Less is More: Token-Efficient Video-QA via Adaptive Frame-Pruning and Semantic Graph Integration
-supplementary material-

Shaoguang Wang[1] Weiyu Guo[1] Ziyang Chen[1] Yijie Xu[1] Xuming Hu[1,2*] Hui Xiong[1,2*]

[1]The Hong Kong University of Science and Technology (Guangzhou), China
[2]The Hong Kong University of Science and Technology, Hong Kong SAR, China

swang440@connect.hkust-gz.edu.cn, xuminghu@hkust-gz.edu.cn, xionghui@ust.hk

A. Generalizability Studies on *VSLS**

A Note on *VSLS.** In our generalizability study, we use a specific, publicly available version of the *VSLS* [2] codebase from May 2025, which we denote as *VSLS**. It is important to note that this version differs slightly from the final published version of *VSLS*. The primary difference lies in the sampling strategy: this earlier version normalizes the relevance scores into a probability distribution before sampling, whereas the final version performs a direct Top-K selection on the raw scores. We use this specific, frozen version to ensure the reproducibility of our experiments presented here.

Experimental Setup. This section provides the complete component-wise ablation results for applying our framework to the output of *VSLS**.

The methods compared are:

- ***VSLS**** **Baseline**: The two high-cost methods using a fixed number of frames (32 or 8) via '*VSLS**'.
- **Matched-Budget Strategies**: Four low-cost methods that all operate on the same, drastically reduced number of frames determined by our ***AFP*** algorithm for each Video-QA instance.
  - Uniform (Matched): Naive uniform sampling.
  - *VSLS** (Top-N, Matched): A strong baseline that truncates the *VSLS** list.
  - ***AFP*** only (Ours): Our pruning method without the graph.
  - ***AFP + Graph*** (Ours): Our full proposed method.

The detailed results are presented in Table 1.

Analysis from Tabular Data. The results in Table 1 robustly confirm that our framework acts as a versatile refinement module for *VSLS**. First, our full ***AFP + Graph*** method achieves a highly favorable efficiency-performance trade-off. The performance gains are particularly striking on open-source models. For instance, on LONGVIDEOBENCH with LLaVA-Video-7B (from 8 frames), our method boosts the accuracy on short videos from 50.0% (*VSLS** Baseline) to a remarkable **72.0%**, a **+22 point** absolute improvement—while using only 2.2 frames. Second, the superiority of our frame selection is evident. Under a strictly matched frame budget, '***AFP*** only' (e.g., on VIDEOMME with GPT-4o from 32 frames) demonstrates clear performance advantages over '*VSLS** (Top-N, Matched)', indicating that our intelligent clustering selects a more representative set of frames. In addition, the semantic graph proves its crucial role. The consistent and significant performance leap from '***AFP*** only' to '***AFP + Graph***' across all settings validates our rationale that the graph provides the essential semantic scaffolding to unlock the MLLM's full potential.

Visual Analysis of Efficiency-Performance. This trade-off is best visualized in the efficiency-performance plot in Figure 1, which focuses on the challenging **GPT-4o on VIDEOMME (from Top-32)** scenario. The plot clearly illustrates ***AFP + Graph***'s superiority. The *VSLS** Baseline resides in the bottom-right, representing high cost and moderate performance. Within the highly competitive low-cost zone on the left, our full ***AFP + Graph*** method distinguishes itself, positioned highest in the desirable top-left quadrant. It not only significantly outperforms the '*VSLS** (Top-N, Matched)' baseline (57.55% vs. 51.29%) but also surpasses the '***AFP*** only' approach (54.18%), demonstrating the crucial performance boost provided by the semantic graph.

B. Generalizability Studies on *T**

Experimental Setup. To further validate the universal applicability of our framework, we conducted a comprehensive set of experiments applying it to another SOTA keyframe selector, *T** [5]. The comparison follows the same structure as

*Corresponding authors.

Table 1. **Complete component-wise ablation study results on the LONGVIDEOBENCH and VIDEOMME datasets using *VSLS** as the upstream selector.** For each model and initial frame setting, we compare our full method against various baselines. **Bold** indicates the best performance among the four matched-budget strategies for each accuracy column.

