Q-Frame: Query-aware Frame Selection and Multi-Resolution Adaptation for Video-LLMs

Supplementary Material

A. Limitations

Q-Frame enhances query-aware video understanding, but it depends on pre-trained models, lacks explicit temporal modeling, and operates within a fixed token budget. Its evaluation primarily relies on benchmarks, which highlight the need for validation in real-world applications. Future work could focus on adaptive frame budgeting, multi-modal fusion, and end-to-end optimization.

B. Experimental Details

B.1. Dataset Details

LongVideoBench [29] is a newly introduced benchmark aimed at evaluating long-term video-language understanding for MLLMs. It comprises 3,763 web-collected videos of varying lengths (up to one hour), all featuring subtitles and encompassing a wide array of themes. This dataset is designed to assess models' capabilities to process and reason with detailed multimodal information from long-form video inputs and is notably comprehensive. It includes 6,678 human-annotated multiple-choice questions spread across 17 fine-grained categories. In this paper, we focus on the validation set without subtitles, which consists of 1,337 question-answer pairs and has an average video duration of 12 minutes.

MLVU [41] is a new dataset designed to evaluate Long Video Understanding (LVU) performance. It addresses the limitations of existing benchmarks by providing longer video durations, a variety of video genres (including movies, surveillance footage, and cartoons), and multiple evaluation tasks. With an average video duration of 12 minutes, the benchmark includes 2,593 tasks across nine categories, delivering a thorough assessment of MLLMs' capabilities in comprehending long videos.

Video-MME [7] is a dataset designed to enhance video understanding for Multimodal Large Language Models (MLLMs). It comprises 900 videos spanning 6 visual domains, with durations ranging from 11 seconds to 1 hour, capturing diverse contextual dynamics. All videos are manually annotated by experts, resulting in 2,700 question-answer pairs, ensuring high-quality data for model evaluation. Experiments on Video-MME will be conducted both with and without subtitles to assess the impact of multimodal inputs.

Parameter	Value
Type	Azure
Model	GPT-4o
Version	2023-12-01-preview
Deployment	GPT-4o

Table 7. Parameter configuration of the API-based model GPT-4o.

Model	Round	#Frames	Acc(%)
GPT-4o [21]	1	8	27.4
	2	8	<u>28.6</u>
	3	8	28.3
+ Q-Frame	-	8	29.3

Table 8. Comparing GPT-4o with and without Q-Frame as an additional module on MLVU benchmark. The <u>underlined</u> results are adopted in the manuscript.

B.2. Repeated Experiment on GPT-4o

As shown in Table 2, the experimental results of GPT-4o on the MLVU benchmark are quite unusual. Therefore, we conducted several rounds of experiments to verify the results. The parameter configuration of the API-based model GPT-4o is detailed in Table 8. The results of the multiple rounds of experiments are presented in Table 10. <u>The underlined results are adopted in the manuscript</u>. This result appears to be intentional, although the specific cause remains under investigation. It might be related to how the API is invoked. The experimental conclusion indicates that, under the same experimental conditions, Q-Frame is also effective for this closed-source model.

B.3. Detailed experimental results on subtasks

To further analyze Q-Frame's impact, we report its performance across multiple subtasks on LongVideoBench, MLVU, and Video-MME. The results indicate that Q-Frame consistently enhances model performance across most subtasks, particularly in query-dependent tasks that require adaptive frame selection.

On LongVideoBench, Q-Frame improves performance in most categories, achieving notable gains in temporal and object-centric tasks. However, some subtasks show minor drops in performance, suggesting that further optimization may be needed for specific scenarios.

