A DYNAMIC TEMPORAL SEGMENTATION FOR LONG VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have achieved remarkable success in video understanding, yet scaling them to long videos remains a major challenge due to severe temporal redundancy and limited frame budgets. Some existing methods rely on uniform temporal segmentation, implicitly assuming that semantic content varies uniformly over time, which misallocates frame budgets and misses critical moments. In this paper, we propose DyQCA, a Dynamic Segmentation framework for keyframe selection in long video understanding. DyQCA detects temporal boundaries by comparing the semantic differences of adjacent frames, partitions the video into dynamic segments, and allocates the frame budget across segments by jointly modeling query relevance and content deviation, followed by anchor-centric keyframe selection within each segment. The proposed DyQCA is training-free, model-agnostic, and can be seamlessly plugged into existing Video-LLMs. Extensive experiments demonstrate that DyQCA consistently outperforms uniform sampling and recent keyframe selection baselines, *e.g.*, 67.3% on LongVideoBench and 76.2% on MLVU under a 64-frame budget, demonstrating that dynamic semantic segmentation offers a simple yet effective path toward more accurate long video understanding.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Li et al., 2024; Bai et al., 2025a; Wang et al., 2025b) have recently achieved remarkable success in video understanding, enabling a wide range of applications such as video question answering (Song et al., 2026), retrieval (Wu et al., 2025), and grounding (Li et al., 2025a). By leveraging large-scale pre-training and strong cross-modal alignment, these models can reason over visual content and textual queries in a unified manner. However, scaling MLLMs to long video understanding remains a major challenge. Long videos are inherently temporally redundant, and encoding dense frame sequences produces an excessive number of visual tokens that quickly exceed the context budget of current MLLMs.

More critically, when the input is truncated or aggressively subsampled, semantically critical moments may be discarded, leading to degraded reasoning performance. As a result, how to select a compact yet informative subset of frames under a strict frame budget has become a central problem for long video understanding.

Existing approaches to this problem can be broadly categorized into several families. Uniform sampling (Zhang et al., 2024) is simple and efficient but query-agnostic, and often selects redundant or irrelevant frames for a specific query. Adaptive keyframe selection methods (Liu et al., 2025a; Tang et al., 2025; Zhang et al., 2025) score frames individually based on query relevance or content importance, but they do not explicitly model how events are organized over time, and thus may miss evidence that spans multiple segments. Token pruning methods (Bolya et al., 2022; Chen et al., 2024; Ju et al., 2026) operate at the token level after encoding frames. While effective at reducing token count, they may discard informative tokens from less-relevant frames and are orthogonal to frame selection.

More recently, segment-level methods (Sun et al., 2025; Peng et al., 2026) explicitly partition videos into temporal segments before selection, but some of them rely on uniform temporal segmentation, implicitly assuming that semantic content varies uniformly over time. This assumption rarely

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

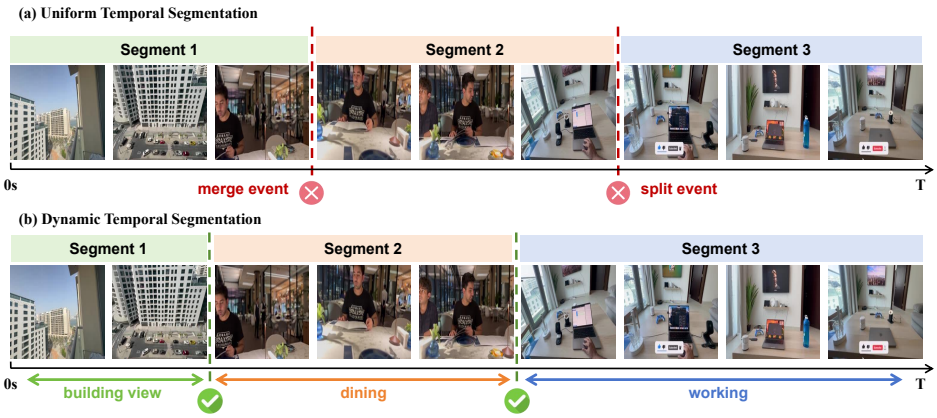

Figure 1: Uniform vs. Dynamic. Uniform temporal segmentation places boundaries at fixed intervals, which may split a single event or merge multiple events into one segment. In contrast, dynamic segmentation aligns boundaries with semantic transitions, producing more coherent segments.

holds in real-world videos, where events are irregularly distributed and semantic boundaries are not aligned with fixed temporal intervals, as shown in Fig.1.

