QSVideo: Query-Conditioned Semantic Temporal Retrieval for Video Understanding

Wei Ao, Lan Wang, and Vishnu Naresh Boddeti

Michigan State University, East Lansing MI 48824, USA
{aowei,wanglan3,vishnu}@msu.edu

Abstract. The performance of vision-language models (VLMs) in video understanding declines with increasing video duration, as video moments unrelated to the query confuse their language components. Multimodal retrieval has emerged as a critical component of video understanding, addressing this challenge by localizing key visual evidence. However, existing multimodal retrieval methods suffer from biased relevance estimation, limited diversity, and temporal collapse. In this paper, we propose QSVideo, a unified framework that systematically addresses relevance, diversity, and temporal modeling in video retrieval. We first introduce a query-conditioned semantic ranker, QSRanker, which reformulates arbitrary questions into retrieval-friendly queries and estimates structured relevance along object, action, and location dimensions. Building upon this, we design QSRetrieval to jointly optimize relevance and diversity for more informative frame selection. Moreover, we propose temporal alignment strategies tailored for both long and streaming videos to improve evidence recall. Extensive experiments on long and streaming video benchmarks demonstrate that QSVideo greatly enhances video VLM performance under strict frame limit constraints. The code is available at https://github.com/human-analysis/QSVideo.

Keywords: Video Understanding · Multimodal · MLLMs· VLMs

1 Introduction

Video understanding involves answering a question using the information from a video. Although vision-language models (VLMs) demonstrate strong video understanding capability [14, 27, 38, 39, 71], their performance often degrades as video duration increases [59]. Naively feeding more frames is not a practical solution to this problem, as it does not necessarily improve accuracy while significantly increasing the GPU memory footprint. Humans rarely consult the video uniformly. Instead, they form a goal-driven hypothesis from the question and actively search for supporting visual evidence in the video. They focus on semantically relevant moments, avoid repeatedly examining similar clips, and scan different time periods to gather complementary clues. This behavior naturally balances *relevance*, *diversity*, and *temporal coverage*, motivating a retrieval-centric design for video understanding.

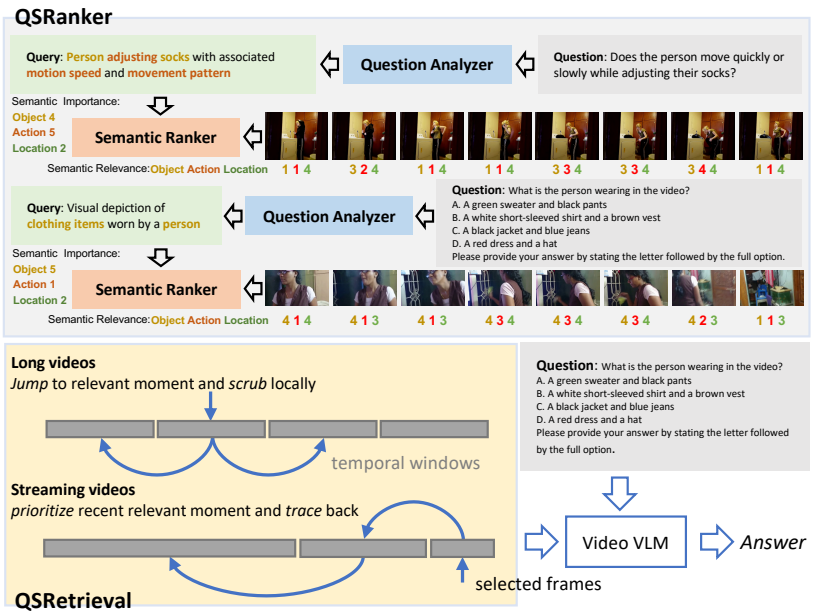

Fig. 1: `QSVideo`: query-conditioned semantic temporal retrieval for long and streaming videos. `QSVideo` includes `QSRanker`, `QSRetrieval` and a video VLM. `QSRanker` proposes a *Question Analyzer* (LLM) to understand arbitrary questions and transcribe original questions to retrieval-friendly queries. The question analyzer analyzes the questions and estimates the importance of Object, Action and Location. `QSRanker` designs a *Semantic Ranker* to score frames in terms of relevance to the query along Object, Action and Location dimensions. `QSRetrieval` is a novel search framework that balances relevance, diversity, and temporal coverage, inspired by the human video understanding process. We can seamlessly integrate `QSRanker`, `QSRetrieval` with different video VLMs to greatly improve their video understanding ability without scaling up models or expanding frames.

Despite recent progress, question-to-frame retrieval still suffers from three fundamental issues. First, **biased relevance**: directly using the original question as a query [6, 24, 25, 57, 73, 74] can yield unreliable relevance estimates, especially for complex formats such as multiple-choice questions. Second, **limited diversity**: ranking frames by relevance alone often results in near-duplicate evidence with low information gain [6, 25, 73]. Third, **temporal collapse**: existing methods frequently concentrate selections to a narrow temporal region, although key evidence may appear at distant timestamps in long videos.

To address these challenges, we propose `QSVideo`, a query-conditioned retrieval framework for video understanding. `QSVideo` consists of two key components – `QSRanker` and `QSRetrieval` as shown in Fig. 1. `QSVideo` discovers visual evidence by explicitly balancing *relevance*, *diversity*, and *temporal coverage* under a strict frame budget. Motivated by human video processing (question analysis

→ evidence discovery → understanding/reasoning), `QSRanker` understands and transcribes the questions to queries and evaluates query-to-frame semantic *relevance*. The distance between visual evidences are taken into account to improve the *diversity* of selections. `QSRetrieval` maximizes the *temporal coverage* for different videos: long video (locate globally relevant anchors and then search locally) and streaming videos (recency-first traversal).

We extensively benchmark `QSVideo` on challenging long [59] and streaming video [36] benchmarks. Under strict frame budgets (8/16/32 frames), `QSVideo` achieves state-of-the-art performance with an efficient 8B video VLM on StreamingBench [36] and outperforms the backbone VLM by 6.9-7.1 pp in accuracy on LVBench [59], and remains plug-and-play across multiple VLM backbones with consistent gains. `QSVideo` achieves this with only modest computational overhead and can be efficiently accelerated via batched inference, making it well-suited for real-world deployment. Our contributions are:

1. `QSVideo`, a query-conditioned retrieval framework that selectively attends to informative visual evidence and greatly boosts the performance of video VLMs;
2. `QSRanker` to understand and transcribe arbitrary questions into retrieval-friendly queries and estimate structured semantic relevance along Object, Action, and Location dimensions;
3. `QSRetrieval` to explicitly balance relevance, diversity, and temporal coverage, with tailored temporal alignment for both long and streaming videos.

2 Related Work

Multimodal Large Language Models (MLLMs) have bridged the gap between visual perception and linguistic reasoning by integrating pre-trained vision encoders with Large Language Models [39, 86]. Early influential models like LLaVA-1.5 [37] and MiniGPT-4 [86] utilized simple linear layers or MLPs to project image tokens, demonstrating remarkable zero-shot capabilities. Recent iterations, such as LLaVA-NeXT [30] and Insight-V [16], have further refined these architectures to support higher-resolution inputs and more complex reasoning tasks. Extending MLLMs from static images to dynamic video content introduces significant challenges, primarily regarding temporal coherence and the token explosion problem, where high frame counts exceed the LLM's context window. Early attempts such as VideoChat [32], Video-ChatGPT [44], and Video-LLaVA [35] primarily relied on sparse frame sampling to fit video data into existing image-based pipelines. To address the limitations of long-video comprehension, recent research has moved toward more sophisticated compression and integration strategies. Video-LLaMA [76] utilizes a Q-Former to bootstrap vision-language alignment, while LLaMA-Vid [34] introduces efficient visual compression to handle extended sequences. Scaling of multimodal pre-training models like InternVL2.5 [11], Qwen2.5-VL [3] also demonstrate superior reasoning across images, videos, and real-time streaming inputs.

