This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Wavelet-based Frame Selection by Detecting Semantic Boundary for Long Video Understanding

Wang Chen[1], Yuhui Zeng[1], Yongdong Luo[1], Tianyu Xie[1]
Luojun Lin[2], Jiayi Ji[1], Yan Zhang[1], Xiawu Zheng[1*]
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University
[2]College of Computer and Data Science, Fuzhou University

Abstract

*Frame selection is crucial due to high frame redundancy and limited context windows when applying Large Vision-Language Models (LVLMs) to long videos. Current methods typically select frames with high relevance to a given query, resulting in a disjointed set of frames that disregard the narrative structure of videos. In this paper, we introduce **W**avelet-based **F**rame **S**election by Detecting **S**emantic **B**oundary (**WFS-SB**), a training-free framework that presents a new perspective: effective video understanding hinges not only on high relevance but, more importantly, on capturing semantic shifts—pivotal moments of narrative change essential to comprehending the holistic storyline of videos. However, direct detection of abrupt changes in the query-frame similarity signal is often unreliable due to high-frequency noise arising from model uncertainty and transient visual variations. To address this, we leverage the wavelet transform, which provides an ideal solution through its multi-resolution analysis in both time and frequency domains. By applying this transform, we decompose the noisy signal into multiple scales and extract a clean semantic change signal from the coarsest scale. We identify the local extrema of this signal as semantic boundaries, which segment the video into coherent clips. Building on this, WFS-SB comprises a two-stage strategy: first, adaptively allocating a frame budget to each clip based on a composite importance score; and second, within each clip, employing the Maximal Marginal Relevance approach to select a diverse yet relevant set of frames. Extensive experiments show that WFS-SB significantly boosts LVLM performance, e.g., improving accuracy by **5.5% on VideoMME, 9.5% on MLVU, and 6.2% on LongVideoBench**, consistently outperforming state-of-the-art methods. Our code is available at https://github.com/MAC-AutoML/WFS-SB.*

*Corresponding author: zhengxiawu@xmu.edu.cn.

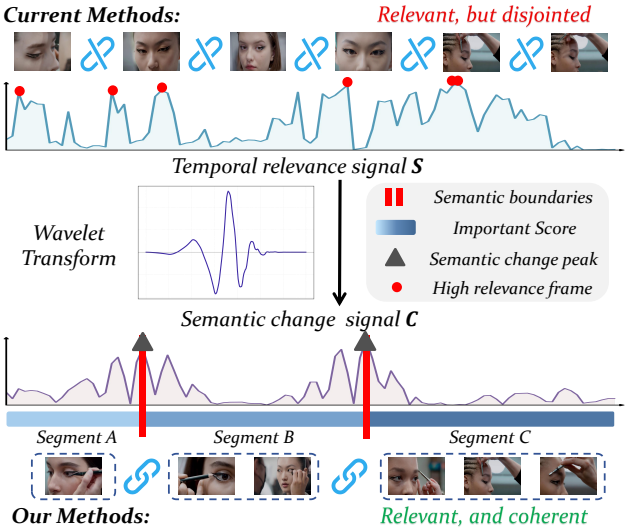


Figure 1. A comparison of frame selection strategies, illustrated with the query: "What is the process of applying makeup around the eyes in the video?". (Top) Current approach: Selects scattered, high-relevance frames (e.g., any frame with eyes), which fails to preserve the procedural order. (Bottom) Our approach: Employs a wavelet-based method to first segment the video into semantically coherent segments (e.g., applying eyeliner, shaping eyebrows) and then samples frames from each segment. This preserves the semantic integrity essential for process comprehension.

1. Introduction

The remarkable success of Large Language Models (LLMs) [1, 34] has catalyzed the development of powerful Large Vision-Language Models (LVLMs) capable of sophisticated multimodal reasoning [16, 35, 48], with recent extensions into the video domain [15, 22]. While demonstrating impressive performance on short clips, applying these models to long-form videos confronts a fundamental bottleneck: the massive number of frames clashes with limited computational budgets and fixed-size context windows.

Directly processing every frame is infeasible, necessitating a strategy to condense the visual stream. Consequently, identifying a concise subset of representative frames, known as keyframe selection, has become a critical prerequisite for deploying LVLMs on real-world, long-form videos. While alternatives like extending context windows [17, 44] or video-to-text summarization [5, 21, 24] exist, they often incur prohibitive computational overhead or risk significant information loss. Frame selection, therefore, presents a more direct and balanced approach, but it begs the question: what constitutes an optimal selection?

Existing methods typically select frames with the highest relevance scores to a given query. This approach, however, disregards the narrative structure of the video, akin to understanding a book's plot by selecting its most beautifully illustrated pages. The resulting frame collection, though individually relevant, is disjointed and fails to capture causality, development, or process. For instance, as illustrated in Figure 1, existing methods often retrieve scattered frames showing eyes but fail to reflect the sequential procedure. In this paper, we introduce **W**avelet-based **F**rame **S**election by Detecting **S**emantic **B**oundary (**WFS-SB**), a training-free framework that presents a new perspective: effective video understanding should not only focus on "which frames have high relevance," but rather on "when do the story's chapters change?". These chapter change moments manifest as abrupt shifts in the query-frame similarity signal (e.g., transitioning from "cooking" to "eating"). However, this raw signal is often fraught with high-frequency noise from model uncertainty or incidental visual variations, making the direct detection of these boundaries unreliable.

Confronted with this challenge, we reframe boundary detection as a signal processing problem. The wavelet transform emerges as an ideal tool for analyzing this non-stationary and noisy relevance signal. Its capacity for multi-resolution analysis in both time and frequency domains allows us to effectively distinguish meaningful, macro-level semantic shifts from spurious, high-frequency noise. By applying a multi-level wavelet decomposition, we extract a clean semantic change signal from the coarsest scale, whose local extrema precisely identify the temporal semantic boundaries. These boundaries effectively segment the video into coherent sections, akin to chapters in a story.