To address this issue, we advocate segmenting videos where their semantic content actually changes, rather than partitioning them by fixed temporal intervals. Intuitively, a video can be viewed as a sequence of semantic states, and transitions between states often correspond to event boundaries. If such boundaries can be detected, the resulting segments will be more semantically homogeneous internally and more distinctive externally, which in turn allows frame budgets to be allocated more precisely and keyframes to be selected more effectively.

However, realizing this idea is non-trivial. First, semantic changes must be measured in a way that is both sensitive to event boundaries and robust to noise. Second, the segmentation criterion must generalize across diverse video domains, from short clips to hour-long videos. Third, the segmentation must remain training-free and lightweight, so that it can be seamlessly integrated into existing Video-LLMs without additional supervision or fine-tuning.

Motivated by these observations, we propose DyQCA, a Dynamic Temporal Segmentation framework for Query- and Content-Aware keyframe selection in long video understanding. DyQCA first extracts frame-level semantic features and measures local semantic variation using the difference between the adjacent frames. A temporal boundary is declared when the difference exceeds a threshold, *i.e.*, which we find to offer the best cross-dataset generalization among several candidate criteria.

The video is then partitioned into dynamic segments accordingly, replacing the uniform segmentation used in prior work. Within this dynamic structure, DyQCA then estimates the information contribution of each segment by jointly modeling query relevance and content deviation, allocates the frame budget proportionally to segment contribution, and performs anchor-centric keyframe selection within each segment (Peng et al., 2026). The whole pipeline is training-free and model-agnostic, and can be plugged into existing Video-LLMs without architectural modification.

To validate the proposed method, we apply DyQCA to three MLLMs, including LLaVA-Video-7B (Zhang et al., 2024), InternVL-3.5-8B (Wang et al., 2025b), and Qwen3-VL-8B (Bai et al., 2025a), and conduct extensive experiments on four long video understanding benchmarks, *i.e.*, Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2025), and LVBench (Wang et al., 2025a). The experimental results not only show its consistent superiority over uniform sampling and recent keyframe selection baselines, *e.g.*, 0.6% improvement on LongVideoBench and 1.0% on MLVU under a 64-frame budget, but also confirm its strong generalization across backbones, demonstrating that dynamic semantic segmentation offers a simple yet effective path toward more efficient long video understanding.

Overall, our contributions are three-fold:

- We identify the limitation of uniform temporal segmentation in long video understanding and propose a semantic-aware dynamic segmentation paradigm that aligns segment boundaries with event transitions, providing a new perspective for frame-budget allocation in Video-LLMs.
- We term this paradigm DyQCA, a training-free framework that combines feature differences for dynamic segmentation, joint query-relevance and content-deviation modeling for budget allocation and keyframe extraction.
- We demonstrate the effectiveness and generality of DyQCA through extensive experiments on four benchmarks and three Video-LLM backbones, achieving consistent improvements over uniform sampling and keyframe selection methods, while remaining lightweight, model-agnostic, and plug-and-play.

2 Related Work

2.1 Video Understanding with MLLMs

MLLMs have achieved remarkable progress in vision-language understanding by extending LLMs with visual perception modules and instruction tuning (Alayrac et al., 2022; Liu et al., 2023). Building on this, recent Video-LLMs represent videos as sequences of visual tokens extracted from sampled frames, enabling tasks such as video question answering, retrieval, and captioning (Zhang et al., 2023; Li et al., 2025b). However, scaling these models to long videos remains a major challenge. Long videos are inherently temporally redundant, and encoding dense frame sequences produces too many visual tokens, quickly exceeding the context budget of current MLLMs. Naively increasing the number of input frames therefore leads to higher computational cost or truncated visual context. This limitation motivates efficient video representations that select a compact yet informative subset of frames, which we refer to as keyframe selection.

2.2 Keyframe Selection for Video-LLMs

More recently, several lines of work have explored efficient video representations under limited frame or token budgets. Segment-based methods explicitly partition videos into temporal segments and perform selection or summarization at the segment level, *e.g.*, through temporal segmentation (Sun et al., 2025; Peng et al., 2026) or wavelet-based boundary detection (Chen et al., 2026). While these methods improve efficiency, most of them still rely on uniform temporal segmentation, implicitly assuming that semantic content varies uniformly over time. Adaptive keyframe selection methods (Liu et al., 2025a; Tang et al., 2025; Zhang et al., 2025) instead score frames individually based on query relevance or content importance, without explicitly segmenting the video. They improve query awareness, but may overlook the intrinsic temporal structure of events. Token pruning (Bolya et al., 2022; Chen et al., 2024; Ju et al., 2026) operates at the token level after encoding, which is effective in reducing tokens, but may discard informative tokens from less-relevant frames.

