QuoTA: Query-oriented Token Assignment via CoT Query Decouple for Long Video Comprehension

Yongdong Luo[1*], Wang Chen[1*], Weizhong Huang[1], Shukang Yin[2], Haojia Lin[1], Jinfa Huang[3], Chaoyou Fu[4], Jiayi Ji[1], Xiawu Zheng[1†], Jiebo Luo[3]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China
[2]Independent Researcher
[3]University of Rochester
[4]Nanjing University

Abstract

Recent advances in long video understanding typically mitigate visual redundancy through visual token pruning based on attention distribution. However, while existing methods employ **post-hoc** low-response token pruning in decoder layers, they overlook the input-level semantic correlation between visual tokens and instructions (query). In this paper, we propose QuoTA, an **ante-hoc** training-free modular that extends existing large video-language models (LVLMs) for visual token assignment based on query-oriented frame-level importance assessment. The query-oriented token selection is crucial as it aligns visual processing with task-specific requirements, optimizing token budget utilization while preserving semantically relevant content. Specifically, (i) QuoTA strategically allocates frame-level importance scores based on query relevance, enabling one-time visual token assignment before cross-modal interactions in decoder layers, (ii) we decouple the query through Chain-of-Thoughts reasoning to facilitate more precise LVLM-based frame importance scoring, and (iii) QuoTA offers a plug-and-play functionality that extends to existing LVLMs. Experimental results demonstrate that implementing QuoTA with LLaVA-Video-7B yields an average performance improvement of **3.2%** across six benchmarks (including Video-MME and MLVU) while operating within an identical visual token budget as the baseline.

Code — https://github.com/MAC-AutoML/QuoTA

Introduction

With the emergence of advanced Large Language Models (LLMs), researchers have expanded their capabilities to video comprehension (Jin et al. 2023; Li et al. 2023; Lin et al. 2023; Zhang et al. 2024b; Zhang, Li, and Bing 2023; Chen et al. 2024d,a), establishing the domain of Large Video-Language Models (LVLMs). Recent studies (Zhang et al. 2024a; Yin Song et al. 2024; Shang et al. 2024; Wang et al. 2024d; Xue et al. 2024) focus on extending the reasoning context capacity of LVLMs, primarily through fine-tuning approaches for enhanced long video understanding. However, empirical evidence from long-context LVLMs (Zhang et al. 2024a) and (Shen et al. 2025) demonstrates performance degradation when frame sampling rates increase, suggesting that merely augmenting frame quantities introduces information redundancy while imposing greater computational demands on complex reasoning tasks.

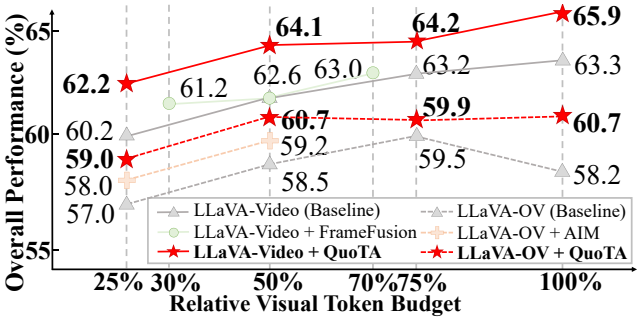

Figure 1: Comparative analysis of Video-MME (Fu et al. 2024a) when implementing attention-based token assignment methods AIM (Zhong et al. 2024) and FrameFusion (Fu et al. 2024b), alongside our proposed query-oriented QuoTA within LLaVA-Video-7B (Zhang et al. 2024c) and LLaVA-OV-7B (Li et al. 2024a) across varied relative visual token budgets. QuoTA exhibits consistent performance enhancement across diverse token budget configurations.

Recent works (Zhong et al. 2024; Tao et al. 2024; Fu et al. 2024b; Wang et al. 2024c) have addressed visual token reduction by analyzing attention value distributions across model layers, thereby providing more informationally efficient tokens for long video comprehension tasks. DyCoke (Tao et al. 2024), for instance, demonstrates that attention score distributions among visual tokens exhibit significant sparsity, enabling the dynamic elimination of less-attended visual tokens within the KV cache. However, as **post-hoc** methods that compress visual tokens either during or after their interaction with textual tokens within the decoder

*These authors contributed equally.
†Corresponding author: zhengxiawu@xmu.edu.cn.

layers, they face substantial limitations: **(i) Neglected task-specific token relevance.** Human visual cognition inherently focuses on query-relevant content when viewing video, highlighting the necessity of query-oriented token selection to align visual processing with task objectives while ensuring efficient token allocation and semantic coherence. Conversely, visual tokens with high attention-weight responses primarily reflect inter-token associative strength rather than their direct relevance to the query.

As demonstrated by (Wu et al. 2024b; Zhong et al. 2024), cross-modal interactions mainly occur in the early fusion stages, with later layers showing intra-modal patterns. Additionally, FrameFusion (Fu et al. 2024b) reveals that token importance significance fluctuates across layers. These observations suggest that high-response visual tokens likely represent intra-modal relationships rather than query-relevant semantic features, potentially rendering attention-based token reduction counterproductive for task-specific comprehension tasks. **(ii) Propagation of sequential reduction errors.** The hierarchical token reduction strategy based on attention patterns exhibits vulnerability to error accumulation, wherein suboptimal selection decisions implemented at initial layers adversely affect subsequent stages.

