This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Shaojie Zhang[1*] Jiahui Yang[1*] Jianqin Yin[1†] Zhenbo Luo[1‡] Jian Luan[1]
[1]MiLM Plus, Xiaomi Inc.
zhangshaojie5@xiaomi.com, yangjiahui1@xiaomi.com, jianqinyin4@gmail.com
luozhenbo@xiaomi.com, luanjian@xiaomi.com

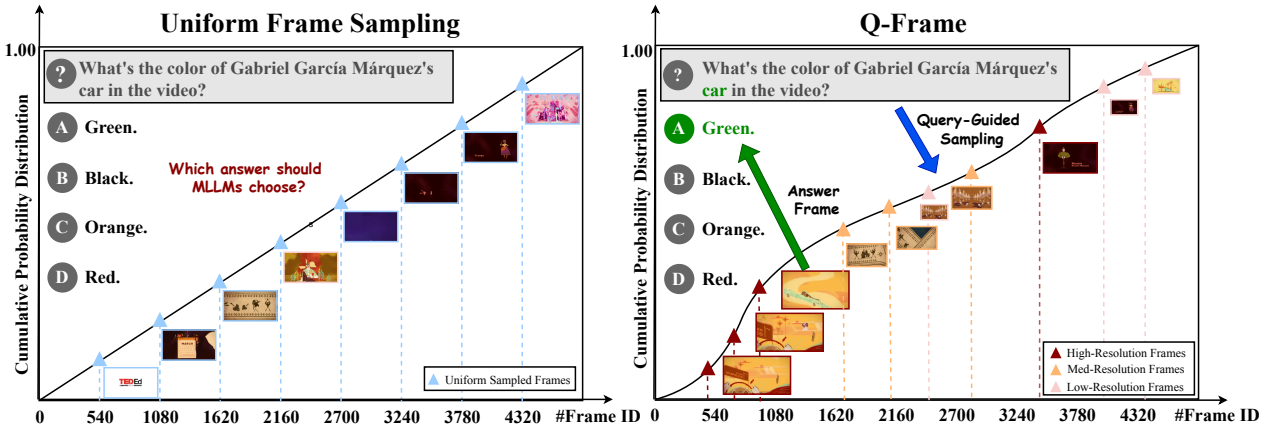

Figure 1. Comparison of uniform frame sampling and proposed **Q-Frame** sampling. Uniform Frame Sampling selects frames at fixed intervals, leading to sparse and potentially irrelevant frame selections that disrupt temporal continuity. In contrast, Q-Frame dynamically selects query-aware frames and adapts their resolution, ensuring that the most relevant frames are selected with optimized resolution to preserve crucial visual details. This adaptive approach enhances the Video-LLMs' ability to understand long-form videos more efficiently and effectively, addressing the limitations of traditional frame sampling methods.

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated significant success in visual understanding tasks. However, challenges persist in adapting these models for video comprehension due to the large volume of data and temporal complexity. Existing Video-LLMs using uniform frame sampling often struggle to capture the query-related crucial spatiotemporal clues of videos effectively. In this paper, we introduce Q-Frame, a novel approach for adaptive frame selection and multi-resolution scaling tailored to the video's content and the specific query. Q-Frame employs a training-free, plug-and-play strategy based on a text-image matching network like CLIP, utilizing the Gumbel-Max trick for efficient frame selection. Q-Frame allows Video-LLMs to process more frames without exceeding computational limits, thereby preserving critical temporal and spatial information. We demonstrate Q-Frame's effectiveness through extensive experiments on benchmark datasets, including MLVU, LongVideoBench, and Video-MME, illustrating its superiority over existing methods and its applicability across various video understanding tasks.

† Corresponding author; ‡ Project leader; * Equal contribution.

1. Introduction

Multimodal Large Language Models (MLLMs) have achieved remarkable performance in a wide range of visual understanding tasks, showcasing their potential across various domains [1, 21, 27]. Building upon these advancements, MLLMs are now poised to tackle more complex challenges, such as video understanding, which requires not only capturing spatial details but also effectively modeling temporal dynamics [5, 11, 14, 20].

Video understanding, however, presents unique challenges due to the inherently large number of video frames and the continuous nature of video content. Despite progress in video understanding, current methods suffer from inherent limitations due to the temporal redundancy in videos and the restricted context length within MLLMs. A typical video contains a large number of frames—e.g.,

a 3-minute video at 24 frames per second (fps) results in approximately 4,320 frames. Modern MLLMs remain constrained by their context—VideoLLaMA2 [5] supports only 2,000 tokens, while VILA-V1.5 [15] extends this to roughly 4,000 tokens. It becomes clear that using all frames for understanding is computationally infeasible. Previous approaches have relied on uniform frame sampling, where a fixed subset of frames is selected at regular intervals. While this method is straightforward, it has several drawbacks: it results in **sparse frame selections** that break temporal continuity, it follows a **static and query-agnostic** sampling strategy, and it **treats all frames equally**, which is inefficient and suboptimal for fine-grained tasks [14, 20, 28, 39].

