) established that

¹Code and protocol: https://github.com/WeMWish/lost_at_the_end_code

position dependence is a property of long-context reading, with follow-ups attributing it to attention-bias asymmetry (Hsieh et al., 2024), showing direction varies by model family (Hutter et al., 2025), and reporting that Qwen-2.5 7B exhibits stronger positional bias than the Llama-3 family in pure-text RAG (Cuconasu et al., 2025). Our pattern is primacy-dominant, contrasting with Liu et al. (2024)’s U-shape at 13B+ and reported recency at 7B; we are partially consistent with their recency component on one of six cells (Qwen3-VL-8B $\times$ E-VQA; §4.1), with no recency rebound on the other five. None of these works extend to multimodal readers, where image conditioning adds an axis of confound and KB-scale retrieval changes the distractor distribution; whether the U-shape transfers, persists in a different shape, or fails entirely in multimodal KB-VQA is the question we answer.

Multimodal position bias. Several characterizations exist for the multimodal setting, none in deployed KB-VQA. Tan et al. (2024) and Tian et al. (2025) report order sensitivity in multimodal LLMs but outside KB-grounded reading; the text-side ancestor is Lu et al. (2022), who found that reordering the examples in a prompt can move accuracy from near-chance to near state-of-the-art. Yao et al. (2025) report a U-shaped Position Sensitivity Index of 2 to 11 pp across MS-MARCO, ChartQA, and VEGA at 2 to 19 distractors, a text-only, an image-only, and a mixed-modality benchmark; we differ in setting (deployed KB-VQA), distractor configuration (9 deterministic top-50 distractors at controlled positions), and statistical methodology (paired-bootstrap confidence intervals, not a single point estimate). Hu et al. (2025) note qualitatively that LVLMS prioritize early candidates on the same KB-VQA benchmarks we use and motivate a top-1-only generation pipeline; we differ in design intent, measuring the effect and locating its mechanism instead of routing around it. Wu and Albreiki (2026) concurrently characterize position bias on the retriever side, a separate question from the reader-side effect we study. Closest in spirit is Wang et al. (2024), whose MM-NIAH benchmark inserts synthetic text and image content into long real interleaved documents and tests whether the model can find it, reporting that a lost-in-the-middle pattern is present in MLLMs. We differ in what supplies the context and in what varies. They insert synthetic content into real documents and vary its depth and the document’s length. Ours is the actual top-50 output of a strong KB-VQA

retriever, and we vary only the gold passage’s position, with distractor identity and ordering held bit-identical within question. The two are complementary, and their setting is a natural next place to run our protocol (Limitations).

3 Probing Protocol

The probe builds on a standard frozen KB-VQA pipeline (Figure 1): a multi-vector retriever produces top-50 candidate passages from a fixed knowledge base, and a vision-language reader generates an answer from the image, the question, and a k -passage subset of those candidates. No component is fine-tuned. Three commitments shape the design choices below: the reader’s response to gold position must be isolated from retrieval rank, distractor identity, image conditioning and scoring noise; every effect must hold at deployment scale; and every measurement must be cross-replicable across reader families, benchmarks, and the scoring matchers a deployer might apply.

Deployment-scale instantiation. We evaluate on the two canonical knowledge-based VQA benchmarks: INFOSEEK (Chen et al., 2023), with a 98K-passage knowledge base, and E-VQA (Mensink et al., 2023), with a 51K-passage knowledge base, both drawn from the M2KR release (Lin et al., 2024). From each we draw a seed-locked subset of 500 question ids; INFOSEEK contributes 496 after excluding four questions with too few distractors, E-VQA contributes 500. These knowledge bases are Wikipedia-derived and contain tens of thousands of passages; we scope our claims to that regime and do not extrapolate to corpora orders of magnitude larger. Retrieval uses PreFLMR ViT-G (Lin et al., 2024), a multi-vector ColBERT-style retriever (Khattab and Zaharia, 2020) that reports state-of-the-art results on M2KR; we take its top-50 candidates per query. The three readers are Qwen2.5-VL-7B (Bai et al., 2025b), InternVL3-8B (Zhu et al., 2025), and Qwen3-VL-8B (Bai et al., 2025a), run frozen and decoded greedily with a short-answer length limit, all sharing the same fixed system prompt and numbered passage block (App. U). Precision, library versions and compute are documented in App. V.

Within-question position protocol. The probe itself is a within-question paired design. For each question we construct three k -passage prompts that are bit-identical except for the index of the gold passage: it occupies slot 0 in the *first* cell, slot

$\lfloor (k-1)/2 \rfloor$ in the *middle* cell, and slot $k-1$ in the *last* cell. The other $k-1$ slots are filled with distractors that are identical in identity and ordering across the three cells per question (App. S). Each question therefore yields an exactly matched set of first, middle, and last results, which is what the paired bootstrap resamples. The gold passage is the highest-ranked positive in the PreFLMR ViT-G top-50; when no top-50 candidate is positive, we fall back to the dataset’s first-listed positive for that question (the case for 142 of the 496 INFOSEEK questions; we report a fallback-gold sensitivity analysis in App. E). The distractor pool is the remaining top-50 non-positives: the hardest, topically near-gold negatives a strong retriever would surface in deployment.

Scoring and statistics. For INFOSEEK we port the evaluation code released alongside Chen et al. (2023) by its first author, which the benchmark repository links but labels community-contributed. E-VQA does have a specified metric: Mensink et al. (2023) define evaluation in terms of BEM (Bullian et al., 2022), a learned answer-equivalence model, and the M2KR downstream setup of Lin et al. (2024) adopts the same choice. What the M2KR release we consume does *not* ship is a scorer implementation or a BEM checkpoint. We therefore report three lexical matchers (normalized substring, word-boundary, exact equality), treat substring as canonical, and give sensitivity in App. D; both E-VQA verdicts we re-score hold under all three. Lexical matching is stricter than a learned equivalence model, so our gaps are conservative: relaxing INFOSEEK scoring to lenient containment *widens* the gap (App. D). Our text-only NQ-Open control runs (App. M, App. O) use Liu et al.’s released scoring code on a seed-locked sample of 500 of their 2,655 paragraph-answer queries. Per-cell accuracy uses Wilson 95% intervals; between-cell deltas use a paired bootstrap ($B=10,000$, fixed seed), and a delta is *separable* when its interval excludes zero. Effect-size verdicts come from a five-way classifier whose numeric bands were fixed in a pre-registration note before each experiment it covers (App. C). Our subsets match the full released splits on all gated covariates (App. B).

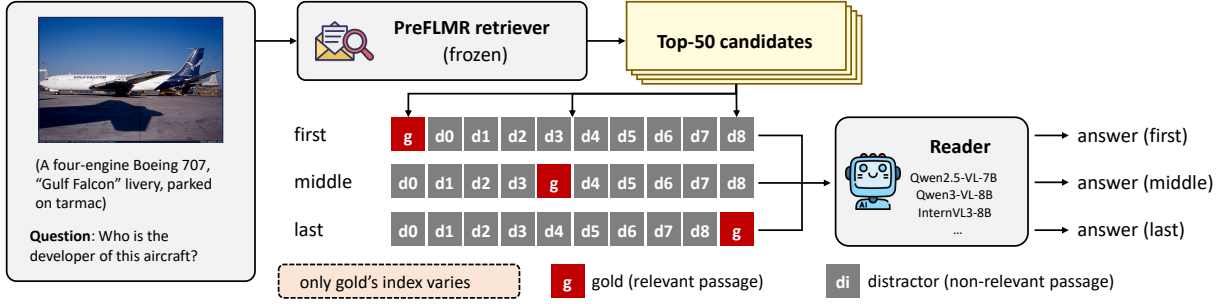


Figure 1: **Pipeline and position-permutation protocol.** The image and question are encoded by a frozen PreFLMR ViT-G retriever, which returns 50 candidate passages. From these we construct three 10-passage prompts that are bit-identical except for the index of the gold passage g (red): *first* (index 0), *middle* (index 4), *last* (index 9). Nine distractors $d_0, \dots, d_8$ (grey) are identical across cells per question. A frozen reader (Qwen2.5-VL-7B, InternVL3-8B, or Qwen3-VL-8B) generates one answer per cell.

4 Experiments

4.1 Primacy, Not U-Shape

Finding 1: Primacy, not U-shape

Gold-at-first beats gold-at-last by 16 to 26 pp on all six combinations of reader and benchmark; middle and last sit close together, far below first. The shape is primacy-dominant, not the U reported for text-only long-context LLMs.

We first establish the effect on Qwen2.5-VL-7B reading top-10 INFOSEEK passages, varying only the gold passage’s slot among *first*, *middle*, and *last* (Figure 1). Accuracy drops sharply with gold position: 45.2% at first, 32.3% at middle, 28.6% at last ($n=496$; Wilson CIs as error bars in Figure 2). The first-vs-last gap of +16.5 pp serves as the comparator anchor below. Restricting to the 354 questions with a gold in the PreFLMR ViT-G top-50 raises absolute accuracy by 2.9 to 4.1 pp across the three $k=10$ cells (and by at most 5.2 pp across the full k -sweep) but preserves the shape (App. E).

Shape across k on all six cells. Figure 2 sweeps $k \in \{1, 3, 5, 10, 20\}$ on all six reader-by-benchmark combinations. At $k=1$ all three conditions are the same prompt (50–72%; Table 7): even with gold alone the reader is the binding constraint. As k grows every panel separates the same way — *first* decays gracefully while *middle* and *last* drop steeply and collapse together below it. We state the pattern as “first $\gg$ middle $\approx$ last” rather than the stricter “first $>$ middle $>$ last” because the middle-vs-last ordering is reader- and dataset-specific and can flip within CI noise (App. L, App. N). The first-vs-last gap stays large and same-signed on all six. *No recency rebound* appears at $k=20$ on five of six cells; the sixth is the exception discussed next.

