QMA-RAG: QUERY-CONDITIONED MULTI-GRANULARITY CONTEXT CONSTRUCTION FOR LONG-VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-granularity retrieval offers local observations and broader context for long-video question answering, but several retrieved nodes may refer to the same source frames. Constructing a limited visual context therefore requires deciding how these semantic views jointly affect which observations are selected. We introduce QMA-RAG, which combines multi-granularity support on shared source clips before selecting the visual context. A reusable graph links atomic, event, and macro descriptions to their supporting clips. Question-conditioned retrieval supplies candidates whose scores are fused on these clips; selection balances the resulting support with temporal needs and repeated coverage. Selected clips are converted into a bounded set of unique frames with original temporal information and retrieved structured support. Across MLVU, Video-MME, and LongVideoBench with three answering backbones, the complete system improves Uniform inference in all nine settings by 3.80 percentage points on average. On MLVU, construction from identical candidates improves over exact frame deduplication by 1.66 points. Removing cross-level fusion while retaining the full selector reduces accuracy by 1.93 points, supporting the role of aggregated semantic support in source selection. Answering-input and component controls examine how the resulting context and its supporting design choices contribute to answering.

1 INTRODUCTION

Long-video question answering must recover useful observations within a limited model context. Compression reduces the visual sequence, while retrieval selects relevant content (Song et al., 2024; Shen et al., 2025a;b). Yet relevance alone does not specify a sufficient answering context. Brief interactions require local detail, ordering distant activities needs multiple temporal anchors, and understanding a video's theme requires distributed coverage. Evidence must match both the question and the temporal arrangement of supporting observations (Figure 1).

These different evidence needs motivate retrieving multiple semantic views: a brief observation, its surrounding activity, and the broader scene. Such views can be complementary yet refer to overlapping source content. An event description and an atomic match may support the same clip without providing two distinct visual observations. This creates a *retrieval-to-context gap*: the semantic unit retrieved need not be the unit charged to the visual budget. Figure 1 links this gap to the need for question-relevant evidence.

Hierarchical representations and query-adaptive retrieval provide selective access to long videos (Wang et al., 2025; Shen et al., 2025b; Xue et al., 2025). Given their overlapping semantic evidence, the remaining decision is which source observations to include in the answering context. Removing duplicate frames enforces uniqueness, but does not specify how local matches and broader contextual support should jointly influence that decision.

We introduce **QMA-RAG**, which combines multi-granularity support on shared source clips *before selecting the visual context*. A reusable graph preserves the source clips behind atomic, event, and macro views. After question-conditioned retrieval, these mappings bring the candidates into a common clip space. Cross-level fusion determines the support for each clip, and selection balances that support against repeated coverage and the question's temporal needs. The selected clips 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.



Figure 1: **Motivation of QMA-RAG.** Left: fixed acquisition mismatches query-specific evidence needs. Right: semantic overlap repeats source frames. Center: query-conditioned acquisition and source alignment construct a bounded context. Video frames are real; answer contrasts and the $B = 4$ budget are illustrative.

then converted into unique frames under the visual budget. Original timestamps, interval gaps, and retrieved structured support accompany those frames for answering.

Our contributions are:

- **Source-aligned multi-granularity context construction.** We connect complementary semantic views to a common source-clip selection process. Shared mappings retain their provenance, fusion combines their support before selection, and frame construction realizes the selected evidence under a unique-frame budget.
- **Mechanism-level validation and cross-model evaluation.** With identical candidates, QMA-RAG improves exact deduplication by 1.66 points on MLVU; removing fusion with the selector retained costs 1.93 points. Answering-input, component, and planner controls examine the contribution and scope of the design. Complete-system gains average 3.80 points over Uniform across three benchmarks and three answering backbones.

2 Related Work

Long-video representations and context construction. Memory and compression extend the video context available to answering models (Song et al., 2024; Ren et al., 2024; Shen et al., 2025a); complementary work scales dense input (Chen et al., 2025), maintains compressed online state (He et al., 2024), or uses instruction-conditioned memory (Diko et al., 2025). Coherent clip selection and hierarchical expansion preserve temporal structure under input limits (Sun et al., 2025; Wang et al., 2025). QMA-RAG links semantic views at three granularities to their original source clips, allowing shared support to be combined before constructing the answering context.

Retrieval and query-adaptive visual access. Graph-based systems retrieve clips and structured context for downstream reasoning (Shen et al., 2025b; Cao et al., 2026). Other methods adapt retrieval, visual tokens, sampling, resolution, or tools to the query (Xue et al., 2025; Shi et al., 2025; Jeoung et al., 2024; Lin et al., 2026; Wang et al., 2026; 2024). AKS balances relevance and coverage (Tang et al., 2025); DIG routes global and localized queries to different selectors (Li et al., 2026); A.I.R. combines adaptive sampling with iterative VLM-based frame analysis (Zou et al., 2026). Our focus is how overlapping semantic evidence is combined before source selection under a visual budget. QMA-RAG uses common source mappings to fuse cross-level support, control repeated coverage, and connect the selected frames to temporal and structured answering support.

3 Method

3.1 Problem Setup and Overview

Given a video V, question q, and answer options $\mathcal{O}$, we construct an answering context with at most B source frames. A **source frame** $V[i]$ is a sampled observation with its original timestamp; a **source clip** c groups consecutive sampled frames. An **evidence node** describes a clip, event, or longer region and retains links to its supporting clips. Nodes are the semantic retrieval units; source clips are the common units on which their support is combined and selected. The resulting context contains unique frames $\mathcal{F}_q$ and accompanying support $\mathcal{S}_q$: original timestamps, interval gaps, and available retrieval-derived structured evidence. The answering interface is

$$\hat{y} = f_{\mathrm{VLM}}(q, \mathcal{O}, \mathcal{F}_q, \mathcal{S}_q), \qquad |\mathcal{F}_q| \le B. \tag{1}$$

The central operation is to combine support *before source selection* (Figure 2). Multi-granularity memory supplies semantic views with shared source mappings. Retrieved views are projected onto their supporting clips, so local and contextual relevance jointly affect which clips enter the context. Frame construction then enforces the budget on unique source indices. Question conditioning specifies the evidence needs used by this process; it does not change the shared source mappings.