Model	**LongVideoBench**																
	TOS	S2E	E3E	S2A	SAA	O3O	T3O	T3E	O2E	T2O	S2O	TAA	T2E	E2O	SSS	T2A	SOS
VILA-V1.5 [15]	26.0	61.3	52.1	59.1	47.2	53.0	44.6	41.1	56.3	50.0	45.8	46.3	56.9	52.3	23.7	49.4	55.6
+ **Q-Frame**	31.5	66.7	54.3	73.9	47.2	40.9	40.5	34.2	59.8	56.6	56.9	42.7	63.1	67.7	24.7	59.5	56.8
GPT-4o [21]	34.2	58.2	48.7	67.8	67.7	45.2	45.1	62.8	45.9	28.9	55.6	63.1	70.5	55.6	48.5	50.0	56.9
+ **Q-Frame**	39.7	67.1	67.1	69.0	63.4	47.9	47.6	57.4	54.1	29.9	63.0	72.3	76.1	58.3	54.5	66.7	66.2
Qwen2-VL-Video [27]	39.7	66.7	59.6	56.8	51.4	40.9	45.9	39.7	56.3	48.7	51.4	47.6	52.3	63.1	35.1	49.4	59.3
Qwen2-VL [27]	31.5	61.3	60.6	53.4	52.8	54.5	48.6	41.1	59.8	60.5	55.6	47.6	58.5	78.5	36.1	51.9	60.5
+ **Q-Frame**	41.1	71.0	66.0	76.1	51.4	56.1	48.6	41.1	63.2	61.8	65.3	51.2	64.6	70.8	35.1	62.0	66.7

Table 9. Performance of subtasks on LongVideoBench. Red fonts represent positive results compared to the Baseline, and blue fonts represent negative results.

Model	**MLVU**								
	TR	TR	VS	NQA	ER	PQA	SSC	AO	AC
VILA-V1.5 [15]	78.8	52.0	0.0	56.6	40.9	45.8	0.0	29.7	21.4
+ **Q-Frame**	73.1	46.5	0.0	80.3	44.9	56.8	0.0	30.5	33.5
GPT-4o [21]	40.2	24.5	0.0	30.7	27.6	26.7	0.0	25.5	24.8
+ **Q-Frame**	39.0	24.5	0.0	28.7	32.1	27.7	0.0	26.6	25.7
Qwen2-VL-Video [27]	78.4	64.0	0.0	64.5	50.6	55.1	0.0	44.8	26.2
Qwen2-VL [27]	83.3	58.5	0.0	61.4	56.8	59.9	0.0	45.9	20.4
+ **Q-Frame**	78.8	57.0	0.0	80.3	63.1	69.0	0.0	53.3	40.8

Table 10. Performance of subtasks on MLVU. Red fonts represent positive results compared to the Baseline, and blue fonts represent negative results.

Model	**Video-MME**$_{(wo/w\ subs)}$					
	Reasoning	Recognition	Perception	Counting	OCR	Information Synopsis
VILA-V1.5 [15]	46.5 / 49.8	48.3 / 51.9	58.9 / 62.6	35.1 / 35.1	43.9 / 51.2	58.8 / 71.4
+ **Q-Frame**	47.3 / 51.4	46.5 / 54.5	61.9 / 64.3	36.6 / 36.6	51.1 / 59.0	60.1 / 72.4
GPT-4o [21]	62.6 / 63.3	60.6 / 64.0	64.6 / 66.7	36.6 / 46.6	56.8 / 60.4	81.4 / 83.3
+ **Q-Frame**	64.9 / 66.0	60.6 / 65.6	67.0 / 70.6	38.8 / 44.6	67.2 / 67.6	83.0 83.3
Qwen2-VL-Video [27]	51.0 / 55/6	53.0 / 56.8	62.8 / 64.6	30.2 / 35.1	56.8 / 63.3	65.6 / 79.6
Qwen2-VL [27]	50.5 / 56.2	54.9 / 57.4	65.9 / 67.3	30.6 / 37.3	56.8 / 65.5	65.9 / 79.6
+ **Q-Frame**	53.3 / 56.5	59.2 / 61.3	69.8 / 70.4	38.8 / 42.9	68.3 / 67.6	71.5 / 81.7

Table 11. Performance of subtasks on Video-MME. Red fonts represent positive results compared to the Baseline, and blue fonts represent negative results.

For MLVU, Q-Frame significantly boosts results in non-uniform video understanding tasks such as NQA and PQA, demonstrating its effectiveness in capturing relevant frames. However, performance variations are observed in some retrieval-based tasks, likely due to the nature of the pre-trained vision-language model guiding frame selection.

In Video-MME, Q-Frame significantly improves perception, recognition, and OCR tasks, where frame relevance and resolution are critical. It also enhances reasoning and counting accuracy, though performance gains tend to be task-dependent.