The one exception: a sign reversal on Qwen3-VL-8B $\times$ E-VQA. On this single panel, last sits above middle — by 3.0 pp at $k=10$ and 5.2 pp at $k=20$, both separable (App. N). The reader is not simply recency-biased: in text-only NQ-Open it shows the *opposite* tail (middle $>$ last by 3.6 pp, separably; App. O). We do not commit to a mechanism; candidates include Qwen3-VL-8B’s architecture (DeepStack encoder, interleaved-MRoPE, longer pretraining context) and E-VQA’s longer answer surface. The first-vs-last gap here remains large (+18.2 pp at $k=10$ and +20.4 pp at $k=20$), so the reversal sits on top of the primacy story rather than in place of it, and it is partially consistent with the recency component Liu et al. (2024) report at text-LLM scale.

Cross-cell generalization. Table 2 reports the first/middle/last protocol at $k=10$ on all six reader-by-benchmark combinations. Every first-vs-last delta is large, same-signed, and separable from zero, ranging from +16.1 pp (Qwen3-VL-8B INFOSEEK) to +26.4 pp (Qwen2.5-VL-7B E-VQA); all six fall within a factor of two of the anchor (App. C). The two largest are Qwen2.5-VL-7B and InternVL3-8B on E-VQA (1.6 and 1.4 $\times$). Exact-match scoring brings both within 3 pp of the anchor (App. D), so E-VQA’s free-form answers widen them through our substring matcher.

Scale: the gap holds from 3B to 32B, but attenuates at 32B. Reader capacity is the obvious alternative explanation — a stronger reader might simply use late evidence better. We hold the protocol fixed and swap only the reader, adding Qwen2.5-VL-3B and Qwen2.5-VL-32B on INFOSEEK at $k=10$ (Table 1). The gap is separable at every scale but does not shrink steadily with it: +14.5 pp at 3B, +16.5 pp at 7B, +10.5 pp at 32B. The 32B gap

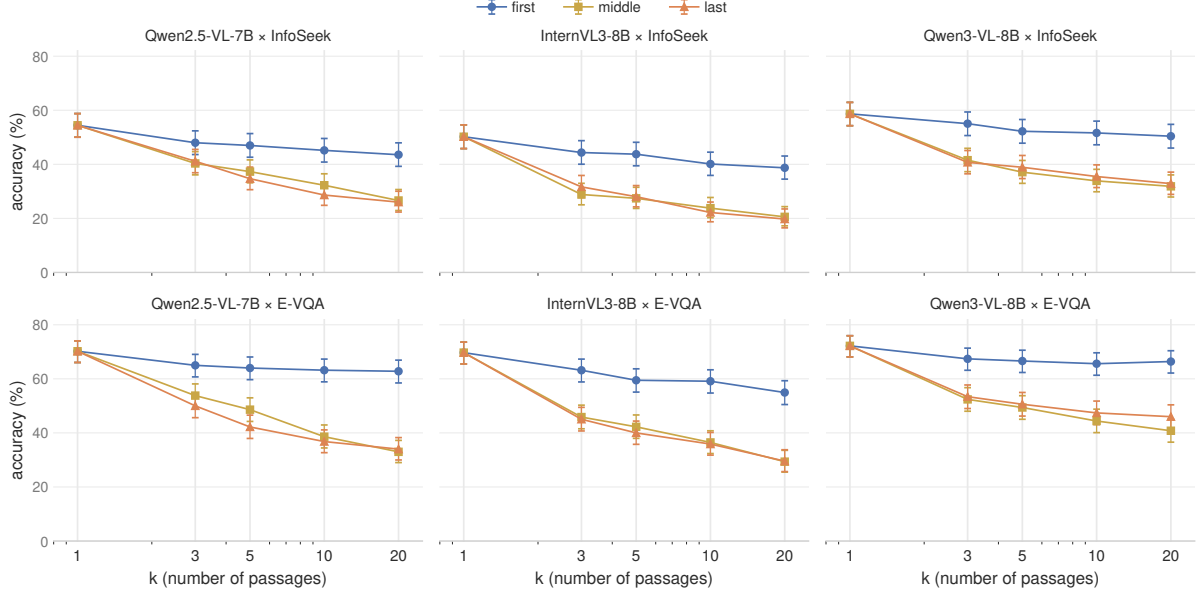


Figure 2: **Position effect across k on six reader $\times$ benchmark cells.** Top row: INFOSEEK (Qwen2.5-VL-7B, InternVL3-8B, Qwen3-VL-8B, left to right); bottom row: E-VQA (same column ordering). *First* decays gracefully on all six panels; *middle* and *last* drop steeply and collapse together below it. The middle-vs-last ordering on Qwen3-VL-8B E-VQA separably inverts at higher k (last > middle by 5.2 pp at $k=20$, paired-bootstrap CI [+2.6, +8.0]); the other five panels show no recency rebound at $k=20$. Error bars are Wilson 95% CIs.

sits separably below the 7B anchor (paired gap-of-gaps -6.05 pp [$-10.28, -2.01$]), whereas 3B-vs-7B is not separable (-2.02 pp [$-6.25, +2.22$]), so we claim attenuation only at 32B. The 32B reader is more accurate at every position, but it gains $+8.3$ and $+8.7$ pp at middle and last against only $+2.6$ pp at first: scale helps the later slots more than the first, narrowing the gap without closing it. The failure signature of §4.3 is unchanged: distractor-lift errors (answers taken from a distractor rather than the gold) still grow with gold position ($20.8 \rightarrow 37.6 \rightarrow 41.8\%$) and the chosen distractor still sits at median PreFLMR ViT-G rank 1–2. Scale is therefore a partial and expensive mitigation, not a fix. The 32B row is INFOSEEK-only: the matching E-VQA run reached 77.4 GB of peak allocation against the 75 GB ceiling we set on an 80 GB H100 and was aborted after 50 of 1,500 prompts. We declined to quantize or reduce image resolution, since either would introduce a precision confound absent from every other cell, so we report the 32B result on one benchmark.

Robustness. The re-scored E-VQA verdicts hold under the word-boundary and strict-equality matchers; excluding fallback-gold rows shifts deltas by at most 1.83 pp; multi-answer scoring flips zero rows. Vanilla PreFLMR ViT-G top-10 (no gold injection) gives 43.0% E-VQA / 29.8% INFOSEEK

Table 1: **Reader scale on INFOSEEK at $k=10$ ($n=496$).** Protocol identical across rows; only the reader changes. Accuracies in %, Δ in unrounded pp with paired-bootstrap 95% CIs. Gap-of-gaps against the 7B anchor is -2.02 ($-6.25, +2.22$) at 3B (not separable) and -6.05 ($-10.28, -2.01$) at 32B (separable).

Reader	first	middle	last	$\Delta(\text{first} - \text{last})$
Qwen2.5-VL-3B	40.7	26.6	26.2	$+14.5$ ($+10.9, +18.1$)
Qwen2.5-VL-7B (anchor)	45.2	32.3	28.6	$+16.5$ ($+12.7, +20.6$)
Qwen2.5-VL-32B	47.8	40.5	37.3	$+10.5$ ($+7.5, +13.9$)

for Qwen2.5-VL-7B, in range of published M2KR systems.

Insight 1: Distribute training answers across positions

VLM instruction-tuning corpora should include multi-passages retrieval examples with gold answers spread across all positions, not concentrated in the single-passage or top-1 format that dominates current recipes (Liu et al., 2023; Dai et al., 2023).

Table 2: **Cross-benchmark generalization of the first-vs-last position effect at $k=10$.** Paired-bootstrap 95% CIs use 10,000 resamples and a fixed seed; all six deltas fall in the same-magnitude band against the anchor. Per-cell accuracies are rounded to 1 dp; Δ is computed from unrounded values and may differ from displayed-value subtraction by up to 0.1 pp.

Reader	Benchmark	first (%)	last (%)	Δ (pp)	95% CI	ratio
Qwen2.5-VL-7B	INFOSEEK ($n=496$)	45.2	28.6	+16.5	(+12.7, +20.6)	1.00 $\times$ (anchor)
Qwen2.5-VL-7B	E-VQA ($n=500$)	63.2	36.8	+26.4	(+22.2, +30.6)	1.60 $\times$
InternVL3-8B	INFOSEEK ($n=496$)	40.1	22.2	+17.9	(+14.1, +22.0)	1.09 $\times$
InternVL3-8B	E-VQA ($n=499$)	59.1	35.9	+23.2	(+19.0, +27.5)	1.41 $\times$
Qwen3-VL-8B	INFOSEEK ($n=496$)	51.6	35.5	+16.1	(+12.5, +20.0)	0.98 $\times$
Qwen3-VL-8B	E-VQA ($n=500$)	65.6	47.4	+18.2	(+14.2, +22.2)	1.10 $\times$

4.2 Primacy Is a Text-Side Property

Finding 2: Removing the image does not shrink the gap

On matched KB-VQA data with only the image removed, the first-vs-last gap is at least as large as in the multimodal setting on all three reader-by-benchmark cells tested. Primacy is a property of the reader’s text-position prior; multimodal conditioning does not introduce it.

How much of the multimodal effect is inherited from the reader’s underlying text-mode behavior? A text-only ablation on a *different* corpus cannot answer this: it varies corpus and modality together, leaving the difference unattributable to either. We therefore hold the corpus fixed and vary only whether the reader receives the image. A cross-corpus amplification claim made in an earlier version is withdrawn for this reason (Limitations). **Matched-corpus control.** We re-run three reader-by-benchmark cells holding everything fixed except whether the reader receives the image: same M2KR questions, same gold passage, same distractors in the same slots, same decoding, same scorer. Prompts are bit-identical to the multimodal runs minus the image block, and no retrieval is repeated. Removing the image does not shrink the first-vs-last gap on any cell; the point estimate grows on all three (Table 3).

The direction is consistent; the interaction separates on one cell of three. We test the difference of gaps directly, resampling questions jointly across all four conditions. It is separable only for Qwen2.5-VL-7B $\times$ INFOSEEK (+6.05 pp [+1.81, +10.28]); Qwen2.5-VL-7B $\times$ E-VQA (+3.40 [−0.80, +7.60]) and InternVL3-8B $\times$ INFOSEEK (+1.21 [−3.02, +5.44]) straddle zero. This supports the claim that primacy does not depend on the image. The stronger claim that the image *dampens* primacy would require the interac-

Table 3: **Matched-corpus text-only control.** Same questions, same gold passage, same distractors in the same slots, same decoding and same scorer; the only change is whether the reader receives the image. Per-cell accuracies in %; Δ in unrounded pp. Removing the image does not shrink the gap on any cell. The paired-bootstrap difference between the two gaps, computed over the shared question set, is +6.05 (+1.81, +10.28) for Qwen2.5-VL-7B $\times$ INFOSEEK, +3.40 (−0.80, +7.60) for Qwen2.5-VL-7B $\times$ E-VQA, and +1.21 (−3.02, +5.44) for InternVL3-8B $\times$ INFOSEEK: same-signed on all three cells, separable on one.

