AGGREGATE GROUNDING SCORES HIDE CITATION TURNOVER

Anonymous authors
Paper under double-blind review

ABSTRACT

Grounding benchmarks typically score whether a citation agrees with the benchmark's evidence annotation under a single presentation. Such marginal scores do not reveal whether the same annotation-correct citations remain correct under content-preserving re-presentation: losses on some questions can be offset by gains on others, leaving aggregate quality nearly unchanged. We formalize this as an identification problem and introduce the Paired Citation Retention Audit (PCRA), which pairs each question's citation outcome across presentations while holding the reader and available evidence fixed. On 3,330 NExT-GQA questions, two 8B video readers lose 24.2% and 42.0% of their initially correct temporal citations under reordering, corresponding to 3.82% and 5.78% of all question–order pairs, while net hit-rate changes remain near zero. The loss is far larger than under identical-input repeats and minimal perturbation controls, and it persists under exhaustive order coverage, across reader families and packet providers, and at larger model sizes. The same pattern appears in text RAG, where six HotpotQA readers lose 4.0–12.6% of their initially correct paragraph citations. Presentation variation also affects reader comparison. On small labelled video sets, averaging grounding quality across multiple presentations improves ranking reliability, whereas paired retention measures whether previously correct grounding decisions are preserved across presentations. The two therefore serve complementary roles: presentation-averaged quality stabilizes reader comparison, whereas PCRA exposes citation turnover that marginal scores cannot identify.

1 INTRODUCTION

Recent work on grounded generation and retrieval-augmented generation evaluates models not only by answer correctness, but also by whether generated claims are appropriately **supported by evidence and attributed to their sources** (Gao et al., 2023; Liu et al., 2023; Kamalloo et al., 2023; Song et al., 2025; Xia et al., 2025). In retrieval-augmented generation (RAG), citations connect generated claims to external sources that users can inspect and verify. Grounded video question answering imposes a related requirement: models are evaluated on whether they localize the temporal or spatio-temporal evidence associated with an answer, rather than on answer correctness alone (Lei et al., 2020; Xiao et al., 2024). When the cited material does not support the answer, the citation fails to provide the intended target for verification (Liu et al., 2023; Xiao et al., 2024).

A separate line of work shows that model behavior can depend on **input order and presentation**. Language models are sensitive to the ordering of information in long contexts and multiple-choice options (Liu et al., 2024; Gupta et al., 2024; Zong et al., 2024), listwise ranking exhibits position and permutation effects (Tang et al., 2024; Bito et al., 2026), and multi-image and multimodal systems can change their selections when visual inputs are reordered (Tian et al., 2025; Yao et al., 2025). In RAG, retrieval permutations can alter generated answers, while changes in retrieval configuration can alter the evidence delivered to the model (Zhang et al., 2026; Ponnuraj, 2026). Yet these results leave a more specific grounding question open: **if a citation is judged correct for a question once, will it remain correct when the same evidence is reordered?**

Standard grounding scores do not answer this question. They are typically **marginal**: they summarize citation quality under one presentation of the evidence. Two evidence orders can therefore achieve nearly identical hit rates while being correct on different questions. An initial hit may be lost

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

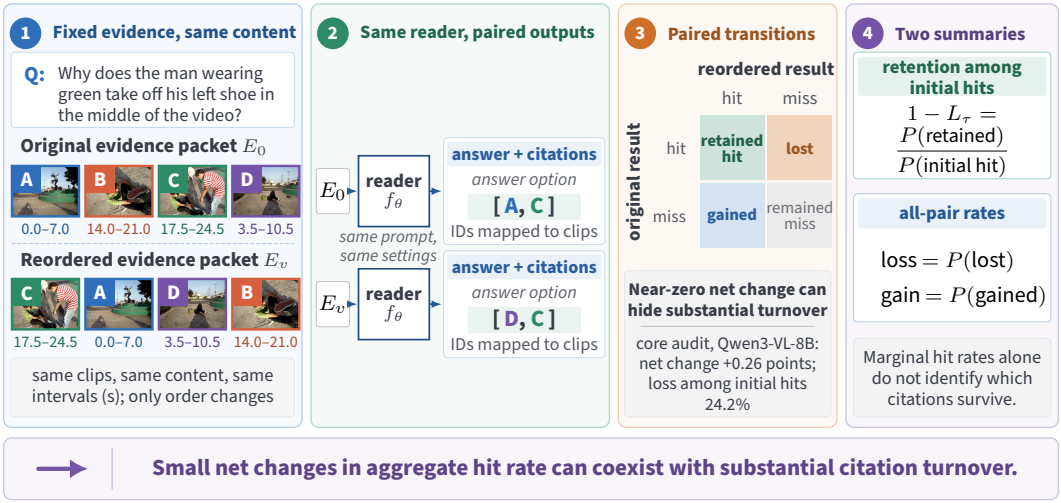

Figure 1: Stable aggregate grounding can hide citation turnover: the PCRA audit, with actual NExT-GQA frames. Outputs in (2) are schematic, with illustrative cited IDs; a hit satisfies $t \geq \tau$. The core-audit values in (3) are Qwen3-VL-8B's (Table 1).

on one question and offset by a gain on another, leaving the aggregate rate nearly unchanged. Stable aggregate grounding therefore does not imply that the same questions remain correctly grounded.

This distinction matters because citations are the evidence references that users and evaluators actually inspect. After the evidence is reordered, a user revisiting the same question may be directed to a different source or video interval even when aggregate benchmark quality is unchanged. For evaluation, this creates an **instance-level reproducibility problem**: a system can preserve its overall score without preserving the grounding decisions that produced it. The same failure mode also arises under an operational pipeline change in our experiments: §6 shows that changing feature-extraction batch size can reorder retrieval packets without changing their membership and, in turn, change the cited source. This is an operational example within the studied pipeline rather than an estimate of production prevalence.

This question is distinct from two adjacent evaluation problems. Evaluation-variance studies examine how benchmark scores and model rankings vary across prompts and other evaluation conditions (Sclar et al., 2024; Mizrahi et al., 2024; Polo et al., 2024; Madaan et al., 2024; Heineman et al., 2025; Li et al., 2025), whereas negative-flip work asks whether correct predictions survive changes to the model itself through updating, retraining, or compression (Fard et al., 2016; Yan et al., 2021; Dutta et al., 2024). Neither directly measures whether previously correct citations are retained when the reader and evidence set are held fixed and only their presentation changes.

Our central claim is that **stable aggregate grounding does not imply instance-level reproducibility under content-preserving re-presentation: gains on some questions can mask losses of previously correct citations on others.**

We organize the study around three questions: **(RQ1)** whether marginal grounding scores identify retention of previously correct citations across content-preserving presentations; **(RQ2)** how much citation turnover occurs under fixed-evidence reordering, how it compares with repeat and perturbation controls, and whether such reorderings can arise from retrieval-pipeline changes; and **(RQ3)** how presentation variation affects reader comparison, and how presentation-averaged quality and paired retention complement one another in grounding evaluation.

To answer these questions, we introduce the **Paired Citation Retention Audit (PCRA)**. PCRA pairs citation outcomes for the same question across content-preserving presentations while holding the reader and evidence set fixed, and decomposes the resulting pairs into retained hits, losses, gains, and remaining misses. §3 shows formally why marginal hit rates do not identify retention. Identical-input repeats and minimal perturbations provide baselines for serving variation and non-order input sensitivity.

Empirically, fixed-evidence reordering reveals a cancellation pattern that marginal grounding scores obscure: previously correct citations are lost on some questions while gains elsewhere leave aggregate quality nearly unchanged. The turnover is much larger than under identical-input and non-order perturbation controls, persists across video and text grounding, and is only partly explained by retrieval-rank and position cues. Presentation variation also affects reader comparison: averaging quality across presentations improves ranking reliability when labelled questions are scarce, whereas paired retention measures whether successful grounding decisions are preserved. Together, these results motivate treating presentation-averaged quality and paired retention as complementary components of grounding evaluation.

Throughout, citation correctness denotes agreement with benchmark evidence annotations. PCRA therefore measures annotation retention rather than semantic evidence sufficiency; model-assessed support is analyzed separately, and human evidence-sufficiency judgments remain outside the present study.

Contributions.

1. **Identification.** We show that marginal grounding scores do not identify whether previously correct citations survive content-preserving re-presentation, even when those marginal scores are known exactly.
2. **Hidden citation turnover.** Across video and text grounding, fixed-evidence reordering can remove many previously correct citations while offsetting gains leave aggregate scores nearly unchanged. The turnover is larger than under repeat and perturbation baselines and persists across readers, datasets, packet providers, and model sizes.
3. **Evaluation design.** We show that presentation variation affects both citation continuity and reader comparison. Presentation averaging improves ranking reliability when labelled questions are scarce, while paired retention measures instance-level citation continuity that marginal quality does not identify.

2 Related Work

Grounding, attribution, and evidence localization. Work on grounded generation evaluates not only answer quality but also the evidence used to support an answer, including citation correctness, verifiability, and attribution quality in text RAG (Gao et al., 2023; Liu et al., 2023; Kamalloo et al., 2023; Song et al., 2025; Xia et al., 2025). In video QA, benchmarks evaluate temporal or spatio-temporal grounding of answers (Lei et al., 2020; Xiao et al., 2024; Meng et al., 2026; Huang et al., 2026b; Wang et al., 2026b), while video retrieval and long-video systems identify relevant intervals or frames for downstream reasoning (Jeong et al., 2025; Zhu et al., 2026; Pan et al., 2026; Zheng et al., 2024; Min et al., 2024; Huang et al., 2026a). Citation correctness and attribution faithfulness need not coincide (Wallat et al., 2024). PCRA instead asks whether a citation that satisfies the benchmark grounding criterion remains correct when the available evidence is held fixed but re-presented; this is an annotation-retention question, not a direct test of semantic support.

Presentation sensitivity and evidence continuity. A large literature documents sensitivity to input position and order in long-context language models, listwise ranking, multi-image VLMs, and multimodal RAG (Liu et al., 2024; Tang et al., 2024; Bito et al., 2026; Tian et al., 2025; Yao et al., 2025). The closest studies ask related but different questions. Zhang et al. (2026) examine whether RAG answers change across permutations of retrieved context; Ponnuraj (2026) measures evidence continuity across retrieval configurations; and Vishwakarma et al. (2026) study which sources answer engines cite first. Ning & Li (2026) audit answer churn as a RAG corpus changes, using repeated inference as a control. **Our setting instead holds both the reader and the delivered evidence set fixed**, varies only how that evidence is presented, and conditions on citations that were initially annotation-correct. This distinction matters when more than one source can satisfy the grounding criterion: a change in source identity alone does not establish that a previously correct citation has been lost.

Paired stability and evaluation variance. Conceptually, our audit is also related to prediction churn and negative flips, which track correct predictions that become incorrect after model updates or compression (Fard et al., 2016; Yan et al., 2021; Dutta et al., 2024). We adopt the same paired perspective for grounding, holding the model and available evidence fixed while varying only their

presentation. Behavioral invariance tests similarly ask whether predictions survive label-preserving input transformations (Ribeiro et al., 2020); our setting differs by conditioning on initial grounding correctness and scoring the cited evidence itself. A separate evaluation-variance literature shows that prompts, option order, and other evaluation conditions can alter aggregate scores and model rankings (Gupta et al., 2024; Zong et al., 2024; Sclar et al., 2024; Mizrahi et al., 2024; Polo et al., 2024; Madaan et al., 2024; Heineman et al., 2025; Alzahrani et al., 2024; Li et al., 2025). We separate two consequences of presentation variation in grounding: **instance-level loss of previously correct citations**, measured by paired retention, and **instability in reader rankings**, for which we study presentation-averaged quality as a complementary evaluation strategy.

3 PCRA: Paired Citation Retention Audit

3.1 Fixed-evidence pairing and transitions

Figure 1 summarizes the audit. For a question q, let $E_0 = (u_1, \ldots, u_n)$ be the delivered evidence packet in provider order and $E_v = \pi_v(E_0)$ a permutation that keeps identifiers, content and intervals. The reader, prompt and decoding settings are fixed. Let c_0, c_v be the cited sources, mapped back from emitted identifiers, and $t_0, t_v \in [0, 1]$ their correctness scores: in video, the best temporal IoU with an annotated window; in text, whether the cited paragraph is a gold supporting paragraph. A hit satisfies $t \geq \tau$ ($\tau = .5$); we call such citations annotation-correct, which is agreement with the benchmark's evidence annotation, not established semantic support (§7). We report

$$L_\tau = \Pr(t_v < \tau \mid t_0 \geq \tau), \qquad F = \Pr(c_0 \neq c_v),$$
$$\Delta R@\tau = P(\text{miss} \to \text{hit}) - P(\text{hit} \to \text{miss}).$$

Retention is $1 - L_\tau$; the all-pair loss $P(t_0 \geq \tau)L_\tau$ is the negative flip rate (Yan et al., 2021). Natural-output runs regenerate answers and citations; we also report the unchanged-answer subgroup. Identical-input repeats resend the original serialized input and separate order effects from serving variation. Uniform replacement, which draws the reordered citation uniformly from the packet, gives a reference loss for a reader that ignores which source it cited before.

Aggregate hit rates do not identify retention. Retention cannot in general be recovered from the two marginal hit rates, even when both are known exactly. Let $H_0 = \mathbf{1}[t_0 \geq \tau]$ and $H_v = \mathbf{1}[t_v \geq \tau]$ with marginal hit rates $p_0 = P(H_0 = 1) > 0$ and $p_v = P(H_v = 1)$. Writing $x = P(H_0 = 1, H_v = 1)$, the four transition probabilities x, $p_0 - x$, $p_v - x$ and $1 - p_0 - p_v + x$ are nonnegative, which gives the Fréchet–Hoeffding bounds (Nelsen, 2006) $\max(0, p_0 + p_v - 1)/p_0 \leq 1 - L_\tau \leq \min(p_0, p_v)/p_0$. When $0 < p_0 = p_v \leq .5$, the marginal rates permit any retention from zero to one. The gap is not estimation noise that more questions would remove: marginal scores discard the pairing needed to tell which annotation-correct citations were preserved.

Uncertainty. Confidence intervals use 10,000 paired bootstrap draws over video clusters (questions in text), with each audit's sampling weights. Comparisons between readers use one shared resample, so pairwise intervals are paired. Reported scores are multiplied by 100.

3.2 Instantiations

Video. NExT-GQA packets contain four clips retrieved with `LanguageBind_Video_V1.5_FT` (Zhu et al., 2024). Each clip is served as an identifier, its bounds and a contact sheet of four timestamped frames (16 frames per question, no audio); the reader returns an answer option and JSON citations, and we score the first resolvable cited clip, falling back to retrieval top-1 when none resolves. Clip identifiers carry the retrieval rank. The core audit serves Qwen3-VL-8B and InternVL3-8B once on all 3,330 eligible validation questions (561 videos): the original order, three reorderings, an identical-input repeat, a one-character whitespace change, and the original and reordered packets under opaque identifiers, all shuffled into one queue (Appendix C.5). Tests described as fixed in advance were locally hash-recorded before inference; later analyses of previously evaluated data are exploratory unless stated otherwise (Appendices A and B).

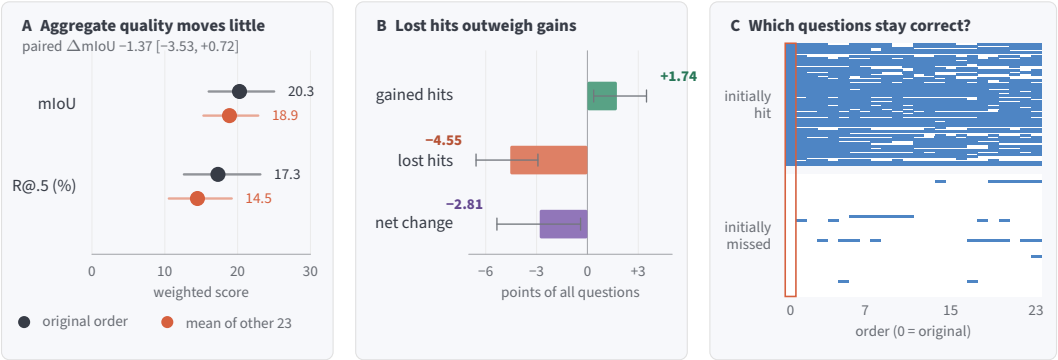

Figure 2: Aggregate quality and citation continuity measure different outcomes (Qwen3-VL-8B, exhaustive 120-question audit). **A**: original-order quality against the mean over the other 23 orders. **B**: newly correct and newly incorrect citations and their net, as points of all questions. **C**: every sampled question under all 24 orders, 60 per historical stratum; filled cells satisfy tIoU $\geq .5$ and the outlined column is the original order. A and B use population weights; bars are 95% video-cluster intervals.

	Qwen3-VL-8B	InternVL3-8B
Initial hit rate (%)	15.8	13.8
Initial-hit questions	526	458
Loss among initial hits (%)		
Reordering	**24.2** [21.4, 27.1]	**42.0** [38.7, 45.3]
answer unchanged	24.2	41.2
opaque identifiers	35.1 [31.8, 38.4]	47.3 [43.7, 50.8]
Whitespace change, same order	3.8	12.4
Identical-input repeat	0.6	2.0
All-pair loss / gain (%)	3.82 / 4.08	5.78 / 5.49
Net $\Delta R@.5$ (points)	+0.26 [−0.48, +1.00]	−0.29 [−1.30, +0.73]

Table 1: Core video audit in one serving run on 3,330 NExT-GQA validation questions (561 videos). Loss conditions on initial hits (tIoU $\geq .5$) over three reorderings; the whitespace change and the repeat keep the original order; opaque identifiers replace the rank-bearing clip IDs in both orders. Separate-run estimates and more cells: Tables 35 and 9; Appendix C.5.

Text. We sample 1,000 HotpotQA distractor questions. Each packet is the native ten Wikipedia paragraphs (two gold) with random paragraph identifiers that are identical across orders. Readers return an answer and paragraph and sentence citations; a hit cites a gold supporting paragraph first. Each question has its dataset order, three seeded permutations and an identical-input repeat (Appendix G).

4 Hidden Turnover Behind Stable Aggregate Scores

Aggregate stability can hide citation turnover. Fixed-evidence reordering can remove many previously annotation-correct citations even when aggregate grounding quality changes little. In the core audit, Qwen3-VL-8B and InternVL3-8B lose 24.2% and 42.0% of their initial temporal hits, while their net changes in $R@.5$ are only $+0.26$ and -0.29 points because losses on some questions are offset by gains on others (Table 1). Restricting to pairs with unchanged answers leaves the loss essentially unchanged, at 24.2% and 41.2%. The key empirical point is therefore not merely order sensitivity, but cancellation across questions: previously correct citations can be lost while new hits elsewhere leave the aggregate score nearly unchanged.

Loss under reordering is much larger than variation under identical-input repeats or minimal perturbations that leave evidence order unchanged. Identical-input repeats lose only 0.6% and 2.0% of initial hits, and a whitespace change loses 3.8% and 12.4% (Figure 3A). Replacing rank-bearing identifiers with opaque identifiers increases rather than removes the loss, to 35.1% and 47.3%, indi-

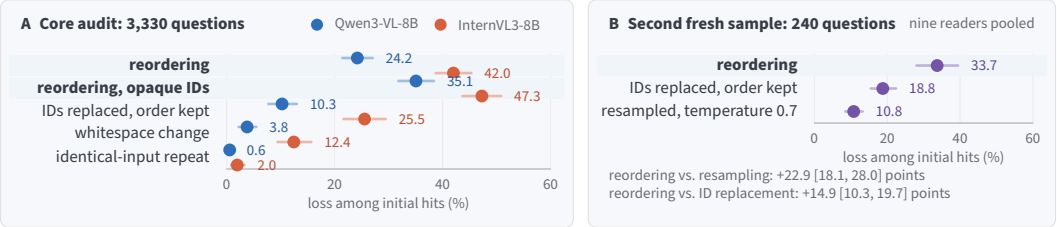

Figure 3: Loss among initial hits by change to the presentation; shaded rows change the order; bars are 95% video-cluster intervals. **A**: core audit (Table 1; Appendix C.5). **B**: second fresh sample (Appendix J.2).

cating that retrieval-rank information in the identifiers helps readers preserve some of their original hits. The same qualitative pattern appears on the NExT-GQA test split, VideoZeroBench, fixed-answer evaluation, additional packet providers, and Qwen3-VL-32B (Table 4; Appendices C.5, E.1, K.6 and K.7). Taken together, these controls indicate that the turnover is associated with how fixed evidence is presented, rather than being explained by repeated inference alone.

Larger models can be more stable, but the effect is not universal. Within InternVL3, increasing model size from 8B to 78B improves both initial grounding quality and retention. On the matched 350-question cohort, the 78B reader hits 53.7% of questions with an available correct candidate, compared with 47.1% for the 8B reader, and reduces initial-hit loss from 43.9% to 34.0% (Appendix C.6). The reduction is substantial, but a third of initially correct citations are still lost. The same monotonic pattern is not established across model families: Qwen3-VL-32B does not show a corresponding reduction relative to Qwen3-VL-8B (Appendix H). Retention loss is therefore not confined to smaller models.

Turnover persists across the full order space. An exhaustive audit over all 24 orders leads to the same conclusion (Figure 2; Appendix D.3). Across 120 stratified questions, Qwen3-VL-8B loses 26.3% of its initial temporal hits, while two identical-input repeats produce no selected-source changes. On the all-question denominator, losses exceed gains (4.55% versus 1.74%), but the more important observation is that order changes repeatedly alter which questions remain correctly grounded. The effect is not limited to isolated or fragile hits: 85% of initial hits come from questions that are correct in at least half of the 24 orders, and those questions still lose 17.3% of them.

Observed loss is also far below the 67.5% expected under uniform replacement. The reader therefore preserves prior hits far more often than a random selector, but retention remains incomplete.

Retrieval rank explains part of the instability, not all of it. Provider order carries retrieval-rank information, and readers make use of that cue. Reorderings that move the top-ranked clip away from the first position lose 29.2% of initial hits, compared with 16.0% when it remains first (Appendix C.7). Position therefore accounts for an important component of the effect. It does not, however, eliminate the phenomenon: even when the top-ranked clip stays first, readers still lose roughly one sixth of their initial hits.

The output analysis points in the same direction. Reordering an otherwise unchanged citation list accounts for only 5.5% and 3.6% of source flips; most source changes involve a different cited set (Appendix C). The instability therefore reflects changes in which evidence is selected, not merely changes in the ordering of an otherwise unchanged citation list.

The same metric blind spot appears in text RAG. The effect is smaller on HotpotQA, where grounding accuracy is much higher, but the same pattern remains. Across 6 text readers, the original order cites a gold paragraph on 79.7–94.0% of questions, yet reordering loses 4.0–12.6% of those initial hits while net hit-rate changes remain within 1.9 points of zero (Table 36). Identical-input repeats lose at most 0.7%. Alternative citation definitions change the magnitude but do not remove it, and a fresh disjoint 1,000-question sample again satisfies the pre-specified turnover criteria for all 6 readers (Appendices G and J.3). Loss–gain cancellation therefore appears under markedly different grounding regimes.

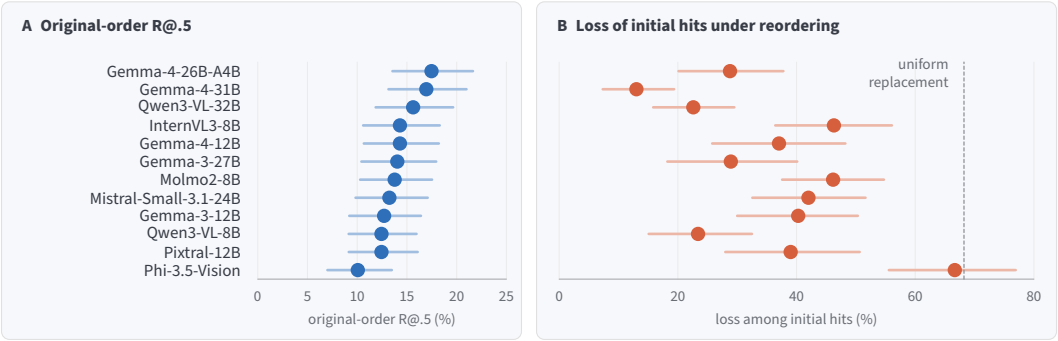

Figure 4: Video readers on identical packets (240 NExT-GQA validation questions, 186 videos; 12 readers). **A**: original-order $R@.5$ (sort order). **B**: loss among initial hits over three reorderings; dashed: uniform replacement. Bars: 95% video-cluster intervals. Shared-hit pairwise comparisons: Table 39.



Figure 5: Split-half reliability of reader rankings: the videos (text: questions) are split into two halves at random, readers are ranked by aggregate quality in each half, and the Spearman correlation between the two rankings is averaged over 1,000 splits; the horizontal axis gives the set size before splitting. One order uses the original presentation; four orders average each question over the original and three reorderings. Exploratory reanalysis of existing responses; intervals and twelve-reader values in Appendix I.

Reordering induces more turnover than resampling or relabelling. A second fresh video sample separates order sensitivity from more general output variability. Across 9 readers, reordering loses 33.7% of initial hits, compared with 10.8% under temperature-0.7 resampling of the original presentation and 18.8% under identifier relabelling (Figure 3B; Appendix J.2). The corresponding paired differences are +22.9 and +14.9 points. Stochastic generation and identifier changes therefore account for part of citation instability, but changing the evidence order produces a substantially larger loss of previously correct citations.

5 Comparing Readers Under Presentation Variation

Presentation variation affects not only whether an individual citation survives, but also how reliably readers can be compared. We evaluate 12 open VLMs from six model families on the same 240 packets, prompt, scorer, and evidence presentations (Appendix H). Original-order $R@.5$ ranges from 10.1% to 17.5%, whereas loss among initial hits ranges from 13.0% to 66.7% (Figure 4). These measures capture related but distinct properties: aggregate quality measures how often a reader is correct, whereas retention measures whether its successful grounding decisions persist across presentations.

Presentation averaging brings aggregate quality into closer agreement with retention. Retention agrees only moderately with original-order aggregate quality (Kendall's $\tau_b = 0.33$) but much more closely with quality averaged over the same four presentations ($\tau_b = 0.76$); the same qualitative relationship appears on a fresh sample (Appendices J.1 and I). The stronger agreement after averaging suggests that much of the discrepancy reflects presentation-specific variation in single-order scores, rather than a wholly separate reader property.

Single-order rankings can be unreliable on small evaluation sets. Splitting 120 video questions into halves of 60 gives single-order split-half Spearman correlations of 0.38 and 0.58 across the two video samples. Averaging each question over four presentations raises these values to 0.69 and 0.80 (Figure 5). The benefit decreases as the number of labelled questions grows and is negligible at 1,000 text questions.

A fresh, pre-specified comparison further clarifies the source of this gain. Reliability increases as presentations are added, but four presentations on 120 questions do not match one presentation on 240 questions (0.74 versus 0.82). Averaging over identifier relabellings or temperature samples yields reliability similar to averaging over reorderings (Appendix J.2). The reliability gain therefore appears to come from averaging over presentation-specific variation, rather than from reordering itself.

The same distinction determines when repeated presentations justify their inference cost. At a fixed budget of 240 video calls, evaluating 240 different questions once separates more reader pairs than evaluating 60 questions under four presentations (33% versus 25%; Appendix I). Repeated presentations are therefore most useful when labelled questions, rather than inference calls, are the limiting resource.

Averaged quality and paired retention serve different purposes. Shared-hit retention can distinguish reader pairs that a single-order aggregate leaves unresolved, but an aggregate averaged over the same presentations separates at least as many pairs overall (Appendices J and I). Presentation-averaged quality is therefore better suited to reader comparison, whereas paired retention measures whether grounding decisions that were already correct are preserved across presentations. Retention is best interpreted as a citation-invariance diagnostic alongside averaged quality, rather than as a replacement ranking statistic.

6 From controlled reordering to operational consequences

Position systematically affects evidence selection. Controlled placement experiments reveal a strong dependence on where evidence appears. Holding candidate membership fixed and moving one shared temporal-hit clip through the packet changes its selection rate sharply: placing it first rather than last raises selection by 49.8 and 30.4 points for InternVL3-8B in four- and eight-candidate packets, and by 27.7 and 19.6 points for Qwen3-VL-8B (Appendices D.10–D.11). The same directional effect appears in text: when a permutation moves the initially cited gold paragraph later rather than earlier, loss rises from 5.6% to 10.2%, with the effect reproduced on a fresh sample (Appendix J.3).

The controlled intervention can arise from a pipeline configuration change. Reordering need not be introduced deliberately by an evaluator. In the studied retrieval pipeline, changing only the feature-extraction batch size from 2 to 8 changes retrieval scores enough to reorder 2.83% of packets while preserving their membership. Among the 80 reordered pairs in which both citations resolve, the selected source changes 27.5% of the time. Extrapolating that conditional rate to the observed reorder frequency gives approximately 0.78% of all questions in this configuration (Appendix K.10).

The 0.78% figure is configuration-specific and is not an estimate of production prevalence. The more general result is that an implementation-level change can alter which citation a user receives even when the available evidence set remains unchanged.

Stabilizing input order does not by itself preserve previously correct citations. Canonical sorting eliminates order variation by committing to a deterministic presentation, but it does not guarantee preservation of citations that were correct under the previous order (Appendices K.1 and K.12). We also evaluated five mitigation strategies, including multi-call voting and retention-oriented fine-tuning; none established a reduction in initial-hit loss under its stated comparison (Table 59). These results separate three objectives that are often conflated in grounded systems: deterministic presentation, high average grounding quality, and preservation of previously correct citations.

7 SCOPE AND LIMITATIONS

PCRA measures citation invariance under a specified grounding criterion; in this paper, that criterion is agreement with benchmark evidence annotations. It does not by itself establish whether a temporal hit-to-miss transition corresponds to a loss of human-perceived evidential support. Automatic support judgments on 451 answer-preserving citation changes provide mixed evidence: Gemma-4-31B finds a higher support-loss rate among temporal-loss pairs, whereas Pixtral-12B does not, and agreement between the two evaluators is modest ($\kappa = 0.39$; Appendices K.14 and K.15). We therefore treat annotation retention and model-assessed support as distinct constructs; human evidence-sufficiency judgments remain unmeasured (Appendix K.16).

Several analyses have additional limitations. Test annotations had been accessed before the reported follow-ups, although no audit design choice used test outcomes (Appendix A). Some reader-panel and reliability analyses reuse previously studied validation data and are exploratory, and the 78B comparison covers one reader from a single model family and does not isolate parameter scale from serving configuration. Protocols described as fixed in advance were hash-recorded locally rather than registered externally. The central results in Table 1, the fresh perturbation comparison, the reader panel, and the 78B audit use fresh inference with corrected inputs; Appendix M maps historical pipeline defects to the exclusions, corrections, or reruns used for the reported analyses.

8 CONCLUSION

Stable aggregate grounding does not imply instance-level reproducibility. Marginal grounding scores can remain nearly unchanged while previously annotation-correct citations are lost and new hits appear on other questions. Because marginal scores discard the pairing across presentations, they cannot determine whether the same correct citations are preserved.

PCRA directly measures this paired quantity. Across video and text grounding, fixed-evidence representation reveals a common pattern: losses of previously correct citations are partly offset by gains elsewhere, so small net changes in aggregate quality can mask substantial citation turnover. Citation turnover is much larger than identical-input variation, persists across exhaustive order changes and diverse readers, and remains even when retrieval-rank cues are partly preserved. We also show that altering one retrieval-pipeline configuration can reorder the same evidence and change the resulting citation, providing an operational route to the controlled intervention.

The results also clarify the role of presentation variation in model evaluation. Averaging grounding quality across presentations reduces single-presentation variability and improves the reliability of reader rankings when labelled questions are scarce. Paired retention addresses a complementary question: whether citations that were already correct remain correct after the evidence is represented. It is therefore best viewed as a citation-invariance diagnostic alongside, rather than in place of, aggregate quality.

For label-limited grounding evaluation, we recommend reporting presentation-averaged quality together with paired retention and identical-input repeat controls. Our measurements concern agreement with benchmark evidence annotations, not established human evidence sufficiency; whether evidential support itself is preserved remains a separate evaluation target. **Grounding evaluation should therefore assess not only how often citations are correct, but also whether correct citations remain correct when the same evidence is re-presented.**

AI USE STATEMENT

For this revision, generative AI assistance was used to review the manuscript, refine the interpretation and design of follow-up analyses, implement analysis and evaluation code, and edit the text. AI-assisted artifacts were subjected to numerical consistency, input-integrity and document-build checks. No human audit labels were generated or replaced with AI outputs. The authors are responsible for the submitted claims, code and disclosure of AI use.

REFERENCES

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

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, 2024.

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Ethan Bito, Yongli Ren, and Estrid He. Position bias undermines preference consistency in listwise LLM-based reranking. *arXiv preprint arXiv:2608.03091*, 2026. URL https://arxiv.org/abs/2608.03091.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems*, 2024.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and iterate: Reducing prediction churn. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/dc5c768b5dc76a084531934b34601977-Abstract.html.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Grunde-McLaughlin_AGQA_A_Benchmark_for_Compositional_Spatio-Temporal_Reasoning_CVPR_2021_paper.html.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. Changing answer order can decrease MMLU accuracy. *arXiv preprint arXiv:2406.19470*, 2024.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. *arXiv preprint arXiv:2508.13144*, 2025.

Ji Huang, Barry Devereux, and Hui Wang. Your VLM already knows when: Training-free temporal grounding by asking yes or no. *arXiv preprint arXiv:2608.08315*, 2026a. URL https://arxiv.org/abs/2608.08315.

Linpeng Huang, Weixing Chen, Zexin Chen, Yang Liu, and Liang Lin. EG-VQA: Benchmarking verifiable video question answering with grounded temporal evidence. *arXiv preprint arXiv:2606.24797*, 2026b. URL https://arxiv.org/abs/2606.24797.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. VideoRAG: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21278–21298. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1096. URL https://aclanthology.org/2025.findings-acl.1096/.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. HAGRID: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*, 2023. URL https://arxiv.org/abs/2307.16883.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8211–8225. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.730. URL https://aclanthology.org/2020.acl-main.730/.

Yiyang Li, Yonghuang Wu, Ying Luo, Liangtai Sun, Zishu Qin, Lin Qiu, Xuezhi Cao, and Xunliang Cai. Instance-level randomization: Toward more stable LLM evaluations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3411–3425, 2025.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.467. URL https://aclanthology.org/2023.findings-emnlp.467/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. MUSEG: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35549–35561. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1644. URL https://aclanthology.org/2026.acl-long.1644/.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

Jiahao Meng et al. VideoZeroBench: Probing the limits of video mllms with spatio-temporal evidence verification. *arXiv preprint arXiv:2604.01569*, 2026. URL https://arxiv.org/abs/2604.01569.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. MoReVQA: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2404.06511.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12, 2024.