However, these methods do not explicitly align the temporal structure of the video with semantic content. This motivates dynamic temporal segmentation, which we present in the next section.

3 Method

3.1 Overview and Problem Formulation

Given a video uniformly sampled at 1 FPS, we denote the frame sequence as $\mathbb{X} = \{x_1, x_2, ..., x_N\}$ where N is the total number of frames. With a query q, the goal is to select some key-frames

$$\mathbb{K} \subset \mathbb{X},\ |\mathbb{K}| \ll N, \tag{1}$$

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

$$\hat{y} = \text{Video-LLM}(\mathbb{K}, q). \tag{2}$$

The key challenge is that the frame budget N' is typically much smaller than N, and the informative frames are irregularly distributed across the video. A good selection strategy must therefore decide where to spend the budget and which frames to keep within each region.

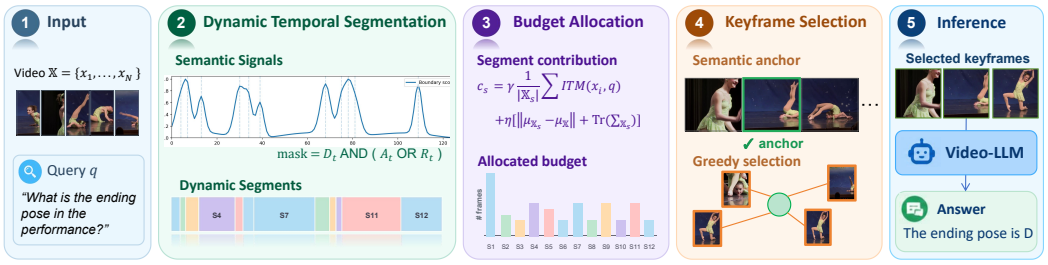

Figure 2: Overview of DyQCA. Given a video and a query, DyQCA first partitions the video into dynamic segments by detecting boundaries via semantic change signals. It then estimates each segment's information contribution by jointly modeling frame-query matching and content deviation, allocates the frame budget accordingly, and selects keyframes within each segment through anchor-centric selection. The resulting keyframe set is fed into a Video-LLM for answering. The entire pipeline is training-free and model-agnostic.

As illustrated in Fig.2, DyQCA consists of three stages, *i.e.*, dynamic temporal segmentation, budget allocation and keyframe selection. The entire pipeline is training-free and model-agnostic, and it can be plugged into existing Video-LLMs without architectural modification.

3.2 Dynamic Temporal Segmentation

Prior keyframe selection methods typically partition a video into S uniformly spaced segments (Peng et al., 2026), either explicitly or implicitly, and then allocate frame budgets across these segments. This design is simple and efficient, but it rests on an assumption that rarely holds in practice Real-world videos contain irregularly distributed events, and semantic transitions do not align with fixed time intervals.

As a result, uniform segmentation often places boundaries inside a semantically coherent event or splits a single event across multiple segments. Both cases corrupt the segment-level statistics used for budget allocation. For segments containing multiple events, relevance scores become overly averaged. Conversely, for segments representing only a portion of an event, content statistics are unstable. Consequently, the frame budget is misallocated, and the selected keyframes may miss the most informative moments. This observation motivates us to segment videos where their semantic content actually changes, rather than at fixed temporal intervals.

To address the above issue, we propose DyQCA, which performs dynamic temporal segmentation. Specifically, DyQCA measures the semantic variation along the temporal axis, and then identifies boundaries where the semantic changes are significant. After partitioning the video into segments, DyQCA passes the resulting dynamic segments to the subsequent budget allocation and keyframe selection stages (Peng et al., 2026). We detail each step below.

