SELECT, COMPRESS, REINVEST: A CONTROLLED STUDY OF VISUAL-TOKEN ALLOCATION IN LONG-VIDEO MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video language models cannot look at every frame: an hour sampled once per second is 3,600 images, and a system keeps only a small fixed slice of that pool. Which frames survive that slice is usually treated as a preprocessing detail; we test whether it should be. Published selectors make the comparison hard because they change the frame scorer, the prompt boundary, the resolution policy, and the answering model all at once. We hold each fixed and vary one decision at a time—selection, spatial compression, and reinvestment of the savings—across six training-free selection rules, three long-video benchmarks, and two answering models. Selection is the largest single lever: on LongVideoBench's hour-long bin, eight query-selected frames beat sixteen uniformly spaced ones by 6.9 points, and Orthogonal Matching Pursuit—an unmodified, decades-old sparse-approximation algorithm—matches or comes within a point of every purpose-built selector we compare it against, across all three benchmarks. Compression is close to free: halving each frame's spatial budget at fixed timestamps costs at most 0.44 points. Reinvestment is where that budget turns back into accuracy: spending the freed tokens on twice as many compressed frames, at a measured cost within 5% of the original eight, returns a further 1.5 to 3 points; compression only pays off once its savings are spent this way. Along the way, an implementation bug in our own AKS baseline and a 0.07–3.74-point gap between two harnesses running the same published rules at the same budget show why these comparisons need to happen inside one controlled harness rather than across papers.

1 INTRODUCTION

A video language model cannot look at a whole video. Decoded at one frame per second, a ten-minute clip is 600 images and an hour-long one is 3,600; a long-video system typically keeps only a small fixed slice of that pool: eight frames in some systems, thirty-two or sixty-four in others, never the whole thing. Everything the model will ever know about the video passes through that handful of frames, so the rule that picks them is not a preprocessing detail. It is the first and tightest bottleneck in the pipeline.

The default rule is still uniform sampling: take a fixed number of frames spread evenly across the timeline, regardless of what was asked. We test at eight, the one budget at which a published Qwen3-VL-8B/LongCLIP baseline exists to match against (Table 6, Appendix G). Our clearest single result is that this default is expensive. On hour-long LongVideoBench videos, eight frames chosen for their relevance to the question beat sixteen uniformly spaced frames by 6.9 points (p=.0011), and the same pattern holds one budget higher on ten-minute videos. Half the frames, better answers. Spending an *answerer visual-token* budget on *which* frames to keep buys more than spending it on *how many*; the selection stage has its own upstream decoding and encoding cost, unmeasured here and scoped in §5.1 rather than included in that budget.

Many recent systems already replace uniform sampling with query-aware retrieval, temporal coverage, diversity, or learned evidence scores (Tang et al., 2025; Sun et al., 2025; Zhu et al., 2025b; Wang et al., 2026a; Peng et al., 2026), and others jointly adapt frame choice and spatial resolution (Zhang et al., 2025a; Chen et al., 2026a; Wang et al., 2026b) or prune tokens once frames are

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.

inside the vision stack (Li et al., 2026; Hong et al., 2026), reporting large end-to-end gains. What they do not settle is a narrower question: holding the scorer, the frame budget, and the answering model constant, which selection rule actually supplies the most useful evidence? Published tables cannot answer this, because those components move together—a selector can look strong because it ships with a better encoder, a dynamic-resolution system because it fits in more frames rather than allocating resolution well. Comparing two numbers from two papers compares two experiments, not two ideas.

This paper isolates the pieces: we fix one frame scorer, prompt boundary, answering model, and evaluation harness, then intervene on a single decision at a time. **Select**: change which timestamps are chosen, holding frame count and resolution fixed. **Compress**: hold timestamps fixed and shrink the spatial budget per frame. **Reinvest**: spend the recovered budget on more timestamps rather than sharper ones. Because each step is a paired comparison on the same questions, the difference it produces is attributable to that step alone; a fourth comparison swaps the scorer itself, to check the resulting ordering is not an artifact of the encoder we happened to freeze.

The selection rule at the centre of these interventions is deliberately old: Orthogonal Matching Pursuit (Pati et al., 1993), a greedy sparse-approximation algorithm from 1993, repeatedly takes the candidate most correlated with what the query still needs, then projects that direction away before choosing again. We adopt it unmodified and untuned, a methodological choice rather than a concession: an off-the-shelf rule has no hyperparameters for us to fit to these benchmarks, so a gap between it and a competing selector reflects the controlled comparison rather than unequal effort, and how well it does is itself a measurement of how much recent purpose-built selectors gain from their subset rules as opposed to their scorers and pipelines.

The three interventions rank consistently across the settings we tested, though estimated in separate contrasts rather than one factorial design, so the ordering is a synthesis rather than an identified effect ranking: selection is the large lever (5.7 to 11.8 points over uniform sampling depending on benchmark, sometimes beating uniform inputs with twice as many frames), compression is nearly free but not by itself useful (at most 0.44 points at fixed timestamps), and reinvestment converts that free budget into a real gain of 1.5 to 3 points. None of this is universal—the benefit depends on which model reads the frames and which benchmark asks the question, and OMP's later picks drift toward content that is visually novel but irrelevant—and we report those boundaries rather than smoothing them.

Contributions.

- **A matched comparison of six training-free selectors** under one scorer, prompt boundary, frame budget, and answerer harness; OMP stays within one point of LDDR's stage-1 selector.
- **Selection substitutes for frame count.** OMP significantly outperforms uniform sampling while using half as many frames, in two disjoint LongVideoBench duration bins.
- **A decomposition of the fixed-budget allocation decision.** Roughly halving the per-frame spatial budget preserves accuracy, bounded by a two-one-sided-test interval rather than a bare null result; reinvesting the savings in additional keyframes then improves long-video QA.
- **The ranking survives a scorer swap.** Replacing LongCLIP with SigLIP changes 67–84% of the selected frames yet leaves the Uniform $\prec$ Top-k $\prec$ OMP ordering intact on LongVideoBench-600 s.
- **Aggregate accuracy can conceal implementation error.** A padding branch in our own AKS port made that baseline select global top-k at every $k < 32$; correcting it moved roughly 99.5% of selected frames yet changed LVBench accuracy by 0.07 points, caught by a cross-arm overlap check.
- **Explicit boundaries on these claims**, from paired tests, two answerer families, three benchmarks, a residual-geometry diagnostic, and a purposive failure audit, unevenly: the equivalence result covers one pooled setting and the audit is exploratory.

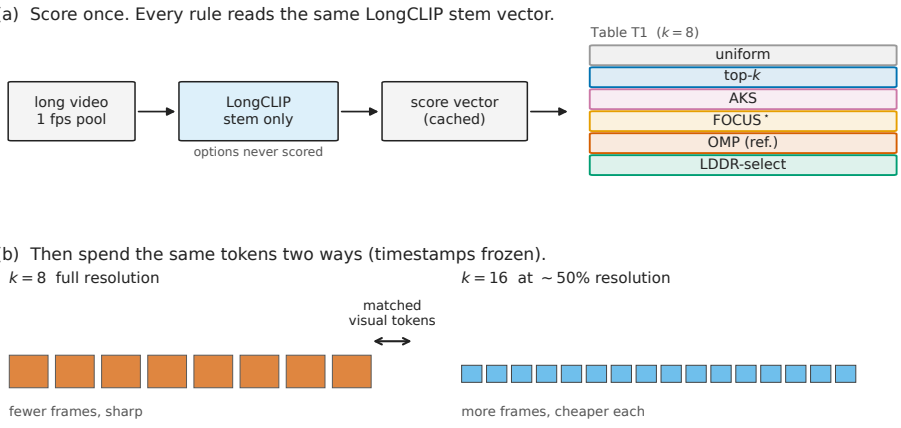


Figure 1: Controlled protocol. (a) A 1 fps pool is encoded once with LongCLIP on the question stem; every k=8 selector reads the same cached scores. (b) Reinvestment: eight full-resolution frames vs. sixteen compressed frames at audited visual-token cost within 5%. The fixed-timestamp compression step (§5.3) is not diagrammed. FOCUS*: schedule replay on dense LongCLIP scores, not the original ITM scorer.

2 Related Work

Query-aware frame selection. Training-free selectors combine relevance, timeline coverage, and diversity in different proportions; the latter two are inherited from classical subset selection—maximal marginal relevance (Carbonell & Goldstein, 1998) and determinantal point processes (Kulesza & Taskar, 2012)—which we also run directly as selector arms (Appendix C). AKS partitions the timeline for coverage (Tang et al., 2025); FOCUS treats clips as bandit arms (Zhu et al., 2025b); MDP3 combines relevance, diversity, and sequentiality via a DPP (Sun et al., 2025); AdaRD-Key and Adaptive Greedy pursue the same tradeoff (Zhang et al., 2025b; Huang & Zhu, 2026); QCA and EFS use per-segment quotas or query-relevant anchors (Peng et al., 2026; Chen et al., 2026b); ReQuest and query-conditioned evidential sampling add uncertainty-triggered computation and frame-level evidence estimates (Kim et al., 2026; Wang et al., 2026a). These differ in scorer, candidate pool, and frame budget as well as subset rule, confounding the rule with its pipeline.