Specifically, the inherent sparsity in uniform sampling is particularly problematic for long videos. As video length increases, the fixed number of sampled frames becomes increasingly sparse, disrupting the temporal flow and potentially omitting critical transitions and subtle details. This is detrimental for tasks such as object counting and temporal grounding [7]. Furthermore, uniform frame sampling is inherently static, meaning that the same combination of frames is adopted for each question, regardless of the query's particular needs. Some efforts, such as text-frame matching [13] and frame ranking [33], have been introduced to break this static selection. However, these methods fundamentally fail to capture the intricate temporal dependencies and dynamic contextual relationships inherent in advanced video comprehension tasks. In the end, the distribution of information density in video is often not uniform, and uniform frame sampling fails to consider the varying information density across frames. Downsampling high-resolution frames can lead to a loss of fine-grained details, which negatively impacts the Video-LLMs' ability to capture spatial information. While maintaining the native resolution of each frame will significantly increase the consumption of computing resources. This one-size-fits-all approach may fail to capture spatiotemporal clues in videos.

To address these limitations, we propose **Q-Frame**, a novel framework for query-adaptive frame selection and resolution adaptation in Video-LLMs. As illustrated in Figure 1, Q-Frame dynamically selects the most relevant frames from a video and adjusts their resolutions according to both the video content and the given query. In contrast to prior approaches that treat all frames uniformly, Q-Frame introduces a dynamic strategy that makes frame sampling *query-aware* and computationally efficient. This strategy is derived from a discrete probability distribution parameterized by the output of a text–image matching model (e.g., CLIP [22]). By leveraging the Gumbel-Max trick [8], we can sample from this distribution to obtain an optimal combination of frames without resorting to complex optimization or any additional training, making Q-Frame easy to implement. Importantly, as shown in Figure 2, Q-Frame allows a Video-LLM to process a greater number of informative frames without exceeding the computational budget, thus preserving crucial visual details that would be lost with uniform sampling. Moreover, Q-Frame is model-agnostic and plug-and-play, requiring no fine-tuning of the Video-LLM and can be applied to both open-source models [10, 27] and closed-source API-based models [21].

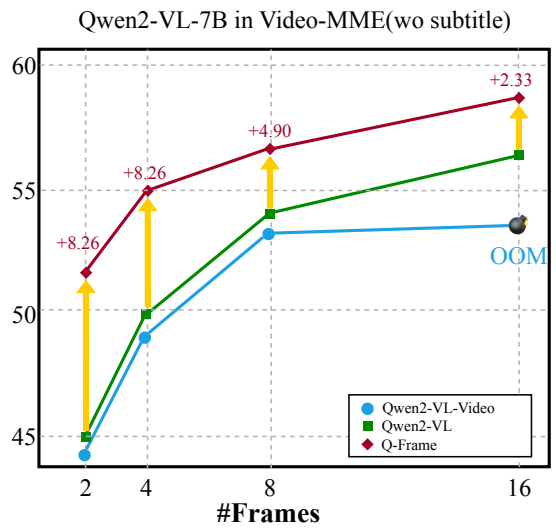

Figure 2. Overall accuracy (%) Video-MME (without subtitles) for Qwen2-VL, comparing uniform sampling and Q-Frame selected from 128 frames. Note that Qwen2-VL-Video is the video understanding model, and Qwen2-VL is a multi-image understanding model based on different activation weights.

The main contributions of this paper are as follows:

1. We introduce **Q-Frame**, a novel framework for query-aware frame selection and multi-resolution frame scaling that improves long-form video understanding by focusing on the most relevant visual content for a given query.
2. We propose a training-free, plug-and-play mechanism that leverages a CLIP-based vision-language model to guide frame selection and resolution assignment *without* any additional model training or fine-tuning.
3. We validate the effectiveness of Q-Frame through extensive experiments on three benchmark datasets (MLVU [41], LongVideoBench [29], and Video-MME [7]), demonstrating its superiority over previous methods and its practical usability for video understanding tasks.

2. Related Work

Video Large Language Models. Early video understanding approaches primarily relied on handcrafted features or conventional deep networks with limited capacity to capture temporal dynamics and semantic richness in visual content [26]. Recent advances in Multimodal Large Language Models (MLLMs) have revolutionized video un-

derstanding. The Video-LLMs can be categorized into the following three categories. Firstly, VideoChat [35], LLaVA-Hound [38], InternVideo2 [28], ShareGPT4Video [3], Video-LLaVA [14], and LLaVA-Video [39] leverage synthesized video-instruction data (e.g., video captions and open-ended QA) to enhance fine-grained action understanding and narrative tracking. Secondly, Video-ChatGPT [20], MovieChat [25], TimeChat [23], Video-LLaMA [35], and Video-CCAM [6] develop efficient feature compression methods for holistic video semantics, employing temporal aggregation and contrastive spatial-temporal modeling. Finally, LongVA [37], LongVILA [32], and Kangaroo [19] extend the LLM's context length to accommodate extended visual token sequences for long video analysis.