To address this issue, we propose QuoTA, an **ante-hoc** approach seamlessly integrated with existing LVLMs (termed *based LVLM*) for visual tokens assignment before cross-modal interactions on decoder layers, enhancing query-specific visual token capture. Specifically, QuoTA implements parallel video frame evaluation utilizing the zero-shot capabilities of a lightweight LVLM (termed *scoring LVLM*) to generate query-relevance scores, subsequently employing these metrics as discriminative criteria for visual token assignment. Our strategy enhanced cross-modal interactions while minimizing redundancy between visual and text tokens during the early fusion phases, thereby augmenting performance. To enhance scoring precision, we prompt the *based LVLM* to decouple the query into a more interpretable question with Chain-of-Thoughts (Wei et al. 2022) reasoning. These reformulated questions subsequently prompt the *scoring LVLM* in generating a query-specific relevance score for each sampled frame. After that, the visual tokens then proceed via one of three approaches based on the importance score: (i) bilinear interpolation, (ii) adaptive pooling, and (iii) dynamic token merging for spatial redundancy minimization. Furthermore, since the elevated information density afforded by our query-oriented assignment protocol, we implement duration-dependent dynamic frame sampling to optimize the extraction of salient visual information. Notably, QuoTA can adapt to a given token budget.

We conduct experiments across diverse video understanding benchmarks, including Video-MME (Fu et al. 2024a), MLVU (Zhou et al. 2024), LongVideoBench (Wu et al. 2024a), VNBench (Zhao et al. 2024), MVBench (Li et al. 2024b), and NeXT-QA (Xiao et al. 2021). Results demonstrate that LLaVA-Video-7B (Zhang et al. 2024c) and LLaVA-OneVision-7B (Li et al. 2024a), augmented with QuoTA in a plug-and-play manner, achieve an average performance improvement of 3.2% and 2.5% across all six benchmarks while maintaining equivalent computational requirements to their original baseline. Furthermore, as shown in Figure 1, QuoTA outperforms recent state-of-the-art approaches AIM (Zhong et al. 2024) and FrameFusion (Fu et al. 2024b) across varying visual token budgets when applied to LLaVA-Video-7B (Zhang et al. 2024c) and LLaVA-OneVision-7B (Li et al. 2024a), while maintaining consistent performance improvements over the baseline regardless of token budget, which is set to 12,544 visual tokens in total.

In summary, our contributions are as follows:

- **We design a plug-and-play pipeline for existing LVLMs:** QuoTA provides a training-free solution applicable to diverse LVLMs, enhancing long video understanding by assigning visual tokens based on text instruction (query) relevance. This approach offers a more elegant and direct methodology compared to conventional attention-based analytical techniques.
- **We propose CoT-driven query decouple for query-oriented frame scoring:** QuoTA employs Chain-of-Thoughts to decouple query into a specific-designed question, enabling high-quality scoring of video frames.
- **Our QuoTA setting a new state-of-the-art:** Integration of QuoTA with LLaVA-Video-7B yields a 3.2% average performance improvement across six benchmarks, achieving the best results in five video benchmarks, including Video-MME and MLVU, among 7B LVLMs.

Related Work

Large Video-Language Models

Recent advances in large language models (LLMs) have sparked interest in developing video understanding systems. Video-ChatGPT (Maaz et al. 2023) processes videos by extracting frame-level features through spatial and temporal pooling. VideoChat (Li et al. 2023) combines textual descriptions with video appearance embeddings. Video-LLaVA (Lin et al. 2023) aligns image and video encoders using a shared projector to map representations into a common language space. LLaVA-NeXT-Video (Zhang et al. 2024b) extends LLaVA-NeXT (Liu et al. 2024a) through video-specific fine-tuning. Recent research has focused on expanding context window sizes for long video understanding. LongVA (Zhang et al. 2024a), Video-XL (Shu et al. 2024), and LongVILA (Xue et al. 2024) leverage LLMs' long-text comprehension capabilities through continuous training. However, they may suffer from performance degradation with excessive frame sampling due to video content redundancy and model capacity constraints.

Efficient Visual Modeling

Previous research in multi-modal LLMs assigns visual tokens to reduce spatial redundancy. FastV (Chen et al. 2024b) reduces them at a particular selected layer based on the distribution of attention. Recent works (Fu et al. 2024b; Tao et al. 2024; Zhong et al. 2024; Wang et al. 2024c) extend it by reducing the redundancy of spatiotemporal information for long video understanding. For example, AIM (Zhong et al. 2024) merges similar tokens and then preserves the

