

Query-Adaptive Diversity for Long-Form Video Question Answering

Jonghee Kim¹ and Jinyoung Moon^{1,2*}

¹ Electronics and Telecommunications Research Institute (ETRI)

² University of Science and Technology (UST)

Abstract. Agent-based long-form video question answering selects frames iteratively by query relevance, but across successive turns it can repeatedly retrieve the same frames. To reduce this, we accumulate the selected frames into a cross-turn *memory pool* and score each turn with Maximal Marginal Relevance (MMR) over it — and find that the MMR diversity penalty is *relevance-blind*: it measures redundancy by frame-to-frame similarity alone, ignoring how relevant each frame is to the query. For questions that require multiple frames of the same recurring event — e.g., counting or temporal ordering — it therefore discards genuine evidence and can perform *worse* than applying no diversity term. To address this, we propose a *query-adaptive* diversity penalty: the penalty on a candidate is reduced in proportion to its own query relevance, retaining query-relevant recurrences while suppressing irrelevant duplicates. We evaluate this on MLVU and LVBench under a shared Qwen3.5-9B backbone and a minimal retrieval agent: on both benchmarks the relevance-blind pool falls *below* plain top- K retrieval, whereas the query-adaptive penalty surpasses it, with the improvement concentrated on counting and ordering questions.

Keywords: Query-Adaptive Diversity · Maximal Marginal Relevance · Long-Form Video Understanding · Frame Selection · Video Question Answering

1 Introduction

Long-form videos — lectures, movies, surveillance footage — routinely exceed the context window of a multimodal large language model (MLLM): only a small subset of frames can be selected to represent the video. *Agent-based video question answering* makes this selection iterative: at each turn the agent issues a natural-language query to a frame index, inspects the returned frames, and reformulates the next query based on the frames retrieved so far.

Per-turn top- K retrieval can repeatedly retrieve the same frames: because successive queries are related, independent turns tend to return near-duplicates of the same high-relevance moments, leaving the MLLM with fewer distinct frames than the nominal frame budget allows. To reduce this, we accumulate

* Corresponding author.

the selected frames into a cross-turn *memory pool* and score each turn with Maximal Marginal Relevance (MMR) [1] over it — a cross-turn use of MMR that, to our knowledge, is new to agent-based video QA (Sec. 3.2). However, the MMR redundancy penalty is *relevance-blind*: it defines redundancy by frame-to-frame visual similarity alone, irrespective of how relevant each frame is to the query. For a question whose answer requires *multiple* frames of the same recurring event — “count the times the presenter changes slides”, “in what order do the three characters enter?” — those frames are genuinely similar, and MMR suppresses them as duplicates, discarding the evidence the question requires. Empirically, naive MMR can score *below* plain top- K retrieval on long-form video for exactly this reason (Sec. 4.2).

We therefore propose a *query-adaptive* diversity penalty: the penalty on a candidate is reduced in proportion to its own query relevance, which retains recurring frames relevant to the query while suppressing irrelevant duplicates (Sec. 3.3). This differs from weakening the diversity term through the MMR trade-off weight λ : raising λ relaxes the penalty for *every* candidate, reintroducing irrelevant duplicates as noise, whereas the query-adaptive form relaxes it only where the query-frame relevance is high.

We situate the mechanism in a deliberately *minimal retrieval agent* — each turn it either issues a retrieval query with an optional time-range or commits an answer, with no self-reflection or answer-critique steps — and any accuracy change attributes to the frame-selection rule rather than to a more elaborate agent policy. We evaluate on MLVU [16] and LVBench [12] under a shared Qwen3.5-9B backbone.

Contributions.

- We introduce a cross-turn memory pool scored with MMR to prevent the agent from re-selecting frames across turns — to our knowledge new to agent-based video QA — and identify its failure mode: being relevance-blind, it suppresses the recurring-event frames some questions require, and on long-form video can therefore underperform plain top- K retrieval.
- We propose a *query-adaptive* pooled diversity penalty, retaining query-relevant recurrences while still suppressing irrelevant duplicates. On both MLVU and LVBench this enables the pooled retriever to outperform plain top- K , which the relevance-blind pool does not, with the improvement concentrated on counting and ordering questions.

