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

Supplementary Material

This supplementary material provides extended technical details, additional experiments, and in-depth analyses that complement the main paper. The content is organized as follows:

Section 6 – Limitations and Future Work. We discuss three main limitations of WFS-SB and outline four promising research directions for extending wavelet-based frame selection.

Section 7 – Further Explanation of the Method. We present the detailed mathematical formulation of our adaptive peak detection algorithm.

Section 8 – Additional Experimental Results. We report extended experimental analyses, including comparisons with additional baseline methods (Section 8.1), performance across different frame budgets and models (Section 8.2), ablation studies on frame sampling rates (Section 8.3), extended VLM comparisons (Section 8.4), hyperparameter sensitivity analysis (Section 8.5), and qualitative visualizations (Section 8.6).

6. Limitations and Future Work

6.1. Limitations

While WFS-SB achieves strong performance across multiple benchmarks and architectures, we acknowledge three main limitations.

Computational Overhead of ITM Feature Extraction. The primary bottleneck is extracting Image-Text Matching (ITM) scores using BLIP-2. As reported in Table 7 of the main paper, this accounts for approximately 79% (19.4 seconds) of preprocessing time. Although the wavelet transform itself is highly efficient ($O(N \log N)$) and boundary detection operates in near-constant time, the upfront cost of computing dense query-frame similarity scores can be prohibitive for extremely long videos or real-time applications. This limitation is shared with other vision-language alignment methods and becomes pronounced for multi-hour videos.

Dependency on Vision-Language Feature Quality. Our framework relies on the quality and calibration of ITM scores from a pretrained vision-language model. While Table 3 demonstrates robustness across multiple VLMs (BLIP-ITM, BLIP-2-ITM, CLIP, SigLIP), WFS-SB performance is bounded by the semantic understanding capabilities of the underlying VLM. Poorly calibrated scores—due to domain shift, adversarial perturbations, or out-of-distribution content—may cause the wavelet analysis to miss meaningful semantic transitions. This underscores the importance of selecting an appropriate VLM for the target domain.

Sensitivity to Extreme Temporal Structures. Videos with very rapid scene cuts (e.g., advertisements, montages) may trigger excessive boundary detection, leading to over-segmentation. Conversely, videos with extended low-relevance periods interrupted by brief high-relevance moments may cause the segment filtering mechanism to inadvertently discard important short segments. While our default hyperparameters ($\eta = 1.2$) work well across diverse benchmarks, domain-specific tuning may be necessary for extreme cases.

6.2. Future Work

Building on the identified limitations, we propose four promising research directions for extending wavelet-based frame selection.

Efficient ITM Score Approximation. To address computational overhead, future work could explore: (1) *Distillation-based approximation*, where a smaller model is trained to predict ITM scores from visual features; (2) *Adaptive FPS sampling*, using lower frame rates for longer videos to reduce preprocessing time while maintaining performance; (3) *Sparse ITM computation*, where only a subset of frames are scored initially, and additional frames are queried adaptively based on wavelet analysis. These strategies could significantly reduce preprocessing time.

Learned Wavelet Kernels for Video Decomposition. While Table 5 demonstrates robustness across different wavelet families, all tested wavelets are hand-designed. An intriguing direction is learning *data-driven wavelet bases* optimized for video semantic boundary detection through end-to-end training on video segmentation datasets, where wavelet filters are parameterized as learnable convolutional kernels. This could capture domain-specific temporal patterns more effectively than generic wavelets.

Multi-Query and Open-Ended Video Understanding. Our current formulation assumes a single query per video. Future work could extend WFS-SB to *multi-query scenarios* by jointly analyzing temporal relevance signals for all queries and selecting a unified frame set that maximizes coverage. For open-ended video understanding without explicit queries, the framework could be adapted to select keyframes based on *self-supervised semantic change detection* using frame-to-frame similarity or anomaly detection.

Extension to Multimodal Signals. While our focus is

visual-textual alignment, many videos contain rich audio and subtitle information. Future work could extend wavelet-based boundary detection to *multimodal temporal signals*, fusing audio energy, speech transcripts, and visual ITM scores into a unified relevance signal. This could enable more robust boundary detection in scenarios where visual cues alone are ambiguous (e.g., dialogue-heavy scenes where semantic transitions are signaled primarily by speech).

7. Further Explanation of Our Method

Section 3.3.1 briefly described the peak detection process for identifying semantic boundaries in the change intensity signal $c_t = |\tilde{s}_t|$. Here we provide the complete mathematical formulation and algorithmic pseudocode.