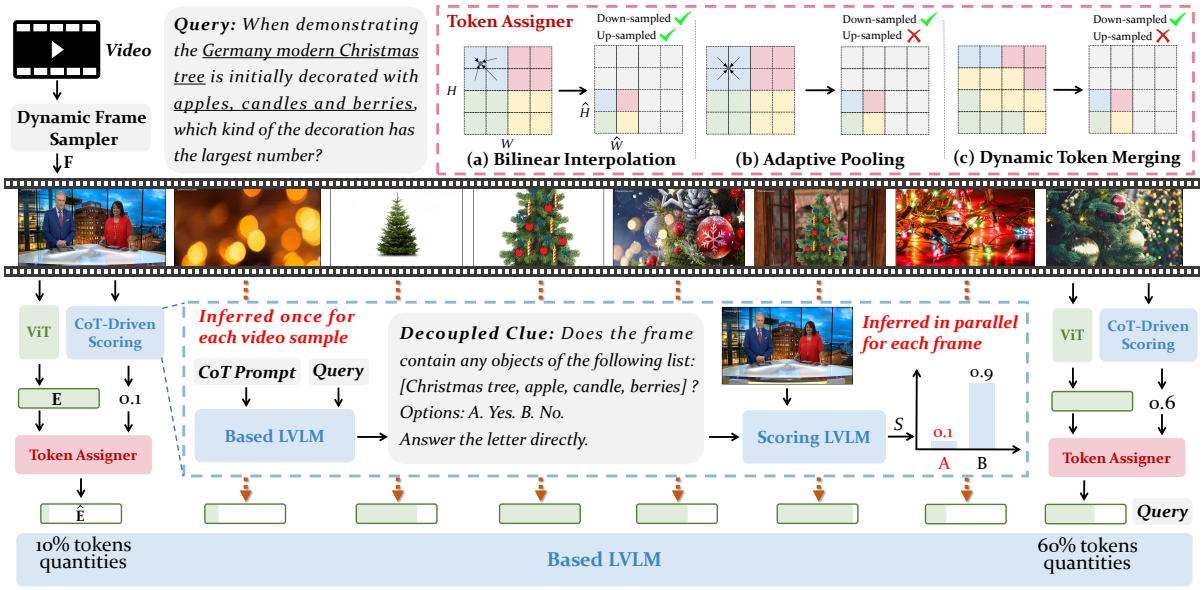

Figure 2: Framework of QuoTA. Initially, a dynamic frame sampler extracts T frames from the video based on duration, which are subsequently processed by $\texttt{ViT}$ to generate visual embeddings $\mathbf{E}$. Then, the based LVLM decouples the query using Chain-of-Thoughts (Wei et al. 2022) into a clue to generate frame-wise importance scores through scoring LVLM in parallel, thus evaluating the relevance to the query. Finally, a token assigner rescales the frame embeddings to $\hat{\mathbf{E}}$ based on these scores.

important tokens hierarchically. However, they rely on attention weight analysis, which only reflects token-wise associations rather than query-specific relevance; this indirect measurement may not accurately capture task-specific token importance. Moreover, as mentioned in FrameFusion (Fu et al. 2024b), token importance is inconsistent across different layers; thus, hierarchical reduction is susceptible to error propagation from early suboptimal decisions. Other works (Shen et al. 2024; Zhang et al. 2025; Wang et al. 2024a; Yang et al. 2024; Li et al. 2024c; Choudhury et al. 2025; Lee et al. 2024) integrate token assignment into the architecture design and enhance model capabilities through training. However, the training-needed schema makes them inflexible.

Method

Preliminaries

For a given video $\mathbf{V}$, a frame sampler extracts T frames $\mathbf{F} = \{\mathbf{F}_i\}_{i=1}^{T}$, which employ uniform sampling at fixed intervals for computational efficiency. Features of each frame are then obtained via $\mathbf{E} = \{\mathbf{E}_i\}_{i=1}^{T} = \texttt{ViT}(\mathbf{F})$, where $\texttt{ViT}$ represents a transformer-based visual encoder (e.g., CLIP-L (Radford et al. 2021)), and $\mathbf{E}_i \in \mathbb{R}^{(H\times W)\times C}$ denotes the visual tokens corresponding to the i-th frame, with H, W representing spatial dimensions and C the feature dimension. In standard LVLM, the video features $\mathbf{E}$ alongside the user query $\mathbf{Q}$ are processed to generate the output $\mathbf{O} = \texttt{LVLM}(\mathbf{E}, \mathbf{Q})$. QuoTA enhances this process by first assessing frame-level importance, which subsequently guides token assignment within each frame to produce refined features $\hat{\mathbf{E}}$. The final output is formulated as $\mathbf{O} = \texttt{LVLM}(\hat{\mathbf{E}}, \mathbf{Q})$. Figure 2 illustrates the overall architecture of QuoTA.

LVLM-Based Frame Scoring

Given the inherent information redundancy in sampled video frames, it is advantageous to compress query-irrelevant visual tokens while directing the LVLM's attention toward keyframes. While text-image similarity scores from CLIP (Radford et al. 2021) could theoretically assess query-frame relevance, CLIP's known bias toward physical entity nouns often yields suboptimal frame importance assessments in reasoning scenarios with a complex query. Instead, in QuoTA, we leverage the robust multi-choice reasoning capabilities of LVLMs by formulating a binary choice question for each frame. This question is processed by a lightweight scoring LVLM, which produces an importance score S (a scalar) derived from the probability of selecting option "A". The naive binary-choice prompt structure fed into the scoring LVLM for each frame follows the format:

> Question: Does this frame contain any information to answer the given query: {query}?
> A. Yes. B. No.
> Answer the letter directly.

To optimize the trade-off between scoring accuracy and efficiency, we employ Qwen2-VL-2B (Wang et al. 2024b) as our lightweight scoring LVLM, which facilitates **parallel inference** across all video frames with minimal latency and modest GPU resource requirements. Upon obtaining the importance score for each sampled frame, we normalize these values to serve as allocation guides for token assignment. The normalized importance score for each frame is computed as $S_n = \{S_n^i\}_{i=1}^{T} = S^i / \sum_{j=1}^{T} S^j$.

CoT-Driven Query Decouple

Employing the naive prompt yields only marginal improvements, as the scoring LVLM tends to optimistically presume that frames contain adequate information. This results in homogeneous importance weights across frames, compromising discriminative capacity. To address this, we implement a Chain-of-Thoughts (Wei et al. 2022) approach that decouples the query that enhances frame differentiation quality, as validated in our ablation studies. Considering that LVLMs demonstrate superior capability in identifying physical entities, QuoTA utilizes the based LVLM to decouple the original query into a structured object list. Specifically, we implement a three-step prompting protocol that directs the base LVLM to decouple the query, extracting concrete physical entities to interrogate the scoring LVLM, which markedly reduces hallucination. This Chain-of-Thoughts process encompasses (i) assessing the necessity for entity decoupling, (ii) transforming the original query into a structured object list when appropriate, and (iii) refining this list by eliminating abstract concepts. When the based LVLM determines entity recognition is warranted, we query the scoring LVLM using the following binary-choice prompt:

> Question: Does the frame contain any objects of the following list: {object_list}?
> A. Yes. B. No.
> Answer the letter directly.

Otherwise, we leverage the naive binary-choice prompt. We also conduct other decouple strategies for query-oriented keyframes scoring, which details in the ablation study.

