

MERTA: Multi-Expert Retrieval for Token Allocation in Video Large Language Models

Bo Gao¹, Donghui Feng¹, Qi Wang², Qunshan Gu², Zhengxue Cheng¹, Li Song¹

¹Shanghai Jiao Tong University, ²Ant Group

Correspondence: zxcheng@sjtu.edu.cn, song_li@sjtu.edu.cn

Abstract

Video Large Language Models (VideoLLMs) excel in video understanding but suffer from significant computational overheads due to extensive visual tokens. Existing mainstream solutions including visual token compression and query-based Retrieval-Augmented Generation (Video RAG), face notable limitations: the former typically pursues task-agnostic compression of global video semantics, lacking query awareness, while the latter fails to optimize visual token construction and relies on costly, rigid annotation data. To address these challenges, we propose **MERTA** (Multi-Expert Retrieval for Token Allocation), a plug-and-play framework that integrates selective multi-expert collaborative retrieval with query-aware token allocation. MERTA selectively leverages lightweight experts to capture multimodal semantics, adaptively identifying key frames by measuring the relevance among queries, visual content, and auxiliary annotations. It then dynamically allocates the visual token budget, assigning higher budgets to key frames while keeping the total visual token budget constant. MERTA integrates seamlessly without requiring architectural modifications, demonstrating superior efficiency and performance over existing methods across multiple video understanding benchmarks.

1 Introduction

Recently, video large language models (VideoLLMs) have demonstrated remarkable performance in video understanding and reasoning tasks (Zhang et al., 2023; Shu et al., 2025; Wang et al., 2024, 2025). However, since a video is inherently composed of multiple consecutive frames, the number of visual tokens it contains is substantially larger than that of a single image. Consequently, how to achieve a more efficient representation of the numerous video frames—so as to enhance the model’s capability to solve video understanding

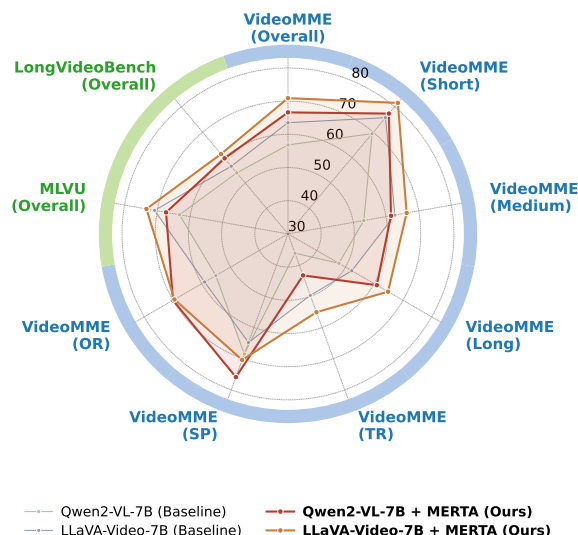


Figure 1: Comparison of the performance of MERTA with LLaVA-Video-7B (Zhang et al., 2024b) and the Qwen2-VL-7B (Wang et al., 2024) baseline across multiple benchmarks, including VideoMME (Overall, Short, Medium, Long, Temporal Reasoning, Spatial Perception, Object Reasoning), MLVU (Overall), and LongVideoBench (Overall).

tasks during inference—has emerged as a critical challenge for the field (Liu et al., 2025b; Song et al., 2024).

To address this issue, existing research primarily focuses on two types of approaches: visual token compression methods (Cao et al., 2025a) and query-based retrieval-augmented generation (RAG) video understanding methods (Zhang et al., 2025).

Visual token compression methods aim to mitigate redundancy in visual representations while preserving the complete semantics of the video as much as possible (Shao et al.). Early approaches exploited spatial-temporal locality through neighborhood pooling (Luo et al., 2023; Maaz et al., 2024; Wang et al., 2024), while more recent works such as VidCom² (Liu et al., 2025a) leverage global semantic similarity to prioritize retaining semanti-

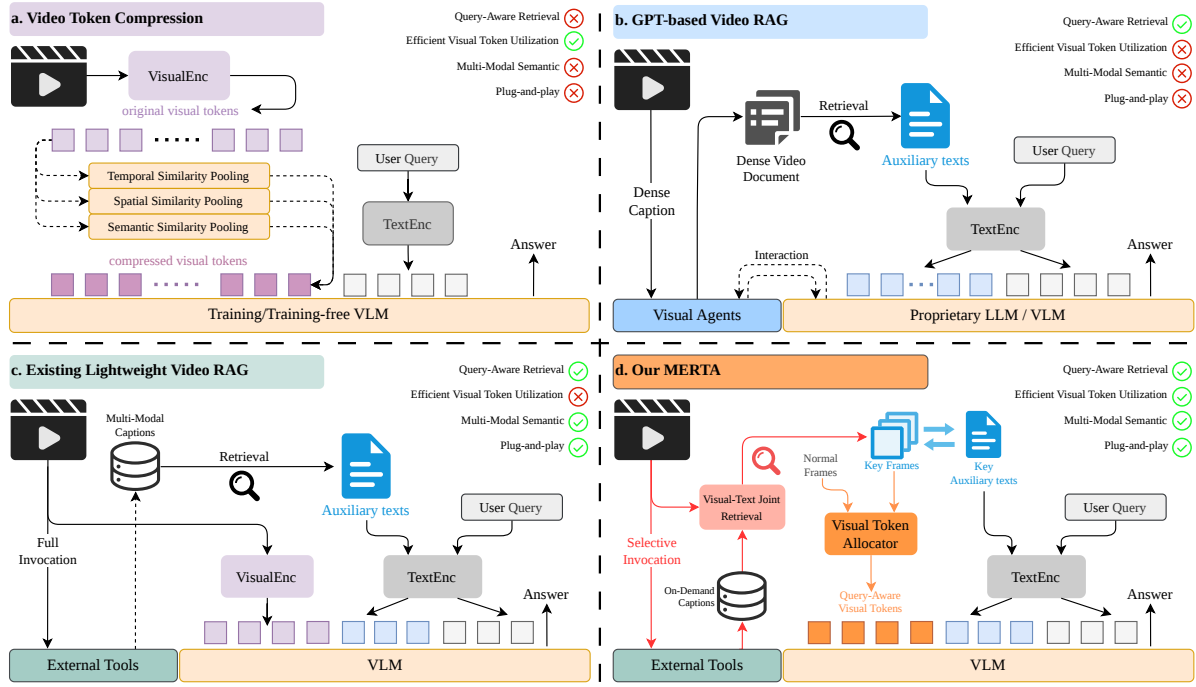


Figure 2: The figure illustrates the mainstream approaches in current video understanding research, including video token compression methods, GPT-based approaches, lightweight Video-RAG methods, and our proposed MERTA framework. MERTA provides a training-free pipeline that is compatible with arbitrary Large Vision-Language Models (LVLMs). By leveraging multimodal fusion-based joint retrieval and adaptive visual token allocation, MERTA achieves superior performance compared with existing mainstream approaches.

cally unique tokens. However, these methods are inherently **task-agnostic**; they mechanically preserve global semantics without adapting to specific queries, which often limits their performance on task-specific objectives. Furthermore, they operate solely at the visual level, ignoring the rich multi-modal context of videos.