3.2 Multi-Granularity Evidence with Shared Source Mappings

Memory is built once from video-associated information, including available subtitles, without questions or answer options. The answering stream is sampled at 1 fps and partitioned into atomic clips of up to 64 frames. **Atomic nodes** retain local observations. **Event nodes** describe contiguous activity, potentially spanning atomic boundaries. **Macro nodes** aggregate adjacent events into broader activity and scene context. Relations within levels and membership links across levels organize these views into a reusable graph.

For each node n, a source mapping σ records its supporting clips $\mathcal{C}(n)$ and their original frame indices. An atomic node maps to its own clip; an event maps to every intersecting atomic clip; a macro node inherits the union of its supporting clips. Event boundaries are estimated separately, so the levels need not form disjoint temporal partitions. The mappings distinguish several descriptions of one source from evidence at different moments. Boundary extraction and event aggregation are given in Appendix B.3.

Question-conditioned candidates. The planner supplies a retrieval configuration π_q: per-level candidate budgets, a temporal operator, and a local/global/hybrid coverage mode. The evaluated implementation uses fixed deterministic language rules on the question and options; Section 4.5

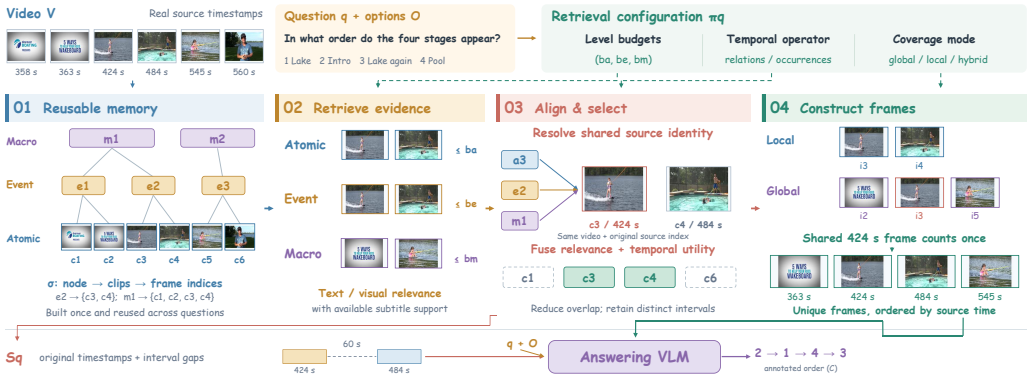

Figure 2: **From multi-granularity support to source-frame context.** Memory preserves the source clips behind semantic nodes. After question-conditioned retrieval, shared mappings allow cross-level support to be fused before clip selection. Frame construction converts selected support into at most B unique frames with original temporal information. The ordering example is illustrative.

tests an LLM replacement under question-only acquisition. Atomic retrieval uses text and visual similarity with attribute and available subtitle support. Event retrieval combines pooled context with strong child matches; macro retrieval uses broader semantic support. Each candidate carries a relevance score and its source mapping. Thus the following selection stage can use both a localized match and the activity or scene that gives it context.

3.3 Fusing Support Before Source Selection

Combining evidence for a common clip. Retrieved nodes are projected onto their supporting clips through σ. For each candidate clip, fusion combines atomic relevance, the strongest supporting event and macro scores, and attribute and subtitle evidence. Parent proposals are first downweighted according to overlap between their supporting clip sets, limiting repeated support from nearly identical regions. This yields a clip score that accounts for the semantic views referring to that clip, rather than treating each view as a separate visual candidate.

In Figure 2, atomic, event, and macro evidence refer to the clip at 424 s. Their combined support affects its rank relative to a clip from another moment, while its frames still occupy one set of source indices. Fusion can therefore change *which content is selected*. Final frame deduplication ensures uniqueness after that decision. Section 5.1 compares these construction choices using identical candidates and separately removes fusion with the selector retained.

An event or macro region may support several clips. Its score contributes through the recorded mappings, while clip selection determines which observations receive local coverage. Thus broad semantic context can guide selection without inserting every frame it describes into the answer.

Selecting complementary source clips. Starting from an empty set, selection greedily adds clips that balance fused relevance and temporal support against overlap with selected intervals and repeated use of the same supporting parent. The question's temporal operator determines how dispersed support is organized: occurrence grouping separates distinct repetitions from multiple observations of one occurrence, while action-wise localization retains separate intervals for actions in an ordering relation. These operations preserve relevant moments when several high-scoring candidates cover the same activity. Scores, coefficients, and temporal settings are detailed in Appendix B.3. The candidate cap limits selection; the unique-frame budget is enforced next.

3.4 Constructing the Bounded Answering Context

Local coverage uses frames from selected clips, global coverage spreads samples across the video, and hybrid coverage combines both, as specified by π_q. We merge source indices, remove duplicates, enforce B, and order retained frames by original time. If the union exceeds the budget, local indices are kept first and remaining capacity is filled with spread-out global-only samples. Empty local support falls back to global coverage. A frame supported by several nodes or sampling paths consequently occupies only one position in $\mathcal{F}_q$.

The context retains original time alongside visual content. Observations at 424 s and 484 s preserve their 60 s gap even if adjacent in the packed input. The accompanying $\mathcal{S}_q$ contains timestamps, interval gaps, and available retrieval-derived structured support. The main answering configuration enables verification. Section 5.2 holds selected frames fixed to measure the incremental effects of temporal information, structured support, and verification on answering.

4 Experiments

4.1 Experimental Setup

We evaluate on MLVU (Zhou et al., 2025) (2,174 questions), Video-MME (Fu et al., 2025) (2,700), and LongVideoBench (Wu et al., 2024) (1,337). We report question-level accuracy over each complete evaluated set, with failed or unparseable predictions counted as incorrect. At inference, benchmark category labels are not model inputs. The answering backbones are Qwen2.5-VL-7B, LongVU-7B (Shen et al., 2025a), and InternVL3.5-8B.

Rerun protocol and comparison scope. All reported baseline accuracies are obtained from our own runs on the evaluated sets, rather than copied from the cited papers. Except for the explicit frame-budget sweep, Uniform, Vgent, AKS, DIG, and QMA-RAG share a maximum of 320 frames in the final answering input. The answering backbone is matched within each comparison, while actual frame counts and auxiliary computation follow each method's pipeline. Uniform answers from uniformly sampled frames without graph retrieval or retrieval-derived structured evidence. Vgent (Shen et al., 2025b) is the graph-based comparator; AKS (Tang et al., 2025) and DIG (Li et al., 2026) are non-graph selectors evaluated with Qwen2.5-VL-7B. Each retains its evidence-selection pipeline under the shared cap. These are complete-system comparisons, including each method's evidence acquisition and answering stages.