Based on this segmentation, WFS-SB implements a two-stage selection strategy: first, adaptively allocating a frame budget to each clip based on a composite importance score; and second, within each clip, employing the Maximal Marginal Relevance approach to select a diverse yet relevant set of frames. Extensive experiments demonstrate that WFS-SB significantly boosts LVLM performance across multiple benchmarks. For instance, when applied to LLaVA-Video-7B with 8 input frames, it improves accuracy by **5.5% on VideoMME, 9.5% on MLVU, and 6.2% on LongVideoBench**, consistently outperforming state-of-the-art methods.

Our contributions are threefold:

- We introduce a new perspective for frame selection that prioritizes capturing semantic shifts over merely high-relevance. We formulate it as a signal processing problem and employ wavelet transform to robustly detect semantic boundaries from noisy relevance signals.
- We propose WFS-SB, a training-free framework that advances frame selection from isolated relevance sampling to a semantic structure-aware process, incorporating semantic segmentation, adaptive budget allocation, and intra-clip diversification.
- We conduct extensive experiments demonstrating that WFS-SB achieves state-of-the-art performance on multiple long-video benchmarks, validating the superiority of our semantic-shift-first approach.

2. Related Work

Efficient Architectures and Strategies for Long Video Understanding. The evolution from early CNN-RNN models to contemporary Large Vision-Language Models (LVLMs) [2, 16, 48] reflects a persistent drive to process increasingly complex visual data. To specifically address long videos, researchers have pursued two primary avenues. The first centers on architectural innovations, such as hierarchical attention or memory mechanisms that expand model context windows [17, 28, 44]. The second involves strategic abstraction, including converting video into text summaries [5, 20, 21, 36] or pruning and merging visual tokens [14, 19, 27, 30]. While powerful, these approaches often introduce significant computational overhead or risk discarding critical visual details during abstraction. This positions lightweight pre-processing, particularly frame selection, as a vital and pragmatic complementary strategy.

Applications of the Wavelet Transform. The wavelet transform's capacity for localized time-frequency analysis has made it indispensable across numerous disciplines. It is a cornerstone of signal and image processing for tasks like denoising and compression, as seen in the JPEG 2000 standard [9, 33, 42]. Its multi-scale nature is also well-suited for analyzing complex physical phenomena, such as turbulence and seismic signals [7]. Recently, wavelets have been integrated into machine learning, serving as powerful feature extractors and enhancing deep learning architectures [29]. In this work, we creatively adapt the wavelet transform to a new domain: identifying semantic boundaries within a query-relevance signal for video frame selection.

Frame Selection for Long Video Understanding. Frame selection methods offer a practical solution to the constraints of fixed context windows in LVLMs. While early techniques relied on low-level, query-agnostic cues like shot boundaries [23, 26], modern approaches are more so-

phisticated. We broadly divide them into training-based and training-free methods. Training-based approaches learn a selection policy but can be data-hungry and computationally expensive [3, 11, 40]. Consequently, training-free methods have gained traction for their flexibility. Prominent strategies include inverse transform sampling (BOLT [18]), Markov decision processes (MDP3 [31]), joint relevance-coverage optimization (AKS [32]), subgraph selection (KFC [6]), dynamic resolution processing (Q-frame [45], VQOS [38]), temporal object-centric search (T* [39]), and iterative reasoning (A.I.R. [50]). However, most existing methods still focus on sampling individually relevant frames, often failing to capture the video's overarching semantic structure. Our work, WFS-SB, addresses this gap by first employing a wavelet-driven, training-free approach to detect semantic boundaries, thereby preserving the video's structural integrity before selection.

3. Method

3.1. Problem Formulation

Given a video $\mathcal{V}$ consisting of T frames and a user query q, the task of frame selection is to extract a concise yet representative frame subset $\mathcal{F} = \{f_1, f_2, \ldots, f_K\}$, where the indices are chronologically ordered and $K \ll T$. The value of K is generally determined in advance, considering both the context length constraints of large vision-language models (LVLMs) and the specific demands of downstream applications or user preferences. The ultimate objective is to choose $\mathcal{F}$ such that, when it serves as visual context together with the query q, the LVLM can produce the most accurate and informative response possible.

To quantify the semantic relevance of each frame to the query, we first uniformly subsample the video at 1 frame-per-second (fps), yielding a sequence of N frames. For each sampled frame f_t, we compute an Image-Text Matching (ITM) score using a pretrained vision-language model:

$$s_t = \mathcal{M}(q, f_t), \quad t = 1, \ldots, N, \tag{1}$$

where $\mathcal{M}$ denotes the ITM head of BLIP-2 [13]. Notably, we only use its ITM head for score computation, without involving the large-scale decoder part. These scores form the foundation for our subsequent analysis.

3.2. Modeling Temporal Relevance Signal

Rather than treating the ITM scores $\{s_t\}_{t=1}^{N}$ as isolated relevance measurements, we adopt a signal processing perspective: these scores collectively form a "temporal relevance signal"—a continuous waveform that encodes how the video's semantic alignment with the query evolves over time. This reframing enables us to leverage powerful tools from signal analysis to identify structural transitions in video content. However, this temporal relevance signal exhibits three challenging characteristics:

Non-stationarity. This is the most critical feature of the signal. The dynamic nature of video content dictates that the signal's statistical properties change drastically over time. A semantic segment highly relevant to the query (e.g., "a person hugging another") might occur within a few seconds, causing a sharp local fluctuation in the signal. In contrast, during irrelevant segments, the signal may remain stable at a low level. Traditional global analysis tools like the Fourier transform are unsuitable as they cannot capture the temporal location of these changes.