Reader $\times$ Benchmark	Image	first	middle	last	Δ (first–last)
Qwen2.5-VL-7B $\times$ INFOSEEK ($n=496$)	yes	45.2	32.3	28.6	+16.5
	no	41.1	17.1	18.5	+22.6 (+18.5, +26.6)
Qwen2.5-VL-7B $\times$ E-VQA ($n=500$)	yes	63.2	38.6	36.8	+26.4
	no	66.0	36.6	36.2	+29.8 (+25.4, +34.2)
InternVL3-8B $\times$ INFOSEEK ($n=496$)	yes	40.1	23.8	22.2	+17.9
	no	37.1	17.1	17.9	+19.2 (+15.5, +23.0)

tion to separate on more than one cell.

Where the image acts. On Qwen2.5-VL-7B $\times$ INFOSEEK the cost of removing the image falls very unevenly across positions — *middle* loses 15.12 pp and *last* 10.08 pp, against only 4.03 pp at *first* — and the image-at-end ablation of §4.3 breaks a near-subset of the very same questions, suggesting the two manipulations are graded doses of one mechanism in which visual grounding partly protects evidence outside slot 0. This is an INFOSEEK-specific hypothesis, not an established cause: on E-VQA every per-position difference straddles zero. Full per-position intervals and the row-level overlap statistics are in App. H.

Primacy is present in a different corpus too. All three readers also show first-vs-last primacy on the NQ-Open text-only protocol (+7.4, +5.2, +6.4 pp; App. M, App. O), evidence that the effect is not an artifact of one corpus.

Scope and reader-family direction. Primacy is not a multimodal-RAG artifact in these readers. Position-bias *direction* at the 7B–8B scale is nonetheless reader-family-specific: Liu et al. (2024) report Llama-2-7B as solely recency-biased on the same NQ protocol, while all three readers we test are primacy-biased, and Cuconasu et al. (2025) independently report that Qwen-2.5 7B exhibits strong positional bias in text RAG. We therefore scope the text-side claim to the open-source instruction-tuned VLM families we tested; whether it holds for Llama-3-based VLMs or proprietary models is open (Limitations).

Insight 2: Mitigate on the text side, not the vision side

The gap survives with no visual input and anchors on a text slot, so mitigation should target the reader’s text-position prior, not the multi-modal interface. Position-aware fine-tuning, data rebalancing and attention calibration developed against text-LLM primacy (Hsieh et al., 2024; Xiao et al., 2024) are the direct candidates to transfer.

4.3 Primacy Is Anchored at Prompt Slot 0

Finding 3: Anchored at prompt slot 0

Primacy is anchored at prompt slot 0 of the instruction-tuned reader. Two ablations rule out image-token proximity and retrieval similarity as primary drivers, and a multi-gold test shows evidence placed after slot 0 is largely wasted. The underlying mechanism (Su et al., 2024; Press et al., 2022; Xiao et al., 2024) remains open (Limitations).

What is the reader doing wrong at later gold positions, and which axis drives it? A typical INFOSEEK case: asked who operates Dresden Hauptbahnhof with gold at slot 9, the reader returns a Munich U-Bahn operator, lifted from the PreFLMR ViT-G rank-1 distractor about Munich (further samples ship with the code; App. T). The pattern is general — when the reader is wrong, its answer is most often substring-present in whichever passage sits at slot 0 or 1.

Failure-bucket diagnostic. We bucket every error in the anchor cells by string-relationship to the prompted passages: *extraction-failed* (prediction in gold only), *distractor-lift* (in a distractor but not gold), *ambiguous* (both), *hallucinated* (neither). Distractor-lift grows monotonically with gold position ($30.1 \rightarrow 53.3 \rightarrow 56.5\%$) while extraction-failed drops ($36.4 \rightarrow 17.9 \rightarrow 13.8\%$), the other two buckets staying roughly constant (App. F): the modal failure flips from “did not use the right passage” to “used a wrong passage instead.” Among distractor-lift rows the chosen distractor’s median PreFLMR ViT-G rank is ≤ 1 in every cell, which under rank-order placement means the first or second distractor slot. Two readings fit: a *position* story (the reader locks onto slot 0) and a *similarity* story (it locks onto rank-1 wherever placed). The next two ablations rule out proximity and similarity in turn.

Image-at-end ablation rejects proximity as pri-

Table 4: **Distractor-shuffle cross-tabulation.** In dissociable cells B+C, slot-0 selection runs well above the $1/9$ chance baseline while rank-1 selection sits at chance. Mechanism replicates on E-VQA.

Cell	rank-1 at	n (dl)	slot 0	rank 1
<i>Qwen2.5-VL-7B</i> $\times$ INFOSEEK ($n=496$)				
A (rank-order)	slot 0	199	0.558	0.558
B (reverse)	slot 8	190	0.384	0.037
C (shuffle)	random	208	0.466	0.188
B+C mean			0.425 ($3.83\times$)	0.112 ($1.01\times$)
<i>Qwen2.5-VL-7B</i> $\times$ E-VQA ($n=500$)				
A (rank-order)	slot 0	213	0.601	0.601
B (reverse)	slot 8	217	0.521	0.055
C (shuffle)	random	197	0.462	0.168
B+C mean			0.491 ($4.42\times$)	0.111 ($1.00\times$)

mary. If image-token proximity to early passages drove primacy, moving the image after the passage block should shrink or invert the gap. Pre-registered rule: a delta at or above $0.75\times$ the anchor rejects proximity. Result on the anchor cell ($n=496$): first-vs-last $+18.55$ pp, $1.12\times$ the anchor. Proximity is therefore not the primary driver. A post-hoc middle-vs-last sign flip does emerge — $\Delta(\text{middle} - \text{last})$ inverts from $+3.6$ to -3.6 pp, separately (App. I) — so proximity plays a secondary modulating role on adjacent slots.

Distractor-shuffle ablation rejects similarity. The position- and similarity-anchoring readings are confounded in the default protocol because the rank-1 distractor sits at slot 0. We dissociate them with a controlled shuffle on the gold-at-last cell ($n=496$, INFOSEEK, Qwen2.5-VL-7B): *Cell A* (rank-order baseline; rank-1 at slot 0), *Cell B* (reverse-rank; rank-1 at slot 8, rank-9 at slot 0), *Cell C* (per-question shuffle; rank-1 at a random slot). Table 4 and Figure 5 cross-tabulate distractor-lift errors by (i) which slot’s passage was chosen and (ii) that passage’s PreFLMR rank. In the dissociable cells B and C, slot-0 selection runs at $3.83\times$ chance ($1/9$) while rank-1 selection runs at $1.01\times$ (App. J). Cell B is decisive: with the least-similar (rank-9) distractor at slot 0, the reader picks it 73 times; with the most-similar (rank-1) distractor at slot 8, it picks just 7. The reader follows the slot, not the similarity.

The mechanism replicates on E-VQA (Table 4, lower panel): slot-0 selection runs at $4.42\times$ baseline while rank-1 sits at exactly $1.00\times$. On neither benchmark does any shuffle variant separately improve on the deployment default (App. P), so this is a mechanism result, not a recommended intervention.

Slot-0 privilege, not decay toward the end. “Lost

at the end” and “slot 0 is privileged” predict the same first-vs-last gap but differ on what a *second* correct passage is worth. We separate them with a pre-registered multi-gold experiment on the 240 INFOSEEK questions having at least two top-50 positives that both contain the answer (rules and thresholds fixed before execution; App. A). Six cells place golds at slots $\{0\}$, $\{4\}$, $\{9\}$, $\{0, 9\}$, $\{4, 9\}$, $\{0, 4\}$, with the remaining slots filled from the same ranked distractor list; since $k=10$ is fixed, a two-gold cell holds eight of the nine distractors used in the one-gold cells (Table 5). A second gold at slot 9 is worth +2.92 pp $[-0.42, +6.25]$ when slot 0 already holds one (C4–C1) — not separable — and +4.17 pp $[+1.25, +7.50]$ when the first gold sits at slot 4 instead (C5–C2). The pre-registered contrast C5–C3 gives +7.08 pp $[+3.33, +10.83]$, which also promotes the first gold from slot 9 to slot 4. One gold at slot 0 also beats two at $\{4, 9\}$ by +9.17 pp $[+4.17, +14.58]$. The registered rule ordered C4–C1 below C5–C3, which holds and is robust (McNemar $p=0.14$ vs $p=0.0002$); a post-hoc test of their *difference* gives +4.17 pp $[+0.42, +8.33]$, which we flag as fragile rather than headline, since $p \approx 0.04$, its lower bound sits one question above zero, and it fails a two-test Bonferroni adjustment. We report the null as a bound, not a zero: with a minimum detectable effect of 4.83 pp at $n=240$, a later redundant gold contributes under 6.25 pp against a +16.25 pp gap on the same questions. Failure buckets agree — a gold at slot 0 cuts distractor-lift from 85 errors to 22 — and of the 129 correct answers under golds at $\{0, 9\}$, only 6 are uniquely attributable to the slot-9 passage. We therefore adopt the slot-0 framing: the reader privileges the first passage slot, and what follows is under-used whether it is a distractor or a second copy of the answer.

What this experiment does not show. The two golds are nearly always different chunks of one Wikipedia article (240 of 240 share a title; median content token-Jaccard 0.16; two pairs are byte-identical, App. G), the M2KR INFOSEEK split supplies *pseudo*-gold passages rather than curated oracles, and E-VQA as distributed in M2KR has no multi-positive questions at all. This tests positional privilege under *redundant* evidence, not multi-hop reasoning where two distinct passages must be combined; App. G gives the full construction and its caveats.

