FRAMES DON'T ADD UP: A FACTORIAL CAUSAL AUDIT OF CONTEXTUAL EVIDENCE UTILITY IN VIDEO-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple-choice video question answering systems often select frames by scoring each independently for relevance, change, or salience. Yet the answerer sees a set, not a ranking. *Does a frame retain its value beside the other selected frames?* We test this assumption through controlled downstream interventions. Our FACTORIAL EVIDENCE AUDIT matches the visual budget, surrounding frames, candidate slots, pixel processing, prompt, and answer scoring across candidate contrasts, measuring individual and joint effects across roughly one million calls, six open video-language models, and three evidence-grounded benchmarks. Annotation-defined frames are causally useful, exceeding matched outside-evidence frames in 17 of 18 model–benchmark settings. But their values do not add independently: evidence pairs interact more strongly than matched outside-evidence pairs across all 18 settings, an effect that survives singleton-scale adjustment and controls for order, slots, context, budget, and decoding. Zero is the exact additive prediction; deterministic no-op controls return exactly zero and repeated baselines match. Singleton scale predicts interaction magnitude, but none of the tested saturation corrections reliably improves set selection; fitted corrections worsen it, whereas measured pair terms cut selection regret by 58%. Common pointwise proxies track downstream utility poorly, and non-negligible higher-order interactions remain beyond pairwise terms. While set-aware selection via plain DPP improves mean downstream log odds over strong ranking baselines, the best-performing selector varies by answerer and benchmark. Evidence matters. But fixed-budget video evidence selection is a set problem, not a ranking problem.

1 INTRODUCTION

Even when a retriever delivers the right evidence frames, the downstream model may not benefit: it receives a complete frame input, not a ranked list, and that input can saturate, interfere, or change in value with context and order. Many selection pipelines nevertheless treat frame selection as an independent ranking problem: scoring each candidate for relevance, visual change, or salience, occasionally followed by a heuristic diversity penalty (Liu et al., 2025). This ranking formulation is attractive because it is simple, cacheable, and answerer-agnostic. But it leaves a foundational question unanswered: *does the value of an evidence frame survive composition with the other selected frames?*

List-wise and adaptive selectors already recognize that a useful frame set should balance relevance, diversity, temporal coverage, or iterative feedback. Our question is prior to choosing such an objective: does a fixed answerer's value for a fixed-budget frame set decompose into context-independent frame scores at all, and what selection loss remains when those singleton scores are measured directly from the answerer rather than approximated by a retrieval proxy? There are concrete reasons to expect that it does not. Two relevant frames can address the same question aspect, so their joint effect is smaller than the sum of their individual contributions (saturation or redundancy). One frame can become misleading or introduce visual ambiguity in the presence of another (interference). A weaker frame may only help beside a specific cue (complementary facilitation). These are properties of the complete answerer input, not of a frame in isolation. Existing benchmarks establish that frame choice matters (Ok & Lee, 2026), that long videos stress retrieval (Wang et al., 2025a; Fu et al., 2025a), and

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.

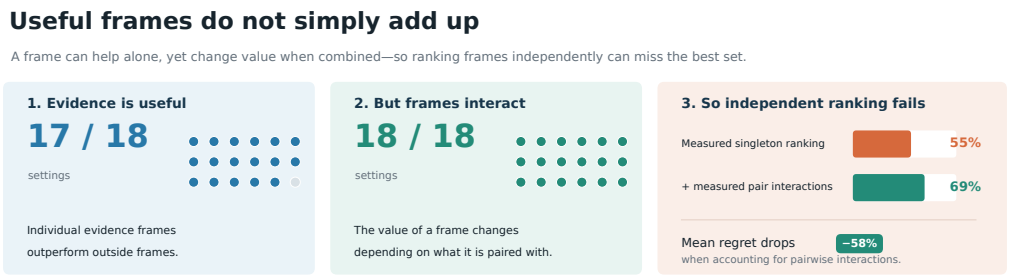


Figure 1: **Main findings at a glance. (1) Evidence is useful:** Annotation-defined evidence frames outperform matched outside frames in 17/18 settings. **(2) Frames interact:** Evidence pairs exhibit stronger non-additivity than outside pairs in all 18 settings. **(3) Independent ranking can miss the best combination:** Selecting three of four evidence candidates, with outside references filling unused slots and eight frames always shown, measured singleton ranking recovers an optimal-value set in 55.2% of records. Adding measured pair interactions raises recovery to 68.6% and cuts mean selection regret by 58.2%.

that some questions demand multi-evidence integration (Ben Ami et al., 2026). We ask a different question: whether independently measured evidence values add up after retrieval.

To answer it we run a factorial causal audit. For every item, candidate pair, and shared context, we evaluate four matched eight-frame inputs: neither candidate, either candidate alone, and both together (Figure 2). Their difference-in-differences isolates pair interaction while holding the question, answer choices, visual budget, context frames, and candidate slots fixed. In parallel, single-candidate insertion interventions measure downstream utility directly rather than equating utility with semantic similarity. We run the audit on a deterministic 500-question holdout from each of LVBench, NExT-GQA, and HERBench, using six open models from four families, with 1,108,740 calls in the core audit. A separate all-eligible census covers those benchmarks and Video-MME.

The answer is consistent and consequential (Figure 1). We call each model–benchmark combination an *evaluation setting*; the full grounded matrix has $6 \times 3 = 18$ settings. Annotated evidence is useful in 17/18 settings, but evidence/evidence pairs are more strongly non-additive than outside/outside pairs in all 18. Signed interaction is negative in 17/18 settings, suggesting frequent saturation or interference, while our primary claim deliberately concerns magnitude because positive complementarity also occurs. Interaction varies with interval topology, and three- and four-way components remain after pairwise accounting in a smaller exhaustive subset study. Pointwise proxies recover these downstream effects poorly. A matched list-wise DPP improves average utility, but no selector wins universally and annotated-evidence fraction rarely identifies the winner.

Contributions. In summary, we **(1) formalize contextual downstream frame utility** through controlled input interventions at a fixed visual budget, rather than relying on surrogate retrieval scores; **(2) introduce the FACTORIAL EVIDENCE AUDIT**, a factorial protocol that isolates individual from pair and higher-order composition while holding context, slots, pixels, and prompt fixed; **(3) establish consistent non-additivity** across 18 model–benchmark settings and perturbation controls, showing that the evidence-pair gap persists after adjusting for singleton scale, candidate slot, and frame order; **(4) quantify selection regret and show constructive gains**: in the four-candidate audit, singleton ranking misses an optimal-value triple in 44.8% of records and pairwise information cuts regret by 58.2%, a gain none of the tested singleton-scale corrections recovers, while list-wise DPP improves average downstream utility without measuring interactions; and **(5) support reproducibility** with frozen manifests, 1.1M+ analysis-ready outcomes, and scripts for reproducing and checking the reported numbers (`https://anonymous.4open.science/r/frames-dont-add-up-9CC5`).

Terminology. We say *set composition* rather than the overloaded "compositional reasoning" (logical structure in questions): we ask only whether a VLM's answer quality, at a fixed frame budget, decomposes over the frames it was shown.

2 Related Work

Video QA and temporal evidence. NExT-QA/GQA and long-video benchmarks test reasoning over dispersed temporal evidence (Xiao et al., 2021; 2024; Mangalam et al., 2023; Wu et al., 2024; Zhou et al., 2025; Wang et al., 2025a; Fu et al., 2025a). Yet single-frame bias can mask weak temporal integration (Lei et al., 2023). HERBench isolates multi-evidence fusion (Ben Ami et al., 2026), and TempCore tests sensitivity to frame replacement (Ok & Lee, 2026). We measure how a fixed answerer's downstream value changes when evidence frames are combined under a matched visual budget.

Frame and token selection. Selection methods learn where to look or which frames to combine (Wu et al., 2019; Zhao et al., 2023), rank keyframes (Faure et al., 2025; Tan et al., 2024; Tang et al., 2025), and compress visual tokens (Shen et al., 2024; Fu et al., 2025b). Closer to our setting, MDP^3 combines relevance, diversity, and sequential allocation (Sun et al., 2025); Frame-Voyager learns combinations from answerer-loss rankings (Yu et al., 2025); and A.I.R. iteratively samples and validates candidates (Zou et al., 2026). We ask whether singleton scores and common proxies recover the downstream set value these selectors seek to improve.

Set objectives and attribution. Maximal marginal relevance and DPP balance relevance with diversity (Carbonell & Goldstein, 1998; Kulesza & Taskar, 2012); submodular guarantees concern suitable surrogate objectives and do not automatically describe the answerer's value (Nemhauser et al., 1978; Lin & Bilmes, 2011). Shapley values, Harsanyi dividends, and functional ANOVA decompose contributions and interactions (Shapley, 1953; Harsanyi, 1982; Hooker, 2007). We use related discrete differences, but measure interventions on complete video inputs rather than post-hoc attribution.

3 What a Frame Is Worth

Before we can ask whether frame values add up, we need a notion of value that belongs to the answerer rather than to a retriever. This section defines one, then writes down precisely what a pointwise selector assumes and what would falsify it.

3.1 Measuring utility at the answerer

Let q denote a multiple-choice question with gold option y^*. When the model is shown the ordered frame input S, we score

$$L(S; q) = \log \frac{p(y^* \mid S, q)}{1 - p(y^* \mid S, q)}, \tag{1}$$

where $p(\cdot)$ is renormalized over valid option letters, so $1 - p(y^* \mid S, q) = \sum_{y \neq y^*} p(y \mid S, q)$ and L is equivalently the gold-option log probability minus the log-sum-exp of incorrect-option log probabilities. Positive values favor the gold option over all alternatives combined; a one-unit gain multiplies these odds by e. We use gold-vs-rest log odds because it is continuous, answer-set aware, and immune to unconstrained generation parsing; accuracy and normalized gold probability serve as secondary scales across all main contrasts. We suppress q where fixed.

Individual utility. We measure what a candidate frame is worth by putting it in front of the answerer. For candidate i and a fixed seven-frame outside-evidence context C_7, we insert i at the fourth position, so that $L(C_7)$ scores seven frames and $L(C_7 + i)$ scores eight:

$$U_i(C_7) = L(C_7 + i) - L(C_7). \tag{2}$$

Utility is answerer- and context-specific by construction. We evaluate three fixed, non-overlapping outside-evidence contexts, averaging them for main candidate comparisons while retaining context-specific records. Note what is deliberately *not* called utility: neither annotation membership nor embedding similarity.

