HFS: Holistic Query-Aware Frame Selection for Efficient Video Understanding

Anonymous Author(s)
Submission Id: 8437

Abstract

Key frame selection is essentially a set-level optimization problem: the quality of the selected subset depends on the interactions among frames, rather than the score of any single frame. Existing methods generally exhibit three major limitations. Point-wise methods score each frame independently and ignore inter-frame dependencies. Although the training-free set-level methods explicitly model the inter-frame relationships, their selection criteria are fixed and cannot be adapted through downstream task feedback. Learnable methods can leverage data-driven training; however, they lack an explicit, differentiable set-quality objective and rely on offline-generated supervision signals. To address these limitations, we propose an end-to-end trainable and task-adaptive framework for frame selection. A Chain-of-Thought approach guides a Small Language Model (SLM) to generate task-specific implicit query vectors, which are combined with multimodal features to enable dynamic, query-aware frame scoring. We further formulate a continuous set-level objective function that jointly accounts for relevance, coverage, and redundancy, enabling differentiable set-level optimization via Gumbel-TopK for selecting optimal frame combinations. Finally, we employ a student-teacher mutual learning strategy, in which the student selector (SLM) and teacher reasoner (MLLM) are trained to align their frame-importance distributions via KL divergence. Combined with cross-entropy loss, this design enables fully end-to-end optimization, eliminating reliance on static pseudo-labels. Experiments across multiple benchmarks, including Video-MME, LongVideoBench, MLVU, and NExT-QA, demonstrate that our method significantly outperforms existing frame-selection approaches.

CCS Concepts

• **Computing methodologies** → **Visual content-based indexing and retrieval**; **Video summarization**.

Keywords

Video understanding, frame selection, multimodal large language model

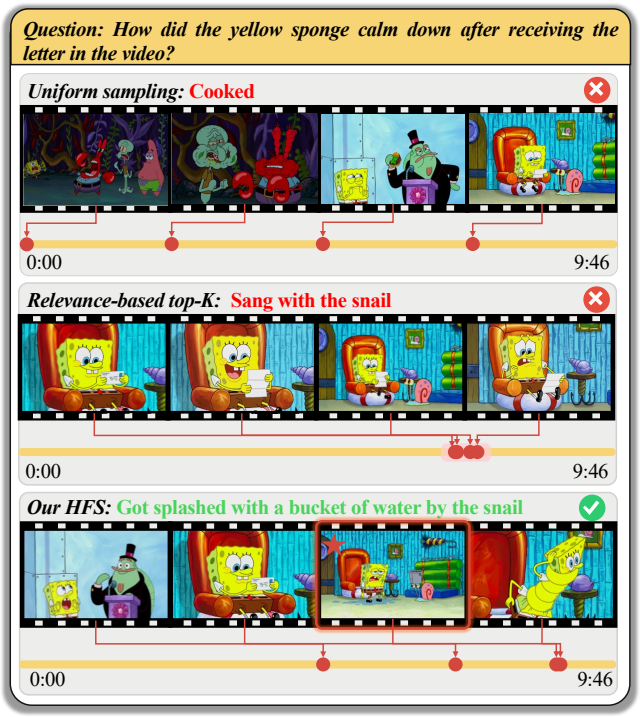

Figure 1: Comparison on an event reasoning task. The task requires the model to locate a sparse, critical moment within a nine-minute video. (a) Uniform sampling sampled frames that completely missed the critical information, leading to a wrong prediction. (b) Relevance-based top-K selection identifies frames that are highly relevant to the query. However, it selects highly redundant frames while overlooking the key event necessary to answer the question. (c) Through CoT-guided query understanding and set-level optimization, our HFS approach effectively suppresses redundancy and accurately locates the action of being splashed with water, guiding the model to produce the correct answer.

1 Introduction

Multimodal Large Language Models (MLLMs) have made significant progress in visual tasks [1, 16, 20], including image understanding, video analysis, and cross-modal reasoning. Unlike static images, video data contains dense temporal information and frames with a high degree of redundancy, which can overwhelm the model's context window and hinder effective processing. This mismatch leads to context overflow when processing lengthy videos, with even moderately sized clips often exceeding the model's processing capabilities. Consequently, selecting informative key frame subsets from raw video has become an essential step in video understanding tasks.

Researchers have proposed various frame selection strategies to address this challenge. One class employs heuristic sampling methods [18, 49], such as uniform sampling at fixed intervals. Another approach introduces learnable scoring mechanisms, assigning importance scores to each frame before applying top-K selection to obtain a critical frame subset [11, 32]. Beyond per-frame scoring, Frame-Voyager [44] formulates frame selection as a combinatorial ranking problem, and FFS [3] introduces a flexible selection operation that learns to determine how many frames to retain for each

query. Further work leverages MLLM to generate spatial-temporal pseudo-labels offline [11], which are then used to supervise the training of lightweight selectors. Although these methods are effective in specific scenarios, fundamental limitations remain.

Frame selection is inherently a set-level problem, as the quality of the selected subset depends on the interactions among frames rather than individual frame relevance alone. However, existing methods have not yet fully addressed this problem. Most methods perform frame-independent relevance scoring [32, 47]. Some methods introduce temporal-coverage heuristic rules [18, 49] or random sampling during selection to alleviate redundancy, but these strategies rely on temporal uniformity or sampling randomness, and fail to explicitly model semantic interactions between frames or provide a differentiable set-quality objective to optimize the subset quality as a whole. The recent training-free set-level method [31] models inter-frame diversity and query correlation through a query-conditioned kernel. However, its query understanding depends on a frozen contrastive VLM text encoder, lacks deep task-intent comprehension, and relies on fixed kernel-based scoring that cannot be adjusted through downstream task feedback. Other learnable methods [11, 44] train selectors using offline-generated pseudo-labels or combinatorial ranking-based selection, but their supervision signals are generated offline before training, disconnected from the downstream task goals, and lack an explicit differentiable set-quality objective to constrain the subset quality. Figure 1 shows a typical failure case: a relevance-based method selects redundant frames corresponding to the same event while omitting the decisive moment required to answer the question.

In order to bridge these gaps, we propose HFS, a Holistic, query-aware Frame Selection framework that, to the best of our knowledge, is the first to integrate a differentiable set-quality objective into end-to-end training for frame selection. HFS contains a lightweight Small Language Model (SLM) acting as a student frame selector, and an MLLM serving as a teacher video reasoner. Inspired by set-function optimization [9, 34, 40], the set-quality objective serves as a regularizer during training, enabling the model to learn what constitutes an effective frame subset directly from the downstream task feedback. This design enables the model to learn what constitutes an effective frame subset directly from the downstream task feedback. To provide accurate scoring inputs for set-level optimization, we design a Chain-of-Thought (CoT)-guided query generation mechanism [39]: SLM understands task intent through CoT guidance, generates a variety of task-specific implicit query vectors, and integrates them with multimodal features through a LoRA adapter [10]. To achieve fully end-to-end training without reliance on static offline supervision, we adopt a mutual learning mechanism [50], where the student selector and teacher reasoner align the frame-importance distributions through KL divergence, jointly optimized with the cross-entropy loss of downstream tasks. This unified objective eliminates reliance on fixed pseudo-labels. Together, these three components, task-aware query understanding, differentiable set-level optimization, and student-teacher mutual learning, form a unified framework in which frame selection and downstream reasoning are jointly optimized, effectively closing the long-standing gap between frame selection quality and task performance.