Long-form and Streaming Video Understanding. Recent advancements in long-form and streaming video understanding have introduced diverse innovations to handle extensively long temporal sequences and real-time inputs efficiently [5, 7, 9, 15, 26, 28, 29, 36, 48, 50, 51, 54, 57, 62, 63, 69, 70, 75, 77]. Early efforts to overcome context window limitations utilized recurrent mechanisms, such as VideoLLaMB [62], which pioneered Recurrent Memory Bridges to propagate information across sequential segments. Streaming video understanding has emerged as a distinct task focusing on online videos, where processing continuously arriving frames allows for real-time interaction in applications like virtual reality and robotics. Unlike offline models, streaming VLMs must simultaneously memorize the past, perceive the current scene, and predict future states [26]. To optimize computational overhead, InfiniPot-V [29] enforces a length-independent memory cap through in-place KV cache compression, while ProVideLLM [7] achieves sub-linear scaling via interleaved summaries. Specialized memory structures have also been proposed, such as Flash-VStream [77], which maintains vision feature centers as semantic memory, and StreamForest [75], which adaptively organizes frames into a Persistent Event Memory Forest based on temporal distance and content similarity. For hour-scale reasoning, Video-XL [51] further condenses context into compact Visual Summarization Tokens.

Retrieval and Frame Selection. Effective information retrieval and selective frame or token sampling are critical for isolating task-relevant evidence from dense video content, ranging from query-agnostic selection to query-cognitive relevance ranking [2, 6, 8, 14, 22, 24, 25, 39–41, 43, 46, 47, 55, 64, 67, 68, 71, 72, 74, 78–81, 84]. While many models like LLaVA-Video [80] and SEAL [57] adopt uniform sampling to maximize temporal diversity, newer methods prioritize query-relevant frames. BOLT [41] utilizes inverse transform sampling and recursive temporal binning to identify informative frames without model retraining. Logic-in-Frames [22] reformulates this as a semantic-logical search across temporal relations. Alternatively, SeViLA [74] and FRAG [25] use VLMs as cross-modal rankers to score frame relevance based on specific queries. However, off-the-shelf VLMs often lack consistency in relevance scores as they are not specifically fine-tuned for cross-modal ranking tasks. To address this, FFS [6] and TimeSearch-R [46] learn optimized policies to dynamically select frames with self-verification. Furthermore, T* [72] recasts temporal search as a spatial grid task, and FlexSelect [81] identifies semantically critical visual tokens by analyzing internal attention patterns.

3 QSVideo: Relevance, Diversity and Temporal Coverage

We first describe our problem of interest, namely **query-to-frame retrieval for video understanding**. Then, we propose the query-conditioned semantic ranker (`QSRanker`) in § 3.1 and introduce automated frame-level video annotation and training in § 3.2. Finally, we present our solution to our problem of interest, `QSRetrieval`, in § 3.3. `QSRanker` followed by `QSRetrieval` can locate key visual clues to facilitate following video VLMs. `QSRanker` and `QSRetrieval`

are agnostic to video VLMs. In this paper, video VLMs enhanced by `QSRanker` and `QSRetrieval` are referred to as `QSVideo` (Fig. 1).

Problem statement. Given a query q and a series of video frames $\mathcal{V} = \{v_i\}_{i=0}^{T_V}$, the objective is to find a subset $A \subseteq \mathcal{V}$ such that (1) A provides the necessary visual information relevant to the query, (2) A has maximum visual diversity (*i.e.* the minimum visual redundancy), and (3) $|A| = K$. We formulate it as:

$$\begin{aligned} \max_{A \subseteq \mathcal{V}} \left\{ \sum_{v_i \in A} \text{Relevance}(q, v_i), \sum_{\substack{v_i, v_j \in A \\ i \neq j}} \text{Diversity}(v_i, v_j) \right\} \\ \text{s.t.} \quad |A| = K, \quad \sum_{i \in A} \text{T}(v_i) \geq \Gamma \end{aligned} \tag{1}$$

where we further introduce the temporal coverage measure $\sum_{i \in A} \text{T}(v_i)$ to mitigate *temporal collapse*, where frame selection concentrates on a narrow temporal region. The temporal coverage measure reduces near-duplicate evidence and increases effective information density. Broader coverage improves evidence recall under a limited frame budget, which is essential for long-range reasoning in long videos, where critical visual clues may be temporally dispersed.

3.1 QSRanker: Query-Conditioned Semantic Ranker

Given a query q and a series of frames uniformly sampled from a video $\mathcal{V} = \{v_i\}_{i=1}^{T_V}$, the goal is to *score* frames based on their relevance with respect to the query q. Frames with higher relevance scores provide more visual evidence required to answer the query. Existing works, *e.g.* SeViLA [74], FRAG [25], Frame Selector [24], Qwen3-VL-Reranker [33], prompt pretrained VLMs to generate "yes" or "no" for each query-frame pairs. This approach has three key drawbacks: (1) they do not analyze and leverage *hints* from the query, (2) a single binary score cannot represent the semantic relationship between a query and frames, and (3) they adopt original questions as queries, likely leading to biased relevance because of wrong options from multiple-choice questions. SEAL [57] introduces three different models to detect scenes, objects, and actions, and ignores the semantic understanding ability of VLMs.

We introduce **Q**uery-conditioned **S**emantic **Ranker**, dubbed `QSRanker`, that integrates a *question analyzer* and a *semantic ranker*. The question analyzer transcribes arbitrary questions to retrieval-friendly queries while exploring hints from questions. The semantic ranker scores query-frame pairs by their Object-, Action-, and Location-level semantic relevance. We describe each module in detail below.

Question Analyzer. In real-world applications, questions may appear in diverse formats (e.g., open-ended or multiple-choice) and span heterogeneous tasks such as recognition, reasoning, and temporal grounding. Directly using the original question as the retrieval query can lead to *incorrect relevance estimation*. For instance, in multiple-choice settings, distractor options introduce misleading

semantic cues; if a ranker attends to these incorrect options, it may assign high relevance scores to irrelevant frames. In contrast, when humans are asked a question, they first identify the key evidence required to answer it. Inspired by this process, we introduce a question analyzer to interpret arbitrary questions and transcribe them into retrieval-friendly queries. Our question analyzer extracts the core intent of the query while filtering out noisy or task-specific artifacts, producing a clean and retrieval-oriented representation.

The relative importance of OBJECT, ACTION, LOCATION evidence varies significantly across queries. For example, some queries emphasize specific entities, while others focus on dynamic events or physical locations. As a result, effective frame retrieval requires adapting the contribution of different semantic aspects to the query. Our question analyzer captures this query-dependent preference and yields important weights along OBJECT-, ACTION-, and LOCATION-dimensions.