U_i is an *insertion* effect (mixing candidate value with moving from seven to eight frames), so a standalone U_i is not a fixed-budget quantity. However, this common baseline is identical across candidates in the same context and cancels from every reported contrast:

$$U_e(C_7) - U_o(C_7) = L(C_7 + e) - L(C_7 + o), \tag{3}$$

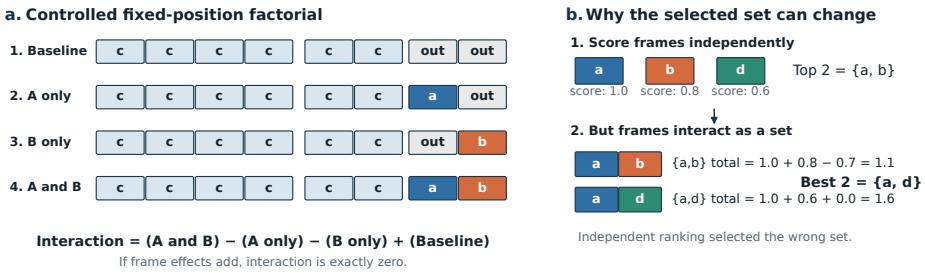


Figure 2: **Controlled factorial design and selection consequence. (a)** Six context frames and two candidate slots remain fixed. Across four matched model calls, candidate frames a and/or b replace baseline outside-evidence reference frames. The interaction is the difference-in-differences, which is identically zero under additivity. **(b)** Illustrative selection failure: scoring frames independently selects $\{a, b\}$, but a negative pair interaction ($I_{ab} = -0.7$) makes $\{a, d\}$ the true best set (Equation 6).

which compares alternative eighth frames under an identical eight-frame budget. Pair, higher-order, and selector contrasts hold the frame budget fixed within each comparison.

Pair interaction. Figure 2 shows the matched four-call construction. Six context frames and two candidate positions stay fixed, and each candidate position holds a real outside-evidence reference frame whenever its candidate is absent, so all four calls present exactly eight real video frames. Let B be that complete baseline input, including both references, and let $B[S]$ replace the designated references for candidates in S. For candidates i, j, the two-slot factorial intervention is

$$I_{ij}(B) = L(B[\{i,j\}]) - L(B[\{i\}]) - L(B[\{j\}]) + L(B). \tag{4}$$

$I_{ij} = 0$ means the two effects simply added in that context. Negative values mean the joint gain fell short of the sum of the individual gains (saturation, redundancy, or interference); positive values mean it exceeded them (complementarity); Appendix A.22 shows examples of both. Because these mechanisms coexist and would partly cancel in a signed average, our primary statistic is the interaction magnitude $|I_{ij}|$, which discards sign but not effect size. We compare evidence/evidence pairs against outside/outside pairs drawn from the same source video. "Outside" means outside every annotation-defined interval; "boundary" means one second before an interval start or after its end. Mixed pairs give a graded topology audit.

Higher orders and selector debt. Pairs are only the second term. For four candidates we evaluate all 2^4 subsets and apply a Möbius decomposition: informally, the coefficient of a subset T is the value unique to having every member of T together, once the contributions of all strict subsets have been subtracted. This separates constant, first-, second-, third-, and fourth-order components. We also ask whether a single scalar summary—a selector's "interaction debt," the summed pointwise utilities of its chosen frames minus the realized utility of the set—predicts correctness beyond annotated-evidence fraction, relevance, coverage, model, and benchmark.

3.2 Why pointwise scoring cannot be assumed

The audit is a direct test of a set-function assumption, so it is worth writing that assumption down. Fix a question, answerer, complete ordered baseline B, and candidate slots, and define

$$F_B(S) = L(B[S]) - L(B), \qquad S \subseteq V. \tag{5}$$

Here V is the candidate pool and $B[S]$ retains references outside S, so $|S|$ never changes the displayed frame count. A purely pointwise account assumes $F_B(S) = \sum_{i \in S} s_i$ for candidate scores s_i that do not depend on which other frames were selected. Let $\Delta_i F(S) = F(S \cup \{i\}) - F(S)$ and $\Delta_{ij} F(S) = \Delta_i \Delta_j F(S)$. Equation 4 is exactly $\Delta_{ij} F_B(\varnothing)$: what we measure is the pointwise assumption's own residual.

Additivity characterization. On a finite candidate pool, $F(S) = F(\varnothing) + \sum_{i \in S} s_i$ for all S if and only if $\Delta_{ij} F(S) = 0$ for every S and distinct $i, j \notin S$. The forward direction is immediate. Conversely, vanishing mixed differences make $\Delta_i F(S)$ invariant to every other member of S, so adding

elements along any chain telescopes into fixed singleton increments. A single nonzero controlled interaction is therefore already a certificate against a globally additive pointwise representation. Our experiment collects such certificates at scale, over controlled evidence and non-evidence pairings.

Proposition 1 (Exact two-frame regret decomposition)**.** *For two active candidates, let $u_i = F_B(\{i\})$ and write $F_B(\{i,j\}) = u_i + u_j + I_{ij}$, suppressing B below. If A is the top-two pair under u and D is any other pair, then*

$$F(D) - F(A) = \underbrace{\sum_{i \in D} u_i - \sum_{i \in A} u_i}_{\leq 0} + \underbrace{I_D - I_A}_{\textit{composition advantage}} . \tag{6}$$

Pointwise ranking is optimal exactly when no alternative pair's interaction advantage exceeds its singleton-utility deficit.

Proposition 1 establishes an exact condition under which pointwise ranking fails: without an assumption bounding the composition advantage $I_D - I_A$, singleton quality alone provides no nontrivial guarantee on set regret (as illustrated in Figure 2b). Diversity objectives can help, but their relationship to downstream composition must be tested rather than assumed.

Contextual utility identity. Let $m(T)$ be the Möbius coefficient of T, so $F(S) = \sum_{T \subseteq S} m(T)$. For $i \notin S$, its marginal utility in context S is

$$\Delta_i F(S) = m(\{i\}) + \sum_{\varnothing \neq T \subseteq S} m(T \cup \{i\}). \tag{7}$$

This makes a point that is easy to miss. Within a fixed reference input, changing a candidate's companions changes its utility through the pair and higher-order terms involving it. That is why we measure individual utility across several contexts and then audit pair and higher orders separately. Proofs and a construction with identical pointwise scores are in Appendix A.18.

4 Audit Protocol

We call the protocol below the Factorial Evidence Audit. Its governing constraint is that comparisons between candidate types differ only in the frame under test: the visual budget, surrounding frames, candidate slots, pixel processing, prompt, and answer scoring are held fixed. The raw singleton effect in Equation 2 deliberately compares seven versus eight frames; Equation 3 shows why candidate-type contrasts still have an equal eight-frame budget.

4.1 Benchmarks and evidence provenance

We evaluate three grounded benchmarks: NExT-GQA (official question-linked intervals; Xiao et al. 2024), HERBench (curated multi-evidence segments; Ben Ami et al. 2026), and LVBench (reference windows as broad localization evidence; Wang et al. 2025a). Point annotations are expanded by ± 2 s for sampling. Privileged sampling within these regions defines our "annotation oracle" control. Video-MME lacks temporal annotations and enters only the external-validity census (Fu et al., 2025a).

The causal track evaluates a deterministic hash-stratified holdout of 500 eligible questions per grounded benchmark, frozen without model outputs or manual filtering. The separate census evaluates all technically eligible questions: 1,549 LVBench, 5,482 of 5,553 NExT-GQA (71 excluded due to video decoding failures), 1,971 HERBench, and 2,700 Video-MME questions.

4.2 Models, inputs, and inference

The answerers comprise Qwen3-VL-4B/32B (Bai et al., 2025), InternVL3.5-8B/38B (Wang et al., 2025b), Perception-LM-8B (Cho et al., 2025), and VideoLLaMA3-7B (Zhang et al., 2025). All receive identical decoded RGB frames (without multi-crop expansion) and multiple-choice prompts. Selector and oracle inputs are chronological; factorial audits fix candidate positions across matched 4-call sets, with order perturbed explicitly in Section 6. Bare option letters are scored via deterministic constrained decoding.

For individual utility, each question contributes two frames sampled inside annotated evidence intervals, two one-second boundary frames, and four outside-evidence frames. Outside frames are deterministically spread over the non-evidence portion of the same video (between 2% and 98% of its duration), avoiding a cross-video control. Each candidate is inserted into three fixed seven-frame outside-evidence contexts. For pair interactions, six shared outside-evidence frames occupy positions 1–6, and the two candidates replace fixed outside-evidence references at positions 7 and 8, identically for every pair type. The four-candidate audit likewise places four shared frames at positions 1–4 and its candidates in fixed slots 5–8. Section 6 moves the candidates to positions 1–2, exchanges them, and re-sorts all frames chronologically. Each pair requires the four calls in Equation 4. The main pair types are evidence/evidence, evidence/boundary, evidence/outside, boundary/outside, and outside/outside, where x/y denotes the categories of the first and second candidates. The primary factorial, higher-order, and selector experiments use eight frames; budget controls are labeled separately. Singleton baselines show seven frames, as explained in Section 3.

4.3 Uncertainty, aggregation, and deterministic checks

The source video is the resampling unit; questions and contexts from the same video stay together. Primary contrasts use "source-video-equal" means: average within video, then across videos. The exhaustive selection audit retains question/context-equal means, with cluster intervals targeting that same weighting (Appendix A.7). We use 10,000 cluster-bootstrap resamples for 95% intervals and 10,000 paired sign-flip draws for tests, with Benjamini–Hochberg FDR correction (Benjamini & Hochberg, 1995) across the stated setting families. Equal-setting summaries weight each model–benchmark combination equally.

Checking the additive zero. Measurement noise can inflate an absolute-value statistic such as $|I_{ij}|$. Here scoring is deterministic: no-op interventions leave the input unchanged and return exactly zero interaction, and repeated baselines match across all 1,350 paired records (Section 6). Thus the tested execution shows no repeat variability to explain the gap. Zero remains the exact prediction of an additive set function; the observed interactions describe departures from that prediction under our fixed scoring interface.

Sample size and scale. The 500-question holdouts reflect compute constraints fixed prior to inference. The core causal audit analyzes 1,108,740 calls (Table 3). For the primary pair contrast, cluster-bootstrap 95% intervals have a median half-width of 0.091 (range 0.013–0.401), with all lower bounds strictly positive. The gap remains positive across all 18 settings under source-video subsampling at 25%, 50%, and 75% (Appendix A.3).

What each track may and may not support. The all-eligible selector census extends coverage beyond the holdout, while including its eligible questions; it estimates no factorial interactions and is never pooled with the causal track. Higher-order and extended-control audits evaluate Qwen3-VL-32B, InternVL3.5-38B, and Perception-LM-8B (the nine representative settings). The expanded higher-order audit evaluates 57,600 calls across 200 questions per benchmark.

5 Results

We report three findings. First, annotated evidence helps, though retrieval proxies poorly predict its value. Second, that value changes when frames are combined. Third, these interactions can change which set a selector should choose.

5.1 Evidence helps, but relevance is not utility

Figure 3 (left) validates the intervention before we lean on it: annotation-defined evidence exceeds outside-frame utility in 17/18 evaluation settings, with a median source-video-equal difference of $+0.47$ gold-vs-rest log odds. Sixteen of those settings have source-video-clustered intervals excluding zero, and all sixteen survive adjusted sign-flip tests; the two exceptions are InternVL3.5-8B and VideoLLaMA3 on HERBench. The annotations therefore carry real downstream information, though how much varies with answerer and benchmark.