Roger B. Nelsen. *An Introduction to Copulas*. Springer, 2 edition, 2006. doi: 10.1007/0-387-28678-0.

Jingjie Ning and Xueqi Li. Same agent, different answers: A repeat-aware audit of corpus-induced answer churn in retrieval-augmented QA. *arXiv preprint arXiv:2608.22856*, 2026.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 708–718, 2020. URL https://aclanthology.org/2020.findings-emnlp.63/.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. TimeSearch-R: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gqb1hvuGcj.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Punitha Ponnuraj. Document overlap is not evidence continuity: Measuring retrieval jitter in citation-based RAG evaluation. In *Proceedings of the Workshop on Evaluating Evaluations (EvalEval)*, pp. 221–226, San Diego, CA, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.evaleval-1.35. URL https://aclanthology.org/2026.evaleval-1.35/.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. LongVideo-R1: Smart navigation for low-cost long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2602.20913.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2410.17434.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4c88827decab6c046b881a2c3a99c76f-Abstract-Conference.html.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14918–14937, 2023. URL https://aclanthology.org/2023.emnlp-main.923/.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 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 (Volume 1: Long Papers)*, pp. 2327–2340, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.129. URL https://aclanthology.org/2024.naacl-long.129/.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2502.21271.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Rahul Vishwakarma, Shushant Kumar, and Ratnesh Jamidar. What gets cited: Competitive GEO in AI answer engines. *arXiv preprint arXiv:2605.25517*, 2026. URL https://arxiv.org/abs/2605.25517.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. Correctness is not faithfulness in RAG attributions. *arXiv preprint arXiv:2412.18004*, 2024. URL https://arxiv.org/abs/2412.18004.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 959–975, 2025a. doi: 10.18653/v1/2025.findings-emnlp.50. URL https://aclanthology.org/2025.findings-emnlp.50/.

Qiuchen Wang et al. VimRAG: Navigating massive visual context in retrieval-augmented generation via multimodal memory graph. *arXiv preprint arXiv:2602.12735*, 2026a. URL https://arxiv.org/abs/2602.12735.

Shijie Wang, Honglu Zhou, Ziyang Wang, Ran Xu, Caiming Xiong, Silvio Savarese, Chen Sun, and Juan Carlos Niebles. Evidence-backed video question answering. *arXiv preprint arXiv:2607.11862*, 2026b. URL https://arxiv.org/abs/2607.11862.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2403.10517.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b. URL https://arxiv.org/abs/2405.19209.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. Ground every sentence: Improving retrieval-augmented LLMs with interleaved reference-claim generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 969–988. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.55. URL https://aclanthology.org/2025.findings-naacl.55/.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xiao_Can_I_Trust_Your_Answer_Visually_Grounded_Video_Question_Answering_CVPR_2024_paper.html.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14299–14308, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yan_Positive-Congruent_Training_Towards_Regression-Free_Model_Updates_CVPR_2021_paper.html.

Jiayu Yao, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Yuyao Ge, Zhecheng Li, and Xueqi Cheng. 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*, pp. 15183–15193, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.768. URL https://aclanthology.org/2025.emnlp-main.768/.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-Voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.03226.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hong-Wei Zheng, and Zhiming Zheng. Stable-RAG: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25907–25926, 2026.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.18079.

Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2408.16219.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QmZKc7UZCy.

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. URL https://arxiv.org/abs/2504.10479.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1OQKqLFcbB.

Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62892–62913, 2024.

APPENDIX

READING GUIDE

Part I contains the evidence for each claim in the main text, in the order of the main text. Analyses from earlier versions that the main text does not rely on are in a separate archival supplement. Table 2 maps each claim to its sections.

PART I: EVIDENCE FOR THE MAIN TEXT

A RECORDED EVALUATION PROTOCOLS

The packets, readers and test split of the audits come from an earlier study of a retrieval-side clip selector, whose protocol this section records; the selector itself is not part of this paper (archival supplement). Development used 3358 temporally labelled NExT-GQA validation questions over 567 videos. The gate required a positive 95% CI lower bound on mIoU or $R@1_{\mathrm{IoU}=.5}$ relative to `retrieval_top1_full`, together with no significantly negative change in joint containment. This second condition tests for detected harm; a non-inferiority test would require a specified margin. After the gate passed, we recorded the validation intervals as predictions and named the two

Claim in the main text	Where the evidence is
Data access, recorded protocols and execution order (§3.2)	Appendices A and B
Core video losses, unchanged answers, cited-set changes (§4)	Appendices C, D and E
Exhaustive 24-order audit and reference orders	Appendix D.3
Further providers and Qwen3-VL-32B	Appendices K.6 and K.7
Text replication and text position effect	Appendix G
Reader panel, shared-hit comparisons (§5)	Appendices H and K.8
Ranking reliability, information matching, separation	Appendices I, J.1 and J.3
Fresh-sample confirmations (video, text)	Appendices J.1 and J.3
Order specificity and presentation averaging	Appendix J.2
Controlled clip position (§6)	Appendices D.10 and D.11
Natural trigger and canonical ordering	Appendices K.10, K.1 and K.12
Mitigation attempts	Appendices L and D.13
Human audit status (§7)	Appendix K.16
Integrity issues, reproducibility, generated numbers	Appendices M, N and O

Table 2: Reading guide to the appendix.

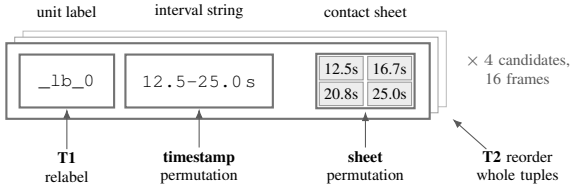

Figure 6: What a candidate tuple is made of, and what each control moves. Every delivered unit reaches the reader as a label, an interval string, and a contact sheet whose frames carry their own rendered times. T1 renames the clips and T2 changes the order they are shown in; both keep each clip's frames and interval together, so the evidence itself does not move (§C.3). The two permutation controls instead split the tuple apart (§S10.7).

readers. The answer-linked test follow-up ran once on 5553 test questions over 990 videos with the same geometry, prompts and decoding settings. Validation intervals quantify the validation estimates; their agreement with test effects is descriptive. Follow-up settings were recorded before their respective runs on already-evaluated data.

Data access and follow-up runs. Earlier exploratory work inspected test annotation files before this protocol was designed. The local artifacts record protocol freezing before test videos, features, packets, predictions and scores were produced. We did not use test examples for design, hyperparameter or model selection. The primary answer-linked evaluation in §3.2 was scored once.

The interface study (Appendix K) re-evaluates a recorded sample of the opened split. It rebuilds contact sheets with the corrected renderer (Appendix N.1) and generates fresh answer-linked outputs. Fixed-answer interfaces then use each reader's saved base answer and the same rebuilt sheets. The second-retriever study rebuilds the packets themselves (Appendix K.5). For Qwen3-VL-8B, answers from the primary and interface studies agree on 91.4% of 859 comparable questions. These follow-ups use newly generated outputs on rebuilt inputs (Supplement S10.3).

The full selector-gate record and validation/test comparison appear in *supplementary development report*, Section D7.

B PROTOCOL RECORDS AND EXECUTION ORDER

The protocol files and their SHA-256 prefixes: test protocol `e7ff324c`, cost-matched protocol `7d1ef1e1`, capability-ladder protocol `e68d8978`, main experiment protocol `014d0892`. Filesystem modification times place the test protocol at 14:01, the test packet at 15:36, the first test reader output at 17:57 and the first test score at 19:49 on the same day, in that order.

These timestamps record local, mutable file history. The workspace has no version-control history or externally timestamped registration. The hashes identify the retained protocol files; Appendix A records prior annotation access and the evaluation sequence.

C CORE-AUDIT DETAILS

This section retains the core-audit results that earlier versions reported in the main text.

C.1 RETENTION ACROSS DATASETS AND INTERFACES

At a 1,024-token output cap, neither reader reaches the limit in the fresh paired control, yet initial-hit loss remains 19.7% for Qwen and 33.3% for InternVL (Appendix E.1).

C.2 CANCELLATION WITHIN INDIVIDUAL QUESTIONS

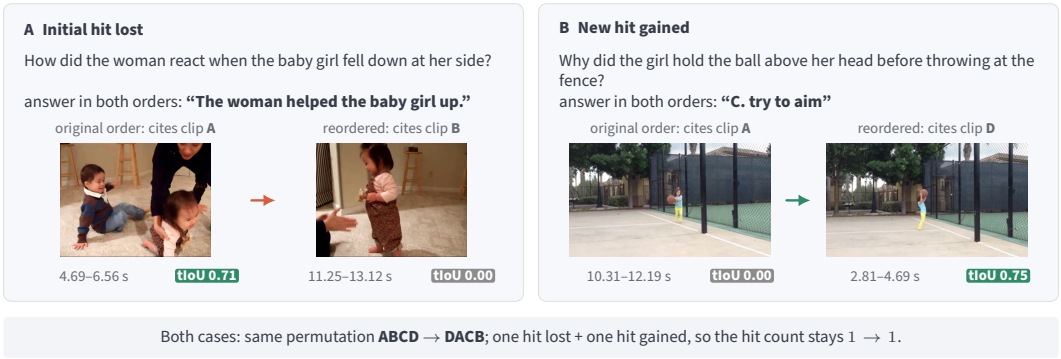

Figure 7: Actual Qwen3-VL-8B outputs with unchanged answer strings under fixed-evidence reordering. Each case (**A**, **B**) shows the source cited under the original order (left) and under the reordering (right), with one served frame per source. Hits satisfy $\mathrm{tIoU} \geq .5$ on mapped source intervals. Full frames, annotations and provenance appear in Figure 10 and Appendix D.14.

Net change balances losses against gains. Figure 7 illustrates cancellation: the same reordering loses one hit and gains another without changing either answer. In the earlier separate-run core audit, the net $R@.5$ changes are $+0.13$ $[-0.60, +0.84]$ and -0.27 $[-1.20, +0.70]$ points, with both CIs including zero.

Lost hits move well beyond the threshold. Lost citations drop by 51.4 and 54.2 tIoU points on average. The continuous drops, zero-overlap replacements and threshold sweeps characterize their movement (Appendix D.1). Lowest-rank reading recovers only a small share: paired reductions on the common initial-hit bases are $+2.43$ $[+1.37, +3.64]$ and $+2.15$ $[+1.22, +3.18]$ points (Table 10).

C.3 CITATION IDENTITY AND CITED-SET CHANGES

Under tuple reordering, signed ΔmIoU is -0.22 $[-0.69, +0.26]$ for Qwen3-VL-8B and -0.56 $[-1.19, +0.08]$ for InternVL3-8B; both 95% CIs include zero. Yet their citations change source on 40.2% and 59.0% of pairs, with mean absolute movements of 6.91 and 10.46 points in the earlier separate runs (Table 35). Table 35 separately tests relabelling, which changes retrieval-rank cues in clip IDs.

The emitted set changes even when the answer is unchanged. We decompose answer, list-order and cited-set changes across 30222 stored records (0 missing; Table 13). With both answer and set fixed, list reordering accounts for only 5.5% and 3.6% of source flips. Most flips preserve the answer while changing the set, whose mean Jaccard overlap is 0.16 and 0.10 in that cohort.

Reading the lowest original retrieval rank removes flips for an unchanged set. Among changed sets, source flips remain at 78.1% and 87.7%. Even with an unchanged answer and a temporal-hit candidate available, the set changes on 44.3% and 62.9% of pairs. Citation-list order therefore explains only part of the movement.

C.4 Reference orders in the exhaustive audit

Across the 24 reference orders, loss estimates range from 13.6% to 31.0%, each conditioned on that order's initial hits; the original-order estimate is toward the upper end. On this 120-question exhaustive sample, choosing two distinct orders uniformly and pooling their transitions gives 23.2 $[16.6,\ 30.6]$% loss among reference hits (Appendix D.3).

Excluding 4 questions with historical index mismatches leaves primary $\tau = .5$ initial-hit counts, loss counts and mean drops unchanged (Appendix D.1).

C.5 Single-run core audit

Why. In the first core runs, original-order and reordered responses came from separate serving runs (the Qwen original run was also resumed midway), and Qwen3-VL-8B's original order had 9.62% parse failures against 4.90% for reorderings (Table 31). Every video packet also names clips by retrieval rank, so a reordered packet shows slot order and identifier rank that disagree. **Design** (recorded and hash-frozen before inference). All 3,330 eligible NExT-GQA validation questions (561 videos; 28 excluded for frame decoding), four LanguageBind clips, corrected contact sheets, greedy decoding and a 1,024-token cap. Ten arms per question: the original order, three seeded reorderings and an identical-input repeat with rank-bearing identifiers; the original order with one whitespace change per clip block (`caption:` to `caption:`); and the original and the same three reorderings with opaque identifiers (one random mapping per question). Figure 6 shows how a reordering moves each clip's frames and interval together. All 33,300 requests per reader were shuffled into one queue served by identical vLLM servers (panel settings with 16 concurrent sequences; InternVL3-8B in its `-hf` conversion, as in the panel). The InternVL run was stopped and resumed twice to add cards; no request was sent twice. **Criterion** C1: under opaque identifiers, reordering loses more initial hits than the whitespace change (interval above zero); C1 is met for Qwen3-VL-8B and met for InternVL3-8B, and the same holds against the repeat (Table 3). Rows marked exploratory were not recorded criteria. Fallback rates differ between the original and reordered arms by $+0.4$ $[-0.1,\ +0.8]$ and -0.3 $[-0.7,\ +0.0]$ points, so the separate-run asymmetry does not recur.

Loss among initial hits (%)	Qwen3-VL-8B	InternVL3-8B
Reordering, rank-bearing identifiers	24.2 [21.4, 27.1]	42.0 [38.7, 45.3]
Reordering, opaque identifiers	35.1 [31.8, 38.4]	47.3 [43.7, 50.8]
Identifiers replaced, order kept	10.3 [7.7, 13.0]	25.5 [21.7, 29.5]
Whitespace change	3.8 [2.2, 5.5]	12.4 [9.4, 15.7]
Identical-input repeat	0.6 [0.0, 1.3]	2.0 [0.8, 3.3]
Rank-bearing minus opaque	-10.9 [-13.9, -7.9]	-5.3 [-8.5, -2.0]
Opaque minus whitespace (C1)	31.3 [27.6, 34.9]	34.8 [30.1, 39.4]
Opaque minus repeat	34.5 [31.2, 37.9]	45.3 [41.5, 49.1]
Opaque, top-ranked clip kept first (expl.)	13.5 [9.5, 17.8]	24.3 [18.5, 30.5]
Opaque, top-ranked clip moved (expl.)	41.4 [37.4, 45.3]	53.5 [49.3, 57.5]
Net $\Delta R@.5$, opaque identifiers	-0.10 [-0.99, $+0.83$]	-1.13 [-2.08, -0.19]
Fallback, original / reordered (%)	6.0 / 5.6	1.2 / 1.5
Parse failures, original / reordered (%)	0.1 / 0.0	1.0 / 1.3

Table 3: Single-run core audit. Brackets: 10,000 video-cluster bootstrap draws. Differences are paired on questions; each loss conditions on its own reference's initial hits.

C.6 Effect of model scale: InternVL3-78B

Design (recorded and hash-frozen before inference). InternVL3-78B (`-hf` conversion, bf16, vLLM pipeline parallel over six cards) received the identical inputs of the single-run audit (Appendix C.5)

Dataset	Reader	Answer-linked (A_0)		Fixed answer (A_2)
		All initial hits	Same answer	All initial hits
NExT-GQA *test*	Qwen3-VL-8B	26.0	26.2	40.9
	InternVL3-8B	47.3	47.7	48.0
VideoZeroBench	Qwen3-VL-8B	26.7	28.0	31.2
	InternVL3-8B	38.9	35.8	50.0

Table 4: Loss among initial temporal hits (%). A_0 regenerates answers and citations; "Same answer" restricts its initial-hit base to unchanged answers. A_2 receives a fixed answer. Table 12 reports full CIs, cohort sizes and common-base comparisons.

for questions whose packet contains a temporal-hit candidate, the only questions on which loss among initial hits is defined. The plan covered all 1,050 such questions and six arms; after 256 responses and before any analysis, throughput (0.07 requests per second on these cards) led to a recorded amendment: a seeded random 350 questions (249 videos) and four arms (original order, two reorderings, whitespace change); InternVL3-8B is compared on the same questions and arms from the single run. Operational changes during serving (memory layout, peer-to-peer settings, grouping each question's arms, and one restart after a worker ran out of memory) changed neither inputs nor decoding; all 6 amendments are listed with the released artifacts. **Criteria**: S1, that reordering loses more initial hits than the whitespace change, held; the recorded reporting rule's "loss persists" (S1 and a lower limit above 10%) held; "loss falls with scale" (8B minus 78B above zero) held.

350 questions with a correct candidate	InternVL3-78B	InternVL3-8B
Initial hit rate (%)	53.7 [48.2, 59.1]	47.1 [41.6, 52.8]
Initial-hit questions	188	165
Loss among initial hits (%)	34.0 [28.6, 39.5]	43.9 [37.9, 50.0]
Whitespace change (%)	5.9 [2.6, 9.7]	9.7 [5.5, 14.4]
Reordering minus whitespace	28.2 [22.1, 34.4]	34.2 [27.9, 40.6]
All-pair loss / gain (%)	18.3 / 12.7	20.7 / 18.3
Net $\Delta R@.5$	-5.6 [-10.6, -0.4]	-2.4 [-8.4, $+3.6$]
8B minus 78B loss; 78B minus 8B hit rate	9.9 [3.7, 16.0]; 6.6 [1.1, 12.2]	

Table 5: InternVL3-8B and InternVL3-78B on identical inputs. Loss averages two reorderings; rates are within questions that have a correct candidate (about a third of all questions), so net changes are larger here than over all questions. Brackets: 10,000 video-cluster bootstrap draws, paired across readers.

Table 5 gives every estimate.

C.7 Retrieval rank, question consistency and fallback

Exploratory reanalysis of stored outputs; no new inference. *Retrieval rank.* In the exhaustive audit (Qwen3-VL-8B, 120 questions, original order as reference), the 5 reorderings that keep the provider's top-ranked clip in the first slot lose 16.0 [9.0, 23.8]% of initial hits and the 18 that move it lose 29.2 [21.8, 37.0]% (paired difference 13.2 [7.3, 19.5] points, 10,000 video-cluster draws with inclusion weights). By Kendall distance from the provider order, loss is 16.1% for the 3 single adjacent swaps (16.1 [9.8, 23.0]) and 32.8% five swaps away; all 23 reorderings together give 26.3 [19.4, 33.6]%. Rank is therefore partly information the reader uses, and reorderings that preserve it still lose hits. *Question consistency.* Grouping questions by their hit rate over the 24 orders, 85 [75, 93]% of initial hits come from questions hit in at least half of the orders; these lose 17.3 [12.5, 22.5]%. Questions hit in at least three quarters of the orders provide 58% of initial hits, 14% of the loss and lose 6.3% themselves: losses concentrate on inconsistently answered questions but are not limited to near-chance hits. *Competence.* Across readers, the Spearman correlation between loss among initial hits and the initial hit rate is -0.52 and -0.83 for the two video panels and -0.94 and -1.00 for the two text panels. *Fallback.* Scoring unresolved citations as misses instead of the provider top-1 clip changes loss by at most 1.9 (recorded) and 0.7 (fresh) points for every reader except Pixtral-12B (22% and 17% unresolved responses; 15.5 and 11.9 points more loss); the reader rankings by loss correlate at 0.92 and 0.98.

Policy	Calls	Fixed order	Exploratory (120)		Fresh (240)	
			Loss (%)	mIoU	Loss (%)	mIoU
Original reference	1	—	0	20.25	0	19.02
Randomly reordered	1	No	26.16	19.02	18.88	19.19
Chronological	1	Yes	26.67	19.80	18.37	20.44
Answer–citation voting	6	No	19.25	20.42	16.84	19.50

Table 6: Qwen3-VL-8B policies within each weighted sample. Loss uses 60/98 initial-hit questions; mIoU uses all questions. Zero reference loss is self-comparison; fixed order does not guarantee repeatable outputs. Full CIs: Tables 29, 27, 28. Per-candidate scoring uses separate cohorts (Supplement S10.3).

C.8 Provider replication and candidate availability

FOCUS-core and VimRAG-derived adapters (Zhu et al., 2026; Wang et al., 2026a) reproduce 25.3–45.8% initial-hit loss across two readers and 200 validation questions, with no losses under identical-input repeats (Appendix K.6).

Retention differences on shared initial hits. On 240 questions, Gemma-4-31B exceeds InternVL3-8B by 3.5 $[1.3, 5.8]$ original-order mIoU points. Its $R@.5$ difference is 2.6 $[-0.5, 5.9]$ points, with an interval that includes zero. Among 40 questions initially hit by both readers, loss is 14.2% for Gemma and 40.8% for InternVL: a paired reduction of 26.7 $[13.3, 40.0]$ percentage points. PCRA quantifies preservation of shared initial hits alongside full-sample quality. Evaluation uses identical packets from 186 validation videos disjoint from the selector-development split. Reader family, size and processing vary together (Appendix K.8).

Candidate availability and reader selection. Low aggregate hit rates reflect both missing temporal-hit candidates and reader selection errors. In the full 3,358-question core audit, hit rates given a candidate with $\mathrm{tIoU} \geq .5$ are 49.6% for Qwen3-VL-8B and 42.6% for InternVL3-8B, versus 30.7% for uniform selection. In the separate matched-reader sample, 31.8% of the weighted population has such a candidate. Conditional hit rates are 53.3% for Gemma and 45.0% for InternVL, versus 30.0% for uniform selection (Appendix K.9).

Stored Qwen3-VL-32B outputs also lose 20.3% of initial hits on full validation (Appendix K.7); Phi-3.5-Vision's advantage over uniform replacement remains uncertain (Appendix D.6).

C.9 Serving policies, quality and call cost

Canonical sorting fixes input order without extra calls. Retrieval-rank sorting reverses artificial permutations; chronological sorting also handles rank changes when membership, IDs and images stay fixed (Appendices K.1 and K.12).

Auditing a fixed policy on fresh questions. Voting selects the plurality answer, then its plurality citation (Tang et al., 2024), using annotation-free schedules and fixed tie-breaking rules. Six calls reduce exploratory loss by 6.9 $[3.8, 10.0]$ points over one random-order call. Fixing this policy before 3,600 fresh responses on 240 questions from 207 new videos yields 2.04 $[-3.61, 7.58]$ points of reduction among 98 initial hits (Table 6). PCRA thus tests whether a serving change preserves hits beyond the exploratory sample; its benefit here remains uncertain.

Fresh mIoU gains remain uncertain (Appendix D.13). Identical-input repetition loses 1/98 hits, versus 18.9% under random reordering. The exhaustive audit's zero-repeat result does not generalize to this execution. Position-balanced voting, pairwise comparison and retention-oriented LoRA training likewise did not establish a loss reduction under their stated designs (Appendix L).

Serving choice and cost. Input invariance, quality and serving repeatability are distinct. Canonical sorting uses one call but can lose original hits. Voting benefits require target-sample validation; per-candidate scoring (A_3) uses n calls with variable quality (Supplement S10.3 and K.2).

Flip class	T1 relabel	T2 reorder
hit → hit	2.4%	2.5%
hit → miss	9.6%	9.0%
miss → hit	9.9%	9.3%
miss → miss	78.2%	79.2%
flip events	2667	4047

Table 7: Qwen3-VL-8B source flips by threshold state before and after re-presentation, as a share of flip events. "hit" means the cited clip agrees with an annotated window at IoU $\geq .5$; it is not a claim that the clip is the semantically correct evidence.

Reader	Cohort	n	Signed Δ	Flip %
Qwen3-VL-8B	all	3358	-0.22 $[-0.69, +0.26]$	40.2
Qwen3-VL-8B	$A_{.5}$=1	1055	-0.48 $[-1.79, +0.82]$	37.3
Qwen3-VL-8B	render-validated	2553	-0.24 $[-0.85, +0.38]$	39.8
Qwen3-VL-8B	both resolved (pairs)	8227	-0.31 $[-0.85, +0.25]$	43.1
InternVL3-8B	all	3358	-0.56 $[-1.19, +0.08]$	59.0
InternVL3-8B	$A_{.5}$=1	1055	-1.64 $[-3.32, +0.11]$	58.3
InternVL3-8B	render-validated	2553	-0.59 $[-1.39, +0.24]$	58.6
InternVL3-8B	both resolved (pairs)	9963	-0.56 $[-1.20, +0.08]$	59.3
resolver transitions under reordering, pairs / flips				
Qwen3-VL-8B	resolved → resolved	8227	3545 (87.6% of flips)	
Qwen3-VL-8B	resolved → fallback	299	84 (2.1%)	
Qwen3-VL-8B	fallback → resolved	768	418 (10.3%)	
Qwen3-VL-8B	fallback → fallback	780	0 (0.0%)	

Table 8: Tuple reordering by subgroup and by resolver transition. $A_{.5}$=1 is defined from the original packet and annotations alone, never from how good the post-transform choice was. "Render-validated" is the timestamp-covered cohort whose covering frames also passed the tile-level render check (Appendix N.1); it is a different set from the time-based cohort and is named separately. In the upper panel, n counts questions except for the rows explicitly marked as pairs; the lower panel counts resolver-transition pairs. Question-based rows average the fixed variants within each question; both-resolved rows average eligible pairs and resample their numerator and denominator together.

D WHAT CHANGES WHEN THE SOURCE CHANGES

Table 7 classifies every source flip by whether the intact and re-presented choices agree with an annotated window at IoU $\geq .5$. The classes measure temporal agreement only; a clip that meets the threshold is not thereby established as semantically the right evidence, so we do not label these classes valid and invalid.

The hit-to-miss and miss-to-hit rates describe changes in $R@.5$; their difference equals the change in that threshold metric when both use all question–variant pairs as the denominator. For mIoU, Table 35 reports signed and mean absolute score changes, distinguishing net movement from per-pair movement. The four classes below are conditional on a source flip and measure temporal threshold agreement, not semantic evidence validity. Most flips move between two clips that both fall below the threshold. The candidates also overlap: mean pairwise interval IoU between the intact and re-presented choice is 78.6 under relabelling and 65.3 under reordering, both on the 0–100 scale.

Table 8 opens the subgroups and the resolver transitions. The flip rate is close to flat across subgroups: restricting to questions where some delivered unit reaches tIoU ≥ 0.5, to questions whose covering frames passed the render check, or to pairs where intact and variant both produced a resolvable citation does not remove it. The pair-conditional row is a mean over eligible pairs, not a question-equal mean, and has no fixed three chances per question, so no any-variant rate is defined for it. Re-presentation also changes whether the reader produces a usable citation at all; of 780 reordering pairs where both sides fell back to the retrieval rank-0 source, none changed source (0 of 1706 such pairs across all three arms), which is the consistency the fallback path should have, and 87.6% of reordering flips occur between two resolved citations rather than through a fallback transition. Reordering serves the intact sheet files in a different order—all 10074 pairs reuse the same files and 0 keep the intact order—so it introduces no rendering difference of its own. Per-question, per-variant records are in `AR02`, `source_change_cohorts.csv` and `resolver_transition_summary.csv`.

Reader	Cohort	first resolvable		lowest rank in set	
		base	lost (%)	base	lost (%)
Qwen3-VL-8B	all pairs	1569	23.2 [20.6, 26.0]	1569	20.8 [18.2, 23.4]
	answer unchanged	1320	22.7 [19.8, 25.7]	1320	20.2 [17.5, 23.0]
	answer unchanged, both resolved	1270	22.7 [19.8, 25.6]	1270	20.1 [17.3, 22.9]
InternVL3-8B	all pairs	1347	42.4 [39.0, 46.0]	1347	40.2 [36.8, 43.8]
	answer unchanged	1259	42.0 [38.5, 45.7]	1259	39.8 [36.2, 43.5]
	answer unchanged, both resolved	1259	42.0 [38.5, 45.7]	1259	39.8 [36.2, 43.5]

Table 9: Retention by cohort and citation-reading rule. Each rule's base is the pairs *that rule* read as a hit on the intact packet, so the two halves are not a paired comparison; Table 10 gives that.

Reader	common base	first resolvable	lowest rank	paired reduction
Qwen3-VL-8B	1566	23.2 [20.6, 26.1]	20.8 [18.3, 23.5]	+2.43 [+1.37, +3.64]
InternVL3-8B	1347	42.4 [39.0, 46.0]	40.2 [36.8, 43.8]	+2.15 [+1.22, +3.18]

Table 10: Retention for both reading rules on the pairs both read as a hit on the intact packet. The reduction is paired—the same bootstrap draws are used for both rules—so its interval is the interval of the difference, not the gap between two independent intervals.

D.1 What a re-presentation costs in retained evidence

Supporting §4. The base is question–variant pairs whose *intact* citation already lay within $\mathrm{tIoU} \geq .5$ of an annotated window, so this is a retention rate and not the hit-to-miss share over all pairs. No new inference: every selection is read back from the stored T2 outputs, and every interval is a video-clustered bootstrap with the project's draws and seed.

All cohorts. Table 9 adds the both-resolved cohort to the two reported in the body. Restricting to pairs where intact and variant each resolved a citation changes little, so the loss is not an artefact of one side failing to produce a readable citation.

The two reading rules on a common base. We compare the rules on pairs that both initially read as temporal hits. Their full initial-hit bases differ (Table 9). For Qwen3-VL-8B, 3 pairs are hits only under first-resolvable reading and 3 only under lowest-rank reading. The corresponding counts for InternVL3-8B are 0 and 0. On the common base, lowest-rank citation reading reduces loss by $+2.43$ $[+1.37,\ +3.64]$ and $+2.15$ $[+1.22,\ +3.18]$ points (Table 10). These reductions remove about a tenth of Qwen3-VL-8B's loss and a twentieth of InternVL3-8B's.

Other thresholds, and how far the lost citations fall. A threshold sweep measures how retention varies with the correctness criterion. Continuous tIoU drops and replacement overlap measure whether lost citations make large temporal moves or remain near the threshold.

Every tIoU in this paper is recomputed from the original interval boundaries against the gold spans, so .3, .5 and .7 are decided by one arithmetic and rounding happens only at printing; the recomputation agrees with the flag stored during the runs on all but 9 of 20148 pairs. Those 9 are not boundary ties—none of them sits at the cut—and the cause is an index, not the arithmetic: the runs recorded a unit index against the four-unit packet while this recomputation reads bounds from the eight-unit rebuild, which orders units differently on a few questions. Counted directly, 66 of 75078 recorded readings do not reproduce their stored tIoU at the same index, on 4 questions; each stored value reappears in the packet at a different index, and 2 of those questions carry a reading that crosses .5, which is where the 9 pairs come from. We report the residue rather than resolving it, because the four-unit packet was not retained. An exploratory exclusion check removes all 4 affected questions from both readers and recomputes first-citation retention at $\tau = .5$. Neither the initial-hit counts, loss counts nor mean drops change: none of the excluded questions belongs to that reconstructed initial-hit cohort. The script verifies the full-cohort point estimates against the original exact-scoring artifact before comparing the exclusion (`AR03`). This bounds the discrepancy's effect on these conditional results; it does not recover the missing historical packet. Because the initially-correct cohort changes with the threshold, Table 11 is a sensitivity analysis and *not* a test that rules out boundary

Reader	Threshold	first resolvable base	first resolvable lost (%)	lowest rank in set base	lowest rank in set lost (%)
Qwen3-VL-8B	tIoU $\geq .3$	3114	18.8 [17.0, 20.5]	3114	17.0 [15.3, 18.8]
	tIoU $\geq .5$	1569	23.2 [20.6, 26.0]	1569	20.8 [18.2, 23.4]
	tIoU $\geq .7$	579	25.7 [21.6, 30.2]	582	21.0 [16.9, 25.2]
InternVL3-8B	tIoU $\geq .3$	2766	32.4 [30.0, 34.8]	2766	31.0 [28.6, 33.4]
	tIoU $\geq .5$	1347	42.4 [39.0, 46.0]	1347	40.2 [36.8, 43.8]
	tIoU $\geq .7$	486	50.4 [44.8, 56.1]	486	47.5 [41.8, 53.2]

Table 11: Retention at three thresholds, every tIoU recomputed from the original interval boundaries. The base is the pairs clearing that threshold on the intact packet, so it shrinks—and changes membership—as the criterion tightens. This is therefore a sensitivity analysis: it shows the effect is not particular to the threshold we report, and it cannot by itself exclude small movements across a boundary. The continuous drops in the text do that.

effects: a citation moving .701 $\rightarrow$ .699 is a loss at .7 and invisible at .3, so a rate rising with τ is consistent with small movements as well as large ones.

The magnitude of the failures is what settles that, and it is measured directly. A lost citation falls by 51.4 [48.8, 53.9] tIoU points on average for Qwen3-VL-8B and 54.2 [52.0, 56.3] for InternVL3-8B; 43.1% and 48.3% of them land on a clip that does not overlap the annotated window at all, and only 14.6% and 10.7% remain above $\mathrm{tIoU} = .4$.

The four later cells. Applying the same definition to the runs of Appendix K reproduces the effect on every cell. Table 12 is the full version of Table 4: it adds the A_1 arm, the base common to every arm in a cell, and the sample sizes. Two things have to be read off it carefully.

First, the arms do not share a track. A_0 and A_1 are *natural output*: the question is answered again under each ordering, so the answer can move with the citation. A_2 is *fixed answer*: one predicted answer is frozen and only the citation interface varies. The table records which is which. For the natural-output arms it also gives the cohort restricted to pairs whose answer did not change, which is the nearest like-for-like comparison available—though it is still not a fixed-answer experiment, because it conditions on an outcome rather than controlling an input. That restriction barely moves the rate (26.0 [21.4, 30.7]% to 26.2 [20.7, 31.9]% on NExT-GQA/Qwen3-VL, 47.3 [41.0, 53.6]% to 47.7 [41.2, 54.2]% on NExT-GQA/InternVL3), so the natural track's answer variation is not what produces its loss.

Second, because the arms select different intervals, an arm's own initially-correct pairs differ from another's, so the table gives both each arm's own base and the base common to the arms in that cell. Even on the common base, an A_2–A_0 difference spans two experimental conditions at once—the selection interface *and* the track—so we report it as descriptive and do not read it as an effect of the interface.

D.2 What changes inside the cited set

Supporting §C.3. A source flip reads one clip out of a citation list, so on its own it cannot say whether the reader re-ordered a fixed set of citations or cited different evidence. We recovered from the stored outputs the answer, the ordered citation list and the de-duplicated cited set for every question–variant pair, with no new reader inference (30222 records, 0 missing; missing rows are reported, never filled with a fallback). The flip rate recomputed this way reproduces the earlier-run flip rates of Table 35 (40.2% and 59.0%), which is the consistency check between the two pipelines.