Multi-scale structure. Semantically coherent segments in a video naturally possess different temporal scales. A rapid action (e.g., "waving a hand") might span only a few frames, manifesting as a sharp, transient pulse in the signal. Conversely, a prolonged process (e.g., "the entire cooking procedure") could extend over hundreds of frames, appearing as a slow, broad peak. We require an analytical tool that can resolve both macroscopic trends and microscopic details simultaneously.

Low signal-to-noise ratio. The raw relevance signal is inundated with substantial noise. This noise originates from several sources: 1) Model uncertainty: Minor perturbations in the internal representations of a VLM for visually similar yet semantically indistinct frames can lead to score jitter. 2) Cross-modal ambiguity: The semantic alignment between a textual query and visual content is inherently ambiguous. 3) Visual artifacts: Factors unrelated to the query's semantics, such as changes in lighting, object occlusions, or camera motion, can cause unintended score fluctuations. This high-frequency noise severely interferes with the identification of true semantic boundaries, making methods based on simple amplitude thresholds or naive gradient detection highly prone to failure.

These three properties collectively motivate our choice of wavelet transform. Wavelets provide a mathematically principled framework that simultaneously addresses all three challenges through multi-resolution decomposition in both the time and frequency domains.

3.3. Wavelet-based Frame Selection

As illustrated in Figure 2, the method integrates three core components in a hierarchical pipeline: 1) Wavelet-Based Semantic Boundary identification—partitioning the video into temporally coherent segments by identifying semantic boundaries through wavelet transform. 2) Adaptive Budget Allocation—computing segment importance scores and distributing frame budget non-uniformly across segments. 3) Diversity-Aware Intra-segment Frame Selection—applying diversity-driven selection within each segment's allocated budget. This hierarchical design preserves narrative structure and ensures local detail and global coherence.

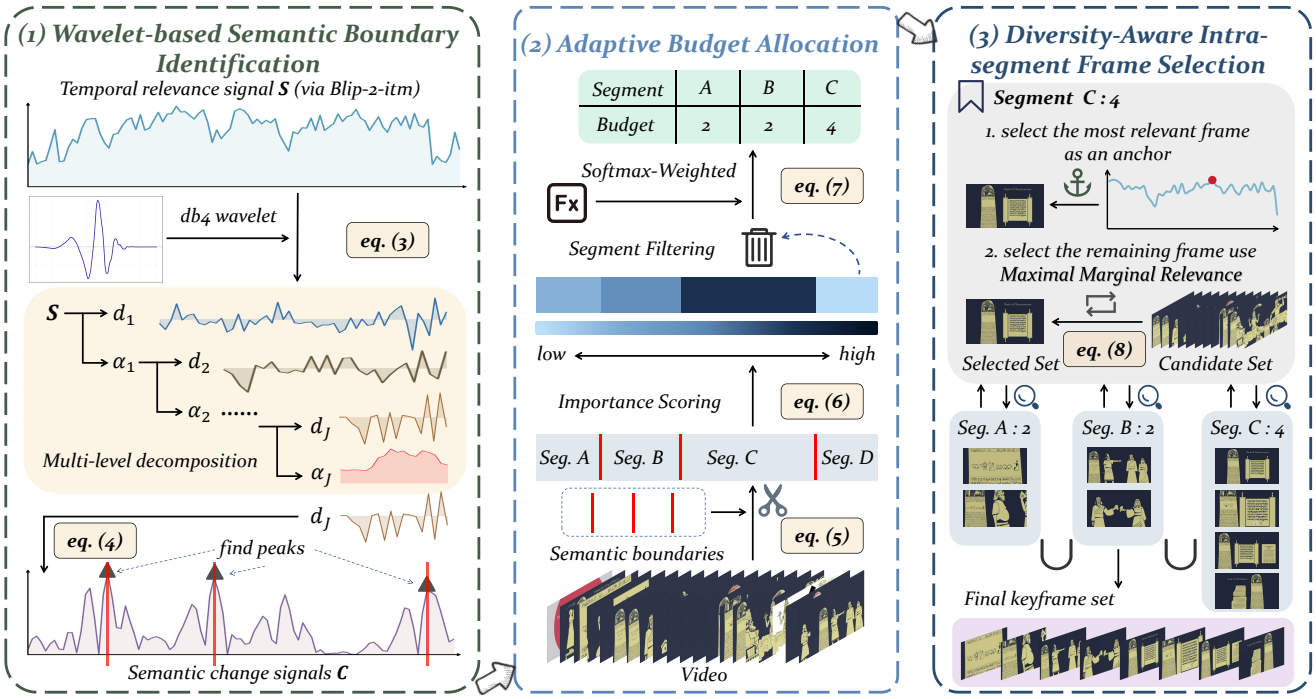


Figure 2. An overview of our proposed WFS-SB framework. The process unfolds in three main stages: (1) Wavelet-based Semantic Boundary Identification: The raw query-frame relevance signal is decomposed using a multi-level Discrete Wavelet Transform. We isolate and reconstruct the coarsest detail coefficients to generate a robust semantic change signal, whose peaks define the boundaries of coherent semantic segments. (2) Adaptive Budget Allocation: A composite importance score is computed for each segment, which guides a softmax-weighted distribution of the total frame budget $K = 8$. (3) Diversity-Aware Intra-segment Selection: Finally, a localized Maximal Marginal Relevance (MMR) selection is performed within each segment to ensure both relevance and diversity in the final keyframe set.

3.3.1. Wavelet-based Semantic Boundary Identification

To address the challenges outlined in Section 3.2, we employ the Discrete Wavelet Transform (DWT). Its multi-resolution analysis is uniquely suited to decompose the similarity score sequence, allowing us to isolate large-scale semantic transitions from transient noise and accurately identify semantic boundaries.