Privileged access does not eliminate composition sensitivity: eight annotation-oracle frames outperform eight uniform frames in all 18 settings on log odds (median difference $+1.04$; accuracy agrees

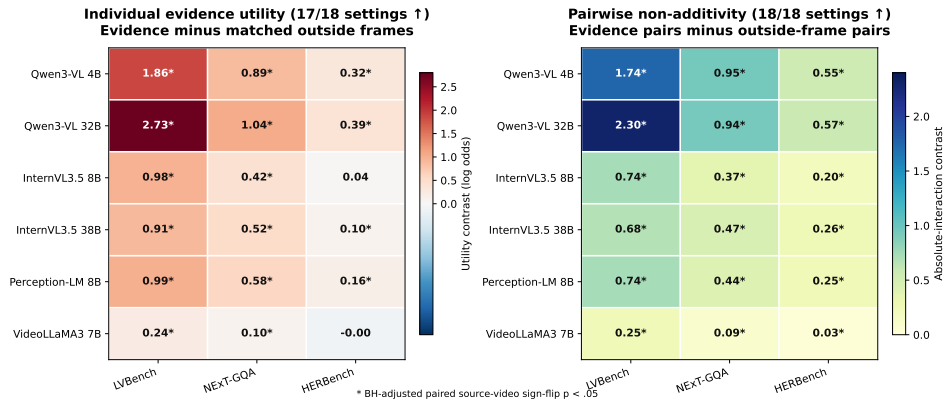


Figure 3: **Primary causal effects across 18 model–benchmark settings.** Each cell shows the source-video-equal mean contrast in gold-vs-rest log odds. **Left (Individual utility):** An annotation-defined evidence frame outperforms a matched outside-evidence frame in 17/18 settings. **Right (Pairwise non-additivity):** Evidence pairs exhibit strictly larger absolute interaction than matched outside-evidence pairs across all 18/18 settings.

in 17/18) and beat a $4 + 4$ oracle/outside mixture in all 18. Presenting oracle frames chronologically outperforms reverse-chronological order in all 18 settings on both log odds (median difference $+0.34$) and accuracy. Even with gold evidence spans, composition and order remain highly consequential.

Common pointwise proxies do not track measured downstream utility well. Across 12 LVBench and HERBench settings (the NExT-GQA proxy run terminated on a temporal-neighbor decode failure), spectral visual saliency has median pooled Spearman correlation $+0.0001$ with utility (range -0.047 to $+0.043$; 95% CIs include zero in 11/12 settings). Its median singleton-utility regret is 0.961 gold-vs-rest log odds: the gap in mean individual utility between its top-two candidates and the utility-ranked top two. This measures ranking loss before composing a joint input. SigLIP relevance aligns better but still incurs positive regret in all 12 settings. Annotation membership itself has median pooled correlation $r = 0.121$ and equal-setting mean regret 0.537 (Appendix Figure 4). Relevance, visual change, and annotation membership are distinct from downstream utility.

5.2 Evidence does not add up

The central result is in Figure 3 (right). Evidence/evidence pairs have larger mean absolute interaction than matched outside/outside pairs in *every* evaluation setting, with a median source-video-equal gap of $+0.513$ gold-vs-rest log odds and positive clustered intervals and adjusted tests in all 18. The direction is unchanged on bounded normalized gold probability (median gap 0.041) and on accuracy (median gap 0.063), so it is not an artifact of the unbounded log-odds scale. Interactions concern the scored output, not an identified internal mechanism (Appendix A.19). These are within-setting contrasts; equal-setting medians summarize their direction without treating a log-odds unit as equally calibrated across answerers. The equal-setting median gap is also positive within every reported question-type stratum (Appendix A.4). Signed interaction is negative in 17/18 settings, consistent with frequent saturation or interference, but complementarity also occurs, which is why our primary claim concerns magnitude rather than a universal sign. Appendix A.22 shows high-magnitude and typical pairs of each sign (Figures 11 and 12); the rare $|I| > 10$ tail (1.02% of evidence pairs) is not needed for the result (Appendix Table 8).

Is that magnitude large enough to matter? One anchor: the unpaired equal-setting mean $|I_{ij}|$ for same-interval evidence pairs is 0.943, roughly twice the median individual evidence advantage of $+0.47$ gold-vs-rest log odds. These two summaries aggregate at different levels and are not a formal contrast, but the comparison makes the point plainly enough: composition effects are the same order of magnitude as the individual-frame signal that pointwise selection is built on.

Is non-additivity merely a scale artifact? Larger singleton effects could explain stronger evidence-pair interactions. We therefore adjust absolute interaction for singleton scale, baseline magnitude, and

temporal distance with source-video-clustered uncertainty. Controlling for singleton absolute sum, maximum effect, signed sum, quadratic scale, baseline magnitude, and temporal distance attenuates the median evidence-pair coefficient by 84.6%, from $+0.507$ to $+0.078$. Yet its clustered interval remains positive, with BH-adjusted $p < .05$, in all 18 settings (Appendix A.5).

Scale is not enough to choose a combination. At fixed set size, a common strictly increasing transform of singleton scores preserves their ranking, as does a strictly increasing transform of their sum: probability-scale additivity and noisy-OR saturation choose the same three-of-four set as singleton ranking in 100% and 99.7% of records. Scale-derived pair penalties, cross-fit on held-out videos, instead *raise* regret by 26% and 47% (interval-supported increases in 6/9 and 5/9 settings; lower in none). Even a penalty strength tuned on the evaluation outcomes themselves cuts regret by less than 0.5%, against 58.2% for measured pair terms (Appendix A.8). The tested corrections miss the pair-specific information that improves selection.

Interaction varies with evidence topology. Same-interval evidence pairs have the largest unpaired mean interaction magnitude (0.943), but the matched same-versus-distinct ordering is heterogeneous: same-interval magnitude is larger in 7/12 eligible settings (median difference $+0.043$). The signed contrast is clearer on NExT-GQA and mixed on HERBench. See Appendix Table 7 and Figure 6.

Pairs are not the end of the story either. In an exhaustive four-candidate audit (200 questions per benchmark, three answerers), third- and fourth-order Möbius terms account for 23.5–27.7% of absolute nonconstant component mass across nine settings. Pair terms therefore cannot reconstruct the entire measured set function; Appendix Figure 5 shows the full order breakdown and source-video-clustered intervals (Appendix A.6).

What this costs a selector. Interaction is not merely present; it changes which combination is best. The exhaustive audit chooses two or three of four evidence candidates; outside references fill unused slots, so all inputs still contain eight frames. Singleton ranking recovers an optimal-value pair in 54.6% of records, at mean regret 0.271 gold-vs-rest log odds and accuracy 70.0%, versus 72.2% for the joint-value oracle. For three selected candidates, adding measured pair terms raises optimal-value recovery from 55.2% to 68.6%, cuts mean regret from 0.326 to 0.136 (a 58.2% reduction), and lifts accuracy from 69.2% to 70.9%, with the exhaustive value oracle at 71.9%. Pairwise accounting reduces regret with positive cluster intervals in 9/9 settings, improves recovery in all nine, and increases accuracy in 7/9. Appendix A.7 gives the matched uncertainty and tie rule.

This diagnostic uses gold-answer measurements unavailable to a deployed selector. For triples, remaining reconstruction error is third order; component mass alone does not quantify the resulting selection regret.

5.3 Selector audit: retrieving evidence is not the same as helping

The causal selector audit compares nine methods: five pointwise or sampling baselines, relevance–diversity, BOLT inverse-transform sampling, MDP^3, and a plain-DPP ablation. MDP^3 combines SigLIP relevance and DPP diversity with temporal segmentation and dynamic allocation; plain DPP uses the same official kernel and greedy MAP rule without that allocation. Every method returns eight chronological frames from the same 64-frame pool; text-conditioned methods use the question alone. We define *annotated-evidence fraction* as the selected-frame share inside an annotated interval; it is fixed independently of answerer outputs, whereas downstream utility is model-specific.

Across the 18 grounded settings, the selector retrieving the most annotated evidence carries a median loss of 0.389 gold-vs-rest log odds to the best method, has a median within-setting Spearman association with utility of -0.042, and loses to uniform in 12/18 settings. It is never the highest-gold-vs-rest log odds method. These counts are descriptive; evidence fraction is not a reliable utility criterion even when a list-wise method occasionally succeeds at both.

Plain DPP leads on macro log odds, while MDP^3 leads on accuracy. Against relevance–diversity, DPP gains $+0.124$ gold-vs-rest log odds (95% CI

Table 1: Selector audit: selected equal-setting means across 18 settings; the supplement reports all nine methods and intervals. Evidence fraction is the selected-frame share inside annotated intervals; coverage is temporal span divided by video duration.

Selector	Acc.	Log odds	Evid. frac.	Coverage
Plain DPP	.533	**.752**	.104	.826
MDP^3	**.536**	.739	.102	.848
Rel.–diversity	.528	.626	.106	.803
BOLT-ITS	.526	.609	.110	.742
SigLIP relevance	.523	.541	**.117**	.695
Uniform	.512	.479	.088	**1.000**

$[+0.049, +0.200]$); against BOLT, it gains $+0.130$ $[+0.023, +0.238]$. The corresponding accuracy gains are less certain: $+0.74$ points $[-0.40, +1.88]$ and $+0.72$ points $[-0.59, +2.06]$. Appendix A.10 examines the opposing accuracy flips and where the log-odds gains accrue. DPP does not select more annotated evidence than SigLIP relevance (0.104 versus 0.117), and the winner varies by setting: DPP, MDP[3], and BOLT lead log odds in 7, 7, and 4 settings.

The overlapping all-eligible census (24 settings, adding Video-MME) shows the same pattern: in all 18 grounded census settings, the evidence-fraction winner never coincides with the log-odds winner. At $K = 16$, the two coincide in only 2/18 causal settings.

Finally, we ask whether this structure compresses into one number: a selector's interaction debt. A controlled regression finds no detectable residual association between debt and correctness beyond standard retrieval metrics; this scalar retains a net sign but loses the identity and separate signs of the interacting pairs.

6 Stress-Testing the Result

We stress-test annotation breadth, frame order, candidate slot, surrounding context, distractor hardness, frame budget, and pixel decoding. The evidence-pair gap remains positive in every setting tested under every condition; Table 2 shows selected controls, and Appendix Table 8 reports the full set.

The gap persists under narrow annotation windows ($\leq 10\%$, $+0.536$), chronological re-sorting ($+0.543$), and SigLIP-hard outside context/reference frames ($+0.344$; Appendix A.13). The gap survives the tested placements: reversing the sequence or exchanging the candidate slots preserves a positive gap in 9/9 settings (Appendix A.14). A post-hoc sensitivity excluding panels with repeated timestamps also retains positive clustered intervals in 18/18 settings (median $+0.469$; Appendix A.15). No-op interactions are zero and all 1,350 baseline repeats match. In a 25-question study, evidence-fraction and utility winners still diverge in 9/9 settings at 32 candidates and 8/9 at 128 candidates. Larger models improve selector accuracy in all six within-family contrasts, but interaction magnitude remains non-monotone (Appendix Figure 9).

Table 2: Selected controls: positive settings and median evidence-pair minus outside-pair $|I_{ij}|$ gap (gold-vs-rest log odds). Rows with /9 use the representative subset.