2 Related Work

Agent-based iterative video QA. Recent systems recast video QA as an iterative evidence-gathering loop over a frame index. VideoAgent [13] interleaves retrieval with self-reflection and planning steps; Video-RAG [4] adds a retrieval-augmented generation layer; DVD [15] performs agentic search with tools over a multi-granular video database. These systems improve the *controller* — how to plan, reflect, and verify — while treating the retrieval mechanism itself as fixed.

Our contribution is orthogonal and complementary: a better frame-selection rule — a query-conditioned diversity penalty over a cross-turn pool — that such controllers could adopt unchanged. We study it with a deliberately minimal agent (§3.1) solely to ensure that accuracy changes attribute to the retrieval rule rather than to the controller.

Frame selectors for video LLMs. A parallel line of work replaces uniform sampling with a learned frame selector as a preprocessing step on top of a fixed end-to-end pipeline. AKS [11] selects keyframes by query-conditioned scoring; Frame-Voyager [14] learns a policy over frame subsets; Apollo [17] and Trial [10] explore alternative selection policies. These selectors run *once* per question, outside any agent loop, and cannot adapt as the agent reformulates its query over turns. Our mechanisms operate *inside* the loop — each retrieval is conditioned on the similarity of the current query to prior queries and on the per-candidate relevance to the current turn.

MMR and diversification variants. MMR [1] and its later variants, together with DPP-based selectors (SeqDPP [2], SH-DPP [9]), diversify frame or document selections but define redundancy independently of the query; conversational retrieval extends diversification to multi-turn settings, yet operates on text passages [5, 6] rather than video frames. To our knowledge, no prior work applies MMR over an accumulating cross-turn frame pool inside an agent-based video-QA loop, nor conditions the redundancy penalty on per-candidate query relevance to admit multiple frames of the same recurring event.

3 Method

3.1 Overall Framework

Fig. 1 summarizes the framework. Given a video V and a multiple-choice question Q , we build a 1 fps CLIP [8] image index $\mathcal{F} = \{(f_i, t_i)\}_{i=1}^N$, where $f_i \in \mathbb{R}^d$ is the L2-normalized CLIP embedding of the frame at time t_i . A single MLLM then drives an agent loop over up to T turns. At each turn s the agent takes one of two actions: (i) issue a retrieval query q_s^{txt} , optionally scoped to a time range $\tau_s = [a_s, b_s]$, and receive the top- K frames from $\text{retrieve}(\cdot)$, which it appends to its context; or (ii) commit its final answer and stop.

```

prior_picks = []
for s = 1..T:
    a = agent(Q, history) # answer, or retrieve ( $\tau_s, q_s^{\text{txt}}$ )
    if a is an answer: return a
    q_s = CLIP_text( $q_s^{\text{txt}}$ )
    F_s = retrieve(q_s, mask= $\tau_s$ , priors=prior_picks)
    history += F_s; prior_picks += [(f, q_s) for f in F_s]
return best-effort answer # max steps reached

```

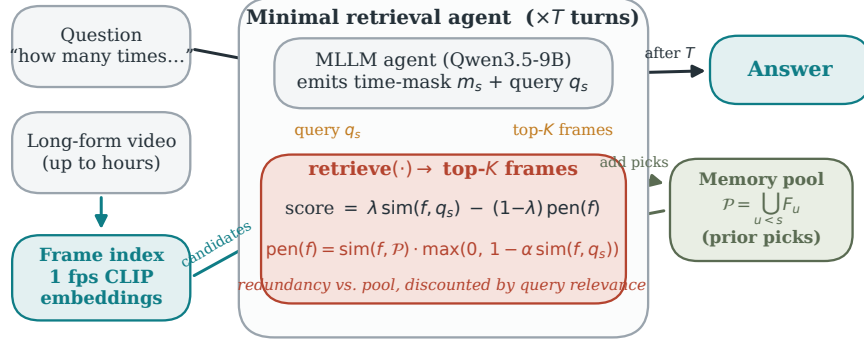


Fig. 1: Method overview. At each turn the agent either commits an answer or issues a retrieval query with an optional time range; $\text{retrieve}(\cdot)$ scores candidates by query relevance minus a query-conditioned redundancy penalty measured against the memory pool of previously selected frames (Eq. 2).