Retrieval-Augmented Generation (RAG) has been adapted for video understanding to enhance reasoning capabilities (Zheng et al., 2025; Luo et al., 2026; Ren et al., 2026). Early "GPT-based agent" methods convert videos solely into massive text, resulting in high latency and loss of visual context (Fan et al., 2024). Recent lightweight variants (e.g., VideoRAG Luo et al., 2026, LVAgent Chen et al., 2025) employ expert models for annotation and retrieval. Despite improvements, they rely on complex multi-step pipelines and often indiscriminately activate all experts for all tasks, introducing unnecessary computational overhead. Critically, existing RAG methods focus solely on retrieving textual auxiliary information while neglecting the **structural optimization of visual token construction**, failing to fully leverage the visual understanding capabilities of Vision-Language Models (VLMs).

To bridge these gaps, we propose **MERTA** (Multi-Expert Retrieval for Token Allocation), a plug-and-play framework that integrates selective multi-expert retrieval with query-aware dynamic token allocation. MERTA selectively invokes lightweight experts to generate auxiliary multimodal annotations. It then fuses relevance scores among the query, auxiliary text, and visual frames to identify key frames. Subsequently, MERTA dynamically allocates the visual token budget: assigning higher budgets to query-relevant key frames while compressing non-key frames. This mechanism mimics human cognitive processes—skipping irrelevant content while focusing on critical segments—thereby optimizing both efficiency and accuracy.

Our main contributions are summarized as follows:

- **Query-aware Token Allocation.** We propose a novel mechanism that selectively leverages multi-expert retrieval to identify query-relevant key frames and dynamically allocates visual token budgets under a fixed total visual token budget, constructing task-customized video representations.

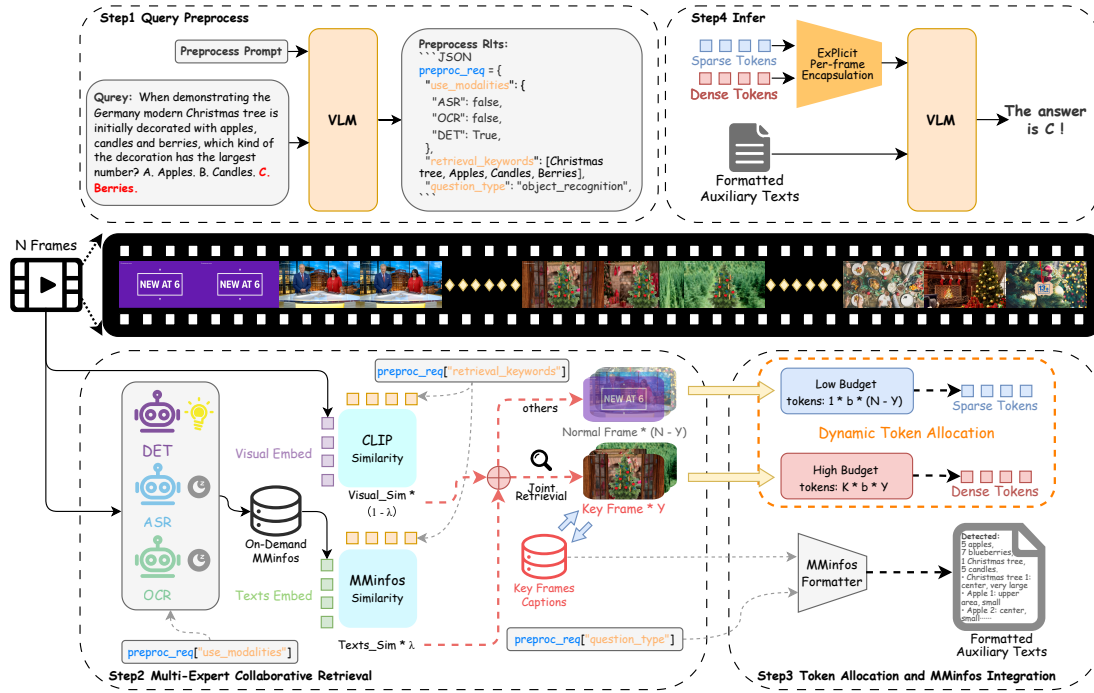


Figure 3: Overview of the proposed MERTA framework. The pipeline first performs query preprocessing, followed by the optional invocation of multiple experts to generate multimodal information (MMinfos). Based on the extracted MMinfos and visual content, key frames are jointly retrieved. Subsequently, dynamic token allocation and integration are applied to the video frames. Finally, during the inference stage, the keyframe auxiliary text, the query, and the token-allocated visual token sequence are jointly fed into the model to generate the final response.

- **Plug-and-Play Framework.** MERTA is a training-free solution that seamlessly integrates into arbitrary Large Vision-Language Models (LVLMs) with minimal computational overhead, consistently improving performance across various video understanding benchmarks.
- **Bridging Retrieval and Compression for SOTA Performance.** MERTA synergizes the semantic retrieval capabilities of RAG with the structural efficiency of visual token compression. By combining query-specific content retrieval with query-aware visual token construction, it outperforms existing visual token compression and RAG-based video understanding methods, establishing a new state-of-the-art (SOTA) in efficient video understanding.

2 Related Work

2.1 Video Large Language Models

Large Vision-Language Models (LVLMs) integrate vision encoders with Large Language Models (LLMs) to achieve robust visual understanding

(Wang et al., 2024; Li et al., 2024). Specialized VideoLLMs have emerged to handle video data, leveraging large-scale datasets for tasks such as multimodal reasoning (e.g., LLaVA-OneVision (Li et al., 2024)) and video instruction following (e.g., LLaVA-Video (Zhang et al., 2024b)). Despite these advancements, the extensive sequence of visual tokens generated from continuous video frames remains a primary bottleneck for efficient deployment.