Control condition	Positive	Median gap
Annotation breadth $\leq 10\%$ of duration	18/18	$+0.536$
Chronological re-sort of the same frames	18/18	$+0.543$
Sequence reversed (candidates at 1–2)	9/9	$+0.428$
Two newly sampled contexts	9/9	$+0.454$
SigLIP-hard background frames	9/9	$+0.344$
Frame budget $K = 16$	9/9	$+0.359$
Direct source-RGB decoding	9/9	$+0.631$

7 Conclusion

Independent relevance asks whether each frame looks useful, but the answerer receives a set whose value can saturate, interfere, complement, or shift with order and context. Across six models, three grounded benchmarks, and roughly one million analyzed answerer calls, annotation-defined evidence is useful yet systematically non-additive. Singleton scale alone is insufficient in our audit: none of the tested corrections reliably improves selection, while measured pair terms cut regret by 58.2%. Plain DPP supplies a constructive result: set-aware proxy structure improves average downstream utility without measuring the answerer's interactions. Its non-universal wins, together with the higher-order audit and controls, show why this is progress rather than a complete solution. These conclusions remain bounded by annotation quality, the fixed-frame multiple-choice interface, and the audit's diagnostic scope (See Appendix A.23 for the limitations).

Where might this lead? One possibility is to shortlist frames by relevance, then refine the selection using estimates of how they work together. Our audit makes that direction worth testing, but does not establish how to estimate those interactions without gold answers, whether they transfer to new contexts, or when their cost pays off. Plain DPP offers an encouraging example of a useful set-aware proxy. Learning what to combine remains an open problem.

8 AI USE STATEMENT

Language models assisted with text editing, manuscript review, and code for reanalyzing experimental outputs. Scientific claims, interpretations, and the final content remain the responsibility of the authors.

9 REPRODUCIBILITY STATEMENT

Analysis-ready results and evaluation code are available at `https://anonymous.4open.science/r/frames-dont-add-up-9CC5`. Appendix A.20 gives additional implementation details.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Dan Ben Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. Herbench: A benchmark for multi-evidence integration in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4505–4514, 2026.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Jaime G Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, volume 98, pp. 290941–291025, 1998.

Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. PerceptionLM: Open-access data and models for detailed visual understanding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 23475–23537, 2025.

Gueter Josmy Faure, Jia-Fong Yeh, Min-Hung Chen, Hung-Ting Su, Shang-Hong Lai, and Winston H Hsu. Hermes: temporal-coherent long-form understanding with episodes and semantics. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22911–22921. IEEE, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025a.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22654–22663. IEEE, 2025b.

John C Harsanyi. A simplified bargaining model for the n-person cooperative game. In *Papers in game theory*, pp. 44–70. Springer, 1982.

Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of computational and graphical statistics*, 16(3):709–732, 2007.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 487–507, 2023.

Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 510–520, 2011.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3318–3327. IEEE, 2025.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Hyunjong Ok and Jaeho Lee. Tempcore: Are video qa benchmarks temporally grounded?, 2026. URL https://arxiv.org/abs/2509.01167.

Lloyd S Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp 3: a training-free approach for list-wise frame selection in video-llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24090–24101. IEEE, 2025.

Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13581–13591. IEEE, 2024.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128. IEEE, 2025.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1278–1287. IEEE, 2019.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214. IEEE, 2024.

Sicheng Yu, CHENGKAI JIN, Huanyu Wang, Zhenghao Chen, Sheng Jin, ZHONGRONG ZUO, XU XIAOLEI, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-voyager: Learning to query frames for video large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LNL7zKvm7e.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Mingjun Zhao, Yakun Yu, Xiaoli Wang, Lei Yang, and Di Niu. Search-map-search: a frame selection paradigm for action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10627–10636. IEEE, 2023.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.i.r.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SZVpOKw0YD.

A SUPPLEMENTARY MATERIAL

CONTENTS

A.1 Experiment map and reporting boundary 13
A.2 Pointwise proxy alignment 13
A.3 Sample-size calibration 13
A.4 Question-type heterogeneity 13
A.5 Utility-adjusted interaction audit 14
A.6 Interpreting higher-order mass 15
A.7 Exhaustive selection: domain, ties, and uncertainty 15
A.8 Can a saturation correction replace measured pair terms? 16
A.9 Candidate-pool coverage and extreme-value calibration 17
A.10 Selector uncertainty 18
A.11 Interaction outcome taxonomy 19
A.12 Pair topology 19
A.13 Matched robustness controls 19
A.14 Candidate slots and placement 20
A.15 Sensitivity to repeated frame timestamps 21
A.16 Controlled selector-debt null 22
A.17 Scale and family comparisons 22
A.18 Proofs and pointwise non-identifiability 23
A.19 Outcome scale and causal scope 23
A.20 Reproducibility details 24
A.21 Model fingerprints and sensitivities 25
A.22 Qualitative case studies 25
A.23 Limitations 26

A.1 EXPERIMENT MAP AND REPORTING BOUNDARY

Table 3 records which data support each claim. The causal and census tracks have separate manifests and are never pooled. Video-MME is census-only because it has no per-question temporal evidence oracle in our protocol.

A.2 POINTWISE PROXY ALIGNMENT

A.3 SAMPLE-SIZE CALIBRATION

As a descriptive sample-size check, we draw 2,000 fixed-seed, without-replacement source-video subsamples at each of 25%, 50%, and 75% of the available clusters. The evidence-pair gap remains positive in every evaluation setting at every fraction; the corresponding equal-setting median gaps are 0.515, 0.512, and 0.513.

A.4 QUESTION-TYPE HETEROGENEITY

We repeat the primary source-video-equal contrast within benchmark-provided question types rather than imposing a shared taxonomy. Its equal-setting median is positive in every reported stratum: all

Table 3: Core audit components and their scope. Counts include only fully completed conditions; the additional perturbation and candidate-density controls are reported separately. Three one-record InternVL3.5-38B uniform-32 fragments are retained in the raw logs but excluded from analyses and totals. The census track is reported without a call count because it uses a different denominator and is never pooled with the causal track.

Component	Purpose and scope	Settings	Model calls
Evidence-access panel	completed dose, oracle, context, and order conditions; grounded holdout	18	124,500
Individual utility	candidate insertion ($7 \rightarrow 8$ frames); grounded holdout	18	243,000
Pair factorial	full pair topology; grounded holdout	18	540,000
Higher order	all 2^4 subsets; 200 items, 3 models	9	57,600
Selector causal	fixed-pool $K = 8$ plus MDP^3/DPP query audit; grounded holdout	18	107,640
Debt test	controlled regression; 68 video clusters	9	36,000
Total (core causal audit)			**1,108,740**
Selector census	all retained items; four benchmarks	24	—

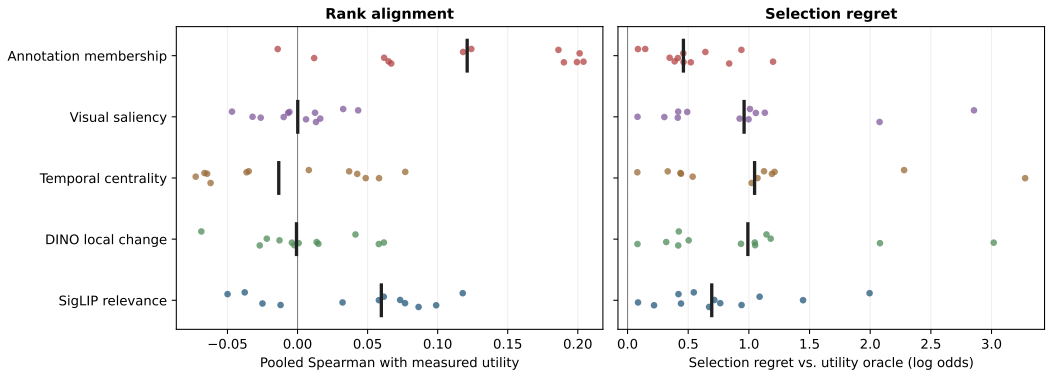


Figure 4: Pointwise proxies versus measured utility in 12 LVBench/HERBench settings. Dots are settings; black bars are medians. Visual and semantic scores show weak pooled rank alignment (left) and positive singleton-utility regret (right), measured before evaluating any joint selection.

six HERBench task types, all five native NExT-GQA types, and six prespecified LVBench tag-present strata (whose labels can overlap). On NExT-GQA, the temporal-type gap exceeds the causal-type gap in all six answerers (median gaps 0.596 versus 0.347 gold-vs-rest log odds); the direct temporal-minus-causal contrast has a source-video-bootstrap interval above zero in 3/6 settings. This suggests stronger composition sensitivity for temporal questions on NExT-GQA, not a benchmark-general ordering. The smallest cells—NExT-GQA TP and LVBench summarization—contain 8 and 10 source videos, respectively, and are retained only as descriptive strata.

A.5 Utility-adjusted interaction audit

The full adjustment regresses $|I|$ separately in each evaluation setting on an evidence/evidence indicator, singleton absolute sum, singleton absolute maximum, signed singleton sum, squared absolute sum, baseline magnitude, and candidate temporal distance. Absolute imbalance is omitted because it equals twice the absolute maximum minus the absolute sum; including all three would make the specification redundant. Uncertainty is clustered by source video, and tests are BH-corrected across the 18 settings within this full specification. The evidence-pair coefficient is positive, has a 95% interval excluding zero, and is BH-significant in 18/18 settings; its median is $+0.07799$ and its range is $+0.01044$ to $+0.17536$. The corresponding unadjusted median is $+0.50719$, an 84.6% attenuation. This regression-sample median differs from the $+0.513$ source-video-equal aggregate in Figure 3 because it is estimated on pair–context records. Because the outcome is an absolute

difference-in-differences, $+0.078$ means that evidence status is associated with an interaction farther from zero by 0.078 log-odds units under the measured controls; it is not a directional gain in answer log odds or accuracy. Adding benchmark-provided question-type fixed effects (native NExT-GQA and HERBench types; overlapping LVBench tag indicators) and question length leaves the coefficient essentially unchanged: median $+0.079$, range $+0.010$ to $+0.201$, with positive intervals and BH-significance in 18/18 settings. The residual persists after adjustment for these measured question attributes; this does not exhaust question complexity. Singleton scale therefore accounts for much of the unconditional *magnitude* contrast, while the positive residual persists under these measured controls. Magnitude is not identity: Appendix A.8 shows that the tested scale-based corrections do not reliably improve selection.

A.6 Interpreting higher-order mass

For a fixed reference input B, the coefficients are

$$m_B(T) = \sum_{R \subseteq T} (-1)^{|T|-|R|} F_B(R). \tag{8}$$

The reported order-k fraction sums $|m_B(T)|$ over all measured question/context records and subsets with $|T| = k$, then divides by the corresponding sum over all nonempty subsets. This is a ratio of aggregate masses, not a mean of per-record fractions. Both the coefficients and their fractions depend on the reference frames, candidate slots, and outcome scale.

Third- and fourth-order terms occupy 5 of the 15 nonconstant subsets of four candidates, so a quarter of the absolute component mass is not disproportionate *per subset*. Equal shares, however, are not the relevant null: for an additive answerer every higher-order component is exactly zero. We do not claim that higher orders dominate; the result is that 23.5–27.7% of measured nonconstant mass lies beyond what a pairwise expansion can represent. This is absolute coefficient mass under the specified reference input, not explained variance or a fraction of accuracy loss.

For uncertainty, we resample source videos and retain all questions, contexts, and subset masks belonging to each sampled video. Each draw recomputes the ratio of summed absolute mass at an order to total nonconstant mass, preserving the original mass-ratio estimand. The 200 questions per benchmark come from 48 HERBench, 91 LVBench, and 184 NExT-GQA source videos. The combined third/fourth-order 95% intervals range from an outer lower endpoint of 0.220 to an outer upper endpoint of 0.293. These intervals account for shared-video dependence without changing the point estimates in Figure 5.