Overall, these results confirm that Q-Frame effectively boosts Video-LLM performance by improving query-aware frame selection and resolution adaptation. The most significant impact is noted in tasks that require fine-grained temporal and spatial understanding.

B.4. Additional Ablation Study

CLIP-like Model: CLIP's text encoder uses an absolute positional embedding restricted to 77 tokens, creating a strict limit on input token numbers [22]. Therefore, to encode longer queries, we utilize Long-CLIP [37] to overcome this

Model	Acc(%)
-	53.5
CLIP [22]	57.8
SigLIP [34]	57.9
Long-CLIP [37]	**58.4**

Table 12. Ablation study of the CLIP-like Model in CQR.

τ	Acc(%)
0.5	57.6
0.8	**58.4**
1.0	58.2
1.2	58.0

Table 13. Ablation study of the temperature parameter τ in QFS.

limitation. Table 12 compares different CLIP-like models for Cross-modal Query Retrieval (CQR). We observe that replacing the baseline model ("-" with no dedicated retrieval module) with standard CLIP [22] improves accuracy from 53.5% to 57.8%, indicating the importance of a robust vision-language alignment for retrieving semantically relevant frames. SigLIP [34] further boosts performance to 57.9%, suggesting that more advanced training objectives can better capture cross-modal similarities. Notably, Long-CLIP [37] achieves the highest accuracy of 58.4%, demonstrating that architectures specifically tailored for handling longer inputs and richer contexts are particularly beneficial for video retrieval. Overall, these results highlight the crucial role of a well-optimized vision-language backbone in effectively guiding frame selection for Q-Frame.

Temperature Parameter: Table 13 shows the effect of varying the temperature parameter τ in Query-adaptive Frame Selection (QFS). Lower values of τ (e.g., 0.5) create a more peaked distribution, which may overlook relevant frames due to excessive exploitation, leading to a slightly lower accuracy of 57.6%. As τ increases, the model achieves a better balance between exploration and exploitation, reaching a maximum accuracy of 58.4% at $\tau = 1.2$. These findings indicate that a moderate level of randomness introduced by the Gumbel-Max trick is crucial for identifying diverse frames; however, an excess of randomness can weaken the selection of important frames. Therefore, adjusting τ is vital for optimizing QFS performance within Q-Frame.

Candidate Frames: We set the number of candidate frames T to 128 to ensure comparability with FRAME-VOYAGER [33]. Moreover, we conduct an ablation study to verify the impact of the number of candidate frames on Q-Frame. As shown in Table 14, we fix the number of sampled frames to 8, and for different experimental settings of the number of candidate frames, Q-Frame can achieve stable improvement.

Candidate Frames	Sampled Frames	Latency(ms)		LongVideoBench
		Embedding	Sampling	
64	8	76.7	87.1	56.8
	4 + 8 + 32	73.5	89.4	57.7
128	8	123.8	178.6	57.6
	4 + 8 + 32	120.1	182.8	58.4
256	8	218.0	361.1	57.8
	4 + 8 + 32	220.6	365.6	58.6

Table 14. Ablation study of candidate frames and overhead.

Overhead: Although Q-Frame doesn't require additional training, it introduces preprocessing overhead in the inference phase. The ablation study of the overhead of Q-Frame is shown in Table 14. The overhead of Q-Frame focuses on calculating embeddings and performing sampling (including resolution allocation). As candidate frame embeddings can be computed in parallel, the computational cost of this part is relatively manageable. Additionally, as the number of candidate frames increases, the burden brought by MRA is almost negligible. Compared to Video-LLM's processing time, Q-Frame can introduce about 5 times more effective frames with minimal impact on time.

B.5. Additional Case Analysis

We present a comparison visualization of uniform sampling and our Q-Frame. As illustrated in Figures 6, 7, and 8, Q-Frame more accurately identifies the video frame where the answer is located, providing a more reliable source of information for Video-LLMs. Furthermore, due to the introduction of the Gumbel-Max trick, the sampling in QFS continues over time, which solves the sparsity problem caused by sampling to some extent.

Bad Case Analysis: However, as illustrated in Figure 9, neither our Q-Frame nor the uniform sampling can correctly handle such a temporal reasoning task. By design, Q-Frame selects a sparse set of frames based on their semantic relevance to the query, guided by a CLIP-based matching score. This selection process doesn't preserve the sequential structure or causal relationships between events that are critical for effective temporal reasoning. Consequently, for the temporal reasoning task, even though the selected frames may be individually relevant, they often fail to capture the transitional moments or event boundaries necessary to reconstruct the full temporal logic required to answer the query accurately.