Keyframe Selection. Keyframe selection is a fundamental technique for identifying representative frame combinations that balance content coverage and reduce redundancy. In the context of early action recognition, AdaFrame [31] pioneered reinforcement learning (RL)-based adaptive frame selection, later extended by [30] through multi-agent decision-making. Dynamic video analysis has led to the development of motion-guided approaches such as MGSampler [40] and its probabilistic variant PMGSampler [2], which prioritize motion-salient frames. Most Video-LLMs [14, 20, 28, 39] adopt query-agnostic uniform sampling, often resulting in incomplete relevant content capture. Recent advances introduce query-conditioned strategies. KeyVideoLLM [13] implements query-aware top-K semantic retrieval, whereas Frame-Voyager [33] employs loss-driven combinatorial optimization. Nevertheless, existing methods face persistent challenges in temporal fragmentation and rigid resolution constraints during frame representation.

Any-Resolution Processing. MLLMs typically encode visual inputs as discrete visual tokens for cross-modal alignment. Seminal works [16, 17] establish this paradigm through fixed-resolution vision encoders, inherently constraining input flexibility. Subsequent approaches [4, 12, 18] address native resolution preservation by partitioning images into fixed-grid patches, yet retain structural rigidity and disregard spatial dependencies during tokenization. A breakthrough emerges with Qwen2-VL's [27] Native Dynamic Resolution framework, enabling adaptive visual token generation across varying input resolutions. Although any-resolution techniques have been well-researched for Image-LLMs, the dynamic resolution adaptation of Video-LLMs remains unexplored.

3. Method

3.1. Overview

This work focuses on Video Question Answering (VQA) as the primary task while ensuring generalization to various video understanding scenarios, such as summarization and temporal grounding [33]. The Video-LLM pipeline can be formalized as $(\mathcal{V}, q) \Rightarrow a$, where the input video $\mathcal{V} = \{\text{Frame}^i\}_{i=1}^{D}$ with D frames, and q represents the textual query.

The proposed Q-Frame framework, as illustrated in Figure 3, adapts dynamically to select query-aware frame sequences with resolution scaling. It enables Video-LLMs to process extended frame sequences while maintaining a fixed computational budget through two key mechanisms: 1) content-aware frame selection based on cross-modal query relevance and 2) resolution scaling that optimally allocates visual tokens across spatial dimensions. Specifically, Q-Frame is composed of Cross-modal Query Retrieval (CQR), Query-Aware Frame Selection (QFS), and Multi-Resolution Adaptation (MRA). CQR focuses on retrieving the most semantically relevant frames from the video based on the textual query, ensuring that only meaningful visual information is considered. QFS is designed to adaptively select frames based on their relevance to the query, enhancing efficiency by concentrating on the most important temporal segments. MRA aims to optimize computational resources by assigning varying resolutions to frames, preserving fine details in important frames while reducing costs for less critical ones.

3.2. Cross-modal Query Retrieval

To effectively answer fine-grained associations of objects, actions, and scenes, it is necessary to retrieve key frames that semantically match the query. We first downsample the original video $\mathcal{V}$ to obtain a candidate frame sequence $\mathcal{F} = \{\text{Frame}^j\}_{j=1}^{T} \subset \mathcal{V}$ using uniform sampling, where $T \ll D$ specifies the target frame budget.

We leverage pre-trained CLIP-like models that align millions of image-text pairs, offering excellent zero-shot capabilities for cross-modal retrieval [22]. The pre-trained Vision-Language Model (VLM) maps the visual input $\mathcal{F}$ and textual input q into a shared semantic space:

$$\begin{aligned} Q &= \text{VLM}_{\text{text}}(q) \in \mathbb{R}^d, \\ F &= \text{VLM}_{\text{vision}}(\mathcal{F}) \in \mathbb{R}^{T \times d} \end{aligned} \tag{1}$$

where $\text{VLM}_{\text{vision}}(\cdot)$ and $\text{VLM}_{\text{text}}(\cdot)$ denote the vision and text encoders of the VLM, respectively. To compute the semantic relevance of each frame, we compute the similarity as the inner product between the query and frame embeddings:

$$I = QF^{\mathrm{T}} \in \mathbb{R}^{1 \times T} \tag{2}$$

3.3. Query-aware Frame Selection

CLIP-like models are primarily trained on image-text pairs, which inherently neglect the temporal correlations present in video frames. To address this, we introduce a probability-guided sampling strategy based on the Gumbel-Max trick