QMA-RAG uses the memory and context-construction procedure in Section 3. Its memory is constructed without questions or answer options. Unless stated otherwise, the source stream is sampled at 1 fps into 64-frame atomic clips; acquisition retains at most 20 candidates and selects at most $B = 320$ frames from up to five local clips. Policy and component studies use Qwen2.5-VL-7B on MLVU; additional acquisition controls cover all 1,337 LongVideoBench validation questions with the same backbone. Main QMA-RAG accuracy and cost use verification together with selected frames, temporal information, and retrieved structured support. The answering-input controls in Section 5.2 separate their effects. Appendices A–C provide task results and the earlier verifier-off profiling.

4.2 Augmenting Answering Backbones

QMA-RAG improves Uniform inference in all nine settings by 2.63–5.20 points (3.80 on average; Table 1). MLVU gains are consistent across Qwen2.5-VL-7B, LongVU, and InternVL3.5 (+5.11, +4.92, and +5.20), and every Video-MME and LVB pairing also improves. These results support question-conditioned context construction across answering backbones and benchmark distributions.

4.3 Comparison with Graph-Based RAG

QMA-RAG exceeds Vgent in all nine settings by 1.60 points on average (Table 1), ranging from +0.33 on LongVU/Video-MME to +2.58 on LongVU/MLVU. LVB gains range from +1.12 to +2.39.

Table 1: Accuracy (%) under matched answering backbones. LVB denotes LongVideoBench. Uniform denotes uniform-frame backbone inference. Each Δ cell reports two gains: blue versus Uniform / orange versus Vgent, in percentage points computed from displayed accuracies. Bold marks the best accuracy per group.

Backbone	System / comparison	MLVU	Video-MME	LVB
Qwen2.5-VL-7B	Uniform	69.64	71.11	61.85
	Vgent	72.68	73.33	62.45
	QMA-RAG	**74.75**	**74.41**	**64.70**
	Δ	+5.11 / +2.07	+3.30 / +1.08	+2.85 / +2.25
LongVU-7B	Uniform	65.55	59.63	50.41
	Vgent	67.89	63.15	51.31
	QMA-RAG	**70.47**	**63.48**	**53.70**
	Δ	+4.92 / +2.58	+3.85 / +0.33	+3.29 / +2.39
InternVL3.5-8B	Uniform	69.50	66.52	61.11
	Vgent	73.14	68.11	63.05
	QMA-RAG	**74.70**	**69.15**	**64.17**
	Δ	+5.20 / +1.56	+2.63 / +1.04	+3.06 / +1.12

Table 2: Non-graph frame-selection comparison with Qwen2.5-VL-7B (accuracy, %). Methods retain their native pipelines; bold marks the best result.

Method	MLVU	Video-MME	LVB
AKS	72.22	71.67	60.96
DIG	72.72	72.41	64.40
QMA-RAG	**74.75**	**74.41**	**64.70**

These are complete-system results. Appendix A gives category trade-offs. Section 5.1 tests source-aligned construction on identical candidates; the remaining analyses examine answering inputs and the evidence organization supporting this mechanism.

4.4 Comparison with Non-Graph Selection Systems

AKS balances prompt relevance and coverage; DIG routes global and localized queries to different selectors. Neither constructs a persistent video graph. Table 2 compares complete systems under the same Qwen2.5-VL-7B answering model.

Under the shared final-frame cap, the complete QMA-RAG answering context exceeds DIG by 2.03 points on MLVU and 2.00 on Video-MME. It improves over AKS by 2.74 points on Video-MME. LVB's 0.30-point margin over DIG is small, and task-level trade-offs remain (Appendix A.3). These comparisons include each system's accompanying evidence and answering procedures. Section 5.2 distinguishes the effect of QMA's selected frames from the gains of temporal information, structured support, and verification.

4.5 Question-Conditioned Acquisition Controls

Table 3 summarizes LVB controls with Qwen2.5-VL-7B and a 320-frame cap. Fixed policies retain graph retrieval and jointly set the retrieval recipe and frame coverage. Appendix B gives their definitions and complete task-wise results.

Table 3: LVB acquisition controls (overall accuracy, %). All 1,337 questions remain in the denominator. Simple-Q uses question-only global/temporal/local routing with hybrid fallback.

Configuration	LVB Overall (%)
Fixed global	62.00
Fixed local	62.75
Fixed hybrid	62.00
Simple-Q	59.76
Router-Q	63.20
Rule-Q	63.65
LLM-Q	63.05
Full QMA-RAG	**64.70**

Question-dependent and fixed policies. QMA-RAG exceeds the strongest fixed policy by 1.95 points on LVB and by 2.81 on MLVU (Appendix B.1). The MLVU-developed configuration is held fixed across the reported LVB controls. Fixed global retains graph retrieval and can supply structured support despite using uniform frames; Uniform bypasses this retrieval path. The fixed-policy controls therefore assess complete retrieval-and-coverage recipes within the system.

Acquisition inputs and planner replacement. Router-Q removes options only from routing, whereas Rule-Q removes them throughout acquisition. Rule-Q obtains 851/1,337 (63.65%), above the separately evaluated Uniform baseline (61.85%). Replacing only its planner with Qwen2.5-VL-7B gives 843/1,337 (63.05%). Both use the same memory, fusion, selection, and frame-construction procedures, supporting compatibility with an alternative planner. Simple-Q also changes policy design and reaches 59.76%. Complete task results appear in Tables 12 and 13.

4.6 Accuracy and Computational Cost

Table 4 pairs main-system accuracy with offline indexing and online answering costs, both in seconds per video minute. Online time is normalized by the duration of the queried video and includes verification for QMA-RAG. Uniform has no offline index; the graph-based methods construct reusable memory. Appendix C retains the earlier verifier-off profile and its stage breakdown.

Table 4: Accuracy and cost with Qwen2.5-VL-7B on MLVU. Both cost columns are in s/video-minute. Online cost includes verification for Full QMA-RAG. Uniform requires no offline index.

Method	Accuracy (%) ↑	Offline ↓ (s/video-minute)	Online ↓ (s/video-minute)
Uniform	69.64	N/A	3.80
Vgent	72.68	24.14	5.06
QMA-RAG	**74.75**	22.12	4.50

QMA-RAG improves accuracy by 2.07 points over Vgent while reducing normalized online cost by 11.1% (5.06 to 4.50) and offline cost from 24.14 to 22.12. Relative to Uniform, the 5.11-point accuracy gain comes with 18.4% more normalized online time and a one-time indexing cost. Memory reuse avoids rebuilding the index for each question; retrieval and evidence preparation remain online.