2.2 Token Compression for LVLMs

Visual token compression aims to mitigate redundancy in visual representations while preserving the complete semantics of the video as much as possible (Shao et al.). Early approaches exploit spatial-temporal locality by pooling over adjacent frames and neighboring spatial regions (Luo et al., 2023; Maaz et al., 2024; Wang et al., 2024). More recent works, such as VidCom² (Liu et al., 2025a), focus on global semantic similarity, retaining unique tokens while discarding redundant ones. However, these methods are largely task-agnostic; they optimize for full semantic preservation rather than relevance to specific queries, which limits their per-

formance on task-specific objectives. Furthermore, they operate solely at the visual level, failing to leverage the rich multimodal context inherent in video data.

2.3 Video RAG for LVLMs

Retrieval-Augmented Generation (RAG) enhances LLMs by retrieving query-relevant content (Zheng et al., 2025) and has been adapted to the domain of video understanding (Luo et al., 2026; Ren et al., 2026; Arefeen et al., 2024; Ma et al., 2025). Early "GPT-based agent" methods rely on closed-source VLMs/LLMs to convert videos into massive amounts of text for retrieval and reasoning, resulting in loss of visual context and high computational costs (Fan et al., 2024). Recent lightweight approaches, such as VideoRAG (Luo et al., 2026) and LVAgent (Chen et al., 2025), employ multiple expert models to annotate videos and retrieve query-relevant segments. Although more efficient than agent-based methods, they typically involve complex and rigid multi-step pipelines. For example, VideoRAG indiscriminately activates all experts for every query, introducing unnecessary overhead. Most critically, existing video RAG methods focus on retrieving textual annotations while neglecting the structural optimization of visual token construction, rendering them unable to fully exploit the visual encoding capabilities of VLMs—a gap that our work aims to bridge.

3 Method

We propose **MERTA**, a plug-and-play, training-free framework that enhances VideoLLM efficiency through query-aware keyframe selection and dynamic token allocation. As illustrated in Figure 3, our pipeline consists of four stages: (i) **Query Preprocessing**, where the VLM analyzes the input query to generate control commands; (ii) **Multi-Expert Collaborative Retrieval**, which selectively invokes lightweight experts to identify query-relevant keyframes; (iii) **Token Allocation & Auxiliary Integration**, which allocates a token budget to each frame based on relevance and formats the relevant multimodal annotations as auxiliary text; and (iv) **Inference**, which integrates the visual tokens with frame indices and feeds them together with the auxiliary text into the VLM to generate the final result. The following subsections detail each component.

3.1 Query Preprocessing

To minimize system complexity, we leverage the inherent natural language processing capabilities of the backbone VLM to preprocess the input query, avoiding the need for external LLMs. This step processes only the short query text input and generates a set of downstream commands, `downstream_cmds`, to guide subsequent modules. These commands include:

- **use_modalities**: A binary flag controlling the activation of specific expert modules (e.g., ASR, OCR, DET). True enables the module, while False bypasses it, ensuring computational resources are allocated only to relevant modalities.
- **retrieval_keywords**: Inspired by VideoRAG (Luo et al., 2026), we extract concise keywords from the query to facilitate precise vector similarity retrieval.
- **question_type**: To optimize the utility of generated auxiliary text, the VLM categorizes the query into predefined types (e.g., *summary*, *object recognition*, *counting*, *ocr*, *temporal/spatial reasoning*). This classification guides the formatting of expert outputs into task-specific auxiliary contexts that better assist the VLM in reasoning.

3.2 Optional Multi-Expert Collaborative Retrieval

To balance efficiency and compatibility, we uniformly sample N frames from the input video V as candidates. Unlike traditional methods that exhaustively annotate all modalities (Luo et al., 2026; Chen et al., 2025), MERTA selectively activates expert models based on the `use_modalities` flags from query preprocessing. This adaptive mechanism improves annotation efficiency and reduces computational overhead by skipping unnecessary processing.

Specifically, for each frame f_i of V , we invoke lightweight experts only for activated modalities M_u : Whisper (Radford et al., 2023) for ASR, EasyOCR (JaidedAI, 2023) for OCR, and DETR-ResNet-50 (Carion et al., 2020) for object detection. The resulting multimodal annotations form the frame-specific text database T_i :

$$T_i = \bigcup_{m \in M_u} \text{Expert}_m(f_i) \quad (1)$$

where $\text{Expert}_m(\cdot)$ denotes the textual output for modality m . To ensure compatibility with text-based retrieval, OCR and detection outputs are formatted into natural language descriptions (e.g., “An image containing the following on-screen text: (*ocr_texts*)” and “An image detected with the following objects: (*detected_objects*)”).

Annotation Lifecycle and Lazy Caching. The annotations T_i produced by Equation 1 are query-independent: they depend only on the video and the modality. Each modality of a video is therefore annotated exactly once, at indexing time, and written to a **persistent per-video, per-modality cache**; every subsequent query reads from this cache and never re-invokes the expert, exactly as in the offline-indexing cost model of conventional Video RAG (Luo et al., 2026). The `use_modalities` flags operate *on top of* this cache and are lazy: within a query they only skip the experts that are not needed and never trigger re-annotation, so selective invocation can only reduce, never increase, the total annotation work. On Video-MME, where each video carries three questions, the union of the activated modalities costs 36.3 s per video (12.1 s per query), compared with 56.7 s per video (18.9 s per query) for exhaustive full-modality annotation. Appendix C reports the per-expert breakdown, the measured activation rates, and the end-to-end comparison against the Video RAG state of the art.

Subsequently, we compute two types of similarity scores for each frame f_i with respect to the retrieval keyword set Q (retrieval_keywords):

Visual Similarity. We employ CLIP (Radford et al., 2021) to measure visual relevance. To enhance robustness against prompt variations, we average cosine similarities across multiple prompt templates $\mathcal{T}$, details of the prompt templates are provided in the Appendix D.1. The template-averaged similarity for a keyword kw is:

$$\bar{S}_{\text{img}}(f_i, kw) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} S_{\text{img}}(f_i, \tau_t(kw)) \quad (2)$$

where $\tau_t(\cdot)$ applies the t -th template. The overall visual similarity for frame f_i is then computed by averaging over all keywords in Q :

$$\text{Sim}_{\text{vis}}(f_i, Q) = \frac{1}{|Q|} \sum_{kw \in Q} \bar{S}_{\text{img}}(f_i, kw) \quad (3)$$

Text Similarity. We utilize BGE-M3 (Chen et al., 2024a) to compute semantic alignment between the