Table 13 crosses the two things that can change. The matrix is defined on pairs whose answer and citations read back on both sides (8205 and 9963 of 10074); pairs where a side did not resolve, or where the answer or the set could not be read, are excluded and counted rather than folded into a cell. Reading the same outputs under an order-invariant rule—cite the member of the set with the lowest original retrieval rank—separates the two: it drives flips to exactly zero wherever the cited set is unchanged, because that is what the rule does, and leaves 78.1% and 87.7% in the cell where the answer is the same and the set is not; Table 14 gives the two readings over all pairs. Sorting the cited set removes one component of the instability and not the one that is moving.

			own base		common base		same answer	
Cell	Arm	Track	n	lost (%)	n	lost (%)	n	lost (%)
NExT-GQA / Qwen3-VL-8B	A_0 answer-linked	natural	172	26.0 [21.4, 30.7]	115	24.3 [18.9, 30.1]	142	26.2 [20.7, 31.9]
	A_1 lowest rank	natural	173	24.5 [19.9, 29.2]	115	22.0 [16.5, 27.7]	143	24.7 [19.3, 30.5]
	A_2 direct	fixed	158	40.9 [35.6, 46.4]	115	38.8 [33.0, 45.0]	—	—
NExT-GQA / InternVL3-8B	A_0 answer-linked	natural	143	47.3 [41.0, 53.6]	102	37.9 [30.7, 45.2]	137	47.7 [41.2, 54.2]
	A_1 lowest rank	natural	143	45.0 [39.0, 51.2]	102	36.9 [30.0, 44.1]	137	45.5 [39.1, 51.9]
	A_2 direct	fixed	132	48.0 [41.5, 54.4]	102	47.7 [40.6, 54.9]	—	—
VideoZeroBench / Qwen3-VL-8B	A_0 answer-linked	natural	30	26.7 [15.6, 38.7]	25	25.3 [13.0, 39.1]	28	28.0 [15.1, 42.3]
	A_1 lowest rank	natural	30	26.7 [15.6, 38.7]	25	25.3 [13.0, 39.1]	28	28.0 [15.1, 42.3]
	A_2 direct	fixed	32	31.2 [20.8, 42.2]	25	25.3 [14.7, 37.0]	—	—
VideoZeroBench / InternVL3-8B	A_0 answer-linked	natural	24	38.9 [25.0, 53.6]	16	33.3 [16.7, 51.0]	20	35.8 [20.0, 52.6]
	A_1 lowest rank	natural	22	36.4 [20.8, 53.0]	16	33.3 [14.6, 53.3]	19	37.7 [20.2, 56.5]
	A_2 direct	fixed	26	50.0 [38.5, 61.3]	16	43.8 [31.1, 57.8]	—	—

Table 12: Retention on the four cells of Appendix K. *Track* says what was held fixed while the run was made: A_0 and A_1 re-answer under each ordering, A_2 does not. "Own" uses each arm's own initially-correct pairs and "common" the pairs every arm read as a hit on the intact packet; "same answer" restricts the natural-output arms to pairs whose answer did not change, and is not available for A_2, which never varies its answer. Every arm loses initially-correct citations. A_2 loses more than A_0 on the common base, but the two differ in track as well as interface, so that gap is descriptive and not an interface effect. The VideoZeroBench bases are small and the intervals correspondingly wide.

	Qwen3-VL-8B					InternVL3-8B				
Pairs where…	pairs	first	canon.	Jaccard	$\lvert\Delta\rvert$	pairs	first	canon.	Jaccard	$\lvert\Delta\rvert$
answer same, set same	3897	5.7	0.0	1.00	0.8	3450	6.2	0.0	1.00	0.7
answer same, set changed	3634	80.9	78.1	0.16	14.3	5959	89.8	87.7	0.10	16.1
answer changed, set same	262	1.1	0.0	1.00	0.3	194	2.6	0.0	1.00	0.4
answer changed, set changed	412	90.3	89.1	0.08	16.1	360	93.9	93.6	0.05	18.2
same answer, $A_{.5}$=1	*2494*	*37.7*	*34.3*	*0.63*	*13.7*	*2949*	*58.5*	*55.7*	*0.43*	*21.6*

Table 13: Tuple reordering, decomposed by what changed in the stored output. "first" is the source flip rate under the paper's rule (first resolvable citation) and "canon." under an order-invariant rule (lowest original retrieval rank in the cited set), both as a percentage of the pairs in that row; Jaccard is the mean overlap of the intact and variant cited sets and $|\Delta|$ the mean absolute change in mIoU. The order-invariant rule is zero by construction wherever the set is unchanged, which is why the two set-changed rows carry nearly all of it. The last row is not a fifth cell but a cut through the first two: same answer, both sides resolved, and a delivered unit in the original packet reaching $\mathrm{tIoU} \geq 0.5$. Excluded from the matrix and counted instead: 1869 and 111 pairs where a side did not resolve or the answer or set could not be read.

The reordering changes which clips are cited, under an unchanged answer. The same-answer–set-changed cell holds 3634 and 5959 pairs, with mean set overlap 0.16 and 0.10 and per-pair movement of 14.3 and 16.1 mIoU points. Restricting further to pairs where the original packet contained a unit that could reach $\mathrm{tIoU} \geq 0.5$—so a good citation was available and the reader's answer did not change—the cited set still changes on 44.3% and 62.9% of pairs, with per-pair movement of 13.7 and 21.6 points against 6.9 and 10.5 over all pairs. That cut is not a fifth cell of the matrix and mixes the two same-answer rows, so it is not directly comparable with either. Cells in which the answer changed are a small minority (262+412 and 194+360 pairs) and are reported separately rather than mixed in.

These are descriptive rates over pairs. They say what moved, not that a changed citation set is a worse one: the matrix measures citation identity and temporal agreement, and §K.4 and Appendices K.3 and K.4 are where support itself is assessed.

D.3 Exhaustive coverage of four-candidate orders

Complete order coverage yields 40.8 [34.9, 46.5]% source flips and 26.3 [19.4, 33.6]% loss of initially correct citations, with no selected-source changes in the 240 identical-input repeat pairs (Table 15).

We evaluate every candidate order on 120 questions from 113 videos in the corrected 1,000-question NExT-GQA test follow-up. Qwen3-VL-8B receives all 24 permutations and two extra copies of the original order, for 3,120 fresh responses. Each tuple keeps its immutable clip ID, full interval and

Tuple reordering (T2)	Qwen3-VL-8B	InternVL3-8B
first resolvable citation (the paper's rule)		
source flip (% of all pairs)	40.2	59.0
of pairs whose set was unchanged (%)	5.49	6.04
signed ΔmIoU	-0.22 $[-0.69, +0.26]$	-0.56 $[-1.19, +0.08]$
lowest original retrieval rank in the set		
source flip (% of all pairs)	36.2	55.5
of pairs whose set was unchanged (%)	0.00	0.00
signed ΔmIoU	-0.13 $[-0.58, +0.33]$	-0.52 $[-1.14, +0.10]$
pairs whose set changed (%)	55.1	63.8
mean set Jaccard	0.52	0.42

Table 14: Two reading rules applied to the same outputs. Signed changes compare variant and intact inputs within each rule; their intervals include zero. The lowest-rank rule eliminates same-set flips, while changed cited sets account for the remaining source changes (Table 13). Each block reports rates over all question–variant pairs and over the unchanged-set subgroup separately. Per-pair records are in `research_upgrade_v6/p0/`.

four-frame contact sheet. The input audit verifies all 24 distinct orders and exact input equality for the repeats.

Sampling and estimation. Archived original-order outputs define two strata: 173 temporal hits at $\mathrm{tIoU} \geq .5$ and 827 misses, including unresolved outputs under the original-rank fallback. We sample 60 questions from each stratum without replacement, using seed 20260919. Inclusion weights are the stratum population divided by its sample size. Weighted estimates therefore target the recorded follow-up population. The reference for every contrast is a designated *fresh* original-order output; 60 sampled questions are temporal hits under this reference. We average the 23 alternative orders within each question and then apply the inclusion weights. Conditional rates divide weighted event counts by weighted eligible-pair counts. Intervals use 10,000 paired video-cluster bootstrap draws with fixed inclusion weights; all questions from each sampled video remain together. These are approximate weighted cluster intervals without a finite-population correction.

Serving and repeated inputs. All conditions for a question use the same GPU and replica configuration. We balance sampling strata over three identical Qwen3-VL-8B-Instruct replicas and shuffle each replica's request schedule before inference. The checkpoint revision is `0c351dd0`; decoding is greedy, with a 384-token cap and seed 20260919. The servers use vLLM 0.16.0, PyTorch 2.9.1 and transformers 4.57.6, with prefix caching disabled. Four audit requests per replica share an engine capacity of eight sequences with other work. The two original-order repeats measure variation under this shared serving configuration. Each of the 24 orders has one serving realization; extra repeats cover the original order. After the shared servers shut down with 2,019 responses saved, we restored the same checkpoints, software and server settings on the same GPUs and ports. The remaining 1,101 requests resumed from the frozen manifest; saved responses were retained and verified byte for byte.

Source changes and temporal-hit retention. Table 15 compares all alternative orders with the three schedules used in the earlier corrected follow-up, evaluated afresh on the same sampled questions. It also reports the two identical-input repeats. The paired permutation-minus-repeat difference is $+40.8$ $[+34.9, +46.5]$ percentage points for source flips and $+26.3$ $[+19.4, +33.6]$ points for temporal-hit loss. Among pairs with the same valid answer, the source-flip rate is 43.1 $[36.9, 49.0]$% and temporal-hit loss is 26.6 $[18.9, 34.8]$%. Restricting to two resolved citations gives 45.2 $[39.1, 50.9]$% and 26.8 $[19.3, 34.6]$%, respectively. Repeat source and temporal-hit changes have degenerate empirical bootstrap intervals because no such event was observed. The emitted cited set still changes on 5.3 $[1.5, 10.1]$% of repeat pairs.

Question-level coverage. At least one alternative changes the selected clip on 89.1 $[82.5, 95.0]$% of questions. The mean number of distinct selected clips across all 24 orders is 2.82 $[2.62, 3.03]$. Drawing two distinct observed orders uniformly gives 45.2 $[39.7, 50.6]$% source disagreement, which does not privilege the original order. Among fresh initial temporal hits, at least one alternative loses the hit on 75.0 $[64.3, 85.2]$%, and the mean worst observed drop is 46.3 $[39.1, 53.4]$

Measure	All 23 alternatives	Earlier 3 schedules	Identical-input repeats
Source flips (%)	40.8 [34.9, 46.5]	48.2 [40.0, 56.6]	0.0 [0.0, 0.0]
Temporal-hit loss (%)	26.3 [19.4, 33.6]	28.9 [20.5, 37.4]	0.0 [0.0, 0.0]
Signed Δ tIoU	−1.37 [−3.53, +0.72]	−1.21 [−4.19, +1.79]	+0.00 [+0.00, +0.00]
Absolute Δ tIoU	6.78 [5.03, 8.77]	7.78 [5.28, 10.77]	0.00 [0.00, 0.00]
Cited-set changes (%)	52.0 [44.8, 58.9]	55.4 [46.7, 64.1]	5.3 [1.5, 10.1]

Table 15: Exhaustive orders and identical-input repeats on the same 120 stratified questions. Entries are inclusion-weighted estimates with 95% video-cluster bootstrap intervals. Temporal-hit loss conditions on the fresh original-order citation satisfying $\mathrm{tIoU} \geq .5$. All comparisons retain unresolved responses through the original-rank fallback.

tIoU points. These any-event summaries have 23 opportunities; the per-order comparisons in Table 15 provide the matched comparison with two repeats.

Reference choice and symmetric temporal transitions. We repeat the conditional-loss calculation with each of the 24 observed orders as the reference, always comparing it with the other 23 orders. Each reference defines its own initial-hit cohort (41–60 questions). Loss ranges from 13.6% to 31.0%, with median 22.9%; the original reference ranks 19 th of 24 in ascending loss. This finite-design range describes reference sensitivity, not a sampling interval or a comparison on a common initial-hit cohort.

A symmetric comparison draws two distinct orders uniformly for each question. Let $M = 24$ and k_q count the orders producing a temporal hit. Averaging the ordered pairs within question gives reference-hit probability $p_q = k_q/M$ and loss probability $\ell_q = k_q(M - k_q)/(M(M - 1))$. Using the original inclusion weights w_q, pooled loss conditional on a reference hit and symmetric hit disagreement are

$$L_{\mathrm{sym}} = \frac{\sum_q w_q \ell_q}{\sum_q w_q p_q}, \qquad D_{\mathrm{sym}} = \frac{2\sum_q w_q \ell_q}{\sum_q w_q}.$$

Thus $D_{\mathrm{sym}} = 2\bar{p} L_{\mathrm{sym}}$, where $\bar{p}$ is the weighted mean of p_q. They summarize the same transitions with different denominators. Pooled conditional loss weights references by their hit mass; it differs from the equal mean of the 24 reference-specific conditional rates.

Pooled conditional loss is 23.2 [16.6, 30.6]%, symmetric hit disagreement is 6.8 [4.4, 9.5]%, and marginal hit rate is 14.6 [10.7, 19.3]%. All-pair loss and gain are each 3.4 [2.2, 4.8]%; retained hits and unchanged misses are 11.2 [7.9, 15.3]% and 82.0 [76.3, 86.8]%, respectively. Loss equals gain by construction under symmetric order sampling. Its zero net change is not an empirical invariance result for any particular serving-policy change. The original-order analysis measures loss of hits achieved by the recorded policy; the symmetric analysis characterizes the full observed order set.

This exploratory reanalysis uses the same 120 questions and 113 videos, fallback rule, inclusion weights and 10,000 video-cluster bootstrap draws. Orders remain together within each video; order pairs are not independent samples. All 24 reference estimates, initial-hit counts and nominal CIs are in `AR04`. Raw-response reconstruction and explicit pair enumeration reproduce these estimates.

How much does selecting three orders matter? The earlier three schedules exceed the full-sweep flip estimate by +7.4 [+2.1, +12.7] percentage points. Their temporal-hit loss differs by +2.6 [−2.7, +8.1] points. Full coverage confirms the phenomenon and refines its magnitude. There are $\binom{23}{3} = 1{,}771$ subsets of three alternative orders. Enumerating all of them gives source-flip rates from 16.7% to 52.0%, with a central 95% range of 26.7–49.6%. Temporal-hit loss ranges from 14.4% to 34.4%, with a central 95% range of 18.9–32.8%. These ranges describe the finite choice of three orders. The sampling intervals in Table 15 instead resample videos. Figure 8 displays both comparisons.

Outputs and reproducibility. The original-order references contain 12 unresolved outputs and 6 length-capped outputs out of 120. The alternative orders contain 236 and 140, respectively, out of 2,760; the repeats contain 20 and 9 out of 240. These counts can overlap. Every response contributes to the declared resolver, while both-resolved analyses isolate successful citation emission. The analysis plan and input hashes were saved before these fresh runs, after the earlier paper results

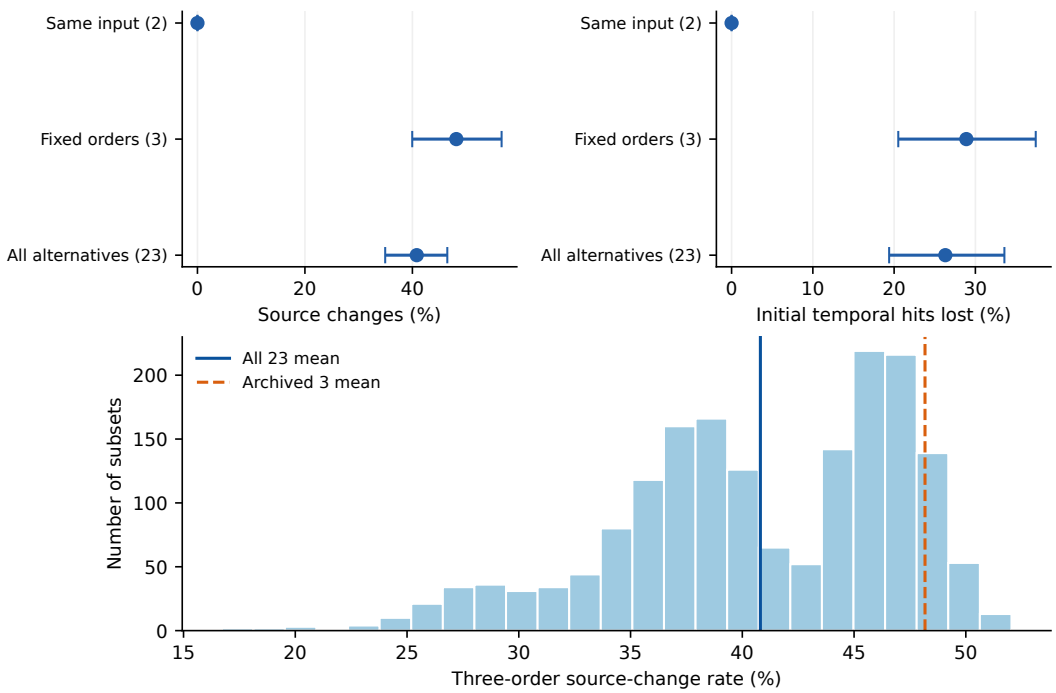

Figure 8: Complete order coverage with identical-input repeats. Top: weighted source flips and temporal-hit losses with 95% video-cluster intervals. Bottom: source-flip rates for all 1,771 three-order subsets; lines mark the mean over all 23 alternatives and the three earlier schedules evaluated afresh.

Displayed slot	Selected slot (%)	Initial-hit loss (%)	Reference source retained (%)
First	42.9 [38.1, 47.7]	17.1 [10.1, 25.0]	74.1 [64.9, 82.5]
Second	16.8 [14.9, 18.6]	28.1 [19.9, 36.6]	61.0 [52.4, 69.1]
Third	18.5 [16.3, 20.7]	30.1 [21.6, 39.0]	60.4 [51.1, 69.2]
Last	21.9 [19.4, 24.5]	28.1 [19.6, 37.0]	62.7 [53.3, 71.5]

Table 16: Exploratory position analysis of the same exhaustive responses. Selected-slot rates use all 24 orders on 120 questions. The other columns group 23 alternatives by the position of the original selected source on the same 60 initial-hit questions; slot denominators are 318, 356, 354 and 352 responses, normalized within question. Values use inclusion weights and 95% video-cluster CIs.

were known. Scripts, raw responses, per-question scores, all three-order subsets and figures are retained in `AR05`.

D.4 Display position in the exhaustive audit

The existing exhaustive outputs show a preference for the first displayed clip. Across all 24 orders, first-slot selection is 42.9 [38.1, 47.7]%, compared with 21.9 [19.4, 24.5]% for the last slot. Each of the four candidates occupies each slot exactly six times per question, so the comparison balances candidate identity and content. Table 16 reports the complete position profile.

Reference-clip placement and temporal-hit loss. We retain the 60 fresh original-order hits from 58 videos. For each question, we group the other 23 orders by the displayed position of its original selected clip. Each slot contains five or six alternatives; we average within question and slot before applying the original inclusion weights. This keeps the initial-hit question cohort identical across positions. Loss rises from 17.1% with the reference clip first to 28.1% with it last, a paired difference of +11.1 [+2.7, +19.6] points. On both-resolved pairs the difference is +11.4 [+3.3, +19.8] points. Selecting another clip can preserve the temporal hit, so source retention and temporal-hit retention differ.

Hit candidates m_q	Questions	Initial hits questions/videos	Reader loss (%)	Uniform loss (%)	Reader − uniform (pp)
All	120	60/58	26.3 [19.4, 33.6]	67.5 [64.3, 70.5]	−41.2 [−47.8, −34.1]
0	52	0/0	–	–	–
1	50	43/42	31.6 [23.3, 40.5]	75.0 [75.0, 75.0]	−43.4 [−51.7, −34.5]
2	17	16/16	12.0 [2.8, 23.9]	50.0 [50.0, 50.0]	−38.0 [−47.2, −26.1]
3	1	1/1	26.1*	25.0*	1.1*
4	0	0/0	–	–	–

Table 17: Exploratory candidate-availability analysis on existing Qwen3-VL-8B exhaustive outputs. Counts are unweighted; rates use sampling weights on the same initial-hit cohort. Intervals are nominal 95% paired video-cluster bootstrap intervals. Uniform loss is an exact expectation: its constant value within each m_q stratum gives a zero-width interval. *The three-hit-candidate stratum contains only one question/video; only its point estimates are shown. Dashes denote undefined retention, not zero loss.

Analysis status and reproduction. This is an exploratory reuse of already evaluated outputs, with 10,000 paired video-cluster bootstrap draws and nominal intervals. Each order has one observed serving realization. The placement comparisons describe behavior under complete tuple permutations, which also rearrange the other candidates; they do not isolate an attention-level mechanism. Unresolved citations retain the original-rank fallback. The artifact reports fallback-excluded selection, both-resolved loss, and a full-24 sensitivity that includes the guaranteed reference hit. It preserves every item–candidate–position selection count and input/output hash. `AR06` reproduces the published all-23 flip and loss estimates before computing these quantities.

Holding the other clips' relative order fixed. A matched-background reanalysis groups the 24 existing outputs by the relative order of non-reference clips and moves the original selected clip through all four positions. We exclude the entire background containing the original-order output, leaving five matched backgrounds per question without a guaranteed reference hit. On 60 initial-hit questions, first-to-last placement increases loss from 16.3% to 27.7%, a paired difference of 11.3 [2.7, 20.0] points. Among 43 questions with exactly one hit candidate, the difference is 15.8 [4.5, 27.0] points. These are exploratory, question-averaged contrasts on existing Qwen outputs with video-cluster intervals; `AR07` contains the reanalysis.

D.5 Random replacement and the number of hit candidates

Comparison on the same initial-hit cohort. We reuse the exhaustive audit's 120 questions and fresh original-order outputs. A hit candidate has $\mathrm{tIoU} \geq .5$; let m_q count such candidates among the four clips for question q. An independent uniform replacement selects each clip, including the original source, with probability $1/4$. Its expected temporal-hit loss is $1 - m_q/4$. We average this exact expectation over the reader's initial hits with the same inclusion weights w_q used for its observed loss:

$$L_{\text{uniform}} = \frac{\sum_q w_q H_{q0}(1 - m_q/4)}{\sum_q w_q H_{q0}}, \qquad L_{\text{reader}} = \frac{\sum_q w_q H_{q0} \frac{1}{23} \sum_{v=1}^{23}(1 - H_{qv})}{\sum_q w_q H_{q0}}.$$

This comparison preserves the reference hits rather than drawing a new initial-hit cohort for the random selector. Each question contributes its mean over the 23 alternative orders. The primary analysis retains the original fallback rule.

Results and interpretation. Across 60 initial-hit questions, reader loss is 26.3 [19.4, 33.6]%, compared with 67.5 [64.3, 70.5]% for uniform replacement; the paired difference is −41.2 [−47.8, −34.1] points (Table 17). With one hit candidate, reader loss is 31.6%; with two, it is 12.0%. The three-candidate result is a single-case description. These strata describe candidate availability and do not isolate its causal effect. The reader preserves temporal hits more often than independent random replacement, yet reordering loses some hits it had already achieved.

On both-resolved pairs, reader and uniform losses are 26.8 [19.3, 34.6]% and 66.7 [63.2, 69.9]%, respectively, with paired difference −39.9 [−47.0, −32.3] points. This sensitivity weights the random expectation by the same eligible-pair fractions as the observed loss. Over all sampled questions,

Quantity	Estimate [95% CI]
Original-order hit rate	10.0 [7.1, 13.3]%
Initial-hit loss (38 initial-hit questions)	57.9 [46.1, 69.3]%
Initial-hit loss, unchanged answer	56.0 [43.9, 67.7]%
Identical-input repeat loss	0.0 [0.0, 0.0]%
Source flip under reordering	71.4 [67.6, 75.3]%
Source flip under identical-input repeat	2.5 [0.6, 5.1]%
Signed tIoU change	−0.1 [−2.4, 2.1] pp
Net $R@.5$ change	−1.2 [−4.1, 1.4] pp

Table 18: Phi-3.5-Vision replication on the weighted validation sample. Reordering uses three paired alternatives per question. Loss conditions on original temporal hits; source flip and aggregate changes use all questions. Intervals resample videos.

weighted original-order reader hit rate is 17.3 $[12.6,\ 23.1]\%$, versus the uniform selector's expected 8.7 $[6.3,\ 11.6]\%$; these use an all-question denominator.

Alternative random policies have different meanings. A source sampled once and then reused across all orders retains every one of its initial temporal hits by construction, regardless of its absolute grounding quality. For independent uniform draws at both reference and alternative orders, retention instead equals $\mathbb{E}[(m_q/4)^2]/\mathbb{E}[m_q/4]$, conditioning on the random selector's own hits. Neither that quantity nor the reader's aggregate hit rate is the baseline in Table 17. The ledger and exact expectations are generated by `AR08`; source hashes and nominal 10,000-draw video-cluster intervals are retained. This is an exploratory reuse of existing outputs, with no new reader inference.

D.6 A further reader family: Phi-3.5-Vision

Phi-3.5-Vision (Abdin et al., 2024) also loses initial temporal hits under fixed-evidence reordering (Table 18). On the same 240 validation questions from 187 videos used for the packet-size follow-up, we run fresh inference with the original order, three reorderings and one identical-input repeat. The four-candidate packets, corrected contact sheets and prompt content are fixed. The pinned checkpoint uses its native chat template and four crops per image, greedy decoding, a 16,384-token context limit and a 1,024-token output cap. The artifact records the model revision, runtime, input hashes and all 1,200 outputs.

All 38 original hits have resolvable citations in their three alternatives, so both-resolved loss equals the primary estimate. The unchanged-answer estimate is 56.0 $[43.9,\ 67.7]\%$. Of 1,200 responses, 1 reaches the output limit and 1 uses fallback. Identical-input repeats change some source identities but lose none of the original hits. The signed tIoU and net hit-rate intervals include zero while conditional loss is substantial, extending the paired-retention finding to this additional reader family. Uniform replacement on the same initial hits has expected loss 63.2 $[59.3,\ 66.9]\%$; the observed-minus-uniform difference is -5.3 $[-16.4,\ 5.6]$ points. The common sample and weighting follow Appendix D.7; the complete analysis is in `AR09`.

D.7 Retention with four and eight candidates

Fresh paired design. We evaluate Qwen3-VL-8B and InternVL3-8B on the same 240 validation questions from 187 videos with nested LanguageBind top-4 and top-8 packets. The first four clips retain their identifiers, intervals and four rendered frames across packet sizes. Within each size, original order, three reorderings and one identical-input repeat give 4,800 fresh responses across readers. Eight-candidate permutations induce the relative order of the shared four clips; rejection sampling ensures three distinct, non-original permutations at both sizes. These are three coupled schedules, not exhaustive eight-candidate coverage.

The corrected frame bank supplies all eight candidates for 3,330 validation questions. We sample 120 questions from each of two strata: at least one $\mathrm{tIoU} \geq .5$ candidate in the top four (population 1,050), or none (2,280). Estimates retain inverse inclusion weights; this is an exploratory follow-up on previously accessed validation data. We verify source-frame hashes, sheet dimensions, interval

Reader	K	Initial-hit questions	Loss (%)	Both-resolved loss (%)	Source flip (%)	Repeat loss (%)
Qwen3-VL-8B	4	67	21.4 [14.6, 28.8]	22.2 [15.1, 29.8]	45.2 [40.0, 50.4]	0.0 [0.0, 0.0]
	8	67	22.3 [15.2, 29.7]	22.9 [15.7, 30.4]	50.0 [45.2, 54.8]	0.0 [0.0, 0.0]
InternVL3-8B	4	54	32.1 [24.1, 40.4]	32.5 [24.4, 40.9]	62.7 [58.3, 66.8]	0.0 [0.0, 0.0]
	8	53	61.0 [50.6, 71.1]	61.0 [50.6, 71.1]	80.4 [76.9, 83.8]	0.0 [0.0, 0.0]

Table 19: Fresh candidate-count follow-up. Each loss conditions on its own original-order hits; both-resolved loss further requires resolvable citations in both outputs. Source flip uses all pairs. Counts are unweighted questions; rates use sampling weights. All intervals are nominal 95% video-cluster intervals. Zero observed repeat events do not establish deterministic serving.

Reader	Common hits questions/videos	$K = 4$ loss (%)	$K = 8$ loss (%)	$K8 - K4$ (pp)	Resolved $K8 - K4$ (pp)
Qwen3-VL-8B	56/51	16.7 [10.1, 24.1]	20.2 [13.6, 27.8]	3.6 [−2.4, 9.4]	3.7 [−2.6, 9.8]
InternVL3-8B	41/38	30.1 [20.0, 40.5]	60.2 [48.3, 71.5]	30.1 [20.2, 40.5]	28.9 [18.9, 39.4]

Table 20: Conditional loss on the shared original-hit cohort. The difference is paired by question; it is not the difference between the distinct conditional cohorts in Table 19. Nominal 95% video-cluster intervals.

membership and unchanged content across orders. The resulting 1,612 sheets contain four frames each; the source audit covers 6,448 distinct video-frame positions.

Each question stays on the same GPU for both packet sizes within reader. Both sizes use the same 32,768-token context cap, eight-image admission limit, greedy decoding and final 1,024-token output cap. A preliminary 384-token run produced truncated citation lists; its 100 responses are preserved as an excluded input/output-integrity pilot. The common cap was increased before retention analysis, and every final request was rerun. Final truncation counts are reported in the artifact. Seeds, pinned model revisions, runtime versions and input/output hashes accompany the results.

Separate and shared initial-hit cohorts. Table 19 reports each packet size's original-hit cohort. For a paired comparison of conditional loss, we additionally restrict both sizes to the same questions whose fresh original output is a hit at both sizes (Table 20). The separate both-resolved initial-hit bases in Table 19 contain 194/195 pairs for Qwen ($K4/K8$) and 160/159 for InternVL. Alternatives are averaged within question before weighting, and the same 10,000 video-cluster bootstrap draws are used across packet sizes. The resolved sensitivity also requires that both paired outputs resolve at both sizes. Repeat source-flip rates are 0.8 [0.0, 2.3]/1.4 [0.0, 3.3]% for Qwen and 0.0 [0.0, 0.0]/0.0 [0.0, 0.0]% for InternVL ($K4/K8$). Source changes can preserve temporal-hit status; repeat loss is reported separately.

Availability, quality and interpretation. The sampling strata fix $K = 4$ temporal-hit candidate availability at 31.5% in the eligible population; the weighted $K = 8$ estimate is 45.8 [39.1, 52.9]%. The paired original-order mIoU changes ($K8 - K4$) are 0.8 [−1.4, 3.1] points for Qwen and −0.8 [−3.1, 1.6] for InternVL. Losses persist at eight candidates for both readers. On the shared initial-hit cohort, InternVL loss increases by 30.1 [20.2, 40.5] points, while Qwen's difference is 3.6 [−2.4, 9.4] points and its interval includes zero. At eight candidates, the all-question hit-to-miss and miss-to-hit rates are 4.3% and 5.0% for Qwen, and 8.7% and 6.9% for InternVL. Their net hit-rate changes are 0.7 [−2.2, 3.7] and −1.8 [−5.4, 1.7] points. Gains on other questions thus offset losses without restoring the lost initial hits. Increasing packet size also adds evidence and visual tokens; the comparison measures this combined change in the input. The full PCRA reports retain baseline hit rates, all-pair losses and gains, canonical-reading loss, cited-set changes, unchanged-answer cohorts, fallback rates and every output. The archive is `AR10`; its input/output-integrity pilot is kept separately from the final analysis.

Uniform replacement on the same initial hits. To interpret packet expansion, we reuse these outputs and count each packet's temporal-hit candidates, m_{qK}. Independent uniform replacement,

including the original source, has expected loss $u_{qK} = 1 - m_{qK}/K$. Its conditional reference is

$$U_K = \frac{\sum_q w_q H_{q0,K} u_{qK}}{\sum_q w_q H_{q0,K}}.$$

Table 21 compares this expectation with observed loss, retaining identical initial-hit cohorts and inclusion weights. For the cross-size comparison, we condition both policies on $H_{q0,4} = H_{q0,8} = 1$ and estimate $(L_8 - L_4) - (U_8 - U_4)$ using the same video-bootstrap draws. This is a descriptive comparison with a reference policy, not a causal adjustment for candidate difficulty or visual tokens.

Reader	K	Initial hits questions/videos	Observed loss (%)	Uniform loss (%)	Observed − uniform (pp)
Qwen3-VL-8B	4	67/62	21.4 [14.6, 28.8]	66.8 [63.8, 69.6]	−45.4 [−52.7, −37.7]
Qwen3-VL-8B	8	67/59	22.3 [15.2, 29.7]	81.6 [79.9, 83.3]	−59.4 [−66.7, −51.6]
InternVL3-8B	4	54/47	32.1 [24.1, 40.4]	64.8 [61.1, 68.3]	−32.7 [−39.8, −25.2]
InternVL3-8B	8	53/49	61.0 [50.6, 71.1]	79.6 [77.3, 81.8]	−18.6 [−28.2, −9.1]

Reader	Common hits questions/videos	Observed change $L_8 - L_4$ (pp)	Uniform change $U_8 - U_4$ (pp)	Difference of changes (pp)
Qwen3-VL-8B	56/51	3.6 [−2.4, 9.4]	16.3 [14.4, 18.2]	−12.7 [−19.0, −6.5]
InternVL3-8B	41/38	30.1 [20.2, 40.5]	16.8 [14.3, 19.3]	13.3 [3.2, 23.6]

Table 21: Uniform replacement references for the fresh K4/K8 audit. The upper panel uses size-specific initial hits; the lower uses questions initially correct at both sizes. Each question contributes its mean over three reorderings. Uniform loss is an exact per-question expectation; its intervals describe video sampling uncertainty. All intervals are nominal 95% paired video-cluster intervals.

Both readers retain more initial hits than uniform replacement at each size. On common initial hits, InternVL's observed increase exceeds the uniform increase by 13.3 [3.2, 23.6] points, while Qwen's difference of changes is −12.7 [−19.0, −6.5] points. Thus the change in the uniform reference does not numerically match the observed reader changes. Both-resolved differences of changes are −12.7 [−19.2, −6.2] and 12.2 [2.1, 22.5] points, respectively.

Strata by available temporal hits. Table 22 reports m_{qK} strata with their question and initial-hit counts. The one- and two-hit strata contain most initial hits. Strata with only one or two initial-hit videos are descriptive. These strata also differ in clip content and question difficulty; they do not isolate the causal effect of adding an incorrect candidate. The per-question ledger, all estimates and independent verification are in `AR11`.

D.8 Candidate count at a matched visual-token budget

We compare four and eight candidates with 16 source frames and the same measured number of visual tokens per question (Table 23). The same 240 validation questions from 187 videos receive nested LanguageBind packets, three coupled reorderings and an identical-input repeat at each size. Both readers use greedy decoding and a 1,024-token output cap, yielding 4,800 fresh responses. Each candidate is still served as one contact sheet.