The $\text{retrieve}(\cdot)$ step contains our contributions (Sec. 3.2–3.3); otherwise the agent policy is deliberately minimal — no self-reflection, verification, planning, or answer-critique steps — and any accuracy change therefore isolates the frame-selection rule rather than a more elaborate agent policy. A time range τ_s , when supplied, restricts retrieval to frames within $[a_s, b_s]$. The following two subsections define what $\text{retrieve}(\cdot)$ returns: a cross-turn pool with MMR (Sec. 3.2) and query-conditioned redundancy (Sec. 3.3).

3.2 Cross-Turn Frame Pool with MMR

At turn s , standard per-turn top- K retrieval scores each frame f against q_{now} alone and returns the K highest-scoring frames. Across turns, the returned sets from turn $1, \dots, s-1$ do not influence turn s , and the retriever may return the same frames repeatedly. This is most severe in an agent loop, where follow-up queries are by design semantically related to earlier ones.

We introduce a cross-turn frame pool $\mathcal{P}_s = \bigcup_{u < s} F_u$ that accumulates the frames selected in all prior turns, and score the candidates at turn s with Maximal Marginal Relevance [1] over this pool:

$$\text{MMR}(f) = \lambda \cdot \text{sim}(f, q_{\text{now}}) - (1 - \lambda) \cdot \max_{f_u \in \mathcal{P}_s} \text{sim}(f, f_u). \quad (1)$$

The redundancy term now discounts candidates that visually resemble a frame the agent has already inspected, whether that frame was selected in the current turn (single-turn diversity) or in any prior turn (cross-turn deduplication).

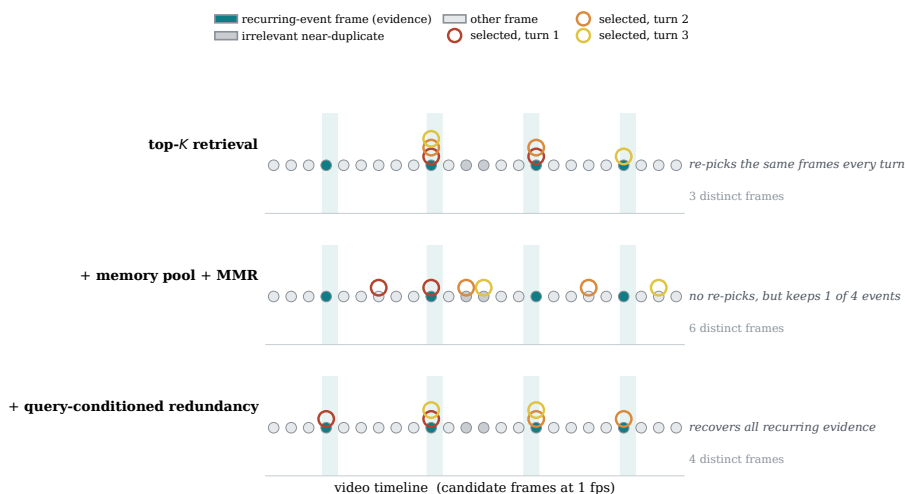


Fig. 2: Frame selection across agent turns for a recurring-event question (shaded bands: recurring instances; rings: selections, coloured by turn). **Top-K** re-selects the same high-relevance frames every turn (stacked rings); **memory pool + MMR** prevents re-selection but suppresses the recurring instances as duplicates; **query-conditioned redundancy** retains them while preserving the no-re-selection behaviour.