The main contributions of this paper are as follows:

1. We propose a frame selection framework that integrates a differentiable set-quality objective into end-to-end training, jointly optimizing frame selection with downstream task objectives.
2. We design a CoT-guided query generation mechanism that produces task-specific implicit query vectors for deep task-intent understanding, along with a mutual learning paradigm that eliminates dependence on static offline supervision.
3. Experiments on Video-MME, MLVU, LongVideoBench, and NExT-QA show that our method significantly outperforms existing approaches, specifically, improving MLVU M-AVG by up to 4.0 percentage points and Video-MME overall accuracy by 2.6 percentage points over the baseline, while surpassing the strongest training-free method by up to 2.0 points.

2 Related Work

Video Understanding Using MLLM. Early MLLMs extended image understanding capabilities to video inputs through visual alignment [18]. ShareGPT4Video [4] and LLaVA-Video [49] synthesize dense video descriptions, while MVBench [17] provides a comprehensive evaluation benchmark, and InternVideo2 [37] further scales unified multimodal video representations. Long-form video understanding has been addressed using sparse memory mechanisms [30], temporal modeling [27], context length extension [5], and adaptive compression [28]. Recent studies also explore fine-grained spatial understanding [15, 45], enhancing regional-level comprehension in videos.

Video Frame Selection Using MLLM. Processing all frames in long videos incurs high computational costs [11], while uniform sampling often discards critical information. Consequently, recent methods have shifted toward adaptive frame selection. AKS [32] balances relevance and coverage through a recursive algorithm, M-LLM [11] introduces spatial-temporal pseudo-labels, and FFS [3] enables flexible frame count determination of the number of retained frames. VideoTree [38] proposes hierarchical clustering for coarse-to-fine extraction, while Frame-Voyager [44] and Q-Frame [47] explore query-aware selection mechanisms with dynamic adaptation. MDP^3 [31] leverages determinantal point processes (DPPs) and dynamic programming within a Markov decision framework to perform training-free list-wise frame selection. In summary, the relevance scoring strategies of AKS and Q-Frame remain per-frame independent, although coverage heuristics or random sampling are introduced, these methods lack a differentiable set-quality objective. MDP^3 realizes set-level selection through DPPs, but relies on frozen VLM features and a predefined kernel, which prevents adaptation through downstream task feedback. Although M-LLM and Frame-Voyager can learn from data, their supervision signals are generated offline and similarly lack an explicit differentiable set-quality objective. In contrast, our method incorporates a differentiable set-quality objective as a regularizer within end-to-end training, enabling joint optimization with downstream tasks through online mutual learning.

3 Holistic Query-Aware Frame Selection

3.1 Overview

The overall architecture of our holistic frame selection (HFS) framework is illustrated in Figure 2. The design principle of HFS is to

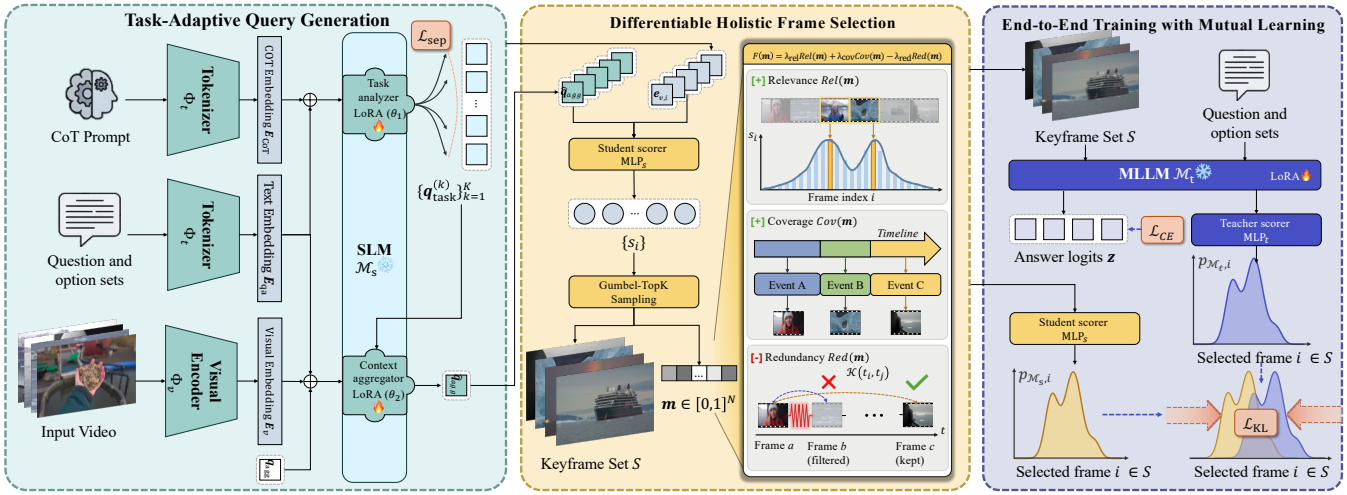

Figure 2: Overall architecture of the proposed HFS framework. 1) In the task-adaptive query generation stage, the student model performs CoT reasoning to generate task query vectors, which are aggregated with video and text features to form a context-aware query vector. 2) Differentiable holistic frame selection scores each frame and selects key frames via Gumbel-TopK sampling, using a set-level objective as regularization. 3) In the teacher-reasoning stage, the selected frames are fed into the teacher model, with the teacher and student aligning their frame-importance distributions through KL divergence.

formulate frame selection as a set-level optimization problem that can be jointly trained end-to-end with downstream tasks. This requires three components: (i) CoT-guided query understanding, which goes beyond shallow feature matching to capture task intent; (ii) a differentiable set-quality objective that evaluates the selected subset holistically; and (iii) online training signals from downstream tasks to guide and refine the frame selection process. Let the input video be $\mathcal{V} = \{\mathcal{V}_i\}_{i=1}^{N}$, where $\mathcal{V}_i$ is the i-th frame and N is the total number of frames. We obtain the video embedding matrix $\boldsymbol{E}_v$ using a pre-trained visual encoder Φ_v, followed by a linear projection layer $\boldsymbol{W}_v$, as follows:

$$\boldsymbol{E}_v = [\boldsymbol{e}_{v,1}, \ldots, \boldsymbol{e}_{v,N}]^\top \in \mathbb{R}^{N \times d}, \tag{1}$$

where $\boldsymbol{e}_{v,i} = \boldsymbol{W}_v \Phi_v(\mathcal{V}_i) \in \mathbb{R}^d$ is the d-dimensional embedding vector of the i-th frame. For the text input, let Q denote the question and $\{O_j\}$ denote the set of answer options. We use the language model's embedding layer Φ_t to transform the tokenized text into the text embedding matrix $\boldsymbol{E}_{\text{qa}}$, as follows:

$$\boldsymbol{E}_{\text{qa}} = \Phi_t([Q; \{O_j\}]) \in \mathbb{R}^{L \times d}, \tag{2}$$

where L is the length of the tokenized text sequence.