Semantic change measurement. To detect semantic boundaries in a training-free manner, DyQCA first extracts frame-level features $f_{t=1}^{N}$ using a pre-trained vision-language encoder, where N is the number of frames uniformly sampled at 1 FPS. We then measure local semantic variation along the temporal axis using three complementary signals. The semantic difference is defined as

$$D_t = \|f_t - f_{t-1}\|_2, \tag{3}$$

which captures the magnitude of semantic change between two adjacent frames. To make the measurement robust to the direction of change, we define the semantic change acceleration as the maximum of the forward and backward differences of D_t:

$$A_t = max(D_t - D_{t-1}, D_t - D_{t+1}), \tag{4}$$

which captures the acceleration of semantic change, *i.e.*, how abruptly the variation changes. Similarly, the semantic change rate is defined as

$$R_t = max(\frac{D_t - D_{t-1}}{D_{t-1}}, \frac{D_t - D_{t+1}}{D_{t+1}}), \tag{5}$$

which measures the relative speed of semantic change.

Intuitively, D_t is large whenever the visual content changes, including gradual within-event motion, whereas A_t and R_t are large only when the change becomes abrupt, which is more characteristic of a true event boundary. The three signals are therefore complementary, *i.e.*, D_t provides sensitivity to semantic variation, while A_t and R_t provide selectivity for genuine transitions.

Boundary detection. Based on these signals, DyQCA identifies a set of boundaries. A frame is declared a boundary if its semantic difference exceeds a threshold and at least one of the two change-sensitive signals also exceeds its threshold:

$$t \in \mathbb{B} \iff D_t \geq \lambda_1 \; and \; (A_t \geq \lambda_2 \; or \; R_t \geq \lambda_3), \tag{6}$$

where λ_1, λ_2 and λ_3 are thresholds for D_t, A_t and R_t, respectively. This criterion requires a boundary to exhibit both a sufficiently large semantic change and a sufficiently abrupt or rapid change, which suppresses spurious boundaries caused by gradual within-event variation.

Dynamic segmentation. Let the selected boundaries be

$$\mathbb{B} = \{b_1, b_2, ..., b_{S-1}\}, \quad 1 < b_1 < b_2 < ... < b_{S-1} < N, \tag{7}$$

which partition the video into S dynamic segments:

$$\mathbb{X} = \bigcup_{s=1}^{S} \mathbb{X}_s, \; \mathbb{X}_s \bigcap \mathbb{X}_{s'} = \emptyset \text{ for } s \neq s'. \tag{8}$$

Here, the number of segments S is a user-specified budget that controls the granularity of segmentation, while the locations of the boundaries are determined by the video content.

This is in contrast to uniform segmentation, where both the number and the locations of boundaries are content-agnostic. The resulting segments are more semantically homogeneous internally and more distinctive externally, which provides a better foundation for the subsequent budget allocation and keyframe selection stages.

One might consider using shot boundary detection or scene segmentation to obtain content-aware segments. However, such methods typically rely on domain-specific detectors or additional supervision, and their granularity is often too coarse or too fine for frame-budget allocation. In contrast, our segmentation criterion is training-free, model-agnostic, and operates directly on the same features used for keyframe selection, requiring no additional modules or annotations. It therefore integrates naturally into the overall DyQCA pipeline.

3.3 Budget Allocation

Given the dynamic segments $\mathbb{X}_{s=1}^{S}$ from Sec.3.2, we apply query-content-aware budget allocation (Peng et al., 2026) to estimate each segment's information contribution by jointly modeling semantic matching and content deviation:

$$c_s = \gamma \cdot \underbrace{\frac{1}{|\mathbb{X}_s|} \sum_{x_i \in \mathbb{X}_s} \text{ITM}(x_i, q)}_{\text{semantic matching}} + \eta \cdot \underbrace{[\|\mu_{\mathbb{X}_s} - \mu_{\mathbb{X}}\| + \text{Tr}(\Sigma_{\mathbb{X}_s})]}_{\text{content deviation}}, \tag{9}$$

where $\mu_{\mathbb{X}_s}$ and $\Sigma_{\mathbb{X}_s}$ denote the mean and covariance matrix of the segment $\mathbb{X}_s$, $\mu_{\mathbb{X}}$ is the mean of global video, and $\text{Tr}(\cdot)$ denotes the trace of a matrix.

Then the allocated budget is

$$q_s = \lfloor \frac{c_s^{\tau}}{\sum_j^S c_j^{\tau}} \cdot N' \rfloor, \tag{10}$$

with remaining frames assigned to the highest-scoring segments if $\sum q_s < N'$.

The segments are now content-driven rather than uniformly spaced. Therefore, semantic matching and content deviation are computed over semantically coherent regions. This makes the budget allocation more accurate, which means a segment containing a single event receives scores that reflect that event, rather than an average over multiple unrelated events. Sec.4 shows that this change alone yields consistent gains over uniform segmentation.