Adaptive Multi-level Decomposition. We apply the DWT to the ITM score sequence $\{s_t\}_{t=1}^{N}$ using the Daubechies-4 (db4) wavelet, which offers a favorable trade-off between smoothness and compact support. The DWT recursively filters the signal, splitting it at each level j into low-frequency approximation coefficients a_j and high-frequency detail coefficients d_j. The key parameter is the decomposition level J, which we set adaptively based on the sequence length N:

$$J = \max\left(1, \lfloor \log_2 N \rfloor - l\right), \tag{2}$$

where l is a drift factor controlling the decomposition depth, set to 3. This adaptive strategy ensures the analysis scale is appropriate for the signal's duration; it uses a larger J for longer videos to focus on coarse, stable components and a smaller J for shorter videos to retain crucial temporal details. The final DWT output is a multi-scale representation of the original signal:

$$\text{DWT}(s_t) = \{a_J, d_J, d_{J-1}, \ldots, d_1\}. \tag{3}$$

Semantic Change Feature Extraction. Semantic boundaries are characterized by significant, large-scale shifts in the ITM score, rather than by high-frequency fluctuations. These slow-varying but substantial changes are captured by the detail coefficients at the coarsest scale, d_J. To isolate this specific feature, we perform a reconstruction using the Inverse Discrete Wavelet Transform (IDWT), where only the d_J coefficients are retained and all others are zeroed out:

$$\tilde{s}_t = \text{IDWT}(\{\mathbf{0}, d_J, \mathbf{0}, \ldots, \mathbf{0}\}). \tag{4}$$

The resulting signal, $\{\tilde{s}_t\}$, represents the component of the ITM score that reflects coarse-grained temporal variations, effectively acting as a robust indicator of semantic change while naturally suppressing high-frequency noise concentrated in finer-scale coefficients (e.g., d_1, d_2).

Semantic Boundary Detection. The moments of most significant semantic change correspond to local extrema in the magnitude of the change indicator signal, $\tilde{s}_t$. We therefore define a change intensity signal as $c_t = |\tilde{s}_t|$ and apply a

peak detection algorithm to it. This process identifies the temporal indices of significant boundaries while filtering out spurious peaks from residual noise using adaptive height and prominence thresholds. Let $\mathcal{B} = \{b_1, \ldots, b_M\}$ be the set of detected boundary indices. These boundaries partition the video into $M+1$ Temporal Semantic Segments:

$$\mathcal{G} = \{[1, b_1], [b_1 + 1, b_2], \ldots, [b_M + 1, N]\}, \quad (5)$$

where each segment represents a temporally coherent unit with stable relevance to the query. These segments form the structural basis for subsequent frame selection.

3.3.2. Adaptive Budget Allocation

Once the temporal semantic segments are established, we allocate the total frame budget K according to the semantic importance of each segment. This adaptive strategy ensures that semantically critical segments receive dense coverage, optimizing the trade-off between global video understanding and fine-grained local detail.

Segment Importance Score. We compute a composite importance score, $\text{Imp}(\mathcal{G}_i)$, for each temporal semantic segment $\mathcal{G}_i$ by integrating multiple relevance cues:

$$\text{Imp}(\mathcal{G}_i) = w_d \cdot \frac{|\mathcal{G}_i|}{N} + w_a \cdot \bar{s}_i + w_m \cdot s_i^{\max} + w_v \cdot \frac{\sigma_i^2}{\sigma_{\text{global}}^2}, \quad (6)$$

where $|\mathcal{G}_i|/N$ represents the normalized segment duration, $\bar{s}_i$ is the mean relevance score, $s_i^{\max}$ is the max relevance score, and $\sigma_i^2/\sigma_{\text{global}}^2$ is the ratio of the relevance score variance within the segment to the global relevance score variance. These terms collectively measure a segment's duration, average importance, peak significance, and internal content diversity. The coefficients w_d, w_a, w_m, w_v are weighting factors, set to $0.4, 0.2, 0.3, 0.1$ by default.

Segment Filtering. To concentrate the budget on salient content, we prune segments with low importance scores. Specifically, we discard any segment $\mathcal{G}_i$ for which $\text{Imp}(\mathcal{G}_i)$ falls below an adaptive threshold τ, defined as τ = mean(Imp) - $\eta \cdot$ std(Imp). The parameter η controls the filtering aggressiveness and is set to 1.2.

Softmax-Weighted Budget Distribution. We then employ a softmax function over the importance scores to assign a proportional budget, k_i, to each remaining segment $\mathcal{G}_i$:

$$k_i = \left\lfloor K \cdot \frac{\exp(\text{Imp}(\mathcal{G}_i))}{\sum_j \exp(\text{Imp}(\mathcal{G}_j))} \right\rfloor, \quad (7)$$

To ensure the allocation is exact ($\sum k_i$ = K), any frames remaining from the floor operation are assigned greedily to the segments with the largest fractional parts.

3.3.3. Diversity-Aware Intra-segment Frame Selection

With a budget k_i allocated to each temporal semantic segment $\mathcal{G}_i$, the final step is to select the specific frames from within that segment. This localized selection strategy is critical: it prevents frames from one segment being suppressed by visually similar frames from another segment, while maintaining frame diversity within each segment.

For each segment $\mathcal{G}_i$, we first select its most relevant frame as an anchor. We identify the index of this frame, t_{anchor}, and initialize the set of selected indices for the segment, $\mathcal{T}_i$, with it: $t_{\text{anchor}} = \arg\max_{t \in \mathcal{G}_i} s_t$, $\mathcal{T}_i = \{t_{\text{anchor}}\}$. To select the remaining $k_i - 1$ frames, we employ a localized Maximal Marginal Relevance (MMR) approach [4]. This method iteratively identifies the next best frame index, t^*, by balancing relevance and diversity:

$$t^* = \underset{t \in \mathcal{G}_i \setminus \mathcal{T}_i}{\text{argmax}} \left[\lambda \cdot s_t - (1 - \lambda) \cdot \max_{t' \in \mathcal{T}_i} \text{sim}(f_t, f_{t'}) \right], \quad (8)$$

where $\text{sim}(f_t, f_{t'})$ is the cosine similarity between the visual embeddings of candidate frame f_t and a previously selected frame $f_{t'}$. The parameter λ (set to 0.5) balances the trade-off between relevance and diversity. After finding the optimal index t^*, it is added to the set $\mathcal{T}_i$. This process is repeated until $|\mathcal{T}_i| = k_i$. The final selected keyframe subset for the video is the union of frames corresponding to all selected indices across all segments.