5 Analysis

We analyze QMA-RAG with Qwen2.5-VL-7B on all 2,174 MLVU questions. We first test whether combining cross-level support before selection improves context construction, then separate the contributions of answering inputs and examine the supporting retrieval and temporal design choices.

5.1 Cross-Level Support Before Selection

Table 5 fixes retrieval candidates, memory, planner output, answering backbone, accompanying text, and verifier setting. Exact deduplication continues through candidates until the budget or candidates

are exhausted. The no-fusion control retains candidates from all levels, source mappings, the full selection mechanism, and frame construction, while disabling cross-level support aggregation.

Table 5: Context construction on identical retrieved candidates, with matched answering settings and a 320-frame cap (MLVU overall accuracy, %).

Context construction	Overall (%)
Concatenation / budget truncation	71.80
Exact source-frame deduplication	73.09
No cross-level fusion + full selection	72.82
Full source-aligned fusion and selection	**74.75**

Exact deduplication improves accuracy from 71.80% to 73.09%; full fusion and selection adds 1.66 points to reach 74.75%. Removing fusion while retaining the full selector reduces accuracy to 72.82%, a loss of 1.93 points and 0.27 points below exact deduplication. Retaining the selector alone therefore does not recover the complete method's performance. All retrieval levels remain available in this control: it tests how their support informs source selection. The branch ablations below separately test the value of having those complementary views available.

5.2 Answering Context and Verification

The input control fixes QMA-selected frames and progressively adds original timestamps and interval gaps, retrieved structured support, and verification. Uniform frames provide a separate visual baseline. Verification is disabled for the first three QMA rows and enabled for Full.

Table 6: Answering-input controls on MLVU with Qwen2.5-VL-7B. QMA rows progressively add answering information to the same selected frames (accuracy, %).

Answering input	Accuracy (%)
Uniform frames only	69.64
QMA-selected frames only	72.21
+ timestamps / interval gaps	72.68
+ retrieved structured support	74.10
+ verifier (Full)	**74.75**

QMA-selected frames alone reach 72.21%, 2.57 points above Uniform. Temporal information adds 0.47 points; structured support adds a further 1.42, reaching 74.10% without verification. The verifier adds 0.65 points to reach 74.75%. Thus visual selection improves Uniform, while temporal and structured evidence further strengthen the answering context. These results attribute the full-system gains to their combined use.

5.3 Component Ablation

Table 7: Complete MLVU component ablations across all tasks (accuracy, %). All 2,174 questions remain in each denominator.

Configuration	Count	Needle	Order	Ego	PlotQA	Anomaly	Topic	Overall
Full QMA-RAG	58.74	81.69	75.29	60.23	77.74	77.50	88.59	**74.75**
w/o atomic clip branch	38.35	79.44	72.97	58.52	76.44	74.50	85.93	70.98
w/o event/macro branch	41.26	80.56	76.83	57.39	75.88	75.50	86.31	71.71
w/o occurrence grouping	42.72	80.00	75.29	60.23	77.74	74.50	88.59	72.68
w/o overlap penalty	42.72	80.28	76.45	60.51	77.55	75.50	87.83	72.86
w/o action-wise localization	58.74	80.00	72.20	60.23	77.74	74.50	88.59	73.83
w/o query router	43.20	79.72	64.48	59.66	74.58	75.00	88.21	70.52

Complementary evidence granularities. Removing atomic retrieval reduces overall accuracy from 74.75% to 70.98%; removing event/macro retrieval reduces it to 71.71% (Table 7). Both losses exceed three points, supporting the use of local observations together with aggregated context. Count is especially affected, dropping by 20.39 and 17.48 points: resolving repeated activity benefits from

access to both individual observations and their event context. Atomic access improves every reported task relative to its removal. Removing aggregated evidence raises Order by 1.54 points but lowers the other six tasks, showing an aggregate benefit with task-dependent trade-offs.

Source overlap and temporal organization. Removing overlap control costs 1.89 points overall and 16.02 on Count. The effect is concentrated on repeated activity, where redundant support can leave distinct occurrences unrepresented. Removing occurrence grouping also lowers Count to 42.72% and costs 2.07 points overall. These operations respectively limit redundant coverage and organize repeated activity. Action-wise localization has a smaller overall effect (0.92 points), with its clearest losses on Order (3.09) and Anomaly (3.00), consistent with the need to retain specific temporal support. Overlap removal improves Order by 1.16 points, showing that suppressing redundancy can also discard useful nearby context; the full setting balances that trade-off across questions.

Conditioning the evidence budget on the question. Removing query routing gives the largest aggregate loss, 4.23 points, including 15.54 on Count and 10.81 on Order but only 0.38 on Topic. This pattern associates query conditioning with dispersed occurrences and temporal relations. With routing retained, removing either granularity branch or overlap control also reduces accuracy. The component results show how question-conditioned candidates and complementary semantic views support the source-selection mechanism tested in Section 5.1.

5.4 Budget and Parameter Sensitivity

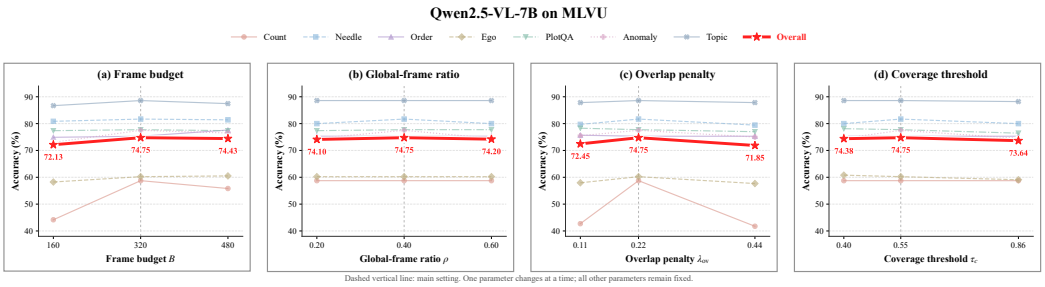

Figure 3: MLVU parameter sensitivity with Qwen2.5-VL-7B across all seven tasks and Overall (red). Each sweep varies one parameter while fixing the others. Dashed vertical lines mark the main configuration: $B = 320$, $\rho = 0.40$, $\lambda_{\text{ov}} = 0.22$, and $\tau_c = 0.55$.