Dynamic Visual Token Assignment

After obtaining the normalized scores $S_n = \{S_n^i\}_{i=1}^T$ for all frames, we calculate the target token quantities $N = \{N_i\}_{i=1}^T$ for each frame, where $N_i = S_n^i \times N_t$, with N_t representing the total visual token budget. Notably, individual frame token quantities N_i may exceed N_t/T. We established N_t according to the empirically optimal frame configuration of the based LVLM to ensure experimental consistency. For instance, LLaVA-Video (Zhang et al. 2024c) demonstrates optimal performance with 64 frames at 196 tokens per frame; consequently, N_t was standardized at $64 \times 196 = 12,544$. This methodological decision was implemented for multiple reasons: (i) Equivalent token counts enable direct comparative assessment of QuoTA's efficacy while controlling computational variables; (ii) Since the base LVLM was trained with this optimal frame configuration that generalizes effectively across most scenarios, maintaining consistent total token budget typically enhances performance and transferability. Subsequently, we assign visual tokens for each frame according to the target token quantities N using a dynamic token assigner. We examine three distinct dynamic token assigners as follows:

(1) Bilinear Interpolation. A straightforward approach involves employing bilinear interpolation to resize the feature maps. For the i-th input frame embeddings $\mathbf{E}_i \in \mathbb{R}^{(H\times W)\times C}$ and its corresponding target token quantities N_i, we compute optimal spatial dimensions $\hat{H}_i$ and $\hat{W}_i$ that satisfy $\hat{H}_i \times \hat{W}_i$ being closest to but not exceeding N_i:

$$\begin{aligned} &\hat{H}_i = \hat{W}_i = \lfloor \sqrt{N_i} \rfloor \\ &\text{if } \hat{H}_i \times \hat{W}_i < N_i : \\ &\qquad \begin{cases} \hat{H}_i \leftarrow \hat{H}_i + 1 & \text{if } (\hat{H}_i + 1) \times \hat{W}_i \leq N_i \\ \hat{W}_i \leftarrow \hat{W}_i + 1 & \text{otherwise.} \end{cases} \end{aligned} \tag{1}$$

Then, bilinear interpolation is applied to transform the original embeddings $\mathbf{E}_i$ into $\hat{\mathbf{E}}_i \in \mathbb{R}^{(\hat{H}_i \times \hat{W}_i)\times C}$:

$$\hat{\mathbf{E}}_i = \texttt{B\_Interpolate}(\mathbf{E}_i, [\hat{H}_i, \hat{W}_i]) \tag{2}$$

(2) Adaptive Pooling. Application of pooling operations represents an intuitive approach. However, as a down-sampling technique, it requires the target token quantities to satisfy $N_i \leq H \times W$. Consequently, the normalized weights S_n require further processing. For the i-th input frame whose target token allocation exceeds spatial constraints (i.e., $N_i > H \times W$), we redistribute their excess weights to other frames proportionally. Let $\mathcal{L} = \{i | S_n^i \cdot N > H \times W\}$ denote frames exceeding the limit, the adjusted weights are:

$$\hat{S_n^i} = \begin{cases} \frac{H\times W}{N_t} & \text{if } i \in \mathcal{L} \\ S_n^i + \frac{S_n^i}{\sum_{j\notin\mathcal{L}} S_n^j} \sum_{k\in\mathcal{L}} (S_n^k - \frac{H\times W}{N_t}) & \text{otherwise} \end{cases} \tag{3}$$

The adjusted target quantities of tokens for the i-th frame is subsequently derived as $\hat{N}_i = \hat{S_n^i} \times N_t$. We compute optimal spatial dimensions $\hat{H}_i$ and $\hat{W}_i$, employing the methodology delineated in Equation 1. Finally, adaptive pooling is applied to transform the original embeddings $\mathbf{E}_i$ into $\hat{\mathbf{E}}_i$:

$$\hat{\mathbf{E}}_i = \texttt{A\_AvgPool2d}(\mathbf{E}_i, [\hat{H}_i, \hat{W}_i]) \tag{4}$$

(3) Dynamic Token Merging. It implements a token merging (down-sampling) operation predicated on cosine similarity metrics between distinct tokens within a visual representation, as introduced in ToMe (Bolya et al. 2023). We compute the optimal spatial dimensions $\hat{H}_i$ and $\hat{W}_i$ utilizing methodologies analogous to Adaptive Pooling, subsequently apply the k-th Bipartite Soft Matching alternately along rows and columns to transform the original embeddings $\mathbf{E}_i$ into $\hat{\mathbf{E}}_i$:

$$\hat{\mathbf{E}}_i = \texttt{B\_SoftMatching}(\mathbf{E}_i, [\hat{H}_i, \hat{W}_i]) \tag{5}$$

Dynamic Frame Sampling

In QuoTA, considering that the information redundancy in the video can be mitigated through dynamic visual token assignment, we accommodate additional input frames within the LVLM to capture more potential critical content. Consequently, we implement uniform frame sampling with adaptive quantity parameters determined by video duration. Specifically, for a video spanning t seconds, the sampled frame count T is calculated according to:

$$T = T_{base} + \min(\lfloor \frac{t}{3600} \times \alpha \rfloor, \alpha) \tag{6}$$

where T_{base} is the base number of frames (e.g., 96), and α is a hyperparameter that controls the upper bound of additional frames (e.g., 64). This formulation ensures that long video sequences receive proportionally increased sampling density to capture salient information while maintaining computational efficiency by capping the maximum additional frames at α. The frames are then uniformly sampled across the video timeline at intervals of t/T seconds.