4. Experiments

4.1. Experimental Setup

Benchmarks and Backbone Models. We evaluate WFS-SB on three long-video QA benchmarks: VideoMME [8] (900 videos, avg. 17 min, 2,700 QA pairs), MLVU [47] (2,174 questions across 7 categories, avg. 11 min), and LongVideoBench (LVB) [37] (1,337-pair validation set, avg. 12 min). Subtitles are not used. To demonstrate generalizability, we integrate WFS-SB with four state-of-the-art LVLMs: LLaVA-OneVision-7B [10], LLaVA-Video-7B [46], Qwen2.5-VL-7B [2], and InternVL3-8B [49], spanning diverse architectures and training philosophies.

Baselines. We benchmark WFS-SB against a suite of state-of-the-art frame selection methods, including Frame-Voyager [40], KFC [6], BOLT [18], AKS [32], FrameOracle [11], MDP^3 [31], and A.I.R. [50]. For methods with publicly available code (†), we re-evaluate them in our controlled environment to ensure a direct comparison. For others ($*$), we cite the results reported in their original publications under comparable settings.

Implementation Details. Candidate frames are uniformly sampled from each video at 1 frame per second (1 FPS). We use BLIP-2-ITM-ViT-g [13] to compute query-to-frame matching scores. The wavelet decomposition drift factor l is set as 3 with Daubechies-4 (db4) wavelet. Budget allocation weights are w_d=0.4, w_a=0.2, w_m=0.3, w_v=0.1, The filtering aggressiveness factor η is set to 1.2, and the MMR parameter is $\lambda = 0.5$. All experiments are conducted with

Table 1. Comparison of VLMs and various frame selection methods on VideoMME, MLVU, and LVB. We report accuracy scores (%). **Bold** indicates best performance and underline indicates second-best performance in each model-dataset group. †Reproduced results, *Reported results, ‡Training-based methods, §LVLM-dependent iterative method.

Model	Method	Size	Frame	VideoMME [8]			MLVU [47]			LongVideoBench [37]		
				Base	+Method	Δ	Base	+Method	Δ	Base	+Method	Δ
LLaVA-OV [10]	Frame-Voyager*‡ [40]	7B	8	53.3	57.5	+4.2	58.5	65.6	+7.1	-	-	-
	KFC* [6]	7B	8	53.3	55.4	+2.1	58.5	66.2	+7.7	54.5	55.6	+1.1
	BOLT* [18]	7B	8	53.8	56.1	+2.3	58.9	63.4	+4.5	54.2	55.6	+1.4
	AKS† [32]	7B	8	54.1	58.2	+4.1	58.6	62.9	+4.3	54.2	58.4	+4.2
	FrameOracle*‡ [11]	7B	8	53.8	57.5	+3.7	58.4	62.9	+4.5	54.3	56.0	+1.7
	WFS-SB†	7B	8	54.1	**59.3**	**+5.2**	58.6	**67.2**	**+8.6**	54.2	**59.8**	**+5.6**
LLaVA-Video [46]	KFC*	7B	8	55.9	57.6	+1.7	60.5	**66.9**	+6.4	54.2	56.5	+2.3
	BOLT*	7B	8	56.0	58.6	+2.6	-	-	-	-	-	-
	AKS†	7B	8	56.2	60.1	+3.9	57.4	64.2	+6.8	54.9	59.6	+4.7
	FrameOracle*‡	7B	8	55.9	58.9	+3.0	60.5	63.4	+2.9	54.2	56.9	+2.7
	WFS-SB†	7B	8	56.2	**61.7**	**+5.5**	57.4	**66.9**	**+9.5**	54.9	**61.1**	**+6.2**
Qwen2.5-VL [2]	AKS†	7B	32	61.2	64.0	+2.8	59.7	67.2	+7.5	58.9	63.2	+4.3
	MDP^{3}* [31, 50]	7B	32	60.8	63.8	+3.0	59.3	66.2	+6.9	58.1	60.0	+1.9
	A.I.R.* [50]	7B	≤32§	60.8	**65.0**	**+4.2**	59.3	67.5	+8.2	58.1	61.4	+3.3
	WFS-SB†	7B	32	61.2	64.4	+3.2	59.7	**70.4**	**+10.7**	58.9	**64.4**	**+5.5**
InternVL-3 [49]	AKS†	8B	32	65.6	66.3	+0.7	**68.4**	74.2	+5.8	58.5	61.5	+3.0
	MDP^{3}*	8B	32	65.6	66.8	+1.2	68.4	74.0	+5.6	58.3	60.9	+2.6
	A.I.R.*	8B	≤32§	65.6	**68.2**	**+2.6**	68.4	74.5	+6.1	58.3	62.8	**+4.5**
	WFS-SB†	8B	32	65.6	67.4	+1.8	68.4	**74.8**	**+6.4**	58.5	**62.9**	+4.4

frame budgets $K \in \{8, 16, 32, 64\}$ on NVIDIA A800 GPUs with 80 GB of memory. We report accuracy measured by the LMMs-Eval toolkit [43].