Question: **What's the color of Gabriel García Márquez's car in the video?**

A. Green.

B. Black.

C. Orange.

D. Red.

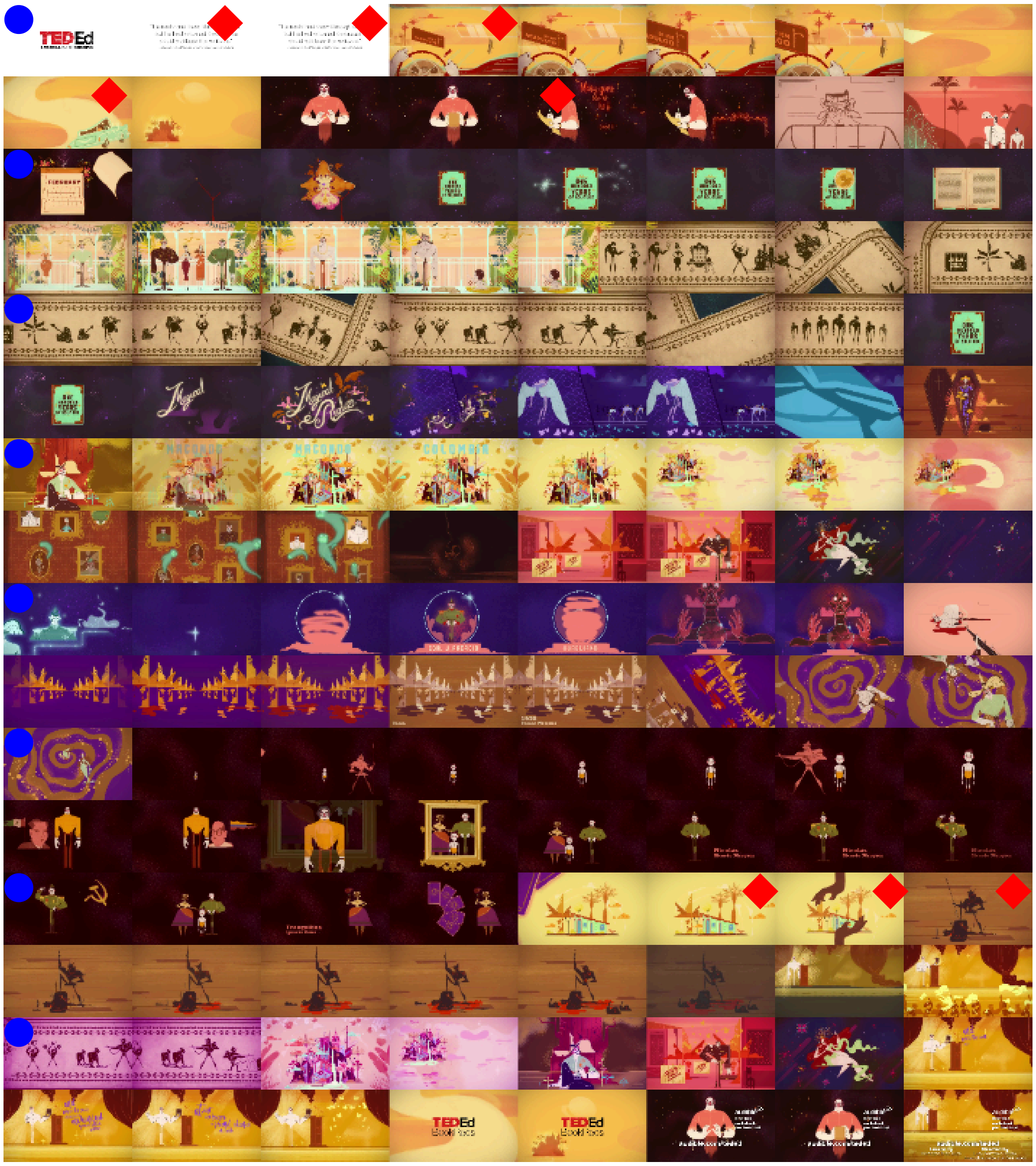

Figure 6. Case analysis from Video-MME. ●: uniform sampling; ◆: Q-Frame. Video-LLMs with Q-Frame can capture the keyframes of the car and provide answers correctly.

