QCA: Query- and Content-Aware Keyframe Selection for Long Video Understanding

Jun Peng[†1], Baiyang Song[†1], Jie Li[1], Hui Li[1], Yiyi Zhou[1], Rongrong Ji[1], and Yonghong Tian[*2,3]

[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, P.R. China
[2] School of Electronic and Computer Engineering, Peking University, P.R. China
[3] Peng Cheng Laboratory, Shenzhen, P.R. China
{pengjun.cn, lijie.32}@outlook.com, songbaiyang@stu.xmu.edu.cn, {hui, zhouyiyi, rrji}@xmu.edu.cn, yhtian@pku.edu.cn

Abstract. Video understanding is often plagued by severe temporal redundancy, where processing dense frame sequences is both semantically inefficient and computationally expensive. This challenge is further amplified when only a small subset of frames is truly relevant to the given query. In this paper, we propose a Query- and Content-Aware (QCA) keyframe selection framework that can select a compact yet information-rich set of frames from long videos. QCA first partitions the video into temporal segments and estimates the information contribution of each segment by jointly modeling query relevance and content deviation, and dynamically allocates keyframe budget to each segment. Within each segment, QCA anchors on the most query-relevant frame and iteratively incorporates additional frames to maximize diversity while maintaining high semantic relevance to the query. Crucially, our method requires no additional training and can be seamlessly integrated into existing Video-LLMs. Extensive experiments across multiple long video understanding benchmarks demonstrate that our proposed approach achieves state-of-the-art performance and has strong generalization ability. For instance, QCA achieves 67.8% on LongVideoBench using 128 frames, while GPT-4o achieves 66.7% using 256 frames. Our codes are available in GitHub.

Keywords: Video Understanding · Keyframe Selection · Training-Free

1 Introduction

Large Language Models (LLMs) [6,11,35] and Multimodal LLMs (MLLMs) [3,21,31] have recently demonstrated remarkable capabilities in unified vision-language understanding, enabling complex reasoning over images and videos [3, 17, 31]. By leveraging large-scale pre-training, these models have significantly advanced in a wide range of video understanding, *e.g.*, video question answering [12], retrieval [33], and grounding [24].

[†]Equal contribution. [*]Corresponding author.

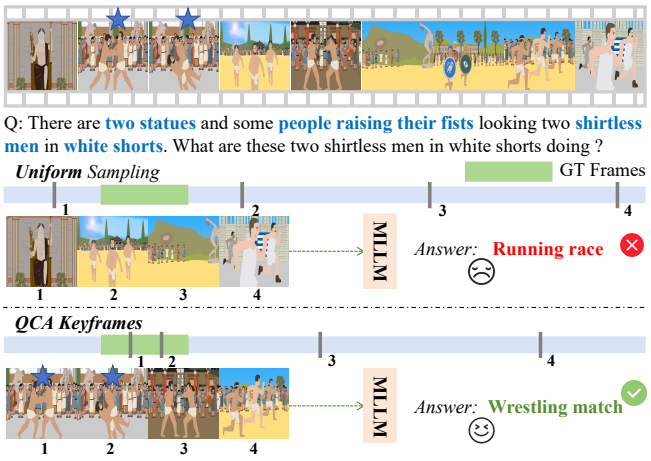

Fig. 1: Uniform sampling may overlook semantically critical moments due to the temporal redundancy in videos, whereas our query- and content-aware selection prioritizes informative and diverse frames, enabling more reliable long video understanding under limited frame budgets.

However, despite their strong representational power, applying LLMs or MLLMs to long-form video remains a challenge. Long videos are inherently temporally redundant, and naively encoding dense frame sequences leads to substantial computational inefficiency. More importantly, when video frames are converted into visual tokens, long videos can easily exceed the token budget limitations of current MLLMs, resulting in truncated input or degraded reasoning performance. Under such constraints, selecting a small yet informative subset of frames is often more effective than processing dense but redundant frame sequences.

Existing approaches typically address this issue through uniform sampling [40], attention- or relevance-based selection [2, 10]. However, uniform sampling often overlooks semantically critical moments, as illustrated in Fig.1. Attention-based mechanisms offer fine-grained modeling, but they remain computationally expensive for long videos. And relying solely on query-frame similarity may overlook video content diversity, leading to redundant frame selection or insufficient information.

Therefore, effective long video understanding with MLLMs requires query- and content-aware keyframe selection. Intuitively, different temporal segments of a video contribute unequally to a given query, and selected frames should be both semantically relevant and content-diverse within each segment.

Motivated by this insight, we propose a **Q**uery- and **C**ontent-**A**ware keyframe selection framework, which is denoted as **QCA**. It first estimates the information contribution of each temporal segment by jointly modeling the semantic relevance and the content deviation. Then dynamically allocate keyframe budgets across segments, followed by an intra-segment keyframe selection strategy that

balances semantic alignment and content diversity. Moreover, QCA is training-free and can be seamlessly integrated into existing Video-LLMs.

Extensive experiments on multiple long video understanding benchmarks show that our proposed QCA achieves the SOTA performance, *e.g.*, 66.9% on LongVideoBench [32] and 70.1% on Video-MME [12]. Besides, it achieves average performance gains of 4.00% on LLaVA-Video [40] and 4.78% on Qwen3-VL [3], validating its generalization ability across different cutting-edge Video-LLMs.

Our main contributions are as follows:

- We formulate query-conditioned keyframe selection for long video understanding as a joint relevance–diversity modeling and frame allocation problem under a limited frame budget.
- We propose QCA, a query- and content-aware keyframe selection framework that integrates segment-level contribution estimation, adaptive budget allocation, and anchor-centric greedy frame selection.
- Extensive experiments demonstrate consistent improvements across benchmarks, Video-LLMs, and VL embeddings, demonstrating its superiority and generalization ability.