4.2. Comparison with State-of-the-Art

Table 1 presents comprehensive results across all benchmarks, LVLM architectures, and frame budgets. WFS-SB demonstrates **powerful plug-and-play enhancement**, delivering consistent improvements without requiring model fine-tuning or architectural modifications—average gains of 3.9% on VideoMME, 8.8% on MLVU, and +5.4% on LongVideoBench across all configurations validate its universal applicability. Our method achieves **state-of-the-art accuracy with superior efficiency**, outperforming both training-based methods (Frame-Voyager, FrameOracle) and iterative approaches (A.I.R.) while maintaining training-free deployment and single-pass inference. Most notably, WFS-SB exhibits **robust performance across diverse benchmarks**, with particularly strong gains under tight frame budgets (e.g., for LLaVA-Video-7B at K=8, gains are 9.5% on MLVU, 6.2% on LongVideoBench and 5.5% on VideoMME), demonstrating that our wavelet-based frame selection effectively adapts to varying content complexity and temporal scales.

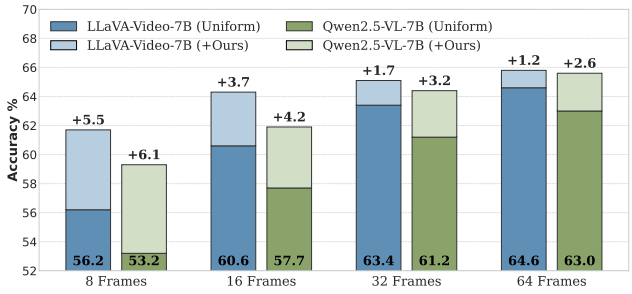


Figure 3. Performance comparison across different frame budgets K on VideoMME. WFS-SB consistently outperforms uniform sampling across all LVLMs and budget settings.

4.3. Ablation Study

Different Sampling Frames. Figure 3 illustrates the performance of WFS-SB across frame budgets {8, 16, 32, 64} on VideoMME with multiple backbones. We observe consistent improvements over uniform sampling (+1.2% to +5.5%), with particularly pronounced gains at smaller budgets (K=8, 16). This validates that boundary-aware selection becomes increasingly critical when frame resources are constrained, demonstrating strong generalization across different VLMs and budget sizes.

Different LVLM Scales. To evaluate scalability, we test

Table 2. Ablation of different LVLM scales using Qwen2.5-VL on VideoMME. Accuracy (%) is reported. Results show performance with K=16 frames (left) and K=32 frames (right). Δ column shows improvement gains for 16 and 32 frames respectively.

Model Scale	Baseline	+WFS-SB	Δ
Qwen2.5-VL-3B	54.1/57.8	58.4/60.0	+4.3/+2.2
Qwen2.5-VL-7B	57.7/61.2	61.9/64.4	+4.2/+3.2
Qwen2.5-VL-32B	60.2/62.9	63.0/66.1	+2.8/+3.2
Qwen2.5-VL-72B	63.3/66.2	66.1/68.9	+2.8/+2.7

Table 3. Ablation of different VLMs for query-frame similarity scoring with Qwen2.5-VL-7B (K=16). Accuracy (%) is reported. Although BLIP-ITM achieves highest performance on VideoMME and LVB, we select BLIP-2-ITM as default because it provides more consistent average benefits across diverse settings.

VLM Scorer	VideoMME	MLVU	LVB
Uniform	57.7	56.2	57.0
BLIP-ITM [12]	**62.8**	67.7	**63.2**
CLIP-VIT-B [25]	61.7	66.5	60.6
SigLIP-so400m [41]	61.9	66.4	61.9
BLIP-2-ITM (Ours) [13]	61.9	**67.9**	62.5

WFS-SB on Qwen2.5-VL at four scales (3B, 7B, 32B, 72B). Table 2 shows consistent improvements with average gains of 3.3%, 3.7%, +3.0%, and 2.8% respectively. The consistent gains validate that WFS-SB's benefits are orthogonal to model scale, making it equally valuable for both resource-constrained deployments (smaller models) and performance-critical applications (larger models).

Different VLMs for Query-Frame Similarity. We investigate different VLMs for computing query-frame relevance. Table 3 compares four widely-used VLMs: BLIP-ITM, CLIP-VIT-B, SigLIP-so400m, and BLIP-2-ITM. All VLMs demonstrate consistent performance gains when integrated with WFS-SB (even on CLIP-VIT-B, there is a 3.6% improvement), validating the robustness and generalization capability of our wavelet-based framework across different feature extractors. While BLIP-ITM achieves the highest performance on VideoMME (62.8%) and LongVideoBench (63.2%), we select BLIP-2-ITM as default because it provides more consistent average benefits across diverse settings. The greater success of the BLIP series highlights the benefit of dedicated image-text matching due to its explicit image-text alignment objective.

Component Ablation. Table 4 systematically evaluates each component's contribution, with emphasis on wavelet-based boundary detection's critical role. Replacing DWT with naive alternatives—local minima or gradient-based methods—causes substantial degradation (61.9→60.8 and 61.9→61.2 on VideoMME; 67.9→64.6 and 67.9→66.8 on MLVU), with local minima suffering -3.3% on MLVU. This validates our core hypothesis: wavelet transform's multi-scale decomposition inherently suppresses high-frequency noise in the similarity signal, isolating genuine semantic transitions while filtering spurious fluctuations. In contrast, raw methods suffer from noise amplification, causing false boundary detections and fragmented segmentation. Beyond DWT, removing adaptive budget allocation (-0.5%) or MMR selection (-1.2% and -5.2%) also degrades performance, confirming all components synergistically contribute (+4.2% on VideoMME, +11.7% on MLVU).

Table 4. Ablation of different components with Qwen2.5-VL-7B (K=16). We remove each component to measure its contribution. Accuracy (%) is reported. B.A. means Budget Allocation.

Configuration	VideoMME	MLVU
Uniform Sampling	57.7	56.2
WFS-SB (Our full method)	**61.9**	**67.9**
w/o DWT (use raw local minima)	60.8	64.6
w/o DWT (use raw gradient)	61.2	66.8
w/o Adaptive B.A. (Average B.A.)	61.6	67.4
w/o MMR (topK selection)	60.9	66.7
w/o MMR (uniform selection)	59.2	62.7

Table 5. Ablation of wavelet families and decomposition level l with Qwen2.5-VL-7B (K=16). Accuracy (%) is reported.