3.4 KEYFRAME SELECTION

Within each dynamic segment $\mathbb{X}_{s=1}^{S}$, we apply semantic anchor-centric keyframe selection (Peng et al., 2026) that balances semantic relevance and content diversity. Specifically, we first choose the most query-relevant frame as a semantic anchor to initialize the set of keyframes $\mathbb{K}$:

$$\mathbb{K}_s = \{\arg\max_{x_i \in \mathbb{X}_s} \text{ITM}(x_i, q)\}. \tag{11}$$

We then build a set of candidates using frames with semantic relevance exceeding a certain threshold:

$$\mathbb{C}_s = \{x_j \in \mathbb{X}_s | \text{ITM}(x_j, q) > \psi_s \cdot r^*\}, \tag{12}$$

where r^* is the anchor's semantic matching score. Then we iteratively add the frame that maximizes its distance to the current keyframe set

$$\mathbb{K}_s \leftarrow \mathbb{K}_s \bigcup \{\arg\max_{x_i \in \mathbb{X}_s \setminus \mathbb{K}_s} \sum_{x_j \in \mathcal{K}_s} \phi(x_i, x_j)\} \tag{13}$$

until $|\mathbb{K}_s| = q_s$. The final keyframe set is $\mathbb{K} = \bigcup_{s=1}^{S} \mathbb{K}_s$

Because the dynamic segments are semantically homogeneous, the anchor frame is a more reliable representative of the segment, and diversity-based expansion operates within a coherent event rather than across unrelated events. This reduces the risk of selecting frames that are diverse but irrelevant, and improves the information density of $\mathbb{K}$ under the fixed budget.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate DyQCA on four widely used long video understanding benchmarks. LongVideoBench (Wu et al., 2024) focuses on referred reasoning questions over videos up to one hour, where answers depend on long-range temporal evidence. Video-MME (Fu et al., 2025) (w/o subs) spans short-, medium-, and long-term videos across diverse visual domains, providing a comprehensive benchmark for general video understanding. MLVU (Zhou et al., 2025) contains videos ranging from 3 minutes to 2 hours and covers many distinct evaluation tasks. LVBench (Wang et al., 2025a) targets extreme-length videos, including TV series, sports broadcasts, and surveillance footage, and emphasizes long-range information extraction.

Backbones. We integrate DyQCA into three representative Video-LLMs, *i.e.*, LLaVA-Video (Zhang et al., 2024), InternVL-3.5 (Wang et al., 2025b), and Qwen3-VL (Bai et al., 2025a). To test scalability, we further report results with Qwen3-VL-30B-A3B-Instruct. All backbones receive the selected keyframes as visual input and the query as text input, and are used in a zero-shot manner without any fine-tuning.

Implementation details. Following prior work (Peng et al., 2026), we uniformly sample frames at 1 FPS before keyframe selection to reduce computational cost. We use BLIP-2 (Li et al., 2023) as the default encoder for both frame features f_t and image-text matching $\text{ITM}(\cdot, \cdot)$. The default frame budget is set to $N' = 64$ unless otherwise stated. The thresholds $\lambda_1, \lambda_2, \lambda_3$ are selected via grid search, γ and η are both set to 0.5 by default. All experiments are conducted on $8\times$3090 GPUs.

4.2 MAIN RESULTS

4.2.1 COMPARISON WITH KEYFRAME SELECTION METHODS

Tab.1 compares DyQCA with uniform sampling, Top-k selection, and recent keyframe selection baselines, including AKS, Q-Frame, OneClip-RAG, and QCA under a 64-frame budget. With Qwen3-VL-8B, DyQCA achieves 67.3 on LongVideoBench, 69.3 on Video-MME, 76.2 on MLVU, and 50.4 on LVBench. It improves over uniform sampling by 7.3% on LongVideoBench and 7.3% on MLVU, and outperforms Top-k, AKS, and Q-Frame on all benchmarks. Even on LVBench, DyQCA remains competitive and still outperforms uniform sampling by a large margin. Compared with OneClip-RAG, which requires additional training, DyQCA achieves superior performance without any fine-tuning on all benchmarks where both are available.

Table 1: Comparisons on LongVideoBench, Video-MME, MLVU, and LVBench. We evaluate our approach against uniform sampling, Top-k selection, and recent frame selection baselines across three backbones: LLaVA-Video-7B, InternVL-3.5-8B, and Qwen3-VL-8B with a 64-frame budget. † denotes our re-implementation.