2 Related Work

2.1 MLLMs for Video Understanding

Multimodal Large Language Models (MLLMs) have recently achieved significant progress in video understanding by extending LLMs with visual perception modules. Early works such as Flamingo [1] demonstrated the effectiveness of aligning pretrained vision encoders with frozen language models for multimodal reasoning. Subsequent models, including LLaVA [21] and its variants, further improved vision–language alignment through instruction tuning, achieving strong performance across a wide range of multimodal tasks.

Building on these advances, recent studies have adapted MLLMs to video understanding by representing videos as sequences of visual tokens extracted from sampled frames. Representative approaches, such as Video-LLaMA [37] and VideoChat [19], allow MLLMs to process temporal visual information together with textual queries, enabling applications including video question answering, retrieval, and captioning.

However, most existing Video-LLMs rely on uniform frame sampling to compress videos into a limited set of frames. While simple and efficient, this strategy often struggles with long-form videos due to severe temporal redundancy and the limited context length of language models. Consequently, naively increasing the number of input frames leads to a higher computational cost or truncated visual context. These limitations motivate the need for more efficient video representations, such as query-aware keyframe selection.

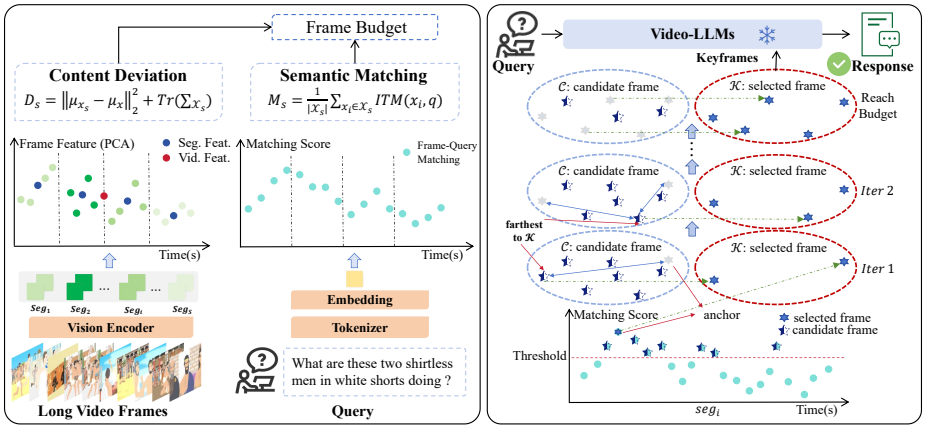

Fig. 2: The proposed QCA consists of: (a) Inter-Segment Keyframe Allocation, where the frame budget for each segment is dynamically determined by considering semantic alignment and visual content deviation; (b) Intra-Segment Keyframe Selection, which iteratively adds the frame with the maximum aggregate distance to the current keyframe set $\mathcal{K}_s$ from a relevance-filtered candidate set $\mathcal{C}_s$.

2.2 Keyframe Selection for Video-LLMs

To address the scalability challenges of video inputs, extensive research has focused on designing efficient video representations for Video-LLMs. A common strategy is uniform sampling [17, 40], which reduces the number of frames while maintaining coarse temporal coverage. However, such methods are query-agnostic and often select redundant or irrelevant frames for specific queries.

In addition, some works [13, 20] have explored relevance-based frame filtering, where frames are ranked according to their semantic similarity to a given query and only the top-ranked frames are retained. While these approaches improve query awareness, they typically overlook the intrinsic structure and diversity of video content, leading to suboptimal coverage of important temporal segments. Clustering- and diversity-based methods [2, 25, 39] have also been investigated to select representative frames, but they are usually query-independent and may fail to capture query-specific evidence.

More recently, several studies have proposed segment-based video representations to better handle long videos under limited frame or token budgets. These methods aim to summarize or compress video content before feeding it into language models, for example, through temporal segmentation [26, 42], adaptive keyframe selection [9, 22, 27], or token pruning [5, 7, 16].

In contrast to prior works, our approach jointly models query relevance and content structure to select a compact yet informative set of keyframes. By dynamically allocating a limited frame budget across temporal segments and performing semantic-anchored and diversity-aware selection within each segment,

our method provides an efficient video representation that is particularly well-suited for long video understanding operating under frame budget constraints.

3 Method

Rather than introducing new frame-scoring primitives, our goal is to provide a structured framework that integrates query relevance, content representativeness, and budget allocation for efficient keyframe selection in long videos. Fig.2 illustrates the overview of our proposed QCA, which first adaptively assigns the keyframe budget to different temporal segments by modeling semantic matching and content deviation. QCA then iteratively selects keyframes from the candidate set, satisfying both semantic alignment with the query and semantic diversity between selected keyframes.

3.1 Problem Formulation

Given a video uniformly sampled at 1fps, the resulting frame sequences are denoted as

$$\mathcal{X} = \{x_1, x_2, ..., x_N\}, \tag{1}$$

where N is the total number of frames. Given a query q, a natural language question, our goal is to select a compact keyframe subset

$$\mathcal{K} \subset \mathcal{X}, |\mathcal{K}| = N', N' \ll N \tag{2}$$

such that $\mathcal{K}$ preserves the most query-relevant and informative visual content for the downstream video understanding task:

$$answer = MLLMs(\mathcal{K}, q) \tag{3}$$

3.2 Temporal Segmentation

To capture coarse temporal structure, we divide the video $\mathcal{X}$ into S uniformly spaced segments:

$$\mathcal{X} = \bigcup_{s=1}^{S} \mathcal{X}_s, \mathcal{X}_s \cap \mathcal{X}_{s'} = \emptyset. \tag{4}$$

where S is set to 12 by default. Intuitively, a larger S provides finer temporal granularity but reduces the number of frames allocated per segment, while a smaller S may overlook short but critical events.