Table 5: **Multi-gold cells** (Qwen2.5-VL-7B $\times$ INFOSEEK, $k=10$, $n=240$, pre-registered). Each cell places one or two answer-bearing gold passages at the listed slots; the remaining slots are filled from the same ranked distractor list, so a two-gold cell holds eight of the nine distractors used in the one-gold cells. A second gold separately helps only when slot 0 is not already gold. Construction and caveats: App. G.

Cell	golds @	acc (%)	Cell	golds @	acc (%)
C1	$\{0\}$	50.8	C4	$\{0, 9\}$	53.8
C2	$\{4\}$	37.5	C5	$\{4, 9\}$	41.7
C3	$\{9\}$	34.6	C6	$\{0, 4\}$	51.3

Insight 3: Rerank for rank 1, not recall@ k

Reranker training should explicitly reward placing the correct passage at rank 1 and penalize placing it anywhere else; recall@ k objectives reward the wrong target here, since a frozen reader privileges slot 0.

4.4 Retrieval-Side Fixes Don’t Help

Finding 4: Retrieval-side fixes don’t help on a frozen reader

MMR diversification, oracle reranking, and rank-based distractor reordering all fail to improve on the deployment default within paired-bootstrap CIs. With the slot-0 anchoring of §4.3, recall@ k is the wrong metric for deployed KB-VQA.

All three act only on what the reader sees: diversification of the retrieved set, reranking that pushes gold toward the start of the context, and reordering of the distractors themselves. We focus on training-free retrieval-side interventions because they are the cheapest to deploy and align with the design space surveyed by Hu et al. (2025). Reader-side interventions are out of scope for a frozen-reader paper but are natural next steps.

MMR diversification does not help. We replace top- k PreFLMR ViT-G retrieval with MMR (Carbone and Goldstein, 1998) over the top-50 PreFLMR ViT-G candidates, sweeping $\lambda \in \{0.3, 0.5, 0.7\}$ at $k \in \{5, 10\}$ (six cells, INFOSEEK, Qwen2.5-VL-7B). The pre-registered KILL rule — registered separately from the five-way classifier of App. C — abandons the intervention if the best MMR gain over vanilla is under 1 pp at every (k, λ) . The maximum point gain is +1.00 pp, at the threshold rather than under it, and every paired-bootstrap CI on $\Delta(\text{MMR} - \text{vanilla})$ overlaps zero;

Figure 8 in App. R shows the flat curves.

Oracle reranking does not separably help. On the gold-in-top-10 subset ($n=254$) we compare *native* (gold at its natural PreFLMR ViT-G rank), *first* (forced to index 0) and *last* (index 9): accuracies 47.64%, 50.79%, 42.13%. First-vs-native is +3.15 pp, directionally positive but not separable at this sample size; last-vs-native is −5.51 pp, separable on the negative side. We read this as suggestive of asymmetry rather than a characterized result — the promotion side is underpowered at $n=254$ (App. Q).

Rank-based reordering does not help either. Reversing the distractor ranking against the rank-order deployment default (Cells B and A of §4.3) gives +2.42 pp [−0.60, +5.65] on INFOSEEK and −3.20 pp [−6.60, +0.00] on E-VQA; neither is separable (App. P).

Insight 4: Invest in reader-side fixes, not more retrieval

Engineering investment should shift from retrieval (larger pools, diversification, reordering) to reader-side training and inference-time intervention; no retrieval-side variant we tried improved on the deployment default.

4.5 What Reader-Side Mitigation Would Look Like

Our results are diagnostic, so a fair question is what a practitioner should do with them. We propose no method, but the measurements constrain the design space enough to name three directions.

Text-side calibration. The most actionable consequence of §4.2 is negative: interventions aimed at the image encoder or the vision–language projection target the wrong component. The transferable candidates are those developed against text-LLM positional bias — attention calibration that rescales per-passage attention by position-debiased relevance (Hsieh et al., 2024), and the attention-sink account of early-token dominance (Xiao et al., 2024) — both inference-time and reader-internal, matching where we locate the effect. **Rank-1 reranking objectives.** §4.3 quantifies how little a correct passage is worth outside the front slot: with a gold already at slot 0, a second answer-bearing gold at slot 9 adds under 6.25 pp against a +16.25 pp gap. An objective rewarding gold anywhere in the top k optimizes a quantity the reader largely ignores; list-wise permutation-aware ranking (Tang et al., 2024) and attention-based rerankers (Chen et al., 2025)

already operate on whole orderings rather than on passages in isolation. **Scale as a partial, costly mitigation.** Going from 7B to 32B shrinks the gap by 6.05 pp but leaves it separable, leaves the failure signature intact, and at $k=10$ on E-VQA exceeded our memory ceiling on a single 80 GB accelerator. **Why the protocol is the contribution.** Each direction is a hypothesis our probe can adjudicate, because the gold-position design supplies what an aggregate benchmark cannot: a within-question paired contrast whose only difference is a permutation, so an intervention’s effect on position dependence is measured directly rather than inferred from an accuracy delta that also moves with retrieval quality. Concurrently, Jung and Wang (2026) propose exactly the text-side calibration above. We have not evaluated it and make no priority claim, but it is the natural first candidate to run through the protocol we release.

5 Conclusion

This paper provides the first controlled probe of reader-side position dependence in deployed multi-modal KB-VQA. Across three open-source 7B/8B readers on two benchmarks, and five models from 3B to 32B, the shape is primacy, not U; the effect survives removing the image on matched data, making it a text-side property of the reader; primacy is anchored at prompt slot 0, to the point that a second answer-bearing passage placed later is largely wasted; and three training-free retrieval-side interventions fail to improve on the deployment default. Scale attenuates the gap at 32B without closing it or changing its signature. The practical consequence is that $\text{recall}@k$ measures something the reader largely ignores, and the remaining headroom sits on the reader side. We release the protocol as a controlled instrument for evaluating any candidate reader-side or training-time intervention.

Limitations

Scope of empirical claims. Our findings hold within: open-source instruction-tuned VLMs from 3B to 32B (Qwen2.5-VL-7B, InternVL3-8B, Qwen3-VL-8B, plus Qwen2.5-VL-3B and Qwen2.5-VL-32B on INFOSEEK); contexts at $k \leq 20$; English KB-VQA over Wikipedia-derived bases of 10^5 passages; greedy decoding; one image per question. We do not claim proprietary VLMs (e.g. GPT-4V (OpenAI, 2023)) or other prompt formats would behave the same; Tian et al. (2025) and Yao et al. (2025) both report proprietary-vs-open-source differences. Our readers span three families, and Cuconasu et al. (2025) report that Qwen-2.5 7B is more positionally biased than Llama-3 in text RAG, so family plausibly matters; Llama-3-based VLMs are untested here. The no-recency-rebound claim holds on five of six $k=20$ cells, the sixth being the Qwen3-VL-8B $\times$ E-VQA reversal reported in §4.1, for which we commit to no mechanism.

One task family, and a modest context budget. Both benchmarks are single-image KB-VQA over Wikipedia-derived text, and neither requires combining two distinct passages. So we cannot separate retrieved from parametric knowledge — for well-covered entities a reader may not use the context at all, which would attenuate rather than manufacture a position effect, but is uncontrolled here — and $k=20$ is far short of what frontier readers accept, so we cannot say whether the shape persists at 10^2 – 10^3 candidates. Long-horizon personal-memory QA with multi-gold evidence pools (Mei et al., 2026) and long interleaved multimodal documents (Wang et al., 2024) would address both at once, under an oracle-plus-noise construction that keeps accuracy clear of the floor where a position probe has no room to move. We ran neither, and did not test the strongest recent open readers (e.g. Qwen3.6-27B).

Mechanism attribution is partial. Section 4.3 rules out image-token proximity and retrieval similarity as primary drivers and narrows the effect to slot 0, without identifying the underlying mechanism. Instruction-tuning data is a recurring candidate: Liu et al. (2024, §4.3) considered it but found the U-shape in both base and instruct MPT-30B, and Lu et al. (2022) show demonstration order alone can swing in-context performance from chance to SOTA. We find the hypothesis consistent with our data but do not test it. The matched-corpus control says where *not* to look — not the vision

encoder or the vision–language interface — but a positive account needs attention- or representation-level probes we leave to future work.

A claim withdrawn between review and publication. The version reviewed at ARR contained a further finding: that the multimodal setting amplifies text-mode primacy by 2.2 – $4.5\times$. Reviewers correctly observed that the supporting ablation varied corpus (NQ-Open vs. M2KR) together with modality. We ran the prescribed matched control, it did not support the claim, and we withdrew it; §4.2 reports the control in its place. We record it because the reviewed version is public. What replaces it is weaker than the reverse claim would be: the control establishes that primacy does *not depend on* the image, not that the image dampens it. The text-minus-multimodal interaction is same-signed on all three cells but separable on one, on E-VQA no per-position difference separates from zero, and the control covers three reader-by-benchmark cells rather than all six (Qwen3-VL-8B has no matched text-only run). The position-local-protection reading should be treated as an INFOSEEK-specific hypothesis from row-level overlap, not an established mechanism.

Sample size and multiplicity. Per-cell accuracies use seed-locked subsets of 486–500 questions, chosen for paired-bootstrap tractability across many cells; on smaller slices (the $n=254$ oracle-rerank and $n=240$ multi-gold cells) the noise floor is higher, and the directional first-vs-native oracle gain does not clear it. We report roughly fifty between-cell deltas without family-wise or false-discovery correction. This does not touch the headline effects — every KB-VQA first-vs-last gap at $k=10$ clears zero by at least 7 pp and would survive any standard correction, as would the null-supported claims, which correction only strengthens — but it does confine our exposure to boundary-positive results, and we mark the one that matters (the post-hoc difference-of-increments interval in §4.3) as failing even a two-test Bonferroni adjustment.

Bucket diagnostic is heuristic, and mitigations tested are training-free. The four-bucket diagnostic is a string-match device, not a causal account: it conflates surface-form with semantic failures, cannot attribute a prediction to one source when it appears in several passages, and has paraphrase blind spots. Re-bucketing under a stricter normalizer (App. F) leaves the qualitative pattern intact, but the quantitative shares should be read accord-

ingly; it also works at passage-slot granularity, so whether the answer’s location *within* a passage matters is outside our scope. Separately, we do not claim that learned rerankers (Yan and Xie, 2024), attention-based rerankers (Chen et al., 2025), or self-reflective selectors (Cocchi et al., 2025) cannot help — they reshape what the reader sees in ways our oracle protocol does not test. We claim only that simple training-free retrieval-side manipulations on a frozen reader do not close the gap.