Accuracy rises from 72.13% at 160 frames to 74.75% at 320 and remains close at 74.43% with 480 frames (Figure 3). At 480, Order gains 2.32 points while Count and Anomaly lose 2.92 and 1.00, so added capacity redistributes rather than improves aggregate performance. Both ρ alternatives remain within 0.65 points of the main setting.

Both weaker and stronger overlap penalties reduce accuracy. Changing τ_c from 0.55 to 0.40 redirects 83 fallback allocations and yields 74.38%, while 0.86 yields 73.64%. The sweeps favor a moderate context and overlap penalty, with limited sensitivity to the remaining thresholds.

6 Conclusion

QMA-RAG combines multi-granularity support on shared source clips before selecting the visual context. Its reusable memory preserves the origins of complementary semantic views, enabling fusion and selection to operate on the same clips that supply the final frames. Matched-candidate controls show gains over exact deduplication and a contribution from fusion with the selector held fixed. Answering-input controls distinguish visual selection from temporal information, structured support, and verification. Across three benchmarks and three backbones, the complete system improves Uniform and is competitive with graph and non-graph alternatives. These results support using semantic context to guide source selection while accounting for unique observations in the final visual budget.

AI USE STATEMENT

We used generative AI tools to assist with drafting and revising manuscript text, improving English expression, and refining the organization and clarity of the paper. We also used these tools for literature search and feedback on methodological descriptions, experimental design, and the interpretation of results. AI-generated suggestions were treated as assistance rather than authoritative evidence. The authors are responsible for reviewing AI-assisted content, verifying references and technical claims, and checking that reported results are supported by the experimental records. We take full responsibility for the final manuscript, including all text, claims, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Sections 3 and 4.1 describe the QMA-RAG pipeline, evaluated benchmarks, answering backbones, and metrics. Complete task results and fixed-policy controls are documented in Appendices A–B; Section 5 reports matched-candidate, answering-input, component, and parameter analyses. Section 4.6 gives accuracy and cost comparisons; Appendix C retains the detailed verifier-off online profile. All reported evaluations retain failed predictions in the denominator.

REFERENCES

Zongsheng Cao, Anran Liu, Yangfan He, Jing Li, Bo Zhang, and Zigan Wang. ViG-RAG: Video-aware graph retrieval-augmented generation via temporal and semantic hybrid reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 48–56, 2026. doi: 10.1609/aaai.v40i1.36963.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. LongVILA: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. ReWind: Understanding long videos with instructed learnable memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13734–13743, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514, 2024.

Sullam Jeoung, Goeric Huybrechts, Bhavana Ganesh, Aram Galstyan, and Sravan Bodapati. Adaptive video understanding agent: Enhancing efficiency with dynamic frame sampling and feedback-driven reasoning. *arXiv preprint arXiv:2410.20252*, 2024.

Jialuo Li, Bin Li, Jiahao Li, and Yan Lu. Divide, then ground: Adapting frame selection to query types for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11369–11380, 2026.

Kuanwei Lin, Wenhao Zhang, and Ge Li. VideoRouter: Query-adaptive dual routing for efficient long-video understanding. *arXiv preprint arXiv:2605.05848*, 2026.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, 2024.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599, 2025a.

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. In *Advances in Neural Information Processing Systems*, 2025b.

Yumeng Shi, Quanyu Long, and Wenya Wang. Static or dynamic: Towards query-adaptive token selection for video question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10759–10771. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.545.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232, 2024.

Guangyu Sun, Archit Singhal, Burak Uzkent, Mubarak Shah, Chen Chen, and Garin Kessler. From frames to clips: Efficient key clip selection for long-form video understanding. *arXiv preprint arXiv:2510.02262*, 2025.

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

Tianyue Wang, Xuying Wu, Yuxiang Ma, Ruiming Liang, Jiaxuan Kang, Yanchao Hao, Zheng Wei, Leigang Qu, Haiyun Guo, and Jinqiao Wang. Routing before looking: Query-adaptive evidence acquisition for long-form video understanding. *arXiv preprint arXiv:2608.20805*, 2026.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision (ECCV)*, pp. 58–76, 2024.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3283, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Zhucun Xue, Jiangning Zhang, Xurong Xie, Yuxuan Cai, Yong Liu, Xiangtai Li, and Dacheng Tao. AdaVideoRAG: Omni-contextual adaptive retrieval-augmented efficient long video understanding. In *Advances in Neural Information Processing Systems*, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, 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://proceedings.iclr.cc/paper_files/paper/2026/hash/1332893b662f655660c9abdf793230cf-Abstract-Conference.html.

A Complete Task-Wise Results

Overall accuracy is question-weighted, with execution and parsing failures counted as incorrect. The main Qwen2.5-VL-7B configuration answers 1,625/2,174 MLVU questions correctly (74.75%). Its reference row is shared by the policy, ablation, and sensitivity studies.

A.1 MLVU

Table 8 gives the complete category breakdown for the three answering backbones under Vgent and QMA-RAG.

Table 8: Complete MLVU accuracy (%).

Backbone	System	Count	Needle	Order	Ego	PlotQA	Anomaly	Topic	Overall
Qwen2.5-VL-7B	Vgent	59.22	79.15	71.81	60.80	74.03	75.00	86.69	72.68
	QMA-RAG	58.74	81.69	75.29	60.23	77.74	77.50	88.59	74.75
LongVU-7B	Vgent	58.25	69.58	50.97	57.67	71.99	77.00	88.21	67.89
	QMA-RAG	57.77	76.34	53.67	60.80	75.32	75.00	88.59	70.47
InternVL3.5-8B	Vgent	65.05	77.46	62.93	61.08	77.37	75.00	89.73	73.14
	QMA-RAG	66.50	76.34	66.80	63.07	81.08	75.50	88.59	74.70

Category sizes: Count 206, Needle 355, Order 259, Ego 352, PlotQA 539, Anomaly 200, Topic 263; total 2,174.

A.2 LongVideoBench

Table 9 consolidates all 17 task categories as correct-answer counts. Vgent is the external baseline for all three answering backbones. The LongVU task counts sum to 718/1,337 (53.70%), which is the reporting basis used throughout. Differences computed from unrounded counts can differ by 0.01 points from differences between displayed percentages.

Table 9: LongVideoBench correct-answer counts. The Overall row additionally reports accuracy (%) in parentheses; N is the category sample count.