3.3 Inter-Segment Keyframe Allocation

Semantic Matching. We first measure the average semantic matching degree $M_s \in [0, 1]$ between frames in $\mathcal{X}_s$ and the given query:

$$M_s = \frac{1}{|\mathcal{X}_s|} \sum_{x_i \in \mathcal{X}_s} ITM(x_i, q), \tag{5}$$

where $ITM(\cdot,\cdot)$ denotes Image-Text Matching function, *e.g.*, BLIP-2 [18]. This term refers to the relevant video segment to the query at the semantic level.

Content Deviation. We define the segment content deviation as

$$D_s = \|\mu_{\mathcal{X}_s} - \mu_{\mathcal{X}}\|_2^2 + Tr(\Sigma_{\mathcal{X}_s}), \tag{6}$$

where $\mu_{\mathcal{X}_s}$ and $\Sigma_{\mathcal{X}_s}$ denote the mean feature and the covariance matrix of $\mathcal{X}_s$, respectively. The first term measures how much a segment differs from the whole video, encouraging the selection of visually distinctive segments that are more likely to contain informative events. The second term, trace of the covariance matrix, captures intra-segment variation, where larger values indicate rich visual diversity. In practice, both terms are normalized to mitigate scale mismatch.

Keyframe Allocation. The final information contribution score of segment $\mathcal{X}_s$ is defined as a weighted sum:

$$c_s = \alpha \cdot M_s + \beta \cdot D_s, \tag{7}$$

where α and β are hyperparameters. And we respectively apply a softmax normalization to M_s and D_s across segments to ensure comparable scales before fusion. We compute the contribution weights via

$$w_s = \frac{{c_s}^{\tau}}{\sum_{j=1}^{S} {c_j}^{\tau}}, \tag{8}$$

where τ controls the smoothness of allocation. Then the number of keyframes allocated to segment $\mathcal{X}_s$ is

$$q_s = \lfloor w_s \cdot N' \rfloor, \sum_{i=1}^{S} s_i = N^{'} \tag{9}$$

Since the floor operation may yield fewer frames than the target budget, the remaining frames are assigned to the highest-scoring segments. Specifically, if a deficit of n Specifically, if a deficit of n highest-scoring segments, ensuring that the final frame count matches the target budget.

3.4 Intra-Segment Keyframe Selection

With each segment $\mathcal{X}_s$, we anchor the most query-relevant frame and iteratively incorporate additional frames to maximize diversity while maintaining high semantic relevance.

Semantic Anchor. We first select the frame with the highest relevance to the query that is added to the keyframe set $\mathcal{K}_s$

$$\mathcal{K}_s = \{\arg\max_{x_i \in \mathcal{X}_s} ITM(x_i, q)\}. \tag{10}$$

Semantic Constraint. Let R^* denote the matching score of the anchor frame. We construct a candidate set

$$\mathcal{C}_s = \{x_j \in \mathcal{X}_s | ITM(x_j, q) \geq \gamma \cdot R^*\}, \tag{11}$$

in which γ controls the size of the candidate pool by filtering frames whose relevance is below a fraction of the anchor score. Smaller γ increases diversity but introduces noise, while larger γ improves relevance but reduces candidate diversity. We set $\gamma = 0.7$ by default.

Content Diversity. We iteratively expand $\mathcal{K}_s$ by selecting a frame that maximizes its distance to the current keyframe set:

$$\mathcal{K}_s = \mathcal{K}_s \cup \{\arg \max_{x_i \in \mathcal{C}_s \setminus \mathcal{K}_s} \sum_{x_j \in \mathcal{K}_s} \phi(x_i, x_j)\}, \tag{12}$$

where ϕ is a distance metric, *e.g.*, Euclidean distance. The process repeats until $|\mathcal{K}_s| = q_s$. Notably, γ can be reduced to broaden the range of the candidate set if $|\mathcal{K}_s| + |\mathcal{C}_s| < q_s$ initially. The final keyframe set is obtained by aggregating selections from all segments:

$$\mathcal{K} = \bigcup_{s=1}^{S} \mathcal{K}_s, \ |\mathcal{K}| = N' \tag{13}$$

The greedy strategy incrementally selects frames that maximize marginal information gain, allowing efficient approximation of diverse and representative keyframe sets.

4 Experiments

4.1 Evaluation Benchmarks

To evaluate the effectiveness of our proposed method, we conduct experiments on four widely used benchmarks for long video understanding.

LongVideoBench [32] highlights the referred reasoning questions that are posed on varying-length videos up to an hour long on diverse themes. These questions are dependent on long frame input and cannot be well-addressed by a single frame or a few sparse frames.

Video-MME [12] (w/o subs) spans 6 primary visual domains with 30 subfields to ensure broad scenario generalizability and encompasses both short-, medium-, and long-term videos, ranging from 11 seconds to 1 hour.

MLVU [41] is constructed from a wide variety of long videos, with lengths ranging from 3 minutes to 2 hours, and includes nine distinct evaluation tasks.

LVBench [30] comprises publicly sourced videos, including TV series, sports broadcasts, and everyday surveillance footage, and encompasses a diverse set of tasks aimed at long video comprehension and information extraction.

4.2 Implementation Details

We investigate LLaVA-Video [40], InternVL-3.5 [31] and Qwen3-VL [3] as our baselines. These MLLMs typically receive 64 uniformly sampled video frames, which may result in the loss of key information. For a fair comparison, we also

Table 1: Performance comparison with existing keyframe selection methods on LongVideoBench, Video-MME, MLVU, and LVBench. We evaluate our approach against uniform sampling, Top-k selection, which selects the most image-text matched frames, and recent frame selection baselines across three MLLM backbones: LLaVA-Video-7B, InternVL-3.5-8B, and Qwen3-VL-8B with 64 frames budget. † denotes our re-implementation. **Bold** indicates the best performance for each backbone. "-" denotes unavailable results. "*" denotes trained.

