

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

Yongdong Luo<sup>1</sup>, Wang Chen<sup>1</sup>, Xiawu Zheng<sup>1</sup>, Weizhong Huang<sup>1</sup>, Shukang Yin,  
Haojia Lin<sup>1</sup>, Chaoyou Fu<sup>2</sup>, Jinfa Huang<sup>3</sup>, Jiayi Ji<sup>1</sup>, Jiebo Luo<sup>3</sup>, Rongrong Ji<sup>1</sup>  
<sup>1</sup>Xiamen University   <sup>2</sup>Nanjing University   <sup>3</sup>University of Rochester

## 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. Extensive 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. Codes are open-sourced at [github link](#).

## 1. Introduction

With the emergence of advanced Large Language Models (LLMs), researchers have expanded their capabilities to video comprehension [2, 5, 11, 14, 17, 44, 47], establishing the domain of Large Video-Language Models (LVLMs). Recent studies [25, 34, 40, 43, 45] focus on extending LVLMs' reasoning context capacity, primarily through fine-tuning approaches for enhanced long video understanding. However, empirical evidence from long-context LVLMs

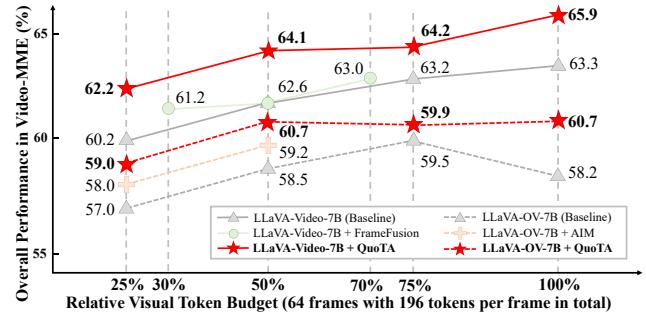


Figure 1. Comparative analysis of Video-MME [7] when implementing attention-based token assignment methods AIM [50] and FrameFusion [9], alongside our proposed query-oriented QuoTA within LLaVA-Video-7B [48] and LLaVA-OV-7B [13] across varied relative visual token budgets. QuoTA demonstrates superior efficacy while exhibiting consistent performance enhancement across diverse token budget configurations relative to the baseline.

[45] and [27] demonstrate performance degradation on Video-MME [7] when frame sampling rates increase, suggesting that merely augmenting frame quantities introduces information redundancy while imposing greater computational demands on complex reasoning tasks.

Recent works [9, 30, 33, 50] have addressed visual token reduction through the analysis of attention value distributions across model layers, thereby providing more informationally efficient tokens for long video comprehension tasks. DyCoke [30], 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 layers, they face substantial limitations: **(i) Neglected task-specific token relevance.** Human visual cognition inherently focuses on query-relevant content when processing video-text tasks, highlighting the necessity of query-oriented token selection to align visual processing with task objectives while ensuring efficient token allocation and se-

mantic coherence. Conversely, visual tokens with high attention-weight responses primarily reflect inter-token associative strength rather than their direct pertinence to the query. As demonstrated by [38, 50], cross-modal interactions mainly occur in early fusion stages, with later layers showing predominantly intra-modal patterns. Additionally, FrameFusion [9] 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 token selection in subsequent computational stages.

To address this issue, we propose QuoTA, an **antehoc** 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 an external lightweight LVLM (termed *scoring LVLM*) to generate query-relevance scores, subsequently employing these metrics as discriminative criteria for visual token assignment. Our query-oriented token assignment strategy enhanced cross-modal interactions while minimizing redundancy between visual and text tokens during 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 [36] reasoning. These reformulated questions subsequently prompt the *scoring LVLM* in generating a query-specific relevance score for each sampled video 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 (more sampling frames in longer videos and vice versa) 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 [7], MLVU [51], LongVideoBench [37], VN Bench [49], MV Bench [15], and NeXT-QA [39]. Results demonstrate that LLaVA-Video-7B [48] and LLaVA-OneVision-7B [13], 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 [50] and FrameFusion [9] across varying visual token budgets when applied to LLaVA-Video-7B [48] and LLaVA-OneVision-7B [13], while maintaining consistent performance improvements over the baseline regardless of token budget, which is set to 64 frames with 196 tokens per frame (12,544 visual tokens in total).

In summary, our contributions are as follows:

- **We design a versatile plug-and-play pipeline for existing LVLMs:** QuoTA provides a training-free solution applicable to diverse LVLMs, enhancing long video understanding performance 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.

## 2. Related Work

### 2.1. Large Video-Language Models

Recent advances in large language models (LLMs) have sparked interest in developing video understanding systems. Video-ChatGPT [21] processes videos by extracting frame-level features through spatial and temporal pooling. VideoChat [14] combines textual descriptions with video appearance embeddings. Video-LLaVA [17] aligns image and video encoders using a shared projector to map representations into a common language space. LLaVA-NeXT-Video [47] extends LLaVA-NeXT [18] through video-specific fine-tuning. While these models show promise, they still struggle with analyzing long-form videos.

### 2.2. Long Video Understanding

Recent research has focused on expanding context window sizes for long video understanding. LongVA [45], Video-XL [28], and LongVILA [40] 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. Several works [10, 20, 29, 35] have employed LLMs with tools to process video as structured text for question-answering. However, these methods are limited by extensive processing times and reliance on proprietary models (e.g., GPT-4o).