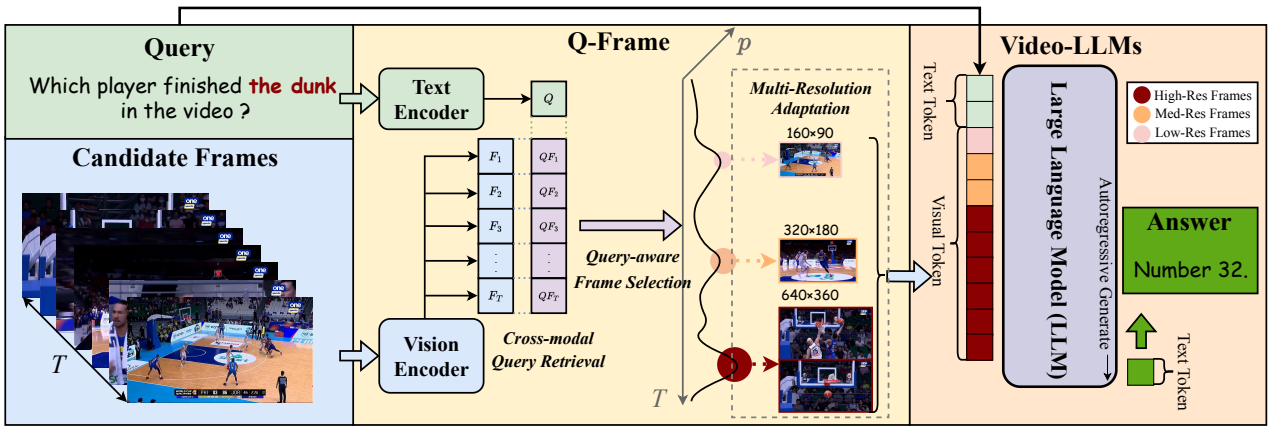

Figure 3. The overall framework of Q-Frame. Q-Frame is composed of Cross-modal Query Retrieval (CQR), Query-Aware Frame Selection (QFS), and Multi-Resolution Adaptation (MRA). CQR focuses on retrieving the most semantically relevant frames from the video based on the textual query, ensuring that only meaningful visual information is considered. QFS is designed to adaptively select frames based on their relevance to the query, enhancing efficiency by concentrating on the most important temporal segments. MRA aims to optimize computational resources by assigning varying resolutions to frames, preserving fine details in important frames while reducing costs for less critical ones. It should be noted that the preprocessing strategy of Video-LLMs is different, and the MRA in the dotted line is not applicable to every model.

[8], which efficiently samples frames according to their query-relevant matching intensities. This strategy not only ensures diversified frame selection but also dynamically adjusts the probability ordering, enhancing the robustness of the sampling strategy.

First, we convert the matching intensity I into a probability distribution using temperature scaling:

$$\pi = \text{Softmax}(I/\tau) \tag{3}$$

which indicates the probability that each frame is selected, where τ is the temperature parameter controlling the sharpness of the distribution.

Subsequently, independent Gumbel noise is injected into the logarithm of the probability distribution to perturb the log-probabilities:

$$\begin{aligned} g &= -\log\left(-\log \epsilon\right), \epsilon \in U[0,\ 1]^T \\ p &= \log \pi + g \end{aligned} \tag{4}$$

where g follows a standard Gumbel distribution. Using the Gumbel-Max trick, we draw discrete samples from the categorical distribution:

$$\text{idx}^{\text{select}} = \{i \mid \text{rank}(i) \leq K\} \tag{5}$$

Here, $\text{rank}(\cdot)$ sorts the frames by the value of p, and $\text{idx}^{\text{select}}$ retrieves the indices of the top-K frames after noise injection. This sampling method combines the exploration-exploitation trade-off, allowing for a more flexible and robust frame selection.

3.4. Multi-Resolution Adaptation

Many existing Video-LLMs process all frames at the same resolution [14, 20, 39]. However, frames that contain fine details (e.g., small objects or textual overlays) benefit from higher resolution, while less critical frames can be processed at lower resolutions to reduce computational cost. To address this, we propose a multi-resolution strategy that adapts frame resolutions based on their relevance to the query.

Using the query-aware sampling strategy from Section 3.3, we assign different resolutions to frames according to their relevance. Let K, M, and N be hyperparameters that define resolution thresholds based on token budgets and empirical analysis. The high, medium, and low-resolution frames are selected as follows:

$$\begin{aligned} \text{idx}^{\text{high}} &= \{i \mid \text{rank}(i) \leq K\} \\ \text{idx}^{\text{mid}} &= \{i \mid K < \text{rank}(i) \leq M\} \\ \text{idx}^{\text{low}} &= \{i \mid M < \text{rank}(i) \leq N\} \end{aligned} \tag{6}$$

For each frame Frame^i, the resolution r_i is assigned as follows:

$$r_i = \begin{cases} r^{(3)} & \text{if } i \in \text{idx}^{\text{high}}, \\ r^{(2)} & \text{if } i \in \text{idx}^{\text{mid}}, \\ r^{(1)} & \text{if } i \in \text{idx}^{\text{low}}, \end{cases} \tag{7}$$

where $r^{(1)}$, $r^{(2)}$, and $r^{(3)}$ correspond to low, medium, and high resolutions, respectively. To transform the frame resolutions, we downsample the frames as follows: $r^{(1)} = 4r^{(2)} = 16r^{(3)}$.

This resolution adaptation has two key benefits: It ensures that critical details are preserved in high-resolution frames, and by reducing the resolution of less important frames, it minimizes the computational load, enabling the Video-LLM to process more frames within the same token budget.