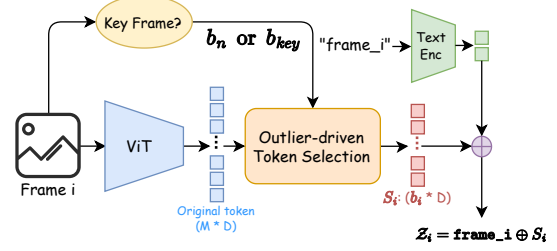


Figure 4: The figure illustrates the dynamic visual token allocation and integration process. First, a token budget is assigned to each frame based on whether it is identified as a key frame. Then, outlier-driven token selection is performed according to the allocated budget. Finally, to distinguish tokens from adjacent frames, the explicit textual identifier “frame_i” is tokenized and concatenated with the corresponding visual tokens of that frame.

keyword set Q and textual annotations T_i . For each activated modality $m \in M_u$, let $A_m \in T_i$ be the corresponding annotation. We first average the similarity across keywords for each modality:

$$\bar{S}_{\text{txt}}(A_m, Q) = \frac{1}{|Q|} \sum_{kw \in Q} s(A_m, kw) \quad (4)$$

where $s(\cdot, \cdot)$ denotes cosine similarity in the BGE-M3 embedding space. The final text similarity is derived by averaging across all activated modalities:

$$\text{Sim}_{\text{text}}(T_i, Q) = \frac{1}{|M_u|} \sum_{m \in M_u} \bar{S}_{\text{txt}}(A_m, Q) \quad (5)$$

Frame Selection. Let $S_v = \text{Sim}_{\text{vis}}(f_i, Q)$ and $S_t = \text{Sim}_{\text{text}}(T_i, Q)$. The comprehensive relevance score is computed as:

$$\text{Fsim}_i = \lambda S_v + (1 - \lambda) S_t \quad (6)$$

where $\lambda \in [0, 1]$ balances modal contributions. Key frames $Y = \{f_i \mid \text{Fsim}_i > \theta\}$ are selected using a threshold θ , with the constraint $|Y| \leq N/4$ to ensure efficiency.

3.3 Token Allocation and Auxiliary Text Integration

Given a video $V = \{F_1, \dots, F_N\}$, we extract visual tokens using the frozen VLM encoder, yielding $X_i \in R^{M \times D}$ for each frame. Under a fixed global budget B , we dynamically allocate tokens based on query relevance to preserve semantically critical details.

Table 1: Comparison of MERTA with existing Video RAG methods on Video-MME (Fu et al., 2025)(w/o Subtitles), MLVU (Zhou et al., 2025), and LongVideoBench (Wu et al., 2024). We compare against proprietary (DrVideo Ma et al., 2025, E-VRAG Xu et al., 2025) and open-source baselines (VideoRAG Luo et al., 2026, TV-RAG Cao et al., 2025b). MERTA uses the same backbone, frames, and visual token budgets as baselines. **#TextToken** is auxiliary text token size. MERTA achieves superior performance with fewer tokens.

Method	Param	Frame	#VisualToken	#TextToken	Video-MME				MLVU	LVB
					All	Short	Med	Long		
Proprietary Methods										
DrVideo (GPT-4)	-	0.2FPS	-	-	-	-	-	51.7	-	-
E-VRAG	7B	64	-	-	65.4	-	-	-	70.2	63.1
Open-Source Baselines										
Qwen2-VL + VideoRAG	7B	64	16k	2.0k	65.5	74.4	62.3	59.8	65.1	56.7
Qwen2-VL + TV-RAG	7B	64	16k	2.1k	65.7	74.2	62.9	60.0	65.7	55.4
LLaVA-Video + VideoRAG	7B	64	13k	2.0k	69.0	81.2	64.6	61.2	72.4	58.7
LLaVA-Video + TV-RAG	7B	64	13k	2.1k	69.3	81.2	65.0	61.7	72.6	58.8
Our Methods										
Qwen2-VL + MERTA (ours)	7B	64	16k	1.5k	66.6	77.3	61.5	60.9	67.3	59.7
LLaVA-Video + MERTA (ours)	7B	64	13k	1.6k	70.9	81.5	66.3	64.8	73.3	61.4
Qwen2.5-VL + MERTA (ours)	7B	64	16k	1.5k	67.3	77.9	66.4	57.7	68.1	61.3

Visual Token Budget Allocation. We assign distinct budgets to key and non-key frames. Let b_n be the budget for normal frames and $b_{key} = K \cdot b_n$ for key frames ($K > 1$). The total budget constraint is:

$$(N - |Y|) \cdot b_n + |Y| \cdot (K \cdot b_n) = B, \quad (7)$$

yielding:

$$b_n = \frac{B}{(N - |Y|) + K|Y|}, \quad b_{key} = K \cdot b_n. \quad (8)$$

Each frame F_i is assigned $b_i \in \{b_n, b_{key}\}$ based on its membership in the key frame set Y .

Outlier-driven Token Selection. Inspired by VidCom² (Liu et al., 2025a), we employ an outlier-based strategy to retain semantically distinctive tokens. For each frame $X_i = \{t_{i,1}, \dots, t_{i,M}\}$, we compute the semantic center $c_i = \frac{1}{M} \sum_{j=1}^M t_{i,j}$ and the cosine similarity $s_{i,j} = \cos(t_{i,j}, c_i)$. The outlier score is defined as $u_{i,j} = 1 - s_{i,j}$. We select the top- b_i tokens with the highest scores:

$$S_i = \text{Top-}b_i(u_{i,j}), \quad (9)$$

where S_i constitutes the compressed token set for F_i .

Auxiliary Text Integration. We integrate auxiliary textual information from lightweight experts (ASR, OCR, DET) via a query-type-aware reformulation mechanism. The query is classified into question_type, guiding the transformation of raw

annotations $\{T_i^{raw}\}$ into a unified natural language sequence:

$$\tilde{T} = \phi(\{T_i^{raw}\}_{i=1}^N, q_{type}), \quad (10)$$

where $\phi(\cdot)$ formats structured data into context-aligned text. Detailed formatting templates for each question_type are provided in Table 8 of Appendix D.2.

3.4 Inference

Frame-wise Token Packaging. To ensure that tokens from different frames are properly separated, we organize the compressed tokens in chronological order and explicitly insert clear frame boundary markers (embedded as text), see Figure 4 for details. The representation for the i -th frame is:

$$\mathcal{Z}_i = \text{frame_i} \oplus S_i, \quad (11)$$

where frame_i indicates the temporal index. The complete visual sequence is $\mathcal{Z}_{vis} = \mathcal{Z}_1 \oplus \dots \oplus \mathcal{Z}_N$.