Selection with spatial adaptation and token reduction. A second line couples frame choice to resolution: Q-Frame pairs CLIP-based selection with multi-resolution scaling (Zhang et al., 2025a); LDDR combines a linear-DPP selector with Group-DPP importance (Chen et al., 2026a); DAFS jointly optimizes candidate-pool size and per-frame token budget from MLLM attention (Wang et al., 2026b). A joint gain here has at least three possible sources—better timestamps, better resolution policy, or more temporal coverage—and none of these papers separates the three. A third line cuts cost after visual encoding instead of at selection time (AdaCodec (Hou et al., 2026), AdaptToken (Qi et al., 2026), Vista-LLM and MoPrune (Li et al., 2026; Hong et al., 2026)); concurrent work AdaAlloc (An & Grauman, 2026) poses our question directly, asking how to split a fixed budget between global context and high-resolution local evidence, though no preprint or code was available at the time of writing. We ask a coarser but experimentally separable version upstream of all of them: before any token-level pruning, does a long-video answerer gain more from sharper retained frames or from additional moments?

3 Method: Three Controlled Interventions

3.1 The experimental unit

Every comparison in this paper is paired: the unit is a single benchmark question answered twice, once under each of two input policies, with everything else held constant, which lets us count the

questions each policy wins rather than compare two accuracy numbers that may differ for unrelated reasons. The three interventions differ only in what the two policies are allowed to change: **selection** changes the timestamps while holding frame count, scorer, and resolution fixed; **compression** holds the timestamps fixed and changes the per-frame spatial budget; **reinvestment** spends the budget saved by compression on additional timestamps. A fourth intervention sits outside that sequence and tests the design itself: replacing the scorer while holding the selection rule, budget, and answerer fixed asks whether the ordering the first three produce is a property of the rules or of one encoder (§5.2).

3.2 One scorer, shared by every rule

Videos are decoded at 1 fps; each candidate frame and the question stem are encoded once with LongCLIP (Zhang et al., 2024), and the resulting embeddings and stem similarities are cached, so every selector reads the same cache and no rule benefits from a better encoder than another.

Answer options never enter selection, a convention rather than a neutral choice, so we measure the cost (Appendix I): fused-query scoring changes 41.9% of top-k's picks and 53.3% of OMP's on LongVideoBench-600 s, and those are also better frames (OMP $+3.88$, $p{=}.033$), a shift a pre-registered 3600 s replication reproduces; a foreign-options control recovers 56% of the gain, ruling out that it belongs to the specific answer set. We nonetheless score with the stem alone throughout, following the official AKS implementation (Tang et al., 2025); the two-to-four-point gap is the price of that deliberately conservative choice.

3.3 Selection rules

The matched comparison contains uniform sampling, cosine top-k, AKS (Tang et al., 2025), FOCUS* (Zhu et al., 2025b), OMP, and LDDR-select (Chen et al., 2026a). Every rule consumes our LongCLIP stem scores, native only for LDDR-select: AKS and FOCUS score with BLIP-ITM in their published form, so both are reproductions under a substituted scorer. FOCUS* also replays the published clip-bandit schedule against the dense LongCLIP vector rather than the original budgeted ITM process, the substance of the method (Appendix A); LDDR-select is the stage-1 Linear-DPP selector only, a published, separately tabulated configuration rather than a truncation we invented.

Let $E \in \mathbb{R}^{N\times d}$ contain L2-normalized frame embeddings and let q_0 be the normalized stem embedding. At each step OMP takes the frame most correlated with the current residual, then subtracts the span of everything already selected from the query:

$$\begin{aligned} b_r &= \arg\max_{i \notin B_{r-1}} e_i^\top q_{r-1}, \\ q_r &= q_0 - \mathrm{Proj}_{\mathrm{span}\{e_j : j \in B_r\}}(q_0), \end{aligned} \tag{1}$$

where $B_r = B_{r-1} \cup \{b_r\}$ and $B_0 = \emptyset$. In its original signal-processing setting this is sparse reconstruction, the residual shrinking as selected atoms explain more of the target; whether that survives in a contrastive image–text embedding space is an empirical question that §E tests directly, not an assumption. What the projection reliably does here is suppress directions already covered by earlier picks.

3.4 Spatial budget and reinvestment

For the fixed-timestamp intervention, frames are resized before reaching the model's processor (Qwen3-VL's own internal resize is disabled so ours is the only one applied); the main compressed arm targets a mean spatial fraction of about 0.53 (D@53), residual-proportional across the selected frames (definitions in Appendix H). Two controls test what the schedule's *shape* contributes: a flat split at the same mean fraction, and a reconstruction of LDDR's stage-2 Group-DPP importance on identical OMP-8 timestamps.

The reinvestment arm selects $k{=}16$ timestamps with an approximately 50% per-frame cap, a different quantity from the 0.53 mean above; because the comparison only means something if the two arms really cost the same, we audit tokens empirically rather than assume the resize ratio transfers—a script reproduces Qwen's smart-resize rules from each video's dimensions, and the sixteen-frame

Method	LongVideoBench	Video-MME	LVBench
Uniform	.5654 (−5.69)	.5637 (−5.85)	.3454 (−11.81)
Top-k	.6028 (−1.95)	.5704 (−5.18)	.4319 (−3.16)
OMP	.6223	**.6222**	.4635
AKS	.5916 (−3.07)	.6059 (−1.63)	.4287 (−3.48)
FOCUS*	.5819 (−4.04)	.5578 (−6.44)	.3983 (−6.52)
LDDR-select	**.6320** (+0.97)	.6193 (−0.29)	**.4693** (+0.58)

Table 1: Matched selector comparison at k=8, Qwen3-VL-8B. Cells are accuracy; parentheses give the difference from OMP in points. FOCUS*: schedule replay (Appendix A); LDDR-select: stage 1 only.

arm consumes 0.996 and 0.984 times the tokens of the eight-frame arm in the 600 s and 3600 s bins, so the LongVideoBench result is conservative: the winning arm is also cheaper. The same audit on Video-MME and LVBench gives 1.023 and 1.008 (Video-MME short 1.046), so there it costs slightly more.

4 Evaluation Setup

We evaluate on LongVideoBench (Wu et al., 2024) (complete official validation set, n=1337), Video-MME (Fu et al., 2025) (n=2700), and LVBench (Wang et al., 2024) (n=1549), abbreviated LVB and V-MME where column width requires it; LongVideoBench is additionally stratified into 15, 60, 600, and 3600 s duration bins, enabling the duration analysis in §5.1. The primary answerer is Qwen3-VL-8B-Instruct (Qwen Team, 2025) via lmms-eval (EvolvingLMMs-Lab, 2024), greedy decoding, subtitles disabled; InternVL3-2B/8B (Zhu et al., 2025a) provide a second open-model family, and GPT-5-mini checks selected contrasts on the same saved timestamps.

Frame budgets are k=8 for selection and k=16 for reinvestment throughout; main claims concern paired accuracy differences under matched inputs, not absolute leaderboard position.

Statistics. Accuracy comparisons use exact two-sided McNemar tests (McNemar, 1947) on paired correctness outcomes. Compression claims need more than a non-significant McNemar result, so where we claim two arms are interchangeable we report two one-sided tests (TOST) with 90% confidence intervals (Schuirmann, 1987), which bound the effect rather than merely failing to find it; the $\pm$2/$\pm$3/$\pm$4-point margins were chosen after seeing the runs rather than preregistered, so we report both the narrowest margin the data support and the adjacent margin they do not. Tests are contrast-specific and uncorrected for multiplicity.

5 Results

5.1 Choosing frames beats having more of them

The most direct evidence that selection matters is that it substitutes for frame count: on 3600 s LongVideoBench videos, OMP with eight frames reaches .5461 while uniform sampling with sixteen frames reaches .4770, a 6.9-point gain on half the input (p=.0011); the same pattern holds one budget higher, OMP-16 reaching .6578 against .6044 for uniform-32 on 600 s videos (p=.018). Both contrasts compare a policy that is cheaper *in answerer visual tokens* against a more expensive one, and the cheaper policy wins—this accounting covers the answerer only, since producing the selected set also requires decoding at 1 fps and encoding the whole pool with LongCLIP, a cost we do not measure and a single-query-per-video deployment would not amortise. Doubling the frame count is the obvious way to spend more compute on a long video, and in these two bins it is the worse way.