Ethical Considerations

Potential risks. This is a characterization study: we measure how existing frozen vision-language readers respond to ordered retrieved context, and release no new models or datasets. The findings (a primacy anchor on prompt slot 0) describe an existing property of deployed KB-VQA systems and do not expose a new attack surface; any practitioner deploying such a system can read this work and adjust their pipeline accordingly. We do not foresee significant misuse risk beyond what is inherent in publishing any analysis of deployed-system behavior.

Artifact licenses and intended use. All models and datasets used are publicly released, all but the 3B reader and E-VQA under a permissive license, and our use (academic measurement of position effects on standard KB-VQA benchmarks) is consistent with each artifact’s intended use. Models: PreFLMR ViT-G (MIT) (Lin et al., 2024); Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-32B-Instruct and Qwen3-VL-8B-Instruct (Apache 2.0) (Bai et al., 2025b,a); InternVL3-8B (Apache 2.0) (Zhu et al., 2025); Qwen2.5-VL-3B-Instruct (Qwen Research License, non-commercial research use) (Bai et al., 2025b). Datasets: M2KR benchmark (MIT) and INFOSEEK (Apache 2.0) (Lin et al., 2024; Chen et al., 2023); Encyclopedic-VQA (released through the `google-research` repository, no explicit dataset LICENSE file; image content drawn from iNaturalist 2021 (CC-BY) and Google Landmarks v2 and textual content derived from Wikipedia under CC-BY-SA) (Mensink et al., 2023); the NaturalQuestions-Open multi-document protocol of Liu et al. (2024) (MIT).

Data content. INFOSEEK and Encyclopedic-VQA are derived from Wikipedia/Wikidata and curated visual-entity datasets. We did not perform an independent PII or offensive-content audit on top of the curation already conducted by the original dataset

authors; our use is read-only and does not modify, redistribute, or augment the underlying corpora. All passages we surface to readers are taken verbatim from the M2KR-released knowledge bases.

Acknowledgements

The authors used AI assistants (LLM-based coding and writing tools) during this project for three purposes: (i) drafting and editing prose, (ii) generating and debugging analysis and plotting code, and (iii) surfacing related work for follow-up reading. All experimental design, statistical methodology, claims, and interpretations are the authors’ own; all AI-suggested code was reviewed and validated against the pre-registered methodology notes (App. A) before producing the numbers reported here.

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, and 1 others. 2025a. [Qwen3-vl technical report](#). *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *arXiv preprint arXiv:2502.13923*.
- Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. [Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 291–305, Abu Dhabi, United Arab Emirates.
- Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Rita Sarto, Marcella Cornia, Lorenzo Baraldi, and Sara Cucchiara. 2024. Wiki-LLaVA: Hierarchical retrieval-augmented generation for multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1818–1826.
- Jaime Carbonell and Jade Goldstein. 1998. [The use of MMR, diversity-based reranking for reordering documents and producing summaries](#). In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.
- Shijie Chen, Bernal Jiménez Gutiérrez, and Yu Su. 2025. [Attention in large language models yields efficient zero-shot re-rankers](#). In *The Thirteenth International Conference on Learning Representations (ICLR)*.

- Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, So-ravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14948–14968.
- Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Augmenting multimodal LLMs with self-reflective tokens for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9199–9209.
- Florin Cuconasu, Simone Filice, Guy Horowitz, Yoelle Maarek, and Fabrizio Silvestri. 2025. Do RAG systems really suffer from positional bias? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 28022–28036, Suzhou, China.
- 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)*.
- Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T. Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics (ACL)*, pages 14982–14995.
- Chan-Wei Hu, Yueqi Wang, Shuo Xing, Chia-Ju Chen, Suofei Feng, Ryan Rossi, and Zhengzhong Tu. 2025. *mRAG: Elucidating the design space of multi-modal retrieval-augmented generation*. *arXiv preprint arXiv:2505.24073*.
- Jan Hutter, David Rau, Maarten Marx, and Jaap Kamps. 2025. Lost but not only in the middle: Positional bias in retrieval augmented generation. In *Advances in Information Retrieval (ECIR)*, volume 15572 of *LNCS*, pages 247–261. Springer.
- Hoin Jung and Xiaoqian Wang. 2026. *The cost of context: Mitigating textual bias in multimodal retrieval-augmented generation*. *arXiv preprint arXiv:2605.05594*.
- Omar Khattab and Matei Zaharia. 2020. *ColBERT: Efficient and effective passage search via contextualized late interaction over BERT*. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.
- Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024. PreFLMR: Scaling up fine-grained late-interaction multi-modal retrievers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5294–5316.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. *Visual instruction tuning*. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Parmajape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. *Lost in the middle: How language models use long contexts*. *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8086–8098.
- Jingbiao Mei, Jinghong Chen, Guangyu Yang, Xinyu Hou, Margaret Li, and Bill Byrne. 2026. *According to me: Long-term personalized referential memory QA*. *arXiv preprint arXiv:2603.01990*.
- Thomas Mensink, Jasper Uijlings, Lluís Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3113–3124.
- OpenAI. 2023. GPT-4V(ision) system card. <https://openai.com/index/gpt-4v-system-card/>.
- Ofir Press, Noah A. Smith, and Mike Lewis. 2022. *Train short, test long: Attention with linear biases enables input length extrapolation*. In *The Tenth International Conference on Learning Representations (ICLR)*.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. *RoFormer: Enhanced transformer with rotary position embedding*. *Neurocomputing*, 568:127063.
- Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. 2024. *Order matters: Exploring order sensitivity in multimodal large language models*. *arXiv preprint arXiv:2410.16983*.
- Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2024. *Found in the middle: Permutation self-consistency improves listwise ranking in large language models*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 2327–2340.

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.

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. 2024. *Needle in a multimodal haystack*. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Kebin Wu and Fatima Albreiki. 2026. *Positional bias in multimodal embedding models: Do they favor the beginning, the middle, or the end?* In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 26948–26955.

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

Yibin Yan and Weidi Xie. 2024. EchoSight: Advancing visual-language models with wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1538–1551, Miami, Florida, USA.

Jiayu Yao, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Yuyao Ge, Zhecheng Li, and Xueqi Cheng. 2025. Who is in the spotlight: The hidden bias undermining multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15183–15193.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. *InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models*. *arXiv preprint arXiv:2504.10479*.

A Pre-registered Methodology Notes

Most experiments were preceded by a methodology note (decision rules, expected effect-size bands, comparators) committed to version control before the run. The notes that exist are: P15a (decision-rule protocol), P16 (allocator probe), P17 (gold-position anchor), P17b (second-model extension), P19 (MMR), P20 (E-VQA replication), P20-VANILLA, P21-TEXT-LIU (text-only NQ ablation on Qwen2.5-VL-7B), P22 (image-at-end), P23 (INFOSEEK shuffle), P24 (E-VQA shuffle), P25

(E-VQA $k=20$), P26 (four-panel k -sweep generalization), P27 (text-only NQ ablation on InternVL3-8B), P34-MULTIGOLD (multi-gold, with a pre-execution amendment restricting the population), and a REVIEW omnibus. Each note records its hypothesis and, where one applies, its decision rules verbatim. One registered element was not executed: the P34 note pre-registers the five-way classifier of App. C for the multi-gold contrasts, and the runner defines but never calls it, so the multi-gold verdicts in §4.3 rest on the paired-bootstrap intervals and McNemar tests we report rather than on a classifier label.

Experiments run without a pre-registration note. For completeness we state which experiments have no note rather than implying blanket coverage: P17c (oracle-rerank demo), P28 (transformers version regression), P29 (Qwen3-VL-8B tiers), and the post-response runs P30, P32 and P33 (matched-corpus text-only controls), P31 (3B INFOSEEK), P35 (32B INFOSEEK) and P36 (32B E-VQA). These inherit the frozen P17/P17b/P20 protocol unchanged, varying exactly one factor — reader identity, library version, or presence of the image — and none of them defines a new decision rule. We flag them explicitly because the pre-registration guarantee applies only to the notes listed above.

B Subset Construction and Audits

Both subsets are drawn with `Generator(PCG64(42))` and audited against the full released split at a ± 5 pp gate. The INFOSEEK audit gates on image-filename prefix, positive-passage count and answer count, with question type and instruction prefix reported but not gated; the E-VQA audit gates on image source, question type, positive-passage count and gold-answer length. Both PASS. The E-VQA draw is taken directly from the released M2KR split; the INFOSEEK draw is taken from a fixed 2,000-question pool used in our earlier k -sweep.

C Statistical Methodology

Per-cell accuracy uses Wilson 95% intervals ($z=1.96$). Between-cell deltas use a paired bootstrap over question ids with $B=10,000$ resamples, a PCG64 generator seeded at 42, and 2.5/97.5 percentiles. “Separable” throughout means the delta’s 95% interval excludes zero. Interaction quantities (a difference of two deltas, e.g. text-vs-image gaps

or cross-model gap-of-gaps) resample question ids jointly across all four cells so the pairing is preserved.

The five-way classifier, with its bands. Given a delta, its interval, and a comparator anchor A , the verdict is: FLAT if the interval contains zero; otherwise CONTRADICTION if the delta’s sign differs from A ; otherwise, with $r = \Delta/A$, AMPLIFIED if $r > 2.0$, PARTIAL if $r < 0.5$, and the same-magnitude band $0.5 \leq r \leq 2.0$ otherwise. We print the thresholds here so the classifier can be reconstructed from the paper alone.

Label provenance. The same-magnitude band was named REPLICATION in the originating pre-registration note (P17b), and the released result artifacts emit that string, apart from the 32B run, which emits GENERALIZATION. Later notes and the version of this paper reviewed at ARR renamed it GENERALIZATION, on the reasoning that most uses compare a *different* reader or benchmark against the anchor rather than repeating the same cell. The thresholds were identical and no verdict differs under either name, but because a rename can read as a reinterpretation of a result, this version simply describes the band by its thresholds and reserves the word for the artifacts that emit it. Readers comparing the paper against released artifacts should expect to see REPLICATION there.