Method	LLM Size	Embedding Model	LongVideo Bench	Video MME	MLVU	LVBench
LLaVa-Video-7B						
+ *Uniform*	7B	-	58.9	64.4	70.8	41.9
+ *Top-k*	7B	BLIP	61.6	63.7	72.8	47.2
+ *AKS* [27]	7B	BLIP	62.7	65.3	71.8	47.6
+ *Q-Frame*† [38]	7B	LongCLIP	61.5	64.7	72.9	47.1
+ *OneClip-RAG* [8]	7B	CLIP*	62.5	65.2	71.2	-
+ *BOLT* [22]	7B	CLIP	62.2	64.6	70.3	-
+ *FRAG* [14]	7B	MLLM(7B)	60.6	63.7	69.2	-
+ *E-VRAG* [34]	7B	LLM+CLIP	**63.1**	65.4	70.2	-
+ *QCA (Ours)*	7B	BLIP	62.9 (4.0↑)	**66.1** (1.7↑)	**74.1** (3.3↑)	**48.9** (7.0↑)
InternVL-3.5-8B						
+ *Uniform*†	8B	-	61.3	61.9	69.9	42.8
+ *Top-k*	8B	BLIP	62.5	61.6	70.4	48.7
+ *AKS*† *[27]*	8B	BLIP	62.9	62.8	70.5	47.9
+ *Q-Frame*† [38]	8B	LongCLIP	62.5	61.7	70.6	48.9
+ *OneClip-RAG*† *[8]*	8B	CLIP*	62.1	61.7	70.2	-
+ *QCA (Ours)*	8B	BLIP	**63.5** (2.2↑)	**63.9** (2.0↑)	**71.3** (1.4↑)	**50.0** (7.2↑)
Qwen3-VL-8B						
+ *Uniform*†	8B	-	63.1	67.6	71.0	43.8
+ *Top-k*	8B	BLIP	64.4	67.4	74.2	50.7
+ *AKS*† *[27]*	8B	BLIP	64.7	68.6	74.2	50.8
+ *Q-Frame*† [38]	8B	LongCLIP	65.8	67.9	74.7	50.7
+ *OneClip-RAG* † [8]	8B	CLIP*	65.3	68.5	71.3	-
+ *QCA (Ours)*	8B	BLIP	**66.9** (3.8↑)	**69.5** (1.9↑)	**75.7** (4.7↑)	**51.8** (8.0↑)

set the number of retrained frames $N' = 64$ in QCA. Following previous research [27], we first sample the frame sequence from the raw video at 1 FPS before performing keyframe selection to reduce computational costs.

For the Image-Text Matching (ITM) function in Eq.5 and the Content Deviation term in Eq.6, we both use BLIP-2 [18] as the default embedding model, in which the text and visual encoders receive the query and frame to measure their matching degree.

Also, we set the weighting coefficients for semantic relevance and content deviation to $\alpha = \beta = 0.5$, and use a unified softmax temperature $\tau = 0.5$ for all softmax operations. All experiments are conducted on 8×A800 80G GPUs.

4.3 Quantitative Results

Comparison to existing keyframe selection methods. Across four long video understanding benchmarks, Tab.1 presents a comprehensive comparison between QCA and several frame selection strategies, including uniform sampling, Top-k selection and recent keyframe selection methods, where Top-k selects keyframes with the highest Top-k image-text matching score.

Table 2: Performance comparison on LongVideoBench, Video-MME, MLVU, and LVBench with state-of-the-art models using uniform sampling across two backbones: Qwen3-VL-8B and Qwen3-VL-30B-A3B-Instruct. † denotes our re-implementation, **bold** means the highest accuracy on different MLLM, and "-" indicates unavailable results.

Method	LLM Size	Frames	LongVideoBench	Video-MME	MLVU	LVBench
GPT-4o [15]	-	256 / 0.5fps	66.7	71.9	64.6	-
Gemini-1.5-Pro [28]	-	-	64.0	75.0	-	33.1
Qwen2.5-VL-72B [4]	72B	1fps	60.7	**73.3**	74.6	47.3
Kimi-VL-16B-A3B [29]	16B	64	64.5	67.8	74.2	-
LLaVA-Video-72B [40]	72B	64	63.9	70.0	74.4	45.5
InternVL3.5-38B [31]	38B	64	65.7	70.9	77.0	-
NVILA [23]	7B	256	57.7	64.0	70.1	-
mPLUG-Owl3 [36]	8B	128	59.7	59.3	70.0	43.5
Apollo [43]	7B	2fps	58.5	61.3	68.7	-
Qwen3-VL-8B	8B	64	63.1	67.6	71.0	43.8
+*QCA (Ours)*	8B	64	66.9	69.5	75.7	51.8
Qwen3-VL-30B-A3B	30B	64	67.2	69.9	72.8	44.0
+*QCA (Ours)*	30B	64	**69.9**	71.4	**77.1**	**52.6**

Without additional training, QCA consistently achieves the best performance across representative Video-LLMs, including LLaVA-Video-7B, InternVL-3.5-8B, and Qwen3-VL-8B, demonstrating strong generalization and robustness. In particular, with Qwen3-VL-8B under a 64-frame budget, QCA achieves 66.9% (3.8% ↑) on LongVideoBench, 69.5% (1.9% ↑) on Video-MME, 75.7% (4.7% ↑) on MLVU, and 51.8% (8.0% ↑) on LVBench. Although the absolute gains appear moderate, they are obtained under strict frame budgets and without additional training, making them particularly valuable for practical long video understanding scenarios.