Method	Avg. Fr.	Long(%)	Med.(%)	Short(%)
Evaluation from Top 8 Keyframes				
Model: GPT-4o				
*VSLS** Baseline	8.0	44.6	46.9	66.0
Uniform (Matched)	2.2	47.6	48.1	62.0
*VSLS** (Top-N, Matched)	2.2	**50.3**	48.1	66.0
AFP only (Ours)	2.2	45.5	51.5	56.0
AFP + Graph (Ours)	**2.2**	47.3	**53.5**	**84.0**
Model: Qwen2.5-VL-7B-Instruct				
*VSLS** Baseline	8.0	**45.8**	**49.2**	54.0
Uniform (Matched)	2.2	38.1	41.9	**70.0**
*VSLS** (Top-N, Matched)	2.2	44.0	47.3	56.0
AFP only (Ours)	2.2	40.2	45.0	68.0
AFP + Graph (Ours)	**2.2**	42.6	45.4	62.0
Model: LLaVA-Video-7B-Qwen2				
*VSLS** Baseline	8.0	42.6	43.5	50.0
Uniform (Matched)	2.2	39.3	46.2	42.0
*VSLS** (Top-N, Matched)	2.2	40.5	42.7	46.0
AFP only (Ours)	2.2	41.1	44.6	44.0
AFP + Graph (Ours)	**2.2**	**45.5**	**49.6**	**72.0**
Evaluation from Top 32 Keyframes				
Model: GPT-4o				
*VSLS** Baseline	32.0	46.1	45.0	76.0
Uniform (Matched)	4.2	47.9	48.8	64.0
*VSLS** (Top-N, Matched)	4.2	**50.9**	**53.8**	64.0
AFP only (Ours)	4.2	45.5	47.7	66.0
AFP + Graph (Ours)	**4.2**	49.4	51.5	**80.0**
Model: Qwen2.5-VL-7B-Instruct				
*VSLS** Baseline	32.0	38.7	42.3	54.0
Uniform (Matched)	4.2	45.8	45.0	68.0
*VSLS** (Top-N, Matched)	4.2	**50.3**	**53.1**	66.0
AFP only (Ours)	4.2	40.2	41.9	60.0
AFP + Graph (Ours)	**4.2**	42.9	46.9	**66.0**
Model: LLaVA-Video-7B-Qwen2				
*VSLS** Baseline	32.0	41.7	48.1	54.0
Uniform (Matched)	4.2	38.1	47.3	52.0
*VSLS** (Top-N, Matched)	4.2	40.2	45.8	52.0
AFP only (Ours)	4.2	41.7	43.5	50.0
AFP + Graph (Ours)	**4.2**	**45.2**	**50.0**	**62.0**

(a) Results on the LONGVIDEOBENCH dataset.

Method	Avg. Fr.	Long(%)	Med.(%)	Short(%)
Evaluation from Top 8 Keyframes				
Model: GPT-4o				
*VSLS** Baseline	8.0	51.7	52.4	56.5
Uniform (Matched)	2.1	52.4	53.4	58.6
*VSLS** (Top-N, Matched)	2.1	50.5	49.8	56.0
AFP only (Ours)	2.1	53.1	51.8	57.3
AFP + Graph (Ours)	**2.1**	**53.5**	**56.5**	**63.8**
Model: Qwen2.5-VL-7B-Instruct				
*VSLS** Baseline	8.0	36.6	39.1	41.1
Uniform (Matched)	2.1	38.5	40.2	42.9
*VSLS** (Top-N, Matched)	2.1	38.1	37.9	41.4
AFP only (Ours)	2.1	39.2	38.8	42.4
AFP + Graph (Ours)	**2.1**	**38.1**	**40.2**	**50.1**
Model: LLaVA-Video-7B-Qwen2				
*VSLS** Baseline	8.0	38.7	41.5	41.3
Uniform (Matched)	2.1	40.6	41.2	38.3
*VSLS** (Top-N, Matched)	2.1	39.2	40.2	38.5
AFP only (Ours)	2.1	38.9	41.0	40.3
AFP + Graph (Ours)	**2.1**	**44.2**	**47.9**	**54.5**
Evaluation from Top 32 Keyframes				
Model: GPT-4o				
*VSLS** Baseline	32.0	51.7	52.4	56.5
Uniform (Matched)	4.3	52.3	52.6	57.1
*VSLS** (Top-N, Matched)	4.3	49.7	50.2	54.1
AFP only (Ours)	4.3	52.5	52.5	57.7
AFP + Graph (Ours)	**4.3**	**53.0**	**56.4**	**63.4**
Model: Qwen2.5-VL-7B-Instruct				
*VSLS** Baseline	32.0	37.9	39.1	**55.8**
Uniform (Matched)	4.3	39.1	43.5	43.6
*VSLS** (Top-N, Matched)	4.3	37.9	40.2	42.5
AFP only (Ours)	4.3	38.3	39.8	42.5
AFP + Graph (Ours)	**4.3**	**39.9**	**44.0**	51.8
Model: LLaVA-Video-7B-Qwen2				
*VSLS** Baseline	32.0	37.4	40.6	42.1
Uniform (Matched)	4.3	39.6	40.6	41.7
*VSLS** (Top-N, Matched)	4.3	38.1	41.1	41.2
AFP only (Ours)	4.3	39.8	40.9	41.5
AFP + Graph (Ours)	**4.3**	**43.8**	**47.9**	**56.7**

(b) Results on the VIDEOMME dataset.

our *VSLS** study, with detailed results presented in Table 2.