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* Tang et al. (2025)	7B	BLIP	62.7	65.3	71.8	47.6
+ *Q-Frame* Zhang et al. (2025)	7B	LongCLIP	61.5	64.7	72.9	47.1
+ *OneClip-RAG* Chen et al. (2025)	7B	CLIP*	62.5	65.2	71.2	-
+ *QCA* Peng et al. (2026)	7B	BLIP	**62.9**	**66.1**	74.1	**48.9**
+ *DyQCA (Ours)*	7B	BLIP	62.7	65.6	**74.7**	48.5
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* Tang et al. (2025)	8B	BLIP	62.9	62.8	70.5	47.9
+ *Q-Frame* Zhang et al. (2025)	8B	LongCLIP	62.5	61.7	70.6	48.9
+ *OneClip-RAG* Chen et al. (2025)	8B	CLIP*	62.1	61.7	70.2	-
+ *QCA* Peng et al. (2026)	8B	BLIP	63.5	**63.9**	**71.3**	**50.0**
+ *DyQCA (Ours)*	8B	BLIP	**65.6**	63.5	70.6	49.5
Qwen3-VL-8B						
+ *Uniform*†	8B	-	62.7	67.4	71.0	43.7
+ *Top-k*	8B	BLIP	64.4	67.4	74.2	50.7
+ *AKS* Tang et al. (2025)	8B	BLIP	64.7	68.6	74.2	50.8
+ *Q-Frame* Zhang et al. (2025)	8B	LongCLIP	65.8	67.9	74.7	50.7
+ *OneClip-RAG* Chen et al. (2025)	8B	CLIP*	65.3	68.5	71.3	-
+ *QCA*† Peng et al. (2026)	8B	BLIP	66.7	69.3	75.2	**50.9**
+ *DyQCA (Ours)*	8B	BLIP	**67.3**	**69.3**	**76.2**	50.4

Table 2: Comparisons with state-of-the-art models using uniform sampling.

Method	LLM Size	Frames	LongVideoBench	Video-MME	MLVU	LVBench
GPT-4o Hurst et al. (2024)	-	256	66.7	71.9	64.6	-
Kimi-VL-16B-A3B Team et al. (2025)	16B	64	64.5	67.8	74.2	-
InternVL3.5-38B Wang et al. (2025b)	38B	64	65.7	70.9	77.0	-
LLaVA-Video-72B Zhang et al. (2024)	72B	64	63.9	70.0	74.4	45.5
Qwen2.5-VL-72B Bai et al. (2025b)	72B	1fps	60.7	**73.3**	74.6	47.3
NVILA Liu et al. (2025b)	7B	256	57.7	64.0	70.1	-
Apollo Zohar et al. (2025)	7B	2fps	58.5	61.3	68.7	-
mPLUG-Owl3 Ye et al. (2024)	8B	128	59.7	59.3	70.0	43.5
Qwen3-VL-8B	8B	64	63.1	67.6	71.0	43.8
+DyQCA (Ours)	8B	64	**67.3**	69.3	**76.2**	**50.4**

With LLaVA-Video-7B, DyQCA consistently outperforms uniform sampling and query-aware keyframe selection baselines across all benchmarks. Compared with prior QCA, DyQCA achieves a higher score on MLVU, which is the benchmark with the longest and most semantically irregular videos. This suggests that content-aware segmentation brings benefit when suffering from temporal redundancy and semantic non-uniformity. On LongVideoBench, Video-MME, and LVBench, DyQCA is slightly lower than QCA, suggesting that for these benchmarks, the uniform segmentation provides a reasonable temporal prior, and the gain from dynamic segmentation is offset by the occasional over-segmentation of stable intervals. A similar trend is observed with InternVL-3.5-8B. The consistent gains across backbones confirm that the benefit of dynamic segmentation is not specific to a single Video-LLM.

4.2.2 Comparison with SOTA Video-LLMs

Tab.2 compares DyQCA with strong proprietary and open-source models. With Qwen3-VL-8B under a 64-frame budget, DyQCA achieves 67.3 on LongVideoBench, 69.3 on Video-MME, 76.2

Table 3: Breakdown on Video-MME with Qwen3-VL-8B under a 64-frame budget.

Qwen3-VL-8B	Video-MME			
	Short	Medium	Long	Overall
+ *Uniform*	79.33	66.00	56.78	67.37
+ *QCA*	**80.56**	**69.11**	58.22	69.30
+ *DyQCA(ours)*	79.89	68.11	**60.00**	**69.33**

Table 4: Breakdown on MLVU with Qwen3-VL-8B under a 64-frame budget.