In this paper, we use a Qwen3-4B [56] as the question analyzer W:

$$(q, w_o, w_a, w_l) = \mathrm{W}(\hat{q}) \tag{2}$$

where $\hat{q}$ is the original question and q is the transcribed question, *i.e.* the query. The question analyzer assigns importance weights w_o, w_a, w_l to OBJECT, ACTION, and LOCATION, respectively. The importance weights are discrete values from $\{1, 2, 3, 4, 5\}$, with higher values indicating greater importance. They are determined solely by the query and reflect how critical each type of semantic evidence is to answering the query. These weights are used to aggregate semantic relevance scores in the following scoring, enabling query-conditioned frame retrieval.

Semantic Ranker. Conditioning on the query, relevant evidence may arise from the presence attributes of specific entities (OBJECT), dynamic events and temporal changes (ACTION), or spatial relations (LOCATION). To model such semantic relevance, we apply a vision–language model that takes a query-frame pair as input and predicts structured semantic relevance scores along the OBJECT, ACTION, and LOCATION dimensions:

$$(o, a, l) = \mathrm{R}_\theta(q, v) \tag{3}$$

where R_θ denotes the VLM ranker with trainable parameter θ, and $(o, a, l) \in \{1, 2, 3, 4, 5\}$ are the predicted relevance scores corresponding to OBJECT, ACTION, and LOCATION respectively. These scores provide an interpretable representation of how each frame supports the query and serve as the basis for frame ranking and selection. Rather than relying on explicit semantic annotations or handcrafted rules, we model relevance implicitly by learning how different semantic aspects of a frame align with the query. In this paper, the semantic ranker is a lightweight pretrained VLM Qwen3-VL-4B-Instruct [4] and is finetuned on a frame-level ranking video dataset.

Scoring. Given the semantic relevance scores $\mathrm{R}_\theta(q, v) = (o, a, l)$ and the query-adaptive weights (w_o, w_a, w_l), we compute a unified ranking score for each frame:

$$\mathrm{Score}(q, v) = \tilde{w}_o \cdot o + \tilde{w}_a \cdot a + \tilde{w}_l \cdot l \tag{4}$$

where $(\tilde{w}_o, \tilde{w}_a, \tilde{w}_l)$ are normalized importance weights via Softmax.

The proposed `QSRanker` decouples question understanding from visual grounding, leverages the question analyzer (LLM) only for estimating evidence importance and generating retrieval-friendly query, and allows the semantic ranker to focus on learning fine-grained visual relevance.

3.2 Training the Semantic Ranker

To align a pretrained VLM with the semantic ranking (Eq. 3), we need to collect video data, annotate videos at the frame level, and fine-tune the pretrained VLM. The resulting semantic ranker can understand the query, the visual clues, their semantic relationship, and how to score them along OBJECT, ACTION, and LOCATION dimensions, explicitly enforcing the correct ranking for query-frame pairs.

Video Data. We construct a fine-tuning dataset, `Video-Ranker-35K`, by randomly sampling 35K videos from LLaVA-Video-178K [80]. For both single-round and multi-round conversations, we retain only one question per video, resulting in 35K video–query pairs. Among them, 28,848 videos are shorter than 30 seconds, and 6,154 are shorter than 60 seconds. `Video-Ranker-35K` covers three video QA formats: open-ended (75.9%), captioning (14.5%), and multiple-choice (9.6%). The videos originate from diverse sources within LLaVA-Video-178K, including YouTube (78.9%) [61, 66, 85], Charades (9.5%) [52], YouCook2 (6.0%) [83], NextQA (2.9%) [65], ActivityNet (2.4%) [17], and Ego4D (0.3%) [20].

Semantic Annotation. For each (video, question) instance, we first prompt the question analyzer to transcribe the question $(\hat{q})$ to a query (q). Then, we uniformly sample frames $\mathcal{V} = \{v_i\}_{i=1}^{T_V}$ and form T_V query-frame pairs $\{(q, v_i)\}_{i=1}^{T_V}$ where $T_V = 8$. The semantic annotation task is to obtain the frame-level semantic relevance scores $\{(o_i, a_i, l_i)\}_{i=1}^{T_V}$.

Manual annotation of fine-grained semantic scores is prohibitively expensive. The state-of-the-art large-scale VLMs (*e.g.* GPT-5 [53], Gemini 3 [19], Qwen3-VL [4], *etc.*) can serve as automatic annotators to generate semantic relevance labels. In this paper, we leverage the strong off-the-shelf VLM Qwen-VL-30B [4] to produce semantic supervision at scale. We prompt Qwen-VL-30B to generate discrete semantic relevance scores, *i.e.* $o, a, l \in \{1, 2, 3, 4, 5\}$ for all query-frame pairs. Therefore, we build the triplet dataset for supervised fine-tuning, $\mathcal{D}_{\text{SFT}} = \{(q, v_i, y_i)_{i=1}^{T_V}\}$ where $y_i = (o_i, a_i, l_i)$.

Supervised Fine-Tuning. We fine-tune the semantic ranker $\mathrm{R}_\theta(q, v)$ via supervised fine-tuning (SFT) [45] on the annotated `Video-Ranker-35K`.

$$\mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}_{(q,v,y)\sim\mathcal{D}_{\text{SFT}}} \Big[\ell_{\text{CE}}(\hat{o}, o) + \ell_{\text{CE}}(\hat{a}, a) + \ell_{\text{CE}}(\hat{l}, l)\Big] \tag{5}$$

where $y = (o, a, l)$ and ℓ_{CE} is the Cross-Entropy loss over the 5 discrete score tokens. After SFT training, the semantic ranker can generate aligned structured relevance scores along OBJECT, ACTION, and LOCATION for individual query-frame pairs.

3.3 QSRetrieval: Relevance–Diversity Temporal Retrieval

Visual Distance. To prevent redundant visual appearance, we need to measure the visual distance of selected frames. To that end, we obtain global visual representations using the vision encoder of the semantic ranker, *i.e.* $g = f(v)$. Specifically, we aggregate embeddings of all vision tokens via mean pooling to produce a global embedding. We define their visual distance as the Euclidean (L2) distance, *i.e.* $d_{ij} = \|g_i - g_j\|_2$. We empirically observe L2 distance is more sensitive to visual change, compared to cosine similarity, which measures only angular difference and discards magnitude information. L2 distance captures both directional and norm variations in the embedding space. L2 distance better reflects subtle appearance and layout changes between frames, making it more suitable for redundancy reduction.

Relevance–Diversity Retrieval. To balance query-to-frame relevance and visual diversity in Eq. 1, we iteratively select a new frame from the complement set $A^c = \mathcal{V} \setminus A$:

$$v^* = \arg\max_{v \in A^c} \left\{ \lambda \cdot \text{Score}(q, v) + (1 - \lambda) \min_{v' \in A} \|f(v) - f(v')\|_2 \right\} \tag{6}$$

where A is the current selection and $\mathcal{V}$ is the whole video. We greedily search for v so that it maximizes the query-to-frame relevance score and the visual distance with respect to selected frames.

Temporal Windowing. Eq. 6 guides the retrieval to attend to both relevance and diversity. However, it neglects temporal coverage introduced by Eq. 1. To encourage temporal coverage, we divide the video frame set $\mathcal{V}$ into K disjoint temporal windows $\{\mathcal{V}_i\}_{i=1}^K$ according to their timestamps, such that

$$\mathcal{V} = \bigcup_{i=1}^{K} \mathcal{V}_i, \quad \mathcal{V}_i \cap \mathcal{V}_j = \emptyset \text{ for } i \neq j \tag{7}$$

After partitioning the whole video into K temporal windows, we restrict the selection within each window to ensure temporal coverage. Specifically, for each window $\mathcal{V}_i$, we select at most one frame by

$$v_i^* = \arg\max_{v \in \mathcal{V}_i} \left\{ \lambda \cdot \text{Score}(q, v) + (1 - \lambda) \min_{v' \in A} \|f(v) - f(v')\|_2 \right\} \tag{8}$$

Temporal-Aware Retrieval. While Eq. 8 enforces window-level coverage, the retrieval across temporal windows should adapt to different videos. In this paper, we focus on two challenging tasks: *long* video understanding and *streaming* video understanding.