The student frame selector $\mathcal{M}_s$ is an SLM that takes the full video frame embedding matrix $\boldsymbol{E}_v$ and the question embedding $\boldsymbol{E}_{\text{qa}}$ as input. It analyzes the global visual-textual context and, through a lightweight student scorer MLP_s, predicts an importance score for each frame in the video, thereby identifying a highly representative key frame subset S.

This subset is then fed, together with the original question, into the teacher video reasoner $\mathcal{M}_t$. The teacher reasoner is an MLLM that performs the final reasoning task based solely on this compact and informative input. To enable mutual learning, $\mathcal{M}_t$ also maintains a lightweight teacher scorer MLP_t that produces its own frame-importance distribution from its internal representations, enabling bidirectional alignment between teacher and student during training.

3.2 Adaptive Query Generation

Some existing video frame selection methods [11] typically rely on a static, learnable query vector to aggregate spatio-temporal information. However, this approach is often inadequate for handling diverse video understanding tasks with varying semantic intents. To overcome this limitation, we design a two-stage adaptive query generation mechanism. This mechanism first decouples multidimensional task intentions from the input text, then deeply integrates these intentions with visual context to provide dynamic and precise guidance for subsequent frame selection.

CoT-guided query vector generation. The objective of this stage is to enable the student selector $\mathcal{M}_s$ to understand the intent behind textual queries. We utilize an SLM with a task analyzer LoRA adapter [10] to specifically process textual information, with parameters denoted as θ_1. To guide the student model $\mathcal{M}_s$ toward deep comprehension of query intent, we define a structured Chain-of-Thought (CoT) prompt [39], denoted as $\mathcal{P}_{\text{CoT}}$. This prompt instructs the model to perform logical analysis based on the question and its candidate options, and to identify key concepts that differentiate these options. As a result, the model is encouraged to generate a rich, step-by-step reasoning trace within its hidden states. The CoT prompt is fed into the text encoder Φ_t to obtain its embedding $\boldsymbol{E}_{\text{CoT}}$, as follows:

$$\boldsymbol{E}_{\text{CoT}} = \Phi_t(\mathcal{P}_{\text{CoT}}) \in \mathbb{R}^{L_p \times d}, \tag{3}$$

where L_p denotes the token length of $\mathcal{P}_{\text{CoT}}$. We then concatenate $\boldsymbol{E}_{\text{CoT}}$ with the question-answer embedding $\boldsymbol{E}_{\text{qa}}$ along the sequence

Table 1: Performance comparison of video understanding models across Video-MME, the test set of MLVU, and the validation set of LongVideoBench. † indicates results evaluated by us under the same experimental setting.

Model	LLM size	# Frames	Video-MME (w.o./w. sub.)				LongVideoBench	MLVU
			Short	Medium	Long	Overall		M-AVG
Avg. Duration			*1.3 min*	*9 min*	*41 min*	*17 min*	*12 min*	*12 min*
ShareGPT4Video [4]	8B	16	48.3/53.6	36.3/39.3	35.0/37.9	39.9/43.6	39.7	33.8
VideoLLaMA2 [6]	7B	8	-	-	-	45.1/46.6	-	45.6
Video-XL [29]	7B	128	64.0/67.4	53.2/60.7	49.2/54.9	55.5/61.0	50.7	45.5
VideoChat2-Mistral [17]	7B	16	48.3/52.8	37.0/39.4	33.2/39.2	39.5/43.8	39.3	-
Kangaroo [21]	8B	64	66.1/68.0	55.3/55.4	46.6/49.3	56.0/57.6	54.2	-
VILA-1.5 [19]	40B	14	-	-	-	-	-	44.2
LongVA [46]	7B	128/256	61.1/61.6	50.4/53.6	46.2/47.6	52.6/54.3	-	41.1
LLaVA-OneVision [14]	7B	32	-	-	-	58.2/61.5	56.3	-
Video-LLaVA [18]	7B	8	45.3/46.1	38.0/40.7	36.2/38.1	39.9/41.6	39.1	30.7
Chat-UniVi-v1.5 [12]	7B	64	45.7/51.2	40.3/44.6	35.8/41.8	40.6/45.9	-	-
Video-CCAM [7]	14B	96	62.2/66.0	50.6/56.3	46.7/49.9	53.2/57.4	-	42.9
Qwen2.5-VL [2]†	7B	16	63.8/68.7	50.3/55.2	45.1/51.3	53.1/58.4	54.5	41.8
+ TopK+Threshold†	7B	16	65.6/69.6	55.1/58.9	47.9/52.4	56.2/60.3	56.5	43.4
+ AKS [32]†	7B	16	65.9/68.7	56.1/59.1	48.7/53.1	56.9/60.3	57.2	43.6
+ Q-Frame [47]†	7B	16	62.8/65.8	54.9/58.4	49.1/53.0	55.6/59.1	57.2	39.6
+ MDP^3 [31]†	7B	16	65.6/69.9	56.1/59.1	49.9/52.9	57.2/60.6	57.0	44.4
+ **HFS**	7B+1.5B	16	**68.9/72.2**	**56.3/60.4**	**53.9/55.0**	**59.7/62.6**	**57.3**	**45.6**
InternVL3 [52]†	8B	16	71.1/75.3	58.8/62.1	52.2/53.8	60.7/63.7	56.7	46.0
+ TopK+Threshold†	8B	16	71.6/75.8	59.8/62.6	51.2/57.4	60.9/65.3	57.2	47.6
+ AKS [32]†	8B	16	70.1/74.0	59.9/61.3	51.6/56.9	60.5/64.1	58.5	48.0
+ Q-Frame [47]†	8B	16	69.0/73.9	59.3/61.8	51.1/58.0	59.8/64.6	57.1	48.6
+ MDP^3 [31]†	8B	16	71.6/76.0	**60.9/63.7**	52.0/57.2	61.5/65.6	59.3	48.0
+ **HFS**	8B+1.5B	16	**73.8/76.8**	59.8/63.0	**56.4/58.7**	**63.3/66.1**	**60.2**	**50.0**

dimension to form the joint input embedding $\boldsymbol{E}_{\text{input}}$, as follows:

$$\boldsymbol{E}_{\text{input}} = [\boldsymbol{E}_{\text{CoT}}; \boldsymbol{E}_{\text{qa}}] \in \mathbb{R}^{(L_p+L)\times d}. \tag{4}$$