3.3 Query-Conditioned Redundancy

MMR (Eq. 1) enforces visual diversity uniformly across all candidates. This is the correct default for questions whose answer depends on a single distinctive frame: two visually similar candidates carry the same information, and the redundancy penalty correctly retains only one of them.

A specific class of questions violates that assumption. When the answer requires *multiple frames of the same recurring visual pattern* — counting how many times an action repeats, ordering visually similar sub-events, or accumulating evidence of a recurring object across timestamps — the “duplicates” are precisely the evidence the question requires. On these questions, the visual-similarity heuristic that MMR uses as a proxy for redundancy is a false positive: similar frames are informative, not redundant. The MMR penalty therefore collapses multiple recurring instances into a single representative frame, which deprives the downstream reasoner of the counting/ordering evidence.

We propose *query-conditioned redundancy*: we make the redundancy penalty depend on the current query in two complementary ways, allowing a plausible recurring instance to survive rather than be suppressed as a duplicate. First, a pooled frame f_u was retrieved at an earlier turn $u < s$ for the query q_u , which need not be related to the current query q_{now} ; we weight its contribution by $w_u = \max(0, \text{sim}(q_{\text{now}}, q_u))$, and frames retrieved for unrelated sub-queries therefore contribute little to the penalty. Second, a candidate that is itself highly relevant to the current query is more likely a relevant recurrence than a redun-

dant duplicate; we therefore discount its penalty in proportion to that relevance. Together, the redundancy term of Eq. 1 becomes

$$\text{pen}(f) = \left(\max_{f_u \in \mathcal{P}_s} w_u \text{sim}(f, f_u) \right) \cdot \max(0, 1 - \alpha \cdot \text{sim}(f, q_{\text{now}})). \quad (2)$$

Setting $w_u=1$ and $\alpha=0$ recovers plain relevance-blind MMR; a larger discount coefficient $\alpha \geq 0$ progressively allows the retriever to co-select similar, highly relevant frames. Crucially the relaxation is query-conditioned — frames irrelevant to the current query keep the full penalty — and therefore does not undo diversity in general, only where the query requires recurring evidence: precisely the regime where the MMR redundancy proxy fails.

Sec. 4.3 sweeps the discount coefficient on LVBench: accuracy is worst under full diversity, peaks at an intermediate discount, and declines again when the discount grows too aggressive; the chosen setting exceeds plain top- K (the no-diversity limit) on both benchmarks (Sec. 4.2). The gain therefore comes from applying the penalty *selectively*, not from weakening it overall, which the global weight λ — trading relevance against diversity for *all* candidates at once — cannot do.

4 Experiments

4.1 Setup

Datasets. We evaluate on two long-form benchmarks. MLVU dev (2174 questions, 1122 videos, 7 task types, tens of seconds to 2 hours) includes recurring-event tasks (counting, ordering) that stress the MMR failure mode. LVBench (1549 questions, 103 videos, mean duration ~ 61 min) is an extreme-long benchmark where frame selection is most consequential.

Implementation Details. Backbone: Qwen3.5-9B [7] accessed via vLLM [3]. Frame index: 1 fps CLIP ViT-B/32 image embeddings, 256 px short side. Agent loop: max steps $T=12$, top- K per retrieve $K=8$ (default; the agent may request a different K per call), $\lambda_{\text{MMR}}=0.5$.

4.2 Main Results

Table 1 presents the central result as a four-stage ladder — uniform sampling, top- K retrieval, a memory pool + MMR, and query-conditioned redundancy — run identically on MLVU and LVBench. All stages share the Qwen3.5-9B backbone, the same 1 fps CLIP index, and the same minimal agent; each row adds a single component, and its accuracy delta is the contribution of that component.

Table 1: Frame-selection ladder under a shared Qwen3.5-9B backbone on two long-form benchmarks. “Uniform” is query-agnostic temporal sampling (64 frames); the remaining rows are the minimal retrieval agent (Sec. 3.1) with progressively richer frame selection, each adding one component. Best per column in **bold**.