Experiments

Datasets

To ensure robustness, we evaluated QuoTA across six datasets: **Video-MME** (Fu et al. 2024a): A widely used benchmark for assessing the ability of LVLMs to handle detailed videos in real-world scenarios that vary in lengths. **MLVU** (Zhou et al. 2024): A large-scale long video benchmark with 9 distinct tasks and diversified lengths, ranging from 3 minutes to 2 hours. **LongVideoBench** (Wu et al. 2024a): A benchmark designed to accurately retrieve and reason over detailed multimodal information from long videos with 17 fine-grained categories. **VNBench** (Zhao et al. 2024): A synthetic benchmark designed to evaluate models' long-context abilities, covering tasks such as retrieval, ordering, and counting. **MVBench** (Li et al. 2024b): A benchmark cross over 20 challenging video understanding tasks, focusing on temporal understanding in dynamic video tasks. **NeXT-QA** (Xiao et al. 2021): A short-video benchmark emphasizing causal and temporal reasoning, challenging models to understand complex sequences.

Implementation Details

We performed all the experiments on NVIDIA A100 40G GPUs. We extend LLaVA-Video (Zhang et al. 2024c) and LLaVA-OneVision (Li et al. 2024a) with our QuoTA at 7B-scale, constrained by available computational resources. Bilinear Interpolation is our dynamic token assigner, offering enhanced flexibility in up- and down-sampling operations while demonstrating superior performance, as evidenced in Table 4. For equitable comparative analysis, we align the total visual tokens budget N_t with each LVLM's original total visual token quantities during inference. To optimize the quality-efficiency trade-off, we use Qwen2-VL-2B (Wang et al. 2024b) as our scoring LVLM. The base frame quantity T_{base} is configured at 96 while the maximum additional frame α is set to 64 across all benchmarks except for VNBench (Zhao et al. 2024) ($T_{base} = 128$, $\alpha = 96$).

Main Results

We evaluate QuoTA implemented within LLaVA-Video (Zhang et al. 2024c) and LLaVA-OneVision (Li et al. 2024a) at 7B-scale, maintaining equivalent computational constraints (total visual tokens budget N_t) as the baseline across three long video understanding benchmarks: LongVideoBench (Wu et al. 2024a), MLVU (Zhou et al. 2024) and Video-MME (Fu et al. 2024a). The empirical outcomes presented in Table 1 demonstrate that QuoTA integration into LLaVA-Video-7B (Zhang et al. 2024c), yielding improvements of 0.8%, 1.1%, and 2.6% on LongVideoBench (Wu et al. 2024a), MLVU (Zhou et al. 2024), and Video-MME (Fu et al. 2024a) (w/o subtitles), respectively. Notably, substantial enhancements manifest in extended-duration video (spanning 30-60 minutes) within Video-MME (Fu et al. 2024a) (47.7% $\rightarrow$ 52.2% for LLaVA-OneVision (Li et al. 2024a), and 51.8% $\rightarrow$ 55.7% for LLaVA-Video (Zhang et al. 2024c)) under "w/o subtitles" conditions, substantiating that our query-oriented token assignment methodology together with dynamic frame sampling strategy, effectively mitigates information redundancy (particularly in long videos) while accentuating salient content, thereby facilitating enhanced model activation and comprehension of complex visual narratives.

Furthermore, we conduct evaluations of QuoTA on two conventional video understanding benchmarks, MVBench (Li et al. 2024b) and NeXT-QA (Xiao et al. 2021), alongside the specifically constructed Needle-In-A-Haystack video benchmark VBNench (Zhao et al. 2024), as illustrated in Table 2. Particularly noteworthy is the substantial enhancement observed on VNBench (Zhao et al. 2024) (44.7% $\rightarrow$ 49.3% for LLaVA-OneVision (Li et al. 2024a) and 54.4% $\rightarrow$ 64.7% for LLaVA-Video (Zhang et al. 2024c)), empirically validating that our query-oriented frame-wise scoring methodology effectively directs the LVLM's attention toward query-relevant keyframes. Additionally, as shown by Figure 1, our proposed QuoTA demonstrates superior efficacy that outperforms recent SoTAs, FrameFusion (Fu et al. 2024b) and AIM (Zhong et al. 2024) at distinct relative visual token budget allocations. Notably, QuoTA establishes new SoTA across five benchmarks.

Ablation Studies

Effect of different components of QuoTA. To investigate the efficacy of QuoTA's components, we conduct ablation experiments with varied configurations to evaluate LLaVA-Video-7B (Zhang et al. 2024c) performance on Video-MME (Fu et al. 2024a) and VNBench (Zhao et al. 2024). As shown in Table 3, when compared to fixed-length sampling (96-frames with $\sim$131 visual tokens per frame, maintaining equivalent token budget N_t as the baseline of 64-frames with 196 visual tokens per frame), dynamic-length sampling (employing adaptive frame sampling 96$\sim$160 frames while preserving consistent token budget N_t) exhibits superior performance only when augmented with our proposed LVLM-based frame scoring methodology for visual token assignment. Additionally, implementing the Chain-of-Thoughts-driven query decoupling technique yields substantial performance enhancement by improving scoring precision, especially in VNBench (Zhao et al. 2024). These empirical findings indicate that increasing frame sampling without discriminative selection provides limited improvement due to query-irrelevant information redundancy.