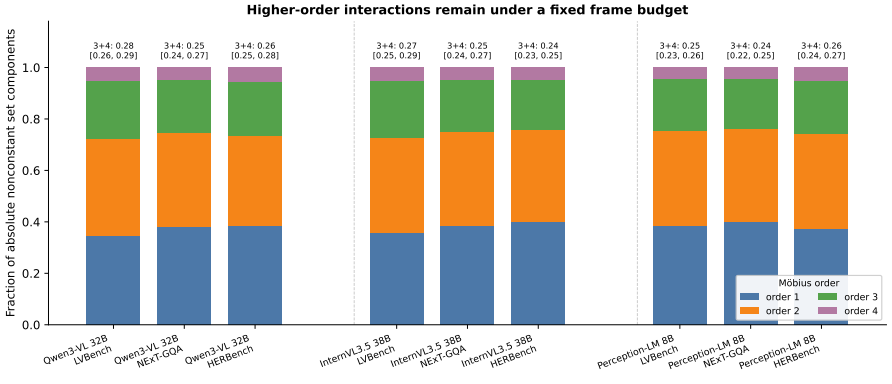


Figure 5: Möbius decomposition of exhaustive four-candidate interventions. Bars show the fraction of absolute nonconstant component mass by order; labels give the combined third- and fourth-order fraction with source-video-clustered 95% intervals.

A.7 Exhaustive selection: domain, ties, and uncertainty

The four evidence candidates occupy designated slots beside four shared outside frames. A candidate left unselected is replaced by its slot-specific outside reference. Selecting two or three candidates

therefore compares six pairs or four triples, respectively, while always showing eight frames. This is a controlled test of composition after privileged evidence retrieval, not an exhaustive search over the separate 64-frame selector pool.

Here the singleton scores are replacement effects $u_i = F_B(\{i\})$, measured within this same factorial grid, not the seven-to-eight-frame insertion effects U_i from Equation 2. For a selected set $\widehat{S}$ of size k, regret is $\max_{T \subseteq V:\,|T|=k} L(B[T]) - L(B[\widehat{S}])$.

We choose combinations by their measured singleton sum, their singleton-plus-pair sum, or their observed gold-vs-rest value. Ties in the selection rule are broken deterministically by candidate indices. *Optimal-value recovery* credits a selected combination whose observed value is within 10^{-10} of the maximum, even if another tied combination has different indices. The value oracle maximizes gold-vs-rest log odds; its reported accuracy is the accuracy of that selection, not a separate accuracy upper bound. For two selected candidates, measured pair terms reconstruct each pair exactly by definition. For triples, third-order terms can still change the ordering; fourth-order terms enter only when all four candidates are active.

The headline statistics first average the two contexts and 200 questions equally within each setting, then average the nine settings equally. Bootstrap draws resample source videos while retaining their questions and contexts, recomputing these question/context-weighted means in each draw. The same video draw is shared across the three answerers within a benchmark. This preserves dependence without changing the target weighting. With 10,000 draws, the three-candidate mean regret reduction is 0.189 gold-vs-rest log odds (95% CI $[0.159, 0.221]$), or 58.2% $[53.0\%, 62.7\%]$. Paired setting-level regret reductions have positive intervals in 9/9 settings; recovery gains have positive intervals in 8/9. Table 4 gives the corresponding macro intervals.

Table 4: Three-of-four candidate selection with eight frames always shown. Equal-setting, question/context-weighted means with source-video-clustered 95% intervals. Regret is in gold-vs-rest log odds; recovery is optimal-value recovery. All scores used to select these combinations are measured with gold-answer access.

Selection rule	Recovery (%)	Mean regret	Accuracy (%)
Measured singletons	55.2 [53.2, 57.2]	.326 [.294, .358]	69.2 [66.4, 71.9]
Singletons + pairs	68.6 [66.6, 70.5]	.136 [.122, .151]	70.9 [68.1, 73.6]
Observed-value oracle	100	0	71.9 [69.1, 74.5]

A.8 Can a saturation correction replace measured pair terms?

Section 5.2 shows that singleton scale accounts for much of the evidence-pair *magnitude* gap. A natural hypothesis follows: a non-linear saturation penalty on pointwise scores might capture most of the interaction without measuring pairs. We test this directly on the exhaustive four-candidate grids of Appendix A.7, where the decision-relevant question is which combination is selected.

Why transforming singleton scores is insufficient. For a fixed number of active candidates, a strictly increasing transform of the singleton sum preserves its set ranking. Maximizing the sum of a common strictly increasing transform of each singleton score also selects the same top-K candidates, under the same tie rule. A non-strict transform can introduce ties, and a general set-dependent function of singleton measurements can change the selection. We test two simple probability-based rules and two fitted pair corrections; this does not exhaust possible functions of singleton measurements.

Rules. At evaluation, all rules use the gold-conditioned singleton utilities u_i and baseline $L(B)$ available to the singleton rule. The two cross-fitted rules additionally learn from measured pair terms on training videos; the outcome-tuned rule uses evaluation outcomes. These information privileges exceed those available at deployment. *Probability-scale additivity* maximizes the summed singleton gains in gold-vs-rest probability. *Noisy-OR saturation* treats each positive probability gain as closing a fixed fraction of the remaining gap to certainty and adds negative gains linearly. The *min penalty* sets $\widehat{I}_{ij} = -\lambda \min(u_i^+, u_j^+)$, a simple redundancy form, with unconstrained λ fit by least squares to measured pair terms. The *scale regression* predicts I_{ij} by ridge regression on singleton sum, minimum, maximum, product, absolute sum, baseline, squared sum, and sum$\times$baseline. Both fitted rules are cross-fit within each setting over five source-video folds, so no video contributes to its own

penalty. The ridge fit standardizes features using its training fold and uses unit ℓ_2 regularization, including the intercept. Finally, an *outcome-tuned* min penalty chooses, per setting, the $\lambda \in [-3, 3]$ (step 0.05) that minimizes that setting's own mean regret, an optimistic within-grid comparator that consumes the evaluation outcomes. Selection ties, recovery, weighting, and the source-video bootstrap (10,000 draws shared across answerers within a benchmark) follow Appendix A.7; the singleton and pairwise selections reproduce the frozen analysis exactly. Bootstrap intervals condition on the frozen fitted predictions and selected tuning values; they do not refit models or retune λ. Setting-level intervals are nominal 95% intervals for this exploratory analysis, not multiplicity-adjusted tests.

Table 5: Saturation corrections versus measured pair terms when three of four evidence candidates are selected (eight frames always shown). Equal-setting means over nine settings with source-video-clustered 95% intervals. Change in regret is relative to singleton ranking (negative is better). "Worse / better" counts settings whose paired regret interval excludes zero in each direction. Singleton and measured-pair rows repeat Table 4.

Selection rule	Mean regret	Change in regret	Recovery (%)	Worse / better
Singleton ranking	.326 [.294, .358]	—	55.2	—
Probability-scale additivity	.326 [.294, .358]	0.0%	55.2	0 / 0
Noisy-OR saturation	.327 [.295, .359]	$+0.3\%$ [-0.1, $+0.9$]	55.2	0 / 0
Min penalty (cross-fit)	.409 [.375, .446]	$+25.7\%$ [$+16.5$, $+36.8$]	48.1	6 / 0
Scale regression (cross-fit)	.480 [.432, .531]	$+47.4\%$ [$+33.8$, $+64.0$]	45.8	5 / 0
Min penalty (outcome-tuned)	.324 [.293, .356]	-0.4% [-1.3, $+0.2$]	55.1	0 / 0
Singletons + measured pairs	.136 [.122, .151]	-58.2% [-62.7, -53.0]	68.6	0 / 9

Results. Table 5 finds no reliable improvement from the tested corrections. The parameter-free saturations select the same set as singleton ranking in 100% (probability additivity) and 99.7% (noisy-OR) of records, so their regret is essentially unchanged. The fitted penalties agree with singleton ranking in only 67–70% of records, and their departures are harmful: regret rises by 26% and 47%, with interval-supported increases in 6/9 and 5/9 settings and no interval-supported improvement. The outcome-tuned penalty, which is allowed to overfit, reduces regret by only 0.4%, and in 8/9 settings its selected λ is negative—a *bonus* for pairing high-utility frames, the opposite of saturation. When two of four candidates are selected, every saturation rule stays within 2% of singleton regret, while measured pair terms reach the optimum by construction.

The selection test separates predicting interaction magnitude from choosing a useful combination. Singleton scale explains much of the magnitude gap in Appendix A.5, but the tested corrections fail to translate it into lower regret. A generic penalty for pairing strong frames can discard a useful combination when that particular pair does not saturate. Measured pair terms retain the identity, sign, and magnitude of each interaction and improve selection substantially. This is evidence against these corrections, not an impossibility result for all predictors built from singleton measurements. The analysis is post hoc and uses only stored four-candidate outcomes (no new model calls).

A.9 Candidate-pool coverage and extreme-value calibration

The causal singleton and factorial audits do not draw candidates from the selector pool. For completeness, within the separate 64-frame selector pool the mean (median) annotated-evidence fractions are 2.76% (1.56%) on HERBench, 2.59% (0%) on LVBench, and 24.38% (15.63%) on NExT-GQA. The fraction of questions with no pool frame inside a sampling interval is 21.0%, 55.4%, and 0%, respectively. This benchmark variation is one reason we do not interpret selector evidence fraction as a benchmark-comparable effect size.

Extreme log-odds interactions are uncommon and are not necessary for the primary result. Among 27,000 evidence/evidence records, the median, 95th, and 99th percentiles of $|I|$ are 0.341, 4.006, and 10.181; 1.02% exceed 10. Among 27,000 outside/outside records, the corresponding values are 0.168, 1.081, and 2.468; 0.06% exceed 10. Removing all $|I| > 10$ records, or trimming or winsorizing each setting at its 99th percentile, preserves a positive evidence-pair gap in all 18 settings. The bounded probability- and accuracy-scale analyses provide separate protection against log-odds tail behavior.

A.10 SELECTOR UNCERTAINTY

Table 6 reports source-video-cluster bootstrap intervals for all nine methods. These are sampling intervals for frozen selections, not random-seed variance; the random selector is deliberately one prespecified draw. MDP^3 uses the official ICCV 2025 implementation at commit `45616806`, and plain DPP shares its SigLIP embeddings, conditional multi-Gaussian kernel, and greedy MAP rule while removing temporal segmentation and dynamic allocation. Both select from the same charged 64-frame pool before any answerer call. The primary query is the question alone, matching the incumbent SigLIP-based selectors; all 2,990 primary selection rows contain eight unique, in-range indices and are presented chronologically.

Table 6: Macro outcome means and source-video-clustered 95% bootstrap intervals across 18 selector settings.