For four candidates, a sheet contains four frames; for eight, it contains two. The four-frame schedule uses offsets 0, 2, 5 and 7 from the corrected eight-frame bank, and the two-frame schedule uses offsets 0 and 7. Tiles are padded to the Qwen processor's 32-pixel spatial stride, making a four-frame sheet twice the height of a two-frame sheet. The actual Qwen visual-token budget ranges from 1,232 to 3,520 across questions and is identical across sizes and orders within *every* question. For InternVL, six image tiles per four-frame sheet and three per two-frame sheet, including thumbnails, give 24 tiles and 6,144 visual tokens at either size. Runtime output records verify these counts for every request. Candidate identities and intervals are nested; increasing candidate count redistributes frames and image tiles across clips and adds clip metadata.

On the common initial-hit cohort, K8 loss exceeds K4 loss by 10.3 [2.2, 18.5] points for Qwen and 23.8 [10.3, 37.1] for InternVL. The both-resolved contrasts are 11.8 [3.4, 20.0] and 23.8 [10.3, 37.1] points. The increases therefore also occur with total frames and measured visual tokens fixed, while frame allocation and candidate content change.

Reader	K	m_{qK}	Questions	Initial hits questions/videos	Observed loss (%)	Uniform loss (%)
Qwen3-VL-8B	4	0	120	0/0	–	–
Qwen3-VL-8B	4	1	88	45/43	25.2 [16.3, 35.4]	75.0
Qwen3-VL-8B	4	2	32	22/21	13.6 [3.5, 25.9]	50.0
Qwen3-VL-8B	8	0	95	0/0	–	–
Qwen3-VL-8B	8	1	86	36/34	26.8 [16.3, 39.5]	87.5
Qwen3-VL-8B	8	2	54	29/27	13.3 [4.6, 24.6]	75.0
Qwen3-VL-8B	8	3	4	2/2	66.7*	62.5
Qwen3-VL-8B	8	4	1	0/0	–	–
InternVL3-8B	4	0	120	0/0	–	–
InternVL3-8B	4	1	88	32/28	42.7 [32.1, 53.3]	75.0
InternVL3-8B	4	2	32	22/20	16.7 [7.2, 27.8]	50.0
InternVL3-8B	8	0	95	0/0	–	–
InternVL3-8B	8	1	86	24/22	73.6 [60.6, 85.9]	87.5
InternVL3-8B	8	2	54	26/24	50.5 [36.1, 65.5]	75.0
InternVL3-8B	8	3	4	2/2	83.3*	62.5
InternVL3-8B	8	4	1	1/1	0.0*	50.0

Table 22: Candidate-availability strata in the fresh K4/K8 audit. Counts are unweighted; loss rates use sampling weights and each reader's original-hit cohort. Dashes denote undefined loss when no initial hit exists. *Fewer than five initial-hit videos: point estimate only. Uniform loss within a stratum is exactly $1 - m_{qK}/K$.

Reader	K	Initial-hit questions	Loss (%)	Resolved loss (%)	Repeat loss (%)
Qwen3-VL-8B	4	62	17.2 [10.8, 24.2]	18.0 [11.4, 25.3]	0.0 [0.0, 0.0]
	8	56	31.2 [21.8, 41.0]	32.8 [23.2, 42.9]	7.0 [0.0, 17.2]
InternVL3-8B	4	52	31.4 [23.3, 39.5]	31.4 [23.3, 39.5]	0.0 [0.0, 0.0]
	8	45	59.4 [47.8, 70.3]	59.4 [47.8, 70.3]	2.0 [0.0, 6.7]

Reader	Common hits questions/videos	Observed change $L_8 - L_4$ (pp)	Uniform change $U_8 - U_4$ (pp)	Difference of changes (pp)
Qwen3-VL-8B	42/39	10.3 [2.2, 18.5]	16.7 [14.5, 18.9]	−6.3 [−15.1, 2.5]
InternVL3-8B	35/33	23.8 [10.3, 37.1]	16.4 [13.6, 19.3]	7.4 [−6.1, 20.8]

Table 23: K4/K8 follow-up with 16 frames and equal visual tokens per question. The upper panel conditions each size on its own original hits; resolved loss also requires both citations to resolve. The lower panel conditions on hits at both sizes. Uniform replacement uses $1 - m_{qK}/K$ on that same cohort. All intervals are nominal 95% paired video-cluster intervals with the original sampling weights.

Repeats at K8 lose 7.0 $[0.0, 17.2]$% and 2.0 $[0.0, 6.7]$% of each reader's own original hits. A post-run check of the common initial-hit cohort finds no repeat hit losses at either size, so subtracting observed repeat loss leaves its paired K8–K4 contrast unchanged. On the full K8 initial-hit cohorts, reordering exceeds repeat loss by 24.2 $[11.5, 35.3]$ and 57.4 $[44.3, 69.3]$ points. Uniform-reference differences of changes are -6.3 $[-15.1, 2.5]$ and 7.4 $[-6.1, 20.8]$ points; both intervals include zero.

There is 1 length stop among the 1,200 Qwen K4 responses and none in the other three cells. Fallback counts are 73/88 for Qwen and 2/5 for InternVL (K4/K8). The artifact `AR12` retains raw outputs, per-request processor settings, quality estimates, uniform references and the repeat check.

D.9 Replication on disjoint validation videos

We repeat the InternVL K4/K8 comparison on 240 questions from 168 videos excluded from the original 240-question follow-up sample. Neither questions nor videos overlap with that sample. The dataset was previously used in core validation analyses; this is a replication on disjoint follow-up videos, not an untouched dataset.

The protocol was recorded before inference. Among remaining questions with verified frames for all eight candidates, we sample 120 with a temporal-hit candidate in the top four and 120 without one, using stratum inclusion weights. The pinned reader, four frames per clip, 1,024-token cap,

Packet	Initial hits	Loss (%)	Uniform loss (%)	Repeat loss (%)
K4	57	45.0 [36.7, 53.5]	64.9 [61.5, 68.3]	0.0 [0.0, 0.0]
K8	47	65.0 [55.6, 74.5]	80.5 [78.3, 82.6]	0.0 [0.0, 0.0]

Common initial-hit cohort	Estimate [95% CI]
K4 loss (%)	45.0 [35.7, 54.3]
K8 loss (%)	64.9 [54.5, 75.0]
Loss increase, $L_8 - L_4$ (pp)	19.8 [10.0, 29.5]
Uniform increase, $U_8 - U_4$ (pp)	16.2 [13.4, 19.1]
Difference of increases (pp)	3.6 [−7.1, 14.0]

Table 24: InternVL3-8B replication on videos excluded from the original follow-up sample. The upper panel conditions each size on its own initial hits. The lower panel uses the 37 common initial-hit questions from 32 videos. Intervals use the fixed sampling weights and video clusters.

nested packets and three coupled permutations match Appendix D.7. Each packet size also has an identical-input repeat. All 2,400 responses are fresh. The primary contrast, specified before the run, is $(L_8 - L_4) - (U_8 - U_4)$ on questions initially correct at both sizes, where $U_K = \mathbb{E}[1 - m_{qK}/K]$ uses the same cohort and weights. We use 10,000 paired video-cluster bootstrap draws. Uniform replacement is a descriptive reference, not a causal adjustment for candidate count.

On the common initial-hit cohort, the K8-minus-K4 loss difference is 19.8 [10.0, 29.5] points; the difference relative to uniform replacement is 3.6 [−7.1, 14.0] points (Table 24). The loss increase reproduces, while its excess over the uniform reference remains uncertain on this sample. Both-resolved contrasts are 22.2 [11.5, 32.8] points for loss and 6.1 [−5.4, 17.4] for the difference relative to uniform replacement. Common-hit repeat losses are 0.0 [0.0, 0.0]% at K4 and 0.0 [0.0, 0.0]% at K8. Length stops number 0 and 0 at K4 and K8. The raw outputs, frozen inputs, sampling exclusions, candidate-hit strata and independently recomputed statistics are retained in `AR13`.

D.10 Equal hit fractions and controlled clip placement

Design. We sample 240 questions from 180 validation videos, uniformly from 455 eligible questions whose top eight clips contain exactly two temporal hits and six misses. One of the two hits is chosen uniformly as the shared clip. K4 contains this clip and three randomly selected misses; K8 contains all eight clips. Shared IDs, intervals and four frames per clip remain identical. Both packets have hit fraction $1/4$, so independent uniform replacement has expected loss 75% at either size. This annotation-informed construction tests placement and packet expansion on a controlled cohort.

For each question, two random orders of the other seven clips define the K8 backgrounds; the K4 backgrounds preserve the relative order of their shared members. We insert the shared hit into every slot, keeping the other clips' relative order fixed within each background. InternVL3-8B generates eight K4 and sixteen K8 intervention outputs, plus a retrieval-rank reference and an identical-input repeat at each size: 6,720 responses in total, with greedy decoding and a 1,024-token cap. The protocol and inputs were recorded before inference in `AR14`.

Packet size and placement answer different questions. The primary K8-minus-K4 contrast averages interventions within question on the 58 common initial-hit questions from 52 videos. Loss is 48.1 [41.3, 54.7]% at K4 and 52.8 [44.4, 61.0]% at K8; their difference is 4.7 [−3.9, 12.6] percentage points (Table 25). Because the uniform reference is identical, this is also the difference of excess losses over uniform replacement. Both-resolved and same-answer differences are 4.5 [−4.1, 12.2] and 4.1 [−4.3, 12.0] points; counting unresolved variants as losses gives 4.2 [−4.2, 11.9] points. The data do not establish an extra packet-size penalty at equal hit fractions.

Position has a clear effect within each packet (Figure 9). Across all 240 questions, first rather than last placement increases selection of the shared hit by 49.8 [43.4, 56.0] points at K4 and 30.4 [24.2, 36.4] points at K8. For the 36 questions from 34 videos whose references at both sizes explicitly cite this same clip, temporal-hit loss rises from 6.9% in the first slot to 69.4% in the last at K4, and from 13.9% to 55.6% at K8. Thus placement can lose an initially selected correct citation

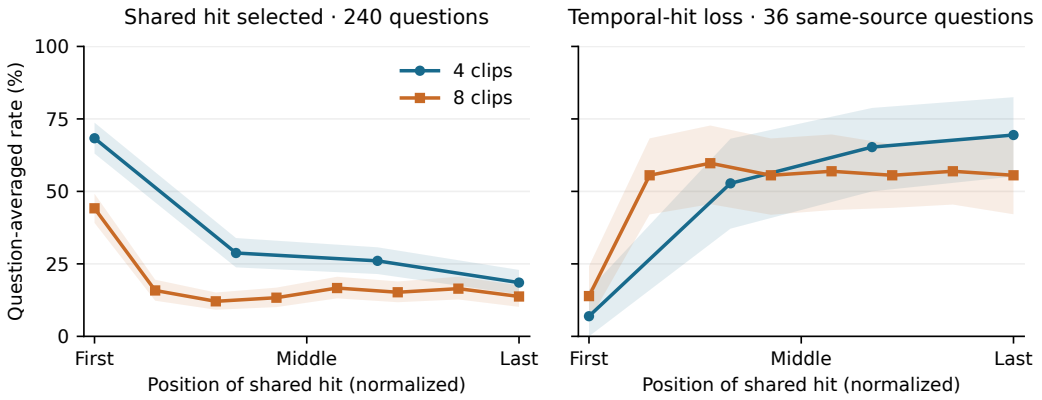

Figure 9: InternVL3-8B: controlled placement of a shared temporal-hit clip at equal hit fractions. Left: selection of that clip across all 240 questions. Right: temporal-hit loss on the same 36 questions that initially cite it at both packet sizes. Each point averages two fixed background orders; bands show pointwise 95% video-cluster intervals. The horizontal axis maps the first and last slots to 0 and 1 for each size. The two panels have different denominators.

Initial-hit cohort	Questions/videos	K4 loss (%)	K8 loss (%)	Difference (pp)
Common hits	58/52	48.1 [41.3, 54.7]	52.8 [44.4, 61.0]	4.7 [−3.9, 12.6]
Same shared source	36/34	48.6 [39.0, 57.7]	51.2 [41.1, 60.7]	2.6 [−7.8, 12.3]

Packet	Slot	Shared-hit selected (%)	Common-hit loss (%)	Same-source hit loss (%)
K4	1	68.3 [63.0, 73.7]	14.7 [7.0, 23.8]	6.9 [0.0, 16.1]
K4	2	28.8 [23.8, 33.9]	47.4 [36.7, 58.2]	52.8 [37.1, 68.2]
K4	3	26.0 [21.5, 30.7]	62.1 [51.1, 72.7]	65.3 [50.0, 78.8]
K4	4	18.5 [14.5, 22.9]	68.1 [57.1, 78.6]	69.4 [55.1, 82.5]
K8	1	44.2 [39.1, 49.2]	22.4 [13.8, 31.7]	13.9 [5.4, 24.2]
K8	2	15.8 [12.3, 19.5]	55.2 [44.3, 65.8]	55.6 [42.0, 68.3]
K8	3	12.1 [9.2, 15.1]	61.2 [49.1, 72.7]	59.7 [45.6, 72.7]
K8	4	13.3 [10.1, 16.8]	58.6 [47.5, 69.3]	55.6 [41.9, 68.3]
K8	5	16.7 [13.1, 20.6]	56.9 [46.8, 66.4]	56.9 [43.5, 69.6]
K8	6	15.2 [11.8, 18.9]	56.0 [46.6, 65.4]	55.6 [44.3, 66.2]
K8	7	16.5 [12.7, 20.6]	56.0 [46.1, 66.1]	56.9 [45.5, 67.8]
K8	8	13.8 [10.2, 17.6]	56.0 [45.0, 67.2]	55.6 [42.1, 68.6]

Table 25: Equal-hit-fraction control. Upper: question-averaged loss over all slots and backgrounds; differences are K8 minus K4. Lower: slot-specific selection on all 240 questions, loss on 58 common initial hits, and loss on 36 initial hits with the same reference source at both sizes. Brackets are 95% video-cluster intervals from 10,000 bootstrap draws. Slot intervals are pointwise.

even while candidate membership and hit fraction stay fixed. Paired first-to-last loss increases are 62.5 [44.7, 78.0] points at K4 and 41.7 [25.0, 56.5] points at K8. These post-hoc differences use the recorded endpoint positions and the same video-cluster draws; they directly quantify the paired change rather than comparing two marginal intervals. Selection of another hit can preserve temporal correctness, particularly at K8.

Repeatability and scope. No response reaches the output cap. On each size's own initial-hit cohort (96 and 99 questions), identical-input repeat loss is 1.0 [0.0, 3.4]% at K4 and 0.0 [0.0, 0.0]% at K8; source flips on all questions are 0.8 [0.0, 2.1]% and 1.2 [0.0, 2.9]%, respectively. K4 has 33 parse failures and 43 fallbacks among 2,400 responses; K8 has 46 and 65 among 4,320. The resolved-output sensitivity above checks the primary contrast. The sample comes from previously accessed validation data and overlaps earlier follow-ups; it is a controlled diagnostic, not a prevalence sample. Across sizes, candidate content and total visual tokens change; within each size, the placement intervention preserves them. The contrasts identify a behavioral dependence on clip

Packet	First−last selection (pp)	First-slot loss (%)	Last-slot loss (%)	Last−first loss (pp)
K4	27.7 [21.3, 34.2]	17.7 [9.2, 27.3]	30.2 [19.2, 41.7]	12.5 [0.0, 25.6]
K8	19.6 [14.1, 25.4]	14.6 [6.1, 25.0]	31.2 [19.8, 42.7]	16.7 [5.0, 28.3]

Table 26: Qwen3-VL-8B on the matched position-control inputs. The selection contrast uses all 240 questions. Temporal-hit losses condition on 48 questions initially citing the same shared hit at both sizes. Endpoints average the two fixed background orders; differences and 95% CIs are paired and clustered by video.

placement, without isolating an internal attention mechanism. Raw responses, frozen inputs, the complete ledger and an independent numerical verification accompany the result artifact.

D.11 Matched Qwen replication of the position intervention

We repeat the control in Appendix D.10 with Qwen3-VL-8B on exactly the same 240 questions from 180 videos. All 6,720 request payloads preserve the InternVL study's candidate IDs, intervals, images and orders; only the reader and its chat template change. Generation uses greedy decoding and the same 1,024-token cap. Question-level averaging and 10,000 paired video-cluster bootstrap draws are unchanged. Reference outputs define each reader's own initial-hit cohorts.

On 73 Qwen common initial-hit questions from 65 videos, loss is 27.6 [21.0, 34.6]% at K4 and 21.4 [15.2, 28.2]% at K8; the paired K8-minus-K4 difference is -6.2 [-12.4, -0.2] points. Both packets retain the same one-quarter hit fraction and 75% uniform expected loss. Table 26 separates selection across all questions from temporal-hit loss on the 48 questions (45 videos) whose two references select the shared hit. These conditional cohorts differ from InternVL's, so their absolute rates are not a paired reader comparison.

Identical-input repeat loss is 0.0 [0.0, 0.0]% on 108 K4 initial hits and 0.9 [0.0, 3.1]% on 110 K8 initial hits. Source flips across all questions are 0.4 [0.0, 1.3]% and 1.2 [0.0, 2.9]%. Length stops number 0 and 3; all responses remain in the primary analysis. Excluding every question with any length stop leaves 238 questions and 48 same-source initial hits. On those hits the last-minus-first loss differences are 12.5 [0.0, 25.6] and 16.7 [5.0, 28.3] points; the common-hit packet-size difference is -6.2 [-12.4, -0.2] points. The artifact records all raw responses, parsing/fallback counts, full slot profiles, input hashes and independent verification in `AR15`.

D.12 Permutation voting without annotation-guided order selection

Policy and comparison. We reuse the exhaustive Qwen3-VL-8B responses on 120 questions from 113 videos. Voting first selects the plurality answer option, then the plurality mapped source among responses with that answer. Answer ties follow option order; source ties use original retrieval rank. Invalid citations retain top-1 fallback; no valid answer yields no answer and the top-1 source. This selects an answer and citation from the same response group. A secondary source-only vote pools citations across answers and is reported separately.

For each question, 256 seeded shuffles of the 23 alternative orders supply two disjoint 11-order pools. The original-order output is excluded. Nested prefixes give budgets of 1, 3, 6 and 11 reader calls. Schedules use question identifiers and order indices, with no annotation, source score or model output entering their construction. The policy and schedules were fixed before computing these exploratory results. We average schedule outcomes within question, retain the exhaustive audit's inclusion weights, and bootstrap videos 10,000 times. Pool realizations are not independent questions.

Retention, quality and cost. Table 27 conditions loss on the same 60 original-order initial hits for every budget. With six calls, answer-then-source voting reduces loss by 6.9 [3.8, 10.0] points and raises mIoU by 1.4 [0.3, 2.6] points over the matched one-call policy. Between-pool source disagreement remains 34.2 [27.7, 40.8]%. The six-call mIoU difference from the recorded original order is 0.2 [-1.9, 2.3] points. Thus the measured quality gain is relative to a randomly reordered one-call policy; improvement over the original order is uncertain. This measures sensitivity to disjoint permutation samples using stored outputs; it does not measure independent serving repeats. Reusing outputs makes this analysis inexpensive, while deployment still requires the stated number

Policy / calls	Initial-hit loss (%)	Reduction (pp)	mIoU	ΔmIoU	Pool disagreement (%)
Answer + source / 1	26.2 [19.2, 33.4]	—	19.0 [15.4, 23.0]	—	45.9 [40.4, 51.2]
Answer + source / 3	21.6 [14.8, 28.8]	4.6 [2.4, 6.7]	19.5 [15.7, 23.7]	0.5 [−0.2, 1.2]	37.3 [31.4, 43.1]
Answer + source / 6	19.3 [12.0, 27.1]	6.9 [3.8, 10.0]	20.4 [16.4, 24.7]	1.4 [0.3, 2.6]	34.2 [27.7, 40.8]
Answer + source / 11	18.7 [11.0, 27.1]	7.5 [3.6, 11.4]	21.0 [16.9, 25.5]	2.0 [0.7, 3.6]	30.7 [23.3, 38.5]
Source only / 1	26.2 [19.2, 33.4]	—	19.0 [15.4, 23.0]	—	45.9 [40.4, 51.2]
Source only / 3	19.9 [13.2, 27.3]	6.2 [4.4, 8.0]	19.4 [15.6, 23.5]	0.4 [−0.3, 1.0]	34.3 [28.6, 39.9]
Source only / 6	17.3 [10.1, 25.2]	8.8 [6.0, 11.7]	19.7 [15.8, 24.0]	0.7 [−0.1, 1.5]	30.9 [24.5, 37.4]
Source only / 11	17.0 [9.3, 25.5]	9.2 [5.6, 12.9]	20.1 [16.1, 24.5]	1.1 [0.1, 2.0]	28.7 [21.2, 36.3]

Table 27: Permutation voting on the same 120 exhaustive-audit questions. Loss conditions on 60 initial hits. Reduction and ΔmIoU compare each policy with its one-call version using the same schedule prefixes. Disagreement compares sources selected from disjoint pools of the stated size. Brackets are paired, weighted 95% video-cluster CIs. Source-only voting can combine citations from different answers.

Policy	Calls	Initial-hit loss (%)	mIoU	$R@.5$
Original reference	1	0.00 [0.00, 0.00]	19.02 [16.37, 21.84]	12.50 [9.91, 15.63]
Randomly reordered	1	18.88 [12.87, 25.51]	19.19 [16.44, 21.98]	12.61 [9.71, 16.00]
Chronological	1	18.37 [10.97, 26.37]	20.44 [17.38, 23.53]	13.29 [10.10, 16.96]
Answer–citation voting	6	16.84 [11.39, 22.82]	19.50 [16.72, 22.37]	13.17 [10.15, 16.65]

Table 28: Fresh Qwen3-VL-8B policy validation: 240 questions, 207 videos and 98 original-hit questions. Random one-call and voting each use two policy replicates per question, losing 37/196 and 33/196 hits, respectively. These are not independent question samples. Values use inclusion weights; brackets give paired 95% video-cluster CIs. Original loss is zero by definition.

of reader calls. All budgets and the secondary policy are retained in `AR16`. Appendix D.13 tests the fixed six-call policy on new videos with fresh inference.

D.13 Fresh validation of the fixed voting policy

Video-disjoint sample and fixed policy. We exclude all 113 videos of the exploratory voting audit from the original 1,000-question NExT-GQA test follow-up. The remaining 784 questions contain 98 historical hits and 686 misses. We include all 98 historical hits and randomly sample 142 misses, with inverse inclusion weights. This gives 240 questions from 207 videos. Historical status determines sampling only; fresh original-order inference defines the 98 initial-hit questions used for retention. The benchmark had been accessed before this follow-up.

Before inference, we record the sample, input hashes, six-call policy and comparisons. We retain the Qwen checkpoint, corrected images, prompt, parser, fallback and 384-token cap of the exhaustive audit. The policy in Appendix D.12 uses the first pair of seeded permutation pools for each new question, taking six orders per pool. The pools are disjoint and exclude original order. Their first calls provide the paired one-call comparators. Original order, chronological order and an identical-input repeat bring the total to 3,600 responses. Policy outcomes are averaged over the two pools within each question before 10,000 weighted video-cluster bootstrap draws. A deployed voting policy uses six calls; the two pools are repeated measurements of it.

Policy effects and repeatability. Table 28 reports all four policies on the same sample. Voting reduces loss by 2.04 $[-3.61, 7.58]$ percentage points relative to random one-call selection. Its mIoU differences are 0.31 $[-0.89, 1.55]$ points from random order and 0.49 $[-1.26, 2.29]$ from original order. The loss-reduction interval permits both increased loss and reduced loss. Fresh validation therefore leaves the benefit uncertain and does not establish a difference from the exploratory effect.

The common sensitivity subset requires a valid answer and citation in the original response and every call of the paired voting pool. Its loss reduction is 1.32 $[-5.06, 7.59]$ points. All responses, including 21 original-order cap terminations and 34 citation fallbacks out of 240, remain in the primary analysis. Identical-input repetition loses 1/98 initial hits (1.02 $[0.00, 3.30]$%) and changes sources on 0.74 $[0.00, 2.15]$% of weighted pairs. Thus exact repeat invariance in the earlier audit does not extend to this execution.

Metric	Original order	Chronological order	Paired change [95% CI]
mIoU	20.25	19.80	-0.45 $[-3.65, +2.91]$
$R@.5$ (%)	17.30	15.44	-1.86 $[-6.04, +3.03]$

Table 29: Chronological input order on the exhaustive audit, with population weights. The original reference is the fresh $p00$ response for the same question.

Measured cost. Weighted mean prompt/completion tokens are 1802.2/167.9 for random one-call selection and 10813.5/1012.2 for six-call voting. Mean summed request durations are 6.66 s and 40.03 s, respectively. Requests ran concurrently; these sums measure service time consumed, not isolated end-user latency. Artifact `AR29` retains frozen inputs, raw outputs, per-pool results, execution costs and independent verification.

D.14 Qualitative examples of temporal-hit turnover

Figure 10 follows two questions through the same tuple permutation in the fresh Qwen3-VL-8B exhaustive audit. Each answer string and answer option remain identical; each output cites one resolvable source. The lost hit in one question is offset by a new hit in the other, preserving their combined hit count while changing which question is correct under the temporal criterion. Both identical-input repeats retain each original selected source. These cases were selected after inspecting outputs to illustrate the two transitions; population estimates come from the complete audit in Table 15. The case manifest preserves raw outputs, emitted timestamps, full source intervals and verified frame hashes in `AR17`.

D.15 Chronological quality from the exhaustive outputs

A chronological sorting rule selects one of the 24 observed orders for every packet. We sort whole tuples by clip start time, end time and immutable identity, then retrieve the response already recorded for that order. The key uses no grounding annotations or reader outputs. This exploratory analysis reuses all 120 questions, their fixed inclusion weights and the same 10,000 paired video-cluster bootstrap draws. 2 packets already have chronological original order, and 9 chronological outputs use the declared fallback; all contribute.

Table 29 reports the paired quality comparison. Chronological serving changes mIoU by -0.45 $[-3.65, +2.91]$ relative to original order and $+0.86$ $[-1.55, +3.39]$ relative to the uniform mean over all 24 orders. Switching from original to chronological order changes the selected clip on 37.24 $[27.38, 47.07]$% of questions and loses 26.67 $[15.79, 38.46]$% of original temporal hits. These are transitions when adopting the rule. After adoption, sorting the same four tuples always produces the same ordered input. The quality intervals permit both gains and losses; no non-inferiority margin was specified. Identical chronological-input repeatability is measured separately in Appendix K.12.

Figure 2 uses the same outputs to decompose hit-rate changes under all 23 alternative orders. Let H_0 and H_π indicate an original and reordered $\mathrm{tIoU} \geq .5$ hit. Averaging alternatives within a question and applying the inclusion weights gives $100\Pr(H_0 = 1, H_\pi = 0) = 4.55$ and $100\Pr(H_0 = 0, H_\pi = 1) = 1.74$. Their difference gives the paired hit-rate change -2.81 $[-5.34, -0.40]$; the conditional loss divides the first probability by $\Pr(H_0 = 1)$. Thus all-question churn and initially correct citation retention have different denominators. The analysis script, chronological-response ledger and figure are in `AR18`.

E Output limits and fallback

An output limit can end generation before the citation JSON is complete. We compare two limits and report results with and without unresolved citations.

E.1 Paired output-cap control

Increasing the output limit removes length stops while retaining temporal-hit losses (Table 30). We rerun both readers on the same 240 validation questions and corrected K4 contact sheets at

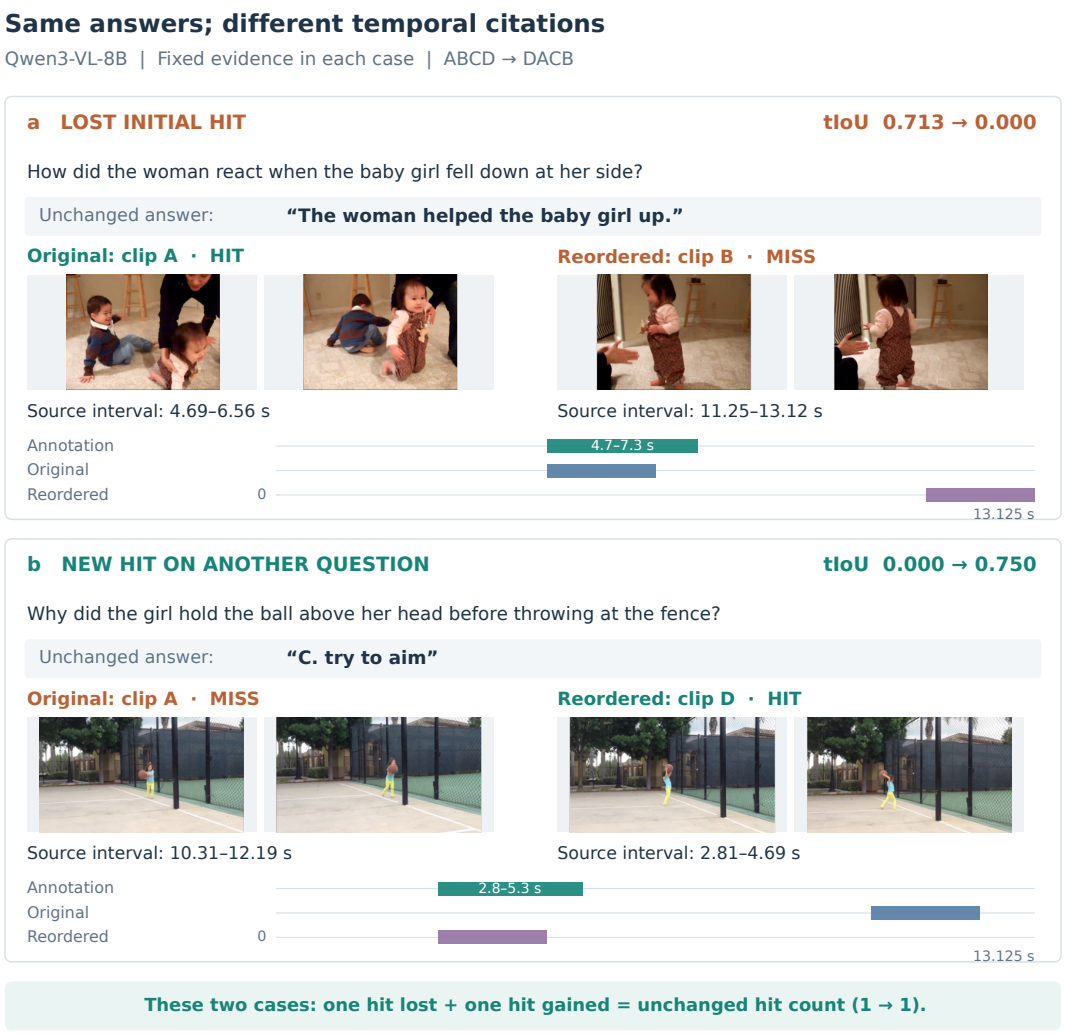

Figure 10: Actual loss and gain with unchanged answers under the same reordering (ABCD → DACB). A–D are local aliases for immutable clip IDs. Two of the four served frames are shown per selected source; bars locate the annotation and full source intervals. As in the audit, tIoU scores the mapped source interval, rather than the generated subspan; hits satisfy $\mathrm{tIoU} \geq .5$. The two selected cases have one hit before and after reordering, but the original hit is lost. They illustrate temporal annotation agreement. Videos: 3070014782 (a), 10864382765 (b).

384 and 1,024 output tokens. Each limit has an original order, the same three reorderings and an identical-input repeat. The two limits for each input share a generation batch; their within-batch order is balanced deterministically. All 4,800 responses are fresh, with stop reasons and output tokens recorded. The sample, population weights and video-cluster bootstrap match Appendix D.7.

Reader	Cap	Initial hits	Loss (%)	Resolved loss (%)	Length	Fallback
Qwen3-VL-8B	384	66	19.7 [13.1, 26.8]	21.1 [14.1, 28.5]	71	126
	1,024	66	19.7 [13.1, 26.8]	20.1 [13.4, 27.3]	0	61
InternVL3-8B	384	54	34.0 [26.1, 42.0]	33.8 [25.5, 42.1]	7	22
	1,024	54	33.3 [25.2, 41.7]	33.8 [25.5, 42.2]	0	15

Table 30: Fresh output-cap comparison. Initial hits count original-order questions; loss conditions on their three reordered pairs. Resolved loss additionally requires both citations to resolve. Length and fallback count generation-limit stops and unresolvable citations among the 1,200 responses in each row. Brackets give video-cluster 95% CIs.

The paired 1,024-minus-384 loss difference on common initial hits is 0.0 [0.0, 0.0] points for Qwen (66 questions) and -0.6 [-2.7, 1.4] for InternVL (54 questions). At 1,024 tokens, both-resolved losses remain 20.1 [13.4, 27.3]% and 33.8 [25.5, 42.2]%. Identical-input repeat losses are 0.0 [0.0, 0.0]% and 1.9 [0.0, 5.7]%. Reordering exceeds identical-input repeat loss by 19.7 [13.1, 26.8] and 31.5 [23.3, 40.0] points. Thus length termination does not explain the losses in this controlled follow-up.

E.2 Legacy records and resolved-only scoring

Output-cap records and resolvable citations. Table 31 audits the original 384-token runs. The saved outputs and runner retain generated text and parsing results, but not generation stop reasons or output-token counts; the inspected run summaries and logs do not recover these fields. Historical truncation frequency is therefore unknown. Parse failure and fallback are reported as observable output properties, not as proxies for length termination.

Reader	Order	Responses	Parse failures n (%)	Fallback n (%)
Qwen3-VL-8B	Original	3,358	323 (9.62)	516 (15.37)
	Reordered	10,074	494 (4.90)	1,079 (10.71)
InternVL3-8B	Original	3,358	19 (0.57)	23 (0.68)
	Reordered	10,074	45 (0.45)	61 (0.61)

Table 31: Available legacy output records. Original runs contain one response per question; reordered runs contain three. Fallback includes parse failures and parsed outputs with no resolvable in-packet citation. Neither column identifies whether generation reached its 384-token cap; termination metadata were not recorded for these responses.

Restricting to unchanged answers with resolvable citations on both sides still gives the retention losses reported in Table 9. These conditional results show loss among resolved outputs; they do not recover missing termination metadata. The excluded 384-token K4/K8 pilot and the complete metadata audit remain in the artifacts; the paired study above tests output-limit sensitivity with fresh inference.

The main results let a reader whose citation cannot be resolved fall back to retrieval top-1. Fallback rows equal retrieval top-1 in the implemented policy, so with $\Delta_{\text{all}} = \text{resolved share} \times \Delta_{\text{resolved}}$ they shrink any measured difference. Table 32 restricts to rows where the reader named a resolvable in-packet unit. These resolved-only results describe the conditional subset on which an interface actually selected a unit; they do not estimate what it would score if every unresolved question were forced to receive a new selection, and because arms resolve on different subsets their resolved-only numbers are not directly comparable to one another. The main results keep the same full cohort for every arm.