This combined embedding $\boldsymbol{E}_{\text{input}}$ is then fed into the SLM $\mathcal{M}_s$, parameterized by θ_1. Under the guidance of the $\boldsymbol{E}_{\text{CoT}}$ component, $\mathcal{M}_s$ generates a final sequence of hidden states $\boldsymbol{H} \in \mathbb{R}^{(L_p+L)\times d}$, which encodes structured reasoning over the augmented text input. We then sample K vectors from the hidden state sequence $\boldsymbol{H}$ at uniformly distributed positions, forming a set of task-specific implicit query vectors $\{\boldsymbol{q}_{\text{task}}^{(k)}\}_{k=1}^{K}$:

$$\begin{aligned} \{\boldsymbol{q}_{\text{task}}^{(k)}\}_{k=1}^{K} &= \boldsymbol{H}[j_k]_{k=1}^{K}, \\ \text{where} \quad j_k &= \lfloor \frac{k-1}{K-1}(L_p + L - 1) \rfloor. \end{aligned} \tag{5}$$

Here, $\boldsymbol{q}_{\text{task}}^{(k)} \in \mathbb{R}^d$ denotes the k-th task query vector, and its index j_k is determined via linear interpolation over the sequence length $L_p + L$. This uniform sampling strategy serves as a simple yet effective heuristic for capturing information from different stages of the CoT-augmented reasoning trace. The resulting set of vectors is intended to capture diverse semantic aspects of the original text query, enabling more expressive and task-aware frame selection.

Query diversification via separation loss. To ensure these K query vectors capture task requirements from distinct angles rather than converging to similar representations, we introduce a separation loss $\mathcal{L}_{\text{sep}}$. This loss function aims to maximize the angular separation between query vectors by minimizing the sum of squared cosine similarities between pairs:

$$\mathcal{L}_{\text{sep}} = \frac{2}{K(K-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \left(\frac{\boldsymbol{q}_{\text{task}}^{(i)} \cdot \boldsymbol{q}_{\text{task}}^{(j)}}{\|\boldsymbol{q}_{\text{task}}^{(i)}\| \|\boldsymbol{q}_{\text{task}}^{(j)}\|} \right)^2. \tag{6}$$

Generation of aggregator query. After generating text-driven task queries, the next step is to fuse them with global visual information to form an aggregator query. We continue using the same SLM $\mathcal{M}_s$, but perform this multimodal fusion task via the context aggregator LoRA adapter, with parameters denoted as θ_2. Specifically, we construct a concatenated sequence $\boldsymbol{E}_{\text{fused}}$ consisting of the video frame embeddings $\boldsymbol{E}_v$, text embeddings $\boldsymbol{E}_{\text{qa}}$, the K task query vectors $\{\boldsymbol{q}_{\text{task}}^{(k)}\}_{k=1}^{K}$ generated in the first stage, and a learnable aggregator query token $\boldsymbol{q}_{\text{agg}} \in \mathbb{R}^d$:

$$\boldsymbol{E}_{\text{fused}} = [\boldsymbol{E}_v; \boldsymbol{E}_{\text{qa}}; \{\boldsymbol{q}_{\text{task}}^{(k)}\}_{k=1}^{K}; \boldsymbol{q}_{\text{agg}}] \in \mathbb{R}^{(N+L+K+1)\times d}. \tag{7}$$

In this extended sequence $\boldsymbol{E}_{\text{fused}}$, the aggregator query token $\boldsymbol{q}_{\text{agg}}$ is placed at the end to integrate information from all modalities and task queries through the self-attention mechanism of the SLM. The hidden state corresponding to this aggregated token in the final

Table 2: Performance comparison on the NExT-QA benchmark. † indicates results evaluated by us under the same experimental setting.

Model	LLM size	# Frames	NExT-QA
Avg. Duration			*44 sec*
LVNet [25]	7B	12	71.1
SlowFast-LLaVA [43]	7B	50	64.2
LLaVA-NeXT-Video [48]	7B	16	62.4
LLaVA-OneVision [14]	7B	32	79.4
Tarsier [35]	7B	8	71.6
NVILA [23]	8B	256	82.2
Video-XL [29]	7B	-	77.2
Oryx-1.5 [22]	7B	64/256	81.8
Qwen2-VL [36]	7B	-	77.6
Qwen2-VL [36] + M-LLM [11]	7B+1.5B	-	78.4
Qwen2.5-VL [2]†	7B	16	77.8
+ TopK+Threshold†	7B	16	78.6
+ AKS [32]†	7B	16	78.8
+ Q-Frame [47]†	7B	16	79.0
+ MDP^3 [31]†	7B	16	79.1
+ **HFS**	7B+1.5B	16	**79.4**
InternVL3 [52]†	8B	16	81.8
+ TopK+Threshold†	8B	16	80.7
+ AKS [32]†	8B	16	80.6
+ Q-Frame [47]†	8B	16	82.1
+ MDP^3 [31]†	8B	16	81.0
+ **HFS**	8B+1.5B	16	**83.1**

output layer of the SLM is taken as the fused query vector $\hat{\boldsymbol{q}}_{\text{agg}}$, as follows:

$$\hat{\boldsymbol{q}}_{\text{agg}} = \text{last}(\mathcal{M}_s(\boldsymbol{E}_{\text{fused}}; \theta_2)) \in \mathbb{R}^d. \tag{8}$$

To overcome the "myopic" nature of traditional top-K selection, which often results in selecting redundant frames, we reformulate frame selection as a set-level optimization problem. Instead of optimizing relevance scores for individual frames, our objective is to directly optimize the quality of the selected set S. Drawing inspiration from set-function optimization [9, 34, 40], which embodies the principle of diminishing returns, we design a differentiable, holistic objective function $F(\boldsymbol{m})$. This principle implies that once an event is adequately covered by frames in S, the marginal benefit of adding more highly similar frames diminishes sharply. Accordingly, our objective balances three complementary factors: maximizing task relevance, ensuring information coverage, and minimizing temporal redundancy.