Motivated by how humans search for evidence in videos, we couple temporal windowing with an adaptive traversal strategy. When answering questions on *long videos*, people typically jump to the most related moment in memory and then search locally around it; for *streaming videos*, people instead prioritize recent context and progressively trace back in time. We instantiate these

two behaviors with different window-ordering and reference-updating rules while keeping the same relevance–diversity scoring.

- **Long Videos.** Let $\mathcal{I} = \{1, \ldots, K\}$ be temporal window indices and maintain a set of available windows $\mathcal{U} \subseteq \mathcal{I}$ (initialized as $\mathcal{U} = \mathcal{I}$). At iteration t, we first choose a *global anchor* from the remaining windows:

$$a^{(t)} = \underset{v \in \bigcup_{i \in \mathcal{U}} \mathcal{V}_i}{\arg\max}\ \mathrm{Score}(q, v) \tag{9}$$

  We denote the corresponding window index of $a^{(t)}$ by $i^{(t)}$. Then, we search locally within a temporal radius r around $i^{(t)}$:

$$\mathcal{N}^{(t)} = \{\, i \in \mathcal{U} \mid |i - i^{(t)}| \le r \,\} \tag{10}$$

  We apply Eq. 8 to search over neighboring windows $i \in \mathcal{N}^{(t)}$ and use the anchor as the reference to compute visual distance:

$$v_i^* = \underset{v \in \mathcal{V}_i}{\arg\max} \Big\{ \lambda\ \mathrm{Score}(q, v) + (1 - \lambda)\, \|f(v) - f(a^{(t)})\|_2 \Big\} \tag{11}$$

  After this local expansion, we mask out visited windows by updating $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{N}^{(t)}$ and repeat the process until K windows are selected. This iterative *anchor-and-expand* scheme naturally recovers evidence that may appear in multiple distant temporal regions.
- **Streaming Videos.** For streaming videos, we traverse windows from recent to past. Starting from the most recent window, we first select

$$v_K^* = \underset{v \in \mathcal{V}_K}{\arg\max}\, \mathrm{Score}(q, v) \tag{12}$$

  We then progressively move to earlier windows $i = K - 1, \ldots, 1$, using the previous selection as the reference, and select

$$v_i^* = \underset{v \in \mathcal{V}_i}{\arg\max} \Big\{ \lambda\ \mathrm{Score}(q, v) + (1 - \lambda)\, \|f(v) - f(v_{i+1}^*)\|_2 \Big\} \tag{13}$$

  This recency-first traversal explicitly prioritizes near-term evidence by construction, while the distance term discourages selecting visually redundant frames across adjacent windows.

4 Experiments

We fine-tune the proposed semantic ranker using SFT on `Video-Ranker-35K`. Training is conducted on 8× NVIDIA A6000 GPUs (48GB VRAM each) with DeepSpeed ZeRO-2 [49] and bfloat16 precision. We adopt parameter-efficient tuning via LoRA [23] (rank=32, α=64). The global batch size is 128 with cosine learning rate decay (base LR 1×10^{-4}, warmup ratio 3%). SFT takes approximately 12 hours for one epoch. During SFT, we freeze the LLM backbone and

Table 1: Experiments on long video understanding (LVBench). LVBench has six types of tasks: entity recognition (ER), event understanding (EU), key information retrieval (KIR), temporal grounding (TG), reasoning (Rea) and summarization (Sum).

Model	LLM	Frames	Overall	ER	EU	KIR	TG	Rea	Sum
Deep Video Discovery [79]	-	-	74.2	73.4	73.3	80.4	72.3	70.6	74.1
Seed1.5-VL-Thinking [21]	200B	-	64.6	65.4	63.4	68.0	53.6	63.7	46.6
AdaReTaKe [60]	72B	≤1024	53.3	53.0	50.7	62.2	45.5	54.7	37.9
GLM-4V-Plus [82]	-	30	38.3	39.9	35.8	34.8	37.7	40.0	32.8
MiniCPM-o 2.6 [71]	8B	16	38.9	37.1	38.2	37.5	32.7	46.3	41.4
InternVL2-40B [13]	34B	16	39.6	37.4	39.7	43.4	31.4	42.5	41.4
Qwen2-VL-72B [58]	72B	48	41.3	38.0	41.1	38.3	41.4	46.5	46.6
TimeMarker [10]	8B	≤128	41.3	42.8	39.1	34.9	38.7	38.2	48.8
MiniCPM-o 2.6 [71]	8B	32	42.0	41.4	40.7	39.2	35.5	48.3	37.9
MiniCPM-o 2.6 [71]	8B	64	42.3	43.0	42.5	39.9	39.1	46.8	31.0
InternVL2.5-78B [12]	72B	16	43.6	43.8	42.0	42.1	36.8	51.0	37.9
`QSVideo` (Ours)	8B	16	45.8	47.3	41.7	49.8	35.5	50.8	32.8
SEAL [57]	34B	16	45.9	47.9	41.3	51.5	32.3	43.3	39.7
GLM-4V-Plus [82]	-	≤300	48.7	46.2	47.8	54.1	**42.7**	46.5	37.9
GPT-4o [27]	-	60	48.9	48.9	**49.5**	48.1	40.9	50.3	**50.0**
`QSVideo` (Ours)	8B	32	**49.1**	**51.1**	45.6	**54.6**	39.6	**53.2**	32.8

fine-tune the vision tower and merger modules, because the semantic ranker mainly requires improving the visual–language alignment between query and frames, which is determined by the vision tower and merger modules. If we fine-tune the LLM on the query-frame dataset `Video-Ranker-35K`, it may affect its strong language understanding ability. In our experiments, `QSVideo` adopts MiniCPM-o 2.6 [71] as the backbone video VLM.

4.1 Long Video Understanding

The long video understanding benchmark LVBench [59] has 103 long videos and around 15 QAs per video, resulting in 1,549 QAs in total. The average video duration is **67 minutes**. LVBench comprehensively includes 6 types of tasks: entity recognition (677), event understanding (647), key information retrieval (291), reasoning (201), temporal grounding (220), and summarization (58). We benchmark `QSVideo` under different frame budgets (16 frames and 32 frames) and compare it with the state-of-the-art video VLMs, as shown in Table 1.

Effective Visual Evidence Discovery. It is challenging to discover relevant visual clues from long videos because visual clues are sparsely distributed. `QSVideo` achieves an overall accuracy of **49.1**%, demonstrating that `QSRetrieval` can effectively identify informative frames for answering diverse questions. `QSVideo` achieves the best performance on several tasks that require accurate evidence discovery, including *entity recognition*, *key information retrieval*, and *reasoning*. These tasks heavily depend on locating relevant moments in long videos. The improvements demonstrate that the proposed query-aware semantic ranking can effectively capture structured visual evidence aligned with the query.