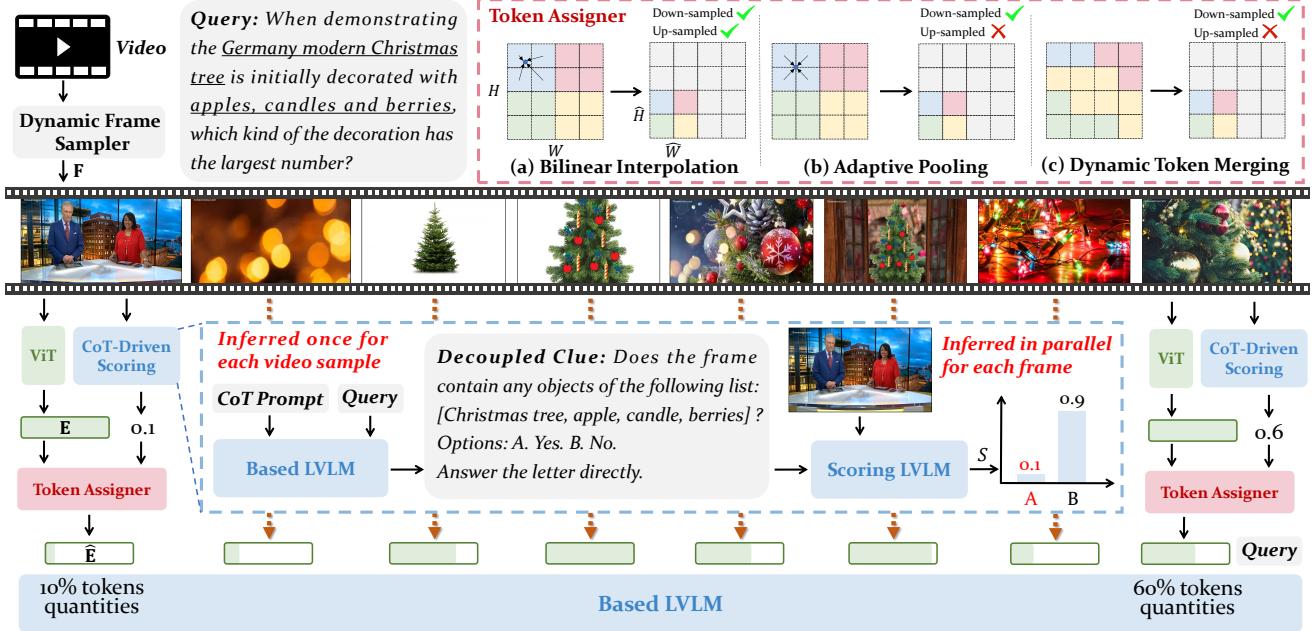


Figure 2. The framework of QuoTA. Initially, a dynamic frame sampler extracts  $T$  frames from the video based on its duration, which are subsequently processed by ViT to generate visual embeddings  $\mathbf{E}$ . Then, the based LVLM decouples the input query using Chain-of-Thoughts [36] reasoning into a decoupled clue to generate frame-wise importance scores through scoring LVLM in parallel, thus evaluating the relevance to the query of each frame. Finally, a token assigner rescales the frame embeddings to  $\hat{\mathbf{E}}$  based on these importance scores.

### 2.3. Efficient Visual Modeling

Previous research in multi-modal LLMs assigns visual tokens to reduce spatial redundancy. FastV [3] reduces them at a particular selected layer based on the distribution of attention. Recent works [9, 30, 33, 50] extend it by reducing the redundancy of spatiotemporal information for long video understanding. For example, AIM [50] merges similar tokens and then preserves the 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 [9], token importance is inconsistent across different layers; thus, hierarchical reduction is susceptible to error propagation from early suboptimal decisions. Other works [6, 12, 16, 26, 31, 41, 46] integrate token assignment into the architecture design and enhance model capabilities through training. However, the training-needed schema makes them inflexible and less generalizable.

## 3. Method

### 3.1. Preliminaries

For a given video  $\mathbf{V}$ , a frame sampler extracts  $T$  frames  $\mathbf{F} = \{\mathbf{F}_i\}_{i=1}^T$ . Conventional approaches typically employ uniform sampling at fixed intervals for computational

efficiency. Features of each frame are then obtained via  $\mathbf{E} = \{\mathbf{E}_i\}_{i=1}^T = \text{ViT}(\mathbf{F})$ , where ViT represents a transformer-based visual encoder (e.g., CLIP-L [23]), 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  $\mathbf{O} = \text{LVLM}(\mathbf{E}, \mathbf{Q})$ . Our proposed 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 thus formulated as  $\mathbf{O} = \text{LVLM}(\hat{\mathbf{E}}, \mathbf{Q})$ . Figure 2 illustrates the overall architecture of our approach.

### 3.2. 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 [23] 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 binary-choice prompt structure fed into the scoring LVLM for each frame follows the format:

#### Prompt for Frame Scoring Based on Source Query

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 computational efficiency, we employ Qwen2-VL-2B [32] 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  $S_n = \{S_n^i\}_{i=1}^T = S^i / \sum_{j=1}^T S^j$ .

### 3.3. CoT-Driven Query Decouple

Employing the prompt described in Section 3.2 yields only marginal improvements, as the scoring LVLM tends to optimistically presume that frames inherently contain adequate information. This results in homogeneous importance weights across frames, compromising discriminative capacity. To address this limitation, we implement a Chain-of-Thoughts [36] approach that decouples the query that enhances frame differentiation quality, as validated in our ablation studies (Section 4.4). Considering that LVLMs demonstrate superior capability in identifying physical entities within visual content, QuoTA utilizes the based LVLM to decouple the original query into a structured object list. Specifically, we implement a three-step prompting protocol (the prompt is detailed in the Appendix) 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:

#### Prompt for Frame Scoring Based on Entity List

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 prompt in Section 3.2. We also conduct other CoT-driven decouple strategies for query-oriented keyframes scoring, which details in Section 4.4.