Frame-selection stage	MLVU LVBench	
Uniform sampling (no retrieval)	67.94	42.67
Top- K retrieval (by query relevance)	69.03	52.55
+ memory pool + MMR	67.02	50.29
+ query-conditioned redundancy	69.60	53.26

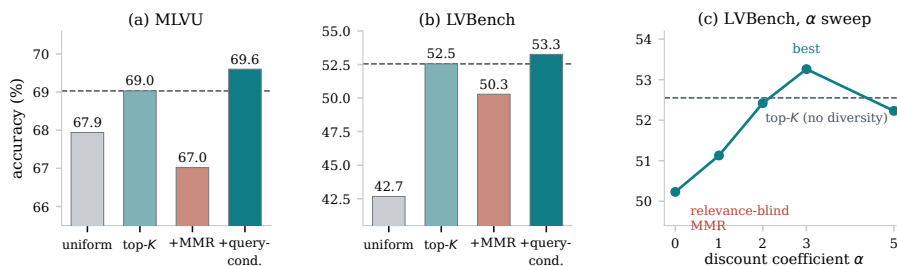


Fig. 3: (a, b) The frame-selection ladder of Table 1: the relevance-blind MMR pool falls below plain top- K retrieval (dashed line); query-conditioned redundancy recovers above it. (c) Sweep of the discount coefficient α on LVBench: worst with no discount, peaking above the top- K line at $\alpha=3$.

A relevance-blind pool degrades accuracy; query-conditioning turns it into a gain. Top- K retrieval already outperforms uniform sampling — by ~ 10 points on hour-scale LVBench ($42.67 \rightarrow 52.55$), where accuracy depends primarily on *which* frames are selected. Adding the memory-pool MMR term then *reduces* accuracy on both benchmarks (LVBench $52.55 \rightarrow 50.29$, MLVU $69.03 \rightarrow 67.02$): being relevance-blind, the diversity penalty discards recurring-event evidence, and accuracy drops below plain top- K . Conditioning the redundancy on the query reverses this, lifting accuracy back above top- K on both ($\rightarrow 53.26$, $\rightarrow 69.60$; $+0.71$, $+0.57$). This dip-then-recover pattern, identical on both benchmarks, is our central empirical result: the same diversity term that degrades accuracy while relevance-blind becomes beneficial once query-conditioned (Fig. 3a,b).

4.3 Ablations

Discount coefficient. Fig. 3c sweeps the discount coefficient on LVBench. Accuracy is worst with no discount (full diversity), rises as the penalty is relaxed for query-relevant candidates, peaks at $\alpha=3$ — exceeding plain top- K (Sec. 4.2) — and declines again at $\alpha=5$, where the discount becomes aggressive enough to erase the diversity benefit. The interior optimum indicates that both extremes

Table 2: MLVU accuracy by task type, isolating the relevance discount: the unweighted pooled MMR penalty with ($\alpha=3$) and without ($\alpha=0$) the discount of Eq. 2. The gain is confined to the recurring-event tasks (*count*, *order*); single-frame task types are essentially unchanged.

	count	order	plotQA	needle	ego	anomaly	topic
MMR	34.5	61.8	72.7	75.5	59.1	69.5	83.3
query-conditioned	46.6	72.2	72.4	75.5	58.0	67.0	83.3
Δ	+12.1	+10.4	-0.4	0.0	-1.1	-2.5	0.0

are inferior: full diversity suppresses recurring evidence, while an excessive discount readmits duplicates. We use $\alpha=3$ throughout.