Table 1 places that result in the wider comparison, with all six rules reading the same cached scores: OMP improves on uniform sampling by 5.69 points on LongVideoBench, 5.85 on Video-MME, and 11.81 on LVBench, decisively on all three ($p<10^{-3}$). LDDR-select is the one rule that keeps pace, staying within a point of OMP on all three benchmarks with no detectable paired difference

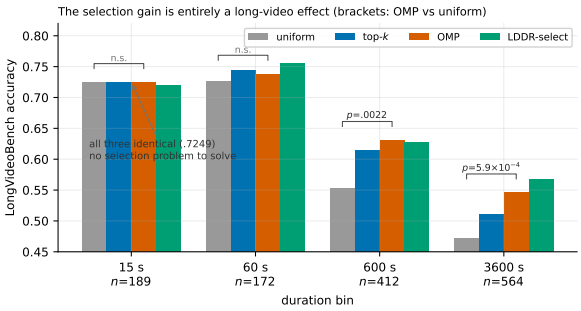


Figure 2: OMP's gain over uniform appears only once the 1 fps pool outgrows the eight-frame budget: identical at 15 s (.7249), +7.8 at 600 s, +7.5 at 3600 s. A threshold, not a duration gradient.

on Video-MME or LVBench (p=.70 and .57), nor on LongVideoBench's two long bins (p=1.00 at 600 s, .18 at 3600 s; a pooled test needs per-item outcomes for the 15 s/60 s bins that were not retained).

The duration split explains when any of this matters: on 15 s clips, uniform sampling, top-k, and OMP return exactly the same accuracy (.7249), since eight frames out of fifteen candidates cover the video regardless of how they are picked (LDDR-select is the exception, differing slightly at this size), but the gains appear once the candidate pool outgrows the budget—7.8 points at 600 s (p=.0022), 7.5 at 3600 s (p=5.9×10^{-4})—and do not keep growing after that, a threshold rather than a duration gradient. (The bins also differ in content and question composition, so this is an observational contrast; isolating pool size from duration would need a controlled truncation experiment we did not run.) Which *kind* of rule is needed is more benchmark-dependent than the headline suggests: on Video-MME, top-k and FOCUS* do not separate detectably from uniform sampling while OMP, LDDR-select, and AKS do (AKS +4.22, p=4.0×10^{-6}), but on LongVideoBench OMP's advantage over plain top-k is small and often undetectable within individual bins—combining query relevance with suppression of already-covered directions is consistently strong here, while any individual simpler heuristic does not transfer on its own (whether that ordering is a property of the rules or of LongCLIP is the subject of §5.2).

5.2 The ranking survives a scorer swap

Every number in Table 1 comes from a single frozen scorer, which leaves open whether the ordering reflects the selection rules or LongCLIP's particular geometry. Replacing LongCLIP with SigLIP-so400m (Zhai et al., 2023), holding everything else fixed, is a substantial intervention rather than a cosmetic one: on the 600 s bin the two scorers agree on only 1.32 of eight frames for OMP and 2.62 for top-k, so 67 to 84% of what the answerer sees changes, yet uniform is beaten by top-k and top-k by OMP under both scorers, and OMP's advantage over uniform stays significant in each case (p=.0024 LongCLIP, p=.00026 SigLIP; full table and caveats in Appendix J). This weakens, without ruling out, the worry that Table 1 is an artifact of LongCLIP.

5.3 Half the pixels cost little, not nothing

Holding the selected timestamps fixed and changing only how many pixels each frame is worth (per-benchmark numbers in Table 8, Appendix H), moving to the compressed arm shifts accuracy by at most 0.44 points in either direction across the three Qwen benchmark aggregates, and a uniform-timestamp control on pooled LongVideoBench long videos moves by 0.31 points (p=.784)—consistent with the stability being a property of the compression rather than of OMP-selected frames specifically, though a non-significant control is not itself an equivalence result. We bound one of these differences properly: on the pooled LongVideoBench long bins (n=976), the compressed OMP arm is 0.72 points *higher* than full resolution, 90% CI $[-0.60, +2.03]$, fitting inside a $\pm$3-point margin (TOST p=.0022) but narrowly failing at $\pm$2 (p=.054); InternVL3-8B clears the tighter $\pm$2 margin under a more aggressive tile reduction (Appendix F).

Answerer / benchmark	k=8 full	k=16 compressed	Δ (pp)
Qwen / LongVideoBench	.6223	.6447	$+2.24$
Qwen / LVB long pool	.5820	.6055	$+2.36$ (p=.0346)
Qwen / Video-MME	.6222	.6378	$+1.56$
Qwen / LVBench	.4635	.4939	$+3.04$ (p=.0009)
GPT-5-mini / LVBench	.4900	.5139	$+2.39$ (p=.039)

Table 2: Reinvestment at audited token ratios of 0.98–1.05 (OMP-selected timestamps, both arms; Video-MME 1.023, LVBench 1.008); the compressed arm approximately doubles temporal coverage. The LongVideoBench and Video-MME overall rows lack a paired test; only the long-bin subset above is matched, and Video-MME was never paired (§7).

Compression being cheap does not mean the allocation *shape* is irrelevant. Controls separate the two (Appendix H): OMP's ranking, useful for choosing frames, does not detectably improve accuracy when used to distribute pixels among them, while a reconstruction of LDDR's stage-2 Group-DPP importance does, by 1.84 points on the pooled long bins (p=.0198). The conclusion is therefore three-part and narrower than "resolution does not matter": coarse compression is close to free, a naive priority-proportional rule adds nothing, and a better-designed importance signal can add something.

5.4 Spending the savings on more frames

The direction of this intervention is not new: Chen et al. (2026a) report fixed-resolution ablations under a single total token budget in which 1024 tokens per frame reach 59.76 on LongVideoBench, 512 reach 61.03, and 256 fall back to 59.31, an interior optimum from halving the per-frame budget. What that ablation does not establish is whether the parity it assumes holds, whether the difference is distinguishable from noise, or whether the effect needs their selector; we address those three questions, so the contribution here is an audited and paired replication rather than a new observation.

Table 2 spends the recovered budget on temporal coverage, comparing eight full-resolution frames against sixteen compressed ones at audited token cost within 5%. LongVideoBench improves by 2.24 points overall and 2.36 on the pooled long bins (p=.0346); LVBench gains 3.04 (p=.0009), Video-MME 1.56, and GPT-5-mini on LVBench 2.39 (p=.039).

We cannot show that this requires OMP: a four-cell selector-by-budget design (Appendix H) gives $+1.13$ points under uniform sampling and $+2.56$ under OMP, an interaction that does not reach significance (p=.2614). The defensible claim is that temporal reinvestment works with OMP, points the same way under uniform sampling, and that its selector-specificity remains open at roughly the one-point scale.

Two further results are reported in the appendix rather than here, for space. Appendix E traces OMP's residual geometry to characterize what its later picks actually add once the query-directed signal is exhausted after roughly eight picks, and connects that trace to a purposive audit of OMP's failures. Appendix F hands the identical OMP and uniform timestamps to a second open answerer family, InternVL3-2B/8B, changing only the model: the selection gain replicates on LongVideoBench and LVBench at both sizes, Video-MME does not cooperate at the 8B size, and the allocation (compression/reinvestment) results transfer more smoothly than selection does. We summarize both findings in the Discussion below.

6 Discussion

The three decisions are not equally worth optimizing. Under a fixed visual-token budget, our interventions rank consistently: choosing frames is the large lever (a small query-focused set beats uniform sampling, sometimes even with twice as many frames), compressing them is close to free but buys nothing on its own, and its accuracy value lies in what the savings are spent on—temporal coverage returns a further 1.5 to 3 points, though compression may still be worth it for latency or memory even unreinvested. The practical ordering is therefore to fix selection first, treat resolution as the slack variable, and reinvest rather than pocket the difference.

An off-the-shelf pursuit algorithm is competitive with purpose-built selectors. The most uncomfortable observation in Table 1 is which rules are close together: OMP has no video-specific machinery and no tuned hyperparameters, yet leads AKS, FOCUS*, and top-k on all three benchmarks while sitting within a point of LDDR's stage-1 selector—evidence about where recent selectors' gains live, not that they are poorly designed. Once the scorer, prompt boundary, frame budget, and answerer are held fixed, the full accuracy range across the five query-aware rules is smaller than the gap from uniform sampling to plain top-k on LVBench (7.10 against 8.65), modestly larger on LongVideoBench (5.01 against 3.74), and dwarfs it on Video-MME (6.44 against 0.67), so on Video-MME especially, *which* rule matters far more than whether it is query-aware at all. The defensible generalization is therefore not "subset rules barely matter"—on one of three benchmarks the between-rule span is the smaller quantity, and on the other two it is not—and the corollary for evaluation practice is concrete: a selector comparison is only informative if the scoring stage is shared, and a paper reporting a new rule should include a classical pursuit or greedy baseline on its own scores before claiming the rule is what helped.