Determinism. Decoding is greedy, so the RNG seed does not affect generation; it fixes only the bootstrap resampling. Run-to-run reproducibility of the predictions themselves depends on GPU, batch composition and kernel backend. We disable the cuDNN SDP backend for stability, and we observed 96.6% bit-identical predictions across a transformers 4.49-to-4.57 upgrade on the anchor cell, so a small amount of residual non-determinism should be expected rather than exact bit-equality. Legacy independent-sample CIs are preserved as `*_legacy_indep` fields for reproducibility audits.

D Scoring Sensitivity

E-VQA: does the matcher choice matter?

We recompute the $k=10$ first-vs-last delta under all three matchers. Qwen2.5-VL-7B: +26.40 [+22.20, +30.60] under `orig` (substring), +26.60 [+22.40, +30.80] under `wb` (word-boundary), +17.80 [+13.80, +21.80] under `exact_only`. InternVL3-8B ($n=499$): +23.25 [+19.04, +27.45], +23.25 [+18.84, +27.66],

+19.24 [+15.23, +23.25]. Every interval excludes zero and every classifier verdict is unchanged, so the effect is not an artifact of matcher leniency. Exact-equality scoring lowers the magnitude, as expected when free-form answers are penalized for surface form. Normalizer-mismatch quantification: 13 of 2,997 pred-side disagreements, 0 verdict flips.

INFOSEEK: does strict scoring create the effect?

The concern runs the other way — that the official exact-match rule marks verbose-but-correct answers wrong, and that those cluster at *first*. We test it by re-scoring under lenient containment, in which a gold answer counts if it appears inside the prediction. Because the choice of lenient rule is arbitrary, we run three variants rather than privileging one: the official INFOSEEK normalizer with and without token-boundary anchoring, and the E-VQA normalizer with anchoring. The gap *widens* under every variant — Qwen2.5-VL-7B from +16.53 to +22.98–23.39 pp, Qwen2.5-VL-32B from +10.48 to +15.32–16.53 pp — so we report the spread rather than a point estimate, since the conclusion does not depend on which matcher is chosen. Taking the boundary-anchored official normalizer as the representative case, the intervals are +22.98 [+19.15, +27.02] and +15.32 [+11.69, +18.95]. Strict scoring therefore understates the position effect.

Naming. This lenient rule is *lexical containment*, optionally anchored to token boundaries. It is not BEM (Bulian et al., 2022), which is a learned BERT-based answer-equivalence model; no such model is loaded anywhere in this work, and we make no claim about how the effect would look under BEM itself. Both analyses in this appendix are produced by a single committed script over the released prediction files.

E Fallback-Gold Sensitivity

Paired-bootstrap clean-subset Δ s for all six deltas after restricting to questions with no fallback-gold rule applied (INFOSEEK $n=354$; E-VQA $n \in \{432, 433\}$). Maximum shift in any delta is 1.83 pp; classifier verdict labels are unchanged. Per-cell accuracies move more than the deltas do, since a shift shared by two cells cancels in their difference: across the thirteen Qwen2.5-VL-7B $\times$ INFOSEEK cells of the k -sweep (one at $k=1$, three at each of $k \in \{3, 5, 10, 20\}$) every cell rises, by at most 5.2 pp (largest: $k=3$ *middle*, 40.32 $\rightarrow$ 45.48); the

three $k=10$ cells rise by 2.9–4.1 pp.

F Bucket Re-analysis Under Stricter Normalizer

We re-bucket all 962 P17 INFOSEEK Qwen2.5-VL-7B error rows at $k=10$ (272 first + 336 middle + 354 last) under a normalizer matching the answer-scoring normalizer. Result: at $k=10$ no error changes bucket and no bucket share shifts. Across the full sixteen-cell re-bucketing (5,004 rows, P17 and P17b), 3 errors flip and the maximum per-cell shift in any bucket is 0.3 pp. The qualitative pattern in Section 4.3 survives normalizer choice. Figure 3 shows the bucket distribution by gold position.

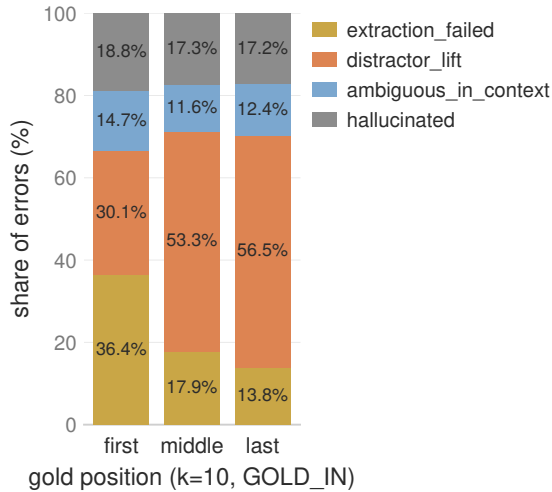


Figure 3: **Failure-bucket distribution shifts with gold position.** Per-cell breakdown of P17 INFOSEEK Qwen2.5-VL-7B error rows at $k=10$ ($n=962$; 272 first, 336 middle, 354 last). As gold moves first $\rightarrow$ last, the modal failure mode flips from `extraction_failed` (36.4% $\rightarrow$ 13.8%) to `distractor_lift` (30.1% $\rightarrow$ 56.5%).

G Multi-Gold Construction and Caveats

Population. Of the 496 INFOSEEK anchor questions, 142 have no positive in the PreFLMR ViT-G top-50 (fallback gold, excluded), 114 have exactly one, and 240 have two or more together with at least eight non-positive distractors. A pre-execution amendment further restricted the analysis population to questions where the answer string is independently present in *both* chosen golds under the official normalizer, since a second gold that does not contain the answer confirms the hypothesis trivially. That check returned 240 of 240,

so the amendment did not change the population. The note also pre-registered a sensitivity slice restricted to the 228 String/Time questions, for which substring answer-presence is most reliable; every verdict is unchanged on that slice (C4–C1 +3.07 [−0.44, +6.58]; C5–C3 +7.46; C1–C5 +8.77). A second gold at slot 4 rather than slot 9 is worth no more once slot 0 is gold: C6–C1 is +0.42 pp [−2.92, +4.17]. The two golds are the first two positives by PreFLMR ViT-G rank; distractors are the top-50 minus the full positive set, exactly as in the anchor.

Redundancy, quantified. All 240 gold pairs share a Wikipedia title. Content token-Jaccard between the two golds has median 0.163, p10–p90 0.109–0.671, and full range 0.076–1.000. Two of the 240 pairs are byte-identical passages that the knowledge base lists under two ids, for which “different chunks” is false; we retain them for population integrity and note that removing them cannot change a null result. An earlier pre-registration note described the spread as a 0.12–0.68 range; that was a p10–p90 interval and is restated as such here.

Annotation noise. The M2KR INFOSEEK split provides pseudo-gold passages selected automatically rather than human-verified oracles. Multi-gold results should be read with that in mind: the positives are noisy, and a benchmark with curated multi-hop evidence would test a different and stronger claim.

Cache reuse. The three single-gold cells are prompt-identical to the corresponding anchor cells restricted to these 240 questions, so their predictions are reused rather than recomputed (720 of 720 cache keys matched, 0 prediction mismatches). Only the three two-gold cells required fresh inference.

H Per-Position Cost of Removing the Image

Detail for the corresponding paragraph of §4.2. Paired-bootstrap differences (text-only minus multimodal) per position cell, over the shared question set, are given in Table 6.

On INFOSEEK the middle and last cells absorb most of the loss, on both reader families. On EVQA no cell moves separably, so the protection reading is INFOSEEK-specific and we do not generalize it.

Row-level overlap with image-at-end. Let S_{end} be the questions answered correctly under the

Table 6: **Per-position cost of removing the image.** Paired-bootstrap differences (text-only minus multimodal) for each position cell, over the shared question set. On INFOSEEK the *middle* and *last* cells absorb most of the loss on both reader families; on E-VQA no cell moves separably.

Cell	first	middle	last
Qwen2.5-VL-7B $\times$ INFOSEEK	-4.03 [-7.26, -0.81]	-15.12 [-18.75, -11.49]	-10.08 [-13.71, -6.45]
Qwen2.5-VL-7B $\times$ E-VQA	+2.80 [-0.20, +5.80]	-2.00 [-5.40, +1.40]	-0.60 [-4.00, +3.00]
InternVL3-8B $\times$ INFOSEEK	-3.02 [-6.85, +0.81]	-6.65 [-10.08, -3.42]	-4.23 [-7.46, -1.01]

image-at-start default but wrongly under image-at-end, and S_{none} the same for image-removed (Qwen2.5-VL-7B $\times$ INFOSEEK). At *middle*, $|S_{\text{end}}|=67$, $|S_{\text{none}}|=87$, and 66 of the 67 lie inside S_{none} (Jaccard 0.750 against a chance level of 0.31). At *last*, $|S_{\text{end}}|=53$, $|S_{\text{none}}|=72$, overlap 52 (Jaccard 0.712, chance 0.27). Moving the image and deleting it therefore damage the same specific questions rather than merely similar numbers of them, which is what motivates reading them as one mechanism at two doses. This is a descriptive row-level observation on one cell; we did not design an experiment to test it.

I Image-at-End Post-Hoc Finding

Under image-at-end, middle accuracy drops disproportionately (21.77% vs. 32.26% under image-at-start, -10.48 pp), while last drops less (25.40% vs. 28.63%, -3.23 pp). The middle-vs-last ordering inverts: $\Delta(\text{middle} - \text{last})$ flips from +3.63 pp at image-start to -3.63 pp at image-end (paired-bootstrap CI [-6.05, -1.41] excludes zero). Image position therefore exerts a mild recency-style protection on adjacent passages but is too small to override the first-position prior, consistent with proximity playing a secondary modulating role on the middle/last contrast while the first-vs-last anchor remains driven by text-position. Figure 4 compares the two templates.