Multimodal Input Construction. The final input $\mathcal{I}$ concatenates the structured visual sequence, auxiliary text, and user query Q :

$$\mathcal{I} = \mathcal{Z}_{vis} \oplus \tilde{T} \oplus Q. \quad (12)$$

This sequence is fed into the VLM for autoregressive response generation $\hat{y} = \text{VLM}(\mathcal{I})$. The entire framework is training-free and plug-and-play.

Table 2: Performance comparison of video token compression methods based on LLaVA-Video-7B (Zhang et al., 2024b) backbone. We evaluate on Video-MME (Fu et al., 2025) (w/o Subs), MLVU (Zhou et al., 2025), and LongVideoBench (Wu et al., 2024) benchmarks. “Average” denotes the mean performance retention across all benchmarks. Note that DyCoke (Tao et al., 2025) requires pruning similar tokens from consecutive 4 frames, thus its retention ratio R cannot be lower than 25%.

Method	Video-MME (w/o Subs)				MLVU	LVB	Average (%)
	Overall	Short	Medium	Long			
<i>Upper Bound</i>							
LLaVA-Video-7B	63.5	75.7	62.7	52.2	70.8	56.6	100.0
<i>Retention Ratio = 25%</i>							
DyCoke [CVPR’25]	56.1	65.8	53.6	48.9	56.9	53.0	87.5
FastV [ECCV’24]	59.3	67.1	60.0	50.8	57.8	51.2	88.5
SparseVLM [ICML’25]	60.1	71.1	59.1	50.1	58.9	54.2	91.2
VidCom ² [EMNLP’25]	61.7	73.0	61.7	50.0	59.0	55.5	92.9
MERTA (ours)	62.8	74.8	62.0	51.6	62.1	55.9	95.1
<i>Retention Ratio = 15%</i>							
FastV [ECCV’24]	51.3	56.4	51.1	46.2	53.8	44.6	78.5
SparseVLM [ICML’25]	55.7	65.0	53.9	48.3	56.2	52.7	86.7
VidCom ² [EMNLP’25]	58.3	68.0	57.3	49.7	56.8	51.5	87.7
MERTA (ours)	61.0	74.4	58.3	50.1	58.0	53.6	90.9

Table 3: Ablation study of MERTA on Video-MME (Fu et al., 2025) (Qwen2.5-VL-7B Bai et al., 2025, 16K visual tokens). TA: Token Allocation; AT: auxiliary texts.

Method	#Text	short	medium	long	avg
anchor	0	70.4	56.2	49.7	58.8
+TA	0	73.0	58.8	52.4	61.4
+TA+AT	1.5K	77.9	66.4	57.7	67.3

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate MERTA on three mainstream video understanding benchmarks: LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2025), and Video-MME (Fu et al., 2025). Detailed dataset statistics are provided in Appendix A.1.

Implementation Details. MERTA is validated on several representative Video LLMs, including LLaVA-Video (Zhang et al., 2024b), Qwen2-VL (Wang et al., 2024), and Qwen2.5-VL (Bai et al., 2025). All experiments are conducted on NVIDIA A100 (80GB) GPUs to ensure consistent hardware conditions. Detailed model information is in Appendix A.2.

Baselines. We compare MERTA against two categories of state-of-the-art methods. For RAG-

Table 4: Sensitivity analysis of the retrieval threshold θ on Video-MME (Fu et al., 2025). We fix $\lambda = 0.4$ and evaluate performance under different thresholds. **#Text-Token** and **Time** denote the total auxiliary text tokens and average inference time per question, respectively.

θ	#TextToken	Time	Overall
0.0	3.1K	15s	66.9
0.1	2.8K	14s	66.7
0.2	2.4K	11s	67.1
0.3	1.5K	8s	67.3
0.4	0.6K	6s	66.6
0.5	0.1K	5s	65.2
1.0	0.0K	5s	61.8

based methods, we include non-open-source works DrVideo (Ma et al., 2025) and E-VRAG (Xu et al., 2025), alongside open-source methods VideoRAG (Luo et al., 2026) and TV-RAG (Cao et al., 2025b). Note that DrVideo employs GPT-4 (Achiam et al., 2023) for reasoning, whereas others utilize 7B-scale models. Comparisons are performed under identical underlying models and total visual token budgets, more introduction can be seen in Appendix B.1. For token compression methods, baselines include FastV (Chen et al., 2024b), SparseVLM (Zhang et al., 2024a), DyCoke (Tao et al., 2025), and VidCom² (Liu et al., 2025a). Following SparseVLM (Zhang et al., 2024a), we adopt the “equivalent retention ratio” for fair comparison. For DyCoke (Tao et al., 2025), which compresses both

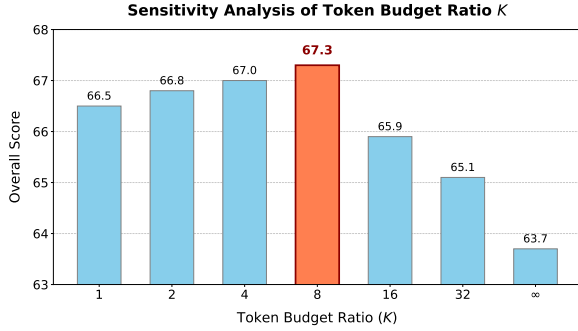


Figure 5: Performance of MERTA under different token allocation ratios between keyframes and normal frames. Here, K denotes the ratio of the token budget per keyframe to that per normal frame. The setting $K = \infty$ indicates that only keyframes are utilized (i.e., the token budget for normal frames is zero).

visual tokens and KV caches, we isolate and evaluate only its visual token compression component (see Appendix B.2).

4.2 Main Results

Comparison with Video-based RAG Methods.

To ensure a fair comparison, we evaluate MERTA under identical settings to mainstream RAG-based methods: same base models, input frame counts, and total visual token budgets. Specifically, we process 64 frames using Qwen2-VL-7B (Wang et al., 2024) and LLaVA-Video-7B (Zhang et al., 2024b), corresponding to visual token budgets of 64×256 and 64×210 , respectively. Unlike existing methods that distribute tokens uniformly, MERTA employs a dynamic allocation strategy.

Results show that MERTA outperforms open-source baselines VideoRAG (Luo et al., 2026) and TV-RAG (Cao et al., 2025b), while introducing less auxiliary text information. This superiority is attributed to two core designs: (1) an “on-demand retrieval” mechanism that minimizes annotation redundancy, and (2) a query-aware dynamic token allocation strategy that prioritizes visual regions that are more relevant to the query rather than distributing tokens uniformly, thereby improving efficiency.