		Qwen2.5-VL-7B		LongVU-7B		InternVL3.5-8B	
Task	N	Vgent	QMA-RAG	Vgent	QMA-RAG	Vgent	QMA-RAG
E2O	65	48	45	30	29	43	47
E3E	94	63	68	61	58	65	69
O2E	87	57	60	45	53	59	62
O3O	66	41	42	34	31	47	41
S2A	88	61	70	49	56	65	63
S2E	93	69	76	58	62	66	65
S2O	72	46	49	32	38	48	52
SAA	72	43	42	35	34	39	41
SOS	81	50	54	40	51	57	67
SSS	97	48	48	42	25	43	39
T2A	79	46	54	42	49	52	53
T2E	65	46	41	48	43	43	40
T2O	76	52	52	33	36	45	51
T3E	73	39	39	34	33	43	44
T3O	74	46	46	31	44	46	44
TAA	82	45	44	33	38	50	44
TOS	73	35	35	39	38	32	36
Overall	1337	835 (62.45)	865 (64.70)	686 (51.31)	718 (53.70)	843 (63.05)	858 (64.17)

A.3 Non-Graph Frame-Selection Baselines

Table 10 expands the Qwen2.5-VL-7B comparison in Table 2. Each method retains its native evidence-selection pipeline, and failures remain in the denominator.

Table 10: Task-wise comparison with non-graph frame selection using Qwen2.5-VL-7B (accuracy, %).

(a) MLVU

Method	Count	Needle	Order	Ego	PlotQA	Anomaly	Topic	Overall
AKS	47.57	**84.51**	59.85	**66.19**	74.95	74.50	87.88	72.22
DIG	50.00	**84.51**	61.78	65.06	76.81	72.00	87.88	72.72
QMA-RAG	**58.74**	81.69	**75.29**	60.23	**77.74**	**77.50**	**88.59**	**74.75**

(b) LongVideoBench

Method	E2O	E3E	O2E	O3O	S2A	S2E	S2O	SAA	SOS	SSS	T2A	T2E	T2O	T3E	T3O	TAA	TOS	Overall
AKS	64.62	69.15	71.26	56.06	72.73	68.82	61.11	**62.50**	**71.60**	43.30	65.82	66.15	65.79	53.42	56.76	48.78	35.62	60.96
DIG	**75.38**	71.28	**72.41**	**68.18**	**80.68**	74.19	63.89	59.72	67.90	47.42	**74.68**	**67.69**	**72.37**	**56.16**	56.76	47.56	36.99	64.40
QMA-RAG	69.23	**72.34**	68.97	63.64	79.55	**81.72**	**68.06**	58.33	66.67	**49.48**	68.35	63.08	68.42	53.42	**62.16**	**53.66**	**47.95**	**64.70**

B Allocation Controls and Implementation Details

B.1 MLVU Allocation Controls

Table 11 reports the MLVU policy controls alongside the LVB summary in Table 3. All profiles use a 320-frame final-answer cap: global uses uniform coverage, local uses retrieved clips, and hybrid uses the predefined local-plus-global profile. Fixed local and hybrid retain three and two failed predictions in the 2,174-question denominator. The query-router ablation in Table 7 is a separate intervention. These are policies inside the retrieval-based system. Fixed global changes visual coverage to uniform sampling; it does not by itself remove upstream evidence acquisition or structured support. Uniform in Table 1 bypasses that retrieval path, so the two rows are not equivalent experimental configurations.

Table 11: Adaptive and fixed retrieval policies on MLVU with Qwen2.5-VL-7B. Values are accuracy (%). Fixed local and fixed hybrid retain three and two failed predictions, respectively, in the denominator.

Policy	Count	Needle	Order	Ego	PlotQA	Anomaly	Topic	Overall
Fixed global	43.20	78.59	66.80	59.94	73.84	75.00	88.21	70.47
Fixed local	42.72	81.13	**76.45**	58.81	77.37	72.00	84.41	71.94
Fixed hybrid	45.15	**81.97**	75.68	57.67	76.07	71.50	86.31	71.90
QMA-RAG	**58.74**	81.69	75.29	**60.23**	**77.74**	**77.50**	**88.59**	**74.75**

B.2 LVB Control Definitions and the Uniform Baseline

The six settings in Table 12 belong to the retrieval-based LVB control suite. Fixed G obtains 829/1,337 (62.00%) with global uniform frames and multi-granularity retrieval. The Uniform baseline in Table 1 is evaluated separately and bypasses retrieval.

Table 12: LVB acquisition controls by task (accuracy, %). Simple-Q is a question-only three-group control (global, temporal, or local, with hybrid fallback) whose match directly fixes retrieval order and allocation. G/L/H denote fixed global/local/hybrid retrieval-and-coverage policies; Fixed G retains graph retrieval and is distinct from Uniform. Task codes follow the LongVideoBench taxonomy (Wu et al., 2024).

Configuration	E2O	E3E	O2E	O3O	S2A	S2E	S2O	SAA	SOS	SSS	T2A	T2E	T2O	T3E	T3O	TAA	TOS	Overall
Full (Q+O)	69.23	72.34	68.97	63.64	79.55	81.72	68.06	58.33	66.67	49.48	68.35	63.08	68.42	53.42	62.16	53.66	47.95	**64.70**
Router-Q	72.31	71.28	67.82	68.18	79.55	73.12	62.50	61.11	62.96	44.33	67.09	61.54	69.74	52.05	58.11	56.10	45.21	63.20
Simple-Q	78.46	65.96	56.32	62.12	75.00	72.04	63.89	58.33	67.90	48.45	50.63	60.00	63.16	43.84	54.05	50.00	45.21	59.76
Fixed G	73.85	67.02	64.37	66.67	68.18	74.19	62.50	58.33	66.67	45.36	55.70	67.69	68.42	53.42	55.41	56.10	52.05	62.00
Fixed L	72.31	71.28	68.97	66.67	78.41	74.19	61.11	59.72	64.20	43.30	68.35	60.00	67.11	53.42	56.76	56.10	42.47	62.75
Fixed H	80.00	73.40	66.67	72.73	75.00	69.89	59.72	61.11	61.73	45.36	58.23	64.62	63.16	50.68	56.76	53.66	42.47	62.00

Fixed retrieval and frame policies. Fixed G uses the `global_topic` intent and retrieves in macro, event, then clip order. Its proposal budgets are 10 clips, 8 events, and 5 macro nodes,

Table 13: Matched Q-only evidence acquisition on LVB (accuracy, %). Rule-Q and LLM-Q differ only in the planner that supplies the acquisition configuration; LLM-Q uses Qwen2.5-VL-7B. Options are restored only for final answering. Task codes are as in Table 12.