The algorithm identifies local maxima in c_t that satisfy two adaptive criteria: *height* and *prominence*. These criteria ensure detected peaks correspond to genuine semantic boundaries rather than noise-induced fluctuations.

Change Intensity Signal. Given the wavelet-reconstructed semantic change signal $\tilde{s}_t = \text{IDWT}(\{\mathbf{0}, d_J, \mathbf{0}, \ldots, \mathbf{0}\})$ (Equation 4), we compute its absolute value to obtain a non-negative change intensity:

$$c_t = |\tilde{s}_t|, \quad t = 1, \ldots, N. \tag{9}$$

The peaks in c_t correspond to moments of maximal semantic transition.

Adaptive Height Threshold. To filter low-magnitude fluctuations, we require peaks to exceed a data-driven threshold:

$$\tau_{\text{height}} = \bar{c} + \alpha \cdot \sigma_c, \tag{10}$$

where $\bar{c}$ is mean change intensity, σ_c is standard deviation, and α is the height factor (default: $\alpha = 0.5$). A peak at index t is retained only if $c_t \geq \tau_{\text{height}}$.

Adaptive Prominence Threshold. Prominence measures how much a peak stands out relative to surrounding valleys. We require prominence to exceed a threshold proportional to the signal's dynamic range:

$$\tau_{\text{prom}} = \beta \cdot (\max_t c_t - \min_t c_t), \tag{11}$$

where β is the prominence factor (default: $\beta = 0.05$). This suppresses broad, low-contrast humps.

Minimum Distance Constraint. To prevent over-segmentation, we enforce minimum temporal separation δ_{min} between consecutive peaks via non-maximum suppression. The minimum distance is adaptively set as:

$$\delta_{\text{min}} = \max\left(5, \lfloor N \times 0.02 \rfloor\right), \tag{12}$$

ensuring at least 5 frames separation for short videos and proportionally larger separation for long videos.

Algorithm 1 Adaptive Peak Detection for Semantic Boundaries

Require: Change intensity signal $\{c_t\}_{t=1}^{N}$, height factor α, prominence factor β
Ensure: Set of boundary indices $\mathcal{B} = \{b_1, \ldots, b_M\}$

Compute signal statistics:
  $\bar{c} \leftarrow \frac{1}{N}\sum_{t=1}^{N} c_t$ ▷ Mean intensity
  $\sigma_c \leftarrow \sqrt{\frac{1}{N}\sum_{t=1}^{N}(c_t - \bar{c})^2}$ ▷ Std deviation
  $R_c \leftarrow \max_t c_t - \min_t c_t$ ▷ Dynamic range
Compute adaptive thresholds:
  $\tau_{\text{height}} \leftarrow \bar{c} + \alpha \cdot \sigma_c$
  $\tau_{\text{prom}} \leftarrow \beta \cdot R_c$
  $\delta_{\text{min}} \leftarrow \max(5, \lfloor N \times 0.02 \rfloor)$
Initialize candidate peak set: $\mathcal{P}_{\text{cand}} \leftarrow \emptyset$
for $t = 2$ **to** $N - 1$ **do**
  if $c_t > c_{t-1}$ **and** $c_t > c_{t+1}$ **then** ▷ Local maximum
    if $c_t \geq \tau_{\text{height}}$ **then**
      Compute prominence p_t of peak at t
      if $p_t \geq \tau_{\text{prom}}$ **then**
        $\mathcal{P}_{\text{cand}} \leftarrow \mathcal{P}_{\text{cand}} \cup \{t\}$
      end if
    end if
  end if
end for
Sort $\mathcal{P}_{\text{cand}}$ by prominence in descending order
Initialize final boundary set: $\mathcal{B} \leftarrow \emptyset$
for each peak $t_p \in \mathcal{P}_{\text{cand}}$ (in sorted order) **do**
  if $\mathcal{B} = \emptyset$ **or** $\min_{b \in \mathcal{B}} |t_p - b| \geq \delta_{\text{min}}$ **then**
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{t_p\}$
  end if
end for
Sort $\mathcal{B}$ in ascending order
return $\mathcal{B}$

Algorithm 1 provides the complete procedural description of the peak detection process.

Hyperparameter Robustness Analysis. Table 8 evaluates robustness to the height factor α and prominence factor β on VideoMME and MLVU with Qwen2.5-VL-7B (K=16).