### 3.4. 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. This budget can be user-defined to accommodate specific computational resource constraints. 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 [48] 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. In this study, we examine three distinct categories of 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  $N_i$ :

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

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

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

**(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:

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

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  $\hat{\mathbf{E}}_i$ :

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

**(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 [1]. 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  $\hat{\mathbf{E}}_i$ :

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

### 3.5. 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:

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

where  $T_{base}$  is the base number of frames (e.g., 96), and  $\alpha$  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  $\alpha$ . The frames are then uniformly sampled across the video timeline at intervals of  $t/T$  seconds.

## 4. Experiments

### 4.1. Datasets

To ensure robustness, we evaluated QuoTA across six datasets: **Video-MME** [7]: A widely used benchmark for assessing the ability of LVLMs to handle detailed videos in real-world scenarios that varying lengths (short, medium, long). **MLVU** [51]: A large-scale long video benchmark with a large wide of 9 distinct tasks and diversified lengths, ranging from 3 minutes to 2 hours. **LongVideoBench** [37]: A benchmark designed to accurately retrieve and reason over detailed multimodal information from long videos with 17 fine-grained categories. **VNBench** [49]: A synthetic

benchmark designed to evaluate models' long-context abilities, covering tasks such as retrieval, ordering, and counting. **MVBench** [15]: A benchmark cross over 20 challenging video understanding tasks, focusing on temporal understanding in dynamic video tasks. **NeXT-QA** [39]: A short-video benchmark emphasizing causal and temporal reasoning, challenging models to understand complex sequences.

### 4.2. Implementation Details

We performed all the experiments on NVIDIA A100 40G GPUs. We extend LLaVA-Video [48] and LLaVA-OneVision [13] 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 5. 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 [32] as our scoring LVLM. The base frame quantity  $T_{base}$  is configured at 96 while the maximum additional frame  $\alpha$  is set to 64 across all benchmarks except for VNBench [49] ( $T_{base} = 128, \alpha = 96$ ).

### 4.3. Main Results

We evaluate QuoTA implemented within LLaVA-Video [48] and LLaVA-OneVision [13] at 7B-scale, maintaining equivalent computational constraints (total visual tokens budget  $N_t$ ) as the baseline across three long video understanding benchmarks: LongVideoBench [37], MLVU [51] and Video-MME [7]. The empirical outcomes presented in Table 1 demonstrate that QuoTA integration into LLaVA-Video-7B [48], yielding improvements of 0.8%, 1.1%, and 2.6% on LongVideoBench [37], MLVU [51], and Video-MME [7] (w/o subtitles), respectively. Notably, substantial performance enhancements manifest in extended-duration video (spanning 30-60 minutes) within Video-MME [7] (47.7%  $\rightarrow$  52.2% for LLaVA-OneVision [13], and 51.8%  $\rightarrow$  55.7% for LLaVA-Video [48]) under "w/o subtitles" conditions, substantiating that our query-oriented token assignment methodology together with dynamic frame sampling strategy effectively mitigates information redundancy (particularly pronounced 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 [15] and NeXT-QA [39], alongside the specifically constructed Needle-In-A-Haystack video benchmark VNBench [49], as illustrated in Table 2. Particularly noteworthy is the substantial enhancement observed on VNBench [49] (44.7%  $\rightarrow$  49.3% for LLaVA-OneVision [13] and 54.4%  $\rightarrow$  64.7% for LLaVA-Video [48]), empirically validating that

| Model                    | Params | Frames  | LongVideo Bench (val) | MLVU (m-avg) | Video-MME (wo/w-subtitles) |                   |                   |                   |
|--------------------------|--------|---------|-----------------------|--------------|----------------------------|-------------------|-------------------|-------------------|
|                          |        |         |                       |              | Short                      | Medium            | Long              | Overall           |
| <i>Proprietary LVLMs</i> |        |         |                       |              |                            |                   |                   |                   |
| GPT-4o [22]              | -      | 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 [24]      | -      | 0.5 fps | 64.0                  | -            | 81.7/84.5                  | 74.3/81.0         | 67.4/77.4         | 75.0/81.3         |
| <i>Open-Source LVLMs</i> |        |         |                       |              |                            |                   |                   |                   |
| LongVA [45]              | 7B     | 128     | -                     | 56.3         | 61.1/61.6                  | 50.4/53.6         | 46.2/47.6         | 52.6/54.3         |
| Video-XL [28]            | 7B     | 128     | 50.7                  | 64.9         | 64.0/67.4                  | 53.2/60.7         | 49.2/54.9         | 55.5/61.0         |
| VITA-1.5 [8]             | 7B     | 16      | -                     | -            | 67.0/69.9                  | 54.2/55.7         | 47.1/50.4         | 56.1/58.7         |
| TimeMarker [4]           | 8B     | 128     | 56.3                  | 63.9         | 71.0/75.8                  | 54.4/60.7         | 46.4/51.9         | 57.3/62.8         |
| AIM [50] (LLaVA-OV)      | 7B     | 32*     | -                     | 69.3         | -/-                        | -/-               | -/-               | 59.2/62.3         |
| LongVILA [40]            | 7B     | 256     | 57.1                  | -            | 69.0/72.9                  | 58.3/64.9         | 53.0/57.4         | 60.1/65.1         |
| LongVU [26]              | 7B     | -       | -                     | 65.4         | -/-                        | -/-               | 59.5/-            | 60.6/-            |
| Qwen2-VL [32]            | 7B     | -       | -                     | 64.8         | -/-                        | -/-               | -/-               | 63.3/69.0         |
| ReTaKe [33] (Qwen2-VL)   | 7B     | -       | -                     | 69.8         | 72.8/-                     | 62.7/-            | 56.2/-            | 63.9/-            |
| NVILA [19]               | 8B     | 256     | 57.7                  | 70.1         | 75.7/77.6                  | 62.2/ <b>69.0</b> | 54.9/63.3         | 64.2/ <b>70.0</b> |
| LLaVA-OV [13]            | 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/ <b>63.9</b> | 60.7/68.0         |
| LLaVA-Video [48]         | 7B     | 64      | <u>58.2</u>           | <u>70.8</u>  | 75.4/77.3                  | <u>62.6/67.7</u>  | <u>51.8/63.6</u>  | <u>63.3/69.5</u>  |
| LLaVA-Video + QuoTA      | 7B     | 64*     | <b>59.0</b>           | <b>71.9</b>  | <b>77.1/79.0</b>           | <b>64.9/68.0</b>  | <b>55.7/62.9</b>  | <b>65.9/70.0</b>  |