Reader	Split	n kept	% kept	ΔmIoU (kept only)	ΔmIoU (all rows)
Qwen3-VL	val	2842	84.6	$+4.50$ [$+3.82$, $+5.16$]	$+3.81$ [$+3.23$, $+4.38$]
InternVL3	val	3335	99.3	$+1.67$ [$+1.11$, $+2.20$]	$+1.66$ [$+1.10$, $+2.19$]
Phi-3.5-Vision	val	2918	86.9	-0.55 [-1.18, $+0.07$]	-0.48 [-1.02, $+0.05$]
Qwen3-VL	test	4652	83.8	$+4.33$ [$+3.74$, $+4.93$]	$+3.63$ [$+3.12$, $+4.15$]
Phi-3.5-Vision	test	4793	86.3	-0.54 [-0.99, -0.09]	-0.46 [-0.85, -0.07]

Table 32: Fallback rows contribute zero difference from retrieval top-1, reducing the magnitude of the full-cohort estimate relative to the resolved-only estimate. InternVL3 keeps 99.3% of rows and its two estimates agree to within 0.01 points.

Reader	Acc.	Sel. unit	mIoU	$R@.5$	ΔmIoU vs. retrieval top-1	Gate
retrieval top-1	—	49.9	17.2	12.0	—	—
Qwen3-VL-4B	59.9	55.1	19.7	14.3	$+2.46$ [$+1.96$, $+2.98$]	GO
Qwen3-VL-8B	63.4	57.8	21.0	15.6	$+3.81$ [$+3.23$, $+4.38$]	GO
same, TP=2	63.5	57.7	21.0	15.8	$+3.83$ [$+3.22$, $+4.41$]	GO
Qwen3-VL-32B	60.2	**60.3**	**22.1**	**16.6**	$+4.90$ [$+4.26$, $+5.52$]	GO
InternVL3-8B	75.8	53.6	18.9	13.4	$+1.66$ [$+1.10$, $+2.19$]	GO
Phi-3.5-Vision	48.1	48.7	16.7	11.4	-0.48 [-1.02, $+0.05$]	NO-GO

Table 33: Five readers on the identical packet and prompt. Within the Qwen3-VL family the gain grows monotonically with scale and the 4B and 32B intervals do not overlap. The TP=2 row is a control: the 32B rung needs two GPUs, so we re-ran 8B under the 32B's exact serving configuration to show that tensor parallelism is not what moves the number.

F ALL READERS ON VALIDATION

Supporting §S10.7. We ran five readers—Qwen3-VL at three scales (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025) and Phi-3.5-Vision (Abdin et al., 2024)—through the same packet on validation (Table 33). These quality measurements use validation data. The selector test follow-up used two readers named in advance; subsequent stability studies use the follow-up designs described in §3.2.

The effect is reader-dependent, and answer accuracy does not explain the ranking. Two of three families gain and one regresses. Within Qwen3-VL the gain rises across 4B, 8B and 32B ($+2.46$, $+3.81$, $+4.90$ mIoU, ends non-overlapping), and re-running 8B under the 32B's two-GPU serving configuration reproduces it to within 0.02 points under that serving control. The most accurate reader, InternVL3-8B at 75.8%, produces one of the smallest gains, and Qwen3-VL-32B gains most while answering *less* accurately than 8B. Decomposing the change in selected-unit overlap into how often a reader overrules the retriever and how much better its choice is when it does, the Qwen gain mainly reflects more frequent overrides (Table 34). This decomposition describes five selected readers using the same temporal annotations as the quality metric.

Paired deltas on $R@1_{\mathrm{IoU}=.5}$: Qwen3-VL $+3.63$ [$+2.80$, $+4.45$], InternVL3 $+1.40$ [$+0.58$, $+2.20$], Phi-3.5-Vision -0.63 [-1.41, $+0.15$]. Reader health on the same runs, as parse rate / strict citation-in-packet / fallback: Qwen3-VL 90.4 / 67.3 / 15.4, InternVL3 99.4 / 63.4 / 0.7, and Phi-3.5-Vision 86.9 / 61.7 / 13.1. The baseline row does not depend on the reader: its localization figures are identical across all of them by construction, which we verify as a consistency check.

Note that the baseline's *joint* figures do differ by reader, and must: a joint metric requires the answer to be correct, and the answer is the reader's. Only the localization half of the baseline is reader-independent.

F.1 A SECOND READER ON TWO OF THE CONTROLS

We repeated the factorial and the re-presentation control with InternVL3-8B on the already-evaluated validation data, using the same packets, prompts, resolver and bootstrap. `K4F4` reuses the existing InternVL3 run, which matches the primary prompt file row for row. The interface comparisons are reported separately in Supplement S10.3.

Reader	Departure rate	Informed lift	Product	Resolved	Δsel-unit	ΔmIoU
Qwen3-VL-4B	20.8%	30.1 pp	6.26	83.3%	+5.21	2.46
Qwen3-VL-8B	34.0%	27.6 pp	9.39	84.6%	+7.95	3.81
Qwen3-VL-32B	45.6%	29.2 pp	13.32	78.2%	+10.42	4.90
InternVL3-8B	34.6%	10.9 pp	3.78	99.3%	+3.75	1.66
Phi-3.5-Vision	30.7%	−4.4 pp	−1.34	86.9%	−1.16	−0.48

Table 34: Departure rate is how often the reader cites a unit other than retrieval top-1, over the rows where it named a unit at all. Informed lift is how much more often its chosen unit overlaps an annotated window than the unit it rejected, on those same rows. Product × resolved share recovers Δsel-unit exactly—that part is an identity—and the product then tracks ΔmIoU across all five readers, which is not.

Contrast / measure	Qwen3-VL-8B	InternVL3-8B
C_4	+1.41 [+0.80, +2.04]	+0.15 [−0.45, +0.77]
C_2	+0.81 [+0.18, +1.46]	−0.39 [−1.04, +0.28]
D_4	−1.49 [−2.02, −0.96]	−0.01 [−0.50, +0.46]
D_8	−2.09 [−2.68, −1.52]	−0.55 [−1.05, −0.05]
I	+0.60 [−0.03, +1.23]	+0.53 [−0.08, +1.15]
E	−0.68 [−1.40, +0.05]	−0.40 [−1.05, +0.28]
Cell means (mIoU)		
K4F4	21.02	18.86
K4F2	19.53	18.85
K8F4g	22.42	19.01
K8F2g	20.34	18.46
Tuple reordering (T2)		
Signed ΔmIoU	−0.22 [−0.69, +0.26]	−0.56 [−1.19, +0.08]
Mean absolute Δ	6.91 [6.49, 7.35]	10.46 [9.92, 11.00]
positive contribution	3.35 [3.04, 3.67]	4.95 [4.56, 5.35]
negative contribution	3.57 [3.24, 3.90]	5.51 [5.08, 5.95]
Source flip (% of pairs)	40.2	59.0
Any-variant flip (% of questions)	66.8	86.2
$R@.5$ miss→hit / hit→miss (%)	3.74 / 3.61	5.40 / 5.67
paired $\Delta R@.5$	+0.13 [−0.60, +0.84]	−0.27 [−1.20, +0.70]
Miss→miss (% of flips)	79.2	79.0

Table 35: Original full-cohort factorial and re-presentation results. The signed change is the difference of the positive and negative contributions and the mean absolute change is their sum. Each reader's contrasts and paired intervals are reported directly; Supplementary Table S16 gives the structurally verified factorial results and Table 61 the three sets side by side. For InternVL3-8B, the reused K4F4 run used an admission limit of 16,384 tokens and the new arms used 32,768, with the model and positional configuration documented in this section. T2 rates distinguish pairs, questions and source-flip events.

Eight-candidate InternVL3-8B prompts require a larger context limit because each contact sheet expands into up to twelve dynamic tiles plus a thumbnail. These packets often exceed the 16,384-token context our primary configuration uses: 40.2% of K8F4g prompts and 36.7% of K8F2g prompts exceed it, the longest reaching 27,461 tokens. The three new factorial arms and the re-presentation arm were therefore run at 32,768, while the reused K4F4 arm is the earlier run at 16,384. That is the only difference between them: model and revision (853e3a79), decoding, dtype, tensor parallelism and prefix caching are identical, and the rotary cache is built from the model's own 32,768-token position limit rather than from the serving limit, preserving the positional basis of the reused arm. The serving limit controls prompt admission; over-long prompts are rejected. K4F4's longest prompt is 11,794 tokens, inside both limits. Every arm is row-complete with no empty generation. Thus the eight-candidate arms use a larger admission limit under the same rendering settings.

The factorial does not carry over. InternVL3-8B's four cell means span 18.46–19.01 mIoU, a range of about half a point against 19.53–22.42 for Qwen3-VL-8B, and its gain over retrieval top-1 is correspondingly smaller. The factorial identity holds for both readers. The Qwen pattern of candidate-expansion gains and frame-thinning losses is not reproduced here, so we report the individual contrasts rather than treating the factor directions as a reader-independent rule.

The re-presentation result does carry over. For InternVL3-8B, no aggregate mIoU change was detected under tuple reordering (-0.56 $[-1.19, +0.08]$), while the cited source changed on 59.0% of question–variant pairs and on 86.2% of questions under at least one of the three orderings—a higher observed flip rate than 40.2% for Qwen3-VL-8B, though we do not test that reader difference. Per-question absolute movement is 10.46 $[9.92, 11.00]$ points, and the share of flips that exchange two clips already below the threshold, 79.0%, matches Qwen3-VL-8B's 79.2% closely. Threshold crossings stay balanced (5.40% up against 5.67% down, paired -0.27 $[-1.20, +0.70]$).

G TEXT RAG REPLICATION

	Original order		Paired over three permutations			
Reader	Answer F1	Hit (%)	Loss among initial hits (%)	Flip (%)	Loss / gain (%)	Net ΔR
Qwen3-32B	74.7	94.0	4.0 [3.1, 5.0]	19.3	3.8 / 2.6	-1.2 [-2.4, 0.0]
Mistral-Small-24B	70.4	90.1	5.7 [4.6, 6.8]	25.3	5.1 / 4.7	-0.5 [-2.1, 1.1]
Qwen3-8B	66.8	88.6	7.6 [6.4, 8.9]	26.2	6.7 / 5.4	-1.3 [-3.1, 0.4]
Gemma-3-12B	61.4	87.3	6.6 [5.4, 7.9]	28.6	5.8 / 5.7	-0.1 [-1.7, 1.6]
Llama-3.1-8B	56.3	84.7	11.3 [9.7, 12.8]	35.6	9.5 / 7.6	-1.9 [-4.1, 0.2]
Ministral-8B	54.9	79.7	12.6 [11.0, 14.3]	36.8	10.0 / 9.3	-0.7 [-2.9, 1.5]
Uniform replacement			80			

Table 36: Text replication on 1,000 HotpotQA distractor questions (ten paragraphs, two gold), three permutations per question. Hit: the first cited paragraph is gold. Loss conditions on initial hits; loss/gain and net ΔR use all pairs. Repeats and further definitions: Appendix G.

Design. 1,000 questions are drawn (seed 2026092401) from the HotpotQA distractor validation set among questions with exactly ten paragraphs and valid supporting-sentence indices. Each packet keeps the ten native paragraphs; every paragraph receives a random four-letter identifier fixed across orders, so identifiers carry no rank cue. Paragraphs show their title and numbered sentences. Readers answer and cite paragraph identifiers and sentence indices in JSON (greedy, 512 tokens). The first resolvable cited paragraph is scored; an unresolved citation is a miss. HotpotQA is a widely used public benchmark, so this is not an untouched test set. The plan and scripts were hash-frozen before inference. Phi-4 was excluded because its local snapshot lacks weights; Qwen3-14B was removed by the authors before its job started to limit one family's share; OLMo-2-7B failed to start twice (its 4,096-token context, then a compilation error) and is a technical failure under the recorded rule.

Secondary definitions. Across readers, sentence-level loss (a cited sentence of the first cited paragraph is a supporting fact) is 5.4–16.5%, loss with unchanged answers is 2.3–5.6%, and identical-input repeats lose 0.0–0.7% of initial hits. The excess of reordering over repeat loss has an interval above zero for 6 of 6 readers. Full support (both gold paragraphs among the cited paragraphs) holds on 6.6–39.1% of questions, and reordering loses 19.9–46.4% of full-support hits. Pooling all four orders as references gives paragraph-level loss of 3.6–12.5%. Loss is 3.6–13.0% on bridge and 5.8–14.6% on comparison questions (exploratory).

Reader comparison. Of 15 reader pairs with at least ten shared initial hits, 0 are hidden with hit rate as the aggregate and 1 with answer F1, and the number of inversions is 0 and 0, respectively. Kendall's τ_b between hit rate and retention is 0.87 $[0.60, 1.00]$, and between answer F1 and retention 0.87 $[0.73, 1.00]$.

Set-based scoring (exploratory). Crediting a question when any cited paragraph is gold, instead of the first cited one, lowers loss among initial hits to 1.8–12.4% across the six readers and both samples. Of the first-cited losses, at most 10% keep the same cited set (a gold paragraph merely listed later) and at most 36% still cite the lost paragraph anywhere; most losses replace the gold paragraph.

Position of the cited paragraph (exploratory). For each initial hit and permutation, the initially cited gold paragraph moves earlier, later or stays in place. Loss per direction uses the same question bootstrap; the later-minus-earlier difference is paired within the resample (Table 37). By the

paragraph's new slot, loss is 6.1% in slots 1–5 and 9.6% in slots 6–10 (3.5 [2.3, 4.8] points). This analysis was added after the frozen analysis.

Reader	Moved earlier (%)	Moved later (%)	Later − earlier (pp)
Ministral-8B	8.7 [6.7, 10.9]	16.1 [13.7, 18.8]	7.4 [4.2, 10.7]
Llama-3.1-8B	8.9 [7.0, 10.9]	14.1 [11.8, 16.6]	5.3 [2.3, 8.2]
Mistral-Small-24B	3.5 [2.4, 4.8]	8.0 [6.1, 9.9]	4.4 [2.3, 6.7]
Qwen3-8B	5.5 [4.1, 7.1]	9.8 [7.8, 11.8]	4.3 [1.9, 6.6]
Gemma-3-12B	5.0 [3.6, 6.5]	8.4 [6.6, 10.3]	3.4 [1.2, 5.7]
Qwen3-32B	2.6 [1.6, 3.7]	5.5 [4.0, 7.1]	3.0 [1.2, 4.8]

Table 37: Loss of initial paragraph hits by the direction in which a permutation moves the cited paragraph; 95% question-bootstrap intervals.

H READER PANEL

Design. The panel reuses the reserved evaluation sample of Appendix K.8: 240 NExT-GQA validation questions from 186 videos, stratified by top-four candidate availability with population weights. Every reader receives the same 1,200 requests: the original order, three seeded non-identity permutations and one identical-input repeat per question, with identical contact-sheet files, task text and scorer. Decoding is greedy with a 1,024-token cap. Gemma-4-31B and InternVL3-8B outputs are those of Appendix K.8; all other readers are fresh inference. The plan, reader list, inclusion rule and analysis were recorded and hash-frozen before inference (released with the artifacts); the analysis script was frozen before any new panel output was read. Table 38 gives every reader's estimates.

Readers and amendments. The recorded list contained fifteen readers. LLaVA-OneVision-1.5-8B was excluded before inference because the installed vLLM does not support its architecture; Qwen2.5-VL-32B and Gemma-4-E4B were excluded because their weights were not available locally. During inference the authors removed Qwen2.5-VL-7B (stopped after 474 responses, which were never analysed) and, before its job started, Qwen3-VL-4B, to limit the share of one model family. InternVL2.5-8B failed to start because its tokenizer file cannot be loaded by the installed sentencepiece version; no runtime setting fixes this, so it is a technical failure under the inclusion rule. Mistral-Small-3.1 (an import incompatibility) and Molmo2-8B (a scheduler token budget below its largest image) each failed once before any response and were retried once with a documented fix, as the rule allows. The panel therefore has 12 readers from six families, 4–32B parameters. No exclusion used quality or retention. All amendments are recorded in `AMENDMENTS.md`. No proprietary API model was evaluated.

Serving settings. New runs use a larger batch (eight concurrent sequences) than the two earlier runs (four). An InternVL3-8B rerun under the new settings, reported separately and excluded from the panel, changes $R@.5$ by 0.00 [0.00, 0.00], initial-hit loss by 0.62 [−2.69, 3.89] and source flips by −1.41 [−3.16, 0.06] points.

Rank agreement. Across the 12 readers, Kendall's τ_b between $R@.5$ and retention is 0.33 [0.09, 0.63], and between mIoU and retention 0.45 [0.18, 0.67] (percentile intervals from the shared resample). Restricting losses to pairs whose citations both resolve gives τ_b 0.39 [0.12, 0.67] with $R@.5$.

Within families. Within each family, the larger model loses fewer shared initial hits by point estimate, but only two of the five differences have intervals that exclude zero. Each loss difference is the larger model minus the smaller one on their shared initial-hit questions, so negative values favour the larger model. Qwen3-VL-32B has a higher $R@.5$ than Qwen3-VL-8B (15.6 versus 12.4), yet their difference is −0.8 [−9.3, 7.9] points (43 questions). For Gemma-3, 27B minus 12B is −10.8 [−23.3, 1.7] (40 questions). For Gemma-4, 31B minus 12B is −20.7 [−32.5, −9.5], 26B-A4B minus 12B is −12.2 [−22.9, −2.0], and 31B minus the 26B mixture-of-experts model is −8.5 [−18.7, 1.5]; the 26B-A4B reader has the highest $R@.5$ in the panel (17.5, against 16.9 for 31B), while its shared-hit loss is higher by a difference whose interval includes zero. Family, size, architecture and image processing vary together, so these are descriptions, not scaling estimates.

Reader	$R@.5$	mIoU	Hit q.	Loss (%)	Both-res.	Uniform	Flip	Fallback	Repeat loss
Gemma-4-26B-A4B	17.5 [13.5, 21.6]	21.3	66	28.8 [20.1, 37.8]	27.9	68.9	43.2	79	3.0
Gemma-4-31B	16.9 [13.1, 21.0]	21.7	64	13.0 [7.3, 19.4]	13.0	69.5	31.7	46	0.0
Qwen3-VL-32B	15.6 [11.9, 19.6]	21.0	59	22.6 [15.8, 29.5]	22.6	69.1	40.5	11	0.0
InternVL3-8B	14.3 [10.6, 18.3]	18.2	54	46.3 [36.4, 56.1]	47.2	67.6	61.5	7	0.0
Gemma-4-12B	14.3 [10.7, 18.2]	19.4	54	37.0 [25.8, 48.2]	35.3	69.0	46.9	57	0.0
Gemma-3-27B	14.0 [10.4, 17.9]	18.2	53	28.9 [18.2, 40.1]	28.9	68.9	44.5	0	1.9
Molmo2-8B	13.8 [10.3, 17.5]	18.5	52	46.2 [37.6, 54.7]	46.4	67.8	56.6	22	0.0
Mistral-Small-3.1-24B	13.2 [9.8, 17.1]	17.9	50	42.0 [32.6, 51.6]	42.0	68.0	62.9	0	0.0
Gemma-3-12B	12.7 [9.2, 16.4]	17.7	48	40.3 [30.0, 50.3]	40.6	67.2	52.0	30	2.1
Qwen3-VL-8B	12.4 [9.1, 16.0]	17.8	47	23.4 [15.1, 32.5]	23.4	67.6	39.9	51	2.1
Pixtral-12B	12.4 [9.2, 16.1]	16.3	47	39.0 [28.0, 50.6]	44.4	68.6	49.7	266	4.3
Phi-3.5-Vision	10.1 [7.0, 13.5]	15.6	38	66.7 [55.6, 76.9]	66.7	66.4	71.5	2	2.6
InternVL3-8B (serving control)	14.3 [10.6, 18.3]	18.2	54	46.9 [37.3, 56.2]	47.8	67.6	60.1	9	1.9

Table 38: Reader panel on identical packets. $R@.5$ and mIoU use the original order; loss conditions on each reader's initial hits over three reorderings; both-resolved restricts to pairs whose citations both resolve; uniform replaces the reordered citation uniformly within the packet. Fallback counts unresolved responses of 1,200; repeat loss uses identical-input repeats. 95% video-cluster CIs from one shared resample.

Pairwise comparisons. For each reader pair, the loss difference is computed on questions both readers hit initially, averaging the three reorderings within a question. A pair is *hidden* when its $R@.5$ difference interval includes zero and its loss difference interval excludes zero, an *inversion* when both exclude zero and the higher-$R@.5$ reader loses more, and *aligned* when both exclude zero in the other direction. Holm adjustment uses bootstrap two-sided p-values over all 66 eligible pairs. With mIoU as the aggregate, 12 pairs are hidden (2 after Holm), 24 aligned and 1 inverted.

Table 39: Reader pairs on shared initial hits, ordered by bootstrap p-value of the loss difference. Δ is first minus second reader.

Reader A	Reader B	Shared	$\Delta R@.5$	Δ loss (pp)	Holm p	Category
Gemma-4-31B	Gemma-3-27B	42	2.9 [0.0, 6.0]	−17.5 [−26.7, −8.5]	0.000	hidden
Gemma-4-31B	Mistral-Small-3.1-24B	41	3.7 [1.0, 6.5]	−30.1 [−40.7, −19.8]	0.000	aligned
Gemma-4-31B	Gemma-4-12B	45	2.6 [0.0, 5.5]	−20.7 [−32.5, −9.5]	0.000	hidden
Gemma-4-31B	Gemma-3-12B	40	4.2 [1.5, 7.2]	−29.2 [−39.5, −18.4]	0.000	aligned
Gemma-4-31B	Phi-3.5-Vision	28	6.9 [3.7, 10.3]	−57.1 [−71.0, −42.3]	0.000	aligned
Gemma-4-31B	Molmo2-8B	43	3.2 [0.5, 5.8]	−34.1 [−44.9, −23.1]	0.000	aligned
Qwen3-VL-8B	Phi-3.5-Vision	28	2.4 [−0.5, 5.2]	−45.2 [−59.3, −31.1]	0.000	hidden
Qwen3-VL-8B	Molmo2-8B	40	−1.3 [−3.7, 1.0]	−23.3 [−35.4, −11.9]	0.000	hidden
Qwen3-VL-32B	Mistral-Small-3.1-24B	43	2.4 [0.0, 4.8]	−20.9 [−30.8, −11.6]	0.000	aligned
Qwen3-VL-32B	Phi-3.5-Vision	29	5.6 [2.4, 8.8]	−47.1 [−60.3, −34.5]	0.000	aligned
Qwen3-VL-32B	Molmo2-8B	41	1.9 [−0.8, 4.6]	−26.0 [−37.3, −15.0]	0.000	hidden
Gemma-3-27B	Phi-3.5-Vision	25	4.0 [0.8, 7.3]	−52.0 [−69.2, −35.1]	0.000	aligned
Gemma-4-12B	Phi-3.5-Vision	29	4.2 [1.3, 7.3]	−32.2 [−47.8, −17.1]	0.000	aligned
Pixtral-12B	Phi-3.5-Vision	24	2.4 [−0.8, 5.6]	−44.4 [−59.4, −29.2]	0.000	hidden
Gemma-3-12B	Phi-3.5-Vision	27	2.6 [−0.3, 5.7]	−32.1 [−48.4, −16.1]	0.000	hidden
Gemma-4-26B-A4B	Phi-3.5-Vision	31	7.4 [4.1, 10.9]	−51.6 [−65.4, −37.0]	0.000	aligned
Gemma-4-31B	InternVL3-8B	40	2.6 [−0.5, 5.9]	−26.7 [−40.0, −13.3]	0.010	hidden
InternVL3-8B	Qwen3-VL-32B	39	−1.3 [−4.6, 1.6]	23.1 [10.1, 36.3]	0.010	hidden
Qwen3-VL-32B	Gemma-3-12B	40	2.9 [0.3, 5.6]	−20.0 [−31.8, −8.7]	0.010	aligned
Mistral-Small-3.1-24B	Phi-3.5-Vision	31	3.2 [0.5, 5.9]	−25.8 [−38.7, −13.6]	0.010	aligned
Mistral-Small-3.1-24B	Gemma-4-26B-A4B	43	−4.2 [−7.1, −1.5]	17.8 [7.4, 28.4]	0.018	aligned
Phi-3.5-Vision	Molmo2-8B	35	−3.7 [−6.0, −1.6]	24.8 [11.1, 37.9]	0.045	aligned
Qwen3-VL-8B	Mistral-Small-3.1-24B	41	−0.8 [−2.8, 1.1]	−17.9 [−28.4, −7.1]	0.070	hidden
InternVL3-8B	Qwen3-VL-8B	37	1.9 [−1.0, 4.8]	18.0 [7.1, 28.7]	0.095	hidden
Gemma-4-26B-A4B	Molmo2-8B	43	3.7 [0.8, 6.7]	−17.8 [−29.1, −6.5]	0.160	aligned
Gemma-3-27B	Molmo2-8B	37	0.3 [−2.5, 3.0]	−19.8 [−32.5, −6.7]	0.189	hidden
Pixtral-12B	Molmo2-8B	32	−1.3 [−4.3, 1.6]	−20.8 [−34.3, −7.1]	0.208	hidden
Gemma-4-31B	Pixtral-12B	30	4.5 [1.1, 8.0]	−18.9 [−33.3, −5.1]	0.320	aligned
Qwen3-VL-32B	Gemma-4-12B	42	1.3 [−1.4, 4.2]	−15.9 [−28.0, −4.2]	0.350	hidden
Gemma-3-12B	Gemma-4-26B-A4B	37	−4.8 [−8.1, −1.6]	14.4 [3.0, 25.9]	0.592	aligned
Gemma-4-12B	Gemma-4-26B-A4B	41	−3.2 [−6.4, 0.0]	12.2 [2.0, 22.9]	0.655	hidden
Qwen3-VL-8B	Gemma-3-12B	38	−0.3 [−2.6, 2.1]	−12.3 [−23.4, −1.9]	0.805	hidden
Gemma-4-31B	Qwen3-VL-32B	48	1.3 [−1.4, 4.0]	−8.3 [−15.7, −1.2]	0.864	hidden
Gemma-4-31B	Qwen3-VL-8B	41	4.5 [1.9, 7.1]	−11.4 [−21.7, −1.6]	0.898	aligned
Gemma-3-27B	Mistral-Small-3.1-24B	38	0.8 [−1.8, 3.4]	−12.3 [−22.2, −1.6]	1.000	hidden
Qwen3-VL-8B	Gemma-4-12B	36	−1.9 [−4.7, 0.8]	−10.2 [−20.2, −0.9]	1.000	hidden
InternVL3-8B	Gemma-3-27B	34	0.3 [−3.1, 3.6]	13.7 [1.1, 25.7]	1.000	hidden
InternVL3-8B	Pixtral-12B	31	1.9 [−1.4, 5.0]	11.8 [0.0, 24.4]	1.000	—
InternVL3-8B	Phi-3.5-Vision	27	4.2 [1.1, 7.3]	−14.8 [−31.5, 1.2]	1.000	—
Gemma-3-27B	Gemma-3-12B	40	1.3 [−1.0, 3.8]	−10.8 [−23.3, 1.7]	1.000	—
Gemma-4-31B	Gemma-4-26B-A4B	51	−0.5 [−3.2, 2.3]	−8.5 [−18.7, 1.5]	1.000	—
Gemma-4-12B	Molmo2-8B	39	0.5 [−1.9, 3.0]	−9.4 [−21.4, 1.5]	1.000	—

Reader A	Reader B	Shared	$\Delta R@.5$	Δ loss (pp)	Holm p	Category
Mistral-Small-3.1-24B	Gemma-4-12B	37	-1.1 $[-3.6, 1.5]$	7.2 $[-1.0, 16.2]$	1.000	—
Mistral-Small-3.1-24B	Pixtral-12B	31	0.8 $[-2.2, 3.8]$	10.8 $[-2.3, 22.9]$	1.000	—
Pixtral-12B	Gemma-3-12B	26	-0.3 $[-3.6, 3.3]$	-12.8 $[-28.0, 2.7]$	1.000	—
InternVL3-8B	Gemma-4-26B-A4B	40	-3.2 $[-6.6, 0.0]$	11.7 $[-2.7, 26.2]$	1.000	—
InternVL3-8B	Gemma-4-12B	37	0.0 $[-2.9, 2.8]$	9.0 $[-2.0, 20.8]$	1.000	—
Gemma-3-12B	Molmo2-8B	40	-1.1 $[-3.3, 1.1]$	-6.7 $[-15.2, 1.7]$	1.000	—
Qwen3-VL-32B	Pixtral-12B	32	3.2 $[-0.3, 6.6]$	-9.4 $[-23.8, 4.3]$	1.000	—
Gemma-4-12B	Pixtral-12B	30	1.9 $[-1.6, 5.3]$	8.9 $[-4.0, 22.2]$	1.000	—
Gemma-3-27B	Gemma-4-12B	39	-0.3 $[-3.1, 2.5]$	-6.8 $[-17.5, 3.8]$	1.000	—
InternVL3-8B	Gemma-3-12B	36	1.6 $[-1.3, 4.5]$	6.5 $[-5.1, 17.4]$	1.000	—
Qwen3-VL-32B	Gemma-4-26B-A4B	45	-1.9 $[-4.8, 1.1]$	-4.4 $[-13.0, 3.8]$	1.000	—
Gemma-3-27B	Pixtral-12B	25	1.6 $[-2.0, 5.1]$	-5.3 $[-17.6, 7.0]$	1.000	—
InternVL3-8B	Mistral-Small-3.1-24B	39	1.1 $[-1.6, 3.7]$	4.3 $[-5.8, 15.0]$	1.000	—
Gemma-3-27B	Gemma-4-26B-A4B	41	-3.4 $[-6.7, -0.3]$	4.9 $[-7.2, 17.9]$	1.000	—
Qwen3-VL-32B	Gemma-3-27B	42	1.6 $[-1.1, 4.4]$	-4.8 $[-17.5, 7.4]$	1.000	—
Qwen3-VL-8B	Gemma-3-27B	34	-1.6 $[-4.5, 1.4]$	-3.9 $[-17.6, 9.5]$	1.000	—
InternVL3-8B	Molmo2-8B	39	0.5 $[-2.1, 3.2]$	2.6 $[-10.2, 15.1]$	1.000	—
Qwen3-VL-8B	Gemma-4-26B-A4B	38	-5.0 $[-8.3, -1.9]$	-1.8 $[-10.8, 7.0]$	1.000	—
Qwen3-VL-8B	Pixtral-12B	30	0.0 $[-3.1, 3.0]$	2.2 $[-12.8, 17.2]$	1.000	—
Pixtral-12B	Gemma-4-26B-A4B	33	-5.0 $[-8.5, -1.6]$	2.0 $[-13.3, 18.7]$	1.000	—
Qwen3-VL-8B	Qwen3-VL-32B	43	-3.2 $[-5.7, -0.9]$	0.8 $[-7.9, 9.3]$	1.000	—
Gemma-4-12B	Gemma-3-12B	37	1.6 $[-1.0, 4.3]$	-0.9 $[-13.1, 10.7]$	1.000	—
Mistral-Small-3.1-24B	Gemma-3-12B	39	0.5 $[-1.6, 2.8]$	0.0 $[-9.7, 9.4]$	1.000	—
Mistral-Small-3.1-24B	Molmo2-8B	44	-0.5 $[-2.4, 1.4]$	0.0 $[-10.3, 10.1]$	1.000	—

I EVALUATION-SET SIZE, INFORMATION MATCHING AND READER SEPARATION

These analyses reuse existing responses; no new inference was run for them. All draw evaluation sets without replacement from a recorded sample (whole videos in video, questions in text), recompute every reader pair's paired difference with 1,000 cluster bootstrap draws and the audit's weights, and count a pair as separated when the 95% interval excludes zero; retention needs at least ten shared initial-hit questions.

Information matching. Shared-hit retention uses the original order and three reorderings of every question, whereas the original-order aggregate uses one presentation. We therefore also average each question's aggregate over the same four presentations. Table 40 reports the share of pairs separated by each statistic. Both statistics that use the reorderings separate more pairs than the single-order aggregate, and the order-averaged aggregate separates at least as many pairs as retention on every full sample. On the full video samples, retention minus the order-averaged aggregate is 1 $[-12, 17]$ and -8 $[-25, 11]$ percentage points (200 outer cluster bootstrap draws), and retention alone separates 11% and 6% of pairs. Separation measures whether a statistic can tell two readers apart, not which difference matters to users.

Ranking reliability. For split-half reliability, the videos (text: questions) are split into halves at random 1,000 times; readers are ranked in each half by the original-order aggregate, the order-averaged aggregate and retention (one minus loss among a reader's own initial hits), and the Spearman correlation between the halves is averaged. On the recorded video sample (all twelve readers; Figure 5 uses the nine present in both samples) these are 0.58, 0.78 and 0.81; on the fresh video sample 0.75, 0.93 and 0.72; in text all three exceed 0.9. Figure 5 traces reliability by the number of labelled questions.

Intervals and reader sets. Figure 5 uses the nine readers present in both video samples and 1,000 splits. Intervals for two contrasts come from 200 outer cluster-bootstrap draws with 200 splits each. Four orders minus one order at 120 questions: $[+0.09, +0.40]$ (recorded) and $[+0.06, +0.35]$ (fresh); at 100 text questions, where four orders reach 0.80 and 0.84, $[+0.10, +0.24]$ and $[+0.09, +0.20]$. Four orders at 120 questions minus one order at 240: $[-0.10, +0.29]$ and $[-0.13, +0.16]$; at 100 versus 250 text questions $[-0.06, +0.06]$ and $[-0.06, +0.02]$. With all twelve recorded readers, reliability at 120 questions is 0.38 for one order and 0.64 for four (gain $[+0.07, +0.37]$).

Fallback and exchangeability. Unresolved citations fall back to retrieval top-1, which does not depend on order, so fallback pairs cannot be lost. Counting fallback responses as misses leaves Kendall's τ_b between the order-averaged aggregate and retention at 0.73–0.89 across the four samples. If a reader's four presentations of a question were exchangeable, loss among initial hits would equal $1 - (c - 1)/3$ averaged over initial hits, where c is the question's hit count over the four presentations. Observed losses differ from this expectation by 2.8 and 2.2 points on average in video (signed -0.1 and -0.8) and by 0.4 and 0.3 in text; rankings by observed and exchangeable loss agree with $\tau_b \geq 0.85$. Retention is thus determined almost entirely by per-question hit consistency. The small signed gaps show only that, averaged over panel readers, the provider order is not markedly better than the three reorderings; for single readers it can be (keeping the top-ranked clip first halves the loss in the exhaustive audit, Appendix C.7).