Compared with both uniform sampling and Top-k selection, QCA consistently delivers superior results across diverse tasks requiring long-range reasoning, semantic localization, and holistic video understanding. Specifically, on LongVideoBench, which emphasizes semantically critical moments, the proposed QCA significantly outperforms uniform sampling, highlighting the limitation of uniform sampling in capturing key events. Meanwhile, on Video-MME, which requires more comprehensive scene understanding, QCA also surpasses Top-k selection. This suggests that selecting frames solely based on high ITM scores may overly focus on locally salient moments while overlooking broader contextual information.

We further compare QCA with recent keyframe selection methods, *e.g.*, AKS [27] and Q-Frame [38]. Our approach almost outperforms these methods across all four benchmarks. Specifically, QCA outperforms AKS by 2.1% on LVBench and outperforms Q-Frame by 2.2% on Video-MME using InternVL-3.5-8B. Compared with FRAG [14] and E-VRAG [34], which rely on larger models for frame selection, QCA achieves comparable or superior performance. For instance, QCA gains a 2.4% improvement on Video-MME over FRAG and a 3.9% improvement on MLVU over E-VRAG using LLaVA-Video-7B. More

Fig. 3: Qualitative comparison of our QCA (left) and Uniform Sampling (right). The green boxes indicate the critical frames successfully selected by our method, which contain the specific evidence required to answer the questions (*e.g.*, the passing dog, the white beard). In contrast, Uniform Sampling fails to capture these informative moments due to its rigid temporal intervals, leading to incorrect answers.

surprisingly, although OneClip-RAG [8] benefits from additional training, QCA still achieves superior performance without any fine-tuning, *e.g.*, gains a 2.9% improvement on MLVU using LLaVA-Video-7B. Importantly, the performance gains of QCA remain consistent across different Video-LLMs, indicating that the proposed method is model-agnostic and can be readily integrated into existing architectures.

Comparison to state-of-art Video-LLMs. Tab. 2 compares QCA with state-of-the-art Video-LLMs using Qwen3-VL-30B-A3B-Instruct and Qwen3-VL-8B as backbones. Although the Qwen3-VL series already achieve strong performance with uniform sampling of 64 frames, incorporating QCA further improves results across all benchmarks.

For example, with Qwen3-VL-30B-A3B-Instruct, QCA outperforms GPT-4o by 12.5% on MLVU and surpasses Gemini-1.5-Pro by 19.5% on LVBench, demonstrating that more informative frame selection can substantially enhance long video reasoning even for strong proprietary models.

Overall, these results indicate that QCA offers a simple yet effective approach for improving long video understanding by selecting a compact, informative, and diverse set of keyframes, without modifying model architectures or requiring additional training.

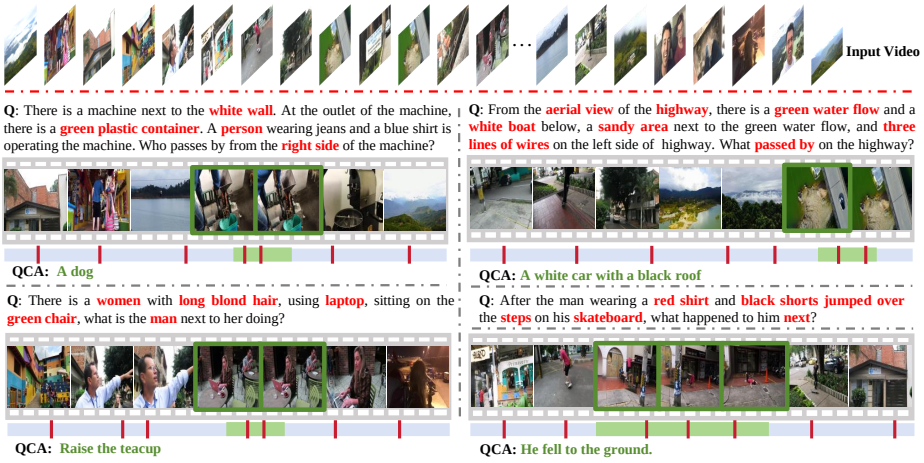

Fig. 4: For the same video, QCA selects different keyframes based on different queries, while ensuring video coverage to prevent missing key information.

4.4 Qualitative Results

Fig.3 and Fig.4 present representative video question answering examples comparing our QCA with uniform sampling. Unlike uniform sampling, which selects a fixed set of frames regardless of the query, QCA dynamically adapts keyframe selection to the question while preserving sufficient global scene coverage.

Beyond identifying frames that are strongly aligned with the query, QCA also retains context frames that provide complementary visual information, even when their direct relevance to the query is weaker. This helps maintain holistic scene understanding under a strict frame budget and prevents errors caused by missing contextual cues. As illustrated, QCA captures key evidence such as the passer-by and salient objects while still preserving the overall scene context needed for disambiguation.

Fig.5 further visualizes the keyframe selection of QCA. The top-right plot shows the frame-query matching score, while the bottom-right plot illustrates the content deviation at the feature level. Based on the average matching score and feature deviation within each segment, QCA adaptively allocates different frame budgets across segments. The resulting keyframes jointly satisfy semantic relevance, content diversity, and temporal coverage.

4.5 Ablation Study

Matching and Deviation. In Tab.3, we first analyze the impact of semantic matching and content deviation by removing them from the full setting. Removing either component consistently degrades performance across all benchmarks, indicating that both query relevance and content representativeness are important for effective keyframe selection.