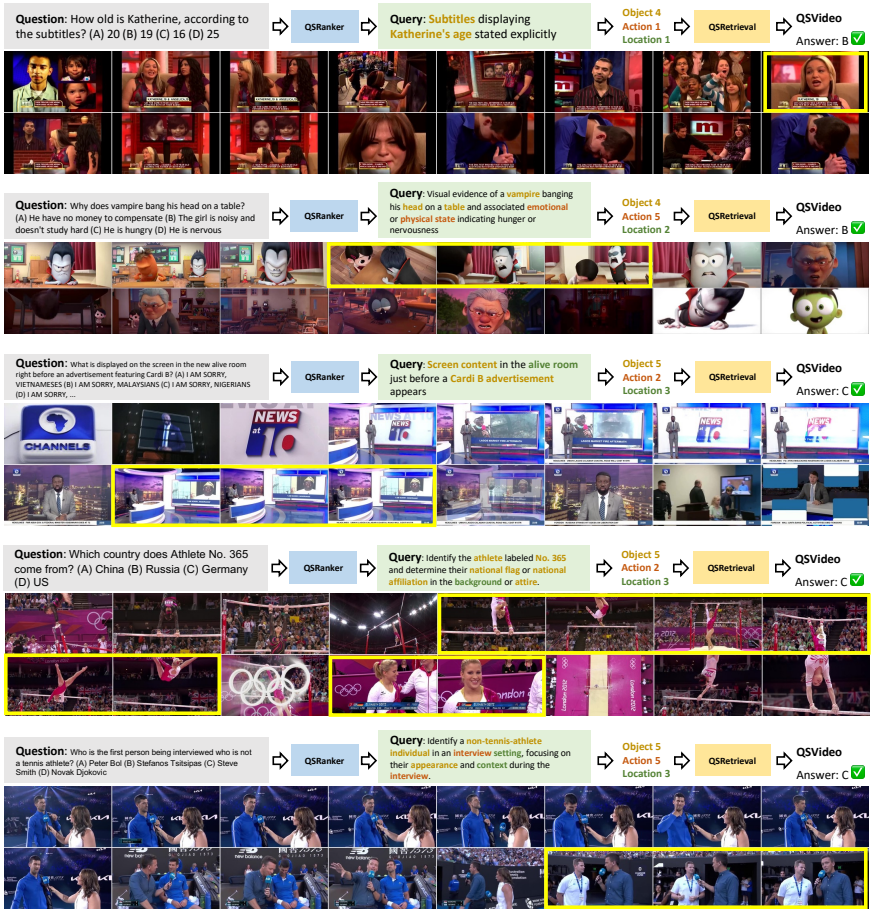

Fig. 2: Visualization of QSVideo on LVBench. The number of selected frames is 16. We highlight frames including important supportive visual evidence.

Boosting Video VLM Performance. The backbone VLM (MiniCPM-o 2.6) reports accuracy of 38.9%, 42.0% and 42.3% using 16 frames, 32 frames and 64 frames. `QSVideo` greatly improves accuracy by +6.9% and +7.1% using 16 frames and 32 frames. It indicates that explicitly modeling query-to-frame retrieval can significantly improve the performance of video VLMs.

Efficient Long Video Understanding. `QSVideo` achieves competitive performance using only an existing 8B VLM and 16/32 frames, while several competing approaches rely on substantially larger models ($\geq$ 34B) or significantly more frames ($\gg$ 32). Our results suggest that locating key visual evidence can be more effective than naively scaling up model size or frame budgets for long video understanding, which requires a dramatic increase in the VRAM footprint.

Table 2: Experiments on streaming video understanding (StreamingBench). StreamingBench real-time benchmark includes ten types of tasks: object perception (OP), causal reasoning (CR), clips summarize (CS), attribute perception (ATP), event understanding (EU), text-rich understanding (TR), prospective reasoning (PR), spatial understanding (SU), action perception (ACP), counting (CT).

Model	LLM	Frames	Overall	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT
MiniCPM-o 2.6 [71]	8B		64.44	67.30	71.09	74.45	71.24	71.43	64.49	55.56	59.35	62.89	35.75
MiniCPM-V 2.6 [71]	8B		66.36	70.30	65.62	79.50	76.80	67.08	67.29	66.67	55.28	59.49	**45.60**
InternVL2 [13]	8B	8	68.52	76.02	60.94	72.24	82.35	68.94	70.72	72.22	66.67	60.62	41.97
LLaVA-OV [31]	7B		72.12	79.29	72.66	82.97	82.35	72.67	69.16	75.00	68.29	69.12	37.31
QSVideo (Ours)	8B		**77.68**	**83.11**	**79.69**	**86.12**	**83.66**	**80.12**	**83.49**	**76.85**	**72.76**	**74.22**	44.04
VILA-1.5 [42]	8B	14	61.54	71.12	57.81	74.68	72.22	70.81	62.62	51.85	51.22	60.34	18.65
MiniCPM-o 2.6 [71]	8B	16	65.12	69.21	77.34	76.66	72.22	72.67	64.17	56.48	58.54	62.89	31.61
MiniCPM-V 2.6 [71]	8B	16	70.24	75.20	71.09	83.28	80.39	68.94	68.85	70.37	59.76	66.29	46.63
InternVL2 [13]	8B	16	70.52	75.20	64.84	75.08	**84.97**	75.16	72.90	73.15	65.85	64.59	42.49
VITA-1.5 [18]	7B	16	70.88	77.66	82.81	82.33	79.34	72.67	71.34	67.59	62.20	69.12	32.12
LLaVA-OV [31]	7B	16	73.76	79.56	79.69	82.65	83.66	73.29	74.77	72.22	69.11	69.97	40.93
Claude 3.5 [1]	-	20	74.04	82.45	73.77	82.43	82.40	76.39	**85.56**	61.68	60.73	67.88	**47.62**
QueryStream [78]	7B	1fps	74.04	82.11	**83.59**	78.23	82.69	75.47	80.06	**79.63**	63.01	67.90	42.55
InfiniPot-V [29]	7B	0.5fps	76.4	-	-	-	-	-	-	-	-	-	-
StreamForest [75]	7B	1fps	77.26	83.11	82.81	82.65	84.26	77.50	78.19	76.85	69.11	75.64	54.40
QSVideo (Ours)	8B	16	**79.36**	**84.47**	82.81	**88.96**	84.31	**81.99**	85.05	75.00	**74.39**	**77.34**	44.56

Visualization. We visualize results of QSVideo on LVBench in Fig. 2. In Fig. 2, we show (1) the original questions, (2) the retrieval-friendly queries generated by QSRanker, (3) the importance weights of Object, Action and Location generated by QSRanker, (4) the selected frames retrieved by QSRetrieval and (5) the answers obtained by QSVideo. From Fig. 2, we observe that (1) QSRanker effectively transcribes the original questions and correctly measures the importance of Object, Action and Location; (2) QSRetrieval effectively locates the visual evidence; (3) QSVideo effectively uses these visual evidence and obtains the correct answers. For example, the question, "Which country does Athlete No. 365 come from? (A) China (B) Russia (C) Germany (D) US", QSVideo first finds the Athlete No. 365 and then figures out the nationality from the screen content.

4.2 Streaming Video Understanding

Unlike *offline* long video understanding, we cannot access to the whole video in streaming video understanding. Because real-world applications require *online* understanding of live videos, *e.g.* AR glasses, live streaming, and autonomous driving. In streaming videos, frames arrive sequentially and future content is not accessible. So, online video VLMs process streaming videos and respond to temporally aligned real-time queries [9, 36, 48, 63].