Configuration	E2O	E3E	O2E	O3O	S2A	S2E	S2O	SAA	SOS	SSS	T2A	T2E	T2O	T3E	T3O	TAA	TOS	Overall
Rule-Q	69.23	67.02	66.67	69.70	73.86	74.19	72.22	58.33	76.54	44.33	65.82	67.69	65.79	45.21	60.81	57.32	47.95	**63.65**
LLM-Q	75.38	64.89	67.82	72.73	77.27	70.97	69.44	63.89	60.49	45.36	56.96	69.23	68.42	52.05	52.70	60.98	46.58	63.05

with a fused-candidate cap of 20. Its final visual input contains up to 320 uniformly distributed global frames and no local clips. Fixed L uses event–clip–macro retrieval with budgets 14/8/2 (clip/event/macro), retains at most 20 fused candidates, and forms its input from up to five local clips. Fixed H uses clip–event–macro retrieval with budgets 14/8/3 and the same fused cap; it combines up to three local clips with up to 128 global samples. Both remain subject to the 320-frame cap. The per-level proposal budgets and fused-candidate cap are retrieval limits, not additional frame allowances.

Retained information paths. All fixed policies retain graph access and the downstream structured-evidence and verification paths. Their usable outputs can accompany the selected frames. Uniform supplies the question, options, and uniformly sampled video directly to the answering backbone. These settings share a frame cap while retaining their respective evidence inputs.

Full uses question and option cues for acquisition. Router-Q removes options only from the router, leaving other acquisition stages intact. Simple-Q uses a question-only global/temporal/local policy with a hybrid fallback; its route directly fixes the retrieval recipe and allocation. The separate Rule-Q/LLM-Q comparison in Table 13 removes options throughout acquisition and changes only the planner, restoring options for final answering.

Scope of the comparisons. These controls compare retrieval-and-coverage policies within the same context-construction system. Fixed G retains upstream retrieval and its structured support, whereas Uniform bypasses that path. The observed differences reflect the policy configuration as a whole.

B.3 Implementation Details

Memory construction. The answering timeline is sampled at 1 fps and grouped into atomic clips of up to 64 frames. EfficientGEBD supplies boundaries on a separate scanning stream. Event boundaries need not align with atomic clips, so each event is mapped to every intersecting clip,

$$\mathcal{C}(e) = \{c \in \mathcal{N}_{\mathrm{a}} : |I_e \cap I_c| > 0\}, \tag{2}$$

where I_n is the source interval of node n. Adjacent events are recursively aggregated into macro regions using a fixed, query-independent criterion over semantic continuity and boundary confidence. Every parent retains the union of its supporting source clips.

Retrieval configuration. The planner supplies the retrieval configuration using fixed deterministic language rules. It maps lexical-semantic indicators of multiplicity, temporal relation, change, and spatial scope to proposal budgets, a temporal operator, and a coverage mode. Explicit scope indicators take precedence; otherwise the coverage fallback uses c_q with $\tau_c = 0.55$. The planner reads the question and options; the subsequent fusion and selection stages rank source clips. Its base configuration is held fixed across the reported LVB controls.

Source-aligned scoring. Atomic retrieval combines text and visual similarity with attribute and subtitle support. Event retrieval combines pooled context, the strongest child match, and support from the best two children; macro retrieval uses pooled text and aggregated lexical support. For parent p, processed in descending relevance order within its level, overlap adjustment is

$$\widetilde{s}_p = \max\left(0, s_p - \lambda_{\mathrm{ov}} \max_{p' \in \mathcal{P}_{<p}} J\big(\mathcal{C}(p), \mathcal{C}(p')\big)\right), \tag{3}$$

where J is Jaccard similarity and an empty maximum is zero. Let E_c and M_c be the strongest adjusted event and macro support for atomic clip c, and let $\mathbf{x}_c$ collect normalized attribute and

subtitle evidence:

$$f(c) = w_{\mathrm{a}} s_c + w_{\mathrm{e}} E_c + w_{\mathrm{m}} M_c + \mathbf{w}_{\mathrm{x}}^{\top} \mathbf{x}_c - \lambda_{\mathrm{p}} r(c). \tag{4}$$

Starting from $S = \emptyset$, selection adds the clip maximizing

$$g_q(c \mid S) = f(c) + d_q(c, S) - \lambda_{\mathrm{r}} R(c, S) - \lambda_{\mathrm{u}} U(c, S), \tag{5}$$

where d_q is temporal utility, R temporal redundancy, and U repeated use of a supporting parent. At most $K = 20$ candidates are retained. The instantiated weights are $w_{\mathrm{a}} = 0.48$, $w_{\mathrm{e}} = 0.27$, $w_{\mathrm{m}} = 0.10$, and $\mathbf{w}_{\mathrm{x}} = (0.08, 0.07)$. Repeated parent support uses $\lambda_{\mathrm{p}} = 0.06\eta$; Eq. (5) uses $\lambda_{\mathrm{r}} = 0.18\eta$ and $\lambda_{\mathrm{u}} = 0.05\eta$, where $\eta = \lambda_{\mathrm{ov}}/0.22$ and the main setting has $\lambda_{\mathrm{ov}} = 0.22$. These values and all tie-breaking rules are shared by the adaptive, fixed-policy, option-removal, and planner-replacement controls.

Frame construction. For T available source samples, frame construction enforces

$$|\mathcal{C}_q| \le K_c, \qquad |\mathcal{F}_q| \le \min(B, T), \tag{6}$$

with $B = 320$ and $K_c = 5$ in the main configuration. Global coverage samples up to 320 frames. Local coverage retains up to five atomic clips and falls back to global sampling if empty. Hybrid coverage combines up to three local clips with 128 global frames ($\rho = 0.40$). If $\mathcal{I}_{\mathrm{loc}}$ and $\mathcal{I}_{\mathrm{glob}}$ are the proposed source indices, frame construction produces

$$\mathcal{F}_q = \{V[i] : i \in \mathrm{SortTime}(\mathrm{Pack}_B(\mathcal{I}_{\mathrm{loc}} \cup \mathcal{I}_{\mathrm{glob}}))\}, \tag{7}$$

where Pack_B removes duplicates and enforces the cap. An oversized union preserves local indices first and fills remaining capacity with spread-out global-only indices; no padding is added after deduplication.