Table 1. Performance on the validation set of LongVideoBench [37], MLVU [51] and Video-MME [7]. By applying QuoTA to LLaVA-Video-7B [48], we observed an average performance improvement of 1.5% across three long video understanding benchmarks while setting new state-of-the-art. \* denotes using the same visual token budget as the baseline. Models in parentheses represent the baselines they used.

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 [9] and AIM [50] at distinct relative visual token budget allocations. Notably, QuoTA establishes new SoTA across five benchmarks.

#### 4.4. 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 [48] performance on Video-MME [7] and VN-Bench [49]. 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 (Section 3.5, employing adaptive frame sampling 96~160 frames while preserving consistent token budget  $N_t$ ) exhibits superior performance only when augmented with our proposed LVLM-based frame scoring methodology detailed in Section 3.2 for visual token assignment. Additionally, implementing the Chain-of-Thoughts-driven query decoupling technique outlined in Section 3.3 yields substantial performance enhancement by improv-

ing scoring precision, especially in VN-Bench [49]. These empirical findings indicate that increasing frame sampling without discriminative selection provides limited improvement due to query-irrelevant information redundancy.

**Effect of different frame-sampling strategies and visual token budget.** As mentioned in Section 3.4 and 3.5, we include several hyper-parameters in QuoTA: (i) The budget of total visual tokens  $N_t$ , (ii) the base number  $T_{base}$ , and (iii) the maximum additional quantities  $\alpha$ . We conduct experiments with different settings of these hyper-parameters when extending LLaVA-Video-7B [48] with QuoTA on Video-MME [7] and VN-Bench [49]. We carefully examined the trade-offs between accuracy and inference time: while larger values of  $T_{base}$  and  $\alpha$  generally lead to better performance, they also increase the inference latency proportionally. As shown in Table 4, setting  $T_{base} = 96$  and  $\alpha = 64$  offers an optimal balance, providing substantial accuracy gains without excessive computational overhead in both Video-MME [7] and VN-Bench [49]. While QuoTA achieves state-of-the-art performance (64.7%) on VN-Bench [49] with  $T_{base} = 128$  and  $\alpha = 96$ , it incurs significantly higher computational overhead. At 50% token budget ( $T_{base} = 96$ ,  $\alpha = 64$ ), QuoTA achieves 64.1% accuracy, surpassing the full-token baseline (63.3%) while reducing inference time. With a more aggressive reduction

| Model                    | MVBench     | NeXT-QA     | VNBench     |
|--------------------------|-------------|-------------|-------------|
| <i>Proprietary LVLMs</i> |             |             |             |
| Gemini-1.5-Pro [24]      | -           | -           | 66.7        |
| GPT-4o [22]              | -           | -           | 64.4        |
| <i>Open-Source LVLMs</i> |             |             |             |
| LongVA [45]              | -           | 68.3        | 41.5        |
| mPLUG-Owl3 [42]          | 54.5        | 76.8        | -           |
| Video-XL [28]            | 55.3        | -           | 61.6        |
| LongVILA [40]            | <b>67.1</b> | 80.7        | <u>63.0</u> |
| LLaVA-OV [13]            | 56.7        | 79.4        | 44.7        |
| LLaVA-OV + QuoTA         | 57.3        | 80.4        | 49.3        |
| LLaVA-Video [48]         | 58.6        | <u>83.2</u> | 54.4        |
| LLaVA-Video + QuoTA      | <u>62.1</u> | <b>83.9</b> | <b>64.7</b> |

Table 2. The overall performance on MVBench [15], VNBench [49] and NeXT-QA [39] at 7B-scale LVLMs with the setting of the original frame rates. By applying QuoTA to LLaVA-Video-7B [48], we observed an average performance improvement of 4.8% across three benchmarks, especially a 10.3% improvement on the Needle-In-A-Haystack benchmark VNBench [49], which set a new state-of-the-art performing better than LongVILA [40], demonstrating QuoTA assist 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        |
| ✓        |         | ✓    | ✓        | <u>64.4</u> | 48.6        |
| ✓        |         | ✓    | ✓        | 64.2        | <b>60.9</b> |
| ✓        | ✓       | ✓    | ✓        | <b>65.9</b> | <u>60.6</u> |

Table 3. Results on combinations of different components in Video-MME [7] and VNBench [49] when using Long-LLaVA-7B [43] 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 in Section 3.3 and 3.4, respectively.

to 25% token budget ( $T_{base} = 64$ ,  $\alpha = 32$ ), it maintains competitive performance with only a 1.1% accuracy drop while achieving  $2.1 \times$  inference speedup.