We benchmark QSVideo on real-time video understanding using StreamingBench [36]. StreamingBench real-time benchmark extensively comprises ten tasks: object perception, causal reasoning, clip summarization, attribute perception, event understanding, text-rich understanding, prospective reasoning, spatial understanding, action perception, and counting. It includes 2,500 QAs from 499 unique videos, which provides a comprehensive evaluation of a VLM's working and episodic memory under streaming settings. The query can be proposed at any time t. We follow the standard evaluation of StreamingBench to set

Table 3: Seamless plug-and-play over different video VLMs. We integrate `QSVideo` with different vodeo VLMs (MiniCPM-V 2.6 [71], InternVL2 [13] and LLaVA-OV [31]) under different frames.

Model	Frames	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	Overall	Δ
MiniCPM-V	8	70.30	65.62	79.50	76.80	67.08	67.29	66.67	55.28	59.49	45.60	66.36	+4.60↑
+QSVideo		76.57	64.06	82.97	81.05	73.29	74.14	73.15	59.35	65.72	45.08	70.96	
MiniCPM-V	16	75.20	71.09	83.28	80.39	68.94	68.85	70.37	59.76	66.29	46.63	70.24	+1.60↑
+QSVideo		77.66	69.53	85.80	81.37	73.29	73.52	74.07	62.60	65.72	41.97	71.84	
MiniCPM-V	32	77.66	68.75	82.02	80.72	68.94	73.52	75.00	59.35	64.59	45.60	70.80	+2.16↑
+QSVideo		78.75	71.09	85.49	83.01	72.67	76.32	75.93	60.57	67.71	45.08	72.96	
InternVL2	8	76.02	60.94	72.24	82.35	68.94	70.72	72.22	66.67	60.62	41.97	68.52	+2.76↑
+QSVideo		78.75	67.19	75.39	85.95	76.40	71.96	70.37	64.63	65.44	44.04	71.28	
InternVL2	16	75.20	64.84	75.08	84.97	75.16	72.90	73.15	65.85	64.59	42.49	70.52	+1.72↑
+QSVideo		77.93	64.84	78.86	85.95	74.53	75.08	75.00	66.26	66.57	43.52	72.24	
InternVL2	32	76.29	67.19	80.44	85.29	72.67	74.77	75.00	64.63	62.89	45.08	71.52	+1.40↑
+QSVideo		79.02	68.75	80.44	86.60	72.67	76.01	74.07	66.26	66.86	44.04	72.92	
LLaVA-OV	8	79.29	72.66	82.97	82.35	72.67	69.16	75.00	68.29	69.12	37.31	72.12	+2.36↑
+QSVideo		80.38	75.00	85.80	83.66	73.29	76.01	75.93	70.33	71.95	37.31	74.48	
LLaVA-OV	16	79.56	79.69	82.65	83.66	73.29	74.77	72.22	69.11	69.97	40.93	73.76	+0.92↑
+QSVideo		82.29	74.22	86.44	83.33	75.16	76.32	73.15	67.48	71.10	40.93	74.68	
LLaVA-OV	32	80.65	78.12	82.33	84.31	70.19	75.08	75.00	66.67	69.97	40.41	73.56	+0.84↑
+QSVideo		81.74	78.12	84.23	81.70	69.57	77.88	74.07	70.33	69.97	41.97	74.40	

the context time to $(t-60, t)$, where t is the query time. We compare `QSVideo` with existing video VLMs under the same frame budgets, 8 frames and 14–16 frames, as shown in Table 2. We also compare with three recent methods that sample frames by 0.5–1 fps.

Retrieval Matters in Short Contexts. One might assume that retrieval is less important in streaming understanding because the context window is already short (60 seconds), and uniform sampling should be sufficient. However, Table 2 shows that `QSVideo` still improves performance by +11.32% and 14.24% using 8 frames and 16 frames compared to the backbone VLM, MiniCPM-o 2.6. Enhanced by `QSRetrieval`, `QSVideo` improves the overall accuracy from 72.12% (LLaVA-OV) to 77.68% with 8 frames, and further achieves SoTA accuracy of 79.36% with only 16 frames. It proves that `QSRetrieval` can effectively discover informative visual evidence within a short temporal context, leading to much better online video understanding.

Better Integrates Visual Clues. We note the improvements are particularly evident on tasks requiring reasoning over multiple visual cues, such as clip summarize, event understanding, and prospective reasoning (Table 2). For instance, under 16 frames, `QSVideo` achieves 81.99% on event understanding, significantly higher than most video VLMs. These tasks require integrating observations that may appear at different moments within the recent context window. By selecting frames that are semantically aligned with the query, `QSVideo` preserves complementary pieces of evidence across time, enabling the model to better connect temporally separated clues when answering streaming queries.

Plug-and-Play over video VLMs. We further apply `QSVideo` to different online video VLMs, including MiniCPM-V 2.6 [71], InternVL2 [13], and LLaVA-OneVision [31], under 8-, 16-, and 32-frame settings. Table 3 shows consistent overall improvements across all backbones. It demonstrates that `QSVideo` is

Table 4: End-to-end latency.

Model	LLM	Candidate	Selected	Overall (%)	Latency (s) ↓	Throughput ↑
SEAL [57]	34B	2045	16	45.9	1702.24	1.2fps
AdaReTaKe [60]	72B	1024	-	53.3	178.15	5.8fps
`QSVideo`	8B	2048	32	49.1	91.17	22.5fps
`QSVideo`	8B	2048	16	45.8	88.73	23.1fps

Table 5: Breakdown of latency.

Acceleration	QSRanker			QSRetrieval		VLM Inference	
	Visual Distance	Question Analyzer	Semantic Ranker	16 Frames	32 Frames	16 Frames	32 Frames
Naive	16.81s	1.15s	608.40s	0.87s	1.10s	8.71s	11.05s
+Batch Inference	16.78s	1.32s	167.07s	0.88s	1.05s	8.68s	10.94s
+Data Parallel	16.78s	1.15s	61.32s	0.89s	1.05s	8.59s	10.87s

model-agnostic and works in a plug-and-play manner. The experiments prove that discovering informative visual clues can facilitate different video VLMs, especially under limited frames. At 8 frames, `QSVideo` improves MiniCPM-V, InternVL2, and LLaVA-OV by +4.60%, +2.76%, and +2.36%, respectively.

4.3 Efficiency

We evaluate the end-to-end latency of `QSVideo` on a long video (video key: Za2Z_JRxCuk) from LVBench with a duration of 34 minutes, as shown in Table 4. `QSVideo` uniformly samples 2048 candidate frames from the original video and selects 16 or 32 frames. SEAL [57] samples frames at 1fps resulting in 2045 candidate frames and selects 16 frames, while AdaReTake [60] compresses 1024 frames. `QSVideo` is faster than SEAL by 18.7×–19.2× and faster than AdaReTake by 2.0×. We list the detailed latency of components in Table 5. `QSVideo` can be easily accelerated from two aspects: (i) lower end-to-end latency due to lightweight VLMs and (ii) higher throughput due to the parallelizable semantic ranker.

5 Ablation Studies

5.1 Ablation Study of λ

We benchmark the effect of λ on LVBench under different frame budgets (Table 6). We observe: (i) the performance is relatively insensitive to λ, which supports "plug-and-play" feature, (ii) there is a strong positive relationship between λ and the performance (Pearson correlation coefficient $r = 0.88$) under 32 frames, which proves that $\lambda = 0.9$ effectively balances relevance and diversity.