Selector	Accuracy [95% CI]	Log odds [95% CI]
Plain DPP	.533 [.514, .553]	.752 [.570, .941]
MDP^3	.536 [.516, .556]	.739 [.553, .932]
Rel.–diversity	.528 [.509, .548]	.626 [.439, .821]
BOLT-ITS	.526 [.506, .546]	.609 [.420, .803]
SigLIP relevance	.523 [.505, .542]	.541 [.351, .739]
Frozen random draw	.513 [.493, .533]	.494 [.303, .692]
Uniform	.512 [.493, .532]	.479 [.296, .665]
DINO diversity	.510 [.490, .532]	.430 [.239, .627]
DINO change	.501 [.481, .522]	.410 [.218, .607]

The prespecified paired family crosses the two new selectors with uniform, SigLIP relevance, relevance–diversity, and BOLT in every setting and outcome; two-sided source-video sign-flip tests are BH-corrected across all $2 \times 4 \times 18$ comparisons within outcome. The macro intervals quoted in the main text jointly resample each benchmark's source videos across its six answerers. They support an average advantage, not universal dominance: plain DPP's setting-level log-odds contrast is positive in 29/36 comparisons with the two strongest incumbents, with 2/36 surviving correction; its accuracy contrast is positive in 23/36, with none surviving the larger setting-level family.

Why log-odds gains move accuracy little. Top-1 accuracy depends on the gold-minus-best-rival margin, $M = \log p(y^*) - \max_{y \neq y^*} \log p(y)$, whereas gold-vs-rest log odds compare the gold option with the sum of all rivals. With more than two options, accuracy is not a fixed threshold on gold-vs-rest log odds. We therefore decompose the paired contrasts with the main-text weighting (items within source video, videos within setting, equal settings) and the same joint source-video bootstrap. Against relevance–diversity, DPP changes correctness on 13.4% $[12.4, 14.3]$ of item–answerer pairs: 7.1% $[6.3, 7.9]$ become correct and 6.3% $[5.7, 7.0]$ become incorrect. The net $+0.74$ points is thus a small difference between two large opposing flows, which is why its interval is wide. The log-odds gain, in contrast, accumulates where it cannot flip the answer. Questions that both selectors miss receive $+0.048$ $[+0.014, +0.082]$ of the $+0.124$ gain (38%). Items whose mean top-1 margin lies within one log-probability unit of the boundary make up 27% of the data but receive only $+0.011$ $[-0.015, +0.036]$ (9%). Against BOLT the pattern repeats: 16.1% of pairs change correctness (8.4% versus 7.7%), jointly missed questions receive $+0.044$ $[+0.008, +0.082]$ of the $+0.130$ gain, and boundary items receive 8%. DPP shifts probability toward the gold option broadly, including on questions where neither eight-frame selection makes it the top choice; top-1 accuracy registers only the part of that shift that crosses the boundary. Margin bins use the mean of both selectors' margins to avoid privileging one arm. These post-hoc, outcome-defined groups describe where the observed gains accrue; they do not identify a causal subgroup effect.

As an interface sensitivity, we repeat MDP^3/DPP after appending the answer choices to the selector query (another 17,940 calls), without changing the selection rule or answerer prompt. The selected sets change substantially—mean overlap with the primary eight-frame sets is 4.53 frames for DPP and 4.39 for MDP^3—but the primary log-odds conclusion does not: the richer-query variant strengthens DPP's contrasts against relevance–diversity to $+0.181$ gold-vs-rest log odds and $+1.57$ accuracy points. Direct question-only-minus-richer-query differences are -0.057 gold-vs-rest log odds $[-0.129, +0.015]$ and -0.83 points $[-1.89, +0.23]$, so neither change is interval-supported. We keep the matched question-only interface primary rather than choosing the query by outcome.

A.11 Interaction outcome taxonomy

We use outcome labels only as descriptions of the measured set function, not as claims about internal model reasoning. Evidence/evidence pairs more often have two positive singleton gains but a joint gain below their sum (14.89%) than outside/outside pairs (9.98%); complementary rescue occurs in 1.59% versus 0.92%. In contrast, the broad destructive-interference rate is essentially identical (5.41% versus 5.39%), and joint value falls below the better singleton less often for evidence/evidence pairs (62.38% versus 72.67%). The taxonomy therefore supports heterogeneous saturation and complementarity, not a claim that evidence uniquely causes destructive interference.

A.12 Pair topology

Table 7 reports average interaction magnitudes and signed interactions across topological pairings of evidence and non-evidence frames. The unpaired macro average is largest for same-interval evidence pairs, but distinct-interval pairs are available in only 12 settings; the matched comparison is reported in the main text and is not uniformly ordered.

Table 7: Pairwise topology, equal-setting macro averages. "Distinct" is available only where annotation structure permits the comparison.

Pair topology	Settings	Pair–context records	I	$\lvert I \rvert$
Evidence/evidence, same interval	18	19,278	−.432	.943
Evidence/evidence, distinct interval	12	7,722	−.121	.719
Evidence/boundary	18	27,000	−.177	.673
Evidence/outside	18	27,000	−.034	.485
Boundary/outside	18	27,000	−.014	.405
Outside/outside	18	27,000	−.010	.327

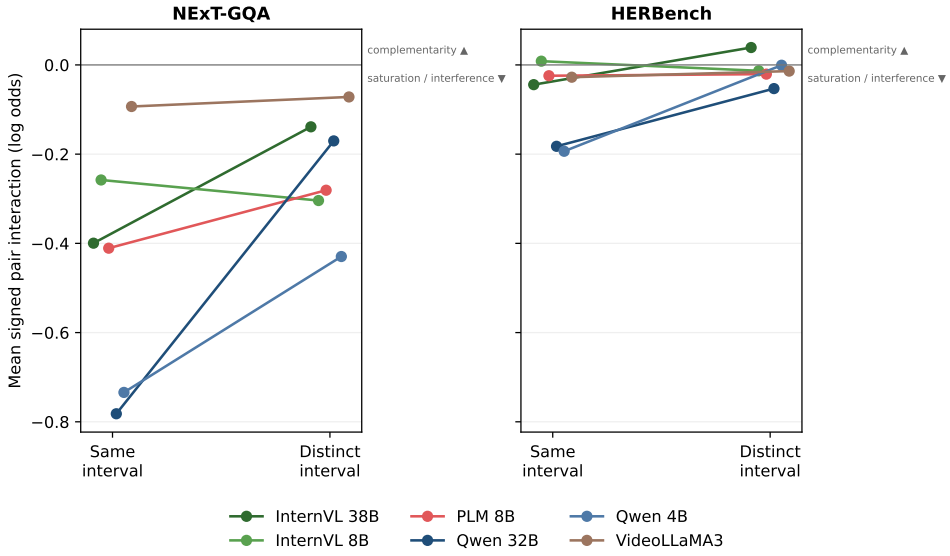


Figure 6: Signed interaction for evidence pairs from the same versus distinct annotated intervals. Same-interval pairs are more negative for 5/6 answerers on NExT-GQA; HERBench is mixed. Lines identify answerers, not repeated trials.

A.13 Matched robustness controls

Table 8 lists every perturbation control; Figure 7 shows per-setting distributions for six of them. Annotation breadth matters because wide evidence windows weaken the meaning of "inside": restricting sampled evidence regions to at most 10% of video duration produces the largest gap among the three

thresholds. These are nested question subsets, not interventions on annotation quality. Chronological re-sorting tests a different presentation rule for the same frames; the gap remains positive in all 18 settings, with every BH-adjusted test at $p < .05$, while ordinal positions can change across cells. The SigLIP-hard condition selects high-relevance outside frames for the six shared context positions and two references, retaining the original candidate identities. It therefore tests robustness to a harder background. Across these and the remaining controls, the gap remains positive in every setting tested.

Table 8: Complete perturbation-control results. Rows report the number of positive settings and the median setting-level evidence/evidence minus outside/outside $|I_{ij}|$ gap (gold-vs-rest log odds). Rows with /9 use the nine representative settings on a 25-question-per-benchmark subset.

Control condition	Positive	Median gap
Annotation breadth $\leq 10\%$ of duration	18/18	$+0.536$
Annotation breadth $\leq 25\%$ of duration	18/18	$+0.513$
Annotation breadth $\leq 50\%$ of duration	18/18	$+0.508$
Chronological re-sort of the same frames	18/18	$+0.543$
Sequence reversed (candidates at 1–2)	9/9	$+0.428$
Candidates exchanged between slots 7 and 8	9/9	$+0.561$
Excluding all records with $\lvert I\rvert > 10$	18/18	$+0.508$
Two newly sampled contexts	9/9	$+0.454$
SigLIP-hard background frames	9/9	$+0.344$
Frame budget $K = 16$	9/9	$+0.359$
Direct source-RGB decoding	9/9	$+0.631$
Boundary window widened to ± 4 s	9/9	$+0.596$

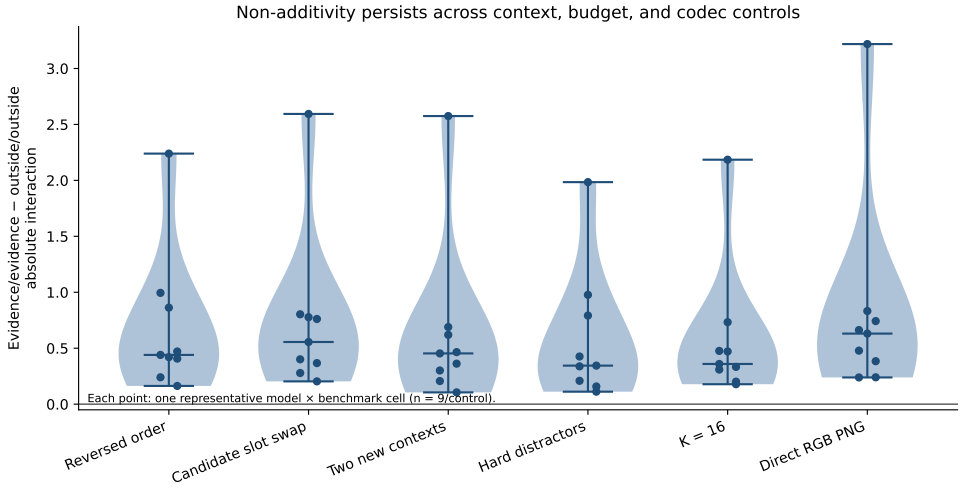


Figure 7: The evidence-pair absolute-interaction gap remains positive under six matched perturbations. Each point is one model–benchmark evaluation setting ($n = 9$ per control); violins summarize settings, not item-level uncertainty.

A.14 Candidate slots and placement

Slot layout. Singleton utility inserts every candidate, whatever its type, at position 4 of 8 in a chronological seven-frame context, so candidate-type contrasts compare frames in the same position. In the pair factorial, six shared outside frames occupy positions 1–6 and the two candidate slots are positions 7 and 8 for every pair type, so the primary evidence/evidence versus outside/outside contrast is position-matched by construction. In the four-candidate audit, four shared outside frames occupy positions 1–4 and the candidates occupy fixed slots 5–8; an absent candidate is replaced by its slot-specific outside reference. In both audits, each context's shared frames and references are drawn from a contiguous block of a deterministic outside-evidence grid spanning 2–98% of the video.

