MuKV: Multi-Grained KV Cache Compression for Long Streaming Video Question-Answering

Supplementary Material

1. Dataset Introduction

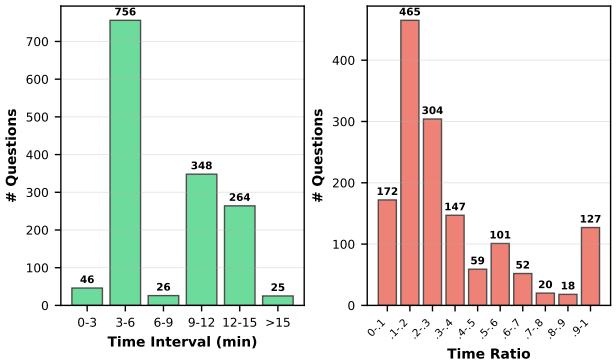
VStream-QA [15] comprises two long-video datasets: RVS-Ego and RVS-Movie. **RVS-Ego** contains 10 egocentric videos with an average duration of 30 minutes, while **RVS-Movie** includes 22 movie videos averaging 1 hour. The distributions of the temporal answer spans and their ratios relative to the question timestamps of both datasets are presented in Fig. 1. The results show that the answer spans and their relative ratios in RVS-Ego are substantially longer than those in RVS-Movie, indicating that RVS-Ego has a higher chance of capturing the answer segment and would result in less redundancy under uniform video sampling for streaming QA. **StreamingBench** [7] does not provide answer span annotations, so we introduce the 10 question categories defined over the subset of 500 videos (on average 10 minutes) in the task of real-time visual understanding:

- **Object Perception (OP):** Detect and identify specific objects, *e.g.*, "What is the person holding right now?".
- **Causal Reasoning (CR):** Analyze event cause-and-effect relationships, *e.g.*, "Why Mr Bean is shocked now?".
- **Clips Summarization (CS):** Summarize main content in specific video clips, *e.g.*, "Which of following best summarize the actions just now".
- **Attribute Perception (ATP):** Identify and categorize object or individual attributes, *e.g.*, "What color is the car directly in front right now?"
- **Event Understanding (EU):** Recognize and describe sequences of events, *e.g.*, "What is happening in the initial scene of the video?"
- **Text-Rich Understanding (TR):** Interpret and explain text-rich content within the video, *e.g.*, "Which team is leading in the racing points?".
- **Prospective Reasoning (PR):** Predict future events based on current video context, *e.g.*, "What might the speaker explain next?".
- **Spatial Understanding (SU):** Understand and describe spatial relationships and locations, "Where is ... now?"
- **Action Perception (ACP):** Identify specific actions in the video, *e.g.*, "What is the person doing now?".
- **Counting (CT):** Count occurrences of specific objects or actions, *e.g.*, "How many times does .. so far?"

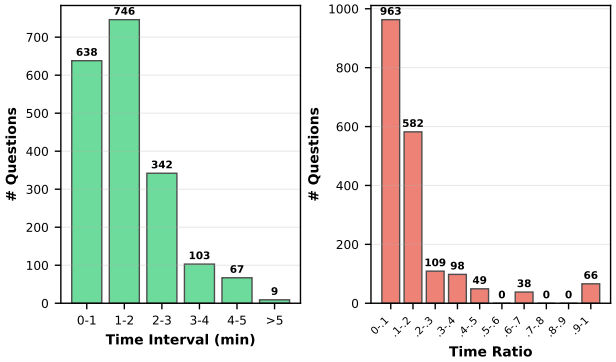
2. Experiments

2.1. Offline VideoQA and Different Backbones

We also extend our method MuKV to the popular offline long VideoQA datasets: Video-MME [3], MLVU [17] and EgoSchema [8]. For a streaming QA setting, we assume that all questions are asked at the end of the videos. The results in Tab. 1 show that MuKV steadily improves over ReKV and other recent streaming QA methods under different backbones, demonstrating the robustness of our multi-grained KV cache compression and semi-hierarchical retrieval approach.