**Effect of different visual token assignment strategy.** In Section 3.4, we proposed three types of dynamic visual token assigners. To evaluate their comparative efficacy, we conducted an ablation study implementing QuoTA within LLaVA-Video-7B [48] on the Video-MME [7] and VNBench [49]. As detailed in Table 5, optimal performance is achieved with bilinear interpolation. Despite the token similarity-based merging approach employed by ToMe

| $T_{base}$                                               | $\alpha$ | $N_t$  | #Time  | V-MME       | VNBench     |
|----------------------------------------------------------|----------|--------|--------|-------------|-------------|
| 16                                                       | 0        | 3,136  | 4.50s  | 60.0        | 30.7        |
|                                                          | 0        | 6,272  | 9.07s  | 62.6        | 40.5        |
|                                                          | 0        | 12,544 | 15.31s | 63.3        | 54.4        |
| <i>Dynamic sampling with 25% budget (3,136 tokens)</i>   |          |        |        |             |             |
| 32                                                       | 32       | 3,136  | 6.59s  | 61.4        | 39.0        |
| 64                                                       | 32       | 3,136  | 7.42s  | <b>62.2</b> | 49.2        |
| 64                                                       | 64       | 3,136  | 8.33s  | 61.5        | <u>49.9</u> |
| 96                                                       | 64       | 3,136  | 9.54s  | <u>62.0</u> | <b>50.1</b> |
| <i>Dynamic sampling with 50% budget (6,272 tokens)</i>   |          |        |        |             |             |
| 32                                                       | 32       | 6,272  | 11.51s | 62.3        | 35.0        |
| 64                                                       | 32       | 6,272  | 12.34s | 63.3        | 47.5        |
| 64                                                       | 64       | 6,272  | 13.29s | 63.8        | 54.6        |
| 96                                                       | 64       | 6,272  | 14.13s | <b>64.1</b> | <b>58.0</b> |
| 96                                                       | 96       | 6,272  | 15.07s | 64.0        | 57.9        |
| <i>Dynamic sampling with 100% budget (12,544 tokens)</i> |          |        |        |             |             |
| 64                                                       | 64       | 12,544 | 20.85s | 64.6        | 56.0        |
| 96                                                       | 64       | 12,544 | 21.68s | <b>65.9</b> | 60.6        |
| 96                                                       | 96       | 12,544 | 22.97s | 64.6        | 60.9        |
| 128                                                      | 96       | 12,544 | 24.82s | 64.1        | <b>64.7</b> |
| 128                                                      | 128      | 12,544 | 26.49s | 64.7        | 63.7        |

Table 4. Results on different frame-sampling strategies and token budgets in Video-MME [7] and VNBench [49]. Gray rows represent baseline LLaVA-Video-7B [48] with fixed frame sampling while others represent dynamic sampling with varying  $T_{base}$  and  $\alpha$  in different budget  $N_t$  when extending the baseline with QuoTA. #Time denotes the average time during inference per sample.

[1], 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 down-sampling 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.** As detailed in Sections 3.2 and 3.3, we involve query decoupling into an object list via Chain-of-Thoughts, followed by LVLM-based frame scoring to determine query-relevant importance. To evaluate their efficacy, we conducted ablation experiments presented in Table 6, indicating that decoupling queries with an emphasis on video event

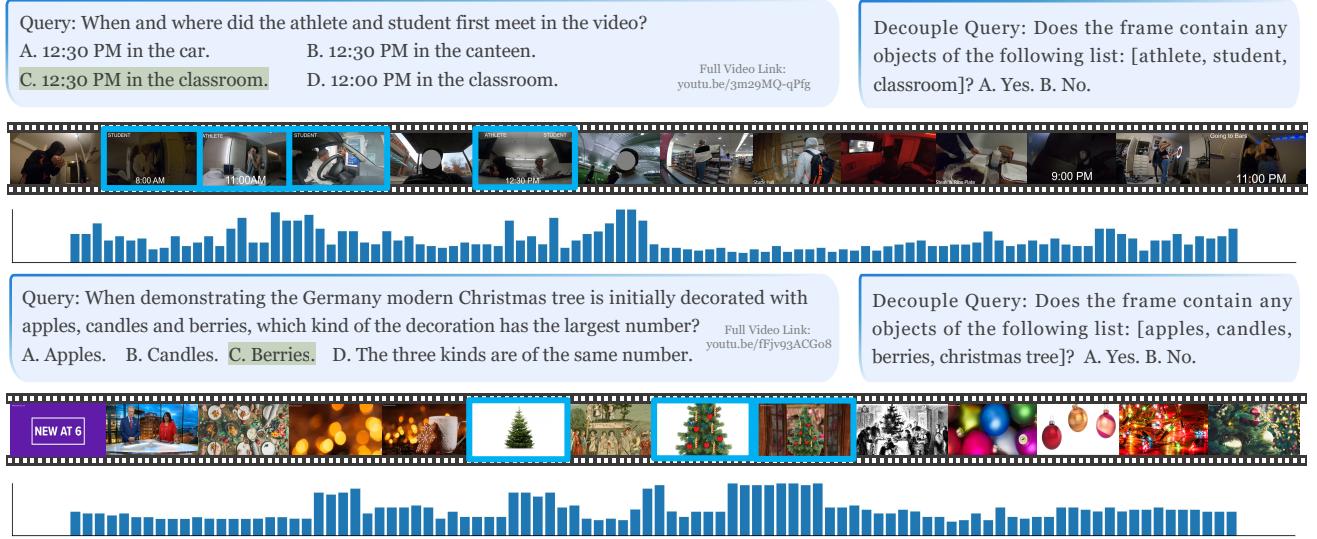


Figure 3. Qualitative result shown in Video-MME [7] benchmark when applying QuoTA with LLaVA-Video-7B [48]. The video frames with a blue border are query-oriented keyframes, and the bar chart shows the normalized scores of QuoTA for each frame.