Two matched harnesses disagree by as much as the effects they measure. LDDR reports a cell matched to ours on answerer family, frame budget, and encoder (Table 6, Appendix G); our reproduction is smaller in eight of nine comparable cells, by 0.07 to 3.74 points—a range covering most of the gaps published selector comparisons are built on, and evidence that cross-paper deltas of this size are fragile rather than that either harness is wrong, since the two harnesses' uniform-sampling floors themselves differ by as much as 6.46 points. The same cross-check caught an implementation bug in our own AKS port that had silently reproduced top-k (a padding branch left every leaf's quota at zero): correcting it changed roughly 99.5% of selected frames while moving the aggregate by as little as 0.07 points on LVBench, which a single-number comparison could not have caught, so we now run a pairwise frame-overlap check between the five rules with comparable frame-index picks (Uniform is never exported to a picks file; Appendix G gives the full table and diagnostic).

Joint allocation systems deserve decomposed reporting, and the mechanism suggests where to look next. This decomposition sharpens how to read systems such as Q-Frame, LDDR, and DAFS (Zhang et al., 2025a; Chen et al., 2026a; Wang et al., 2026b), whose end-to-end gains combine several causal routes at once: our controls find no value in a simple priority-proportional resolution rule, a positive effect from reconstructed LDDR-style importance, and the clearest gain from temporal reinvestment—three answers a single end-to-end number would have blurred into one, and future methods would be easier to build on if they reported the three separately alongside the joint result. OMP itself appears to work in a way its original derivation does not predict (Appendix E): the text query is nearly orthogonal to the frame-embedding cone, so the residual barely shrinks and what remains useful is relevance plus decorrelation—late picks extend multi-scene coverage after the text residual flattens, a genuine benefit, but unconstrained diversity also rewards irrelevant novelty, the usual cause of missed evidence in our failure audit. Relevance floors, event-boundary constraints, and scorers with a smaller modality gap are the natural next interventions.

Conclusions should span durations and answerers. Selection remains a property of the answerer–benchmark pair (Appendix F): InternVL3-8B is nearly insensitive to OMP on medium and long Video-MME yet gains over ten points on LVBench, and OMP's advantage over uniform sampling is exactly zero at 15 s—a claim measured on one benchmark at one duration with one answerer is not safe to generalize, and our own results would have supported at least three different headlines depending on which cell we had chosen to report.

7 LIMITATIONS

Two selectors are controlled reproductions; equivalence margins were chosen after seeing the data; mechanism evidence is encoder-specific. FOCUS* and LDDR-select are the schedule/stage-1 substitutions described in §3, and neither row is a reproduction result for the original method. The Qwen compression result covers pooled LongVideoBench 600 and 3600 s examples within a $\pm$3-point margin and misses $\pm$2; because the margin was selected after inspecting the interval rather than fixed in advance, we treat the result as descriptive rather than confirmatory, and the tests are uncorrected for multiplicity. The small LVBench differences are descriptive, and the

selector-by-budget interaction is underpowered for effects near one point. The residual geometry (Appendix E) comes from frozen LongCLIP stem embeddings, mostly on LongVideoBench-600 s; the selector ordering survives a swap to SigLIP (§5.2), but both encoders share a contrastive image–text objective, and BLIP-ITM heads, MLLM-attention scorers, and subtitle-aware selection remain untested. The model-assisted failure audit examined selected rather than randomly sampled failures, so its counts are exploratory and cannot estimate population rates.

Some secondary results were measured in a second compute environment, and evaluation scope is incomplete. Tables 1–2 come from a single stack (RTX PRO 4500, PyTorch 2.11/CUDA 13.0); the selector-by-budget interaction, allocation-shape contrast, and variant sweep (Appendix C) were measured on L40S GPUs with PyTorch 2.6, with every arm of each contrast—including a re-run of the full-resolution baseline—executed inside that second environment, so no reported difference is computed across stacks. Replicating the baseline in both quantifies rather than assumes the drift (the uniform eight-frame arm reads 0.36 points higher on the second stack in the 3600 s bin, OMP 0.71 lower), so absolute accuracies from these sections are never placed in the same table as Tables 1–2, though paired differences are unaffected. Subtitles are disabled throughout, leaving subtitle-anchored questions partly inaccessible to every selector; GPT-5-mini compression uses a pixel ratio rather than Qwen token accounting; and our rule set is narrower than the comparison table it is matched against—MDP3 and Q-Frame appear in Chen et al. (2026a) but neither has a published number at our exact cell, so neither port could be validated against anything, the condition that let the AKS padding error stand undetected (Appendix G).

8 Conclusion

In our matched harness, which frames reach the answerer is the largest of the three allocation decisions we varied: eight well-chosen frames beat sixteen uniformly spaced ones in LongVideoBench's hour-long bin, and OMP, run unmodified, is enough to get that result, outperforming uniform sampling by 5.7 to 11.8 points and staying within one point of a purpose-built modern selector. Roughly half of the per-frame spatial budget can then be removed with little observed change, and reinvesting those tokens in additional keyframes returns a further 1.5 to 3 points. How far any of this generalizes beyond this harness is exactly what the evaluation-sensitivity results above bound: select relevant frames first, compress them second, and use the savings to see more of the video, though how much it helps depends on the video's length, benchmark, and answering model.

Ethics Statement

The experiments use existing public video-QA benchmarks and frozen open or API models. No new human subjects, personal data, videos, or annotations were collected. Question stems, but not answer options, enter the selector. Benchmark videos remain governed by their original licenses and are not redistributed with the released artifacts.

Reproducibility Statement

Evaluation code, selector implementations, analysis scripts, the selected frame indices for every rule and budget, and per-item predictions permitted by the benchmark licenses are released at `https://anonymous.4open.science/r/long-video-token-allocation-6BD6/`. The release includes the scorer-swap pipeline of §5.2, the fused-query arms of §3, the script that builds Table 6 from the cited papers' reported numbers, and the pre-registration documents for the registered arms, including the 3600 s replication of §3. The release excludes benchmark video files and API credentials. All statistical tests (McNemar, TOST) are specified in §4, and every reported margin, sample size, and discordant-pair count is stated alongside its effect so results can be checked without rerunning the pipeline.

AI Use Statement

Generative AI tools (a large language model coding assistant) were used to help write and debug evaluation scripts, generate figure code, and copy-edit prose drafted by the author. They were not used to design the experiments, choose the selection rules or statistical tests, generate results, or draft

the paper's claims. All code was reviewed and run by the author; all reported numbers come from the author's own experiment runs, not from AI-generated data. The author takes responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Joungbin An and Kristen Grauman. AdaAlloc: Adaptive visual token allocation for long-video question answering. Project page, https://vision.cs.utexas.edu/projects/adaalloc/, 2026. No preprint, PDF, or code publicly available as of 2026-08-21; the project page lists a truncated identifier (arXiv:2605) and no retrievable paper. Cited from the project-page abstract as concurrent unreleased work.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pp. 335–336, New York, NY, USA, 1998. Association for Computing Machinery. doi: 10.1145/290941.291025.

Jingfeng Chen, Jiawen Qian, Wendi Deng, Yinuo Guo, Jiaqi Yu, Sicong Leng, Raghuveer Thirukovalluru, and Bhuwan Dhingra. LDDR: Linear-DPP-based dynamic-resolution frame sampling for video MLLMs. *arXiv preprint arXiv:2605.11477*, 2026a. URL https://arxiv.org/abs/2605.11477.

Wang Chen, Yongdong Luo, Yuhui Zeng, Luojun Lin, Tianyu Xie, Fei Chao, Rongrong Ji, and Xiawu Zheng. Event-anchored frame selection for effective long-video understanding. *arXiv preprint arXiv:2603.00983*, 2026b.

EvolvingLMMs-Lab. LMMs-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Wenhao Hong, Ziyang Wang, Yixin Zhang, and Zilei Wang. MoPrune: Scene-guided motion-aware token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6929–6941, 2026. doi: 10.18653/v1/2026.findings-acl.344. URL https://aclanthology.org/2026.findings-acl.344/. ACL Anthology 2026.findings-acl.344.

Haowen Hou, Zhen Huang, Zheming Liang, Qingyi Si, Chenglin Li, Shuai Dong, Kele Shao, Ruilin Li, Dianyi Wang, Nan Duan, and Jiaqi Wang. AdaCodec: A predictive visual code for video MLLMs. *arXiv preprint arXiv:2606.02569*, 2026.

Yuning Huang and Fengqing Zhu. Adaptive greedy frame selection for long video understanding. *arXiv preprint arXiv:2603.20180*, 2026. URL https://arxiv.org/abs/2603.20180.

Minkuk Kim, Suyong Yun, Young Tae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. ReQuest: Rethinking-based question-aware frame selection for long-form video QA. *arXiv preprint arXiv:2607.01737*, 2026. URL https://arxiv.org/abs/2607.01737.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044.