Calls versus labelled questions. At a fixed budget of 240 video calls on the recorded sample, 240 questions under one order, 120 under two and 60 under four separate 33%, 33% and 25% of pairs, with reliability 0.58, 0.52 and 0.47; text shows the same ordering at 400 and 1,000 calls. Reordering helps when labelled questions, not calls, are scarce.

Retention and order-averaged quality. Kendall's τ_b between reader rankings by the order-averaged aggregate and by retention is 0.76 $[0.45, 0.87]$ and 0.83 $[0.50, 0.94]$ in video and 0.87 $[0.73, 1.00]$ and 0.87 $[0.87, 1.00]$ in text (cluster bootstrap). The two are related by construction, since a reader that keeps its hits across orders has a higher order-averaged hit rate.

		Reader pairs separated (%)			
Sample	Questions	One order	Four orders	Retention	Retention only
Video, recorded	120	21	42	43	13
Video, recorded	180	28	51	50	12
Video, recorded	240	33	59	55	11
Video, fresh	120	32	53	44	13
Video, fresh	180	43	64	53	12
Video, fresh	240	50	75	53	6
Text, recorded	100	36	59	47	6
Text, recorded	250	59	78	68	3
Text, recorded	500	73	89	79	1
Text, recorded	1000	87	93	87	0
Text, fresh	100	41	61	46	5
Text, fresh	250	67	82	70	2
Text, fresh	500	80	91	84	1
Text, fresh	1000	87	93	93	0

Table 40: Share of reader pairs separated by the original-order aggregate, the aggregate averaged over the same four presentations that retention uses, and shared-hit retention; "retention only" counts pairs separated by retention but not by the four-order aggregate. Means over 200 subsamples; full samples are single evaluations. The recorded confirmation tests (Appendix J.1) used the same number of bootstrap draws with another seed, so thresholded counts can differ by a pair (fresh video, 240 questions: 53% here, 56% there). Exploratory reanalysis of existing responses.

J FRESH-SAMPLE CONFIRMATIONS

Both confirmations test analyses first reported as exploratory, on samples disjoint from the ones analysed before, with criteria recorded and hash-frozen before inference.

J.1 VIDEO

The video separation result against the original-order aggregate was an analysis added after the frozen panel analysis. We drew 240 NExT-GQA validation questions (seed 2026092504) from 170 videos outside the reader-panel sample, with the panel's stratification, packets, rendering, prompt, orders and scoring, and ran 9 panel readers; the three slowest two-card readers were omitted for time. Before inference we recorded C1v, that at 120, 180 and 240 questions shared-hit retention separates at least as many reader pairs as $R@.5$, confirmed only if it holds at all three sizes. The NExT-GQA validation split was used in the core audit; these videos are new only to the panel.

C1v is confirmed: retention separates 44%, 53% and 56% of pairs against 33%, 42% and 50% for the original-order $R@.5$; at 60 questions, which had no criterion, the shares are 16% and 18%. Against $R@.5$ averaged over the same four presentations, retention does not separate more pairs (Appendix I). All 9 readers lose less than uniform replacement; loss among initial hits ranges from 16.2% to 53.6% and $R@.5$ from 10.9 to 17.6 (Table 41). Kendall's τ_b between $R@.5$ and retention is 0.69 $[0.20, 0.82]$; of 36 pairs, 9 are hidden (7 after Holm), 10 aligned and 0 inverted. On 41 questions both hit initially, InternVL3-8B loses 25.2 $[14.6, 36.4]$ points more than Gemma-4-31B, and Gemma's $R@.5$ advantage is 4.5 $[1.2, 7.9]$.

The independent recount first disagreed on one question whose best clip has temporal IoU exactly 0.5 (overlap 2.9 s, union 5.8 s): floating-point division placed it just below the threshold, while the analysis uses exact decimal arithmetic. The recount now uses the same exact arithmetic and agrees for every reader.

Reader	$R@.5$	Loss (%)	Uniform (%)	Flip (%)
Gemma-4-31B	17.6 [13.3, 22.2]	16.2 [10.3, 22.6]	66.9	29.4
Qwen3-VL-32B	17.6 [13.5, 22.2]	23.0 [15.8, 30.6]	65.9	41.0
Gemma-4-12B	16.4 [12.3, 21.0]	24.2 [16.4, 32.5]	65.2	43.6
Qwen3-VL-8B	15.2 [11.4, 19.5]	28.1 [20.0, 37.0]	66.8	42.8
InternVL3-8B	13.1 [9.6, 17.0]	38.8 [29.5, 48.1]	63.2	61.4
Gemma-3-12B	13.1 [9.4, 17.2]	24.2 [16.0, 32.8]	65.0	50.8
Pixtral-12B	12.6 [9.1, 16.5]	37.7 [27.9, 47.9]	64.6	56.9
Molmo2-8B	11.9 [8.5, 15.7]	26.7 [17.1, 36.7]	63.0	56.9
Phi-3.5-Vision	10.9 [7.6, 14.6]	53.6 [43.2, 63.0]	63.6	68.5

Table 41: Fresh-sample video confirmation (240 questions, 170 videos), with 95% video-cluster intervals.

J.2 Video, second sample: presentations and perturbations

A second fresh sample of 240 NExT-GQA validation questions (seed 2026092512) comes from 107 videos that appear in neither the reader panel nor the first confirmation (410 videos excluded; overlap 0), with the same stratification, packets, prompt and scoring and the same nine readers. Besides the original order, three reorderings and an identical-input repeat, every question received three temperature samples of the original order (0.7, top-p 0.95, distinct seeds) and three relabellings (original order, greedy, clip identifiers replaced by random tokens and mapped back for scoring). The seed and relabel arms were added by an amendment recorded before any response. One reader's identical-input repeats ran partly on restarted or parallel servers with identical settings.

Tests recorded before inference: A1 (primary), that split-half ranking reliability with 120 questions under four orders is at least that with 240 questions under one, did not hold (0.74 versus 0.82, paired difference -0.07 $[-0.52, +0.32]$); A2, that reliability at 120 questions increases with each added order, held (0.58, 0.68, 0.72, 0.74); A3, that four orders separate more reader pairs than one, held (52% versus 33%). B1 compares reliability at 120 questions when the four presentations are reorderings (0.74), temperature samples (0.68) or relabellings (0.75); reorderings minus samples $[-0.06, +0.21]$, minus relabellings $[-0.13, +0.16]$ (200 outer video-bootstrap draws). B2 compares loss among initial hits, pooled over readers, when the second presentation is a reordering (33.7 $[28.0, 39.5]$%), a temperature sample (10.8 $[8.5, 13.3]$%) or a relabelling (18.8 $[15.5, 22.4]$%); reordering minus sampling is 22.9 $[18.1, 28.0]$ points and reordering minus relabelling 14.9 $[10.3, 19.7]$. Loss among initial hits under reordering spans 19.6–65.3% across readers, and an independent recount agrees for every reader.

Operational notes: Gemma-4-31B's first start failed while the previous tenant's memory was being released and was retried once; one card was returned to other work during the run; Gemma-3-12B was interrupted twice by scheduling errors and resumed from its stored responses, its last requests distributed over five servers with identical settings. Each response records its server.

J.3 Text

The separation result and the text position effect were first found in analyses added after the frozen analyses, on a sample already analysed. We therefore drew 1,000 new HotpotQA distractor questions (seed 2026092502) from the eligible questions that exclude every question of the first sample, and ran the same six readers with the same packets, prompt, arms and scoring. Before inference we

recorded and hash-froze four tests: C1 (primary), that at 100, 250 and 500 questions shared-hit retention separates at least as many reader pairs as hit rate, at all three sizes; C2, that the per-reader turnover criteria hold for a majority of readers; C3, that the pooled later-minus-earlier position difference has an interval above zero; and C4, descriptive rank agreement. Client concurrency was 32 instead of 16 for every reader of this run.

C1 is confirmed: retention separates 46%, 70% and 84% of pairs against 41%, 65% and 79% for hit rate, and 93% against 87% on the full sample. These criteria compare retention with the original-order aggregate; Appendix I compares it with the aggregate averaged over the same presentations. C2 holds for 6 of 6 readers, with loss among initial hits of 4.1–13.2% and full-support loss of 23.4–52.8%. C3 holds: the pooled difference is 4.6 $[3.2,\ 6.0]$ points and is above zero for 5 of 6 readers. C4: Kendall's τ_b between hit rate and retention is 1.00 $[0.73,\ 1.00]$; of 15 pairs, 13 are aligned, 1 hidden and 0 inverted. Unlike the first sample, Llama-3.1-8B shows a net change whose interval excludes zero (3.0 $[0.7,\ 5.2]$ points), because its gains exceed its losses. Table 42 gives every reader.

One change was made after inference: the C1 script computed its result but failed to save it because two boolean values were not JSON-serialisable; casting them to Python booleans was the only edit, recorded in the plan with the frozen file kept alongside.

Reader	Hit (%)	Loss (%)	Net ΔR	R1–R3	Later − earlier (pp)
Qwen3-32B	92.6	4.1 $[3.2, 5.1]$	-0.1 $[-1.5, 1.3]$	yes	1.7 $[-0.1, 3.5]$
Mistral-Small-24B	89.8	5.2 $[4.1, 6.2]$	0.3 $[-1.2, 1.9]$	yes	4.1 $[2.1, 6.2]$
Qwen3-8B	86.3	7.3 $[6.1, 8.6]$	0.3 $[-1.5, 2.1]$	yes	5.3 $[2.9, 7.7]$
Gemma-3-12B	85.9	7.5 $[6.2, 8.8]$	1.1 $[-0.8, 3.1]$	yes	2.7 $[0.2, 5.2]$
Llama-3.1-8B	79.3	10.0 $[8.6, 11.6]$	3.0 $[0.7, 5.2]$	yes	5.1 $[2.4, 8.0]$
Ministral-8B	78.3	13.2 $[11.5, 15.0]$	-0.2 $[-2.5, 2.1]$	yes	8.7 $[5.4, 12.1]$

Table 42: Fresh-sample confirmation: first-cited-paragraph hit rate, loss among initial hits, net hit-rate change, the turnover criteria recorded before inference and the position contrast, with 95% question-bootstrap intervals.

K SELECTION INTERFACES IN FULL

Every arm selects one whole delivered unit under the same citation contract, on the same packet and the same questions, so the arms differ only in how the selection is elicited. `B3` and `B4` score each candidate by the first-token likelihood of `Yes` against `No` and take the argmax, breaking ties toward the lower retrieval rank; they carry no abstention option, which is why their unresolved rate is zero by construction rather than by improvement. The Holm family is the four contrasts declared in the protocol before the runs. The recorded costs are totals over the same 13432 scoring calls (every candidate of every question) for both arms, under one model revision, dtype, context length, memory fraction and logprob setting; the only declared difference is how many candidate images a call carries, which is the interface being compared. They exclude the shared retrieval index, feature extraction and any downstream question answering.

K.1 INPUT CANONICALISATION AND SERVING REPEATABILITY

Sorting candidates by their original retrieval rank restores the input before a T2 permutation (§6). The check covers everything the reader receives: candidate identifiers, intervals, contact sheets and prompt text.

The sorting key is the original rank recovered from each immutable clip identifier, with ties broken by that identifier. Across 19,791 question–permutation pairs, 4 cells and 2 scoring arms, 0 prompts matched before sorting. After sorting, the ordered identifiers (19791), interval bounds and sheet paths (19791), system prompts (19791) and serialized user prompts (19791) all match. Thus $K(\pi(C)) = C$.

Output equality, $F(K(\pi(C))) = F(C)$, additionally depends on repeatable execution. Re-running 100 stored prompts twice under one configuration reproduced 100 generated texts and 100 selections exactly. At a different batch size, 51 of 100 texts and 99 of 100 selections matched. The changed selection resolved cleanly on both sides. Batch composition can change floating-point arithmetic and the outcome of near-tied greedy decisions.

Canonicalisation removes candidate-order variation when its key survives the perturbation. Re-extraction can change retrieval ranks themselves (Appendix K.10); Appendix K.12 therefore evaluates a key based on clip time and immutable identity. Input equality and serving repeatability are checked separately in both experiments.

K.2 Re-presentation stability per interface

The re-presentation control of §C.3 reorders whole candidate tuples. Which interfaces that can reach depends on their prompts: direct selection and joint-context scoring put the packet listing and every sheet in one call, so the ordering is visible to them, while the per-candidate scorers put one candidate in each call, so no candidate-order permutation reaches their prompt. Identical scores additionally require a reproducible execution path. We ran the first group under the same three recorded permutations and carry the second with that note rather than an equal number of calls (955, 999, 384 and 439 questions, three variants each).

Natural-output and fixed-answer tracks. On the *natural-output* track the question is answered again under each ordering, so the answer may change with the citation; `A0` reads the first resolvable citation of that output and `A1` the member of its cited set with the lowest original retrieval rank. On the *fixed-answer* track one answer is frozen and only the citation interface varies. A contrast within a track is controlled; one across them is not, because the natural track carries answer variation the fixed-answer track does not. Table 43 labels each row with its track and we read the two within-track contrasts first.

Sorting the cited set reduces part of the instability. On the natural track the order-invariant reading lowers the flip rate in every cell (-5.8 $[-7.1, -4.5]$, -6.1 $[-7.3, -5.0]$, -1.7 $[-2.9, -0.6]$ points, and -1.0 $[-2.1, +0.3]$ on the fourth, whose interval contains zero) and never near zero: the lowest rate anywhere is 41.0%. This reproduces on these cells what Appendix D.2 found on validation—the list-order component is real but small, and what remains is a changed cited set, which sorting cannot undo. `A1`'s flip rate is zero only when one stored output is reused for every ordering, which is a property of that reuse and not of the rule.

On the fixed-answer track the two order-sensitive interfaces differ by reader and dataset. Joint context flips less than direct selection on NExT-GQA (-6.6 $[-9.2, -4.1]$ points) but the same contrast on VideoZeroBench is $+2.9$ $[-0.9, +6.9]$, whose interval contains zero. Across the tracks direct selection flips more than the answer-linked policy in 3 of four cells ($+15.0$ $[+12.6, +17.5]$, $+5.9$ $[+1.4, +10.2]$ and $+4.6$ $[+1.2, +8.3]$ against -4.3 $[-6.9, -1.8]$ for InternVL3 on NExT-GQA); every one of those differences excludes zero, but they compare a fixed answer against a re-answered one, and the uncontrolled variation sits on the side that flips *less*, so we report the direction and not a clean interface effect.

No aggregate mIoU change was detected on either track: no signed interval excludes zero, the largest is 1.19 points, and the two threshold directions stay within 1.6 points of each other in every cell, so citations that stop clearing $\mathrm{tIoU} \geq .5$ are replaced about as often as they are lost. Flip rates measure source stability. Appendix K and Appendix K.3 report localization quality and model-assessed support for the same interfaces. The packet-reading interfaces here receive native-order inputs; Appendix K.1 evaluates input canonicalisation separately.

K.3 An auxiliary check on answer support

This earlier support screen compares the intervals selected by different policies for the same frozen answer. It complements the temporal measurements in §4 and Appendix K with model assessments of answer support. The later paired screen in §K.4 instead compares the two clips involved in each citation change (Appendix K.4). Both screens use automatic evaluators. Their cohort and comparison definitions follow §3.1; human validation remains an open target (§7).

Design. From the interface study we took the intervals that `B0`, `A1`, `A2`, `A3` and `AM` actually selected for the same frozen predicted answer, de-duplicated the cases where two policies selected the same interval, and put each (answer, interval) pair to a third model from neither reader's family (`google/gemma-4-12B-it`), which scored how well the interval supports that answer without being told which policy produced it. Every interval is rendered from the *original* video at the unit's

	NExT-GQA test		VideoZeroBench	
	Qwen3-VL	InternVL3	Qwen3-VL	InternVL3
source flip (% of question–variant pairs)				
natural output `A0` first cited	46.8	69.0	46.0	68.0
natural output `A1` lowest rank	41.0	62.9	45.1	66.3
fixed answer `A2` direct	61.8	64.7	51.9	72.7
fixed answer `A4` joint	55.1	—	54.8	—
fixed answer `A3`, `AM`	—	—	—	—
within-track paired difference in flip rate (points)				
`A1`−`A0` *(natural)*	-5.8 $[-7.1, -4.5]$	-6.1 $[-7.3, -5.0]$	-1.0 $[-2.1, +0.3]$	-1.7 $[-2.9, -0.6]$
`A4`−`A2` *(fixed)*	-6.6 $[-9.2, -4.1]$	—	$+2.9$ $[-0.9, +6.9]$	—
across tracks, reported as a direction only				
`A2`−`A0`	$+15.0$ $[+12.6, +17.5]$	-4.3 $[-6.9, -1.8]$	$+5.9$ $[+1.4, +10.2]$	$+4.6$ $[+1.2, +8.3]$
mean absolute ΔmIoU				
`A0` / `A1`	8.59 / 7.80	12.77 / 12.11	3.91 / 3.70	5.14 / 4.67
`A2` / `A4`	11.88 / 9.94	11.71 / —	4.59 / 4.95	5.68 / —
signed ΔmIoU [95% CI]				
`A0`	-0.20 $[-1.19, +0.81]$	-0.31 $[-1.54, +0.93]$	$+0.26$ $[-0.64, +1.14]$	$+0.37$ $[-0.67, +1.42]$
`A1`	-0.33 $[-1.31, +0.66]$	-0.28 $[-1.49, +0.94]$	$+0.11$ $[-0.79, +1.00]$	$+0.35$ $[-0.70, +1.41]$
`A2`	-1.19 $[-2.43, +0.04]$	-0.36 $[-1.53, +0.82]$	$+0.28$ $[-0.69, +1.19]$	-0.34 $[-1.48, +0.82]$
`A4`	-0.74 $[-1.90, +0.39]$	—	-0.36 $[-1.52, +0.78]$	—
hit→miss / miss→hit at tIoU $\geq .5$ *(% of pairs)*				
`A0`	4.7 / 4.3	6.8 / 6.2	2.1 / 2.5	2.1 / 2.4
`A1`	4.4 / 3.9	6.4 / 6.0	2.1 / 2.1	1.8 / 1.9
`A2`	6.8 / 6.1	6.3 / 5.8	2.6 / 3.0	3.0 / 2.7
`A4`	5.9 / 5.0	—	3.5 / 1.9	—

Table 43: The same three recorded orderings applied to each citation policy. *Natural output* re-answers the question under each ordering, so `A0` and `A1` carry answer variation as well; *fixed answer* freezes one answer and varies only the interface. Flip and threshold rates use question–variant pairs as the denominator, so the miss→hit rate minus the hit→miss rate is the change in $R@1_{\text{tIoU}=.5}$ on that same denominator—unlike the flip classes in Appendix D, which are conditional on a flip. Signed and absolute changes average a question's three variants before a video-clustered bootstrap. The `A3` and `AM` inputs are invariant to candidate order by construction: each call contains one candidate. Dashes in that row mark unmeasured repeated-inference flip rates. Joint context was budgeted for the Qwen3-VL cells only.

own boundaries, so the grey control is judged on the real clip it selected from grey sheets: the arms differ in how they chose, never in what the judge was shown. 561 tasks over 166 videos. The screen scores a first-token `Yes`/`No` log-odds at temperature 0, read back from the prompt's own logprobs so that no row depends on a top-k list, and it is run twice: over frames sampled across the whole cited unit, and over the contact sheet the reader was actually served. Two gates were fixed in advance and evaluated on a *separate development sample* of 141 tasks whose videos do not intersect the final one—the screen must tell an interval from the item's own video from one taken from a different video, and it must separate at least the local-scoring and grey-control arms ($+2.60$ $[+0.90, +4.15]$ over 60 blocks on 15 videos). The final sample was evaluated once after these gates passed. The gates check sensitivity to video identity and selection policy.

Support scores. Matched intervals score above mismatched ones by $+9.97$ $[+9.08, +10.88]$ points, higher on 90.9% of items. On the arms, local scoring beats the grey control by $+1.92$ $[+1.34, +2.54]$ and retrieval top-1 by $+0.93$ $[+0.36, +1.52]$, direct selection is slightly below local scoring (-0.48 $[-0.91, -0.05]$), and local scoring against the reader's own lowest-rank citation is $+0.36$ $[-0.06, +0.80]$, with an interval containing zero. Scoring the served frames preserves the directions of the local-minus-grey and direct-minus-local contrasts ($+1.84$ $[+1.23, +2.48]$ and -0.63 $[-1.13, -0.16]$). The `A3`−`A1` interval excludes zero only in the served-frame pass ($+0.62$ $[+0.12, +1.14]$), so that comparison depends on which frames the evaluator sees.

The local-scoring versus grey-control contrast has the same direction on support and mIoU (Supplement S10.3). This complements the interface-quality results in Appendix K. Support scores correlate weakly with temporal agreement on the pooled items (Pearson $r = 0.2061$); this descriptive association combines variation in support, evaluator error and dataset composition. Section K.4 instead tests whether the two criteria identify the same citation changes.

Evaluation scope and development. This screen compares policies using model judgments. Gemma uses a different family from the Qwen and InternVL readers in this screen; shared judgment errors remain possible. Human ratings are a separate evaluation target. The artifact record retains three discarded measurement designs: a saturated four-option format, a top-k readout with systematic missingness, and a masking design whose candidate probabilities did not sum to one. The main support results use the paired screen and calibrated threshold in Appendix K.4; this policy comparison supplies supporting context.

K.4 Scoring both sides of a flip for support

Model-assessed support. For answer-preserving citation changes, a blinded evaluator scores each clip against the question and answer, without annotations, pair identity or the other clip. Loss crosses a fixed score threshold downward or changes `SUPPORTED` to `PARTIAL`/`UNSUPPORTED`; unjudgeable and invalid ratings remain unknown. Pooled transitions are descriptive; interface contrasts remain within natural-output or fixed-answer tracks.

Temporal agreement and model-assessed support. Temporal-hit loss and model-assessed support loss identify different events. On the same 120-pair subset, Gemma 4 31B and Qwen3-VL-8B classify 23.9% and 22.2% of temporal-loss pairs as support losses (Table 58). Score-based checks with Gemma 4 12B and Molmo2-8B also distinguish the criteria (Table 46). These comparisons motivate reporting temporal retention and model-assessed support separately; absolute support rates depend on the evaluator and decision rule (Appendices K.15 and K.13).

Appendix K.3 scores what *different policies* chose on one intact packet; this scores the two intervals a *re-presentation* moved between, which is the comparison the rest of the paper leaves open. For the score-based screen below, design, thresholds and decision rules were recorded before scoring, and development checks preceded evaluation.

Sample. The cohort includes all observed answer-preserving citation changes from NExT-GQA *test* under two readers, two policies and 3 candidate orderings. Qwen3-VL-8B contributes 2,776 pairs and InternVL3-8B contributes 3,842, for 6,618 question–variant pairs over 631 videos. The answer-linked policy A_0 contributes 2910 pairs with unchanged generated answers; direct selection A_2 contributes 3708 pairs with the answer frozen. Table 55 gives per-cell counts. A scoring task is a (question, answer text, interval) triple. Sharing identical triples across policies reduces 6,618 pairs to 3,713 tasks. Intervals are re-rendered from the original video at their own boundaries, without audio or overlays.

Judge and blinding. `google/gemma-4-12B-it`, outside both readers' lineage (Qwen3-VL-8B is Qwen3 and InternVL3-8B pairs InternViT with a Qwen2 language model, so the readers already share one). It never sees which policy chose an interval, which side is intact, the annotation, or the other side of the pair. Each interval is scored *alone*, removing candidate order from the judging input. The score is the first-token log-odds of `Yes` against `No` at temperature 0, read from the prompt's own logprobs. Two passes are reported separately and never averaged: frames sampled across the whole cited unit, and the contact sheet the reader was actually served.

Recorded evaluator checks. *G1, annotation recovery*, requires higher scores for the annotated-hit side when exactly one side clears tIoU $\geq .5$. The differences are $+0.976$ $[+0.483, +1.458]$ over 1292 pairs on the whole-unit pass and $+0.517$ $[+0.024, +0.988]$ on the served-frame pass, both excluding zero. *G2, mismatch control*: on a 620-task subset, an interval from the item's own video outscores one deranged to a different video by $+10.40$ $[+9.77, +11.05]$, higher on 93.9% of tasks, reproducing the P4 screen's $+9.97$ $[+9.08, +10.88]$.

A second evaluator (G3). `allenai/Molmo2-8B` rescored the full sample. Its Qwen3-8B backbone shares the readers' broader Qwen language-model family, while its vision stack and judging prompt differ. The agreement analysis was recorded as informational. Both evaluators pass annotation recovery on the 1292 decisive pairs in both frame passes. Their paired differences correlate at Spearman $\rho = 0.44$ for whole units and 0.38 for served frames. Pair-level sign agreement is 63.7% and 60.6%, respectively.

	whole unit		served frames	
Of 6618 flips	judge loss	no judge loss	judge loss	no judge loss
annotation hit→miss	51	620	92	579
no annotation hit→miss	316	5631	484	5463

Table 44: Temporal and model-assessed support losses on the same cohort. Temporal hit-to-miss events occur on 10.14 [8.46, 11.85]% of pairs and evaluator support losses on 5.54 [4.43, 6.72]%. Of evaluator losses, 86.1% (whole unit) and 84.0% (served frames) occur without a temporal hit-to-miss event. On the whole-unit pass, 267 of the 316 such pairs are temporal miss-to-miss, 11 are hit-to-hit and 38 are miss-to-hit. The first row counts temporal losses, including those shared with the evaluator. The conditional rates here and in §4 use their respective loss populations as denominators.

	whole unit		served frames	
	pairs	Δ support	pairs	Δ support
All flips	6,618	-0.202 $[-0.378, -0.021]$	6,618	-0.049 $[-0.206, +0.110]$
by what the annotation says about the pair				
Intact hit, replacement miss	671	-1.351 $[-1.885, -0.833]$	671	-0.873 $[-1.469, -0.294]$
judge-supported transitions over all flips				
supported → supported	5424	82.0%	3539	53.5%
supported → not	367	5.5%	576	8.7%
not → supported	253	3.8%	576	8.7%
not → not	574	8.7%	1927	29.1%

Table 45: Support scored on both sides of a flip, by a judge blinded to policy, side and annotation. Δ support is the paired change in first-token Yes/No log-odds, video-clustered. The row that matters for §4 is the second: on the pairs that *are* the retention loss, support falls and the interval excludes zero, yet of those whose intact side the judge calls supported, 91.6 [87.2, 95.4]% (whole unit) and 79.4 [71.5, 86.7]% (served frames) move to an interval it also calls supported. Over all flips the mean change is small and the two passes disagree on whether it excludes zero, so we report no aggregate support loss.

The aggregate directions agree. In the retention-loss cohort, both evaluators score replacements lower, with intervals excluding zero: -1.351 $[-1.885, -0.833]$ and -0.438 $[-0.576, -0.305]$ for whole units, and -0.873 $[-1.469, -0.294]$ and -0.346 $[-0.485, -0.209]$ for served frames. Across all flips, the whole-unit change is -0.202 $[-0.378, -0.021]$ for Gemma and -0.012 $[-0.055, +0.031]$ for Molmo2; only Gemma's interval excludes zero. These results replicate the retention-loss direction while quantifying evaluator dependence at the pair level. Effect magnitudes use each evaluator's own log-odds scale.

An incidental replicate. Two workers scored 51 tasks twice with identical inputs but different batch composition. The scores are not bit-identical—mean absolute difference 0.026, maximum 0.114 log-odds—which is the same serving nondeterminism Appendix K.1 measures on the readers, and is two orders below the effects read here.

Reading. The primary quantity is the paired difference $\mathrm{score}(\text{replacement}) - \mathrm{score}(\text{intact})$, which needs no threshold. Classifying support transitions requires a threshold: $\tau = -12.25$, fixed on the P4 development sample alone by maximising Youden's J against the different-video control (AUC 0.836). That criterion separates same-video from different-video intervals. Cells below use "judge-supported" for scores above this calibrated threshold.

Interpretation. Temporal agreement and model-assessed support identify different losses in this cohort: 86.1% of evaluator losses occur outside temporal hit-to-miss events. Their cross-tabulation shows disagreement in both directions (Table 44). The conditional rates use different denominators, and their values depend on the low temporal-hit prevalence in this cohort (Appendix K.13). Together, the measures describe temporal alignment and model-assessed support for each transition. Human support remains unlabelled in the `support_labels_*` namespace (Appendix K.3).

	support transition				
Temporal transition	S→S	S→U	U→S	U→U	pairs
hit → hit	149	11	4	4	168
hit → miss	557	51	22	41	671
miss → hit	527	38	26	30	621
miss → miss	4191	267	201	499	5158
all	5424	367	253	574	6618

Observation	Temporal status	S→S	S→U	U→S	U→U	pairs
whole unit	hit → miss	557	51	22	41	671
	any other	4867	316	231	533	5947
served frames	hit → miss	355	92	57	167	671
	any other	3184	484	519	1760	5947

Table 46: Every temporal transition crossed with every support transition, on the whole-unit pass (top), and the same crossing with the three non-hit-to-miss temporal states pooled, for both observations (bottom). The top panel is the grid the collapsed view hides: a support loss under *miss* → *hit* is a pair whose temporal status improved, and one under *hit* → *hit* a pair whose status did not move, so neither is silence. Column totals match the four support-transition counts in Table 45 and the *hit* → *miss* row is the retention-loss stratum. "S" is evaluator-supported at the threshold above, not a human verdict.

	all pairs	both sides cited
pairs	3000	920
source flip (%)	35.33 [32.87, 37.81]	72.72 [69.34, 75.98]
signed ΔmIoU	+0.15 [−0.65, +0.94]	−0.57 [−2.73, +1.57]
mean absolute ΔmIoU	4.87 [4.14, 5.63]	10.95 [9.25, 12.73]
initially-correct retention loss (%)	33.33 [23.96, 43.01]	70.00 [57.14, 81.94]
on how many initially-correct pairs	189	70

Table 47: The re-presentation control on an X-CLIP packet. The left column is the recorded analysis. The right restricts to pairs where both sides named an in-packet unit, and is a diagnostic added after the abstention rate was seen, not a recorded cohort: a pair where neither side cites is retrieval top-1 on both sides and is a forced non-flip, so the left column is diluted by a population that is more than half frozen. Both columns show the same three things the paper reports on the original retriever—the cited clip moves, no change in the aggregate is detected, and citations that were already correct are lost—and the right column shows them larger, on a smaller base. Abstention itself responds to the ordering, running from 47.7% to 58.4% across the four arms.

K.5 The same control on a packet built by a second retriever

Supporting §7. This replication replaces LanguageBind with X-CLIP while holding the remaining evaluation conditions fixed: the same 1000-question recorded sample of Appendix K on NExT-GQA *test*, the same reader and decoding, the same three permutations, the same frame bank, the same reading rule and the same bootstrap. The second retriever is X-CLIP on the identical window geometry, scored with its own video-conditioned similarity rather than a plain dot product, and selected with the same NMS and the same padding rule. Design and decision criteria were recorded before the reader was run.

The two packets really differ. On the sampled questions the two retrievers agree on the top-ranked unit for only 10.3% of questions, and the mean best match between their four delivered intervals is 0.368. The replication therefore changes the delivered evidence.

Candidate quality and abstention. X-CLIP delivers an overlapping unit for 21.5% of the sampled questions against 31.8% for the retriever the paper uses—a gap of 10.3 points, consistent with the earlier selection of that retriever over this one. The consequence appeared only after the reader ran: given a packet with less to cite, the reader abstains on 52.9% of questions against 4.1% on the original packet, and an abstention resolves to retrieval top-1 on both sides of a pair, which cannot flip by construction. The resulting flip rates are conditional on each retriever's candidate quality and abstention rate. We report both the full cohort and pairs with two resolved citations.

Replication result. Both cohorts in Table 47 reproduce the pattern. The full-cohort results meet all three recorded criteria: a flip interval above 15%, a retention-loss interval above zero and a signed-change interval containing zero. The replication covers one reader and split under X-CLIP's lower evidence delivery and higher abstention rates.

K.6 Retention across additional packet-provider adapters

We apply the same paired evaluation to packets from the FOCUS core selector and a VimRAG-derived embedding-search adapter (Zhu et al., 2026; Wang et al., 2026a). FOCUS uses its released selection component with CLIP-L similarities; the VimRAG adapter uses Qwen3-VL embeddings to search within the question's source video. Both receive the question and answer options. The stored packet boundaries and frames come from an earlier provider audit in this project. Each provider contains the same 200 previously evaluated NExT-GQA validation questions from 165 videos. These runs test provider transfer under a common citation interface.

Fresh paired inference. We rebuild timestamped contact sheets from the recorded frames and retain clip IDs and intervals. Three distinct nonidentity orders are sampled per question with seed 20260921; an additional repeat uses exactly the original serialized input. Two VimRAG questions have three candidates, and two units contain three recorded frames; the remaining packets have four candidates and four frames per unit. These differences remain fixed across orders. Qwen3-VL-8B and the official Hugging Face conversion of InternVL3-8B each generate fresh answers for all conditions, yielding 4,000 responses. The checkpoint revisions are `0c351dd0` and `259a3b64`, respectively. Both use greedy decoding, a 384-token cap, disabled prefix caching and the current reader prompt without provider captions. The InternVL conversion and current runtime differ from the historical core-reader setup.

Scoring and uncertainty. The first resolvable cited ID selects a provider unit; an unresolved output falls back to original provider rank zero, independently of display order. Its full unit interval is scored against temporal annotations using the recorded decimal timestamps. We average the three alternatives within question and use 10,000 video-cluster bootstrap draws. Table 48 reports conditional initial-hit loss alongside all-pair loss and gain, identical-input repeats, and paired reordering-minus-repeat contrasts. Both-resolved and unchanged-answer cohorts expose sensitivity to output parsing and answer changes.

All four cells lose initial temporal hits more often under reordering than under identical-input repeats: the paired excess-loss intervals exclude zero. Losses also remain in the both-resolved and unchanged-answer cohorts. The aggregate changes differ across cells; the all-pair loss and gain columns explain how they combine into the net hit-rate change. Paired transitions measure retention across providers even when the aggregate score changes.

Input and execution checks. All 6,390 recorded frame files decode successfully. One dark frame was verified by decoding the source video at nearby timestamps; it is retained. The artifact reports the sensitivity excluding its video and the sensitivity restricted to four-candidate packets. The artifact also records all raw outputs, input hashes, complete decoding settings, parser failures, truncated responses and runtime initialization retries.

K.7 Temporal-hit retention at 32B parameters

Supporting §7. We test whether citation instability persists at 32B parameters by running Qwen3-VL-32B under the same control: the same `T2` prompt file the 8B readers read, byte for byte, the same scorer and the same video-clustered bootstrap, on 3358 validation questions over 567 videos. Design and thresholds were recorded before the reader was run.

Retention and uniform replacement. Reanalysis of the stored 32B outputs gives 20.3 $[17.8, 22.9]\%$ temporal-hit loss: 338 of 1,668 initial-hit pairs, from 556 questions in 331 videos. Independent uniform replacement over the same four candidates loses 67.8 $[66.6, 68.8]\%$ on this cohort; the paired reader-minus-uniform difference is -47.5 $[-50.1, -44.8]$ points. Each question contributes three variants, averaged before 10,000 video-cluster bootstrap draws with the original seed.