Effect of different visual token assignment strategy. To evaluate the efficacy of distinct token assigners, we conducted an ablation study implementing QuoTA within LLaVA-Video-7B (Zhang et al. 2024c) on the Video-MME (Fu et al. 2024a) and VNBench (Zhao et al. 2024). As detailed in Table 4, optimal performance is achieved with bilinear interpolation. Despite the token similarity-based merg-

Model	Params	Frames	LongVideo Bench (val)	MLVU (m-avg)	Video-MME (wo/w-subtitles) Short	Medium	Long	Overall
Proprietary LVLMs								
GPT-4o (OpenAI 2024)	-	384	66.7	64.0	80.0/82.8	70.3/76.6	65.3/72.1	71.9/77.2
Gemini-1.5-Pro (Reid et al. 2024)	-	0.5 fps	64.0	-	81.7/84.5	74.3/81.0	67.4/77.4	75.0/81.3
Open-Source LVLMs								
LongVA (Zhang et al. 2024a)	7B	128	-	56.3	61.1/61.6	50.4/53.6	46.2/47.6	52.6/54.3
Long-LLaVA (Yin Song et al. 2024)	7B	64	-	-	61.9/66.2	51.4/54.7	45.4/50.3	52.9/57.1
Video-XL (Shu et al. 2024)	7B	128	50.7	64.9	64.0/67.4	53.2/60.7	49.2/54.9	55.5/61.0
TimeMarker (Chen et al. 2024c)	8B	128	56.3	63.9	71.0/75.8	54.4/60.7	46.4/51.9	57.3/62.8
AIM (Zhong et al. 2024)	7B	32*	-	69.3	-/-	-/-	-/-	59.2/62.3
LongVILA (Xue et al. 2024)	7B	256	57.1	-	69.0/72.9	58.3/64.9	53.0/57.4	60.1/65.1
LongVU (Shen et al. 2024)	7B	-	-	65.4	-/-	-/-	59.5/-	60.6/-
Qwen2-VL (Wang et al. 2024b)	7B	-	-	64.8	-/-	-/-	-/-	63.3/69.0
ReTaKe (Wang et al. 2024c)	7B	-	-	69.8	72.8/-	62.7/-	56.2/-	63.9/-
NVILA (Liu et al. 2024b)	8B	256	57.7	70.1	75.7/77.6	62.2/**69.0**	54.9/63.3	64.2/**70.0**
LLaVA-OV (Li et al. 2024a)	7B	64	56.3	64.7	70.2/74.0	56.6/64.2	47.7/62.4	58.1/66.9
LLaVA-OV + QuoTA	7B	64*	57.4	69.7	71.1/74.8	58.8/65.2	52.2/**63.9**	60.7/68.0
LLaVA-Video (Zhang et al. 2024c)	7B	64	58.2	70.8	75.4/77.3	62.6/67.7	51.8/63.6	63.3/69.5
LLaVA-Video + QuoTA	7B	64*	**59.0**	**71.9**	**77.1/79.0**	**64.9**/68.0	**55.7**/62.9	**65.9/70.0**

Table 1: Performance on the validation set of LongVideoBench (Wu et al. 2024a), MLVU (Zhou et al. 2024) and Video-MME (Fu et al. 2024a). By applying QuoTA to LLaVA-Video-7B (Zhang et al. 2024c), we observed an average performance improvement of 1.5% across three long video understanding benchmarks while setting a new state-of-the-art. * denotes using the same visual token budget as the baseline. Models in parentheses represent the baselines they used.

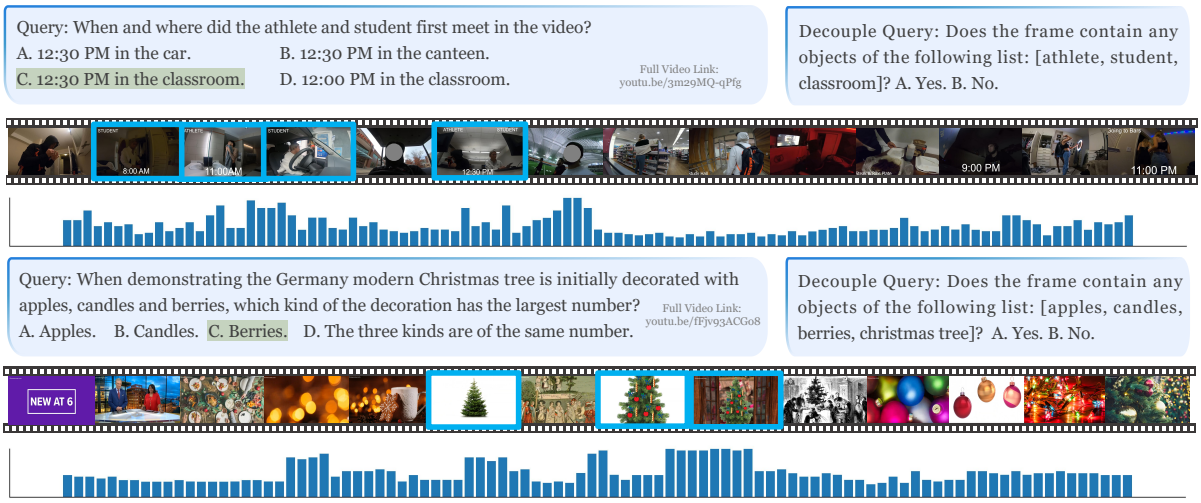

Figure 3: Qualitative result shown in Video-MME (Fu et al. 2024a) benchmark when applying QuoTA with LLaVA-Video-7B (Zhang et al. 2024c). The frames with a blue border are query-oriented keyframes, and the bar chart shows the scores.

ing approach employed by ToMe (Bolya et al. 2023), we contend that such a methodology disrupts spatial coherence in video representations, consequently impeding effective cross-modal interaction during early fusion stages. Similarly, adaptive pooling encounters analogous limitations, as it compromises continuous spatial structure within visual features, potentially degrading spatial attention quality and cross-modal alignment precision. Conversely, bilinear interpolation provides flexibility, supporting both up and downsampling operations while preserving spatial continuity and maintaining stable cross-modal information propagation, facilitating robust feature learning and cross-modal integration. The results further suggest that prevalent attention-based token assignment methodologies, which prioritize token similarity for merging operations, may represent suboptimal strategic approaches.