J Distractor-Shuffle Cross-Tabulation (INFOSEEK)

Figure 5 plots the INFOSEEK cross-tabulation reported numerically in Table 4: for each shuffle cell, the share of distractor-lift errors that selected the passage at slot 0 against the share that selected the passage at PreFLMR ViT-G rank 1. Disclosure: the pre-registration’s verdict picker averages over all three cells, including the degenerate Cell A in which slot 0 and rank 1 coincide, and on that average it returns MIXED-LEANING-POSITION; restricted to the dissociable cells B and C it returns

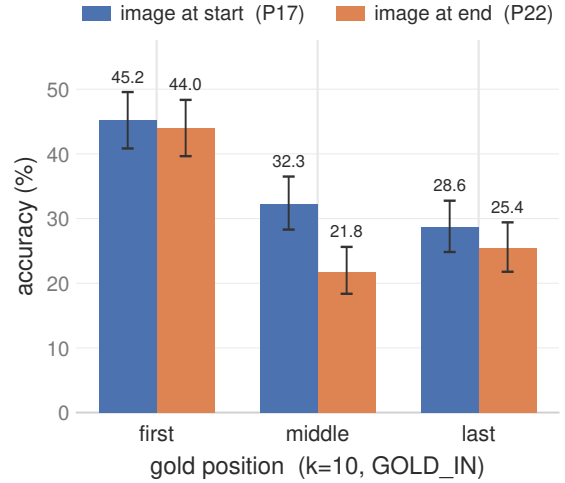


Figure 4: **Image-token proximity is not the primary driver of first-vs-last.** Per-cell accuracy under image-at-start (P17 default) vs. image-at-end. First-vs-last gap preserved (1.12 $\times$ anchor); middle/last ordering inverts under image-at-end (App. I discusses).

POSITION. We report the B+C reading throughput, since Cell A cannot separate the two axes by construction, and the released artifact records both.

K Shuffle Replication on E-VQA (P24)

Figure 6 shows the cross-tabulation for the P24 shuffle replication on Qwen2.5-VL-7B $\times$ E-VQA gold-at-last ($n=500$). The qualitative pattern is identical to the INFOSEEK result (Figure 5): slot-0 anchoring dominates in dissociable cells B+C (4.42 $\times$ chance, vs 3.83 $\times$ on INFOSEEK), while rank-1 selection sits at chance (1.00 $\times$, vs 1.01 $\times$ on INFOSEEK). Aggregate-accuracy comparisons across the three cells against the deployment default are audited in App. P.

L Full k -sweep on Six Reader $\times$ Benchmark Cells (P26 + P29 Tier 2)

Table 7 reports per-cell accuracies for the 6-panel k -sweep summarized in Figure 2. The pattern is

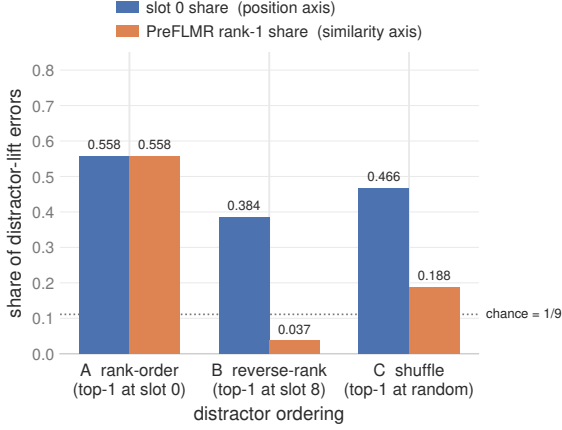


Figure 5: **Reader follows prompt slot, not retrieval rank** (INFOSEEK). Distractor-lift error breakdown for the three shuffle cells (gold fixed at slot 9). Slot-0 selection (blue) dominates in dissociable cells B+C; rank-1 selection (orange) runs at the dashed baseline. Numbers are in Table 4; the E-VQA replication is in App. K.

consistent across all six cells: first decays gracefully, middle and last drop steeply and collapse together; no recency rebound at $k=20$ on five of six cells. The sixth cell (Qwen3-VL-8B $\times$ E-VQA) shows a separable last $>$ middle effect at $k=10$ and $k=20$, discussed in §4.1 and App. N. Small middle-vs-last inversions occur on twelve sub-cells across the sweep, every one is within CI noise except the Qwen3-VL-8B $\times$ E-VQA cells discussed in §4.1, the largest of the remainder being InternVL3-8B INFOSEEK $k=3$ at $+2.82$ pp. InternVL3-8B $\times$ E-VQA incurs OOM exclusions on large prompts at every k , giving per-cell n between 486 and 499 (most affected at $k=20$: $n \in \{486, 487, 487\}$ for first/middle/last after 13–14 exclusions per position cell); all other cells use the full subsets, $n=496$ for INFOSEEK and $n=500$ for E-VQA.

Table 7: **Full k -sweep accuracies (first/middle/last %; one prompt at $k=1$) across six reader $\times$ benchmark cells.**

Reader $\times$ Benchmark	$k=1$	$k=3$	$k=5$	$k=10$	$k=20$
Qwen2.5-VL-7B $\times$ INFOSEEK	54.4	48.0/40.3/41.1	47.0/37.3/34.7	45.2/32.3/28.6	43.5/26.6/26.0
Qwen2.5-VL-7B $\times$ E-VQA	70.2	65.0/53.8/50.0	64.0/48.6/42.2	63.2/38.6/36.8	62.8/33.0/34.0
InternVL3-8B $\times$ INFOSEEK	50.2	44.4/28.8/31.7	43.8/27.4/28.0	40.1/23.8/22.2	38.7/20.6/19.8
InternVL3-8B $\times$ E-VQA	69.7	63.2/45.9/45.1	59.5/42.3/40.0	59.1/36.5/35.9	54.9/29.4/29.6
Qwen3-VL-8B $\times$ INFOSEEK	58.7	55.0/41.5/40.7	52.2/37.1/38.9	51.6/33.9/35.5	50.4/31.9/32.9
Qwen3-VL-8B $\times$ E-VQA	72.2	67.4/52.4/53.4	66.6/49.4/50.6	65.6/44.4/47.4	66.4/40.8/46.0

M Text-only NQ-Open Ablation on InternVL3-8B (P27)

We run the NQ-Open text-only protocol on InternVL3-8B using the identical 500-question subset from P21 (sampled from Liu et al. (2024)’s

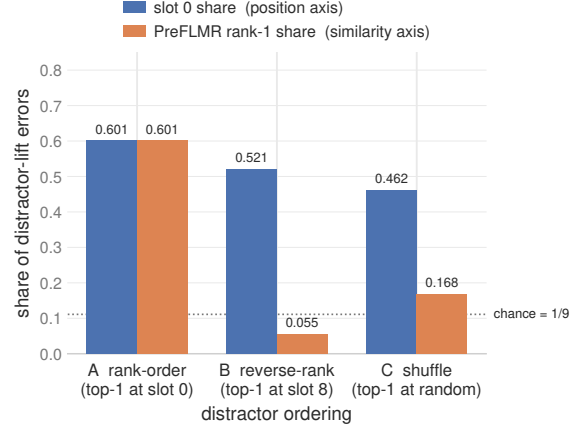


Figure 6: **Shuffle ablation replicates on E-VQA**. Distractor-lift error breakdown for the three shuffle cells on Qwen2.5-VL-7B $\times$ E-VQA gold-at-last ($n=500$). Same protocol as Figure 5; slot-0 dominance is stronger ($4.42\times$ vs $3.83\times$ chance baseline of $1/9$).

2,655 paragraph-answer queries, PCG64 seed 42). Per-cell accuracies (Wilson 95% CI): first 55.00% [50.62, 59.31]; middle 51.20% [46.83, 55.56]; last 49.80% [45.44, 54.17]. Paired-bootstrap deltas ($B=10,000$, seed 42): first–last $+5.20$ pp [$+1.20, +9.40$]; first–middle $+3.80$ pp [$+0.00, +7.60$]; middle–last $+1.40$ pp [$-2.00, +5.00$] (CI includes zero: middle and last are flat, as on Qwen2.5-VL-7B, whose P21 run on the same subset gives first 63.60%, middle 55.40%, last 56.20%, i.e. first–last $+7.40$ pp [$+3.80, +11.00$]). This establishes that primacy is present for InternVL3-8B on a second corpus and a different task format.

N Qwen3-VL-8B Multimodal Replication (P29 Tiers 1, 2)

We replicate the full position-effect protocol on Qwen3-VL-8B-Instruct ($n=496$ INFOSEEK, $n=500$ E-VQA, subsets identical to P17/P20). Tier 1 establishes the $k=10$ row of Table 2: first-vs-last $+16.1$ pp [$+12.5, +20.0$] on INFOSEEK ($0.98\times$ anchor; first 51.61%, middle 33.87%, last 35.48%) and $+18.2$ pp [$+14.2, +22.2$] on E-VQA ($1.10\times$; first 65.60%, middle 44.40%, last 47.40%). Both fall in the same-magnitude band of the pre-registered classifier. Tier 2 extends to $k \in \{1, 3, 5, 20\}$; per-cell accuracies are plotted in the two right-column panels of Figure 2. Pre-registered verdicts for the recency-rebound test on $k=20$:

- Qwen3-VL-8B $\times$ INFOSEEK at $k=20$:

$\Delta(\text{last} - \text{middle}) = +1.0$ pp, paired-bootstrap CI $[-1.0, +3.0]$. No recency rebound.

- Qwen3-VL-8B $\times$ E-VQA at $k=20$: $\Delta(\text{last} - \text{middle}) = +5.2$ pp, paired-bootstrap CI $[+2.6, +8.0]$. *Separable last > middle* (CI excludes zero positively; see §4.1 for framing). At $k=10$ the same effect is present at $+3.0$ pp $[+0.2, +5.8]$.

This is the only cell of six showing a separable last > middle effect in the default protocol.

O Text-only Ablation on Qwen3-VL-8B (P29 Tier 3)

A third NQ-Open text-only run on Qwen3-VL-8B-Instruct (same 500-question subset as P21 and P27). Per-cell accuracies: first 62.0%, middle 59.2%, last 55.6%. Paired-bootstrap deltas: first–last $+6.4$ pp $[+3.2, +9.6]$; middle–last $+3.6$ pp $[+1.2, +6.2]$ (CI excludes zero positively; Qwen3-VL-8B shows separable middle > last in text mode, unlike Qwen2.5-VL-7B and InternVL3-8B, which are flat). The text-mode middle > last on Qwen3-VL-8B inverts in multimodal E-VQA to last > middle by 5.2 pp, the sign reversal reported in §4.1; note that this contrast concerns the *shape* of the middle/last tail rather than the magnitude of the first-vs-last gap, and it is the only cell of six where that tail separably inverts.