MERTA is also cheaper to run: under an identical setup, selective invocation lowers the amortised annotation cost from 20.3 s to 12.1 s per query and added latency from +5.2 s to +3.4 s relative to VideoRAG (Luo et al., 2026), at higher accuracy. Full end-to-end measurements are reported in Appendix C.

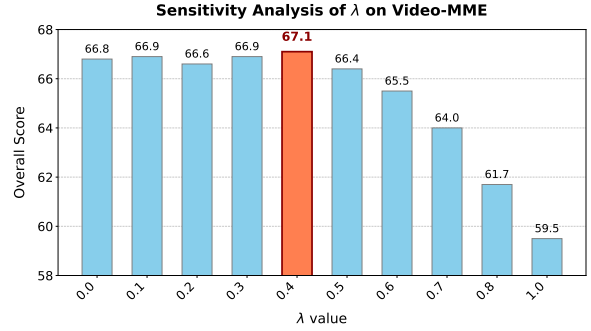


Figure 6: Sensitivity analysis of λ on Video-MME (Fu et al., 2025). We fix the retrieval threshold θ at 0.2 and evaluate MERTA’s performance under different λ values.

Comparison with Token Compression Methods.

MERTA consistently surpasses state-of-the-art token compression methods across various benchmarks. On LLaVA-Video (Zhang et al., 2024b), MERTA outperforms the SOTA method VidCom² (Liu et al., 2025a) by 2.2% and 3.2% at compression ratios of $R = 25\%$ and 15% , respectively. These results validate that dynamically allocating limited token budgets to query-relevant visual regions is more effective than rigidly preserving global semantic content.

4.3 Ablation Studies

Impact of MERTA Components. We perform stepwise ablation studies on Qwen2.5-VL-7B (Bai et al., 2025) (token budget: 16,384) to isolate the contributions of the Token Allocation (TA) module and auxiliary text. Starting from the base model baseline (58.8%), incrementally adding TA and retrieved keyframe auxiliary text yields significant performance gains, reaching 61.4% and 67.3%, respectively. This confirms the effectiveness of both components.

Impact of Keyframe Amplification Factor K .

To achieve dynamic allocation, we assign a per-frame token quota to keyframes that is K times that of normal frames, while keeping the total budget constant. We observe a trade-off in selecting K : small K values diminish the benefit of dynamic allocation, while excessively large K values starve normal frames of tokens, compromising global semantic preservation. For instance, retaining only keyframes ($K \rightarrow \infty$) leads to a significant performance drop (67.3% $\rightarrow$ 63.7%). We evaluate various K values to find the optimal balance, with results detailed in Figure 5.

Impact of Similarity Weight λ and Retrieval Threshold θ . Keyframe selection relies on a comprehensive similarity score F_{sim_i} , defined as the weighted average of text similarity (via BGE-M3) and visual similarity (via CLIP), controlled by weight λ . A frame is selected if $F_{sim_i} > \theta$, where θ serves as the retrieval threshold. This threshold presents a trade-off: high values risk missing relevant keyframes, while low values introduce redundancy and increase reasoning complexity. To optimize these hyperparameters, we conduct a two-step ablation: first, fixing $\theta = 0.2$ to tune λ ; second, fixing the optimal $\lambda = 0.4$ to tune θ . Detailed results are presented in Figure 6 and Table 4.

5 Conclusion

This paper addresses the computational bottleneck caused by redundant visual tokens in Video Large Language Models (Video LLMs). We point out that existing token compression methods are limited by their task-agnostic nature, while current Retrieval-Augmented Generation (RAG) approaches fail to optimize visual token construction. To bridge this gap, we propose **MERTA**, a plug-and-play framework that synergistically integrates multi-expert retrieval with dynamic token allocation. By leveraging selectively invocable lightweight expert models for annotation and retrieval, and dynamically allocating visual token budgets, MERTA highlights relevant information at both the textual annotation and visual token levels. Extensive experiments demonstrate that, as a training-free solution, MERTA achieves state-of-the-art (SOTA) performance with high efficiency across multiple benchmarks. This work validates the effectiveness of combining semantic retrieval with structured visual token optimization, paving the way for scalable long-form video understanding.

6 Limitations

In this work, we propose a plug-and-play framework for efficient query-aware keyframe retrieval and dynamic token allocation, aimed at enhancing VideoLLM performance under fixed visual token budgets. Due to computational resource constraints, we have not yet evaluated the proposed method on larger-scale models, such as LLaVA-Video-72B and Qwen2-VL-72B. Nevertheless, given the simplicity of MERTA and its demonstrated advantages, we expect that the method will yield similar—if not greater—benefits when ap-

plied to larger architectures. This expectation is grounded in the observation that efficient token management becomes increasingly critical as model complexity grows. Future work will involve comprehensive experimental evaluations across different model scales to further validate and explore MERTA’s potential in large-scale scenarios. In addition, the current expert set covers only three modalities—ASR, OCR and object detection—which were chosen because they capture the dominant non-visual cues in the benchmarks studied here; incorporating experts along further dimensions, such as scene-graph generation, pose or depth estimation, and temporal action detection, is a natural extension that we leave to future work. Additionally, we plan to extend MERTA to real-time streaming video understanding settings and more specialized tasks, such as Video Anomaly Detection (VAD), thereby broadening its practical applicability.

Acknowledgments

This work was partly supported by the NSFC(62431015, 62571317, 62501387), the Fundamental Research Funds for the Central Universities, Shanghai Key Laboratory of Digital Media Processing and Transmission under Grant 22DZ2229005, 111 project BP0719010 and Ant Group Research Fund.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024. irag: Advancing rag for videos with an incremental approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4341–4348.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Weidong Cao, Zeqin Feng, Fangwei Sun, Danjun Liu, Yongqiang Xie, and Zhongbo Li. 2025a. A survey on visual token compression for efficient vision-language models. In *2025 5th International Conference on Advanced Algorithms and Neural Networks (AANN)*, pages 735–741. IEEE.