(a) Distribution of answer spans and relative ratios on RVS-Ego.



(b) Distribution of answer spans and relative ratios on RVS-Movie.

Figure 1. The answer spans and their ratios relative to the corresponding question timestamps in RVS-Ego are substantially longer than those in RVS-Movie, meaning that RVS-Ego has a higher chance of capturing the answer segment and thus brings less redundancy under uniform video sampling for streaming QA.

2.2. Hyper-parameters

Frame Sampling Rates. Tab. 2 shows that denser video sampling improves model performance on StreamingBench but harms performance on RVS-Ego. Upon examining the datasets, we find that RVS-Ego exhibits much lower visual appearance variation than StreamingBench, meaning that sparse sampling already captures most key frames,

Table 1. Results comparison on offline long VideoQA benchmarks. Compared results are copied from StreamMem [14]. For the number of memory tokens, we report on a per 30-frame (1 minute) basis to satisfy all dataset, *e.g.*, $5.9K \approx 196 * 30$.

Method	Frames/FPS	# Mem. Tok	MLVU	EgoSchema	VideoMME		
					Medium	Long	All
GPT-4o [4]	–	–	64.6	72.2	70.3	65.3	71.9
MovieChat+ [12]	2048	-	25.8	53.5	-	33.4	38.2
Dispider [10]	1 fps	-	61.7	55.6	53.7	49.7	57.2
LongVA [16]	128	–	–	–	50.4	46.2	52.6
LongVU [11]	400/1fps	–	67.6	58.2	59.5	60.6	-
LLaVA-OV-7B [6]	32	–	64.7	60.1	54.7	46.2	56.9
+ ReKV	0.5 fps	5.9K	68.5	60.7	–	–	–
+ LiveVLM [9]	0.5/0.2 fps	-	66.3	63.0	56.4	48.8	57.3
+ StreamMem [14]	0.5/0.2 fps	6K	66.9	63.0	56.6	50.1	59.4
+ MuKV (Ours)	0.5fps	5.9K	67.8	63.3	57.9	52.1	61.2
LLaVA-OV-0.5B [6]	32	–	50.4	26.8	39.7	46.2	45.4
+ ReKV	0.5 fps	5.9K	53.2	29.6	40.1	47.9	48.2
+ MuKV (Ours)	0.5fps	5.9K	55.2	30.5	41.4	48.5	49.1
Qwen2.5-VL-3B [1]	768	–	63.3	64.4	58.0	47.2	60.3
+ InfiniPot-V [5]	768	6K	62.1	61.8	-	-	59.3
+ StreamMem [14]	4.0/0.5 fps	6K	62.3	62.2	60.1	49.1	59.5
+ MuKV (Ours)	0.5 fps	5.9K	63.0	63.0	61.0	50.0	60.8
Qwen3-VL-4B [13]	768	–	64.8	65.8	60.2	49.5	62.5
+ MuKV	0.5fps	5.9K	66.0	67.0	61.8	51.0	63.6
Qwen3-VL-2B [13]	768	–	64.0	65.0	59.0	48.5	61.5
+ MuKV	0.5fps	5.9K	65.2	66.3	60.8	49.8	62.6
Qwen3-VL-8B [13]	768	–	76.1	66.2	70.1	64.9	71.5
+ ReKV	0.5fps	5.9K	76.9	67.1	70.3	65.2	71.8
+ MuKV	0.5fps	5.9K	78.3	68.7	72.5	67.7	73.3

while denser sampling introduces unnecessary redundancy. Meanwhile, MuKV consistently improves ReKV [2] across all sampling rates, with the performance gains becoming more pronounced at higher FPS values. This further demonstrates the effectiveness of our KV compression mechanism for removing redundancy.