3.5. Inference with Video-LLMs

During inference, we connect our Q-Frame to the conventional Video-LLMs. Suppose we uniformly sample T candidate frames and expect N frames with various resolutions as visual input for Video-LLMs. Then, the textual input and visual input are fed into respective tokenizers to extract tokens. It's important to note that frames with higher resolutions will produce more visual tokens, while frames with lower resolutions will produce fewer tokens. In the end, the textual tokens and visual tokens are connected for LLM's autoregressive generation.

4. Experiments

In this section, we present a comprehensive evaluation of the proposed Q-Frame framework, detailing our experimental setup, main results, and ablation studies to demonstrate its effectiveness and robustness in video understanding. **Due to space limitations, additional experimental details and results are provided in the Appendix.**

4.1. Experimental Setup

Baseline Models: To validate Q-Frame, as detailed in Table 1, we select representative open-source models, namely VILA-V1.5 [15] and Qwen2-VL [1], as well as the closed-source API-based GPT-4o [21]. Q-Frame is integrated with these baselines to assess its performance relative to recent state-of-the-art Video-LLMs. In order to maintain comparability with FRAME-VOYAGE [33], experiments using VILA-V1.5 employ QFS with fixed input frames and resolutions. Additionally, because of the closed-source nature of GPT-4o, the effect of its preprocessing on image resolution remains unknown. Consequently, the GPT-4o's experiments only utilize QFS. Our evaluation includes two experimental settings: one with a fixed number of input frames and another with a fixed token budget.

It is also noteworthy that the video understanding and multi-image understanding modes in Qwen2-VL activate distinct model weights; hence, we refer to them as Qwen2-VL-Video and Qwen2-VL, respectively.

Benchmarks: We evaluate Q-Frame on three widely adopted long-video benchmarks: (1) MLVU [41], a benchmark for multi-task understanding of long videos comprising 2593 tasks across nine categories (average video duration: 12 minutes); (2) LongVideoBench [29], evaluated on a validation set without subtitles, containing 1337 QA pairs (average duration: 12 minutes); and (3) Video-MME [7], which includes 2700 manually annotated QA pairs (average duration: 17 minutes).

Implementation Details: Following the experimental protocol in [33], we uniformly sample 128 candidate frames and apply Q-Frame to select 8 frames (or an equivalent token budget via multi-resolution assignment) as input to the Video-LLMs. All experiments are conducted on LMMs-Eval [36] using 8 H100 GPUs.

Model	Open-Source	Q-Frame	
		QFS	MRA
VILA-V1.5 [15]	✓	✓	×
GPT-4o [21]	×	✓	×
Qwen2-VL [27]	✓	✓	✓

Table 1. Baseline models and experimental configurations for Q-Frame.

4.2. Main Results

4.2.1. Comparison with SOTA Methods

Table 2 presents a detailed comparison of Q-Frame applied to VILA-V1.5 [15], Qwen2-VL [27], and GPT-4o [21], along with several recent Video-LLMs. In both fixed frame and fixed token settings, Q-Frame consistently outperforms uniform frame sampling. For example, Q-Frame significantly boosts performance in Video-MME and MLVU benchmarks, demonstrating its capability to select query-relevant frames and dynamically adjust resolutions adaptively. Compared to the baselines, Q-Frame substantially improves the SOTA performance across all three evaluated benchmarks. We highlight several key contributions:

1. Our Q-Frame empowers Qwen2-VL to achieve the highest overall performance on MLVU, while GPT-4o achieved the best result in other benchmarks, confirming the effectiveness of Q-Frame.
2. The improvements brought by Q-Frame are consistent across different Video-LLMs, whether open-source models like VILA-V1.5 and Qwen2-VL or API-based models like GPT-4o. This proves that Q-Frame is a robust, model-agnostic enhancement that can be seamlessly integrated with existing Video-LLM architectures to improve performance across diverse video understanding tasks.
3. When Q-Frame is applied to Video-LLMs, whether open-source VILA-V1.5 and Qwen2-VL or closed-source GPT-4o, it consistently outperforms uniform frame sampling. This shows that Q-Frame is effective in frame selection and generalizes well across various Video-LLMs.
4. Notably, VILA-V1.5 with Q-Frame selecting 8 frames outperforms FRAME-VOYAGER [33], which employs loss-driven combinatorial optimization and trains a