Zhenyu Li, Zuchao Li, Ping Wang, Lefei Zhang, and Haojun Ai. Vista-LLM: Decoupled query-guided visual token pruning for efficient long-video large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13171–13187, 2026. doi: 10.18653/v1/2026.acl-long.601. URL https://aclanthology.org/2026.acl-long.601/. ACL Anthology 2026.acl-long.601.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers*, 1993.

Jun Peng, Baiyang Song, Jie Li, Hui Li, Yiyi Zhou, Rongrong Ji, and Yonghong Tian. QCA: Query- and content-aware keyframe selection for long video understanding. *arXiv preprint arXiv:2607.00983*, 2026. URL https://arxiv.org/abs/2607.00983.

Haozhe Qi, Kevin Qu, Mahdi Rad, Rui Wang, Alexander Mathis, and Marc Pollefeys. AdaptToken: Entropy-based adaptive token selection for MLLM long video understanding. *arXiv preprint arXiv:2603.28696*, 2026.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP3: A training-free approach for list-wise frame selection in video-LLMs. *arXiv preprint arXiv:2501.02885*, 2025. URL https://arxiv.org/abs/2501.02885.

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

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. LVBench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. Accepted at ICCV 2025.

Yiheng Wang, Lichen Zhu, Yueqian Lin, Yudong Liu, Jingyang Zhang, Hai Li, and Yiran Chen. Query-conditioned evidential keyframe sampling for MLLM-based long-form video understanding. *arXiv preprint arXiv:2604.01002*, 2026a. URL https://arxiv.org/abs/2604.01002.

Yilin Wang, Xiangxi Zheng, Dongxing Mao, Linjie Li, Zhengyuan Yang, Ping Yu, Rui Yan, Yuan Yao, and Alex Jinpeng Wang. Efficient frame selection for long videos at test time with attention-based MLLM selectors. *arXiv preprint arXiv:2607.15689*, 2026b. URL https://arxiv.org/abs/2607.15689.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *European Conference on Computer Vision*, 2024.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-Frame: Query-aware frame selection and multi-resolution adaptation for video-LLMs. *arXiv preprint arXiv:2506.22139*, 2025a. URL https://arxiv.org/abs/2506.22139.

Xian Zhang, Zexi Wu, Zinuo Li, Hongming Xu, Luqi Gong, Farid Boussaid, Naoufel Werghi, and Mohammed Bennamoun. AdaRD-key: Adaptive relevance–diversity keyframe selection. *arXiv preprint arXiv:2510.02778*, 2025b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. *arXiv preprint arXiv:2510.27280*, 2025b. URL https://arxiv.org/abs/2510.27280.

A FOCUS⋆ SCHEDULE REPLAY

Every row in Table 1, including FOCUS⋆, consumes the same dense 1 fps LongCLIP stem-score vector. The original FOCUS method instead treats scoring itself as a budgeted online process, in which clip-arm pulls purchase ITM evidence (Zhu et al., 2025b). Our replay preserves the temporal clip schedule and default hyperparameters but reveals already-computed LongCLIP scores to the bandit. It therefore tests the schedule under a matched scorer, and says nothing about the efficiency or accuracy of the full FOCUS pipeline.

The replay uses 16 s clips, three coarse pulls per clip, α=0.25, a pull budget equal to half the candidate pool, and approximately $k/4$ surviving clips. Unobserved rewards are interpolated from the nearest observed frame, and the random seed is fixed per item. This boundary is why the row carries a star, and why we draw no conclusion from it about FOCUS under its native ITM protocol.

B QUESTION CATEGORIES DO NOT DEFINE A STABLE SELECTOR REGIME

LongVideoBench provides question-category tags, which invite the hypothesis that temporally referred questions need different frame selection from the rest. We test it by comparing OMP with uniform sampling on the temporally referred categories (the T⋆ group) against all remaining categories, without relabeling the data.

Bin	Slice	n	Uniform	OMP	Δ (McNemar)
600 s	temporal T⋆	148	.5608	.5811	$+2.03$, p=.74
600 s	other	264	.5492	.6591	$+10.98$, $p=5.2\times10^{-4}$
3600 s	temporal T⋆	222	.4369	.5270	$+9.01$, p=.013
3600 s	other	342	.4942	.5585	$+6.43$, p=.021

Table 3: OMP minus uniform on official LongVideoBench categories. Percentage-point gains differ by duration; the category interaction is not significant in either long bin (Fisher's exact test on discordant counts, p=.087 at 600 s, p=.73 at 3600 s).

The temporal slice looks resistant to selection at 600 s and responsive at 3600 s, which is the shape of a result that would be easy to over-interpret. It is not one. Fisher's exact test on the discordant counts (OMP-only against uniform-only correct: 20/17 temporal and 48/19 other at 600 s; 40/20 and 53/31 at 3600 s) finds no significant category interaction in either bin (p=.087 and p=.73), and which slice "wins" reverses between them. Many T⋆ stems refer to subtitles or speech while our selector is vision-only and subtitles are disabled, which offers one plausible source of the instability. Two significant-versus-non-significant subgroup tests are not an interaction, so we infer no question-type router from these results.

C SELECTOR VARIANTS

This sweep changes only the subset rule, on the same LongCLIP cache. All variants use the L2-normalized embeddings of Eq. 1. OMP itself is implemented with sequential Gram–Schmidt in float32: each pick is orthogonalized against the stored orthonormal basis, normalized, and removed from the residual query ($q \leftarrow q - (q^\top v)v$, residual not renormalized); a pick whose orthogonal component has norm at most 10^{-6} is kept but adds no direction, chosen frames are masked, and ties go to the lowest index. α-*orth* deflates by $q \leftarrow q - \alpha(q^\top v)v$ (α=1 is OMP, α=0 top-k).

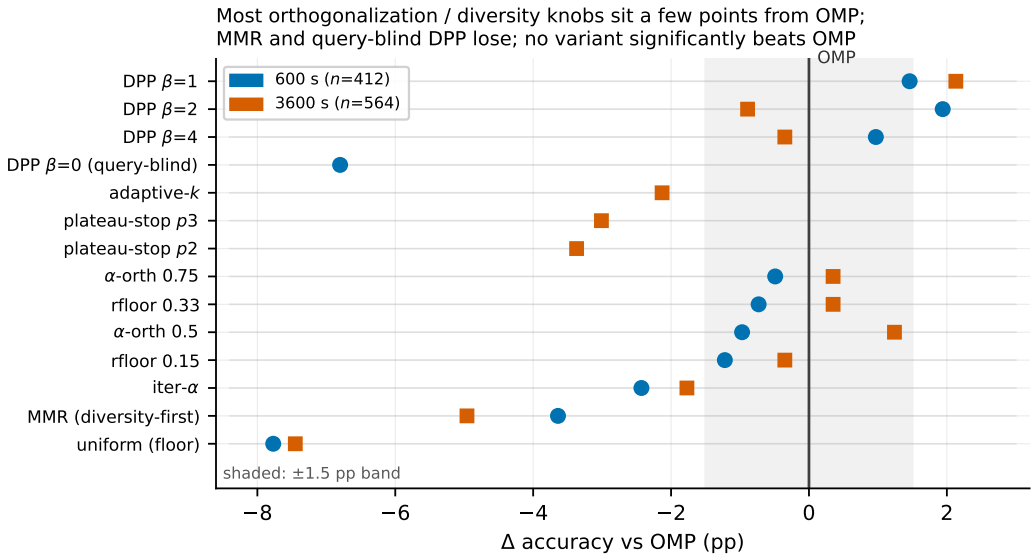


Figure 3: Coverage-gated selector variants against OMP at k=8 on the 600 and 3600 s LongVideoBench bins; rows with one marker were run in one bin. The shaded ±1.5-point band is a visual aid, not a statistical criterion.

iter-α uses $\alpha_r = \mathrm{clip}(g_r/g_1, 0.3, 1)$, with $g_r = e_{b_r}^\top q_{r-1}$ the greedy score. *rfloor f* runs OMP inside the $\max(\lceil fN \rceil, 3k)$ most query-relevant frames. *MMR* keeps OMP's first three picks and adds $\arg\max_i \lambda e_i^\top q_0 - (1-\lambda)\max_{j\in S} e_i^\top e_j$ with λ=0.5. *DPP* is greedy log-determinant MAP (incremental Cholesky) on $L = \mathrm{diag}(r)\, K\, \mathrm{diag}(r)$, $K_{ij} = e_i^\top e_j$, $r_i = \exp(\beta z_i)$ with z_i the per-video standardized query cosine; if the conditional gain reaches 10^{-12} before k picks, the most relevant remaining frames fill the set. *Plateau-stop p* runs OMP up to 16 picks and stops once $\|q_t\|/\|q_0\|$ changes by less than 10^{-3} for p consecutive steps, evaluating the smaller set unpadded (mean k 4.47 and 5.48; 3600 s only). *Adaptive-k* stops before the first pick whose greedy score g_r falls below τ=0.01, keeping between 3 and 8 frames, also unpadded (mean k=5.15); its original generator was not retained, but re-implementing this rule reproduces 97.7% of the stored frame sets (98.3% at 600 s), and the run's arm map confirms τ=0.01 is the plotted setting.