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

Table 5. Results on different token assignment strategies when extending QuoTA with LLaVA-Video-7B [48] on Video-MME [7] and VN Bench [49]. “None” represents the baseline.

identification (decouple prompt elaborated in Appendix) with slight performance decreases both in two 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, as further demonstrated in the Appendix with the sub-tasks evaluation on Video-MME [7]. Furthermore, CLIP [23] 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.

#### 4.5. Qualitative Evaluation

We present qualitative analyses from Video-MME [7] case studies in Figure 6, which illustrates normalized frame scores alongside selected sampled frames, with query-relevant keyframes highlighted by blue borders. As demonstrated, QuoTA enables preferential token allocation to query-relevant keyframes while proportionally reducing to

| Scoring Strategy | Video-MME   |             |             |             | VN Bench    |
|------------------|-------------|-------------|-------------|-------------|-------------|
|                  | S           | M           | L           | O           |             |
| 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  | <b>77.1</b> | <b>64.9</b> | <b>55.7</b> | <b>65.9</b> | <b>64.7</b> |
| LVLM-CoT-Event   | 76.2        | 63.9        | <b>55.8</b> | 65.3        | 64.4        |
| CLIP-CoT-Entity  | 75.8        | 63.6        | 52.8        | 64.0        | 62.7        |

Table 6. Results on different frame scoring strategies when extending QuoTA with LLaVA-Video-7B [48] on Video-MME [7] and VN Bench [49]. “None” represents the baseline.

ken assignment to irrelevant frames of LLaVA-Video-7B [48], effectively mitigating visual redundancy and facilitating more precise task-specific responses to user queries.

## 5. 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.

## References

- [1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. [5](#), [7](#)
- [2] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. [1](#)
- [3] 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*, pages 19–35. Springer, 2024. [3](#)
- [4] Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-lm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024. [6](#)
- [5] Zhe Chen, Jiamnan Wu, Wenhui 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. [1](#)
- [6] Rohan Choudhury, Guanglei Zhu, Sihan Liu, Koichiro Niinuma, Kris Kitani, and László Jeni. Don’t look twice: Faster video transformers with run-length tokenization. *Advances in Neural Information Processing Systems*, 37:28127–28149, 2025. [3](#)
- [7] Chaoyou Fu, Yuhan Dai, Yondong 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. [1](#), [2](#), [5](#), [6](#), [7](#), [8](#)
- [8] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Hetting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025. [6](#)
- [9] Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large visual language models. *arXiv preprint arXiv:2501.01986*, 2024. [1](#), [2](#), [3](#), [6](#)
- [10] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. [2](#)
- [11] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023. [1](#)
- [12] Seon-Ho Lee, Jue Wang, Zhikang Zhang, David Fan, and Xinyu Li. Video token merging for long-form video understanding. *arXiv preprint arXiv:2410.23782*, 2024. [3](#)
- [13] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. [1](#), [2](#), [5](#), [6](#), [7](#)
- [14] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. [1](#), [2](#)
- [15] 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. [2](#), [5](#), [7](#)
- [16] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. [3](#)
- [17] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. [1](#), [2](#)
- [18] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. [2](#)
- [19] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2024. [6](#)
- [20] Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-rag: Visually-aligned retrieval-augmented long video comprehension, 2024. [2](#)
- [21] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. [2](#)
- [22] OpenAI. Gpt-4o system card. <https://openai.com/index/gpt-4o-system-card/>, 2024. [6](#), [7](#)
- [23] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [3](#), [8](#), [2](#)
- [24] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricu, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. [6](#), [7](#)
- [25] Yuzhang Shang, Bingxin Xu, Weitai Kang, Mu Cai, Yuheng Li, Zehao Wen, Zhen Dong, Kurt Keutzer, Yong Jae Lee,