Table 8. Robustness of peak detection to hyperparameter variations. Evaluated on VideoMME and MLVU with Qwen2.5-VL-7B (K=16). Performance varies by only 1.1% or 1.2% across different settings, demonstrating strong hyperparameter insensitivity.

α	β	VideoMME	MLVU
0.5	0.05	61.9	67.9
0.0	0.05	**62.6**	67.2
1.0	0.05	61.5	**68.4**
0.5	0.00	61.8	67.6
0.5	0.1	62.0	67.9

The results demonstrate strong robustness: performance varies by only 1.1% or 1.2% across different hyperparameter settings. This insensitivity to α and β indicates that the

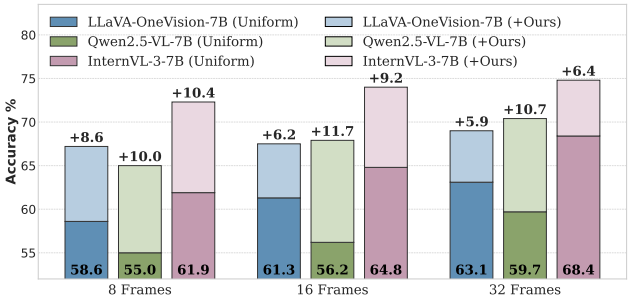

Figure 5. **Performance across different frame budgets on MLVU.** WFS-SB consistently outperforms uniform sampling across all LVLMs and budget settings.

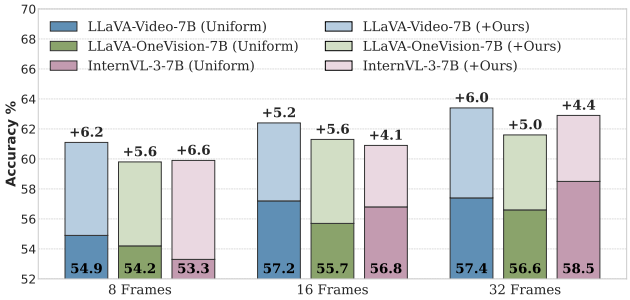

Figure 6. **Performance across different frame budgets on LongVideoBench.** WFS-SB demonstrates robust improvements across multiple architectures and budget constraints.

adaptive thresholding mechanism naturally adjusts to signal characteristics, making the method reliable across diverse video content without requiring task-specific tuning.

8. Additional Experimental Results

This section presents extended experimental results that complement the analyses in the main paper, including comparisons with additional baseline methods, extended frame budget and model ablations, performance under different frame rates, hyperparameter sensitivity studies, and qualitative visualizations.

8.1. Comparison with More Baseline Methods

Table 9 extends the main comparison (Table 1) by evaluating additional recent methods on Qwen2-VL-7B.

8.2. Extended Frame Budget and Model Ablation

Figure 5 and Figure 6 extend the frame budget analysis from Figure 3 to MLVU and LongVideoBench. WFS-SB consistently outperforms uniform sampling across all models and budgets, with particularly strong gains at smaller budgets (K=8, 16).

8.3. Performance Across Different Frame Rates

Our default pipeline samples candidate frames at 1 FPS. To address the ITM extraction bottleneck (79% of preprocessing time, Table 7), we evaluate an adaptive FPS strategy that uses different sampling rates for videos of different durations in VideoMME. Table 10 reports results on VideoMME with Qwen2.5-VL-7B.

The results demonstrate that adaptive FPS strategies significantly reduce computational overhead. By using lower sampling rates for medium and long videos (e.g., 1-0.5-0.25 fps for short-medium-long videos), ITM extraction time decreases from 19.4s to 5.8s (70% reduction) while maintaining comparable or even slightly improved performance. This validates that reducing FPS for longer videos is a practical approach to mitigate the preprocessing time bottleneck without sacrificing accuracy.

8.4. Extended VLM Comparison for Query-Frame Matching

Table 11 extends the VLM comparison from Table 3 by evaluating on LLaVA-Video-7B with K=16 across all three benchmarks.

The results demonstrate WFS-SB's robustness to VLM choice, with all tested models yielding substantial improvements over uniform sampling. Although BLIP-ITM performs better under the settings in Table 3, we select BLIP-2-ITM as default because it provides more consistent average benefits across diverse settings. Specifically, BLIP-2-ITM achieves the best performance, with accuracies of 64.3%, 71.0%, and 62.4% on VideoMME, MLVU, and LongVideoBench, respectively, demonstrating its superior cross-benchmark generalization ability. This reflects BLIP-2-ITM's better calibration from its enhanced image-text alignment objective.