Table 6: Ablation study of λ.

Frames	λ	Overall	ER	EU	KIR	TG	Rea	Sum
16	0.1	46.1	46.7	44.2	49.8	37.3	50.2	32.8
	0.3	45.6	47.3	43.3	49.1	33.6	48.3	31.0
	0.5	46.0	47.6	42.3	51.2	35.9	50.7	31.0
	0.7	45.6	47.0	41.7	50.9	35.0	50.2	31.0
	0.9	45.8	47.3	41.7	49.8	35.5	50.8	32.8
32	0.1	48.2	49.0	45.1	53.6	38.2	51.7	34.5
	0.3	48.3	51.1	44.4	54.3	36.4	51.2	31.0
	0.5	48.2	51.3	44.5	54.0	35.5	52.2	29.3
	0.7	48.7	51.3	45.0	55.0	39.1	51.7	32.8
	0.9	49.1	51.1	45.6	54.6	39.6	53.2	32.8

5.2 Ablation Study of `QSRanker`

We benchmark `QSRanker` on LVBench. 1549 QAs have "time_reference" annotations, which are used as ground-truth for ranking evaluation. `QSRanker` first uniformly samples 2048 frames and then scores frames. Recall@K measures whether at least one of the top-K ranked frames falls within the "time_reference" interval. The baseline is a holistic ranker that only generates a single score. The baseline reports 17.38% Recall@1, 26.81% Recall@5, 32.43% Recall@10, 46.77% Recall@50, 52.78% Recall@100, while `QSRanker` reports 17.51% Recall@1, 27.20% Recall@5, 33.20% Recall@10, 50.65% Recall@50, 58.20% Recall@100. These results demonstrate that `QSRanker` effectively improves evidence ranking.

6 Conclusion

In this paper, we present `QSVideo`, a multimodal retrieval framework to improve video understanding. `QSVideo` designs `QSRanker` to analyze and transcribe arbitrary questions to retrieval-friendly queries, and score frames along Object-, Action-, and Location-dimensions. `QSVideo` further proposes `QSRetrieval` to discover visual evidence by jointly considering relevance, redundancy reduction, and temporal coverage. On LVBench, `QSVideo` achieves the best trade-off between model size (8B), the number of frames (32 frames), and QA accuracy (49.1%). On StreamingBench, `QSVideo` reports very competitive QA accuracy 77.68% and 79.36% by only using 8 and 16 frames. We integrate `QSVideo` with different video VLMs and observe consistent improvements, especially for limited frames.

Acknowledgements

This work was supported in part by the National Science Foundation (award #2500983). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

References

1. Anthropic: Claude 3.5 sonnet (2024)
2. Arnab, A., Iscen, A., Caron, M., Fathi, A., Schmid, C.: Temporal chain of thought: Long-video understanding by thinking in frames. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Bao, X., Xie, C., Tang, H., Weng, T., Wang, X., Zheng, Y., Wang, X.: DynImg: Key frames with visual prompts are good representation for multi-modal video understanding. In: International Conference on Computer Vision (ICCV) (2025)
6. Buch, S., Nagrani, A., Arnab, A., Schmid, C.: Flexible frame selection for efficient video reasoning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 29071–29082 (2025)
7. Chatterjee, D., Remelli, E., Song, Y., Tekin, B., Mittal, A., Bhatnagar, et al.: Memory-efficient streaming VideoLLMs for real-time procedural video understanding. In: International Conference on Computer Vision (ICCV) (2025)
8. Chen, B., Yue, Z., Chen, S., Wang, Z., Liu, Y., Li, P., Wang, Y.: LVAgent: Long video understanding by multi-round dynamical collaboration of MLLM agents. In: International Conference on Computer Vision (ICCV) (2025)
9. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: VideoLLM-online: Online video large language model for streaming video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18407–18418 (2024)
10. Chen, S., Lan, X., Yuan, Y., Jie, Z., Ma, L.: TimeMarker: A versatile video-LLM for long and short video understanding with superior temporal localization ability. arXiv preprint arXiv:2411.18211 (2024)
11. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
12. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
13. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821 (2024)
14. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24185–24198 (2024)
15. Diko, A., Wang, T., Swaileh, W., Sun, S., Patras, I.: ReWind: Understanding long videos with instructed learnable memory. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)

16. Dong, Y., Liu, Z., Sun, H.L., Yang, J., Hu, W., Rao, Y., Liu, Z.: Insight-V: Exploring long-chain visual reasoning with multimodal large language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
17. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: ActivityNet: a large-scale video benchmark for human activity understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 961–970 (2015)
18. Fu, C., Lin, H., Wang, X., Zhang, Y.F., Shen, Y., Liu, X., Li, Y., Long, Z., Gao, H., Li, K., et al.: VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. arXiv preprint arXiv:2501.01957 (2025)
19. Gemini 3 pro model (2025)
20. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4D: Around the world in 3,000 hours of egocentric video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18995–19012 (2022)
21. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1.5-VL technical report. arXiv preprint arXiv:2505.07062 (2025)
22. Guo, W., Chen, Z., Wang, S., He, J., Xu, Y., Ye, J., Sun, Y., Xiong, H.: Logic-in-Frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR) **1**(2), 3 (2022)
24. Hu, K., Gao, F., Nie, X., Zhou, P., Tran, S., Neiman, T., Wang, L., Shah, M., Hamid, R., Yin, B., et al.: M-LLM based video frame selection for efficient video understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13702–13712 (2025)
25. Huang, D.A., Radhakrishnan, S., Yu, Z., Kautz, J.: FRAG: Frame selection augmented generation for long video and long document understanding. arXiv preprint arXiv:2504.17447 (2025)
26. Huang, Z., Li, X., Li, J., Wang, J., Zeng, X., Liang, C., Wu, T., Chen, X., Li, L., Wang, L.: Online video understanding: OVBench and VideoChat-Online. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3328–3338 (June 2025)
27. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
28. Kang, H., Park, Y., Yoo, Y., Choi, Y., Kim, S.J.: Open-ended hierarchical streaming video understanding with vision language models. In: International Conference on Computer Vision (ICCV) (2025)
29. Kim, M., Shim, K., Choi, J., Chang, S.: InfiniPot-V: Memory-constrained KV cache compression for streaming video. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
30. Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., Zhang, Y., Liu, Z., Li, C.: LLaVA-NeXT: Stronger LLMs supercharge multimodal capabilities in the wild (May 2024)
31. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: LLaVA-OneVision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)

32. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: VideoChat: Chat-centric video understanding. Science China Information Sciences **68**(10), 200102 (2025)
33. Li, M., Zhang, Y., Long, D., Keqin, C., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., Zhou, J., Lin, J.: Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026)
34. Li, Y., Wang, C., Ji, J.: LLaMA-VID: An image is worth multiple tokens in large language models. arXiv preprint arXiv:2311.17043 (2023)
35. Lin, B., Zhu, B., Ye, Y., Cui, J., Ning, M., Jin, P., Li, L.Y.: Video-LLaVA: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
36. Lin, J., Fang, Z., Chen, C., Wan, Z., Luo, F., Li, P., Liu, Y., Sun, M.: StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. arXiv preprint arXiv:2411.03628 (2024)
37. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
38. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: LLaVA-NeXT: improved reasoning, ocr, and world knowledge (January 2024)
39. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Annual Conference on Neural Information Processing Systems (NeurIPS) **36**, 34892–34916 (2023)
40. Liu, R., Sun, S., Tang, H., Gao, W., Li, G.: Flow4Agent: Long-form video understanding via motion prior from optical flow. In: International Conference on Computer Vision (ICCV) (2025)
41. Liu, S., Zhao, C., Xu, T., Ghanem, B.: BOLT: Boost large vision-language model without training for long-form video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
42. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., Li, X., Fang, Y., Chen, Y., Hsieh, C.Y., Huang, D.A., Cheng, A.C., Nath, V., Hu, J., Liu, S., Krishna, R., Xu, D., Wang, X., Molchanov, P., Kautz, J., Yin, H., Han, S., Lu, Y.: NVILA: Efficient frontier visual language models. arXiv preprint arXiv:2412.04468 (2024)
43. Ma, Z., Gou, C., Shi, H., Sun, B., Li, S., Rezatofighi, H., Cai, J.: DrVideo: Document retrieval based long video understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
44. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-ChatGPT: Towards detailed video understanding via large language and vision models. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2024)
45. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Annual Conference on Neural Information Processing Systems (NeurIPS) **35**, 27730–27744 (2022)
46. Pan, J., Zhang, Q., Zhang, R., Lu, M., Wan, X., Zhang, Y., Liu, C., She, Q.: TimeSearch-R: Adaptive temporal search for long-form video understanding. In: International Conference on Learning Representations (ICLR) (2026)
47. Pang, Z., Wang, Y.X.: Mr. Video: "MapReduce" is the principle for long video understanding. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)

48. Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., Wang, J.: Streaming long video understanding with large language models. In: Annual Conference on Neural Information Processing Systems (NeurIPS). vol. 37, pp. 119336–119360 (2024)
49. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models. In: International Conference for High Performance Computing, Networking, Storage and Analysis. pp. 1–16. IEEE (2020)
50. Shen, X., Zhang, W., Chen, J., Elhoseiny, M.: Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
51. Shu, Y., Liu, Z., Zhang, P., Qin, M., Zhou, J., Liang, Z., Huang, T., Zhao, B.: Video-XL: Extra-long vision language model for hour-scale video understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
52. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision (ECCV). pp. 510–526. Springer (2016)
53. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai GPT-5 system card. arXiv preprint arXiv:2601.03267 (2025)
54. Suo, Y., Ma, F., Zhu, L., Wang, T., Rao, F., Yang, Y.: From trial to triumph: Advancing long video understanding via visual context sample scaling and self-reward alignment. In: International Conference on Computer Vision (ICCV) (2025)
55. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
56. Team, Q.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
57. Wang, L., Chen, Y., Tran, D., Boddeti, V.N., Chu, W.S.: SEAL: Semantic attention learning for long video representation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
58. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
59. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y., et al.: LVBench: An extreme long video understanding benchmark. arXiv preprint arXiv:2406.08035 (2024)
60. Wang, X., Si, Q., Zhu, S., Wu, J., Cao, L., Nie, L.: AdaReTaKe: Adaptive redundancy reduction to perceive longer for video-language understanding. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 5417–5432 (2025)
61. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: InternVid: A large-scale video-text dataset for multimodal understanding and generation. In: International Conference on Learning Representations (ICLR) (2023)
62. Wang, Y., Song, Y., Xie, C., Liu, Y., Zheng, Z.: VideoLLaMB: Long-form video understanding with recurrent memory bridges. In: International Conference on Computer Vision (ICCV) (2025)
63. Wu, S., Chen, J., Lin, K.Q., Wang, Q., Gao, Y., Xu, Q., Xu, T., Hu, Y., Chen, E., Shou, M.Z.: VideoLLM-MoD: Efficient video-language streaming with mixture-of-depths vision computation. In: Annual Conference on Neural Information Processing Systems (NeurIPS). vol. 37, pp. 109922–109947 (2024)

64. Wu, Z., Wang, X., Huang, L., Xu, T., Peng, P.: A training-free framework for long video understanding via video-query-options similarity. In: International Conference on Learning Representations (ICLR) (2026)
65. Xiao, J., Shang, X., Yao, A., Chua, T.S.: NExT-QA: Next phase of question-answering to explaining temporal actions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9777–9786 (June 2021)
66. Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing high-resolution video-language representation with large-scale video transcriptions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5036–5045 (2022)
67. Xue, Z., Zhang, J., Xie, X., Cai, Y., Liu, Y., Li, X., Tao, D.: AdaVideoRAG: Omni-contextual adaptive retrieval-augmented efficient long video understanding. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
68. Yang, Z., Chen, D., Yu, X., Shen, M., Gan, C.: VCA: Video curious agent for long video understanding. In: International Conference on Computer Vision (ICCV) (2024)
69. Yang, Z., Hu, Y., Du, Z., Xue, D., Qian, S., Wu, J., Yang, F., Dong, W., Xu, C.: SVBench: A benchmark with temporal multi-turn dialogues for streaming video understanding. In: International Conference on Learning Representations (ICLR) (2025)
70. Yang, Z., Zhang, K., Hu, Y., Wang, B., Qian, S., Wen, B., Yang, F., Gao, T., Dong, W., Xu, C.: LiveStar: Live streaming assistant for real-world online video understanding. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
71. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: MiniCPM-V: A GPT-4V level MLLM on your phone. arXiv preprint arXiv:2408.01800 (2024)
72. Ye, J., Wang, Z., Sun, H., Chandrasegaran, K., Durante, Z., Eyzaguirre, C., Bisk, Y., Niebles, J.C., Adeli, E., Fei-Fei, L., et al.: Re-thinking temporal search for long-form video understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
73. Ye, X., Gan, Y., Ge, Y., Zhang, X.P., Tang, Y.: ATP-LLaVA: Adaptive token pruning for large vision language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
74. Yu, S., Cho, J., Yadav, P., Bansal, M.: Self-chained image-language model for video localization and question answering. Annual Conference on Neural Information Processing Systems (NeurIPS) pp. 76749–76771 (2023)
75. Zeng, X., Qiu, K., Zhang, Q., Li, X., Wang, J., Li, J., Yan, Z., Tian, K., Tian, M., Zhao, X., et al.: StreamForest: Efficient online video understanding with persistent event memory. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
76. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
77. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Jin, X.: Flash-VStream: Efficient real-time understanding for long video streams. In: International Conference on Computer Vision (ICCV) (2025)
78. Zhang, K., Yang, Z., Wang, B., Qian, S., Xu, C.: QueryStream: Advancing streaming video understanding with query-aware pruning and proactive response. In: International Conference on Learning Representations (ICLR) (2026)

79. Zhang, X., Jia, Z., Guo, Z., Li, J., Li, B., Li, H., Lu, Y.: Deep video discovery: Agentic search with tool use for long-form video understanding. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
80. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: LLaVA-Video: Video instruction tuning with synthetic data. Transactions on Machine Learning Research (June 2025)
81. Zhang, Y., Lu, Y., Wang, T., Rao, F., Yang, Y., Zhu, L.: FlexSelect: Flexible token selection for efficient long video understanding. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
82. ZhipuAI: GLM-4V-Plus-0111 (2025)
83. Zhou, L., Corso, J.: Youcookii dataset (2017)
84. Zhou, Y., He, Y., Su, Y., Han, S., Jang, J., Bertasius, G., Bansal, M., Yao, H.: ReAgent-V: A reward-driven multi-agent framework for video understanding. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
85. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al.: LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. arXiv preprint arXiv:2310.01852 (2023)
86. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)