- and Yan Yan. Interpolating video-lmms: Toward longer-sequence lmms in a training-free manner. *arXiv preprint arXiv:2409.12963*, 2024. 1
- [26] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 3, 6
- [27] Yunhang Shen, Chaoyou Fu, Shaoqi Dong, Xiong Wang, Peixian Chen, Mengdan Zhang, Haoyu Cao, Ke Li, Xiawu Zheng, Yan Zhang, et al. Long-vita: Scaling large multimodal models to 1 million tokens with leading short-context accuracy. *arXiv preprint arXiv:2502.05177*, 2025. 1
- [28] 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. 2, 6, 7
- [29] Dídac Surís, Sachit Menon, and Carl Vondrick. Viperpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023. 2
- [30] Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. *arXiv preprint arXiv:2411.15024*, 2024. 1, 3
- [31] Han Wang, Yuxiang Nie, Yongjie Ye, Deng Guan Yu, Yanjie Wang, Shuai Li, Haiyang Yu, Jinghui Lu, and Can Huang. Dynamic-vlm: Simple dynamic visual token compression for videollm. *arXiv preprint arXiv:2412.09530*, 2024. 3
- [32] 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. 4, 5, 6
- [33] Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*, 2024. 1, 3, 6
- [34] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal lmms to 1000 images efficiently via hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024. 1
- [35] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*, 2024. 2
- [36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2, 3, 4, 1
- [37] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024. 2, 5, 6
- [38] Qiong Wu, Wenhao Lin, Weihao Ye, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Accelerating multimodal large language models via dynamic visual-token exit and the empirical findings. *arXiv preprint arXiv:2411.19628*, 2024. 2
- [39] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2, 5, 7
- [40] 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. 1, 2, 6, 7
- [41] Chenyu Yang, Xuan Dong, Xizhou Zhu, Weijie Su, Jiahao Wang, Hao Tian, Zhe Chen, Wenhai Wang, Lewei Lu, and Jifeng Dai. Pvc: Progressive visual token compression for unified image and video processing in large vision-language models. *arXiv preprint arXiv:2412.09613*, 2024. 3
- [42] 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 multimodal large language models, 2024. 7
- [43] Yin Song and Chen Wu and Eden Duthie. aws-prototyping/long-llava-qwen2-7b, 2024. 1, 7
- [44] 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. 1
- [45] 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. 1, 2, 6, 7
- [46] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*, 2025. 3
- [47] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 1, 2
- [48] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 1, 2, 4, 5, 6, 7, 8
- [49] Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in a video haystack: A scalable synthetic framework for benchmarking video mllms. *arXiv preprint*, 2024. 2, 5, 6, 7, 8, 1
- [50] Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. Aim: Adaptive inference of multi-modal lmms via token merging and pruning. *arXiv preprint arXiv:2412.03248*, 2024. 1, 2, 3, 6
- [51] 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, 6

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

## Supplementary Material

### 6. CoT-Driven Decouple Prompt

As discussed in the manuscript, we employ a decoupling approach for the user’s query using Chain-of-Thoughts (CoT) reasoning [36] to achieve improved frame-level query-oriented importance scoring. This is accomplished through the use of specially designed decoupled prompts tailored for the language-vision model (LVLM). We have developed two distinct types of Chain-of-Thoughts reasoning: (i) Object List-Based and (ii) Video Event-Based, with the former demonstrating superior performance. As illustrated in the accompanying figure, we begin by clearly defining the task, followed by the inclusion of additional in-context examples to effectively prompt the LVLM. For the CoT-driven decoupling prompt related to the object list, we guide the LVLM to decouple the query by (i) assessing the necessity of entity decoupling, (ii) transforming the original query into a structured object list when appropriate, and (iii) refining this list by eliminating abstract concepts. In the case of the CoT-driven decoupling prompt for video events, we instruct the LVLM to decouple the query by (i) analyzing the type of the original question (e.g., what, who, how, where, when, why), (ii) identifying the key elements to focus on (such as objects, actions, states, and scenes), and (iii) formulating a simple, direct question to ascertain whether a frame contains these key elements.

### 7. Effect of Different Frame Scoring Strategy

We conduct experiments that evaluate the efficacy of query-oriented frame scoring strategies across diverse question types in the Video-MME [7] benchmark and VNBench [49]. Figure 6 reveals insights into our QuoTA underlying different strategies of different sub-tasks in Video-MME [7]:

**Generalization Advantage of Entity-Centric Decomposition.** The strategy based on entities demonstrates superior performance in 9 out of 12 question types compared to the baseline, with notable gains in Object Recognition (+5.1%), OCR Problems (+6.5%), and Spatial Reasoning (+7.1%). This aligns with the hypothesis that decomposing queries into entity lists (e.g., “person, bottle, table” for Object Recognition) establishes stable semantic anchors for cross-video generalization. The strategy’s dominance in Information Synopsis (79.3%) further suggests that entity-based representations mitigate noise from transient visual patterns, enabling robust aggregation of core semantics across temporal spans.

**Dynamic Context Sensitivity of Event-Driven Strategies.**

#### CoT-Driven Decouple Prompt for Object List

```
Quesiton:{query}
To answer this question, please follow the following step to give an object list
for object detection.
Step 1: Do you think it is necessary to detect specific physical entities in the
video? Please only answer "yes" or "no" in this step.
Step 2: If 'yes' in previous step, then what object in the video do you think
needs to be detected to answer this question? Please only output the name of all
the objects need to detect in the form of a Python list without other
information.
Step 3: Filter the object list provided in step 2, make sure they are physical
entities, not abstract concepts. Please only output the name of the filtered
object in the form of a Python list without other information. If all the words
being filtered, please only output null.
# Example 1:
Question: How many blue balloons are over the long table in the middle of the
room at the end of this video?
Your output should be:
Step 1: yes
Step 2: ["blue balloons", "long table", "video", "blue"]
Step 3: ["blue balloons", "long table"]
# Example 2:
Question: In the lower left corner of the video, what color is the woman wearing
on the right side of the man in black clothes?
Your output should be:
Step 1: yes
Step 2: ["the man in black", "woman", "lower left corner"]
Step 3: ["the man in black", "woman"]
# Example 3:
Question: In which country is the comedy featured in the video recognized
worldwide?
Your output should be:
Step 1: no
```

Figure 4. CoT-driven decouple prompt for object list.