- Zongsheng Cao, Yangfan He, Anran Liu, Jun Xie, Feng Chen, and Zhepeng Wang. 2025b. Tv-rag: A temporal-aware and semantic entropy-weighted framework for long video retrieval and understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9071–9079.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.
- Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. 2025. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20237–20246.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 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*, pages 19–35. Springer.
- Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.
- Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer.
- Chaoyou Fu, Yuhao Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118.
- JaidedAI. 2023. [Easyocr: A ready-to-use ocr with 80+ languages](#). GitHub repository.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. 2025a. Video compression commander: Plug-and-play inference acceleration for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1910–1924.
- Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, and 1 others. 2025b. Nvila: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4122–4134.
- Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.
- Yongdong Luo, Xiwu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and 1 others. 2026. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033.
- Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezaatoughi, and Jianfei Cai. 2025. Drvideo: Document retrieval based long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18936–18946.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2026. Videorag: Retrieval-augmented generation with extreme long-context videos. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2390–2401.
- Kele Shao, TAO Keda, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin,

- Yang Sui, and Huan Wang. A survey of token compression for efficient multimodal large language models. *Transactions on Machine Learning Research*.
- Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26160–26169.
- Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.
- Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025. Dycok: Dynamic compression of tokens for fast video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18992–19001.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, and 1 others. 2025. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.
- Zeyu Xu, Junkang Zhang, Qiang Wang, and Yi Liu. 2025. E-vrag: Enhancing long video understanding with resource-efficient retrieval augmented generation. *arXiv preprint arXiv:2508.01546*.
- Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553.
- Rui Zhang, Chen Liu, Yixin Su, Ruixuan Li, Xuanjing Huang, Xuelong Li, and Philip S Yu. 2025. A comprehensive survey on multimodal rag: all combinations of modalities as input and output. *Authorea Preprints*.
- Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and 1 others. 2024a. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.
- Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. 2025. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv preprint arXiv:2503.18016*.
- Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, and 1 others. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13701.

A Benchmark and Model Details

A.1 Benchmark Details

We present a detailed overview of video understanding benchmarks, as described below:

- **Video-MME** (Fu et al., 2025) is a widely used benchmark designed to assess the capability of Large Vision-Language Models (LVLMs) in processing detail-rich videos in real-world scenarios. The dataset is divided into three subsets based on video duration, ranging from 11 seconds to 1 hour.
- **MLVU** (Zhou et al., 2025) focuses on long-video understanding, covering 9 distinct task types with high representativeness. It is constructed from diverse long-video materials, with durations ranging from 3 minutes to 2 hours and an average length of approximately 12 minutes.
- **LongVideoBench** (Wu et al., 2024) aims to evaluate the model’s ability to accurately retrieve and reason over fine-grained multimodal information from long videos. It contains 6,678 human-annotated multiple-choice questions across 17 categories.

A.2 Model Details

We describe the Video Large Language Models (VideoLLMs) evaluated in our study:

- **LLaVA-Video** (Zhang et al., 2024b): Built upon LLaVA-OneVision, this model employs a SigLIP visual encoder and Qwen2 LLM. It is fine-tuned on the large-scale synthetic dataset LLaVA-Video-178K, achieving robust performance in detailed captioning and diverse QA tasks across mainstream benchmarks.
- **Qwen2-VL** (Wang et al., 2024): This model introduces Naive Dynamic Resolution to adaptively convert frames of arbitrary resolutions into visual tokens. By employing Multimodal Rotary Position Embedding (M-RoPE), it effectively handles long-form videos (exceeding 20 minutes) within a unified image-video processing paradigm.
- **Qwen2.5-VL** (Bai et al., 2025): It achieves significant advancements in visual recognition, precise object localization (via bounding boxes or points), robust document pars-

ing (invoices, forms, tables, charts), and long-video comprehension. Key innovations include dynamic resolution processing and absolute time encoding, enabling hour-long video understanding with second-level event localization without traditional normalization techniques. A native dynamic-resolution Vision Transformer with Window Attention reduces computational overhead while preserving native resolution. Qwen2.5-VL also functions as an interactive visual agent capable of reasoning, tool usage, and real-world task execution (e.g., operating computers and mobile devices).

B Baseline Details

We provide detailed introductions of existing RAG-based video understanding methods and token compression methods mentioned in the main text, as follows:

B.1 RAG-based Methods

We provide both closed-source and open-source RAG-based video understanding methods mentioned in the paper.

- **DrVideo** (Ma et al., 2025): A document-retrieval-based system. It converts long video understanding into a long-document understanding task, leveraging LLMs effectively. DrVideo first creates a coarse text-based document from the video to retrieve key frames, then iteratively augments the document with missing information via an agent-based loop, finally making predictions using chain-of-thought reasoning.
- **E-VRAG** (Xu et al., 2025): An efficient retrieval-augmented generation framework. E-VRAG first applies hierarchical query decomposition to pre-filter irrelevant frames, then uses a lightweight VLM for frame scoring. A global statistical distribution strategy mitigates performance loss from the lightweight model. Finally, a multi-view QA scheme enhances information extraction from retrieved frames.
- **VideoRAG** (Luo et al., 2026): To address the limited context window of existing LVLMs in long video comprehension, the authors propose Video-RAG (Video Retrieval-Augmented Generation), a training-free and

cost-effective pipeline. It uses open-source tools to extract visually-aligned auxiliary texts (e.g., audio, OCR, object detection) and integrates them into any LVLM alongside video frames and queries in a plug-and-play manner.

- **TV-RAG** (Cao et al., 2025b): A training-free framework. It introduces two main mechanisms: (i) a time-decay retrieval module that incorporates temporal offsets into similarity computation, and (ii) an entropy-weighted key-frame sampler that selects information-dense frames. By combining temporal and semantic signals, TV-RAG enables dual-level reasoning that can be added to any LVLM without retraining.

B.2 Token Compression Methods

We provide token compression methods mentioned in the paper.

- **DyCoke** (Tao et al., 2025): A two-stage VideoLLM-specific method that first prunes similar tokens temporally, then compresses less-attended visual tokens in the KV cache using LLM attention weights. Due to its reliance on similarity-based compression across divided frame sets, it cannot achieve aggressive one-pass compression. The token compression stage is Flash Attention-compatible, but the KV cache compression requires explicit attention weights and thus remains incompatible with efficient attention operators.
- **FastV** (Chen et al., 2024b): An intra-LLM compression method that performs one-time token pruning based on attention weights of the output token from a selected LLM layer. Its explicit dependence on attention weights makes it incompatible with Flash Attention (Dao et al., 2022).
- **SparseVLM** (Zhang et al., 2024a): An intra-LLM compression method that ranks token importance using text-visual attention maps and prunes tokens via pre-selected text prompts to reduce attention noise. Like FastV, SparseVLM is incompatible with Flash Attention.
- **VidCom**² (Liu et al., 2025a): The authors systematically analyze existing token compression methods and identify two critical issues: information loss from overlooking cross-

frame visual signals, and implementation constraints causing incompatibility with modern architectures. They propose a plug-and-play inference acceleration framework that quantifies each frame’s uniqueness and adaptively adjusts compression intensity across frames.