Qwen3-VL-8B	MLVU							
	Plot QA	Neddle QA	Ego Reasoning	Action Count	Action Order	Anomaly Recognition	Topic Reasoning	Overall
+ *QCA*	75.85	**83.10**	70.45	**59.71**	59.46	77.50	89.35	75.21
+ *DyQCA(ours)*	**78.48**	81.41	**71.88**	59.22	**66.80**	**78.50**	**91.25**	**76.22**

on MLVU, and 50.4 on LVBench. Notably, DyQCA outperforms GPT-4o on LongVideoBench and MLVU, despite using only 64 frames compared to GPT-4o's 256 frames. It also surpasses LLaVA-Video-72B on LongVideoBench and MLVU, and outperforms Qwen2.5-VL-72B on MLVU and LVBench, demonstrating strong frame efficiency.

4.2.3 Breakdown on Video-MME and MLVU

Tab.3 breaks down Video-MME into short, medium, and long subsets under Qwen3-VL-8B. DyQCA achieves 79.89, 68.11, and 60.00 with an overall accuracy of 69.33, outperforming uniform sampling and QCA overall.

The breakdown reveals where dynamic segmentation matters most. On the long subset, DyQCA improves to 60.00, with $+3.22$ and $+1.78$ gains over uniform sampling and QCA, respectively. This is consistent with our motivation, *i.e.*, long videos suffer most from temporal redundancy and semantic non-uniformity, and content-driven segmentation is precisely what allows the frame budget to be concentrated on the most informative events.

Tab.4 reports per-task results on MLVU under Qwen3-VL-8B. DyQCA achieves 78.48 on plotQA, 81.41 on needle, 71.88 on ego, 59.22 on count, 66.80 on order, 78.50 on anomaly recognition, 91.25 on topic reasoning, and 76.22 overall.

The largest gain is on the "order" task, which requires understanding the temporal ordering of events, meaning that correct segment boundaries are crucial. DyQCA also improves plotQA, ego, anomaly recognition, and topic reasoning, showing that dynamic segmentation benefits tasks involving temporal structure, causal reasoning, and event-level understanding. On needle and count, DyQCA is slightly lower than QCA, which is expected because these tasks rely on locating a single frame or counting repeated objects, and may benefit more from uniform segmentation than from dynamic segmentation.

4.3 Ablation Studies

Tab.5 compares boundary criteria on the long subset of Video-MME, the 600-3600s subset of LongVideoBench, and MLVU, under the same frame budget and the same number of segments. The final criterion achieves the best results on Video-MME and MLVU, while $A_t \vee R_t$ alone is best on LongVideoBench, showing that sparse-transition long videos benefit from change-sensitive signals without the D_t constraint. These results indicate a trade-off between sensitivity and selectivity, and suggest that adaptively choosing the criterion based on video statistics is a promising direction.

4.4 Qualitative Results

Fig.3 shows visualization examples of DyQCA, illustrating the effect of content-aware dynamic temporal segmentation. Instead of partitioning videos at fixed intervals, DyQCA places boundaries

Table 5: Ablation on boundary criteria. "Uniform" uses fixed temporal intervals without dynamic segmentation. D_t, A_t, and R_t denote the semantic difference, change acceleration, and change rate.

Criterion	Video-MME(long)	LongVideoBench(600-3600s)	MLVU
Uniform Segment	58.22	62.09	75.21
D_t	58.78	62.30	75.34
A_t	57.67	63.73	75.76
$A_t \vee R_t$	57.67	**64.45**	75.85
$D_t \wedge (A_t \wedge R_t)$	59.67	63.22	75.71
$D_t \wedge (A_t \vee R_t)$	**60.00**	62.81	**76.22**