Which questions benefit. Query-conditioned redundancy should help exactly the questions whose answer requires multiple frames of the same recurring event, and the MLVU per-task breakdown confirms this (Table 2). The gain concentrates on the two multi-instance task types, *count* and *order*; the other five task types, answerable from a single distinctive frame, are essentially unchanged. These are precisely the cases where the MMR heuristic “visually similar $\Rightarrow$ redundant” is a false positive: the recurring instances a count or order question requires resemble one another, and relevance-blind MMR therefore suppresses them while query-conditioning restores them. Figs. 4 and 5 illustrate two complementary failure modes on counting questions. In Fig. 4, top- K re-retrieves the same frames across turns and the relevance-blind pool scatters retrieval over the video; both under-count, while the query-conditioned run concentrates on the recurring instances — co-selecting near-duplicate views the relevance-blind penalty would reject — and answers correctly. In Fig. 5 the target action occurs exactly once: top- K is unaffected, but the relevance-blind penalty pushes retrieval off the single relevant scene and the agent concludes the action never occurs; query-conditioning recovers it.

5 Conclusion

We studied how to select frames for long-form video QA in a minimal retrieval agent. Our finding is that a memory-pool MMR diversity term is relevance-blind and *degrades accuracy* on long-form video — it falls below plain top- K on both MLVU and LVBench by discarding recurring-event evidence — and that conditioning its redundancy penalty on the query corrects this. Query-conditioned redundancy — weighting the cross-turn penalty by query similarity and discounting it by the query relevance of each candidate — recovers the accuracy lost at the pooled MMR step and lifts both benchmarks above plain top- K , where a relevance-blind pool falls below it, with the gain concentrated on count and order tasks. The gains are specific to long-form, recurring-event questions, and the discount coefficient α is a fixed hyperparameter whose optimum is likely video-dependent; a data-driven, per-question α is natural future work.

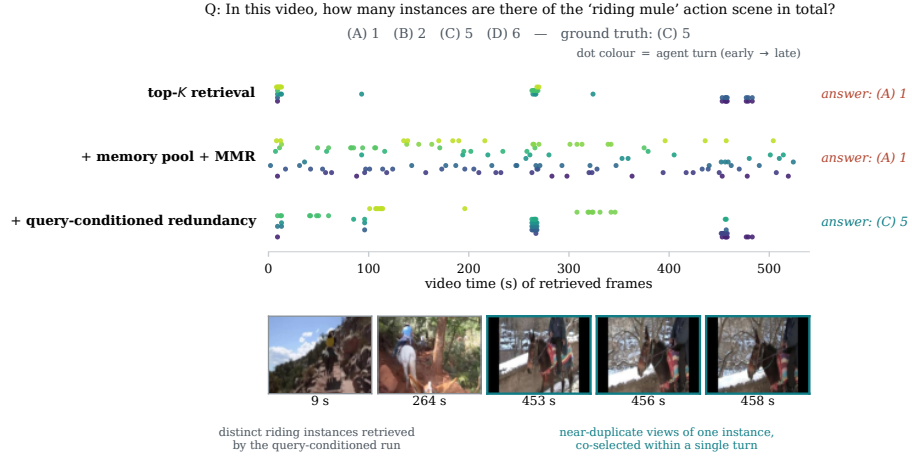


Fig. 4: Qualitative example: an MLVU counting question (ground truth 5). *Top:* retrieval times, one dot per frame, coloured by turn. Top- K re-retrieves the same frames (its first two turns are identical) and the MMR pool scatters retrieval; both undercount. The query-conditioned run concentrates on the recurring segments and counts correctly. *Bottom:* its retrieved frames — distinct instances (left) and near-duplicate views co-selected within one turn (right).