P Operational Audit of Shuffle Interventions

The body uses the Cell A / Cell B / Cell C cross-tabulation to establish that the reader follows prompt slot, not retrieval rank (the mechanism question). Here we report the corresponding aggregate-accuracy comparisons, which speak to a different question: does any shuffle variant improve on the deployment default (Cell A = rank-order, the configuration a normal pipeline would produce)?

INFOSEEK (Qwen2.5-VL-7B, gold-at-last, $n=496$). Per-cell accuracies (Wilson 95% CI): A 28.83% $[25.02, 32.97]$; B 31.25% $[27.33, 35.46]$; C 28.02% $[24.25, 32.13]$. Paired-bootstrap deltas ($B=10,000$, seed=42): $\Delta(B - A) = +2.42$ pp $[-0.60, +5.65]$ (not separable); $\Delta(B - C) = +3.23$ pp $[+0.40, +6.05]$ (separable); $\Delta(C - A) = -0.81$ pp $[-3.43, +1.81]$ (flat).

E-VQA (Qwen2.5-VL-7B, gold-at-last, $n=500$). Per-cell accuracies: A 36.80% (matches the $k=10$ E-VQA last cell exactly, confirming protocol deter-

minism); B 33.60%; C 34.20%. Paired-bootstrap deltas: $\Delta(B - A) = -3.20$ pp $[-6.60, +0.00]$ (flat, point-negative); $\Delta(B - C) = -0.60$ pp $[-3.80, +2.40]$ (flat); $\Delta(C - A) = -2.60$ pp $[-5.80, +0.60]$ (flat).

Reading. The mechanism evidence (Cell B+C slot-0 dominance at $3.83\text{--}4.42\times$ chance, rank-1 at chance) is robust and benchmark-independent. The aggregate-accuracy story is different: on neither benchmark does any shuffle variant separably improve on the deployment default. The $\Delta(B - C) = +3.23$ pp on INFOSEEK that might initially read as “demoting rank-1 helps” is driven partly by Cell C dipping ~ 0.8 pp below default by chance, and disappears against the operationally relevant Cell A baseline.

Scope of the test. Cell B is an extreme intervention: it reverses all rankings, not just demotes rank-1. A surgical variant (“move rank-1 to slot 1 only, leave the rest alone”) was not tested and could plausibly behave differently. We do not recommend any training-free position-based intervention in the body of the paper.

Q Oracle Reranking Visualization

Figure 7 shows the oracle-rerank result described in §4.4. Paired-bootstrap intervals on the $n=254$ subset: first–native $+3.15$ pp $[-0.39, +6.69]$; last–native -5.51 pp $[-9.84, -1.18]$. The first-vs-last contrast on the same subset is $+8.66$ pp $[+3.54, +14.17]$, separable and consistent with Table 2.

R MMR Unit Tests

8/8 PASS on synthetic ground-truth ordering at $\lambda \in \{0, 0.5, 1\}$. The unit tests fix the expected selection order at the extremes ($\lambda=1$ reduces to pure relevance, $\lambda=0$ to pure diversity) and at $\lambda=0.5$, and check that the first pick is the arg-max query similarity regardless of λ .

S Distractor-Pool Determinism

We confirm bit-identical $d_0, \dots, d_{k-2}$ across {first, middle, last} cells per qid. We also verify P24 Cell A predictions are bit-identical to the P20 gold-at-last cell (cache reuse 500/500, distractor SHA256 hash failures = 0).

T Qualitative Error Samples

A sixty-row sample (five rows per bucket $\times$ four buckets $\times$ three cells) of P17 INFOSEEK

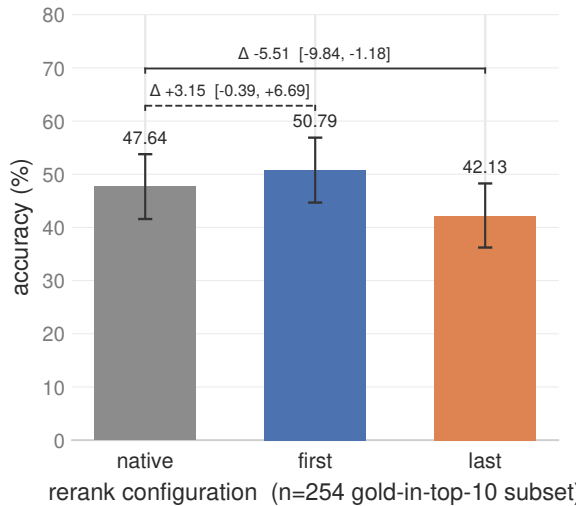


Figure 7: **Oracle reranking does not separably help on a frozen reader.** Qwen2.5-VL-7B on $n=254$ gold-in-top-10 INFOSEEK subset. Solid bracket: paired-bootstrap CI excludes zero. Dashed bracket: CI includes zero. Forcing gold to last yields a separable loss; forcing gold to first yields a directional gain that does not clear the noise floor at this sample size.

Qwen2.5-VL-7B errors ships with the code release at https://github.com/WeMWish/lost_at_the_end_code, under results/p17/qualitative_examples.md.

U Prompts and Decoding

All three readers share one system prompt and one user-message template, defined once and imported by each runner. The system prompt, verbatim and identical across readers:

```
You are a visual question
answering assistant. Use the
retrieved context to answer the
question. Respond with a short
phrase only, no explanation.
```

The user message, verbatim:

```
{question}
Context:
[1] {passage 1 content}
---
[2] {passage 2 content}
...
[k] {passage k content}
```

Passages are 1-indexed and joined by a literal `---` rule on its own line. Passage text is the M2KR `passage_content` field used verbatim, with no re-wrapping. Note that the two benchmarks differ here: INFOSEEK `passage_content` already begins `title:`

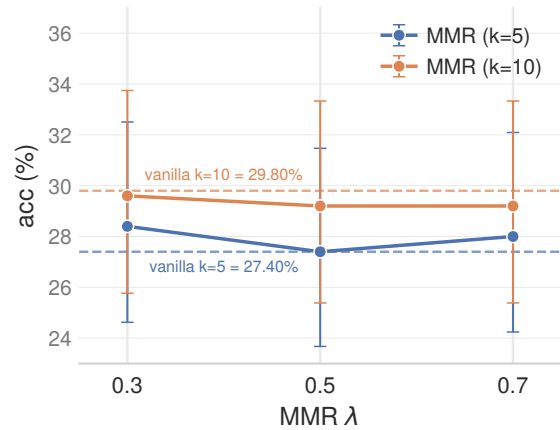


Figure 8: **Diversification doesn't help.** MMR at three λ values fails to improve over vanilla top- k at $k=5$ or $k=10$. Every paired-bootstrap CI on $\Delta(\text{MMR} - \text{vanilla})$ overlaps zero.

... content: ..., whereas E-VQA passages carry no title field and contribute raw content only. There is no explicit answer cue; the assistant turn is opened by the chat template.

Decoding, identical for every cell in the paper. Greedy (`do_sample=False`), `max_new_tokens=32`, no beam search, no top- p , no repetition penalty, `bfloat16` weights. Because decoding is greedy, no sampling seed enters the predictions.

Reader-specific image handling — the only place the three readers differ. Qwen2.5-VL-7B and Qwen3-VL-8B receive the system prompt through the chat template and place the image-token block first in the user message. InternVL3-8B has no per-call system slot in its `model.chat()` path, so the same system text is prepended to the user text and the image is marked with a leading `<image>\n`. Preprocessing follows each model's own recommendation, so that neither is handicapped: Qwen2.5-VL-7B and Qwen3-VL-8B re-size to a 1024 vision-token budget (`min_pixels = 56×56`, `max_pixels = 1024×28×28 = 802,816` for Qwen2.5-VL-7B and `1024×32×32 = 1,048,576` for Qwen3-VL-8B, whose patches are 16 pixels), while InternVL3-8B uses dynamic tiling of up to 12 tiles of `448×448` plus a thumbnail. Only Qwen2.5-VL-7B exposes the image-position toggle used for the image-at-end ablation (§4.3).

Text-only mode (§4.2). The image content block is removed from the user message entirely and no image argument is passed to the processor; everything else, including the system prompt, the passage

block and the decoding configuration, is unchanged. Empirically the resulting prompts are a median 611 tokens shorter on INFOSEEK and 272 on E-VQA than their multimodal counterparts on the same (question, passage ordering) pairs, bounded above by each reader’s vision budget.

Because prompt construction and decoding are identical across position cells by construction, the position effect cannot be a prompt-formatting or sampling artifact.

V Compute and Reproducibility

H100 80GB; bfloat16 weights; torch 2.11. P17–P27 (Qwen2.5-VL-7B and InternVL3-8B experiments) were run under transformers 4.49.0. P29 (Qwen3-VL experiments; see App. N) was run under transformers 4.57.0, which is the minimum version supporting Qwen3-VL. We verified that the library upgrade introduces a small accuracy drift on Qwen2.5-VL-7B (+0.81 pp on the P17 anchor cell, with 96.6% bit-identical predictions); this is well below our smallest reported effect size and does not flip any pre-registered verdict. We accordingly retain transformers 4.49.0 as authoritative for the P17–P27 numbers reported throughout the paper and use transformers 4.57.0 only for Qwen3-VL. The post-response runs P30–P36 ran under torch 2.11.0+cu128; P35 and P36 record transformers 4.49.0 in their released logs, and P30–P34 record no library version, inheriting the repository pin of transformers 4.49.0. Both the cu130 and cu128 CUDA builds of torch 2.11 appear across the experiment set, which is one source of the residual nondeterminism quantified in App. C. Total session compute is approximately \$3–\$5 USD plus \$0.40 for the P21 text-only ablation, \$0.50 for the P24 shuffle replication, \$0.50 for the P25 $k=20$ replication, \$5 for the P26 four-panel k -sweep, \$0.50 for the P27 InternVL3 text-only ablation, and \$1–\$7 for the P29 Qwen3-VL experiments (tier-dependent; see App. N).