Relocating the candidates. On the 25-question-per-benchmark subset used for the nine-setting controls, the same evidence/evidence and outside/outside panels were re-evaluated with the candidates exchanged between slots 7 and 8, and with the full frame sequence reversed, which moves the candidates to positions 1–2. Table 9 compares these placements with the original on identical panels. The gap remains positive in every setting under every placement, with BH-significant positive intervals in 9/9 (exchange) and 8/9 (positions 1–2) settings. There is limited evidence of a differential relocation effect: the difference-in-differences between the evidence-pair and outside-pair change in $|I_{ij}|$ has median -0.03 and is BH-significant in 0/9 settings for positions 1–2 and 1/9 for the exchange. That single case, Qwen3-VL-32B on LVBench, shrinks a gap that remains the largest in the table ($+2.56$). Record-level magnitudes are position-dependent (median Spearman correlation of $|I_{ij}|$ with the original placement 0.42 for positions 1–2 and 0.58 for the exchange), so we make no claim that an individual pair's interaction is invariant to where it appears. The tested placements preserve the positive direction of the evidence-versus-outside gap, not an invariant gap size. Full reversal also reverses the shared context, so it is not an isolated intervention on candidate position.

Table 9: Candidate placement on identical panels (nine representative settings, 25 questions per benchmark, three contexts). "Positive" counts settings with a positive evidence/evidence minus outside/outside $|I_{ij}|$ gap; "BH" counts positive, BH-significant settings (source-video sign-flip tests). Relocation effect is the median over settings of the evidence-pair minus outside-pair change in $|I_{ij}|$ relative to the original placement. The chronological row uses the full 500-question holdout.

Placement	Candidate positions	Positive	BH	Median gap	Relocation effect
Original	7, 8	9/9	9/9	$+0.552$	—
Exchanged	8, 7	9/9	9/9	$+0.561$	-0.029 (1/9 BH)
Sequence reversed	2, 1	9/9	8/9	$+0.428$	-0.026 (0/9 BH)
Chronological re-sort	varies by timestamp	18/18	18/18	$+0.543$	—

Ordinal proximity. The chronological re-sort varies how far apart the two candidates land. Pairs that end up far apart (four or more positions) have smaller $|I_{ij}|$ than adjacent pairs for both pair types: the far-minus-adjacent median is -0.309 for evidence pairs (interval below zero in 8/12 settings with support) and -0.316 for outside pairs (18/18). For outside pairs the adjacent and far groups share the same median temporal distance, but this does not match their full temporal-distance distributions or other panel attributes. These comparisons are descriptive associations, not isolated causal effects of ordinal proximity. Their similar median contrasts motivate the position-matched primary design; the controlled reordering and relocation tests provide the direct robustness evidence. Middle positions (for example 4–5) were not crossed with the pair audit.

A.15 Sensitivity to repeated frame timestamps

The deterministic outside-frame grid can reuse a frame already present in the context. In particular, an outside candidate repeats a shared frame in every context-0 outside/outside panel. The input still has eight frame slots, and the factorial contrast remains well defined, but this repetition changes the candidate's novelty relative to the background.

As a post-hoc sensitivity, we retain an item/context only when all four factorial cells in *both* the evidence/evidence and outside/outside panels contain eight distinct timestamps, rounded to eight decimal places. Applying one eligibility rule to both pair types retains 1,000 panels per type on HERBench, 998 on LVBench, and 994 on NExT-GQA per answerer, spanning 49, 103, and 383 source videos, respectively. We recompute source-video-equal contrasts, 10,000 cluster-bootstrap draws, and 10,000 paired sign-flip draws, with BH correction across the 18 settings.

The evidence-pair magnitude gap remains positive with a positive 95% interval and BH-adjusted $p < .05$ in 18/18 settings; its median is $+0.469$ gold-vs-rest log odds, compared with $+0.513$ in the primary analysis. Even the smallest contrast, VideoLLaMA3 on HERBench, remains positive: $+0.030$ $[+0.016, +0.049]$. This sensitivity removes exact timestamp repetition, not every possible decoded-frame collision or visually near-identical frame. It uses existing calls and leaves the frozen primary analysis intact.

A.16 CONTROLLED SELECTOR-DEBT NULL

This mechanistic subset covers 25 questions per benchmark, two contexts, eight selectors, and three representative answerers—3,600 sets from 68 source-video clusters and 36,000 component calls. Figure 8 tests whether interaction debt explains correctness after controlling for annotated-evidence fraction, summed relevance, temporal coverage, model, and benchmark. It does not: in an unpenalized logistic model with source-video cluster-sandwich covariance, the debt coefficient is $+0.055$ per standard deviation (clustered SE 0.085; 95% CI $[-0.111, +0.221]$, or $[0.895, 1.247]$ on the odds-ratio scale) with $p = .515$. This is an absence of evidence for a residual effect of that scalar, not evidence that composition is irrelevant—the recovery analysis of Section 5.2 retains the identity and sign of every measured pair term, precisely the information a signed sum discards.

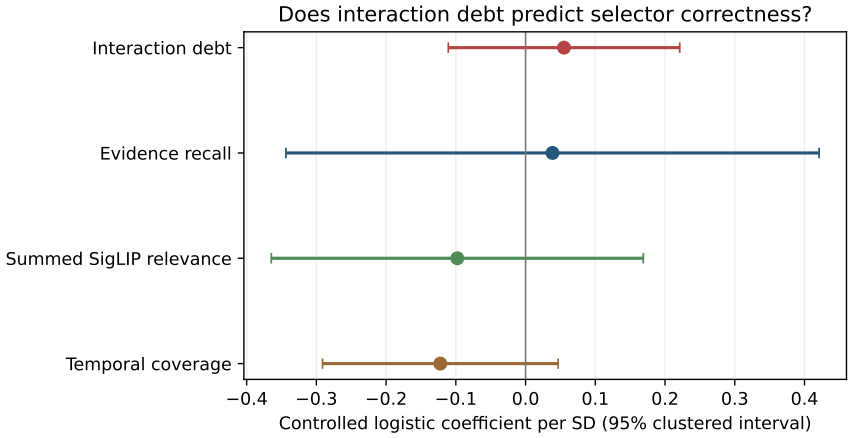


Figure 8: Controlled logistic coefficients per predictor standard deviation, with source-video-clustered uncertainty. The historical plot label "Evidence recall" denotes annotated-evidence fraction, not coverage of all evidence frames. No residual association with correctness is detected for interaction debt (odds ratio 1.057, 95% CI $[0.895, 1.247]$, $p = .515$).

A.17 SCALE AND FAMILY COMPARISONS

Figure 9 breaks down interaction magnitude by model family and parameter scale. Qwen3-VL-32B has larger evidence-pair absolute interaction than Qwen3-VL-4B on two of three benchmarks, including $+0.667$ on LVBench. InternVL3.5-38B is larger than InternVL3.5-8B on two benchmarks and smaller by 0.117 on LVBench. Scale improves selector accuracy in all six within-family benchmark comparisons, but it is not a monotone remedy for set sensitivity.

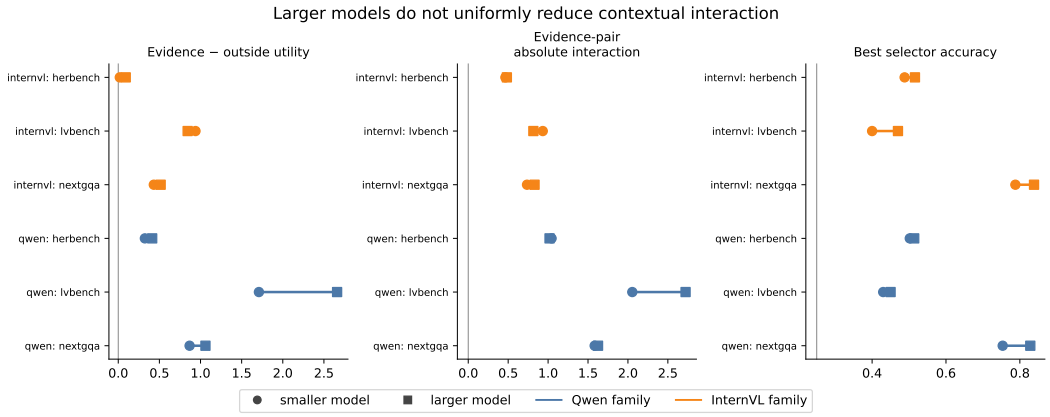


Figure 9: Within-family scale comparisons. Larger models tend to answer better, but evidence interaction does not vanish monotonically with parameter count.

A.18 Proofs and pointwise non-identifiability

Additivity characterization. If F is additive, each marginal $\Delta_i F(S) = s_i$ and every mixed difference vanishes. For the converse, fix i. Any two subsets of $V \setminus \{i\}$ can be connected by single-element additions or removals. The condition $\Delta_{ij} F(S) = 0$ implies that adding or removing j does not change $\Delta_i F(S)$; hence $\Delta_i F(S) = \Delta_i F(\varnothing) = s_i$ for every admissible S. Along an arbitrary ordering $i_1, \ldots, i_k$ of S, telescoping gives $F(S) - F(\varnothing) = \sum_{r=1}^{k} \Delta_{i_r} F(\{i_1, \ldots, i_{r-1}\}) = \sum_{i \in S} s_i$.

Proof of the exact two-frame regret proposition. Substitute $F(A) = \sum_{i \in A} u_i + I_A$ and the analogous expression for D, then subtract. Because A contains the two largest singleton utilities, its singleton sum is at least that of every other pair. Rearranging yields Equation 6 and the stated condition.

Contextual utility identity. Using the Möbius expansion, $F(S \cup \{i\}) - F(S) = \sum_{R \subseteq S} m(R \cup \{i\})$ because all terms not containing i cancel. Separating $R = \varnothing$ yields Equation 7.

Identical pointwise scores, different optima. Consider four candidates with $u_i = 1$ for all i. Define one downstream set function with $I_{12} = 1$ and all other pair interactions zero, and another with $I_{34} = 1$ and all others zero. Every pointwise score, annotation membership, and singleton ranking is identical between the two functions, yet their unique optimal pairs are disjoint. More generally, scaling the favored interaction makes the regret of any fixed tie-breaking pointwise rule arbitrarily large. Pointwise metadata therefore cannot identify the optimal evidence set without restrictions on downstream interactions.

Submodularity, curvature, and greedy limits. A set function $F : 2^V \to \mathbb{R}$ is submodular if $\Delta_i F(S) \geq \Delta_i F(T)$ for all $S \subseteq T \subseteq V \setminus \{i\}$, which is equivalent to $\Delta_{ij} F(S) \leq 0$ for all S and distinct $i, j \notin S$. For a normalized, nonnegative, monotone submodular objective under a cardinality constraint, classical greedy maximization guarantees a $(1 - 1/e)$ approximation. Our audit finds positive mixed differences as well as negative ones; a positive difference rules out submodularity for that measured set function. A guarantee for a submodular selection proxy therefore does not automatically transfer to downstream answerer utility. Such a transfer would require additional assumptions relating the proxy to the answerer's objective.

A.19 Outcome scale and causal scope

The causal claim concerns controlled changes to the answerer's input and their effects on its scored output. Evidence status itself is not randomized: evidence and outside frames differ in visual content, so their contrast does not isolate a causal effect of annotation membership. Likewise, output interactions do not identify internal reasoning mechanisms.

Additivity is outcome-scale dependent. Even if every option logit were additive, $z_y(B[S]) = b_y + \sum_{i \in S} a_{yi}$, the scored outcome

$$L(B[S]) = z_{y^*}(B[S]) - \log \sum_{y \neq y^*} \exp z_y(B[S]) \tag{9}$$

need not be additive when there are multiple incorrect options. Thus some interaction can arise from aggregating competing answers, not only from visual integration. Agreement on bounded probability and accuracy scales rules out reliance on unbounded log-odds tails; it does not make interaction scale-invariant. The selection audit supplies the operational consequence: measured interactions improve selection for the stated gold-vs-rest objective, with accuracy reported separately. It does not locate the interaction inside the model.