C Detailed Efficiency Analysis

C.1 Profiling Protocol

The full-MLVU comparison uses identical question/video IDs, Qwen2.5-VL-7B, and a 320-frame cap. Evaluation runs on gpu09 with eight A100-PCIe-40GB GPUs and eight workers, processing one question per worker at a time. Both runs contain 2,174 predictions and zero execution errors. The U320 control has no graph access or retrieved nodes and near-zero verification time; QMA-RAG uses manifest-verified prebuilt memory with verification off. One-time process and model initialization is excluded.

For stage s, host-side elapsed wall time is normalized as

$$T_s^{\mathrm{norm}} = \frac{\sum_{i=1}^{n} t_{i,s}}{\sum_{i=1}^{n} D_i/60} \quad [\text{s per video minute}], \tag{8}$$

where D_i is source duration in seconds. The question-weighted denominator is 23,054.37 video-minutes; videos with multiple questions contribute once per question. Per-question summaries are not divided by the worker count.

E2E begins before video loading and ends after answering. The internal online timer starts after graph access and includes query preprocessing, retrieval, verification when enabled, evidence preparation, and generation. These application-level timers are not isolated GPU-kernel measurements.

C.2 Latency Distribution and Stages

With prebuilt memory and verification disabled, mean online latency is 45.039 s/question for QMA-RAG and 40.360 for Uniform. QMA-RAG uses 6.4% fewer frames but adds 4.679 s (11.6%) mean latency. These measurements cover all 2,174 MLVU questions and include video loading, evidence preparation, and answer generation. They describe the verifier-off configuration.

Mean latency and frame counts appear in Table 14; Table 15 adds distribution and memory measurements.

Table 14: Online cost with prebuilt memory and verification disabled. Latency is wall time in s/question, not divided by worker count. Uniform uses Qwen2.5-VL-7B with a 320-frame cap; both runs have zero execution errors.

Configuration	Mean frames	Mean	Median	P95
Uniform (U320)	311.7	40.360	29.739	90.024
QMA-RAG, verifier off	291.7	45.039	34.238	92.112

Table 15: Full-MLVU profiling with prebuilt memory and verification off. U320 denotes uniform sampling with a 320-frame cap. Median/P95 are s/question; normalized E2E is s/video-minute. Peak GPU usage is the maximum observed on one device.

Method	Median	P95	Norm. E2E	Peak GPU (MiB)
U320	29.739	90.024	3.8059	27,275
QMA-RAG	34.238	92.112	4.2471	29,069

The E2E increase is 0.4412 s/video-minute (11.59%); core memory access and planning/retrieval contribute 0.0337 s/video-minute. Other work includes query preprocessing, frame construction, and media loading. Peak GPU figures are run-level observations from the shared harness; the U320 control also retains the embedding runtime, so these are not pure-backbone memory footprints.

D Qualitative Examples and Failure Cases

Archived traces illustrate mechanisms and limitations; the examples were selected after evaluation and do not estimate prevalence. Highlighted spans in Figure 4 show selected evidence rather than ground-truth relevance annotations.

Local detail: `mlvu:ego:0000`. For "What did I put in the orange trashcan?", the retrieval configuration requests local detail and Local allocation. Clips $[6, 5, 4, 0, 2]$ include $[384, 408)$ s. Answer A, "a lemon green sponge," matches the annotation. This shows recovery of a supporting span without establishing relevance of every selected clip.

Temporal ordering: `mlvu:order:0027`. Clips $[5, 6, 7, 8]$ cover 320–576 s of the wakeboarding video. The correct answer is C, $2 \rightarrow 1 \rightarrow 4 \rightarrow 3$, but three action estimates share timestamp 320 s and the fourth is 384 s, with multiple positive clips per action. The trace flags `ambiguous_multi_clip_verification` and declines symbolic answering. Final answering succeeds despite ambiguous localization.

Global allocation: `mlvu:topic_reasoning:0002`. For "What type of film is this?" on a 640-s video, the retrieval configuration requests local detail, while final allocation is Global with 320 uniform frames. Answer C, "Romance," is correct. Retrieval configuration, retrieved nodes, and final frame coverage are recorded in the trace.

Routing mismatch and answer failure. For `mlvu:topic_reasoning:0026`, "What event is depicted in the entire video?", a local-detail configuration with Local allocation still produces

Table 16: Duration-normalized stage times (s/video-minute). The core memory subtotal is the arithmetic sum of graph access and planning/retrieval. Broader pipeline spans overlap and should not be summed as disjoint stages.

Component / timer	U320	QMA-RAG
Graph access	0.0000	0.0199
Planning / retrieval	0.0000	0.0138
Core memory subtotal	0.0000	0.0337
Structured verification	0.0000	0.0000
Answer generation	2.1706	1.9586
Internal online span	2.1706	2.6436
Full question E2E	3.8059	4.2471



Figure 4: Trace-grounded MLVU examples. Photographs illustrate retained spans, not the complete input tensor; fractional-second close-ups in (a) include frames between 1-fps samples. Global sampling marks are thinned for display. The figure depicts allocation, not verified localization accuracy.

the correct answer, "Police drug bust." Thus correctness does not certify correct routing. For `mlvu:findNeedle:0016`, clips $[7, 5, 4, 6, 3]$ yield D, "The windshield is clean," instead of C, "The windshield is broken." Without frame-level relevance annotations, this error cannot be assigned specifically to retrieval, visual ambiguity, or answer interpretation.

E LIMITATIONS

The deterministic planner can misclassify ambiguous or unfamiliar question formulations. The 1-fps source stream and fixed atomic clips limit temporal resolution. A short action between sampled frames can remain inaccessible even when retrieval finds the correct clip; selective source resampling at a higher frame rate is a possible extension. Evaluation covers multiple-choice answering with three backbones; transfer to open-ended tasks remains untested. Component ablations and parameter studies use MLVU, with additional planner controls on LongVideoBench; these studies use Qwen2.5-VL-7B. Fusion removal is evaluated with the full selector retained; the input controls measure incremental effects in one addition order. Detailed stage timing is available for the verifier-off configuration. The small margins over some comparators and the eight-answer Rule-Q/LLM-Q difference lack paired uncertainty estimates; detailed planner-latency measurements are unavailable. Appendix D illustrates routing mismatches, ambiguous localization, and answer failures.