Wavelet	Db4	Db4	Db8	Haar	Sym4	Bior3.3
l	3	4	3	3	3	3
VideoMME	**61.9**	60.4	**61.9**	61.3	**61.9**	**61.9**
MLVU	67.9	**68.3**	66.4	68.1	67.6	68.2

Wavelet Family and Decomposition Level. Table 5 demonstrates remarkable robustness across wavelet families and validates our adaptive decomposition level formula (Equation 2). At l=3, multiple families achieve highly consistent performance: Db4, Db8, Sym4, and Bior3.3 all reach 61.9% on VideoMME, while Haar achieves 61.3%—only 0.6% variance despite distinct mathematical properties. This confirms multi-scale signal decomposition, rather than specific basis choice, drives effectiveness. We select Db4 with l=3 for best balance across benchmarks (61.9%/67.9%). The adaptive level formula proves critical: at l=4, VideoMME drops to 60.4% due to over-smoothing, while MLVU improves to 68.3% due to different temporal scales. The drift factor in Equation 2 automatically adjusts depth based on video length, ensuring optimal noise suppression across diverse durations.

Hyperparameter Sensitivity. A critical advantage of WFS-SB is its robustness despite involving multiple hyperparameters. Table 6 analyzes sensitivity to the importance score weights and MMR diversity parameter. Re-

Table 6. Ablation of hyperparameters on VideoMME and MLVU with Qwen2.5-VL-7B (K=16). Accuracy (%) is reported.

w_d, w_a, w_m, w_v	λ	VideoMME	MLVU
0.4, 0.2, 0.3, 0.1	0.5	61.9	**67.9**
0.0, 0.3, 0.3, 0.2	0.5	61.6	64.8
0.5, 0.0, 0.3, 0.2	0.5	61.8	66.8
0.5, 0.3, 0.0, 0.2	0.5	62.0	66.8
0.5, 0.2, 0.3, 0.0	0.5	**62.1**	67.4
0.4, 0.2, 0.3, 0.1	0.3	61.2	67.5
0.4, 0.2, 0.3, 0.1	0.7	61.6	66.5

Table 7. Computational efficiency breakdown on VideoMME with Qwen2.5-VL-7B (avg. N=1040 frames, K=32). Times measured on a single NVIDIA A800 GPU and two Intel Xeon Silver 4314 CPUs (16 cores), averaged over 900 videos.

Component	Time (s)	Complexity
ITM Signal Extraction	19.4	$O(N)$
DWT and Boundary Detection	~ 0	$O(N \log N)$
Budget Allocation	~ 0	$O(M)$
MMR Selection	0.7	$O(NK^2)$
LVLM Inference	4.4	-

moving any single component degrades performance (-0.5% to -3.1% on MLVU), confirming that all four factors—duration, average relevance, peak relevance, and variance—contribute complementary information for segment importance estimation. The relatively small degradation indicates that while each component is necessary, the overall framework is robust to specific weight values. **Most remarkably, we use identical settings across all four LVLM backbones, all three benchmarks, and all frame budgets without task-specific tuning.** The MMR parameter λ exhibits particularly stable behavior: performance varies by only 0.4% on VideoMME across $\lambda \in [0.3, 0.7]$, with optimal balance at λ=0.5. This robustness stems from wavelet decomposition inherently adapting to signal characteristics and budget allocation automatically scaling to segment distributions, enabling deployment with default hyperparameters without expensive grid search.

4.4. Complexity and Time Analysis

Table 7 presents a comprehensive analysis of WFS-SB's computational efficiency. For videos on VideoMME (avg. N=1040 frames), the timing breakdown reveals that ITM signal extraction dominates at 19.4s (79% of total time)—this represents an optimization opportunity as ITM computation could be accelerated through batching, quantization, or distillation of the BLIP-2-ITM model. The wavelet processing pipeline itself (DWT, boundary detection, budget allocation, MMR) adds only 0.7s overhead. Critically, while ITM extraction introduces computational cost compared to naive uniform sampling, this overhead is entirely justified by the substantial performance gains.

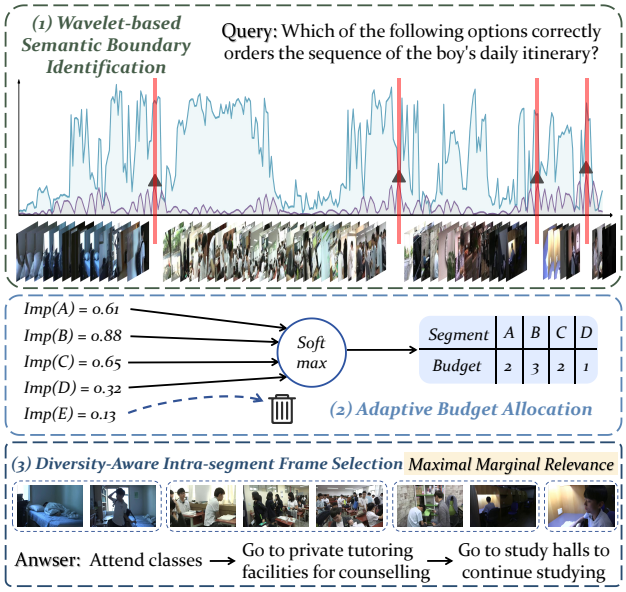


Figure 4. Visualization of our WFS-SB framework on a daily itinerary video. Our wavelet-based method first partitions the video into coherent segments by detecting robust semantic boundaries. Subsequently, after filtering segments based on importance scores, Maximal Marginal Relevance sampling is applied within each to select keyframes that preserve the essential narrative, correctly capturing the sequence of class, tutoring, and self-study.