8.5. Hyperparameter Ablation: Segment Filtering

Section 3.3.2 introduced a segment filtering mechanism controlled by threshold $\tau = \text{mean}(\text{Imp}) - \eta \cdot \text{std}(\text{Imp})$, where η controls filtering aggressiveness. Table 12 analyzes the impact of η on performance.

The results demonstrate two key properties. **Robustness:** Performance varies by only 0.2-0.6% across different η values, indicating the method is insensitive to this hyperparameter. **Effectiveness:** Comparing the no-filtering baseline ("-") with filtered configurations shows that segment filtering consistently improves performance by 0.4-0.9%, as it concentrates the frame budget on high-relevance segments while discarding low-importance regions. The default $\eta = 1.2$ provides a reliable balance.

8.6. Additional Qualitative Examples

Figure 7 provides a qualitative comparison of different frame selection strategies on a video from VideoMME.

Table 9. Extended comparison with additional baseline methods on VideoMME, MLVU, and LongVideoBench using Qwen2-VL-7B (K=8). WFS-SB achieves competitive performance with consistent improvements across all benchmarks (+4.4%, +8.4%, +3.8%). *Same token budget but more frames.

Model	Method	Size	Frame	VideoMME [8]			MLVU [47]			LongVideoBench [37]		
				Base	+Method	Δ	Base	+Method	Δ	Base	+Method	Δ
Qwen2-VL-7B [35]	KFC [6]	7B	8	55	56.7	+1.7	59.6	**65.9**	+6.3	53.4	54.6	+1.2
	Q-Frame [45]	7B	8 (4+8+32)*	53.7	**58.3**	**+4.6**	56.9	65.4	**+8.5**	53.5	**58.4**	**+4.9**
	WFS-SB	7B	8	52.0	56.4	+4.4	55.0	63.4	+8.4	51.7	56.5	+3.8