Model	LLM Size	#Frames	MLVU	LongVideoBench	Video-MME$_{(wo/w\ subs)}$			
					Overall	Short	Medium	Long
Avg. Video Duration			*12min*	*12min*	*17min*	*1.3min*	*9min*	*41min*
Video-LLaVA [14]	7B	8	47.3	-	39.9 / 41.6	45.3 / 46.1	38.0 / 40.7	36.2 / 38.1
Qwen-VL [1]	7B	8	-	-	41.1 / 41.9	46.9 / 47.3	38.7 / 40.4	37.8 / 37.9
VideoChat2 [11]	7B	8	44.5	-	39.5 / 43.8	48.3 / 52.8	37.0 / 39.4	33.2 / 39.2
Chat-UniVi-V1.5 [9]	7B	8	-	-	40.6 / 45.9	45.7 / 51.2	40.3 / 44.6	35.8 / 41.8
VideoLLaMA2 [5]	7B	8	-	-	47.9 / -	56.0 / -	45.4 / -	42.1 / -
LLaVA-NeXT-QW2 [18]	7B	8	-	-	49.5 / -	58.0 / -	47.0 / -	43.4 / -
LongVILA [32]	8B	128	-	-	49.2 / -	60.2 / -	48.2 / -	38.8 / -
LongVA [37]	7B	128	-	-	52.6 / 54.3	61.1 / 61.6	50.4 / 53.6	46.2 / 47.6
Video-XL [24]	7B	128/256	64.9	-	55.5 / 61.0	64.0 / 67.4	53.2 / 60.7	49.2 / 54.9
LLaVA-OneVision [10]	7B	*	64.7	56.3	58.2 / -	- / -	- / -	- / -
Video-CCAM [6]	9B	96	58.5	-	50.3 / 52.6	61.9 / 63.1	49.2 / 52.3	39.6 / 42.4
Fixed input frames:								
VILA-V1.5 [15]	8B	8	46.3	47.1	47.5 / 50.0	57.8 / 61.6	44.3 / 46.2	40.3 / 42.1
+FRAME-VOYAGER [33]	8B	8	49.8	-	50.5 / 53.6	60.3 / 65.0	47.3 / 50.3	43.9 / 45.3
+ **Q-Frame**	8B	8	54.4	51.6	50.7 / 55.0	59.8 / 65.3	48.0 / 53.4	44.2 / 46.2
GPT-4o [21]	-	8	28.6 ♠	53.3	61.9 / 64.5	69.1 / 72.1	62.4 / 63.8	54.0 / 57.9
+ **Q-Frame**	-	8	29.3	**58.6**	**63.8 / 66.5**	**69.9 / 73.6**	**63.8 / 65.2**	**57.6 / 60.8**
Fixed input tokens:								
Qwen2-VL-Video [27]	7B	8	55.6	51.0	53.0 / 58.3	64.1 / 68.2	49.3 / 55.1	45.6 / 51.7
Qwen2-VL [27]	7B	8	56.9	53.5	53.7 / 59.4	65.0 / 69.8	50.7 / 56.2	45.3 / 52.1
+ **Q-Frame**	7B	4 + 8 + 32	**65.4**	58.4	58.3 / 61.8	69.4 / 73.2	57.1 / 61.0	48.3 / 51.1

Table 2. Comparing Video-LLMs with and without Q-Frame as an additional module. The experiments of LongVideoBench were performed on the validation set, excluding interleaved subtitles. Results are shown for the Video-MME benchmark under two setups: without subtitles (w/o sub.) and with subtitles (w sub.). Video-XL uses 128 frames for Video-MME and 256 frames for MLVU. LLaVA-OneVision refers to results from the official report with the well-tuned number of frames. And Video-CCAM uses 96 frames for each benchmark. The lighter the color, the lower the frame resolution. The best results are shown in **bold**. ♠: For more experimental results, refer to the appendix.

frame ranking model, while our Q-Frame offers a straightforward and training-free approach.

4.2.2. Performance Across Video Lengths

As shown in Table 3, the baseline models (VILA-V1.5, GPT-4o, and Qwen2-VL) exhibit a consistent performance drop as the video duration increases, reflecting the difficulty in handling longer videos with fixed computational resources. When Q-Frame is applied, we observe significant improvements across most video lengths, particularly in longer videos, demonstrating the effectiveness of our query-aware frame selection and multi-resolution adaptation mechanisms. Here are some noteworthy insights:

1. As video length increases, Video-LLMs with uniform frame sampling strategies lead to a significant drop in performance due to the inherent temporal sparsity and redundancy. Q-Frame mitigates this by intelligently selecting key frames based on their relevance to the query, yielding notable improvements, particularly for longer videos.
2. Q-Frame's improvement distribution across various baselines is notable. It demonstrates a significant enhancement over VILA-V1.5, particularly in medium-length videos, while GPT-4o and Qwen2-VL show advancements aimed at long videos. This distinction may arise from the differences in the training data strategies of Video-LLMs.
3. The results indicate that Q-Frame significantly enhances baseline performance for most video durations. However, the performance of VILA-V1.5 in the (8s, 15] range demonstrates a slight decrease. This is because uniformly sampling 8 frames in this range results in about one frame per second, making frame selection less significant for very short videos.