Effect of different query-oriented frame scoring strategy. To evaluate the efficacy of query decouple via CoT, we conducted ablation experiments presented in Table 5, indicating that decoupling queries with an emphasis on video

Model	MVB	N-QA	VNB
Proprietary LVLMs			
Gemini-1.5-Pro (Reid et al. 2024)	-	-	66.7
GPT-4o (OpenAI 2024)	-	-	64.4
Open-Source LVLMs			
LongVA (Zhang et al. 2024a)	-	68.3	41.5
mPLUG-Owl3 (Ye et al. 2024)	54.5	76.8	-
Video-XL (Shu et al. 2024)	55.3	-	61.6
LongVILA (Xue et al. 2024)	**67.1**	80.7	63.0
LLaVA-OV (Li et al. 2024a)	56.7	79.4	44.7
LLaVA-OV + QuoTA	57.3	80.4	49.3
LLaVA-Video (Zhang et al. 2024c)	58.6	83.2	54.4
LLaVA-Video + QuoTA	62.1	**83.9**	**64.7**

Table 2: The overall performance on MVBench (Li et al. 2024b), VNBench (Zhao et al. 2024) and NeXT-QA (Xiao et al. 2021) at 7B-scale LVLMs with the setting of the original frame rates. By applying QuoTA to LLaVA-Video-7B (Zhang et al. 2024c), we observed an average improvement of 4.8% across three benchmarks, especially a 10.3% improvement on the Needle-In-A-Haystack benchmark VNBench (Zhao et al. 2024), which set a new SOTA, demonstrating QuoTA assists query-oriented keyframes focusing.

Fix-len.	Dy-len.	Wei.	CoT-Dec.	V-MME	VNBench
				63.3	54.4
✓				64.0	58.7
	✓			63.5	58.4
✓		✓		63.6	49.0
	✓	✓		64.4	48.6
✓		✓	✓	64.2	**60.9**
	✓	✓	✓	**65.9**	60.6

Table 3: Results on combinations of different components in Video-MME (Fu et al. 2024a) and VNBench (Zhao et al. 2024) when using LLaVA-Video-7B (Zhang et al. 2024c) as the based LVLM on QuoTA. **Fix-len.** and **Dy-len.** represent fix sampled 96-frame and dynamic sampled 96∼160 frames with the same token budget N_t as the baseline, respectively. **Wei.** and **CoT-Dec.** denote the LVLM-based frame scoring and CoT-Driven Query Decouple, respectively.

event identification with slight performance decreases in both benchmarks, suggesting that entity-based representations constitute fundamental and generalizable features for video understanding. Notably, when encountering summarization tasks, QuoTA can evenly distribute tokens without forming clusters. Furthermore, CLIP (Radford et al. 2021) resulted in substantial performance degradation, attributable to CLIP's propensity to prioritize visually anomalous frames (e.g., those with overexposure), consequently compromising accurate query-oriented frame selection.

Qualitative Evaluation

We conduct qualitative analyses from Video-MME (Fu et al. 2024a) in Figure 3, which illustrate normalized frame scores alongside selected sampled frames, with keyframes highlighted by blue borders. As demonstrated, QuoTA enables preferential token allocation to keyframes while proportionally reducing token assignment to irrelevant frames of LLaVA-Video-7B (Zhang et al. 2024c), effectively mitigating visual redundancy and facilitating more precise task-specific responses to user queries.

Visual Token Assigner	Video-MME S	Video-MME M	Video-MME L	Video-MME O	VNBench
None	75.4	62.6	51.8	63.3	54.4
Bilinear Interpolation	**77.1**	**64.9**	**55.7**	**65.9**	64.7
Adaptive Pooling	75.7	63.0	53.1	63.9	**64.8**
Dynamic Token Merging	76.4	62.8	54.3	64.5	63.0

Table 4: Results on different token assignment strategies when extending QuoTA with LLaVA-Video-7B (Zhang et al. 2024c) on Video-MME (Fu et al. 2024a) and VNBench (Zhao et al. 2024). "None" represents the baseline.

Scoring Strategy	Video-MME S	Video-MME M	Video-MME L	Video-MME O	VNBench
None	75.4	62.6	51.8	63.3	54.4
LVLM-Based	76.0	63.3	54.0	64.4	49.5
LVLM-CoT-Entity	**77.1**	**64.9**	55.7	**65.9**	**64.7**
LVLM-CoT-Event	76.2	63.9	**55.8**	65.3	64.4
CLIP-CoT-Entity	75.8	63.6	52.8	64.0	62.7

Table 5: Results on different frame scoring strategies when extending QuoTA with LLaVA-Video-7B (Zhang et al. 2024c) on Video-MME (Fu et al. 2024a) and VNBench (Zhao et al. 2024). "None" represents the baseline.

Conclusion

We present **QuoTA**, which presents a training-free framework that employs query-oriented visual token assignment for enhanced long video understanding. By leveraging Chain-of-Thoughts reasoning for query-specific frame importance assessment, we achieved a 3.2% average improvement on LLaVA-Video-7B across six benchmarks while maintaining computational efficiency. The plug-and-play nature of **QuoTA** ensures seamless integration with existing LVLMs without additional training requirements. Our work demonstrates that query awareness and intelligent token assignment are fundamental to addressing the information redundancy challenges in long video understanding tasks.

Acknowledgments

This work is supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. 62576299, No. U21B2037, No. U22B2051, No. U23A20383, No. U21A20472, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), the Natural Science Foundation of Fujian Province of China (No. 2021J06003, No.2022J06001) and the Fundamental Research Funds for the Central Universities.

References

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT but Faster. In *International Conference on Learning Representations*.

Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Lin, B.; Tang, Z.; et al. 2024a. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. *arXiv preprint arXiv:2406.04325*.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024b. 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*, 19–35. Springer.

Chen, S.; Lan, X.; Yuan, Y.; Jie, Z.; and Ma, L. 2024c. TimeMarker: A Versatile Video-LLM for Long and Short Video Understanding with Superior Temporal Localization Ability. *arXiv preprint arXiv:2411.18211*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024d. 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*, 24185–24198.