Reader	Provider	Initial-hit questions	Source flip	Initial-hit loss	Signed ΔtIoU	Repeat loss
Qwen3-VL-8B	FOCUS	45	39.0 [34.0, 43.9]	27.4 [17.6, 38.3]	−2.9 [−5.0, −0.9]	0.0 [0.0, 0.0]
Qwen3-VL-8B	VimRAG	54	43.5 [38.2, 48.9]	25.3 [17.4, 33.9]	−1.8 [−4.3, 0.7]	0.0 [0.0, 0.0]
InternVL3-8B	FOCUS	45	52.2 [47.0, 57.3]	34.8 [23.3, 46.8]	−3.5 [−6.0, −1.1]	0.0 [0.0, 0.0]
InternVL3-8B	VimRAG	48	56.0 [51.5, 60.4]	45.8 [37.2, 55.0]	−2.8 [−5.7, 0.1]	0.0 [0.0, 0.0]

Reader	Provider	All-pair loss	All-pair gain	Net ΔR@.5	Repeat flip	Excess hit loss
Qwen3-VL-8B	FOCUS	6.2 [3.6, 9.1]	2.2 [0.8, 3.9]	−4.0 [−7.4, −0.8]	0.5 [0.0, 1.6]	27.4 [17.6, 38.3]
Qwen3-VL-8B	VimRAG	6.8 [4.4, 9.6]	5.3 [2.9, 8.0]	−1.5 [−5.2, 2.2]	0.0 [0.0, 0.0]	25.3 [17.4, 33.9]
InternVL3-8B	FOCUS	7.8 [4.7, 11.2]	2.5 [1.1, 4.1]	−5.3 [−9.1, −1.8]	0.5 [0.0, 1.6]	34.8 [23.3, 46.8]
InternVL3-8B	VimRAG	11.0 [7.7, 14.5]	6.3 [3.9, 8.9]	−4.7 [−9.4, −0.2]	0.5 [0.0, 1.6]	45.8 [37.2, 55.0]

Reader	Provider	n_{res}	Resolved loss	n_{ans}	Same-answer loss	Set change	Excess flip
Qwen3-VL-8B	FOCUS	129	28.7 [18.6, 39.9]	121	29.8 [19.5, 40.9]	51.7 [46.8, 56.6]	38.5 [33.6, 43.3]
Qwen3-VL-8B	VimRAG	157	26.1 [17.9, 34.9]	144	23.6 [15.3, 32.5]	51.0 [45.7, 56.4]	43.5 [38.2, 48.9]
InternVL3-8B	FOCUS	129	34.1 [22.6, 46.4]	118	33.1 [21.4, 45.5]	63.5 [58.2, 68.7]	51.7 [46.4, 56.8]
InternVL3-8B	VimRAG	142	45.8 [37.1, 54.8]	138	46.4 [38.0, 55.1]	59.5 [55.1, 63.9]	55.5 [51.1, 59.9]

Table 48: Fresh provider-adapter audit with 95% video-cluster bootstrap intervals. Rates are percentages; signed tIoU, net hit-rate changes and excesses are points. Initial-hit loss conditions on the original temporal hits. Each initial-hit question contributes three reordering pairs; n_{res} and n_{ans} count initial-hit pairs with both citations resolved and with unchanged answers, respectively. Excesses subtract identical-input repeat rates using paired video draws. Both providers use the common citation interface.

Tuple reordering (T2)	Qwen3-VL-8B	InternVL3-8B	Qwen3-VL-32B
Signed ΔmIoU	−0.22 [−0.69, +0.26]	−0.56 [−1.19, +0.08]	+0.14 [−0.29, +0.58]
Mean absolute Δ	6.91 [6.49, 7.35]	10.46 [9.92, 11.00]	5.85 [5.44, 6.28]
Initial-hit loss (%)	23.2	42.4	20.3 [17.8, 22.9]
Source flip (%)	40.2	59.0	33.8 [32.5, 35.1]
Any-variant flip (%)	66.8	86.2	56.3
R@.5 gain / loss (%)	3.75 / 3.61	5.44 / 5.68	3.55 / 3.35
Flips that are miss→miss (%)	79.2	79.0	76.6
hit→miss (%)	9.0	9.6	9.9

Table 49: The same reordering control across three readers. At 32B, initial temporal hits are lost despite small net aggregate changes. Initial-hit loss conditions on each reader's own reference hits; core 8B CIs are in Table 9. Differences between reader identity, size and serving configuration are not isolated by this comparison.

Restricting to pairs with both citations resolved gives 20.6 [17.9, 23.6]% loss (288/1,396 eligible initial-hit pairs); restricting to equal nonempty answer-option labels gives 19.9 [17.1, 22.9]% (260/1,305 pairs). Their matched uniform losses are 68.9 [67.8, 70.0]% and 69.0 [67.8, 70.0]%, respectively. These restrictions describe subsets of the generated outputs.

On all validation questions, initial hit rate is 16.6 [15.2, 17.9]%, versus 9.7 [9.0, 10.3]% for uniform selection. A hit candidate is available on 31.4 [29.5, 33.4]% of questions; among those questions the reader initially hits on 52.7 [49.5, 55.9]%. This separates candidate availability from selection quality. Reader loss remains below uniform replacement, while reordering still loses achieved hits.

This exploratory reanalysis uses historical inputs and outputs. Delivered intervals in preserved prompts reproduce the archived scores to their stored precision; the original scorer reproduces the published gain/loss, source-flip and signed-change estimates and CIs. Artifact `AR24` contains source hashes, all denominators, candidate-count strata and independent verification. Stored responses contain no reader logprobs or candidate-score margins, so these results do not measure high-confidence retention.

Instability persists at 32B. The recorded criteria tested whether the flip interval exceeded 15% and whether it fell below the lower 8B rate. Both hold: 33.8 [32.5, 35.1]% of pairs change their cited clip at 32B, below either observed 8B rate. Signed mIoU change is +0.14 [−0.29, +0.58]

and $\Delta R@.5$ is $+0.20$ $[-0.48, +0.87]$, with both intervals spanning zero. Mean absolute change is 5.85 $[5.44, 6.28]$ (Table 49).

Quality and stability across readers. In this comparison, InternVL3-8B has 18.9 mIoU and 59.0% flips; Qwen3-VL-8B has 21.0 and 40.2%; Qwen3-VL-32B has 22.1 and 33.8%. Higher quality coincides with fewer flips across these three readers. Within each reader, however, substantial citation movement coexists with a small signed mean change. For example, Qwen3-VL-8B moves by 6.91 $[6.49, 7.35]$ in absolute terms within a signed interval of -0.22 $[-0.69, +0.26]$ (§C.3). Even the highest-quality reader changes its citation on 33.8 $[32.5, 35.1]$% of pairs. This is the distinction measured in §3.1.

Comparison scope. Reader identity, parameter count and serving configuration vary together. The 32B run uses tensor parallelism 2, while the 8B runs use 1. Appendix K.1 measures serving sensitivity separately. The three-reader comparison establishes persistence at 32B on this split and retriever; a scaling analysis would require matched configurations across more sizes.

K.8 Retention after selecting a reader for grounding quality

Gemma-4-31B achieves higher original-order mIoU and lower retention loss than a fresh InternVL3-8B reference, including on questions initially hit by both readers (Table 50). Remaining losses are measured separately from mean quality. This follow-up selects a reader using grounding quality before measuring retention on different videos.

Development selection and evaluation. We partition the corrected-frame NExT-GQA validation pool by video before inference. Development contains 80 questions from 54 videos; evaluation contains 240 questions from 186 nonoverlapping videos. Each sample contains equal numbers of questions with and without a top-four temporal-hit candidate; weights restore each stratum's question count within its video partition. The recorded gate requires both initial mIoU and $R@.5$ to exceed the fresh InternVL point estimates, at least 90% resolvable citations and at most 5% truncated outputs. Among eligible readers, the highest development mIoU determines selection. Gemma's development mIoU/$R@.5$ are 19.0/15.2, versus 15.3/9.9 for InternVL and 14.7/11.4 for Pixtral-12B. Resolvable outputs number 78, 78 and 63 of 80, respectively. Gemma passes; Pixtral does not. No retention outcome enters this decision.

Both evaluated readers receive the same corrected contact sheets, clip IDs, intervals, task text and orders: one original, three seeded alternatives and one identical-input repeat per question. Greedy decoding uses a 1,024-token cap and two concurrent requests. Model-specific chat templates and image processors differ. Gemma uses two GPUs and InternVL one; each reader's configuration stays fixed across orders. Scoring uses the first resolvable clip ID's complete packet interval, with original top-1 fallback and 10,000 video-cluster bootstrap draws. Artifact `AR25` records checkpoint revisions, settings, sampling weights, raw outputs and independent verification.

What improves, and what remains unstable. The paired evaluation mIoU advantage is 3.5 $[1.3, 5.8]$ points; the $R@.5$ difference is 2.6 $[-0.5, 5.9]$ points and remains uncertain. Gemma's own-hit loss is lower than InternVL's, but these rates condition on different initial-hit sets (64 versus 54 questions). Gemma's 13.0% loss is below its matched uniform-replacement loss and persists with both citations resolved and with unchanged answer-option labels. Its small net hit-rate change combines losses and gains on different questions.

Shared initial-hit comparison. On the shared initial-hit cohort, Gemma loses 17 pairs and InternVL 49; the paired InternVL-minus-Gemma difference is 26.7 $[13.3, 40.0]$ points. This exploratory reanalysis averages three variants within question and retains sampling weights, with 10,000 paired bootstrap draws over all evaluation videos. Jointly resolved pairs give 28.2 $[15.1, 41.5]$ points. Common hits need not cite the same clip: restricting to the 33 questions (32 videos) with the same initial source gives 26.3 $[11.4, 41.2]$ points. Table 50 separates these cohorts; artifact `AR26` records source hashes, cohort IDs and independent raw-output and bootstrap verification. Lower Gemma loss on shared initial hits cannot be attributed solely to different initial-hit sets; the same-source subset also holds the initial clip fixed. Quality and retention improve together

Measure	Gemma-4-31B	InternVL3-8B
Initial mIoU	21.7 [18.8, 24.6]	18.2 [15.4, 21.0]
Initial $R@.5$ (%)	16.9 [13.1, 21.0]	14.3 [10.6, 18.3]
Initial-hit questions	64	54
Lost / initial-hit pairs	25/192	75/162
Initial-hit loss (%)	13.0 [7.3, 19.4]	46.3 [36.4, 56.1]
Both-resolved loss (%)	13.0 [7.3, 19.4]	47.2 [37.3, 56.9]
Same-answer loss (%)	14.1 [7.9, 21.1]	46.0 [34.9, 56.5]
Uniform replacement loss (%)	69.5 [66.7, 72.1]	67.6 [64.1, 70.7]
Reader − uniform (pp)	−56.5 [−62.3, −50.1]	−21.3 [−30.9, −12.0]
Source flip (%)	31.7 [26.6, 36.9]	61.5 [57.3, 65.7]
Signed ΔmIoU	0.7 [−0.7, 2.2]	−2.0 [−4.1, 0.1]
All-pair hit→miss (%)	2.2 [1.2, 3.4]	6.6 [4.4, 9.1]
All-pair miss→hit (%)	3.2 [1.9, 4.6]	4.8 [3.2, 6.4]
Net $\Delta R@.5$ (pp)	1.0 [−0.8, 2.8]	−1.9 [−4.9, 1.0]
Repeat source flips / hit losses	0 / 0	0 / 0
Parse failures / fallback responses	0/46	7/7
Shared initial-hit questions: paired comparison		
Common-cohort loss (%)	14.2 [6.3, 23.0]	40.8 [29.3, 52.5]
Jointly resolved common-cohort loss (%)	13.7 [5.7, 22.7]	41.9 [30.2, 53.5]
Same initial source, common-cohort loss (%)	17.2 [7.8, 27.5]	43.4 [31.1, 56.0]

Table 50: Fresh reader evaluation on the same 240 questions, 186 videos and 1,200 responses per reader. Estimates use sampling weights; counts are raw. The upper block conditions loss on each reader's own initial hits; the lower block fixes the same 40 initially correct questions (39 videos, 120 pairs) for both readers; jointly resolved pairs number 117, and same-source pairs 99. In the upper block, both-resolved eligible pairs number 192/159; same-answer pairs number 177/137. All intervals are 95% video-cluster CIs. Neither reader has truncated outputs. Repeats contain 240 pairs per reader; zero observed events do not establish deterministic serving.

in this comparison. PCRA quantifies both the improvement and remaining loss, which aggregate mIoU alone does not determine.

This is an exploratory follow-up on previously accessed validation data: development and evaluation videos are disjoint from each other, but both overlap earlier studies. The comparison varies reader family, size and processing jointly; it does not isolate a parameter-scaling effect.

K.9 Candidate availability and reader selection

Low aggregate temporal-hit rates combine two components: a packet may contain no hit candidate, or the reader may fail to choose one that is available. A candidate is available here if at least one of the four delivered clips satisfies $\mathrm{tIoU} \geq .5$. Since a selected hit must come from such a packet,

$$\Pr(\text{hit}) = \Pr(\text{available}) \Pr(\text{hit} \mid \text{available}).$$

The all-question miss rate is the sum of candidate unavailability and an available candidate not being selected. This is an accounting identity for delivered clips; both retrieval and clip interval geometry affect availability.

Table 51 applies this decomposition to the matched reader comparison in Appendix K.8. The raw sample contains 120/240 questions with an available candidate; inclusion weights give the eligible-population estimate 31.77%. Gemma and InternVL have identical candidate packets. Their conditional hit rates exceed uniform selection by 23.33 [14.89, 31.76] and 15.00 [6.25, 23.72] points, respectively. Availability constrains aggregate hit rates, while reader selection errors remain among packets that contain a hit.

The separate Qwen32 validation ledger gives the same decomposition on 3,358 questions: availability is 31.42%, and hit rate conditional on availability is 52.70%. We do not pool this historical cohort with the fresh matched-reader sample. Artifact `AR28` provides interval-based verification, counts stratified by the number of hit candidates, resolved-output sensitivity and oracle-candidate mIoU. Availability refers to temporal annotation overlap, not human judgments of semantic support or calibrated reader confidence.

Measure	Gemma-4-31B	InternVL3-8B	Qwen3-VL-32B
Questions / videos	240/186	240/186	3,358/567
Initial-hit questions	64	54	556
Candidate available (%)	31.77 [26.31, 37.35]	31.77 [26.31, 37.35]	31.42 [29.47, 33.35]
Hit given available (%)	53.33 [45.00, 61.62]	45.00 [36.00, 53.91]	52.70 [49.53, 55.93]
Uniform hit given available (%)	30.00 [28.25, 31.94]	30.00 [28.25, 31.94]	30.73 [30.00, 31.49]
Overall $R@.5$	16.94 [13.14, 20.99]	14.30 [10.60, 18.30]	16.56 [15.19, 17.94]
No hit candidate (all-question %)	68.23 [62.65, 73.69]	68.23 [62.65, 73.69]	68.58 [66.65, 70.53]
Available but not selected (all-question %)	14.83 [11.27, 18.64]	17.47 [13.50, 21.65]	14.86 [13.48, 16.27]

Table 51: Temporal-hit availability and selection. Gemma and InternVL share one weighted 240-question sample; Qwen32 uses a separate full-validation cohort. The last two rows sum to the all-question miss rate. All intervals use 10,000 video-cluster bootstrap draws with each source study's weights and random stream. Uniform selection chooses one of the four candidates; its conditional hit rate accounts for the number of hit candidates in each packet.

Core 8B audit. The same decomposition on all 3,358 validation questions (567 videos) uses one original-order response per question, equal question weights, the primary exact-boundary scorer and its retrieval-top-1 fallback. A hit candidate is available on 1,055/3,358 questions (31.42 [29.47, 33.35]%). Qwen3-VL-8B and InternVL3-8B hit 523 and 449 questions, giving conditional hit rates of 49.57 [46.61, 52.57]% and 42.56 [39.35, 45.77]%. Uniform selection reaches 30.73 [30.00, 31.49]%; paired excesses are 18.84 [15.90, 21.80] and 11.82 [8.70, 14.87] points. With resolved original citations only, conditional rates are 51.29 [48.19, 54.43]% and 42.61 [39.41, 45.81]%.

Candidate intervals use the same top-four reconstruction as the primary exact retention analysis. Excluding the 4 known index-discrepancy questions leaves 1,051 available packets and both initial-hit counts unchanged; conditional rates become 49.8% and 42.7%, respectively, with uniform expectation 30.7%. CIs use 10,000 video-cluster bootstrap draws (seed 20260908). Artifact `AR30` retains the question ledger, full decomposition, exclusion sensitivity and verification. This extends the availability analysis to the core readers without treating temporal candidate availability as semantic support or calibrated confidence.

K.10 Does the packet reorder on its own?

Supporting §6, we test whether feature re-extraction changes candidate order in the retrieval pipeline. The experiment uses 3,358 validation questions over 567 videos with retained window features. Design and thresholds were recorded before the features were re-extracted.

Identical extraction settings reproduce the packet. Re-extracting every video with the same model, revision, window geometry, frame sampling and batch size returns embeddings that are bit-identical on 567 of 567 videos, so the delivered packet—set, order and top-ranked clip—is reproduced exactly. This rerun verifies packet repeatability under the tested extraction settings.

Changing batch size changes candidate order. Changing the feature-extraction batch size from 2 to 8, with the model, revision, geometry, frames and preprocessing all held, leaves *no* video bit-identical: the scores move by 1.3×10^{-3} relative at the median and 8.4×10^{-3} at the 95th percentile. Re-running the same selection rule on them, the delivered set changes on 1.10 [0.79, 1.44]% of questions, the same four clips arrive in a different order on 2.83 [2.27, 3.42]%, and the top-ranked clip changes on 0.62 [0.38, 0.88]% (Table 52).

Natural reordering changes the citation. Serving the re-extracted order to the same reader, on the 95 questions it reordered, the cited clip changes on 27.5 [18.3, 37.6]% of the 80 pairs where both sides named an in-packet unit (73 videos; 15 pairs had no readable citation on one side). The signed change is $+0.28$ [-2.56, $+3.54$] mIoU, with an interval including zero, while the mean absolute change is 4.12 [1.52, 7.36]. Multiplying the resolved-pair rate by the observed order-change fraction gives an extrapolated 0.78 [0.52, 1.06]% of the split (about one in 128). This assumes the same flip rate for unresolved pairs and holds the observed trigger fraction fixed in its interval; it is not a directly

	bit-identical videos	set changes	order only	top-1 changes
real re-extraction				
same settings, re-run	567/567	0.00	0.00	0.00
batch size $2 \rightarrow 8$	0/567	1.10	2.83	0.62
full precision	0/567	1.13	3.01	0.66
modelled perturbation, for comparison				
relative 10^{-3}	—	2.50	5.58	1.55
relative 10^{-2}	—	24.10	27.50	13.35

Table 52: What a re-extraction does to the packet, in per cent of questions. *Order only* requires the delivered set to be unchanged, which is the condition the reordering result applies to. The modelled rows apply independent Gaussian noise of the stated relative size to the window scores. The float16 storage-precision reference for the stored query is 5.0×10^{-5}, about twenty-five times smaller than the score movement induced by changing batch size. At a matched perturbation size, independent Gaussian noise predicts 5.58% reordering, compared with 2.83% under re-extraction. Independent noise disturbs the ranking more than the correlated score shifts observed in this re-extraction.

measured all-question churn rate or a lower bound. The same questions under the three recorded permutations flip at 46.4 $[38.1,\ 55.2]$% over 239 pairs; we report that next to it descriptively, because a re-extraction usually swaps adjacent ranks while the experimental permutations are non-identity shuffles, so the two are not interventions of one strength.

Candidate membership changes are a separate outcome. Re-extraction leaves 3,226 questions unchanged, reorders the same four clips on 95 questions, and changes the candidate set on 37 questions. The order experiment uses the middle group, where the delivered evidence stays fixed. The set-changed group is counted separately; its new units lack rendered contact sheets and were not scored by the reader.

Batch size also affects reader execution. Re-serving identical prompts at a different batch size preserves 99 of 100 selections (Appendix K.1). These experiments locate two routes from configuration to citation changes: retrieval can alter the candidate order, and reader execution can alter the output for the same input. Their rates are measured on separate cohorts.

Time-based keys cover retrieval-rank changes. Sorting by original retrieval rank restores an applied permutation (§6). When re-extraction changes that rank, a key based on clip time and immutable identity preserves a common order. Appendix K.12 tests this key and measures serving repeats on the resulting identical inputs.

Observed trigger. A batch-size change produces same-set reordering and citation movement in this pipeline; an otherwise identical extraction rerun preserves the packet. These measurements cover one machine, driver, split, retriever and reader. They identify a natural trigger for the controlled intervention, with occurrence rates specific to this configuration change.

K.11 The estimand of an order-sensitive comparison

Condition comparisons on serving order. The Qwen–InternVL gap remains positive across all four observed schedules, whereas reading-rule order ranges exceed the intact-order CI widths. Both reading-rule mean gaps have CIs spanning zero, and within-schedule sampling accounts for most joint variation. The analysis distinguishes fixed-order CIs, schedule-mean CIs and joint sensitivity intervals (Table 53). These finite-design comparisons do not estimate uncertainty across arbitrary deployment orders.

Table 53 uses an exploratory reanalysis of the stored validation outputs: 3,358 questions, 567 videos, 10,000 paired video-cluster bootstrap draws, seed 20260919. This reanalysis was specified after the earlier ranking results. The four schedules v are the intact packet and three recorded, question-specific non-identity permutations. The four schedules can contain repeated permutations for individual questions.

Comparison	Intact gap [95% CI]	Four-schedule mean [95% CI]	Joint sensitivity interval
Qwen − InternVL	$+2.154$ $[+1.560, +2.754]$	$+2.413$ $[+1.950, +2.894]$	$[+1.669, +3.241]$
Qwen: first − lowest rank	-0.007 $[-0.061, +0.033]$	-0.072 $[-0.187, +0.034]$	$[-0.286, +0.091]$
InternVL: first − lowest rank	-0.007 $[-0.021, +0.000]$	-0.038 $[-0.115, +0.037]$	$[-0.194, +0.096]$

Table 53: Comparison gaps in mIoU points on 3,358 questions and 567 videos. CIs use 10,000 video-bootstrap draws. The joint sensitivity interval pools the four observed schedules with equal weight and describes their combined variation with video sampling. The reading-rule comparisons use the same generated outputs.

Three targets. Let $G(v) = \mathrm{mIoU}(A \mid v) - \mathrm{mIoU}(B \mid v)$ be a comparison under schedule v. A conditional interval estimates $G(v)$ with v fixed. The finite-design mean $\bar{G} = \frac{1}{4}\sum_v G(v)$ averages the four schedules within every bootstrap draw, keeping the same resampled videos across schedules. Finally, the equally weighted mixture of the four conditional bootstrap distributions describes joint sensitivity to video resampling and selection of one of the observed schedules. Its central 95% range is a sensitivity interval for this finite design. A CI for $\bar{G}$ instead uses the within-draw schedule average defined above. Deployment coverage would require a sampled deployment distribution and separate calibration.

Writing $G^*(v)$ for a bootstrap draw, the mixture variance decomposes exactly as

$$\mathrm{Var}_{D^*,V}(G^*(V)) = \mathbb{E}_V[\mathrm{Var}_{D^*}(G^*(V) \mid V)] + \mathrm{Var}_V(\mathbb{E}_{D^*}[G^*(V) \mid V]) .$$

The first term is the mean within-schedule bootstrap variance; the second is variation of the bootstrap means across schedules. The decomposition describes variation under the declared finite distribution.

Sources of comparison variability. Under the intact order, only 5 Qwen questions and 1 InternVL question have nonzero first-minus-lowest-rank tIoU differences. Reordering increases those counts to 93–101 and 64–81. Consequently the intact-order CI is unusually narrow. The observed ranges, 0.100 and 0.079 mIoU, exceed its width; the variance decomposition below quantifies the relative contributions directly. When all four schedules are included, the RMS within-schedule bootstrap SD is 0.089 and 0.065, compared with observed schedule SDs 0.041 and 0.029. The between-schedule component is 17.3% and 15.8% of mixture variance. Within-schedule sampling variation accounts for the majority. For the reader comparison, the between-schedule share is 24.6% and each conditional CI remains positive. Both reading-rule means have intervals spanning zero, with small absolute differences.

Reproduction and scope. The script and artifact are `AR19` and its JSON output. The input hash and all four conditional intervals, nonzero counts and variance components are saved there. Synthetic checks cover zero order variation, zero sampling variation and the variance identity. All questions from a sampled video stay together; no partial-video subsampling is used. Historical artifacts also retain the earlier subsample sign-disagreement and range summaries.

K.12 Chronological canonicalisation on the natural-trigger cohort

This exploratory follow-up uses all 95 RX-3 questions whose candidate membership stayed fixed while extraction batch size changed their retrieval order. Each tuple retains its original immutable clip identifier, interval and contact-sheet bytes. Sorting by (start time, end time, clip identifier) must yield the same observable input from both orders. The input audit also enumerates all 24 permutations of each four-candidate packet. This checks the serializer; the separate repeat experiment below measures output agreement.

The quality and repeat experiment freshly serves original-rank, re-extracted and chronological orders, plus a second copy of each chronological input. It uses Qwen3-VL-8B-Instruct at the same checkpoint revision as RX-3, greedy decoding and a 384-token cap through a shared local HTTP server (vLLM 0.16.0, PyTorch 2.9.1, transformers 4.57.6; prefix caching disabled). Two client requests may run concurrently, with an engine capacity of eight sequences; request order is shuffled with seed 20260919, and batching and kernel execution are not pinned. All arms regenerate the answer, so this is a natural-output comparison. Historical RX-3 outputs are not used as the baseline; the reconstructed original-rank prompts match the archived system text, user text and image references

Freshly served condition	mIoU	Fallback count	Length-capped count
Original retrieval rank	19.49	14	9
Re-extracted order	20.13	11	6
Chronological order	19.31	8	6
Chronological repeat	19.31	7	5

Table 54: All 95 RX-3 questions over 87 videos, re-answered under one fresh HTTP serving configuration. No question is removed for parse failure, unresolved citations or length-capped decoding; failures use the declared original-rank fallback. Counts can overlap. The two chronological rows receive identical inputs but are separate requests. Quality differences and their video-clustered intervals are reported in the text, with each reference specified.

Reader	Track	pairs	questions	videos	correct answer	temporal losses	support losses	audited
InternVL3-8B	fixed answer	1938	880	587	1456	190	116	131
InternVL3-8B	natural output	1904	867	581	1492	190	111	130
Qwen3-VL-8B	fixed answer	1770	834	574	1256	194	82	124
Qwen3-VL-8B	natural output	1006	535	420	651	97	58	66
all		6,618			4,855	671	367	451

Table 55: What was collected. Questions and videos are distinct counts and do not add across rows, because a question can flip under both readers. "Temporal losses" is the denominator of the substitution rate and "support losses" that of the missed-loss rate—the two are different events on different bases, which is why they are counted separately here. "Audited" is the number of pairs drawn for the human audit of Appendix K.16; 0 of them have come back labelled, which is the state this paper reports.

on every cohort question. Original RX-3 sheets are retained; this follow-up isolates order rather than repairing the earlier renderer. Missing images fail the run; every result retains its input hash. The first resolvable immutable candidate ID selects the unit, and unresolved outputs use original rank-0, consistently across arms. Video-clustered intervals include those fallback rows. Quality comparisons use this selected cohort. The input-invariance check holds candidate membership, immutable identifiers and image contents fixed.

Results. The serializer restores identical observable inputs on all 95 original/re-extracted pairs and all 2,280 enumerated permutations, checking text, tuple metadata, ordered images and file hashes. The fresh natural-order comparison gives 23.2 $[14.9, 31.9]\%$ source flips (including fallback selections), with ΔmIoU $+0.64$ $[-1.67, +3.26]$. Chronological order changes mIoU by -0.18 $[-3.30, +2.83]$ against fresh original-rank serving; Table 54 supplies absolute levels and failures. Its repeat gives 1.1 $[0.0, 3.3]\%$ source flips, ΔmIoU $+0.00$ $[+0.00, +0.00]$ and exact text agreement on 69/95 pairs. The repeat uses identical inputs, so its output differences measure serving variability. The single repeat source change crosses from a length-capped, unresolved output using fallback to a resolved citation; both selected intervals have the same tIoU. All observed repeat tIoU differences are zero, which produces a degenerate empirical bootstrap interval. Estimating rare repeat failures requires further independent runs.

The outputs, serving configuration, image hashes, input audit and score artifact are in `AR20`; the analysis script checks completeness of all four conditions. The bootstrap uses 10,000 video-cluster draws with seed 20260919 and preserves pairing across arms. These are nominal intervals for this follow-up, with no multiplicity correction and no externally calibrated non-inferiority margin.

K.13 The cohort, and the same rates under other criteria

We examine how the support results in §K.4 vary with cohort composition, temporal metric and threshold, using the existing scored pairs.

Does it survive a correct-answer cohort? The evaluator judges the answer the reader gave, wrong answers included, while the annotation is tied to the gold answer. Restricting to the 4855 pairs whose evaluated answer matches gold removes that mismatch and moves neither rate materially:

Cohort	Criterion		missed support losses	substitutions
all 6618	tIoU	$\geq .3$	77.4 [68.9, 85.1]	81.3 [76.2, 85.9]
		$\geq .5$	86.1 [78.9, 92.5]	83.0 [77.0, 88.6]
		$\geq .7$	89.1 [82.3, 95.0]	80.1 [71.2, 88.4]
	IoP	$\geq .3$	72.2 [63.3, 80.6]	79.3 [74.0, 84.2]
		$\geq .5$	79.8 [71.5, 87.5]	81.2 [76.1, 86.1]
		$\geq .7$	80.9 [72.2, 89.0]	82.0 [76.4, 87.1]
correct only 4855	tIoU	$\geq .5$	86.3 [77.9, 93.4]	83.4 [76.6, 89.9]
	IoP	$\geq .5$	78.5 [68.0, 88.2]	82.7 [76.5, 88.2]
wrong answer 1763	tIoU	$\geq .5$	85.6 [69.8, 96.7]	81.9 [69.3, 91.5]
	IoP	$\geq .5$	83.5 [70.2, 94.3]	77.4 [67.5, 86.3]

Table 56: Both rates under every binarisation we pre-specified, with video-clustered intervals. Denominators change with the criterion: a looser cut makes more pairs eligible to be a temporal loss, so the substitution rate is over a larger base at $\geq .3$ than at $\geq .7$ (1006 against 346 temporal losses), while the missed-loss denominator—support losses recorded by the evaluator—does not move with it. IoP$\geq .7$ is beyond what NExT-GQA reports and is an extension, not a benchmark figure. Support is the evaluator's throughout; with human labels the same table is recomputed on those.

the missed-loss rate is 86.3 $[77.9, 93.4]\%$ against 86.1 $[78.9, 92.5]\%$ on all 6618 pairs, and the substitution rate 83.4 $[76.6, 89.9]\%$ against 83.0 $[77.0, 88.6]\%$.

Does it survive other criteria? Table 56 recomputes both rates at three tIoU thresholds, under NExT-GQA's own IoP criterion, and without any threshold. The missed-loss rate falls as the threshold loosens—at IoP$\geq .3$ it is 72.2% rather than 86.1%—which is the expected direction, since a looser criterion records more hits and therefore more transitions, but it does not approach zero anywhere in the range. Without a threshold the separation is weak: the paired drop in tIoU ranks a support-loss pair above a non-loss pair only 0.568 of the time (IoP, 0.568), against 0.5 for no information. 2027 of the 6618 pairs have both intervals at tIoU zero, where no threshold and no continuous drop can register anything.

Where the rate comes from. The missed-loss rate is high partly because the initial state is usually already a miss: of the pairs whose intact side the evaluator calls supported, 768 start from a temporal hit and 5023 from a miss. The loss rates from those two starting points are close (8.1% against 6.1%), so the imbalance largely reflects how often citations satisfy the annotation in this cohort. A hit-to-miss flag requires an initial temporal hit, so it covers a small share of these pairs. The conditional rates distinguish this base-rate effect from a difference in support-loss risk.

K.14 Categorical support judgments on the full 451-pair sample

Design. The prepared pool contains 6,172 auditable answer-preserving citation changes. A stratified sample of 451 pairs (661 unique clips, 266 videos) carries exact inclusion weights ($N_{\text{auditable}}/n_{\text{sampled}}$ per stratum). Gemma-4-31B and Pixtral-12B rate each silent, timestamped clip independently of its pair, without annotations or pair identity, as supported, partial, unsupported or unjudgeable. Final ratings were allowed only after development gates on 32 separate videos under four conditions (matched, grey, foreign and within-video clips). Pixtral first failed output-format validation and passed after a schema revision on the same development sample. Intervals use 10,000 video-cluster draws with the inclusion weights. These are automatic labels, not human judgments.

Table 57 gives every estimate. Gemma finds temporal losses enriched for support losses under both definitions; Pixtral does not under either. Most support losses occur without a temporal loss for both judges. The next subsection repeats the comparison with a third judge on a 120-pair subset.

K.15 Categorical support judgments from Gemma and Qwen

Temporal and support losses differ. On the same 120 answer-preserving citation-change pairs, Gemma 4 31B IT and Qwen3-VL-8B-Instruct classify 23.9 $[14.0, 35.1]\%$ and 22.2 $[12.3, 33.4]\%$ of temporal-loss pairs as support losses (Table 58). These estimates come from 19 and 17 observed

(%)	Gemma-4-31B strict	Gemma-4-31B partial-positive	Pixtral-12B strict	Pixtral-12B partial-positive
Support loss | temporal loss	25.1 [16.1, 35.1]	25.0 [16.5, 34.6]	12.4 [4.9, 20.9]	13.4 [6.2, 21.9]
Support loss | no temporal loss	14.0 [8.9, 19.6]	11.9 [6.9, 17.3]	10.3 [5.6, 15.6]	12.1 [7.5, 16.9]
Difference (points)	+11.2 [+0.5, +22.3]	+13.2 [+2.8, +24.2]	+2.1 [−7.1, +11.8]	+1.3 [−7.6, +11.1]
Both supported | temporal loss	20.6 [11.9, 30.3]	26.7 [17.1, 36.9]	7.0 [2.9, 12.1]	22.3 [13.9, 32.1]
Support losses without temporal loss	83.7 [74.0, 90.4]	81.4 [69.4, 89.1]	88.5 [76.6, 95.8]	89.3 [80.2, 95.4]
Raw support losses (pairs)	90		39	
Labels: supported / partial / unsupported		199 / 54 / 408		107 / 116 / 438

Table 57: Automatic support judgments on the full 451-pair sample. "Strict" counts only `SUPPORTED` as support; "partial-positive" also counts `PARTIAL`. A support loss is a supported intact clip whose replacement is not supported. Gemma and Pixtral agree on 77.6% of clips (binary, strict), with Cohen's $\kappa = 0.39$ and positive agreement 51.6%.

losses, respectively. Both clips remain supported on 15.7% and 11.8% of temporal-loss pairs. Other pairs can be unsupported initially, so the complement of the support-loss rate does not measure preserved support. Both judges also identify support losses without temporal hit-to-miss. The two criteria therefore identify different transitions.