MMR underperforms; query-blind DPP (β=0, 600 s only) is indistinguishable from uniform sampling (+0.97, p=.75) and significantly below OMP (−6.80, p=.0072); and query-weighted DPP, relevance pools, and partial orthogonalization all sit in a band around OMP. No DPP setting significantly exceeds OMP in either long bin, and the best β differs between bins (β=2 at 600 s, β=1 at 3600 s), so the figure shows every setting rather than a per-bin maximum. Both plateau-stop rules are significantly worse than fixed k=8 (p=.040 each).

This sweep ran in the secondary compute environment described in §7; its arms are internally paired against each other and are never compared against Tables 1–2.

These nulls do not show that all diversity objectives are equivalent. They show something more specific and more useful: within this scorer geometry and budget range, moving around inside the relevance–diversity family produces less change than replacing uniform sampling or reallocating pixels into frames. That is consistent with the residual analysis in §E: if the query-directed signal is exhausted after roughly five picks, then rules that differ only in how they trade relevance against diversity have little left to differ about.

D Exploratory failure audit

Two independent model-assisted visual passes inspected OMP failures at k=8 through a frame viewer: 41 cases spanning 16 categories in the 600 s bin, and 52 cases spanning 17 categories in

the 3600 s bin. Cases were chosen to cover categories and to include both OMP-only and shared failures; the sample is not random. Reviewers compared the selected timestamps, the rendered frames, the question, the gold answer, and the model's prediction. Each case received one primary tag (Table 4); the 600 s pass recorded its tallies only approximately.

The most frequent tag in both passes was missed evidence, most often because visually distinctive but irrelevant content consumed several of the eight picks: unrelated chapters in compilations, separate storylines in long narrative video, title and subscribe cards, and dark transitions. Because the sample is failure-conditioned, category-balanced by construction, model-assisted, and scored without blinding, inter-rater agreement, or a matched non-OMP control, we cannot call this pattern dominant, OMP-specific, or prevalent. The opposite failure showed up on sequence-of-scenes and cross-scene tracking questions, where dense top-k sampling covered one relevant moment but omitted another required scene. The audit therefore motivates relevance-constrained diversity as a design direction. It cannot estimate population failure rates, and it cannot establish that adding frames benefits any particular question category.

Primary tag	600 s (n=41, approx.)	3600 s (n=52)
Missed evidence	∼20	35
Not answerable from frames	∼10	4
Answerer error, adequate frames	∼6	6
Temporal offset	∼4	5
Redundant picks	∼2	1
Same frames as top-k, answer flips	–	1

Table 4: Primary-tag counts in the purposive failure audit. The 600 s pass logged approximate counts (summing to about 42); counts describe this non-random sample, not population rates.

E Mechanism: what OMP's later picks add

OMP's residual trace turns out not to behave the way its signal-processing origin predicts. We offer the following as a hypothesis consistent with that discrepancy, not as a demonstrated mechanism: nothing here intervenes on residual geometry, so the account below is observational. On both LongVideoBench long bins, the residual norm is 0.972 of its starting value after a single pick and 0.967 after eight, and does not fall further by sixteen. Almost none of the query is ever reconstructed. What does change is the residual's correlation with the frame being selected, which drops from .233 at pick one to .004 at pick eight and to roughly zero by pick sixteen. The query-directed signal is exhausted early.

Yet those later frames are not junk: they keep a cosine near .20 with the original query. The picture is a text query that sits almost orthogonal to the *dominant directions* of the frame-embedding cone—the modality gap of contrastive image–text encoders (Liang et al., 2022), visible directly in Figure 4b—so with an effectively low-rank frame dictionary, eight greedy picks reach only those dominant directions and OMP cannot reduce the residual much whatever it selects. What it can do is avoid repeating itself. Late picks add little new query direction while adding new chances for the answerer to encounter a relevant moment. That is consistent with the reinvestment result, but the two are co-occurring measurements; we did not test mediation.

Coverage helps only while it stays relevant; Appendix D audits where that breaks and finds off-topic novelty as the dominant failure pattern in an exploratory, failure-conditioned sample.

F Cross-model transfer to InternVL3

Table 5 hands the identical OMP and uniform timestamps to InternVL3, changing the answerer and nothing else. The selection gain replicates on LongVideoBench and LVBench at both model sizes. Video-MME does not cooperate: the 2B model benefits at every duration, while the 8B model benefits only on short videos and is flat on medium and long ones despite receiving demonstrably different frames. Capacity alone does not account for it, since the same 8B model gains 10.14 points

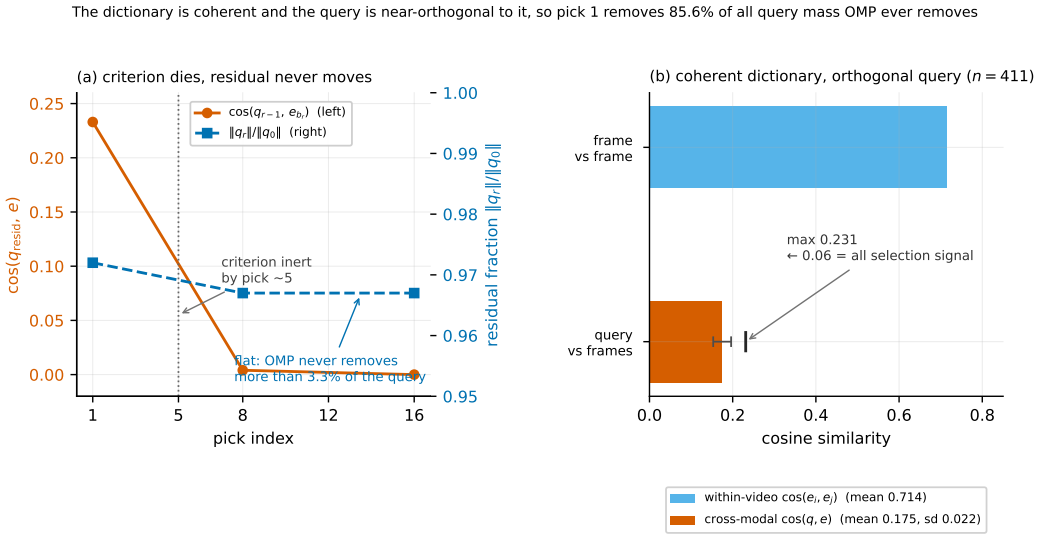


Figure 4: LongCLIP geometry on LongVideoBench-600 s (n=411 of the 412-video bin used elsewhere; one video's LongCLIP embedding was unrecoverable and is excluded here only). (a) Residual correlation vanishes while residual norm changes little; the line connects logged picks 1, 8, and 16 and is not a dense trace. (b) Within-video frames occupy a coherent cone, while the text query is nearly orthogonal to the frame dictionary. In this space, OMP mainly suppresses directions already represented by earlier selections.

Benchmark	n	IVL3-2B Δ	IVL3-8B Δ
Video-MME short	900	+6.00 (p=1.2 × 10^{-4})	+5.44 (p=1.9 × 10^{-4})
Video-MME medium	900	+3.33 (p=.046)	+0.33 (p=.891)
Video-MME long	900	+4.22 (p=.0062)	+0.11 (p=1.000)
LVBench	1549	+9.04 (p=7.2 × 10^{-11})	+10.14 (p=1.2 × 10^{-13})
LongVideoBench	1337	+4.71 (p=1.6 × 10^{-4})	+6.51 (p<10^{-4})

	IVL3-2B		IVL3-8B	
Benchmark	uniform	OMP	uniform	OMP
Video-MME short	.6256	.6856	.6722	.7267
Video-MME medium	.5100	.5433	.5822	.5856
Video-MME long	.4267	.4689	.5033	.5044
LVBench	.3538	.4442	.3680	.4693
LongVideoBench	.5191	.5662	.5550	.6200

Table 5: InternVL3 on identical timestamps and tile budgets. Top: OMP minus uniform accuracy in percentage points. Bottom: absolute accuracies. The flat 8B medium/long Video-MME differences are ties at non-trivial accuracy (95 and 96 discordant pairs each way on long), not a failed run.

on LVBench; a capacity-by-benchmark interaction remains possible and one model family cannot separate the two.