Question: **Which player finished the dunk in the video?**

A. Number 13.

B. Number 32.

C. Number 17.

D. Number 5.

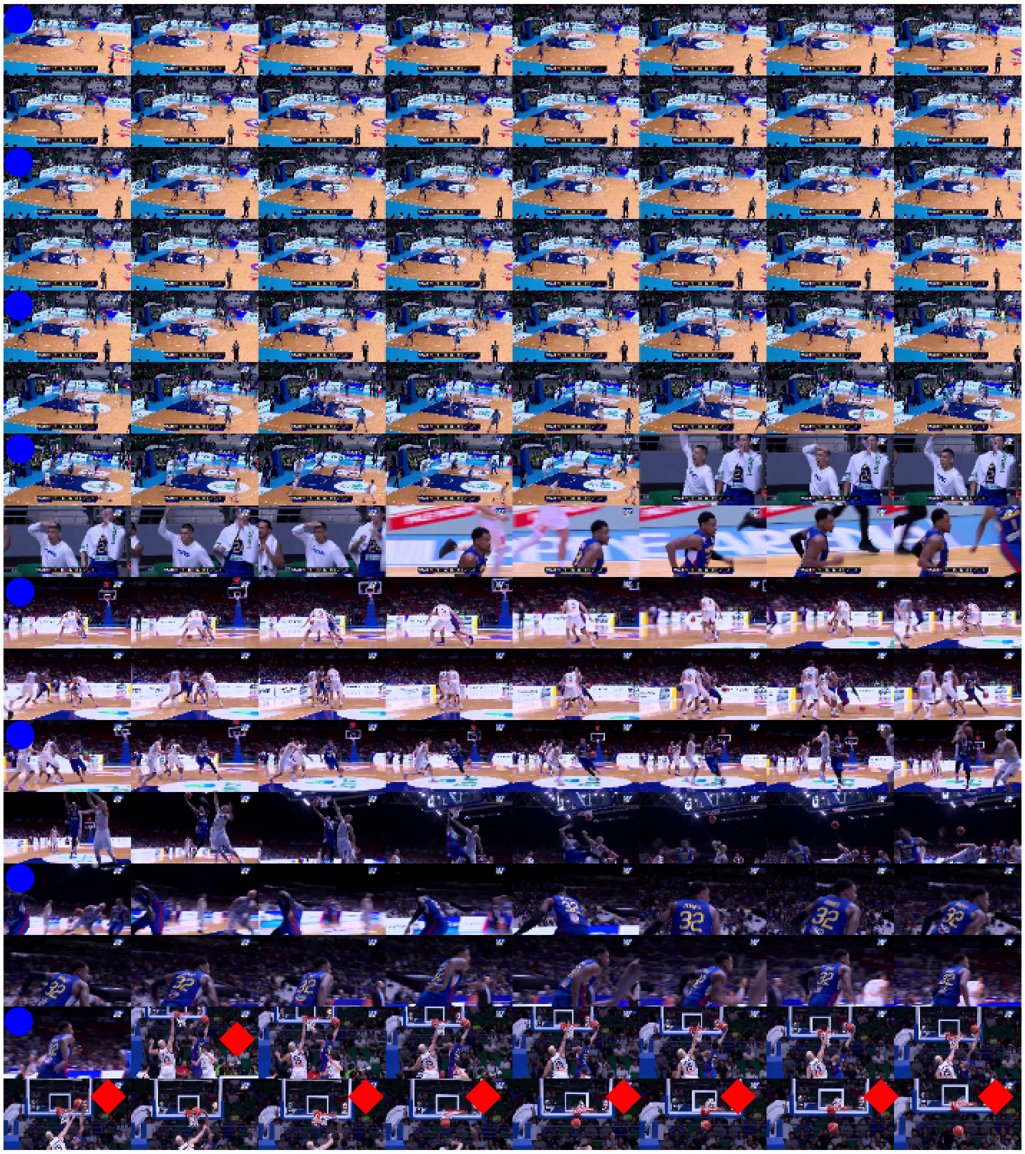

Figure 7. Case analysis from Video-MME. ●: uniform sampling; ◆: Q-Frame. Video-LLMs with Q-Frame can capture the keyframes of the action of dunking and respond accordingly.