Choudhury, R.; Zhu, G.; Liu, S.; Niinuma, K.; Kitani, K.; and Jeni, L. 2025. Don't Look Twice: Faster Video Transformers with Run-Length Tokenization. *Advances in Neural Information Processing Systems*, 37: 28127–28149.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024a. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075*.

Fu, T.; Liu, T.; Han, Q.; Dai, G.; Yan, S.; Yang, H.; Ning, X.; and Wang, Y. 2024b. FrameFusion: Combining Similarity and Importance for Video Token Reduction on Large Visual Language Models. *arXiv preprint arXiv:2501.01986*.

Jin, P.; Takanobu, R.; Zhang, C.; Cao, X.; and Yuan, L. 2023. Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. *arXiv preprint arXiv:2311.08046*.

Lee, S.-H.; Wang, J.; Zhang, Z.; Fan, D.; and Li, X. 2024. Video token merging for long-form video understanding. *arXiv preprint arXiv:2410.23782*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, X.; Wang, Y.; Yu, J.; Zeng, X.; Zhu, Y.; Huang, H.; Gao, J.; Li, K.; He, Y.; Wang, C.; et al. 2024c. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; Li, X.; Fang, Y.; Chen, Y.; Hsieh, C.-Y.; Huang, D.-A.; Cheng, A.-C.; Nath, V.; Hu, J.; Liu, S.; Krishna, R.; Xu, D.; Wang, X.; Molchanov, P.; Kautz, J.; Yin, H.; Han, S.; and Lu, Y. 2024b. NVILA: Efficient Frontier Visual Language Models. arXiv:2412.04468.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

OpenAI. 2024. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Shang, Y.; Xu, B.; Kang, W.; Cai, M.; Li, Y.; Wen, Z.; Dong, Z.; Keutzer, K.; Lee, Y. J.; and Yan, Y. 2024. Interpolating Video-LLMs: Toward Longer-sequence LMMs in a Training-free Manner. *arXiv preprint arXiv:2409.12963*.

Shen, X.; Xiong, Y.; Zhao, C.; Wu, L.; Chen, J.; Zhu, C.; Liu, Z.; Xiao, F.; Varadarajan, B.; Bordes, F.; et al. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

Shen, Y.; Fu, C.; Dong, S.; Wang, X.; Chen, P.; Zhang, M.; Cao, H.; Li, K.; Zheng, X.; Zhang, Y.; et al. 2025. Long-VITA: Scaling Large Multi-modal Models to 1 Million Tokens with Leading Short-Context Accuray. *arXiv preprint arXiv:2502.05177*.

Shu, Y.; Zhang, P.; Liu, Z.; Qin, M.; Zhou, J.; Huang, T.; and Zhao, B. 2024. Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding. *arXiv preprint arXiv:2409.14485*.

Tao, K.; Qin, C.; You, H.; Sui, Y.; and Wang, H. 2024. DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models. *arXiv preprint arXiv:2411.15024*.

Wang, H.; Nie, Y.; Ye, Y.; GuanYu, D.; Wang, Y.; Li, S.; Yu, H.; Lu, J.; and Huang, C. 2024a. Dynamic-VLM: Simple Dynamic Visual Token Compression for VideoLLM. *arXiv preprint arXiv:2412.09530*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Si, Q.; Wu, J.; Zhu, S.; Cao, L.; and Nie, L. 2024c. ReTaKe: Reducing Temporal and Knowledge Redundancy for Long Video Understanding. *arXiv preprint arXiv:2412.20504*.

Wang, X.; Song, D.; Chen, S.; Zhang, C.; and Wang, B. 2024d. LongLLaVA: Scaling Multi-modal LLMs to 1000 Images Efficiently via Hybrid Architecture. *arXiv preprint arXiv:2409.02889*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024a. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*.

Wu, Q.; Lin, W.; Ye, W.; Zhou, Y.; Sun, X.; and Ji, R. 2024b. Accelerating Multimodal Large Language Models via Dynamic Visual-Token Exit and the Empirical Findings. *arXiv preprint arXiv:2411.19628*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xue, F.; Chen, Y.; Li, D.; Hu, Q.; Zhu, L.; Li, X.; Fang, Y.; Tang, H.; Yang, S.; Liu, Z.; et al. 2024. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*.

Yang, C.; Dong, X.; Zhu, X.; Su, W.; Wang, J.; Tian, H.; Chen, Z.; Wang, W.; Lu, L.; and Dai, J. 2024. PVC: Progressive Visual Token Compression for Unified Image and Video Processing in Large Vision-Language Models. *arXiv preprint arXiv:2412.09613*.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. arXiv:2408.04840.

Yin Song et al. 2024. aws-prototyping/long-llava-qwen2-7b.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024a. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Zhang, S.; Fang, Q.; Yang, Z.; and Feng, Y. 2025. LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token. *arXiv preprint arXiv:2501.03895*.

Zhang, Y.; Li, B.; Liu, h.; Lee, Y. j.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024b. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024c. Video Instruction Tuning With Synthetic Data. arXiv:2410.02713.

Zhao, Z.; Lu, H.; Huo, Y.; Du, Y.; Yue, T.; Guo, L.; Wang, B.; Chen, W.; and Liu, J. 2024. Needle In A Video Haystack: A Scalable Synthetic Framework for Benchmarking Video MLLMs. *arXiv preprint*.

Zhong, Y.; Liu, Z.; Li, Y.; and Wang, L. 2024. Aim: Adaptive inference of multi-modal llms via token merging and pruning. *arXiv preprint arXiv:2412.03248*.

Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Xiao, S.; Yang, X.; Xiong, Y.; Zhang, B.; Huang, T.; and Liu, Z. 2024. MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding. *arXiv preprint arXiv:2406.04264*.