C Efficiency Measurement Details

All measurements reported in this appendix were obtained with LLaVA-Video-7B (Zhang et al., 2024b), 64 uniformly sampled frames and a 13K total visual token budget, on a single NVIDIA A100 (80GB) GPU, over the full Video-MME (Fu et al., 2025) set (2,700 questions on 900 videos, i.e. 3 queries per video on average).

Per-Expert Annotation Cost and Activation Rates.

Table 5 reports the wall-clock annotation cost of each expert per video, together with the activation rate measured over the 2,700 questions. The three experts are invoked in parallel, so the annotation time of a video is the maximum over the modalities actually activated rather than their sum; ASR (Whisper) is consequently the dominant bottleneck, and skipping it is where selective invocation pays off most. Taking, for each video, the union of the modalities activated by its queries, selective invocation reduces annotation from 56.7 s to 36.3 s per video, i.e. from 18.9 s to 12.1 s per query at three queries per video.

Table 5: Per-expert annotation cost and measured activation rate over the 2,700 Video-MME questions. Experts are invoked in parallel, so the annotation wall-clock of a video is the maximum over the activated modalities. “Full-modality” annotates every modality for every video; “MERTA (selective)” takes, for each video, the union of the modalities that its queries actually activated.

Expert	Time / video	Activation rate
ASR (Whisper)	56.0 s	49.8%
OCR (EasyOCR)	4.7 s	42.2%
DET (DETR-R50)	4.2 s	90.1%
Full-modality	56.7 s	100%
MERTA (selective)	36.3 s	–

Amortisation over Queries. Because annotations are cached per video and per modality (Section 3.2), the annotation cost charged to a single query is $36.3/n$ s when a video receives n queries, as reported in Table 6. Video-MME corresponds to $n=3$, which is the setting used in the end-to-end

comparison below; deployments that serve more queries per video amortise this cost further, and the residual per-query overhead converges to the +3.4 s spent on query preprocessing, retrieval, token allocation and auxiliary-text injection.

Table 6: Amortised annotation cost per query as a function of the number of queries n served per video, given MERTA’s measured 36.3 s per-video annotation. $n=3$ corresponds to the Video-MME setting.

Queries per video n	1	3	10
Annotation (s / query)	36.3	12.1	3.6

Query-Time Overhead. Beyond annotation, MERTA adds +3.4 s per query for query preprocessing, similarity retrieval, token allocation and auxiliary-text injection, together with +6 GB of GPU memory. Against this cost, accuracy on Video-MME rises from 63.5 (the uncompressed LLaVA-Video-7B baseline) to 70.9.

End-to-End Comparison with the Video RAG State of the Art. Table 7 compares MERTA with the lightweight Video RAG state of the art (Luo et al., 2026) under an identical setup. MERTA is lower on all four cost axes—annotation (12.1 s vs. 20.3 s), GPU memory (6 GB vs. 8 GB), added inference latency (+3.4 s vs. +5.2 s) and auxiliary-text length (1.6k vs. 2.0k tokens)—while being more accurate (70.9 vs. 69.0). Three factors produce this: (i) selective invocation skips unneeded experts within a query, most importantly the ASR bottleneck, and simultaneously yields a cleaner retrieval corpus, lowering retrieval load and noise; (ii) fewer auxiliary-text tokens shorten the LVLm input sequence and its KV cache; and (iii) query-aware token allocation grants key frames a larger share of the fixed visual budget, raising accuracy at no additional token cost.

Relation to Reusable Offline Indexing. Since MERTA’s annotation is itself a form of offline indexing—query-independent, computed once per video and modality, and cached (Section 3.2)—its advantage does not depend on a particular number of queries per video. When a video receives few queries, annotation dominates the total cost, and MERTA’s selective annotation (12.1 s per query) is below the exhaustive annotation of Video RAG (20.3 s per query). When a video receives many queries, annotation amortises away and per-query

Table 7: End-to-end cost and accuracy comparison against the lightweight Video RAG state of the art (Luo et al., 2026) under an identical setup (LLaVA-Video-7B, 64 frames, 13K total visual tokens, A100-80GB, Video-MME with 2,700 questions over 900 videos, i.e. 3 queries per video). Annotation is a cached indexing-time cost and is reported amortised over the queries of each video.

Cost per query	VideoRAG (all experts)	MERTA (selective)
Annotation (amortised)	20.3 s	12.1 s
Extra GPU memory	8 GB	6 GB
Extra inference latency	+5.2 s	+3.4 s
Auxiliary-text tokens	2.0k	1.6k
Video-MME Overall	69.0	70.9

latency dominates, where MERTA’s +3.4 s is again below the +5.2 s of Video RAG.

D MERTA Pipeline Details

D.1 Prompt Templates Details

The prompt templates used in Equation 2 of the main paper are as follows:

1. “a photo of { }”
2. “a picture of { }”
3. “an image of { }”
4. “a scene containing { }”
5. “a frame showing { }”
6. “{ } in the image”

D.2 MMinfos Formatter Details

Please see the table in next page.

Table 8: Reference for the conversion formats of multimodal annotations corresponding to different `question_type` categories. The symbol “-” indicates that no special formatting is applied, and the raw output is used directly.

Question Type	Context Prompt	ASR Text (if activated)	OCR Text (if activated)	DET Text (if activated)
summary	This is a summarization task. Please synthesize all provided content to generate a comprehensive summary.	Detected speech transcription: “...”	Detected on-screen text: “...”	Detected objects in the video: “...”
object recognition	This is an object recognition task. Focus on the target objects, observing their morphology, color, and other attributes.	-	-	Frame i detects m objects of type $object_x$; $object_x_1$ is located in the center, small size...
counting	This is a counting task. Pay attention to the frequency of occurrence of the same object or event.	-	-	Frame i detects m objects of type $object_x$, n objects of type $object_y$...
ocr	This is an OCR task. Pay special attention to the detected on-screen text relevant to the query.	-	Frame i detects on-screen text: “...”	-
temporal reasoning	This is a temporal reasoning task. The video consists of N frames, located at timestamps 0.0s, 5.0s, 10.0s, etc.	Detected speech transcription: “...”	Frame i detects on-screen text: “...”	Frame i detects m objects of type $object_x$, n objects of type $object_y$...
spatial reasoning	This is a spatial reasoning task. Focus primarily on the positional information of the objects.	-	-	Frame i detects m objects of type $object_x$; $object_x_1$ is located in the center, small size...