Context-aware relevance scoring. We aim to generate a relevance score set $\{s_i\}_{i=1}^N$ for each frame, which serves as the basis for subsequent differentiable sampling. An accurate score must be context-dependent. To this end, we utilize the previously generated aggregator query vector $\hat{\boldsymbol{q}}_{\text{agg}}$, which encodes rich information about specific task intent and multimodal context. We concatenate $\hat{\boldsymbol{q}}_{\text{agg}}$ with each frame embedding $\boldsymbol{e}_{v,i} \in \mathbb{R}^d$ and feed the result into a lightweight student MLP scorer MLP_s. This scorer outputs an initial relevance score s_i for each frame:

$$s_i = \sigma(\text{MLP}_s([\boldsymbol{e}_{v,i}; \hat{\boldsymbol{q}}_{\text{agg}}])) \in [0,1], \quad \forall i \in \{1, \ldots, N\}, \tag{9}$$

where σ denotes the sigmoid function, used to constrain the score to the interval $[0, 1]$. The discrete operation of selecting the top-K frames from the score set $\{s_i\}$ is non-differentiable. To relax this process, we employ the Gumbel-TopK technique [13], which introduces Gumbel noise and generates a continuous, differentiable selection mask $\boldsymbol{m} = \{m_i\}_{i=1}^N \in [0,1]^N$ via the Softmax function, where m_i represents the "soft" probability of selecting the i-th frame:

$$\boldsymbol{m}, S = \text{Gumbel-TopK}(\{s_i\}_{i=1}^N, \tau, k_{\text{sel}}), \tag{10}$$

where k_{sel} is the number of frames to select, S is the index set of selected frames, and τ is the temperature parameter, which is gradually annealed during training.

Holistic set objective. We define a continuous objective function $F(\boldsymbol{m})$ to evaluate the quality of the soft-selected set based on three complementary criteria: Relevance (Rel), Coverage (Cov), and Redundancy (Red).

$\text{Rel}(\boldsymbol{m})$ measures the expected cumulative relevance score of the selected frames. It is computed by weighting the original relevance scores s_i with the soft mask $\boldsymbol{m}$:

$$\text{Rel}(\boldsymbol{m}) = \sum_{i=1}^N s_i \cdot m_i, \tag{11}$$

$\text{Cov}(\boldsymbol{m})$ encourages at least one selected frame to achieve a high relevance score, preventing the selection from distributing probability mass uniformly across low-relevance frames. We implement this term using the log-sum-exp function, which serves as a smooth approximation of the maximum, with temperature parameter τ_c:

$$\text{Cov}(\boldsymbol{m}) = \tau_c \log\left(\sum_{i=1}^N \exp\left(\frac{s_i \cdot m_i}{\tau_c}\right)\right). \tag{12}$$

To promote diversity, we introduce a temporal redundancy term $\text{Red}(\boldsymbol{m})$, which is minimized during optimization. We define a temporal similarity kernel function $\mathcal{K}(t_i, t_j)$ that assigns higher redundancy to temporally proximate frames:

$$\mathcal{K}(t_i, t_j) = \exp\left(-\frac{(t_i - t_j)^2}{2\gamma^2}\right), \tag{13}$$

$$\text{Red}(\boldsymbol{m}) = \sum_{i=1}^N \sum_{j=1, j\neq i}^N m_i \cdot m_j \cdot \mathcal{K}(t_i, t_j). \tag{14}$$

We combine these three components with weights to obtain the continuous set-level function $F(\boldsymbol{m})$, as follows:

$$F(\boldsymbol{m}) = \lambda_{\text{rel}} \cdot \text{Rel}(\boldsymbol{m}) + \lambda_{\text{cov}} \cdot \text{Cov}(\boldsymbol{m}) - \lambda_{\text{red}} \cdot \text{Red}(\boldsymbol{m}), \tag{15}$$

where λ_{rel}, λ_{cov}, and λ_{red} are hyperparameters that balance the contributions of relevance, coverage, and redundancy, respectively.

Notably, $F(\boldsymbol{m})$ acts as a differentiable regularizer during end-to-end training, rather than an independent objective requiring formal approximation guarantees at inference time. Unlike the predefined set-quality criteria used in the training-free method, the scores $\{s_i\}$ in $F(\boldsymbol{m})$ are generated by a learnable scorer, which receives downstream gradient signals. This allows the notions of "relevance" and "redundancy" to adapt dynamically to specific tasks. The total loss function $\mathcal{L}_{\text{total}}$ includes the term $-\lambda_{\text{set}} \cdot F(\boldsymbol{m})$. Minimizing $\mathcal{L}_{\text{total}}$ therefore encourages maximization of $F(\boldsymbol{m})$, leading to the selection of more informative and diverse frame subsets.

Table 3: Ablation study on the main components using Qwen2.5-VL-7B-Instruct as the teacher model. Highlighted rows indicate the configuration used in our final model.

Selection Method	CoT-Query	Set Objective	KL-Distill	$\mathcal{L}_{\text{sep}}$	MLVU	LongVideoBench	NExT-QA
Baseline	×	×	×	×	42.2	55.0	78.1
HFS w/o CoT-Query	×	✓	✓	×	43.4	55.3	78.5
HFS w/o Set Objective	✓	×	✓	✓	43.8	55.2	78.3
HFS w/o KL Distillation	✓	✓	×	✓	44.4	55.9	78.7
HFS w/o $\mathcal{L}_{\text{sep}}$	✓	✓	✓	×	44.8	56.1	78.9
HFS	✓	✓	✓	✓	**45.6**	**57.3**	**79.4**

Table 4: Ablation study on the number of generated queries K using InternVL3-8B as the teacher model.

Task analyzer LoRA	Context aggregator LoRA	K	Video-MME
×	✓	0	64.5
$\mathbf{e}_{\text{qa}} \rightarrow \mathbf{q}_{\text{task}}^{(1)}$	✓	1	65.2
$\mathbf{e}_{\text{qa}} \rightarrow \{\mathbf{q}_{\text{task}}^{(k)}\}_{k=1}^{2}$	✓	2	65.7
$\mathbf{e}_{\text{qa}} \rightarrow \{\mathbf{q}_{\text{task}}^{(k)}\}_{k=1}^{3}$	✓	3	**66.1**
$\mathbf{e}_{\text{qa}} \rightarrow \{\mathbf{q}_{\text{task}}^{(k)}\}_{k=1}^{4}$	✓	4	65.9

Table 5: Ablation study on the components of the set quality objective $F(\boldsymbol{m})$ using InternVL3-8B as the teacher model.

Set quality Objective			MLVU	LongVideoBench
Relevance	Coverage	Redundancy		
×	×	×	47.8	58.0
✓	×	×	48.4	58.5
✓	✓	×	49.4	59.3
✓	×	✓	49.0	59.1
✓	✓	✓	**50.0**	**60.2**

3.3 End-to-End Training with Mutual Learning

Training frame selectors using static, offline-generated MLLM pseudo-labels is suboptimal because such supervision cannot adapt to task-specific feedback. To address this limitation, we propose an end-to-end online distillation framework [50] in which both the student scorer MLP_s and the teacher scorer MLP_t are jointly optimized through a mutual alignment objective $\mathcal{L}_{\text{KL}}$.