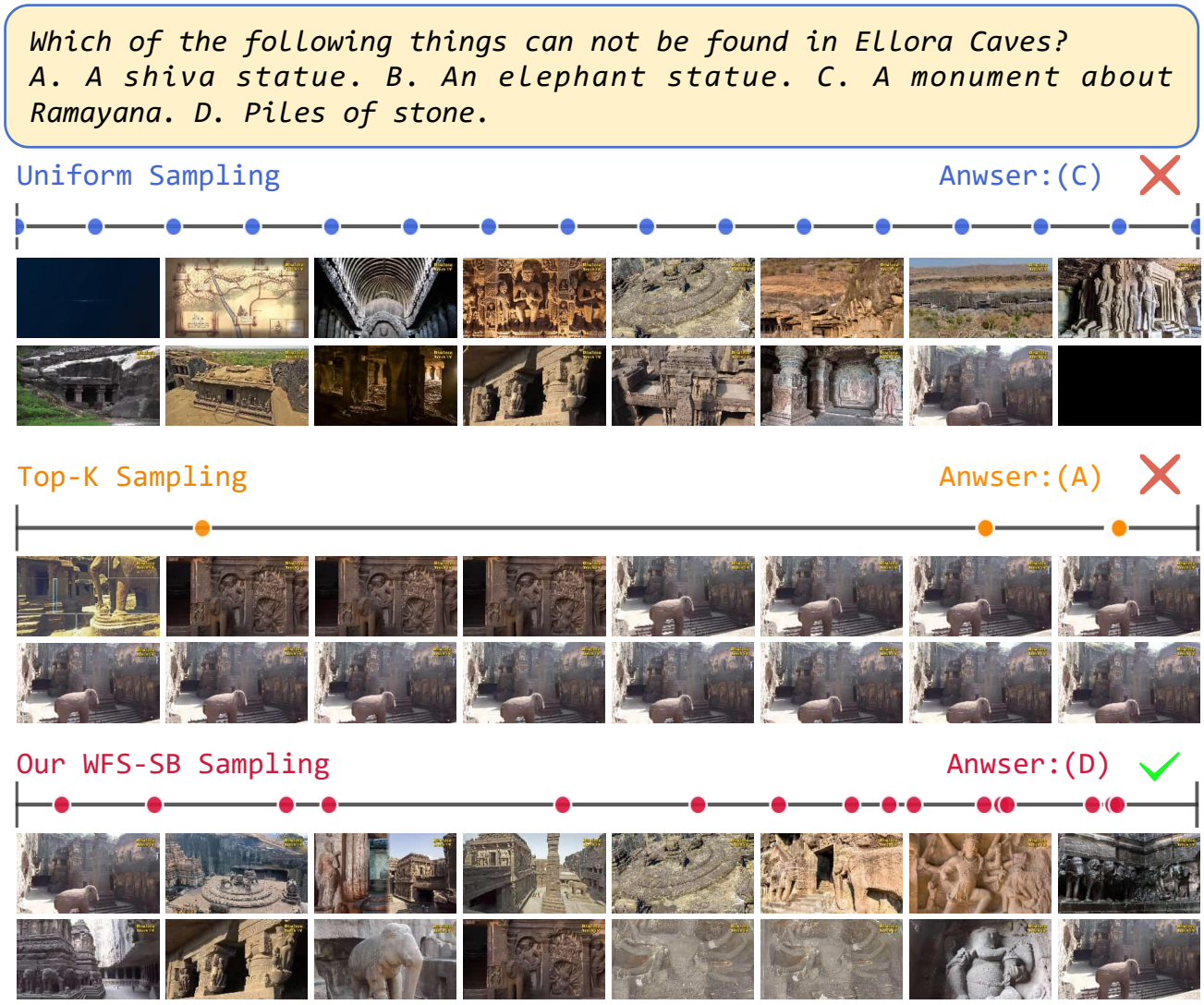

Figure 7. **Qualitative comparison on Ellora Caves video.** Query: *"Which of the following things can not be found in Ellora Caves? A. A shiva statue. B. An elephant statue. C. A monument about Ramayana. D. Piles of stone."* **Top:** Uniform sampling (Answer: C) - Too random, lacks query-specific awareness, missing critical frames. **Middle:** Top-K sampling (Answer: A) - Focuses on highly salient objects (e.g., elephant statue) but misses other equally important yet less visually prominent frames. **Bottom:** Our WFS-SB (Answer: D, correct) - Through wavelet-based boundary detection, effectively extracts semantic boundaries to segment the video. Adaptive budget allocation and intra-segment diversity selection successfully identify keyframes that enable correct question answering.

Table 10. Impact of adaptive FPS sampling on performance and ITM extraction time. Evaluated on VideoMME with Qwen2.5-VL-7B. The notation "1-0.75-0.5" indicates using 1 fps for short videos, 0.75 fps for medium videos, and 0.5 fps for long videos. Adaptive strategies reduce ITM time by 46% (1-0.75-0.5) and 70% (1-0.5-0.25) respectively while maintaining or improving performance. F. represents the number of input frames.

Sampling Rate	**8 F.**	**16 F.**	**32 F.**	**ITM Time (s)**
Uniform Sampling	53.2	57.7	61.2	-
1 fps	59.3	61.9	64.4	19.4
1-0.75-0.5 fps	58.9	**62.0**	64.5	10.5
1-0.5-0.25 fps	**59.4**	61.9	**64.6**	5.8

Table 11. Ablation on different VLMs for query-frame similarity scoring with LLaVA-Video-7B (K=16). We select BLIP-2-ITM as default for its superior performance across all benchmarks.

VLM Scorer	**VideoMME**	**MLVU**	**LVB**
Uniform	60.6	60.9	57.2
BLIP-ITM [12]	63.2	69.5	61.8
CLIP-VIT-B [25]	63.6	67.3	62.2
SigLIP-so400m [41]	62.4	68.9	61.5
BLIP-2-ITM (Ours) [13]	**64.3**	**71.0**	**62.4**

Table 12. Robustness of segment filtering to hyperparameter η. Evaluated on VideoMME and MLVU with Qwen2.5-VL-7B. Performance varies by only 0.2-0.6% across different η values. Filtering is effective: compare "-" (no filtering, all segments retained) vs. others.

η	**VideoMME**	**MLVU**
1.2	61.9	67.9
-	61.5	67.2
1.0	61.7	67.5
1.5	61.9	68.1

This example illustrates the advantages of the WFS-SB Method. Uniform sampling fails due to its random, query-agnostic nature, missing critical content. Top-K sampling exhibits bias toward visually salient objects (e.g., prominent elephant statue), overlooking other equally important but less conspicuous elements necessary for answering the question. In contrast, WFS-SB leverages wavelet-based semantic boundary detection to partition the video into coherent segments, then applies adaptive budget allocation and diversity-aware selection within segments to capture comprehensive coverage of all relevant content, enabling accurate question answering.