Granularity Hyper-Parameters. Tab. 3 studies different compression ratios at KV cache of different granularities. The results show that pruning more KVs at the lower-level granularity often yields better performance, highlighting the strategy of segment-level modeling in video understanding. Tab. 4 studies the trade-off parameter between the first-stage parallel-retrieval score and the second-stage cross-grain hierarchical retrieval score. A smaller λ brings better performance on RVS-Ego but worse performance on RVS-Movie, suggesting that cross-grain retrieval scores are less effective on egocentric videos which have less event-level content variations compared to that on movie videos.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, Hao Jiang, et al. Streaming video question-answering with in-context video kv-cache retrieval. In *ICLR*, 2025. 2, 3

[3] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 1

[4] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[5] Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding. *NeurIPS*, 2025. 2

[6] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[7] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024. 1

[8] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra

Table 2. Streaming VideoQA performance comparison under different video sampling rates (FPS). For evaluation on VStream-QA [15], we use GPT-3.5-turbo. Less inference visual tokens and memory tokens indicate higher efficiency. For the number of memory tokens, we report on a per 300-frame (10 minutes) basis to satisfy all datasets, *e.g.*, $59K \approx 196 * 300$.

Model	**Size**	**FPS**	**#Inf. Tok ↓**	**#Mem. Tok ↓**	**RVS-Ego**	**StreamingBench (Real-Time Understanding)**										
						OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	All
ReKV [2]	7B	0.5	12.5K	59K	56.2	71.3	70.3	69.1	68.6	61.6	55.2	63.9	58.1	55.7	46.3	62.3
ReKV [2]	7B	2	12.5K	236K	54.8	72.0	70.9	66.2	69.2	63.0	55.7	64.1	58.7	56.2	46.4	62.9
ReKV [2]	7B	3	12.5K	354K	53.9	71.0	70.5	65.2	69.2	62.1	55.2	64.5	58.2	55.8	46.6	62.5
MuKV (Ours)	7B	0.5	8.3K	59K	59.5	74.0	78.2	72.2	71.8	64.8	55.5	66.7	58.9	59.4	39.4	64.4
MuKV (Ours)	7B	2	8.3K	118K	57.7	78.4	78.3	82.1	77.7	68.3	57.4	70.1	62.8	61.3	42.0	68.2
MuKV (Ours)	7B	3	8.3K	354K	56.1	82.1	82.8	82.1	80.0	71.9	61.6	74.0	65.4	65.9	43.7	71.4

Table 3. Sensitivity analysis on granularity-specific KV retention (reversed side of compression) ratios (ρ_p, ρ_f, ρ_s). Higher segment retention (ρ_s) leads to better performance, highlighting the importance of segment-level signal cues.

Retention Ratio (ρ_p, ρ_f, ρ_s)			RVSEgo		RVSMovie	
ρ_p	ρ_f	ρ_s	Acc	Score	Acc	Score
0.1	0.1	0.8	**57.9**	**3.89**	**45.2**	**3.34**
0.1	0.8	0.1	55.6	3.80	44.1	3.33
0.8	0.1	0.1	54.7	3.79	43.5	3.30

Table 4. Sensitivity analysis of λ in the semi-hierarchical retrieval module. Smaller λ reduces dependency on the second-stage cross-grain retrieval scores, which slightly improves performance on RVSEgo but declines performance on RVS-Movie.

Method	RVSEgo		RVSMovie	
	Acc	Score	Acc	Score
MuKV rerank (λ={0.7, 0.7, 0})	56.1	3.86	**47.2**	**3.41**
MuKV rerank (λ={0.3, 0.3, 0})	**57.9**	**3.89**	45.2	3.34

Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 1

[9] Zhenyu Ning, Guangda Liu, Qihao Jin, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025. 2

[10] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24045–24055, 2025. 2

[11] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 2

[12] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[13] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2

[14] Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025. 2

[15] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *ICCV*, 2025. 1, 3

[16] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 2

[17] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701, 2025. 1