Downstream task supervision. One objective of this framework is to maximize the accuracy of downstream tasks. In our teacher-student architecture, the MLLM $\mathcal{M}_t$ serves as the teacher. It receives the image data $\mathcal{V}_S = \{\mathcal{V}_i\}_{i \in S}$ corresponding to the key frame indices S selected by the student model $\mathcal{M}_s$, along with the original question text Q and the list of options $\{O_j\}$. The teacher $\mathcal{M}_t$ performs inference based on this sparse input and generates answer logits $\boldsymbol{z} \in \mathbb{R}^C$, where C is the number of candidate answers:

$$\boldsymbol{z} = \mathcal{M}_t(\mathcal{V}_S, Q, \{O_j\}). \tag{16}$$

We compute the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ as one of the objectives:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log(\text{Softmax}(\boldsymbol{z})_c), \tag{17}$$

where y_c is the ground-truth label. It forms a one-hot vector $\boldsymbol{y} \in \{0,1\}^C$, such that $y_c = 1$ when c is the index of the correct answer, and $y_c = 0$ otherwise. Simultaneously, we introduce the set-level objective $F(\boldsymbol{m})$ as a structured regularization term, which directly evaluates the intrinsic quality of the selected set S.

Mutual learning. To enable co-training, we extract frame-level representations from the second-to-last hidden layer of $\mathcal{M}_t$. Vision tokens corresponding to each frame are identified and average-pooled to produce vectors $\boldsymbol{h}_i$, forming $\boldsymbol{H}_{\mathcal{M}_t} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{k_{\text{sel}}}]^\top \in \mathbb{R}^{k_{\text{sel}} \times d_h}$, where d_h denotes the hidden dimension of the teacher model $\mathcal{M}_t$. We also extract the final text token's hidden state $\boldsymbol{h}_{\text{con}}$ as a global context summary. The teacher scorer MLP_t then generates its importance distribution:

$$\boldsymbol{p}_{\mathcal{M}_t} = \text{Softmax}(\text{MLP}_t([\boldsymbol{H}_{\mathcal{M}_t}; \boldsymbol{h}_{\text{con}}])) \in \mathbb{R}^{k_{\text{sel}}}. \tag{18}$$

Correspondingly, the student distribution $\boldsymbol{p}_{\mathcal{M}_s}$ is obtained by smoothing its own frame importance scores $\{s_i\}_{i \in S}$ using a distillation temperature τ_d:

$$\boldsymbol{p}_{\mathcal{M}_s} = \text{Softmax}(\{s_i/\tau_d\}_{i \in S}) \in \mathbb{R}^{k_{\text{sel}}}. \tag{19}$$

We employ the Kullback-Leibler divergence as our alignment loss $\mathcal{L}_{\text{KL}}$. Gradients from $\mathcal{L}_{\text{KL}}$ flow back to both the student scorer MLP_s and the teacher scorer MLP_t, forcing them to co-evolve:

$$\mathcal{L}_{\text{KL}} = \sum_{i=1}^{k_{\text{sel}}} \boldsymbol{p}_{\mathcal{M}_t,i} \log \frac{\boldsymbol{p}_{\mathcal{M}_t,i}}{\boldsymbol{p}_{\mathcal{M}_s,i}}. \tag{20}$$

Overall objective. Our final training objective $\mathcal{L}_{\text{total}}$ is a multi-task loss function that integrates all the above loss components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} - \lambda_{\text{set}} \cdot F(\boldsymbol{m}) + \lambda_{\text{sep}} \cdot \mathcal{L}_{\text{sep}}, \tag{21}$$

where λ_{KL}, λ_{set}, and λ_{sep} are hyperparameters controlling the contribution of each term.

4 Experiments

4.1 Experimental Setup

Datasets. We trained our model using data from two video instruction datasets: 1) 250K annotated samples from VideoChat2-IT [17], covering multiple task types including video description, question-answering, and reasoning. 2) 196K samples from LLaVA-Video-178K [49].

Table 6: Ablation study on the supervision strategy using InternVL3-8B as the teacher model.

Pseudo-label	$\mathcal{L}_{CE}$	$\mathcal{L}_{KL}$	NExT-QA	Video-MME
✓	×	×	82.1	64.4
×	✓	×	82.8	65.3
×	✓	✓	**83.1**	**66.1**

Table 7: Ablation study on the set-level optimization weight configuration using Qwen2.5-VL-7B-Instruct as the teacher model.

λ_{rel}	λ_{cov}	λ_{red}	Video-MME			
			Short	Medium	Long	Overall
0.4	0.4	0.2	72.0	59.9	54.1	62.0
0.5	0.2	0.3	71.9	60.1	54.3	62.1
0.5	0.3	0.2	**72.2**	**60.4**	**55.0**	**62.6**
0.3	0.3	0.4	71.3	59.0	53.2	61.2

We evaluated performance across four benchmarks: 1) Video-MME [8], which is subdivided into short-, medium-, and long-duration subset, and evaluated under both uncaptioned and captioned configurations; 2) the validation set of LongVideoBench [41], featuring an average video length of 12 minutes and containing reasoning tasks tailored for extended videos; 3) the test set of MLVU [51], which encompasses topic reasoning, anomaly detection, and needle QA tasks, with an average video length of 12 minutes; and 4) the test set of NExT-QA [42], with an average video length of 44 seconds.

Baselines. We compare HFS with other frame selection methods, and all methods are evaluated under the same experimental settings. (1) **Uniform sampling**, which samples k_{sel} frames evenly at fixed intervals. (2) **TopK+Threshold**, which uses CLIP to calculate the cosine similarity between each frame and the query, filter out frames below a fixed threshold of 0.85, and then select the top-k_{sel} frames from the remaining candidates. (3) **AKS** [32], (4) **Q-Frame** [47], (5) **MDP**^{3} [31], which are training-free frame selection methods and are directly applied at inference time.

Implementation details. We employ Qwen2.5-1.5B-Instruct [33] as the student frame selector $\mathcal{M}_s$, and Qwen2.5-VL-7B-Instruct [2] together with InternVL3-8B-hf [52] as the teacher video reasoner $\mathcal{M}_t$. For video encoding, we use CLIP ViT-Base-Patch32 [26] to extract visual features from input frames, keeping the visual encoder frozen throughout training. The linear projection layer $\boldsymbol{W}_v$ is jointly trained with the other learnable components. The initial video is uniformly sampled at a resolution of 224×224 to yield $N = 128$ frames, from which the selector identifies $k_{sel} = 16$ key frames.