Question: **What is the color of the pants worn by the young girl who converses with the elderly man on the subway?**

A. Light brown.

B. Black.

C. Pink.

D. Blue gray.

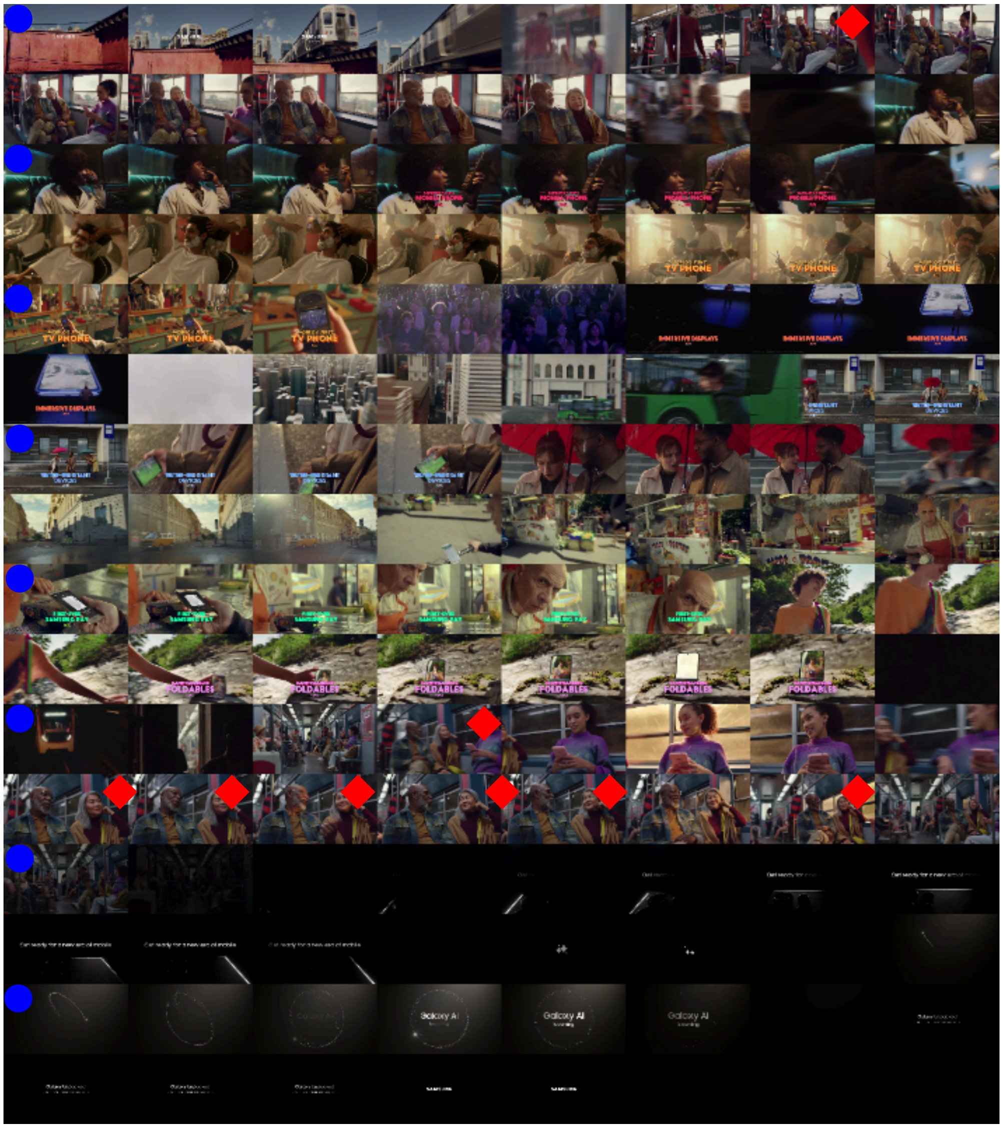

Figure 8. Case analysis from Video-MME. ●: uniform sampling; ◆: Q-Frame. Video-LLMs with Q-Frame can capture the keyframes of the young girl and the elderly man and respond accordingly.