#### CoT-Driven Decouple Prompt for Video Event

```
As a video keyframe selection expert, your task is to design a question for
filtering keyframes. The question should help determine if a frame contains key
information needed to answer the original question.
# Example 1:
Original question: "How does the person open the bottle?"
Analysis:
1. This is a how-type question about an action process
2. Key elements: person, bottle, opening action
3. Need to capture crucial moments of the opening process
[Filtering Question]: "Does this frame show a key moment in the bottle opening
process?"
# Example 2:
Original question: "What color is the car that passes by?"
Analysis:
1. This is a what-type question about object property
2. Key elements: car, passing action, visibility of color
3. Need clear view of the car
[Filtering Question]: "Is the passing car clearly visible in this frame?"
Now, please think through these steps for the given question: {query}
1. Analyze the type of the original question (what/who/how/where/when/why)
2. Identify the key elements to focus on (objects/actions/states/scenes)
3. Design a simple, direct question to determine if a frame contains these key
elements
Output format:
[Filtering Question]: "your designed question"
```

Figure 5. CoT-driven decouple prompt for video event.

While the strategy based on video events underperforms the entity-based one in most categories, it achieves marginal improvements in Counting Problems (50.7% vs. 48.1%)

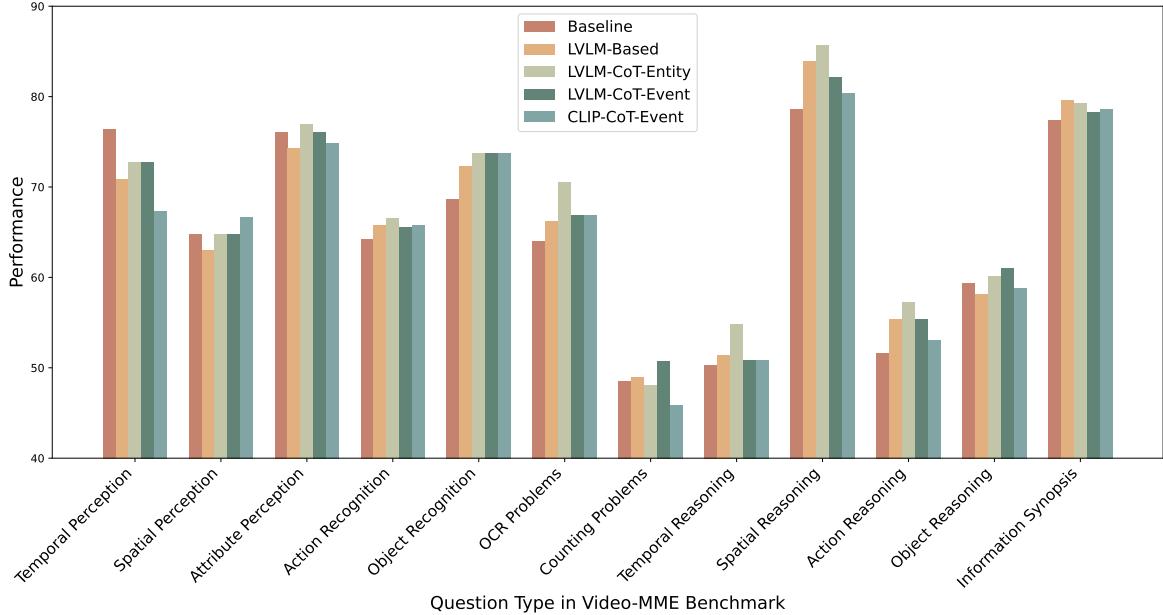


Figure 6. Sub-task results shown in Video-MME [7] benchmark when applying distinct frame scoring strategy of LLaVA-Video-7B [48].

| Scoring Strategy | Count       | Order       | Retrieve    | Overall     |
|------------------|-------------|-------------|-------------|-------------|
| None             | 24.0        | 62.7        | 76.4        | 54.4        |
| LVLM-Based       | 18.9        | 74.7        | 54.7        | 49.5        |
| LVLM-CoT-Entity  | <b>29.7</b> | <b>80.4</b> | 84.0        | <b>64.7</b> |
| LVLM-CoT-Event   | 27.6        | 79.8        | <b>85.8</b> | 64.4        |
| CLIP-CoT-Entity  | 26.7        | 76.0        | 84.7        | 62.4        |

Table 7. Results on different frame scoring strategies when extending QuoTA with LLaVA-Video-7B [48] on VN Bench [49].

and Object Reasoning (61.0% vs. 60.1%). This indicates that event-centric decomposition (e.g., “opening a door → placing an object”) better captures transient interactions critical for counting sequential actions. However, its inferior performance in Temporal Reasoning (50.8%) reveals limitations in modeling long-range dependencies, likely due to fragmented event segmentation that disrupts holistic temporal logic.

**Semantic-Visual Misalignment in CLIP-Based Scoring.** Frame scoring based on CLIP [23] exhibits paradoxical behavior: it achieves competitive performance in low-semantic tasks like Spatial Perception (66.7%) but severely degrades in high-level reasoning tasks (Temporal Reasoning: 50.8%). This dichotomy stems from CLIP’s inherent bias toward visually salient frames (e.g., motion-blurred or overexposed frames) rather than semantically scenes. For instance, in Attribute Perception, CLIP’s focus on anomalous frames reduces its ability to aggregate consistent attribute features across time, resulting in a 2.7% drop com-

pared to the entity-based strategy.

Table 7 also shows distinct frame scoring strategies’ performance under counting, ordering, and retrieving sub-tasks on VN Bench [49], which also demonstrated the superior performance of query-oriented frame scoring strategy with CoT-driven query decoupling.