We perform parameter-efficient fine-tuning using LoRA [10] on three distinct adapters. In the student frame selector $\mathcal{M}_s$, the adapter for CoT-guided query generation (parameterized as θ_1), which generates $K = 3$ task-specific query vectors, employs rank $r_1 = 16$, while the adapter for context aggregation (parameterized as θ_2) adopts rank $r_2 = 16$. In the teacher video reasoner $\mathcal{M}_t$, the adapter for answer generation employs rank $r_3 = 8$. We employ AdamW [24] to optimize all trainable parameters, with a learning rate of 1×10^{-5} and a weight decay of 0.01. The effective batch size is 16.

The continuous set-level objective function balances three terms with weights: $\lambda_{rel} = 0.5$, $\lambda_{cov} = 0.3$, and $\lambda_{red} = 0.2$. For Gumbel-TopK sampling, we initialize the temperature $\tau = 2.0$ and apply exponential decay with a decay factor of 0.999 per step, reaching a minimum value of $\tau_{min} = 0.5$. The smoothing parameter τ_c for the $\mathrm{Cov}(\boldsymbol{m})$ is set to 2.0. The temporal similarity kernel employs a bandwidth of $\gamma = 10.0$ for redundancy calculation. The KL divergence loss weight λ_{KL} linearly increases from 0.1 to 1.0 during the first epoch, serving as a warm-up phase, prioritizing task accuracy while maintaining teacher-student alignment. The distillation temperature τ_d is set to 0.5. The set-level optimization regularization loss weight is set to $\lambda_{set} = 1 \times 10^{-4}$, and the query separation loss weight is set to $\lambda_{sep} = 0.01$.

4.2 Main Results

Table 1 demonstrates the performance of HFS on Video-MME [8], MLVU [51], and LongVideoBench [41]. The upper section lists representative video understanding models for reference, while the lower section compares different frame selection methods under identical experimental settings. On the Video-MME benchmark, HFS exhibits stable gains across short, medium, and long video subsets. Notably, InternVL3 [52] with HFS achieves the best overall performance and leads on both short and long subsets, validating HFS's generalizability and robustness. On MLVU, integrating HFS with Qwen2.5-VL [2] and InternVL3 [52] improves M-AVG scores by 3.8 and 4.0 percentage points, respectively. On LongVideoBench [41], combining the InternVL3 [52] with HFS achieves a score of 60.2%, significantly outperforming all baseline models.

Table 2 further evaluates performance on the NExT-QA benchmark [42], which emphasizes fine-grained temporal and causal reasoning, thereby placing stringent demands on the quality of frame selection. Our InternVL3 [52] combined with HFS achieves an accuracy of 83.1%, surpassing all state-of-the-art methods, including NVILA [23] and Oryx-1.5 [22].

4.3 Ablation Studies

As shown in Table 3, we validated the effectiveness of the four primary components in HFS. The baseline is defined as a model trained end-to-end using only $\mathcal{L}_{CE}$, employing static queries and relying on top-k selections from MLP_s. HFS, which integrates all components, consistently outperforms the baseline across all three benchmarks, demonstrating the overall efficacy of the proposed HFS framework.

Impact of CoT-guided query vectors. We further examine the impact of varying the number of query vectors K in Table 4, using InternVL3-8B [52] as the teacher model for this experiment. When $K = 0$, the model degenerates to using static queries, yielding the lowest performance. As K increases from 1 to 3, performance on Video-MME steadily improves, peaking at 66.1% when $K = 3$. However, performance slightly declined to 65.9% at $K = 4$, likely due to the introduction of redundant information. Based on these results, we set $K = 3$ for all other experiments.

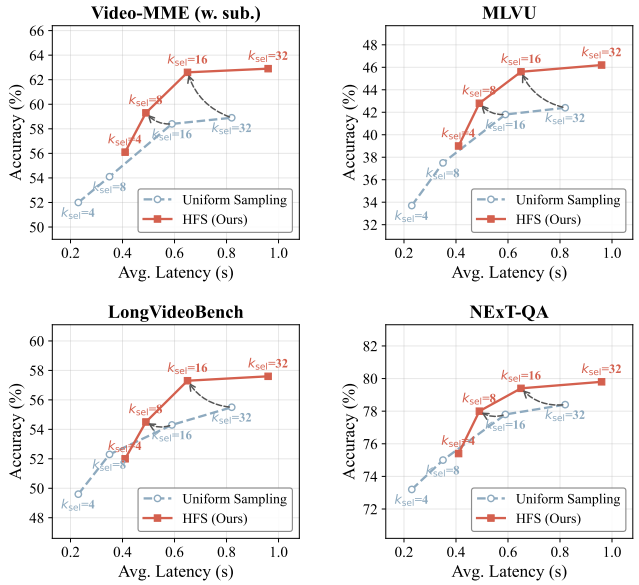

Figure 3: Accuracy-latency comparison of uniform sampling and HFS under different k_{sel} settings (teacher model: Qwen2.5-VL-7B-Instruct). The dotted arrow indicates that HFS with a smaller k_{sel} can exceed the accuracy of uniform sampling with a larger k_{sel}, while having lower inference latency.

Set-level objective analysis. As shown in Table 5, the baseline without $F(\boldsymbol{m})$ achieves 47.8% on MLVU [51]. Adding the relevance term $\text{Rel}(\boldsymbol{m})$ alone yields a 0.6 percentage point improvement. Incorporating the coverage term $\text{Cov}(\boldsymbol{m})$ on top of $\text{Rel}(\boldsymbol{m})$ delivers the greatest marginal gain, indicating that incentivizing coverage is crucial for avoiding information bottlenecks. Simultaneously introducing the redundancy term $\text{Red}(\boldsymbol{m})$ further improves performance to 50.0%. Table 7 confirms that our weight configuration ($\lambda_{\text{rel}} = 0.5$, $\lambda_{\text{cov}} = 0.3$, $\lambda_{\text{red}} = 0.2$) achieves the best overall performance on Video-MME. [8].

Supervision strategy comparison. Table 6 compares three different supervision strategies: 1) training with static pseudo-labels generated offline by using Qwen2.5-VL-7B-Instruct [2]; 2) end-to-end training using only cross-entropy loss from the downstream task; and 3) our complete end-to-end online distillation framework. Training with $\mathcal{L}_{\text{CE}}$ but without the $\mathcal{L}_{\text{KL}}$ distillation loss already outperforms using static pseudo-label supervision. By introducing the distillation loss $\mathcal{L}_{\text{KL}}$, our approach achieves state-of-the-art performance on both NExT-QA [42] and Video-MME [8].

Efficiency analysis. Figure 3 compares the performance and latency of HFS versus uniform sampling under different frame selection counts k_{sel}. Latency was measured using a batch size of 1 on a single NVIDIA H100 GPU to reflect real-world single-sample inference speeds, and includes the student selector overhead. HFS significantly outperforms uniform sampling across all k_{sel} settings. Notably, HFS with $k_{\text{sel}} = 8$ surpasses the accuracy of uniform sampling with $k_{\text{sel}} = 16$, while reducing inference latency from 0.59s to 0.49s per sample.