4.2.3. Performance Across Video Tasks

Figure 4 compares the accuracies of Qwen2-VL-Video, Qwen2-VL, and Q-Frame on six video tasks in Video-MME, including *Reasoning*, *Recognition*, *Perception*, *Counting Problem*, *OCR Problems*, and *Information Synopsis* in the Video-MME benchmark. The results reveal several important insights:

1. Q-Frame enhances performance in *Reasoning* tasks, where temporal context and subtle transitions are crucial.

Model	LongVideoBench			
Video Duration	(8s, 15s]	(15s, 1m]	(3m, 10m]	(15m, 60m]
VILA-V1.5 [15]	56.1	60.5	43.4	42.7
+Q-Frame	$52.9^{-3.2}$	$62.8^{+2.3}$	$53.9^{+10.5}$	$46.1^{+3.4}$
GPT-4o [21]	53.4	64.8	52.4	46.8
+Q-Frame	$60.3^{+6.9}$	$68.0^{+3.2}$	$59.7^{+7.3}$	$54.3^{+7.5}$
Qwen2-VL [27]	53.5	67.6	51.5	45.9
+Q-Frame	$63.0^{+9.5}$	$71.5^{+3.9}$	$58.5^{+7.0}$	$56.4^{+10.5}$

Table 3. Performance of various baselines with and without Q-Frame on LongVideoBench across different video durations.

By adaptively selecting semantically relevant frames and applying multi-resolution adaptation, Q-Frame demonstrates a marked improvement over the baseline.

2. In *Recognition* tasks, Q-Frame shows significant performance improvements, suggesting that its query-aware frame selection effectively boosts the model's capacity to capture distinctive visual spatial details.
3. Q-Frame also enhances performance on *Counting Problem* tasks by adaptively choosing key frames that capture all critical instances and maintain fine-grained details through multi-resolution scaling, which reduces redundancy and minimizes interference from irrelevant frames.
4. For *Perception* and *OCR Problems*, the multi-resolution assignment of Q-Frame maintains fine details, which is essential for accurately identifying small objects or textual elements, resulting in superior performance.
5. In the *Information Synopsis* task, Q-Frame surpasses the baselines by effectively aggregating key scenes to create a thorough understanding of the video's narrative.

Overall, these findings substantiate that Q-Frame's adaptive strategies are particularly beneficial for tasks that demand nuanced temporal and spatial representations, thus validating the design choices underlying our framework.

4.3. Ablation Study

We further explore the contributions of the essential components of Q-Frame through comprehensive ablation studies on LongVideoBench using Qwen2-VL as the backbone.

Effectiveness of Q-Frame: Table 4 presents an ablation study of the overall Q-Frame framework, highlighting the significance of each module. The results demonstrate that the combination of QFS and MRA achieves the highest accuracy (58.4%), confirming that the complete Q-Frame framework significantly enhances performance. Specifically, applying CLIP directly to the Top-K matching frames (without QFS) yields an accuracy of 56.0%, which is an improvement over uniform sampling. Incorporating QFS boosts performance to 57.6%, underscoring the importance of query-aware frame selection. Additionally, using MRA alone results in a lower accuracy (52.58%), indicating that multi-resolution adaptation alone does not fully leverage the benefits of Q-Frame. Therefore, it is evident that both QFS and MRA are essential for achieving optimal performance.

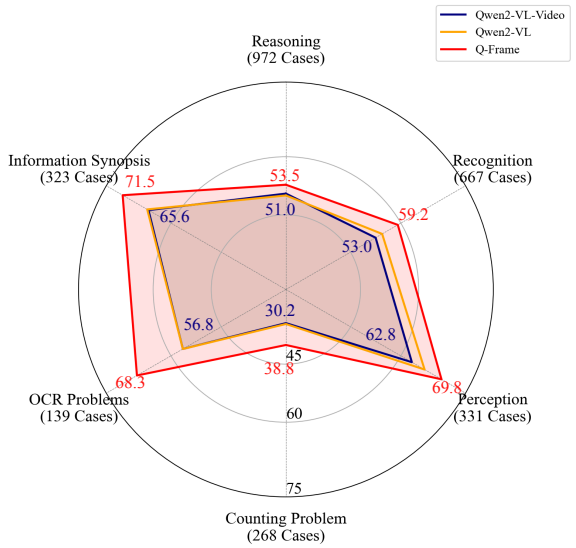

Figure 4. Accuracies (%) of Qwen2-VL-Video, Qwen2-VL, and Q-Frame on six tasks in Video-MME. The maximum results for each task are highlighted.

Sampling			Resolution		Acc(%)
Uniform	CLIP	QFS	Fixed	MRA	
✓			✓		53.5
	✓		✓		56.0
		✓	✓		57.6
✓				✓	52.6
	✓			✓	56.6
		✓		✓	**58.4**

Table 4. Ablation study of Q-Frame. CLIP denotes direct application of cross-modal matching results, and Fixed indicates a fixed resolution setting.

Multi-resolution Strategy: Table 5 highlights the effects of various resolution strategies within the MRA module. The results emphasize the advantages of multi-resolution processing, where allocating a high resolution to key frames and lower resolutions to less important frames enhances performance. The highest accuracy of 58.4% is achieved when all three resolution categories (high, medium, and low) are utilized, illustrating that a flexible resolution strategy is ideal. When only high-resolution frames are employed, the accuracy declines to 57.6%, and relying solely on medium or low resolutions leads to further performance degradation (55.9% and 49.0%, respectively).

These results suggest that a balanced multi-resolution approach offers the best compromise between performance and computational efficiency.