***Question*: How often do the people in the video take water breaks per workout?**

A. 10 minutes.

B. 5 minutes.

C. 15 minutes.

D. No regular breaks.

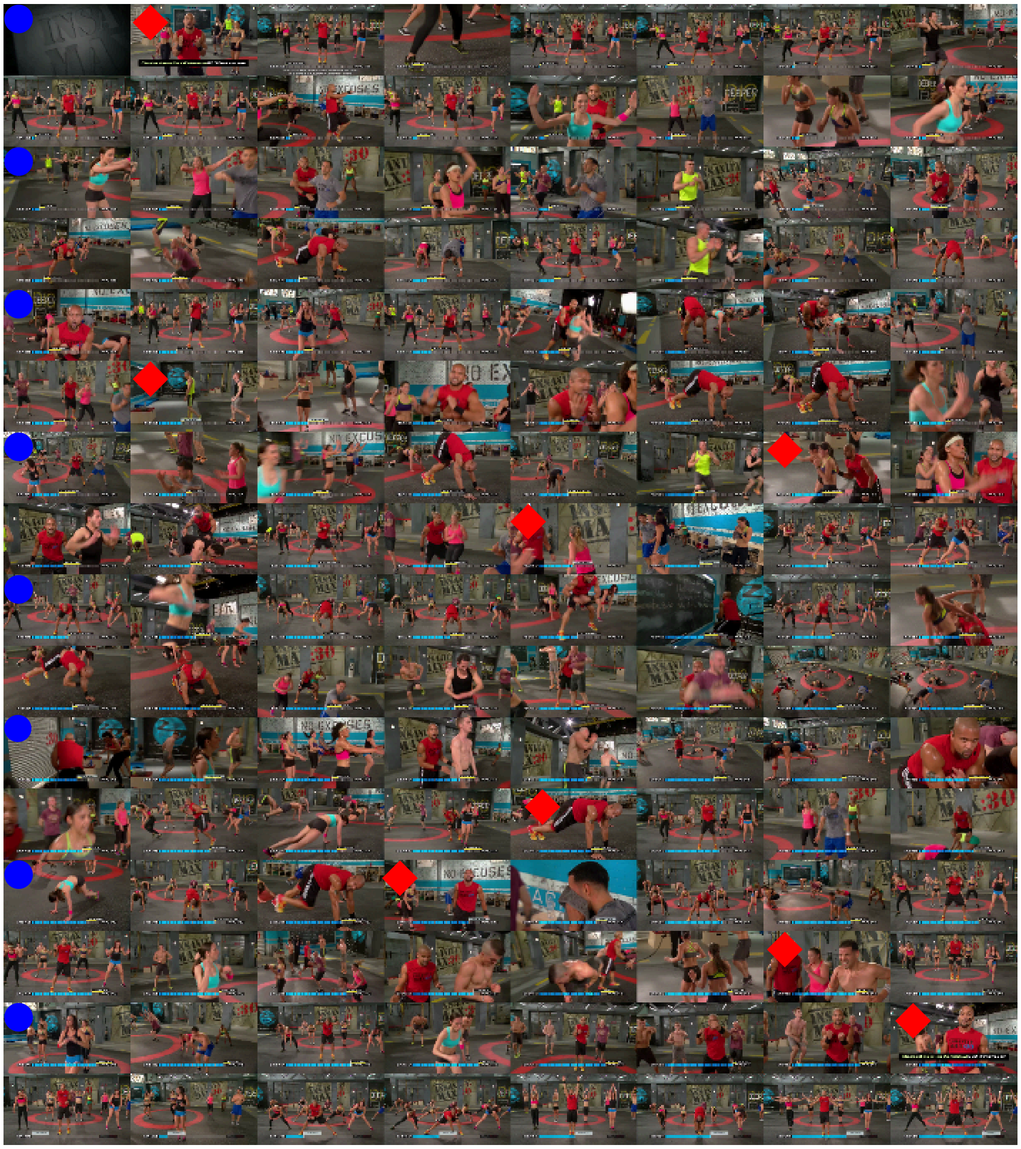

Figure 9. Case analysis from Video-MME. ●: uniform sampling; ◆: Q-Frame. Neither our Q-Frame nor the uniform sampling can correctly handle such a temporal reasoning task.