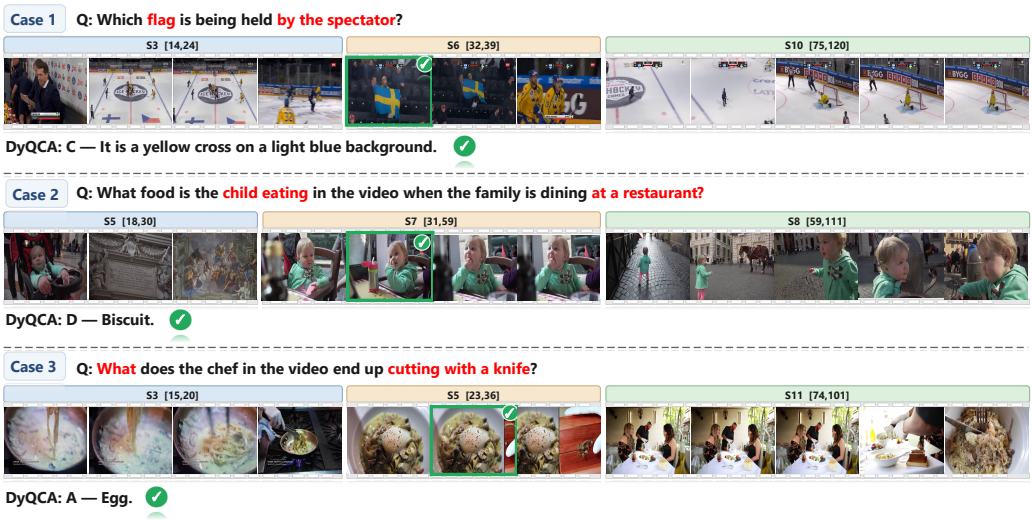

Figure 3: Qualitative results of DyQCA. Each video is partitioned into segments of varying lengths, and selected keyframes are marked with green boxes. In all three cases, the answer depends on a short, query-critical event, *e.g.*, a flag, which uniform sampling may miss. Dynamic segmentation allows DyQCA to localize these events while retaining sufficient context.

where the semantic content changes, allowing the frame budget to concentrate on query-relevant segments while preserving sufficient context. In all three cases, the answer depends on a short, localized event, *e.g.*, the Swedish flag, the biscuit, and the egg, and DyQCA selects the frame containing that evidence, whereas uniform sampling could easily miss it.

5 Conclusion

In this paper, we introduce DyQCA, a training-free framework with dynamic temporal segmentation that detects boundaries via the semantic difference, change acceleration, and change rate, and allocates the frame budget across content-driven segments according to query relevance and content deviation. Extensive experiments on four widely used benchmarks with three Video-LLM backbones demonstrate that DyQCA consistently outperforms uniform sampling and recent keyframe selection baselines, and can surpass much larger models under a smaller frame budget. The gains are most pronounced on longer and more semantically irregular videos, as well as on tasks requiring temporal ordering, which aligns with our motivation that content-driven segmentation is most valuable when suffering from temporal redundancy and semantic non-uniformity. In addition, DyQCA is training-free, model-agnostic, and can be seamlessly integrated into existing Video-LLMs.

AI USE STATEMENT

We used generative AI tools only for organizing and restructuring existing code during the research process. Generative AI tools were not used for code generation, research ideation, methodology development, experimental design, data analysis. All research ideas, technical contributions, experimental results, and the final manuscript were developed and prepared by the authors. We have reviewed all AI-assisted outputs and take full responsibility for the final content of this work.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 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.

Tao Chen, Shaobo Ju, Qiong Wu, Chenxin Fang, Kun Zhang, Jun Peng, Hui Li, Yiyi Zhou, and Rongrong Ji. Towards effective and efficient long video understanding of multimodal large language models via one-shot clip retrieval. *arXiv preprint arXiv:2512.08410*, 2025.

Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. 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.

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*, pp. 24108–24118, 2025.

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.

Shaobo Ju, Baiyang Song, Tao Chen, Jiapeng Zhang, Qiong Wu, Chao Chang, HuaiXi Wang, Yiyi Zhou, and Rongrong Ji. 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.

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.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8592–8603. IEEE, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. 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, 2025a.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4122–4134, 2025b.

Jun Peng, Baiyang Song, Jie Li, Hui Li, Yiyi Zhou, Rongrong Ji, and Yonghong Tian. QCA: Query-and content-aware keyframe selection for long video understanding. *arXiv preprint arXiv:2607.00983*, 2026.

Baiyang Song, Jun Peng, Yuxin Zhang, Guangyao Chen, Feidiao Yang, and Jianyuan Guo. KTV: Keyframes and key tokens selection for efficient training-free video-LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9060–9068, 2026. doi: 10.1609/aaai.v40i11.37862.

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

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

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, et al. Kimi-vl technical report, 2025.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. A large cross-modal video retrieval dataset with reading comprehension. *Pattern Recognition*, 157:110818, 2025.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. *arXiv preprint arXiv:2506.22139*, 2025.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

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*, pp. 13691–13701, 2025.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, 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.