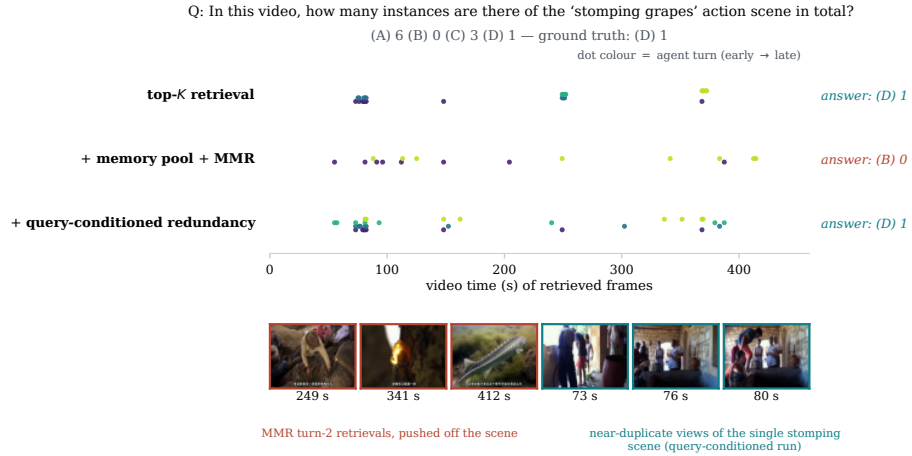


Fig. 5: Second qualitative example: the target action occurs exactly once. Top- K answers correctly, but the relevance-blind MMR penalty drives retrieval off the single relevant scene and the agent concludes the action never occurs; query-conditioning re-admits near-duplicate views of the scene (bottom right; the off-scene MMR retrievals, bottom left).

Acknowledgements

This work was supported by IITP grant funded by the Korea government (MSIT) (No. RS-2020-II200004, Development of Previsional Intelligence based on Long-term Visual Memory Network).

References

1. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for re-ordering documents and producing summaries. In: Proc. 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 335–336 (1998). <https://doi.org/10.1145/290941.291025> 2, 3, 4
2. Gong, B., Chao, W.L., Grauman, K., Sha, F.: Diverse sequential subset selection for supervised video summarization. In: NeurIPS. pp. 2069–2077 (2014) 3
3. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with PagedAttention. In: Proceedings of the 29th Symposium on Operating Systems Principles (SOSP) (2023) 6
4. Luo, Y., Zheng, X., Yang, X., Li, G., Lin, H., Huang, J., Ji, J., Chao, F., Luo, J., Ji, R.: Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In: NeurIPS (2025) 2
5. Min, S., Lewis, M., Hajishirzi, H., Zettlemoyer, L.: Joint passage ranking for diverse multi-answer retrieval. In: EMNLP (2021) 3
6. Mo, F., Mao, K., Zhao, Z., Qian, H., Chen, H., Cheng, Y., Li, X., Zhu, Y., Dou, Z., Nie, J.Y.: History-Aware conversational dense retrieval. In: Findings of ACL (2024) 3
7. Qwen Team: Qwen3.5-Omni technical report. arXiv preprint arXiv:2604.15804 (2026) 6
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 3
9. Sharghi, A., Gong, B., Shah, M.: Query-focused extractive video summarization. In: ECCV (2016) 3
10. Suo, Y., Ma, F., Zhu, L., Wang, T., Rao, F., Yang, Y.: From trial to triumph: Advancing long video understanding via visual context sample scaling and self-reward alignment. In: ICCV (2025) 3
11. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding. In: CVPR (2025) 3
12. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y., Ding, M., Tang, J.: LVBench: An extreme long video understanding benchmark. In: ICCV (2025) 2
13. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: VideoAgent: Long-form video understanding with large language model as agent. In: ECCV (2024) 2
14. Yu, S., Jin, C., Wang, H., Chen, Z., Jin, S., et al.: Frame-Voyager: Learning to query frames for video large language models. In: ICLR (2025) 3
15. Zhang, X., Jia, Z., Guo, Z., Li, J., Li, B., Li, H., Lu, Y.: Deep video discovery: Agentic search with tool use for long-form video understanding. In: NeurIPS (2025) 2

16. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: MLVU: Benchmarking multi-task long video understanding. In: CVPR (2025) [2](#)
17. Zohar, O., Wang, X., Dubois, Y., Mehta, N., Xiao, T., et al.: Apollo: An exploration of video understanding in large multimodal models. In: CVPR (2025) [3](#)