Matched sample. The comparison uses 211 unique question–answer–interval units over 97 videos, drawn from the earlier 451-pair, 661-unit audit. It contains 80 temporal-loss pairs and 40 comparison pairs without temporal loss. The four strata cross temporal loss with the earlier score-based evaluator's loss flag; their sample sizes are 10, 70, 20 and 20 for both, temporal-only, support-only and neither. Gemma ratings are reused on these exact units; Qwen receives fresh requests. This is an exploratory comparison on previously analyzed data. The subset was fixed before observing Qwen judgments, but the earlier judge results were already known.

Judges and visible evidence. Each request shows one question, one proposed answer and chronological frames from one silent interval. It contains no annotations, previous scores, reader identity, policy, original/reordered role or paired clip. We sequentially decode each clip and sample $\min(32, \max(8, \lceil 2d \rceil))$ frames at uniform midpoint indices, where d is its duration in seconds. Frames preserve aspect ratio with longest side 384 pixels. The two judges receive the same cached frames, timestamps and rubric.

The rubric distinguishes `SUPPORTED`, `PARTIAL`, `UNSUPPORTED` and `UNJUDGEABLE`. Full support requires the answer's core action and relation, including a before/after relation or visible basis for a stated cause. Responses contain a label, a short description of visible evidence and supporting frame timestamps. Only `SUPPORTED` is positive in the primary analysis; counting `PARTIAL` as positive is a sensitivity analysis. Both judges use temperature 0, top-p 1, seed 20260920 and a 384-token output cap. Gemma's thinking mode is disabled.

Development checks. Each judge first evaluates 128 controls: matched, grey, foreign-video and same-video alternative intervals from 32 reserved videos. Both pass the recorded checks for output validity, label diversity, absence of false support on grey and foreign clips, and higher ordinal support on matched than foreign clips. All Qwen responses parse successfully and terminate before the output cap. Qwen labels grey frames `UNSUPPORTED` rather than the rubric's `UNJUDGEABLE`; the false-support gate passes, but this separate abstention-label mismatch remains in the development record.

Estimation and reproducibility. Inverse inclusion weights, recomputed for the subset, target 6,172 auditable citation-change pairs after excluding 40 development videos. Intervals use 10,000 video-cluster bootstrap draws. Invalid and unjudgeable ratings remain unknown; all pairs in this comparison are assessable. Repeated units are judged once. `AR27` records the subset, prompts, frame hashes, raw responses, pair ledgers and verification; `AR21` retains the full earlier audit and additional evaluator runs.

Interpretation. The two judges give similar aggregate support-loss rates within the temporal-loss group, but their binary agreement on unique units is 77.3%, with positive agreement 53.8%

Quantity (%, unless indicated)	Gemma 4 31B	Qwen3-VL-8B
Weighted estimates; only SUPPORTED is positive		
Support loss, given temporal loss	23.9 [14.0, 35.1]	22.2 [12.3, 33.4]
Both supported, given temporal loss	15.7 [7.0, 25.8]	11.8 [5.0, 19.9]
Support loss, given no temporal loss	15.3 [1.3, 31.7]	9.5 [0.0, 23.7]
Difference between loss rates (pp)	+8.7 [−11.0, +25.8]	+12.8 [−4.8, +28.6]
Loss among initially supported	61.5 [24.0, 94.3]	32.3 [6.0, 67.4]
PARTIAL also positive (sensitivity)		
Both supported, given temporal loss	20.9 [11.5, 31.8]	41.4 [27.6, 54.6]
Difference between loss rates (pp)	+9.7 [−9.9, +26.8]	+13.9 [−0.7, +27.0]
Raw counts on the same sampled pairs		
Support losses / temporal-loss pairs	19/80	17/80
Support losses / comparison pairs	7/40	2/40
Initially supported pairs	40	33

Table 58: Gemma and Qwen support judgments on the same 120 pairs (211 unique units, 97 videos). Weighted estimates include 95% video-cluster intervals; raw counts are unweighted. Comparison pairs also have changed citations, but no temporal hit-to-miss. These are automatic judgments.

and Cohen's $\kappa = 0.39$. Their between-group risk differences are $+8.7$ $[-11.0, +25.8]$ and $+12.8$ $[-4.8, +28.6]$ points; both intervals include zero. Counting `PARTIAL` as positive raises the fraction of temporal-loss pairs with both clips supported to 20.9% for Gemma and 41.4% for Qwen. The similar strict support-loss rates therefore coexist with differences in which clips receive support. Judge choice and the decision rule affect the measured support rates; Qwen also shares a model family with one audited reader. Human evidence sufficiency remains a separate target.

K.16 Prepared human-audit sample

Human ratings have not been collected. The prepared sample contains 451 pairs and 661 unique question–answer–interval units over 266 videos. It includes 51 pairs flagged as losses by both automatic criteria, 120 temporal-only, 120 judge-only and 160 neither-loss pairs (Table 44). A separate 40 videos are reserved for calibration. The 120-pair, 211-unit Gemma–Qwen comparison uses a subset of this prepared pool (Appendix K.15); those ratings are model judgments, not human labels.

The supplementary `HUMAN_AUDIT_PROTOCOL.md` retains the full rubric, blinded presentation, sampling weights and analysis plan. It specifies two independent raters per silent clip and a third adjudicator; the four labels distinguish full, partial, absent and unjudgeable support. The planned estimates are human support losses missed by temporal hit-to-miss and the fraction of temporal losses with both clips human-supported. Pair-level inclusion weights and video-cluster intervals account for the sample; unjudgeable labels remain unknown. These are planned human measurements, with 0 labelled pairs.

A rating package for the 120-pair, 211-unit subset is prepared: each unit is a silent clip cut to its delivered interval and named by an opaque identifier, presented in a per-rater shuffled order with the rubric, the question and the proposed answer only; reader, order role, paired clip, annotations and automatic scores are withheld. A separate interface shows only disagreements to the adjudicator. The analysis script implements the estimands above with recomputed inclusion weights and Cohen's κ, and was tested on synthetic labels only. It is released with the supplement.

L Mitigation attempts under fixed evidence

We report four studies, covering five policies, that attempt to reduce initial-hit loss while the candidate evidence stays fixed: answer–citation voting, position-balanced and adaptive voting, pairwise candidate comparison and retention-oriented fine-tuning. All use Qwen3-VL-8B, the corrected contact sheets, the exact-interval scorer and the paired estimands of §3. A positive loss reduction favours the mitigation. Brackets are paired 95% video-cluster bootstrap CIs with 10,000 draws. Table 59 summarises the five policies. None established a loss reduction: every primary interval includes zero. In a post-hoc contrast, four random calls lost 6.6 $[-1.0, 14.3]$ points less than one random

Mitigation (comparator)	Calls	Training	q/v/hits	Loss reduction (pp)	ΔmIoU	Gate outcome
Six-call voting (random 1-call)	6 (1)	None	240/207/98	2.04 [−3.61, 7.58]	0.31 [−0.89, 1.55]	Recorded before inference; CI includes 0
Balanced 4-call voting (random 4-call)	4 (4)	None	240/173/91	−1.10 [−8.79, 6.67]	0.09 [−1.48, 1.72]	Recorded primary; CI includes 0
Adaptive 2/4-call voting (random 4-call)	3.07 (4)	None	240/173/91	−1.10 [−8.75, 6.74]	0.05 [−1.53, 1.67]	Recorded secondary; CI includes 0
Pairwise comparison (fresh A_3)	12 (4)	None	120/116/60	5.00 [−8.33, 18.52]	0.20 [−3.91, 4.47]	Exploratory pilot; CI includes 0
Retention LoRA, dev only (augmented SFT)	1 (1)	LoRA	703/112/114	1.75 [−1.98, 5.15]	0.22 [−0.24, 0.65]	Gate not met; test not run

Table 59: Mitigation attempts: five policies from four studies, each against its own comparator; q/v/hits gives questions, videos and initial-hit questions. Calls are reader calls per question for the mitigation and, in parentheses, its comparator; the adaptive value is the weighted mean. Pairwise and retention calls follow a shared frozen answer generation. The first three rows use inclusion weights; retention uses the unweighted development split, and its ΔmIoU averages four evaluation orders. Samples, interfaces and comparators differ, so rows are not a head-to-head comparison.

call. These results do not show that loss cannot be reduced; they report what these designs, samples and budgets did not detect. A zero-containing interval does not establish equivalence either. All outcomes are temporal annotation measures; human support labels remain at 0 (Appendix K.16).

Six-call answer–citation voting. Appendix D.13 gives the design: the six-call policy and its comparisons were recorded before 3,600 fresh responses on 240 questions from 207 videos disjoint from the exploratory audit, with 98 fresh initial hits. The exploratory reduction of 6.9 $[3.8,\ 10.0]$ points was followed by a fresh reduction of 2.04 $[-3.61,\ 7.58]$ points over random one-call selection, and ΔmIoU of 0.31 $[-0.89,\ 1.55]$. Mean prompt/completion tokens per question are 10813.5/1012.2 for voting and 1802.2/167.9 for one call. *Reading:* fresh validation did not establish a reduction.

Position-balanced and adaptive voting. A reanalysis of stored outputs on the 120-question exhaustive sample (60 initial hits) motivated this policy. Four cyclic rotations, which place each candidate once in each slot, lost fewer hits than four random orders from the same 20-order universe (2.72 $[0.71,\ 4.77]$ points). That estimate is exploratory: the policy was constructed with those outputs known. We then recorded the policy, orbit choice, tie-breaking and comparisons before inference. The sample contains 240 NExT-GQA test questions from 173 videos absent from every earlier test follow-up: 180 questions with a top-four hit candidate and 60 without one, weighted to a population of 1,541. Each question received 10 fresh calls: original order, an identical-input repeat, four balanced orders and four random orders (2,400 calls). The rotation orbit containing the original order is excluded. The secondary adaptive policy stops after two rotations when both responses give the same valid answer and source, reusing the balanced responses. Fresh original-order inference defines 91 initial hits over 77 videos. The protocol also named InternVL3-8B as a second reader. Its runs were stopped at the authors' request after 760 responses; those partial outputs are excluded, and the Qwen requests and settings were unchanged.

Balanced voting lost 19.78 $[12.22,\ 27.96]\%$ of initial hits. Random four-call voting lost 18.68 $[11.11,\ 27.10]\%$, and one random order lost 25.27 $[16.88,\ 34.44]\%$. The primary reduction is -1.10 $[-8.79,\ 6.67]$ points, with ΔmIoU of 0.09 $[-1.48,\ 1.72]$. On the subset with resolved outputs in every call, the reduction is -1.27 $[-9.52,\ 6.95]$. The adaptive policy used 3.07 $[2.89,\ 3.24]$ calls, with a reduction of -1.10 $[-8.75,\ 6.74]$ and ΔmIoU of 0.05 $[-1.53,\ 1.67]$. The identical-input repeat lost 0/91 initial hits. Weighted mean prompt/completion tokens per question are 1866.0/154.9 for one call, 7463.8/643.0 for four calls and 5799.2/510.7 for the adaptive policy. *Reading:* on fresh videos, the point estimate is negative; its interval includes both zero and the exploratory estimate.

Pairwise candidate comparison. This is an exploratory development pilot, recorded before inference but designed with earlier results known. It uses the fixed-answer interface: every policy receives the same frozen predicted answer. From 955 NExT-GQA test questions, we drew 60 stored initial hits and 60 misses (120 questions, 116 videos) with inverse inclusion weights. Each of the six candidate pairs is shown in both orders, giving 12 calls per question. A clip wins a pair only when both orders name it; the clip with most net wins is cited, and ties go to the first clip identifier in sorted order. The comparator is freshly executed per-candidate scoring (A_3, 4 calls). Including 288 repeat controls, the pilot made 2,208 requests.

Pairwise selection lost 28.33 $[17.24, 40.32]\%$ of initial hits, and A_3 lost 33.33 $[21.87, 45.45]\%$. This is a reduction of 5.00 $[-8.33, 18.52]$ points, with ΔmIoU of 0.20 $[-3.91, 4.47]$. Only 19.00 $[14.01, 24.66]\%$ of weighted pairs produced a consistent preference in both orders, and 22.21 $[15.76, 28.90]\%$ named opposite clips. Among 1,440 pair responses, 236 were ties, 352 rejected both clips and 36 were invalid. On 59.30 $[48.74, 69.34]\%$ of weighted questions no pair was decided, so the tie rule chose the citation. Repeating the pairs on 24 questions lost 1 of 9 initial hits. Unweighted mean prompt tokens per question, excluding answer generation, are 10,984 for pairwise comparison and 1,950 for A_3. *Reading:* the pilot did not establish a reduction; with most questions settled by the tie rule, it says little about pairwise judgments themselves.

Retention-oriented LoRA training. The protocol was recorded before the first adapter fit; it is an exploratory learning experiment on previously studied data, not a preregistration. The video-level split supplies 2,076 training questions with positive candidate overlap and a development split of 703 questions from 112 videos. Frozen Qwen generates the predicted answer, and each arm selects one of four slots for that answer. Three arms share the same budget: rank-8 LoRA on the language decoder's query and value projections (3,833,856 parameters, vision encoder frozen), one epoch, 519 optimizer steps and 4,152 training views. Plain SFT uses original order. Augmented SFT uses permuted orders. Retention training uses the same permuted views with grounding, initial-hit preservation and consistency losses in equal weight. The principal contrast is retention versus augmented SFT on reordered loss among the 114 frozen-base initial hits. The recorded advancement rule required a positive CI for this reduction and lower CI bounds above -1 point for the order-averaged mIoU and $R@.5$ differences. Only then would the checkpoint be evaluated on a 600-question test sample frozen before development inspection.

The reduction was 1.75 $[-1.98, 5.15]$ points, with ΔmIoU of 0.22 $[-0.24, 0.65]$ and $\Delta R@.5$ of 0.18 $[-0.53, 0.87]$. The loss interval includes zero, so the gate was not met and the test sample was never evaluated. Among the 111 questions both arms initially hit, the reduction is 2.40 $[-0.64, 5.51]$. Relative to the frozen model, every fitted arm raised order-averaged mIoU: retention 1.92 $[1.17, 2.63]$, augmented SFT 1.70 $[0.94, 2.44]$, plain SFT 1.67 $[0.95, 2.35]$. The retention arm's reordered-loss reduction relative to the frozen model was 1.46 $[-4.64, 7.37]$. Adopting a fitted model also lost original-order base hits: 17.54 $[10.34, 25.71]\%$ for retention training and 20.18 $[11.43, 30.10]\%$ for augmented SFT. Each fit took about 75 minutes on one GPU (peak 20.1 GiB); serving adds one citation call after answer generation. *Reading:* training raised aggregate grounding on development videos, but did not establish better retention of the same initial hits than matched augmentation.

Scope. The designs differ in samples, interfaces and comparators, and each has its own evidentiary status: recorded fresh validation, an exploratory pilot or a development-only gate. The balanced-voting, pairwise and retention results passed independent recomputation with 3,519, 3,416 and 42,257 checks, respectively. Their inputs, outputs and protocols are retained with the released artifacts.

M INTEGRITY CHECKS AND CORRECTIONS

Table 60 links each issue to the affected analyses and the checks used for the reported results. Detailed input audits remain with their experiments, as referenced in the table. Startup failures occurred before any test prediction was produced and left the evaluation settings fixed; the supplementary `STARTUP_RECORD.md` retains that operational history.

Issue	Affected analyses	Correction or verification
Incorrect packet metadata	Selector test follow-up	The split field defaulted to `val`, and a hardcoded field incorrectly stated that test was unused. Input checksums identify the test annotations. Original fields are retained with their corrected interpretation; prior access is documented in Appendix A.
Historical packet-index mismatch	First-citation retention at $\tau = .5$	Excluding all 4 affected questions leaves initial-hit counts, loss counts and mean drops unchanged. The missing historical packet remains unrecovered (Appendix D.1).
Contact-sheet rendering	Factorial controls; subsequent exhaustive-order audit	Flagged-question exclusion preserves the relation of CIs to zero for all 24 factorial contrasts (Appendix N.1). The exhaustive audit uses corrected sheets and fresh inference and reproduces source changes and temporal-hit loss (Appendix D.3).
Missing clip index in scorer	Per-clip scoring arm	The original scorer treated every row as unresolved, understating the arm. Corrected index handling is used in the reported analysis.
Ignored bootstrap argument	Delivery-conversion intervals	Bootstrap execution was restored; the reported conversion results use the corrected analysis (§S10.6; Supplement S2).
Incomplete reader files accepted	Ablation and delivery-conversion analyses	Missing rows were counted as retrieval-top-1 fallbacks. The corrected code checks file completeness, and the reported tables use the corrected analyses.

Table 60: Integrity issues and their treatment in the reported results. Exclusion checks, corrected-input inference and analysis-code corrections address the specific results listed in each row.

N Reproducibility details

Supplementary selector development. The separate *Selector development report* preserves retrieval-side rescoring, UniVTG width comparisons, citation-selection variants, learned-selector configurations, development-half window selection, containment controls and historical gate decisions. Its sections and tables use a D prefix. These records characterize selector development; the present appendix retains PCRA inputs, paired results, serving controls and integrity checks. The report is supplied as `selector_development.pdf` with editable source.

Reusable PCRA workflow. The supplementary `pcra/README.md` specifies the packet and response contracts and provides a runnable example. Each packet records a question, video ID, ordered clips and frames; temporal annotations remain in the evaluator. The `python -m pcra` commands `prepare`, `run` and `analyze` freeze inputs, invoke a reader adapter, and export transitions with estimates and intervals. Preparation preserves clip membership, bounds and media hashes across orders. Analysis checks complete response coverage and input hashes, averages alternatives within questions, and resamples video clusters. It reports fallback and unchanged-answer cohorts; repeat contrasts require repeat coverage on every question. The package also provides a CPU-only synthetic example and a regression check against the provider results in Appendix K.6. In a separate Python virtual environment outside the repository, the copied package completes the synthetic prepare–run–analyze workflow and reproduces 62 shared metrics and their intervals on 1,000 archived provider responses (214 checks at tolerance 10^{-9}). This verifies software isolation on the same host; the real-data check reuses archived outputs. The commands and environment are recorded in `AR22`.

Optional support assessment. Temporal PCRA uses annotation overlap directly. For answer-support assessment, a separate LVLM judge receives the question, a fixed proposed answer and one clip, with pair roles and temporal annotations hidden. Appendix K.15 reports the completed

model-assessed audit. These judgments measure model-assessed support; human evidence sufficiency remains a separate target.

Retrieval geometry and development. The default NExT-GQA retriever searches 15 overlapping windows on a 16-bin grid, each covering 12.5% of the video. Temporal NMS at IoU 0.35 retains four citable *units* as the *packet*. `LanguageBind_Video_V1.5_FT` (Zhu et al., 2024) was selected over CLIP-L and X-CLIP during development. Retriever and reader receive the same question and five answer options. X-CLIP replication preserves this geometry (Appendix K.5); candidate-expansion controls vary packet size.

Scoring and secondary metrics. For each citation, we compute temporal IoU between its mapped source-clip interval and the best-matching annotated window, then average these values to obtain mIoU. We also report $R@1_{\mathrm{IoU}=.5}$. Additional metrics measure selected-unit overlap, containment (AnyGC: the citation lies inside an annotated window), and joint answer-and-grounding correctness. AnyGC is the strictest case of NExT-GQA's intersection-over-prediction (IoP) score. Joint metrics follow Acc@GQA with an IoU rather than IoP threshold. Confidence intervals use 10,000 paired bootstrap draws over video clusters. Constructed packets check the scorer (Supplement S9).

supplementary development report, Sections D6 and D3 report sub-window and multiple-citation comparisons (*supplementary development report*, Table D4); Supplement S2 reports coverage ceilings.

N.1 A RENDERING DEFECT IN THE CONTACT SHEETS

Our frame extractor occasionally returns a successful read whose pixels are entirely black, and it carried no blank-frame guard, so such frames were composited into contact sheets and delivered to the reader as if they showed video. The criterion below was fixed by the first audit and is recorded rather than re-chosen: a pixel is dark below luminance 8; a sheet's dark fraction is the share of its dark pixels including the white caption strip, which makes the test stricter and puts an all-black sheet near 0.91 rather than 1; and a video is flagged in a condition when the mean dark fraction over its sheets there exceeds 85/100. The recorded criterion uses input pixels rather than localization scores and is applied unchanged across conditions. By this rule 21 of 567 videos render black in at least one of the four factorial conditions while only 3 do so in all four. The remainder are the problem: the conditions differ in *which* videos were observable, so they are not matched on observability. Every one of these videos decodes normally on a later read, so the cause is transient—most likely decoder state under the parallel build—rather than corrupt source video, and we therefore record them as suspected rendering failures: passing the check is not proof that a sheet is faithful.

We preserve the original inputs and predictions and report the original full cohort alongside the post hoc render-audited subset. Subset estimates reuse the recorded predictions; no corrected input is retrospectively substituted into the original runs. Dropping every question on a flagged video, in all four conditions alike, leaves 3246 of 3358 questions on 546 videos (112 excluded, 3.3%) and defines the post hoc render-audited subset. On that subset every factorial contrast keeps the same relation to zero for both readers and both metrics, over 24 comparisons, and the largest point-estimate shift is 0.061 points for Qwen3-VL-8B and 0.141 for InternVL3-8B—both on $R@.5$ rather than mIoU, where the largest mIoU shift is smaller still. Both readers were served the same sheet files, so one mask applies to both. A stricter variant that flags a video when any single sheet fails the same pixel test changes the set by one video (22 videos, 118 questions), so the subset is not sensitive to that choice. The per-video manifest, the question list and its hash are in `AR23` and `render_audit_summary.json`.

The video-level rule can dilute a failure, since one normal sheet can pull a video's mean below the threshold. Cutting the sheets back into their tiles and testing each rendered frame separately (§S10.6) measures the defect where it acts: of the 16,073 frames whose timestamps fall inside an annotated window, 279 arrived dark and 102 sit on sheets holding fewer tiles than the condition renders, leaving their tile-to-timestamp mapping undetermined. Those frame-level exclusions and this video-level subset are different cuts of the same defect and we report them separately rather than merging them.

	Full cohort 3358 q.	Render-audited 3246 q.	Structurally verified 3183 q.
Qwen3-VL-8B			
C_4	$+1.41$ $[+0.80, +2.04]$	$+1.44$ $[+0.82, +2.09]$	$+1.42$ $[+0.79, +2.09]$
C_2	$+0.81$ $[+0.18, +1.46]$	$+0.85$ $[+0.19, +1.52]$	$+0.90$ $[+0.22, +1.58]$
D_4	-1.49 $[-2.02, -0.96]$	-1.53 $[-2.09, -0.98]$	-1.48 $[-2.06, -0.93]$
D_8	-2.09 $[-2.68, -1.52]$	-2.12 $[-2.74, -1.53]$	-2.01 $[-2.62, -1.42]$
I	$+0.60$ $[-0.03, +1.23]$	$+0.59$ $[-0.05, +1.25]$	$+0.53$ $[-0.13, +1.19]$
E	-0.68 $[-1.40, +0.05]$	-0.68 $[-1.43, +0.07]$	-0.59 $[-1.35, +0.17]$
InternVL3-8B			
C_4	$+0.15$ $[-0.45, +0.77]$	$+0.12$ $[-0.51, +0.74]$	$+0.08$ $[-0.54, +0.69]$
C_2	-0.39 $[-1.04, +0.28]$	-0.50 $[-1.19, +0.20]$	-0.48 $[-1.16, +0.22]$
D_4	-0.01 $[-0.50, +0.46]$	$+0.04$ $[-0.45, +0.54]$	$+0.07$ $[-0.41, +0.56]$
D_8	-0.55 $[-1.05, -0.05]$	-0.57 $[-1.09, -0.05]$	-0.50 $[-1.02, +0.03]$
I	$+0.53$ $[-0.08, +1.15]$	$+0.62$ $[-0.01, +1.28]$	$+0.56$ $[-0.09, +1.18]$
E	-0.40 $[-1.05, +0.28]$	-0.46 $[-1.14, +0.22]$	-0.41 $[-1.09, +0.26]$

Table 61: The same six contrasts on the three nested question sets, ΔmIoU with paired video-bootstrap intervals. The full cohort is the original recorded result; the render-audited subset removes videos flagged by the brightness rule; the structurally verified subset additionally removes questions whose served sheets failed an input-construction check. Each set is applied identically to every arm and to both readers. Narrower sets are post hoc and are reported alongside the original rather than replacing it.

Input construction on the audited set. We separately check frame count, displayed-time agreement and cross-condition pixel correspondence in the served files. Across 24434 arm-by-window rows over 7085 windows, 24260 sheets contain the specified number of frames and 174 contain fewer. For each correctly sized sheet, the rendered caption regions were compared with labels drawn from the eight recorded sampling times; the nearest caption match agrees with the intended time on 24260 sheets. This checks the displayed time labels, not decoded-frame provenance by itself.

The intended two-frame sampling indices are a subset of the four-frame indices. In the served-tile comparison, each two-frame tile has the expected four-frame tile as its nearest match in 12120 comparable rows, with 0 mismatches; the median matched difference is 0.06 grey levels against 25.3 for the nearest alternative. The comparison is over the frame-content regions with the caption strips excluded, on the served tiles at their stored size with no rescaling, under mean absolute difference in 8-bit grey levels, and a tile passes when the expected counterpart is its nearest match among all tiles of that window. This measures cross-density correspondence of served tiles. The render records retain sampling indices and settings; original decoded-frame digests are unavailable. At fixed frame count, shared candidates in the four- and eight-candidate packets have identical compared pixels in 20528 rows (maximum difference 0.00, 0 failures), over the same frame-content regions, so an effect of adding candidates is not carried by different pixels for a shared one.

The 174 short sheets are the renderer's skip path: a frame it cannot read is dropped rather than padded, so the sheet holds half the frames of its condition and its tiles can no longer be tied to timestamps. They touch 63 questions on 10 videos. We exclude those questions once, from all four arms and both readers, giving the structurally verified subset of 3183 questions on 540 videos used in Figure S1 and Supplementary Table S16. Relative to the render-audited subset, 2 of 24 contrasts change their relation to zero and the largest point-estimate shift is 0.133 points (Table 61). The existing brightness rule still flags 41 individual tiles on 25 sheets inside this subset; frame counts being right is a different question from those frames passing the brightness check, and we report both. These construction checks do not certify that every frame passes the brightness check or that the displayed frames are semantically sufficient.

Models. Grounding baseline `UniVTG` (released unified checkpoint, CLIP-only variant) and CLIP ViT-B/32 (Radford et al., 2021) for its features. Retrievers `LanguageBind/LanguageBind_Video_V1.5_FT` rev `5d53aab2` and, for the replication of Appendix K.5, `microsoft/xclip-base-patch32` rev `a2e27a78`. Readers `Qwen/Qwen3-VL-8B-Instruct` rev `0c351dd0`, `OpenGVLab/InternVL3-8B` rev `853e3a79`, `microsoft/Phi-3.5-vision-instruct` rev `12b77fb4`, `Qwen/Qwen3-VL-4B-Instruct` rev `ebb281ec` and `Qwen/Qwen3-VL-32B-Instruct` rev `0cfaf481`. All revisions are pinned, and each is recorded in the run artifact of the step that used it.

Decoding. The original reader runs use temperature 0, top-p 1, a 384-token cap, prefix caching disabled and one generation per question. Follow-up configurations and seeds are recorded with their experiments, including the exhaustive-order audit (Appendix D.3) and categorical judges (Appendix K.15). Greedy decoding and serving repeatability are distinct: repeated-input measurements appear in Appendices K.1 and K.12.

Compute. Feature extraction shared by every arm: 975 GPU-seconds over the validation videos. Readers: 3653 (Qwen3-VL), 8091 (InternVL3), 3492 (Phi-3.5-Vision) GPU-seconds; the 4-frame variant 1456. All retrieval-only arms together require 11 CPU-seconds and no GPU. Single A6000, eager execution.

Scoring. The scorer carries 25 contract tests covering tIoU, containment, overlap and metric assembly. It independently reproduces the retrieval packet's published 144-span oracle (mIoU 44.75 recomputed against 44.7 published); $R@1_{\mathrm{IoU}=.5}$ differs by 0.09 points because the cached oracle values were stored in float16 and three of 3,358 rows sit exactly at the 0.5 threshold. We recompute in float64 from the interval bounds stored as float32. Higher-precision computation retains the input quantization: on the eight-unit test packet 76 of 44,424 unit–question tIoU values sit close enough to .5 that one unit in the last place of a stored bound would move them across it, and 49 of those are exactly at .5, where the recorded $\geq$ convention decides them. We report the exposure rather than re-deriving the bounds, which the packet does not retain at higher precision.

O How the reported numbers were produced

Reported table values, confidence intervals and effect sizes are generated from result artifacts. The supplementary `ARTIFACT_INDEX.md` maps artifact IDs to the original code, inputs, outputs and provenance records. Dates and filesystem paths are retained there. One script reads the experiment JSONs and emits the LaTeX macros used in the manuscript. The manually entered quantities are an early pilot sweep, agreement with a published oracle, contract-test and item counts, and hardware and timing details. The generated `result_registry.csv` links each of the 2656 cited macros to its source artifact. The supplementary material contains the artifact index, every artifact cited as `AR01`–`AR35`, and the development report cited as "supplementary development report". An earlier audit, `AR01`, adds experiment version, cohort, comparison direction and metric for 536 macros. Later macro families retain artifact links in the registry.

Appendix N specifies model revisions, decoding, compute accounting and scorer tests. Appendices A and B record evaluation gates, protocol checksums and execution order. Appendix M links input-integrity issues and analysis-code corrections to the affected results and their verification. The submission package includes standalone PCRA code and a CPU-only reproduction of the four- and eight-candidate follow-up from all 4,800 archived responses. It preserves prompts, intervals, media hashes and model settings, with relocated media paths recorded explicitly. Video pixels and model weights are obtained separately for fresh inference. The workspace has no version-control history; local files support the protocol record. Earlier work accessed NExT-GQA test annotations, as reported in §3.2.

P Ethical considerations and intended use

We evaluate NExT-GQA and VideoZeroBench under their released licences using publicly released open-weight models. Both benchmarks contain web videos. This study collected no new video data or human annotations. Appendix K.16 summarizes the prepared sample and retained audit protocol; execution would require rater recruitment, compensation, informed consent and the running institution's review for video annotation.

The intended use is to measure how reliably a system preserves its video citations when the available evidence stays fixed. Temporal overlap evaluates agreement with benchmark annotations. Claims about evidence that users can verify require human support assessment in the intended application. Reporting these measurements separately helps users interpret grounding scores.

Q EXTENDED RELATED WORK

Grounded and evidence-backed video QA. TVQA+ and NExT-GQA link answers to temporal or spatio-temporal evidence (Lei et al., 2020; Xiao et al., 2024), while AGQA targets compositional spatio-temporal reasoning (Grunde-McLaughlin et al., 2021). Grounded-VideoLLM and VideoZeroBench evaluate localized evidence and its verification (Wang et al., 2025a; Meng et al., 2026), MUSEG targets multi-segment grounding (Luo et al., 2026), and EG-VQA and E-VQA score answer-linked evidence with semantic and temporal components (Huang et al., 2026b; Wang et al., 2026b). These works evaluate, and in several cases train, end-to-end grounding. We take the delivered packet as given and measure whether its emitted citations preserve their identity and temporal correctness when its order changes.

Video retrieval and long-video RAG. Long-video systems reduce context with retrieval, temporal search, keyframe selection and hierarchical routing (Jeong et al., 2025; Zhu et al., 2026; Pan et al., 2026; Qiu et al., 2026; Wang et al., 2026a; Zhang et al., 2025). These systems motivate testing whether changes to evidence delivery preserve correct citations. A supporting analysis separates candidate coverage from selected-citation quality: the `R5` variant increases delivery by 14 points with an mIoU change of $+0.02$ $[-0.26, +0.31]$ (*supplementary development report*, Table D1). The stability experiments hold delivery fixed and track citation changes within the same packet.

Training-free grounding with pretrained models. Training-free grounding provides the methodological background for our selection interfaces. TFVTG generates and selects proposals with a pretrained VLM and treats simple proposal selection as its own baseline (Zheng et al., 2024); MoReVQA decomposes training-free video QA into grounding and reasoning stages (Min et al., 2024); and concurrent work elicits temporal boundaries from a frozen VLM by per-interval yes/no questioning, analysing how emitted interval length relates to event length (Huang et al., 2026a). We use frozen readers to study which delivered clip a system cites and whether that choice remains correct across candidate orders. Appendix K compares direct clip selection with per-clip scoring using the same model and candidate packet. This isolates how the selection interface affects the emitted citation.

Language models as selectors and rerankers. Using a language model to choose among retrieved candidates is an established pattern. In text IR, monoT5 and RankGPT rerank passages with a pretrained sequence model and with a general-purpose LLM respectively (Nogueira et al., 2020; Sun et al., 2023). In video, a substantial line uses an LLM or VLM to decide which frames or clips the model should look at, by agentic frame requests, an adaptive tree over clips, a *trained* querying policy over frame combinations, keyframe selection under a fixed budget, or language-guided spatiotemporal compression (Wang et al., 2024; 2025b; Yu et al., 2025; Tang et al., 2025; Shen et al., 2025). We build on this selection pattern and evaluate the stability of emitted citations while holding the reader and delivered evidence fixed (§2). Bito et al. (2026) shows that recommendation quality can diverge from consistency under equivalent candidate permutations. Vishwakarma et al. (2026) measures source-order effects on first-citation selection in text RAG. We follow citation changes through the emitted set to temporal-hit loss, then measure the offsetting gains that affect aggregate quality. This connects order sensitivity to the persistence of correct evidence for each question.

Attribution and citation evaluation. Text attribution work separates answer correctness, citation completeness and citation entailment (Gao et al., 2023; Liu et al., 2023; Kamalloo et al., 2023; Es et al., 2024). We compare temporal-hit retention with model-assessed support transitions for the same answer (§K.4). A separate supporting analysis shows that containment scores can reward short citations even as temporal overlap falls (*supplementary development report*, Section D6); it motivates measuring temporal correctness with overlap in the stability audit.

Evidence continuity under retrieval changes. Ponnuraj (2026) identifies retrieved chunks by normalized-text hashes and reports pairwise and worst-case overlap across retrieval configurations. We preserve candidate content and boundaries, so a changed citation selects different evidence from the same available set. Temporal-hit retention then measures whether an initially correct citation stays correct.