Resolution			Acc(%)
High	Medium	Low	
✓			57.6
	✓		55.9
		✓	49.0
✓	✓		57.9
✓		✓	58.1
	✓	✓	56.3
✓	✓	✓	**58.4**

Table 5. Ablation study of different resolution strategies within the MRA module.

Frames' resolution allocation: Table 6 analyzes how different numbers of frames with varying resolutions affect performance. To regulate the number of tokens, the quantity of frames for each resolution adheres to these guidelines:

$$K + \frac{M}{4} + \frac{N}{16} = 8 \tag{8}$$

The optimal configuration includes 4 high-resolution frames, 8 medium-resolution frames, and 32 low-resolution frames, achieving an accuracy of 58.4%. Further increasing the number of lower-resolution frames (up to 48) decreases accuracy to 57.4%, highlighting the diminishing returns of adding more low-resolution frames. Moreover, due to the necessary preprocessing in MLLMs, the number of tokens has increased slightly with the increase in number, which is generally acceptable. This suggests that the Q-Frame approach benefits from balancing the number of frames of different resolutions, optimizing computational load while preserving high accuracy.

Frames			Tokens/Video	Acc(%)
K	M	N		
8	0	0	2265.1	57.6
6	6	8	$2304.1^{+1.7\%}$	57.9
6	4	16	$2312.5^{+2.0\%}$	58.3
4	8	32	$2344.8^{+3.5\%}$	**58.4**
4	6	40	$2355.8^{+4.0\%}$	58.2
4	4	48	$2370.1^{+4.6\%}$	57.4

Table 6. Ablation study of frame resolution allocation. "Tokens/Video" indicates the average number of tokens processed per video on LongVideoBench.

Figure 5. Case analysis from Video-MME [7]. Uniform sampling captures only a limited number of frames relevant to the query. While our Q-Frame extracts more relevant frames with a variety of resolutions.

4.4. Qualitative Analysis

Figure 5 illustrates a representative example from our qualitative analysis. In this example, the query "How do the players celebrate scoring the first goal in this video?" is answered correctly as "Kneeling on the field and screaming" when Q-Frame is applied, whereas uniform sampling fails to capture sufficient context. This comparison highlights that Q-Frame effectively selects key frames and adapts their resolutions to preserve critical temporal and spatial details, ultimately enabling more accurate and coherent responses. Additionally, with the same consumption level, our Q-Frame can process more frames and assign appropriate resolutions based on relevance.

5. Conclusion

We introduce Q-Frame, a training-free method for efficient video understanding that dynamically selects query-aware frame sequences and adjusts frame resolutions to optimize performance while minimizing computational costs. By incorporating query-aware frame selection and multi-resolution adaptation mechanisms, Q-Frame outperforms traditional uniform sampling and keyframe selection techniques, showing superior accuracy on benchmarks such as MLVU, LongVideoBench, and Video-MME. Our experiments emphasize the essential roles of QFS and MRA in enhancing frame relevance and optimizing resolution. Q-Frame's model-agnostic, plug-and-play design offers a scalable and effective solution for long-form video understanding, paving the way for future advancements in Video-LLMs integration.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1, 5, 6

[2] Jing Bai, Yuxiang Zhang, Yiran Wang, Zhu Xiao, Yong Xiong, and Licheng Jiao. Robust motion-guided frame sampler with interpretive evaluation for video action recognition. *IEEE Transactions on Mobile Computing*, 2025. 3

[3] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. 37:19472–19495, 2025. 3

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 3

[5] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 2, 6

[6] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024. 3, 6

[7] 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. *arXiv preprint arXiv:2405.21075*, 2024. 2, 5, 8, 11

[8] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 2, 4

[9] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 6

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

[11] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1, 6

[12] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024. 3

[13] Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. Keyvideollm: Towards large-scale video keyframe selection. *arXiv preprint arXiv:2407.03104*, 2024. 2, 3

[14] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. 2024. 1, 2, 3, 4, 6

[15] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 2, 5, 6, 7, 12

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023. 3

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3

[18] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 3, 6

[19] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024. 3

[20] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024. 1, 2, 3, 4

[21] OpenAI. Hello gpt-4o, 2024. 1, 2, 5, 6, 7, 11, 12

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021. 2, 3, 12, 13

[23] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 3

[24] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 6

[25] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 3

[26] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023. 2

[27] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 3, 5, 6, 7, 12

[28] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *Proceedings of the European Conference on Computer Vision*, pages 396–416, 2024. 2, 3

[29] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. 37:28828–28857, 2025. 2, 5, 11

[30] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6222–6231, 2019. 3

[31] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2019. 3

[32] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 3, 6

[33] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. 2025. 2, 3, 5, 6, 13

[34] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 13

[35] 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. 3

[36] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024. 5

[37] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 3, 6, 12, 13

[38] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. 3

[39] 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. 2, 3, 4

[40] Yuan Zhi, Zhan Tong, Limin Wang, and Gangshan Wu. Mgsampler: An explainable sampling strategy for video action recognition. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pages 1513–1522, 2021. 3

[41] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 2, 5, 11