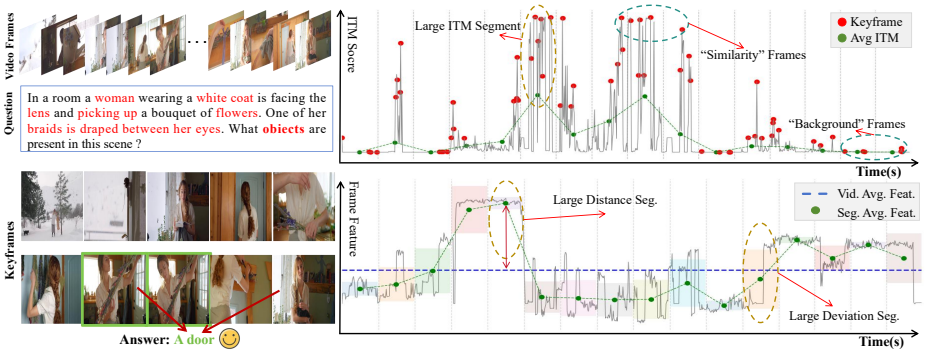

Fig. 5: An example of our QCA keyframe selection. The upper right visualizes the match score between each frame and query, while the bottom right shows the changes in frame content at the feature level. The red dot indicates the selected keyframe, and the green dot indicates the average ITM score or average frame feature within a segment. The shaded area represents the standard deviation within the segment.

Table 3: Ablation study on different components and settings. Results are reported on LongVideoBench, Video-MME, MLVU, and LVBench with Qwen3-VL.

Setting	Matching	Deviation	Anchor	Candidate	Diversity	LongVideoBench	Video-MME	MLVU	LVBench
Full	✓	✓	✓	✓	✓	**66.9**	70.1	**75.8**	**52.4**
w/o D_s	✓	✗	✓	✓	✓	66.3	**70.4**	75.4	51.0
w/o M_s	✗	✓	✓	✓	✓	65.8	69.5	75.1	51.1
w/o D_s, M_s	✗	✗	✓	✓	✓	66.0	68.3	75.2	51.2
w/o Anchor	✓	✓	✗ (random)	✗	✓	63.7	68.2	74.1	47.2
w/o Candidate Set	✓	✓	✓	✗ (all frames)	✓	64.5	69.0	72.4	48.7
w/o Diversity	✓	✓	✓	✓	✗ (top matching)	66.2	68.6	75.3	49.2
Uniform Selection	✓	✓	✗	✗	✗	62.7	67.6	71.8	43.9

Specifically, removing the matching term (or both terms) reduces performance from 66.9% to 65.8% on LongVideoBench and from 70.1% to 68.3% on Video-MME, highlighting the dominant role of semantic relevance in query-driven video understanding. Interestingly, the 2^{nd} row shows that using only M_s still achieves 70.4% on Video-MME, further confirming the importance of query-aware frame selection.

Fig.6 (middle) shows the sensitivity study in α, β and γ, where $\beta = 1 - \alpha$. We can observe that balanced relevance-diversity weighting generally provides the best trade-off. Regarding γ, a smaller γ could introduce a noisy candidate set, while a larger one leads to insufficient diversity.

Anchor, Candidate, and Diversity. We further examine the anchor-centric selection strategy and the construction of the candidate set $\mathcal{K}_s$, as shown in the bottom of Tab.3. When replacing the semantic anchor with a random frame, the candidate set is formed from the remaining frames in the segment. Performance degradation indicates that initializing the selection with the semantic anchor provides a strong semantic reference for subsequent selection.

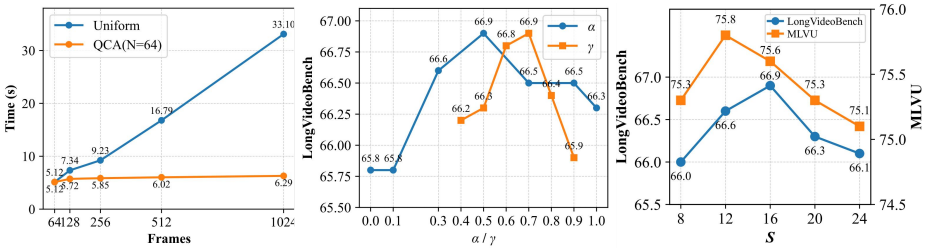

Fig. 6: Time cost (left), sensitivity study in α and γ (middle) and S(right).

Moreover, removing the candidate set and selecting frames from the entire segment leads to performance degradation, particularly on MLVU and LVBench. This result suggests that restricting the search space to high-relevance candidates helps suppress noisy frames and stabilize the diversity modeling process.

To further assess the role of diversity modeling, we replace the diversity-driven selection with a simple top-matching strategy. The performance drop across all benchmarks demonstrates that explicit diversity modeling is necessary to capture temporally dispersed yet semantically relevant events in long videos. In comparison, uniform selection causes substantial performance degradation, confirming that the improvements mainly stem from informed frame selection.

In summary, the components of QCA play complementary roles, *e.g.*, semantic matching ensures query relevance, deviation modeling captures representative content, and diversity modeling improves temporal coverage.

Temporal Segmentation. The number of temporal segments S directly affects the effectiveness of the segment-level scoring. As shown in Fig. 6 (right), the performance first improves and then decreases as S increases. On LongVideoBench, the accuracy rises from 66.0% at S=8 to a peak of 66.9% at S=16, and then drops when S becomes larger. A similar trend is observed on MLVU.

When S is too small, each segment covers a long temporal span and may contain multiple events, causing overly averaged statistics and inaccurate frame allocation. Conversely, when S is too large, each segment contains fewer frames, making the statistics less stable. Therefore, a moderate segmentation granularity provides a better trade-off.

In addition, we explore a simple dynamic strategy in which S increases linearly with video length. However, this naive strategy did not consistently outperform using a fixed S. Simply increasing the number of segments does not necessarily align with the video's underlying semantic structure. This indicates that QCA is relatively robust to a moderate fixed S.