The allocation results transfer more smoothly. On pooled LongVideoBench long videos, InternVL3-8B holds its accuracy through a threefold tile reduction (24 tiles to 8: -0.10 points, p=1.000; 90% CI $[-1.66, +1.45]$, clearing a $\pm$2-point TOST margin at p=.022) and gains 3.38 points when the same total tile budget is spread from 8 to 24 uniformly sampled frames (p=.0099), an effect concentrated in the 600 s bin. Both halves of the select–compress–reinvest story therefore survive a change of answerer, but their magnitudes do not: selection and allocation effects are properties of an answerer–benchmark pair, not constants of a timestamp set.

	Video-MME		LongVideoBench		LVBench	
Rule	pub.	ours	pub.	ours	pub.	ours
uniform (absolute)	57.19	56.37	54.53	56.54	28.08	34.54
AKS†	$+4.29$	$+4.22$	$+5.01$	$+2.62$	$+12.07$	$+8.33$
FOCUS*†	$+1.70$	-0.59	$+5.00$	$+1.65$	$+3.49$	$+5.29$
LDDR-select	$+5.96$	$+5.56$	$+8.90$	$+6.66$	$+14.98$	$+12.39$
OMP	—	$+5.85$	—	$+5.69$	—	$+11.81$

Table 6: Two matched-condition harnesses running the same rules. Published columns are Chen et al. (2026a), Table 1, Qwen3-VL-8B at eight frames with a LongCLIP encoder—the same answerer family, budget, and encoder as ours—where LDDR-select is their stage-1 row. The first line gives absolute uniform-sampling accuracy; all other cells are gains over the uniform row of the same column. † marks rules we run on a substituted scorer (both score with BLIP-ITM natively), and * marks the FOCUS schedule replay of Appendix A. OMP has no published counterpart, and MDP3 and Q-Frame appear in their table but not ours (§7).

G Published-harness comparison and the AKS implementation bug

LDDR is the one prior study that standardises every baseline to a single encoder, and it reports a cell matched to ours on answerer family, frame budget, and encoder. Table 6 places that cell beside our own. Our gains are smaller in eight of the nine comparable cells, by 0.07 to 3.74 points. We do not read this as either harness being wrong. The uniform baselines themselves differ—most starkly on LVBench, 28.08 against our 34.54—so the two harnesses are not measuring from the same floor. A lower floor plausibly leaves more room for a selector to gain, though we did not test that and offer it as a hypothesis rather than an explanation. Prompt template, candidate-pool construction, decoding path, and answer parsing all differ and none of them is specified tightly enough in either paper for a reader to reconcile the two. We also find no clean pattern in the residuals, and in particular no advantage to keeping a rule's native encoder: on Video-MME the closest reproduction is AKS†, within 0.07 points of the published gain despite a substituted scorer, while LDDR-select—the one rule we run on its authors' own encoder—reproduces closest on LongVideoBench (2.24) and LVBench (2.59).

Our first implementation of AKS ended in a padding branch that is absent from the authors' released code. Its recursive partition gives each leaf at depth d a quota of $\lfloor k/2^d \rfloor$ frames, so at $k{=}8$ with the published depth of five every leaf is allotted zero, the recursion returns nothing, and our pad filled the entire selection from global cosine top-k. For every benchmark in an earlier draft of this paper, the AKS row was a duplicate of the top-k row. We caught it because those two rows were the closest pair in Table 1 on all three benchmarks; a pairwise overlap check over selected frame *indices* then showed 98.5% agreement. What is worth reporting is not the bug but how completely aggregate accuracy hid it. Correcting the depth changes roughly 99.5% of the selected frames on every benchmark, yet the aggregate moves by only -1.05 points on LongVideoBench and $+0.07$ on LVBench, while moving $+2.74$ on Video-MME. On LVBench a reader comparing accuracies alone would see two-hundredths of a point between a faithful implementation and one selecting entirely different frames for 99.5% of questions.

Table 7 extends the same check to every pair among the five rules with directly comparable frame-index picks (Top-k, AKS, FOCUS*, OMP, LDDR-select), using mean Jaccard similarity on timestamps rounded to 0.1 s, the same rounding-then-overlap method used above; Uniform is excluded because its frames are sampled live from the raw video by the stock loader and never exported to a picks file, so no comparable index set exists for it.

A single-number comparison cannot detect this class of error, and neither can a per-arm sanity check that only asks whether each row looks individually plausible. What did detect it was comparing arms against *each other*: if two rules that optimise different objectives select nearly the same frames, at least one is not implementing its objective. We now run that overlap check by construction, and we recommend it wherever a table's rows are meant to be distinct methods. The corrected row also reproduces the published Video-MME gain to within 0.07 points against a 2.81-point gap under the

Rule pair	LongVideoBench	Video-MME	LVBench
Top-k / AKS	0.395	0.387	0.348
Top-k / FOCUS*	0.161	0.074	0.043
Top-k / OMP	0.317	0.289	0.229
Top-k / LDDR-select	0.304	0.228	0.157
AKS / FOCUS*	0.131	0.044	0.019
AKS / OMP	0.298	0.270	0.216
AKS / LDDR-select	0.300	0.244	0.176
FOCUS* / OMP	0.133	0.050	0.018
FOCUS* / LDDR-select	0.133	0.046	0.013
OMP / LDDR-select	0.466	0.480	0.454

Table 7: Pairwise frame-overlap check at k=8: mean Jaccard similarity between each pair's selected frame-index sets (timestamps rounded to 0.1s), pooled over all bins of each benchmark. No pair reaches the $\sim$0.99 degenerate overlap the buggy AKS/top-k pair showed; OMP/LDDR-select is the closest pair, consistent with both being query-relevance-plus-diversity rules.

Answerer / benchmark	Full	Compressed	Δ (pp)
Qwen / LongVideoBench	.6223	.6253	$+0.30$
Qwen / Video-MME	.6222	.6178	-0.44
Qwen / LVBench	.4635	.4674	$+0.39$
GPT-5-mini / LVBench	.4900	.4926	$+0.26$ (p=.84)
Qwen / OMP / LVB long pool	.5820	.5892	$+0.72$ (TOST p=.0022 at ± 3)
Uniform / LVB long pool	.5061	.5092	$+0.31$ (p=.784)

Table 8: Fixed-timestamp spatial compression. Qwen rows: saved half-budget arm; GPT row: pixel-ratio control, not a Qwen token count; uniform control: approximately 46% realized per-frame budget. The Qwen/OMP/LVB-long-pool row is the only formally bounded (TOST) contrast, backing the claim in §5.3; the remaining LVBench and Video-MME differences here are descriptive stability checks, not equivalence results.

bug, which is evidence both that the correction is right and that this harness can match a published number when the port is faithful.

The disagreement is itself the measurement. Two carefully controlled harnesses, running the same published rules at the same budget with the same encoder, differ by as little as 0.07 and as much as 3.74 points, a range that covers most of the gaps published selector comparisons are built on. Two harnesses cannot bound the disagreement in general, but they are enough to show that cross-paper selector deltas of this size are fragile, including the numbers this paper's own argument targets, and it is why every claim we make is a paired contrast inside one harness rather than a comparison against a published figure. It also means Table 6 should be read as a measurement of harness disagreement, not as a reproduction verdict on any of these methods.

H Allocation-shape and reinvestment figures

All spatial arms resize each frame before Qwen3-VL's processor with its smart-resize rule (32-pixel grid, $p_{\min}$=200,704 and $p_{\max}$=1,605,632 pixels): a frame with fraction $s_r<1$ is re-sized under the cap $\max(p_{\min}, \lfloor s_r hw \rfloor)$, where $h \times w$ is its full-resolution size, preserving aspect ratio. The residual-proportional schedule D sets $w_r = \max(0, g_r)/g_1$ from OMP's greedy score and $s_r = \mathrm{clip}(F + (1-F)w_r^\gamma, F, 1)$, with floor $F = p_{\min}/(1196 \cdot 1024) \approx 0.164$ and γ found by bisection so that the mean s_r over all frames in the bin equals the target (0.53 for D@53, 0.50 for the k=16 reinvestment arm), giving γ=0.329 and 0.365 for D@53 and 0.154 and 0.204 for the reinvestment arm in the 600 and 3600 s LongVideoBench bins (recovered from the stored per-frame fractions, which agree on a single value in every bin). The flat split sets s_r=0.5 for every frame. E1 (halve-on-drop) starts at s_1=1 and halves s_r, floored at F, whenever $w_{r-1} - w_r \geq 0.005$, landing near 32% of the full-resolution budget.