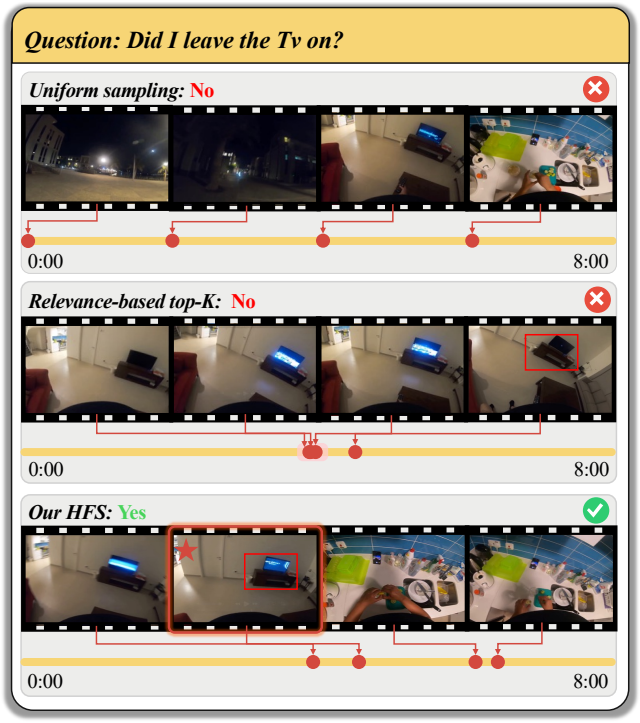

Figure 4: Qualitative comparison on an egocentric reasoning task. Uniform sampling selects frames from outdoor and kitchen scenes that are unrelated to the "TV" target. Relevance-based top-K sampling successfully localizes the "TV" object; however, the final selected frame depicts the "TV off" state. In contrast, our HFS also localizes the "TV" object, but its selected frames consistently capture the "TV on" state.

4.4 Qualitative Analysis

We provide a qualitative analysis in Figure 4. The task involves identifying the running status of a television (TV) in a long egocentric video. Uniform sampling completely fails to capture frames containing the television. Relevance-based top-K sampling locates the TV object but shows it in the 'off' state, as shallow visual similarity cannot distinguish task-relevant state changes. In contrast, HFS leverages CoT-guided queries to infer the need to identify the operational status and selects a frame that correctly shows the "TV on" state.

5 Conclusion

This paper proposes an end-to-end, task-adaptive framework for video frame selection, where a CoT-guided query generation mechanism adapts the selection focus to diverse task intents. The set-level optimization objective jointly enhances relevance, coverage, while suppressing redundancy. Moreover, the mutual learning mechanism between the student selector and the teacher reasoner eliminates the reliance on offline pseudo-labels. Extensive experiments across multiple benchmarks demonstrate that our method consistently outperforms existing approaches.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 23716–23736. https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923

[3] Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2025. Flexible Frame Selection for Efficient Video Reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 29071–29082.

[4] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. 2025. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*.

[5] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. 2025. LongVILA: Scaling Long-Context Visual Language Models for Long Videos. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

[6] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476* (2024).

[7] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. 2024. Video-CCAM: Enhancing Video-Language Understanding with Causal Cross-Attention Masks for Short and Long Videos. arXiv:2408.14023

[8] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. 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 (CVPR)*. 24108–24118.

[9] Michael Gygli, Helmut Grabner, and Luc Van Gool. 2015. Video summarization by learning submodular mixtures of objectives. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3090–3098. doi:10.1109/CVPR.2015.7298928

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[11] Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. 2025. M-LLM Based Video Frame Selection for Efficient Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13702–13712.

[12] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13700–13710.

[13] Wouter Kool, Herke van Hoof, and Max Welling. 2020. Ancestral Gumbel-Top-k Sampling for Sampling Without Replacement. *Journal of Machine Learning Research* 21, 47 (2020), 1–36.

[14] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research* (2024).

[15] Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. 2025. LLaVA-ST: A Multimodal Large Language Model for Fine-Grained Spatial-Temporal Understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8592–8603.

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. JMLR.org, Article 814, 13 pages.

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

[18] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 5971–5984.

[19] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: On Pre-training for Visual Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26679–26689.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *NeurIPS*.

[21] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024. Kangaroo: A Powerful Video-Language Model Supporting Long-Context Video Input. *arXiv preprint arXiv:2408.15542* (2024).

[22] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2025. Oryx MLLM: On-Demand Spatial-Temporal Understanding at Arbitrary Resolution. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

[23] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, et al. 2024. NVILA: Efficient Frontier Visual Language Models. arXiv:2412.04468

[24] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

[25] Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryoo, Donghyun Kim, and Michael S. Ryoo. 2024. Too Many Frames, Not All Useful: Efficient Strategies for Long-Form Video QA.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 8748–8763.

[27] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14313–14323.

[28] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. 2025. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. In *Forty-second International Conference on Machine Learning*.

[29] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26160–26169.

[30] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18221–18232.

[31] Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. 2025. MDP3: A Training-free Approach for List-wise Frame Selection in Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 24090–24101.

[32] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. 2025. Adaptive Keyframe Sampling for Long Video Understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 29118–29128.

[33] Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. https://qwenlm.github.io/blog/qwen2.5/.

[34] Sebastian Tschiatschek, Aytunc Sahin, and Andreas Krause. 2018. Differentiable Submodular Maximization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. 2731–2738.

[35] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. 2024. Tarsier: Recipes for Training and Evaluating Large Video Description Models. arXiv:2407.00634

[36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, et al. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191

[37] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXV*. 396–416.

[38] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025. VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3272–3282.

[39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought

Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[40] Kai Wei, Rishabh Iyer, and Jeff Bilmes. 2015. Submodularity in Data Subset Selection and Active Learning. In *Proceedings of the 32nd International Conference on Machine Learning*. 1954–1963.

[41] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. LongVideoBench: A Benchmark for Long-Context Interleaved Video-Language Understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[42] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9777–9786.

[43] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. 2024. SlowFast-LLaVA: A Strong Training-Free Baseline for Video Large Language Models. *arXiv:2407.15841* (2024).

[44] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Liufang Zhang, and Qianru Sun. 2025. Frame-Voyager: Learning to Query Frames for Video Large Language Models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

[45] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. 2025. Videorefer Suite: Advancing Spatial-Temporal Object Understanding with Video LLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18970–18980.

[46] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024. Long Context Transfer from Language to Vision. arXiv:2406.16852

[47] Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. 2025. Q-Frame: Query-aware Frame Selection and Multi-Resolution Adaptation for Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22056–22065.

[48] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

[49] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-Video: Video Instruction Tuning With Synthetic Data. *Transactions on Machine Learning Research* (2025).

[50] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep Mutual Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[51] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. MLVU: Benchmarking Multi-task Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13691–13701.

[52] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479