Preprocessing Overhead. We also analyze the preprocessing overhead of QCA. As shown in Fig.6 (left), the keyframe selection process introduces only a marginal computational cost. For instance, as the number of input frames increases from 128 to 1024, the ITM matching time grows from 0.595s to 1.165s, while the selection process itself takes only about 0.005s. This overhead is sub-

Table 4: Comparisons with different VL embedding models.

Method	Embedding	LongVideoBench	VideoMME	MLVU	LVBench
Qwen3-VL	-	62.7	67.6	71.0	43.8
+*Top-k*	CLIP	64.1	67.5	73.2	50.4
+*Q-Frame* [38]		64.7	67.8	74.1	50.6
+*AKS* [27]		64.3	68.5	73.7	50.1
+*QCA (Ours)*		**65.5** (2.8↑)	**69.4** (1.8↑)	**75.3** (4.3↑)	**52.4** (8.6↑)
+*Top-k*	BLIP	65.4	67.4	74.2	50.9
+*Q-Frame* [38]		65.8	67.9	74.7	50.9
+*AKS* [27]		64.7	68.6	74.2	50.7
+*QCA (Ours)*		**66.9** (4.2↑)	**70.1** (2.5↑)	**75.8** (4.8↑)	**51.8** (8.0↑)
+*QCA (Ours)*	LongCLIP	**66.2** (3.5↑)	**69.5** (1.9↑)	**75.6** (4.6↑)	**51.9** (8.1↑)

Table 5: Comparison of different frame budgets on LongVideoBench, MLVU and LVBench with Qwen3-VL.

Method	Frame	LongVideoBench			MLVU	LVBench
		Med	Long	Avg		
Uniform	64	65.0	52.8	63.1	71.0	43.8
QCA	16	63.1	56.6	62.7	71.3	46.3
	32	66.0	57.4	64.5	73.3	48.9
	64	67.2	59.2	66.9	75.9	52.4
	128	**69.7**	**59.4**	**67.8**	**76.8**	**53.2**

stantially lower than the cost of processing dense visual tokens with the Video-LLM, demonstrating the efficiency of our approach.

VL Embeddings. Tab. 4 evaluates QCA with different Vision-Language embeddings, *e.g.*, CLIP, BLIP, and LongCLIP, under a fixed frame budget of 64. QCA consistently improves performance across different VL embedding models and benchmarks, indicating that the gains mainly stem from the proposed selection strategy rather than the choice of pre-trained encoder. These results highlight the model-agnostic design of QCA and its ability to generalize across different VL embedding models.

Keyframe Budget. We further analyze the effect of different frame budgets to assess the frame efficiency of QCA. As shown in Tab.5, our method consistently outperforms uniform sampling across all benchmarks even with only 16 or 32 frames, demonstrating that QCA can significantly reduce the number of frames or tokens required by MLLMs while maintaining strong performance.

As the frame budget increases from 16 to 128, QCA shows steady performance improvements on LongVideoBench, MLVU, and LVBench. This trend indicates that QCA can effectively utilize additional frames without introducing severe temporal redundancy. Unlike uniform sampling, which often captures re-

Table 6: Comparison with ForestPrune on VideoMME and MLVU with LLaVA-Video.

Method	In Frames	Prune%	Used Frames	VideoMME	MLVU
LLaVA-Video	64	0	64	64.4	70.8
ForestPrune [16]	128	50%	64	63.7	70.5
ForestPrune [16]	256	75%	64	64.2	72.5
QCA (Ours)	64	0	64	**66.1**	**74.1**

dundant frames, QCA prioritizes informative and diverse keyframes, leading to more efficient use of the available frame budget.

Furthermore, we compare our method with a recent token pruning approach, ForestPrune [16], under an identical effective visual token budget. As shown in Tab.6, QCA consistently outperforms ForestPrune on VideoMME and MLVU. This is likely because QCA proactively selects semantically informative frames before encoding, leading to higher information density within the fixed budget, whereas token pruning operates post-encoding and may discard tokens from less-relevant frames.

Overall, these results demonstrate a favorable trade-off between frame budget and performance, making QCA well-suited for long video understanding under strict computational constraints.

5 Conclusion

In this work, we study the problem of efficient long video understanding with MLLMs under a strict frame budget. We identify temporal redundancy as a key factor in existing Video-LLMs and propose a query- and content-aware framework that dynamically allocates frame budgets across video segments before selecting keyframes. By jointly modeling semantic relevance, content deviation, and diversity among selected frames, our method selects a compact yet informative set of keyframes for subsequent reasoning. Extensive experiments demonstrate that our approach consistently outperforms uniform sampling and recent frame selection baselines across different backbones. In particular, our method achieves strong performance even with substantially fewer frames, highlighting its superior frame efficiency and scalability. Notably, we adopt uniform temporal segmentation for simplicity and efficiency. Incorporating content-aware segmentation, *e.g.*, shot boundary detection, is a promising direction for future work.