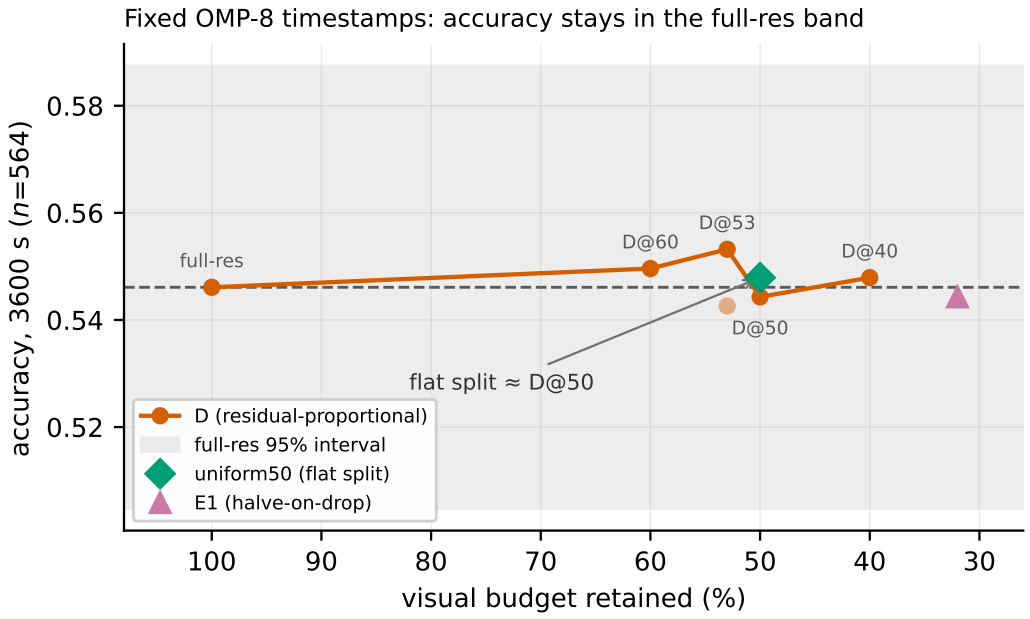


Figure 5: LongVideoBench 3600 s accuracy on fixed OMP-8 timestamps across spatial budgets. The flat 50% control and the residual-proportional schedule are indistinguishable, so OMP rank alone does not provide a useful resolution policy. E1 is a separate halve-on-drop schedule, not a point on the D curve.

The residual-proportional schedule does not beat a flat split at the same mean budget (Figure 5), so OMP's ranking, useful as it is for choosing frames, did not improve accuracy detectably when used to distribute pixels among them; we ran no equivalence test on that contrast, so this is a failure to detect a difference rather than evidence of none. A reconstruction of LDDR's stage-2 Group-DPP importance, applied to identical OMP-8 timestamps, does better, improving on the flat split by 1.84 points on the pooled long bins (p=.0198, 95% CI $[+0.37, +3.32]$). That contrast comes from our own same-environment reconstruction rather than the authors' released implementation, and its 3600 s component is not individually significant, so we treat it as indicative.

To test whether reinvestment requires OMP, we ran all four cells of the selector-by-budget design—uniform and OMP, each at eight full-resolution and sixteen compressed frames—within a single environment, so the interaction is measured without any cross-environment comparison; Figure 6 plots the OMP pair, with the uniform-sampling pair given numerically below. Under uniform sampling the eight-to-sixteen intervention gives $+1.13$ points (p=.3673); under OMP it gives $+2.56$. The interaction between them is 1.43 points and does not reach significance (p=.2614), with OMP gaining on 110 items where uniform does not and uniform gaining on 93 where OMP does not; with 203 discordant pairs the test excludes only large effects, so this is evidence against a *large* OMP-specific interaction rather than proof the two are identical.

I Prompt boundary: fused-query robustness check

Answer options never enter selection in the main results. That is a convention rather than a neutral choice, so we measure what it costs. Scoring against the fused question-and-options string instead of the stem alone changes 41.9% of the frames top-k selects on LongVideoBench-600 s and 53.3% of those OMP selects; OMP is the more sensitive of the two because each pick is chosen to explain a residual component of the query vector, so a perturbation of that vector compounds through the greedy chain. Those are also better frames, though not because of the options themselves. On that bin, with the fused query, OMP reaches .6699 against .6311 for the stem ($+3.88$ points, p=.033, 33 rescued against 17 broken) and top-k reaches .6408 against .6068 ($+3.40$, p=.087). A pre-registered replication on the 3600 s bin gives the same picture: OMP reaches .5691 against .5461 for the stem

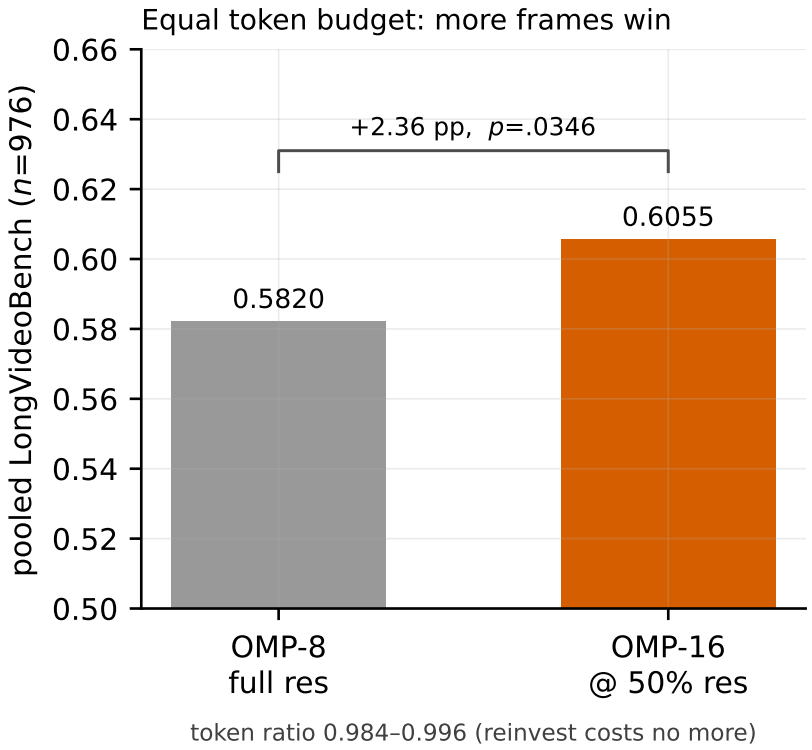


Figure 6: LongVideoBench 600 s and 3600 s pooled: sixteen compressed OMP frames outperform eight full-resolution OMP frames at a measured token ratio no greater than one.

(+2.30 points, n=564, 46 rescued against 33 broken, p=.18), with the fused query moving roughly 69% of the selected frames. Neither long bin resolves the shift on its own at these sample sizes, but the sign and magnitude agree across two disjoint video sets, so we read the prompt boundary as a level shift affecting long videos generally rather than a property of the 600 s bin. A control arm isolates what the answer options contribute. Replacing each item's options with those of a different video—unrelated in content, matched on length, and perturbing 54% of OMP's picks against the real options' 53%—reaches .6529, recovering 56% of the gain; the item's own options add a further 1.70 points beyond that, which is not significant (p=.42). Neither the control against the stem (+2.18, p=.23) nor the fused query against the control separates at this sample size, so we do not claim a mechanism. What the control indicates, without ruling out a smaller effect the test is underpowered to detect, is that most of the gain does not require the answer set to be semantically relevant: a scorer query built from the wrong options recovers 56% of it, moving accuracy nearly as far as one built from the right ones. The consequence for reading the literature is the substantive one: the text handed to the scorer is an uncontrolled axis worth roughly two to four accuracy points across the two long bins—and, on the evidence above, one that does not require the substituted text to be relevant—so a comparison inconsistent about it is measuring the prompt rather than the selector.

J Scorer-swap table, power bound, and the interaction-test caveat

Neither selector's gain in Table 9 depends detectably on which scorer produced it, but with roughly 134 discordant pairs the 90% interval on the OMP scorer effect spans about $[-3.6, +5.6]$ points, so this bounds large scorer effects rather than establishing their absence. The 60 s bin adds nothing either way (largest contrast +3.49, p=.31), expected when eight frames already cover 22% of a median 36-frame pool.

Selector	LongCLIP	SigLIP	SigLIP − LongCLIP
Uniform	.5534		—
Top-k	.6068	.6068	$+0.00$ (p=1.00)
OMP	.6311	.6408	$+0.97$ (p=.73)

Table 9: Scorer swap on LongVideoBench-600 s (n=412), all arms run together and paired on identical items. Uniform sampling is scorer-independent and appears once. The final column is a paired test of whether each selector's gain depends on the scorer.

One reading must be resisted: OMP beats top-k significantly under SigLIP ($+3.40$, p=.049) but not under LongCLIP ($+2.43$, p=.33), which invites the conclusion that OMP suits SigLIP better. The interaction test finds no evidence for that: the two gains differ by 0.97 points (p=.68)—and the two per-scorer p-values differ only because they are different tests over different per-item patterns; a significant result beside a non-significant one is not evidence that the two differ (Appendix B records the same error on question categories).

The claim this supports is bounded: one alternative scorer on two duration bins does not establish scorer-invariance in general, and a scorer built on a different principle—an ITM head, or MLLM attention—could still reorder the table.