A.20 Reproducibility details

The shared answer prompt presents the question and lettered alternatives and requests the option letter only; inference uses deterministic constrained one-token decoding. Qwen inputs use fixed 200,704-pixel processor bounds, InternVL uses one patch per frame, and dynamic multi-crop expansion is disabled. All answerers use bfloat16; exact repository commit hashes and the complete prompt are stored in `configs/reproducibility.json`.

Effective evidence regions. Candidate construction clips intervals to the video duration and expands point annotations to a ± 2-second window before clipping. Evidence sampling, outside-frame exclusion, boundary sampling, and annotated-evidence fraction use these effective regions. This expansion applies to 281/500 HERBench and 89/500 LVBench holdout questions; NExT-GQA has no point annotations in the holdout. These are operational sampling regions, not newly verified frame-level annotations.

The released configuration records exact model revisions, processor settings, prompts, seeds, and runtimes. Candidate construction and causal sampling were frozen before inference; technical decoding exclusions are documented, and post-hoc analysis restrictions are labeled separately. Primary frames are decoded to RGB and shared across model families. Evidence-access and selector calls use chronological order except in explicit order controls. Factorial calls deliberately keep candidate positions and context identities stable across their four matched conditions; a separate timestamp-sorted replication tests natural chronological presentation. The full valid-option distribution was stored during inference. The release retains it where required for the reported score-scale analyses and otherwise provides analysis-ready log odds, probabilities, accuracies, and intervention identities.

For NExT-GQA, dense candidate-pool validation retained 5,482 of 5,553 census questions and logged 71 technical video decode exclusions. The causal selector audit contains 495 technically decodable questions rather than silently assigning outcomes to five failures. InternVL3.5-38B's original intervention runs exceeded an 80GB device at the 32-uniform-frame condition on all three grounded benchmarks and terminated before uniform-4 and uniform-8. A targeted continuation completed those two conditions without repeating earlier calls. The three uniform-32 conditions each retain one pre-failure record; these partial conditions are excluded from every contrast and from Table 3's call total, and no missing value is imputed.

We exclude adaptive ReaSon from the fixed-pool table because its detector/grounder observations and 16-frame policy interface expose additional video information that the fixed 64-frame pool does not charge. A fair comparison therefore requires a separate raw-video observation protocol. No usable public LensWalk implementation was available. AKS is an exact output alias of SigLIP relevance on the frozen pool and is labeled as such in artifacts rather than counted as independent evidence.

Compute accounting. Each preparation or answerer-inference process used one GPU, typically with 12 CPU cores and 96GB host memory; aggregation and validation used 2–4 CPU cores and 16–32GB. The retained records do not provide a reliable device inventory or elapsed-resource ledger, so we report model calls and resource bounds rather than GPU-hours (which we did not properly record). We release portable Bash entry points for preparation, inference, verification, and figure reproduction. Selector comparisons charge the common 64-frame candidate pool and answerer calls uniformly; DINO and SigLIP feature construction is selector-side preprocessing and is not included in the answerer-call totals. Relevance–diversity reuses SigLIP features, MDP^3 and plain

DPP share one SigLIP embedding pass, and BOLT-ITS uses its own preprocessing. This study makes no runtime-superiority claim.

A possible direction for selection. One hypothesis is that a relevance shortlist could make composition-aware refinement affordable. Given M candidates, an estimated objective might take the form $\sum_{i\in S} \widehat{u}_i + \sum_{\{i,j\}\subseteq S} \widehat{I}_{ij}$, subject to $|S| = K$. Here the hats matter: the gold-conditioned quantities measured in our audit are unavailable at deployment. Useful estimates would have to be obtained without gold answers and validated under the intended context and presentation rule. There are $\binom{M}{2}$ pairs—120 for $M = 16$—but this quadratic count describes score construction, not the cost of optimizing the resulting objective. Pairwise estimates can also miss higher-order effects, and shortlisting can discard a useful partner before refinement begins. Whether this approach improves on an existing proxy such as DPP at an acceptable cost remains an empirical question.

Released artifact bundle. The public repository provides deterministic manifests, eligibility and exclusion records, analysis-ready intervention outcomes, per-candidate utilities, per-pair interactions, higher-order components, selector outcomes, cluster-bootstrap summaries, sign-flip tests, paper-facing analysis scripts, and a SHA-256 release manifest verifying data integrity.

A.21 Model fingerprints and sensitivities

Figure 10 breaks the evidence-access interventions down by model and benchmark. The figure is descriptive: it shows that sensitivity to evidence dose, frame order, and the no-frame prior varies across evaluation settings.

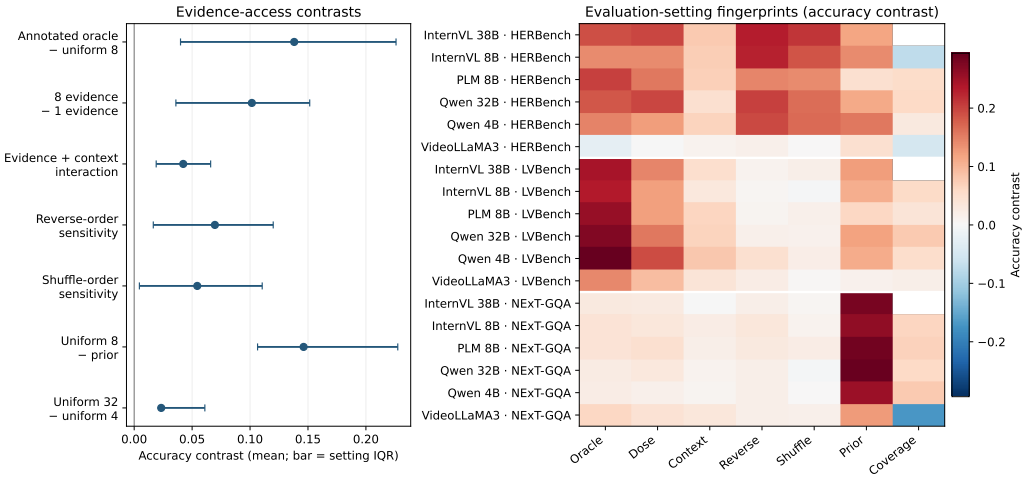


Figure 10: Evidence-access contrasts across models and benchmarks. Left: equal-setting macro summaries for evidence dose, order, and prior-only interventions. Right: the corresponding model–benchmark values. Blank coverage entries mark the three incomplete InternVL3.5-38B uniform-32 conditions and are excluded from the macro summary. Labels identify measured condition contrasts; variation across evaluation settings is descriptive and should not be read as a test of model architecture.

A.22 Qualitative case studies

Figures 11 and 12 show what measured interactions look like, from the tails and from the center of the distribution. "Redundancy" and "complementarity" describe the measured set function, not the answerer's internal reasoning.

Tail cases. Figure 11 shows two deterministically selected high-magnitude cases, one negative and one positive. They illustrate signed output interactions rather than representative effect sizes: $|I| > 10$ occurs in 1.02% of evidence/evidence records and 0.06% of outside/outside records. The primary result does not depend on such cases. Removing them, or trimming or winsorizing each setting at its 99th percentile, preserves a positive evidence-pair gap in 18/18 settings (Appendix A.9).



Figure 11: High-magnitude interactions from the frozen registry. Top: a negative evidence/evidence pair on NExT-GQA; bottom: a positive evidence/evidence pair on LVBench. Each row shows the two candidate frames with their timestamps, the question, answer choices, gold answer, and answerer. I is the difference-in-differences in gold-vs-rest log odds (Equation 4).

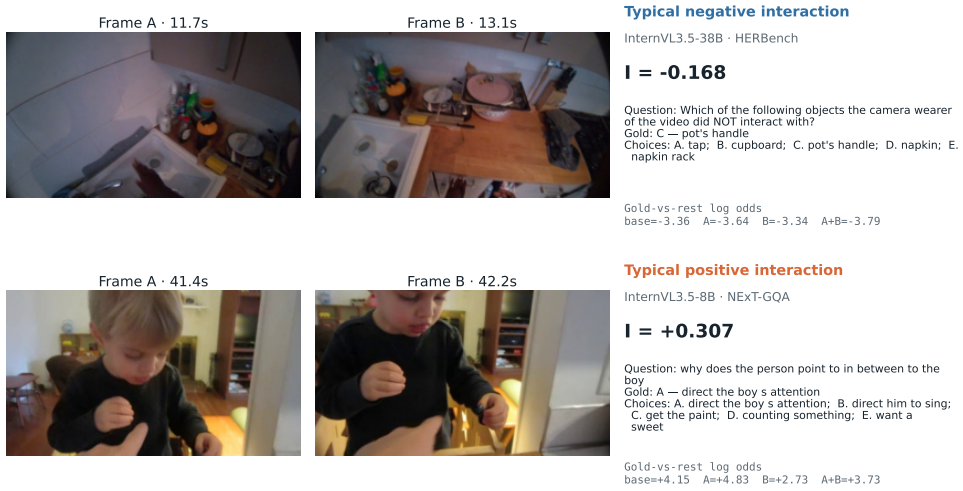


Figure 12: Typical, non-extreme interactions from the frozen registry. Both cases fall between the 25th and 50th percentiles of $|I|$ within their sign and evidence–evidence stratum. For legibility, we show the shortest-question case in each fixed stratum; this secondary presentation rule does not inspect the frames or further rank outcomes. Images are center-cropped to equal display areas. The right column gives the complete answer choices, the answer text, and the gold-versus-rest log odds for the baseline, either candidate alone, and both candidates together. Their difference-in-differences gives the displayed interaction. These examples calibrate ordinary measured interactions rather than explaining the model's internal mechanism.

Typical cases. Figure 12 complements the tails with cases whose absolute interactions fall between the 25th and 50th percentiles, closer to the center of the measured distribution. The released registry provides additional cases across pair types, signs, and magnitude ranges.

A.23 Limitations

Three boundaries define where these conclusions apply.

The evidence annotations are imperfect. Annotation-defined regions vary in precision across benchmarks. LVBench reference windows in particular are localization hints, not exact frame-level proof, and no additional human verification was available to us. The annotation-breadth analysis in

Table 8 reduces this concern by showing the effect is strongest under the tightest windows, but it does not eliminate it. This is also why we never treat annotated-evidence fraction as a benchmark-comparable effect size.

The interface is narrow by design. We study constrained multiple-choice scoring through a fixed decoded-frame interface, and the control this buys us is exactly what limits its reach. Native video tokenizers, adaptive observation policies that choose what to look at next, and open-ended generation all change both the observation process and the outcome measure. Transfer to those settings is untested. Our contexts are likewise frozen, deterministic outside-evidence constructions; the new-context control preserves the direction, but richer natural contexts could shift interaction magnitudes.

The audit is diagnostic. The recovery analysis consumes downstream measurements unavailable at inference time, so it measures the value of composition information rather than showing how to obtain it cheaply. Our primary statistic likewise answers "how far from additive," not "in which direction"; signed results are reported separately. Section 7 sketches a possible selection direction whose accuracy, transfer, and cost remain to be tested.