Acknowledgments

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122701), the National Natural Science Foundation of China under Grant (No.U25B6003 and No.62425101), and the Shenzhen Science and Technology Program under Grant (No.KQTD20240729102051063).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022) 3
2. Arslan, S., Tanberk, S.: Key frame extraction with attention based deep neural networks. arXiv preprint arXiv:2306.13176 (2023) 2, 4
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 1, 3, 7
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 9
5. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022) 4
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 1
7. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: European Conference on Computer Vision. pp. 19–35. Springer (2024) 4
8. Chen, T., Ju, S., Wu, Q., Fang, C., Zhang, K., Peng, J., Li, H., Zhou, Y., Ji, R.: Towards effective and efficient long video understanding of multimodal large language models via one-shot clip retrieval. arXiv preprint arXiv:2512.08410 (2025) 8, 10
9. Chen, W., Zeng, Y., Luo, Y., Xie, T., Lin, L., Ji, J., Zhang, Y., Zheng, X.: Wavelet-based frame selection by detecting semantic boundary for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24052–24061 (2026) 4
10. Dong, W., Zhang, Z., Song, C., Tan, T.: Identifying the key frames: An attention-aware sampling method for action recognition. Pattern Recognition **130**, 108797 (2022) 2
11. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 1
12. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., 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. pp. 24108–24118 (2025) 1, 3, 7

13. Hu, K., Gao, F., Nie, X., Zhou, P., Tran, S., Neiman, T., Wang, L., Shah, M., Hamid, R., Yin, B., et al.: M-llm based video frame selection for efficient video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13702–13712 (2025) 4
14. Huang, D.A., Radhakrishnan, S., Yu, Z., Kautz, J.: Frag: Frame selection augmented generation for long video and long document understanding. arXiv preprint arXiv:2504.17447 (2025) 8, 9
15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) 9
16. Ju, S., Song, B., Chen, T., Zhang, J., Wu, Q., Chang, C., Wang, H., Zhou, Y., Ji, R.: Forestprune: High-ratio visual token compression for video multimodal large language models via spatial-temporal forest modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8326–8336 (2026) 4, 15
17. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024) 1, 4
18. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 6, 8
19. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. Science China Information Sciences **68**(10), 200102 (2025) 3
20. Liang, H., Li, J., Bai, T., Huang, X., Sun, L., Wang, Z., He, C., Cui, B., Chen, C., Zhang, W.: Keyvideollm: Towards large-scale video keyframe selection. arXiv preprint arXiv:2407.03104 (2024) 4
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 1, 3
22. Liu, S., Zhao, C., Xu, T., Ghanem, B.: Bolt: Boost large vision-language model without training for long-form video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3318–3327 (2025) 4, 8
23. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al.: Nvila: Efficient frontier visual language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4122–4134 (2025) 9
24. Soldan, M., Pardo, A., Alcázar, J.L., Caba, F., Zhao, C., Giancola, S., Ghanem, B.: Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5026–5035 (2022) 1
25. Song, B., Peng, J., Zhang, Y., Chen, G., Yang, F., Guo, J.: KTV: Keyframes and key tokens selection for efficient training-free video LLMs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 9060–9068 (2026). https://doi.org/10.1609/aaai.v40i11.37862 4
26. Sun, G., Singhal, A., Uzkent, B., Shah, M., Chen, C., Kessler, G.: From frames to clips: Efficient key clip selection for long-form video understanding. arXiv preprint arXiv:2510.02262 (2025) 4
27. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29118–29128 (2025) 4, 8, 9, 14

28. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024) 9
29. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., Wang, C., Zhang, D., Du, D., Wang, D., Yuan, E., Lu, E., Li, F., Sung, F., Wei, G., Lai, G., Zhu, H., Ding, H., Hu, H., Yang, H., Zhang, H., Wu, H., Yao, H., Lu, H., Wang, H., Gao, H., Zheng, H., Li, J., Su, J., Wang, J., Deng, J., Qiu, J., Xie, J., Wang, J., Liu, J., Yan, J., Ouyang, K., Chen, L., Sui, L., Yu, L., Dong, M., Dong, M., Xu, N., Cheng, P., Gu, Q., Zhou, R., Liu, S., Cao, S., Yu, T., Song, T., Bai, T., Song, W., He, W., Huang, W., Xu, W., Yuan, X., Yao, X., Wu, X., Li, X., Zu, X., Zhou, X., Wang, X., Charles, Y., Zhong, Y., Li, Y., Hu, Y., Chen, Y., Wang, Y., Liu, Y., Miao, Y., Qin, Y., Chen, Y., Bao, Y., Wang, Y., Kang, Y., Liu, Y., Dong, Y., Du, Y., Wu, Y., Wang, Y., Yan, Y., Zhou, Z., Li, Z., Jiang, Z., Zhang, Z., Yang, Z., Huang, Z., Huang, Z., Zhao, Z., Chen, Z., Lin, Z.: Kimi-vl technical report (2025) 9
30. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22958–22967 (2025) 7
31. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 1, 7, 9
32. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024) 3, 7
33. Wu, W., Zhao, Y., Li, Z., Li, J., Zhou, H., Shou, M.Z., Bai, X.: A large cross-modal video retrieval dataset with reading comprehension. Pattern Recognition **157**, 110818 (2025) 1
34. Xu, Z., Zhang, J., Wang, Q., Liu, Y.: E-vrag: Enhancing long video understanding with resource-efficient retrieval augmented generation. arXiv preprint arXiv:2508.01546 (2025) 8, 9
35. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 1
36. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint arXiv:2408.04840 (2024) 9
37. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023) 3
38. Zhang, S., Yang, J., Yin, J., Luo, Z., Luan, J.: Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. arXiv preprint arXiv:2506.22139 (2025) 8, 9, 14
39. Zhang, Y., Zhao, Z., Chen, Z., Ding, Z., Yang, X., Sun, Y.: Beyond training: Dynamic token merging for zero-shot video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22046–22055 (2025) 4
40. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024) 2, 3, 4, 7, 9
41. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding.

In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13691–13701 (2025) 7
42. Zhu, Z., Xu, H., Luo, Y., Liu, Y., Sarkar, K., Yang, Z., You, Y.: Focus: Efficient keyframe selection for long video understanding. arXiv preprint arXiv:2510.27280 (2025) 4
43. Zohar, O., Wang, X., Dubois, Y., Mehta, N., Xiao, T., Hansen-Estruch, P., Yu, L., Wang, X., Juefei-Xu, F., Zhang, N., et al.: Apollo: An exploration of video understanding in large multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18891–18901 (2025) 9