4.5. Visualization Analysis

Figure 4 demonstrates WFS-SB on a daily activities video with distinct yet visually similar events. Our wavelet-based boundary detection identifies key semantic transitions from the noisy similarity signal, segmenting the video into coherent clips. Subsequent stratified sampling ensures selected frames represent each event (class, tutoring, self-study) while preserving chronological narrative, faithfully reconstructing the event sequence and highlighting our method's strength in process-oriented understanding.

5. Conclusion

In this paper, we introduced WFS-SB, a training-free framework that enhances long-video understanding for LVLMs by focusing on narrative structure. Our key innovation is the use of wavelet transform to reliably detect semantic boundaries, dividing the video into coherent clips. A two-stage selection strategy then allocates frames based on clip importance and intra-clip diversity. Extensive experiments demonstrate that WFS-SB significantly outperforms state-of-the-art methods across multiple benchmarks, highlighting the importance of capturing a video's narrative flow.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2025YFE0113500), the National Science Fund for Distinguished YoungScholars (No. 62525605), the National Natural Science Foundation of China (No. U25B2066, No. U22B2051, No. 62576299, No. 62406071, No. 62302411), and the Fundamental Research Funds for the Central Universities.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 5, 6

[3] Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *CVPR*, pages 29071–29082, 2025. 3

[4] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998. 5

[5] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *ECCV*, pages 75–92. Springer, 2024. 2

[6] Bo Fang, YuXin Song, Haoyuan Sun, Qiangqiang Wu, Wenhao Wu, and Antoni B. Chan. Threading keyframe with narratives: MLLMs as strong long video comprehenders. In *ICLR*, 2026. 3, 5, 6

[7] Marie Farge et al. Wavelet transforms and their applications to turbulence. *Annual review of fluid mechanics*, 24(1):395–458, 1992. 2

[8] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, pages 24108–24118, 2025. 5, 6

[9] Maximilian Grobbelaar, Souvik Phadikar, Ebrahim Ghaderpour, Aaron F Struck, Nidul Sinha, Rajdeep Ghosh, and Md Zaved Iqubal Ahmed. A survey on denoising techniques of electroencephalogram signals using wavelet transform. *Signals*, 3(3):577–586, 2022. 2

[10] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. 5, 6

[11] Chaoyu Li, Tianzhi Li, Fei Tao, Zhenyu Zhao, Ziqian Wu, Maozheng Zhao, Juntong Song, Cheng Niu, and Pooyan Fazli. Frameoracle: Learning what to see and how much to see in videos. *arXiv preprint arXiv:2510.03584*, 2025. 3, 5, 6

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 7, 6

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023. 3, 5, 7, 6

[14] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, pages 323–340. Springer, 2024. 2

[15] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, pages 5971–5984, 2024. 1

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pages 34892–34916, 2023. 1, 2

[17] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. In *ICLR*, 2025. 2

[18] Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *CVPR*, pages 3318–3327, 2025. 3, 5, 6

[19] Yongdong Luo, Wang Chen, Xiawu Zheng, Weizhong Huang, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Jiebo Luo, et al. Quota: Query-oriented token assignment via cot query decouple for long video comprehension. *arXiv preprint arXiv:2503.08689*, 2025. 2

[20] Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In *NeurIPS*, 2025. 2

[21] Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *CVPR*, pages 18936–18946, 2025. 2

[22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, pages 12585–12602, 2024. 1

[23] Azra Nasreen and G Dr Shobha. Key frame extraction using edge change ratio for shot segmentation. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(11):4421–4423, 2013. 2

[24] Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryoo, Donghyun Kim, and Michael S. Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. In *European Chapter of the Association for Computational Linguistics*, 2026. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021. 7, 6

[26] C Va Sheena and NK Narayanan. Key-frame extraction by analysis of histograms of video frames using statistical methods. *Procedia Computer Science*, 70:36–40, 2015. 2

[27] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 2

[28] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *CVPR*, pages 26160–26169, 2025. 2

[29] MIMRAMZ Sifuzzaman, M Rafiq Islam, and Mostafa Z Ali. Application of wavelet transform and its advantages compared to fourier transform. *Journal of Physical Sciences*, 13: 121–134, 2009. 2

[30] Baiyang Song, Jun Peng, Yuxin Zhang, Guangyao Chen, Feidiao Yang, and Jianyuan Guo. Ktv: Keyframes and key tokens selection for efficient training-free video llms. *arXiv preprint arXiv:2602.03615*, 2026. 2

[31] Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp3: A training-free approach for list-wise frame selection in video-llms. In *ICCV*, pages 24090–24101, 2025. 3, 5, 6

[32] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *CVPR*, pages 29118–29128, 2025. 3, 5, 6

[33] Chunwei Tian, Menghua Zheng, Wangmeng Zuo, Bob Zhang, Yanning Zhang, and David Zhang. Multi-stage image denoising with the wavelet transform. *Pattern Recognition*, 134:109050, 2023. 2

[34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 5

[36] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *CVPR*, pages 3272–3283, 2025. 2

[37] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS*, pages 28828–28857, 2024. 5, 6

[38] Zhirong Wu, Xiaodong Wang, Langling Huang, Teng Xu, and Peixi Peng. A training-free framework for long video understanding via video-query-options similarity. In *ICLR*, 2026. 3

[39] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Rethinking temporal search for long-form video understanding. In *CVPR*, pages 8579–8591, 2025. 3

[40] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. In *ICLR*, 2025. 3, 5, 6

[41] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 7, 6

[42] Dengsheng Zhang. Wavelet transform. In *Fundamentals of image data mining: Analysis, Features, Classification and Retrieval*, pages 35–44. Springer, 2019. 2

[43] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025. 6

[44] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 2

[45] Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *ICCV*, pages 22056–22065, 2025. 3, 5

[46] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025. 5, 6

[47] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 5, 6

[48] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024. 1, 2

[49] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 5, 6

[50] Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.i.r.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *ICLR*, 2026. 3, 5, 6