Analysis from Tabular Data. The results in Table 2 further solidify our framework's universal applicability. Our full method consistently offers a superior cost-utility trade-off, with gains being most pronounced on open-source models. For instance, when refining the *T** output for LLaVA-Video-7B on VIDEOMME (from 32 frames), our method elevates the accuracy on short videos from 41.6% to **55.0%**, a **+13.4 point** improvement. The quality of frames selected by '***AFP*** only' is demonstrably high. Under the matched-budget setting for GPT-4o on LONGVIDEOBENCH (from 8 frames), '***AFP*** only' (47.3% on Long videos) outperforms both 'Uniform (Matched)' (45.5%) and '*T** (Top-N, Matched)' (46.4%). In addition, the semantic graph provides a significant performance boost, reinforcing our central thesis about the synergy between our two components.

Visual Analysis of Efficiency-Performance. This trade-off is powerfully visualized in the efficiency-performance plot in Figure 2, which highlights the GPT-4o on VIDEOMME (from Top-32) scenario. Similar to the *VSLS** case, the *T** resides in the high-cost, bottom-right region. Within the highly competitive low-cost zone, our full ***AFP + Graph*** method once again establishes itself as the superior choice, occupying the highest point in the desirable

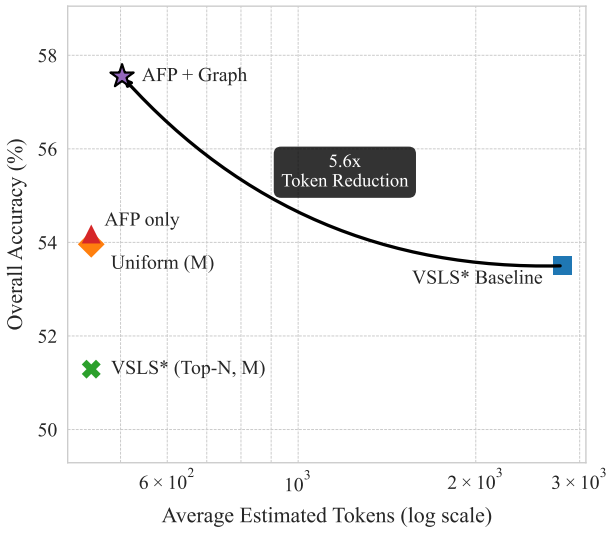

Figure 1. **Efficiency-Performance Trade-off on VIDEOMME with *VSLS**.** This plot visualizes the results for GPT-4o when refining the Top-32 output from *VSLS**. The x-axis (log scale) represents token cost, and the y-axis represents weighted average accuracy. The top-left region is optimal.

top-left quadrant. It significantly outperforms the '*T** (Top-N, Matched)' baseline (56.91% vs. 51.20%) and shows a clear improvement over the '***AFP*** only' approach (53.66%). This parallel success story proves that our framework's effectiveness is not tied to a specific selector but stems from its fundamental principles of intelligent pruning and semantic compensation.

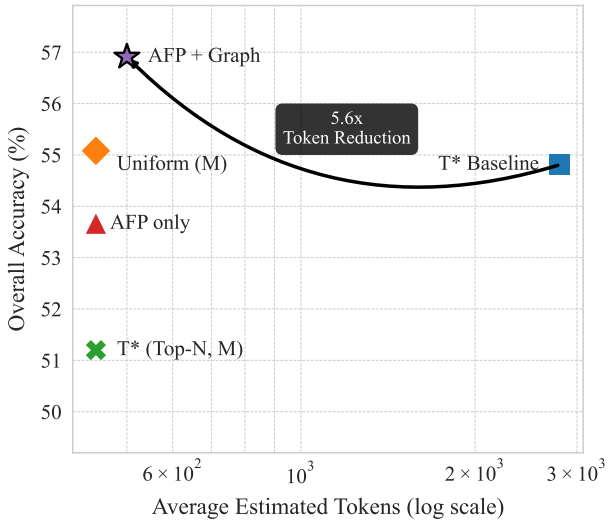

Figure 2. **Efficiency-Performance Trade-off on VIDEOMME with *T**.** This plot visualizes the results for GPT-4o when refining the Top-32 output from *T**. The top-left region is optimal.

C. Additional Visual Evidence for the Prevalence of 'visual echoes'

To further demonstrate that 'visual echoes' are a pervasive challenge not limited to a specific video type, this section provides additional case studies. While our main paper (Figure 2) uses a simple animation to clearly illustrate the issue, here we show that the same pattern of severe frame redundancy occurs in more complex scenarios, including a scientific animation (Figure 3) and a real-world video with significant visual detail (Figure 4).

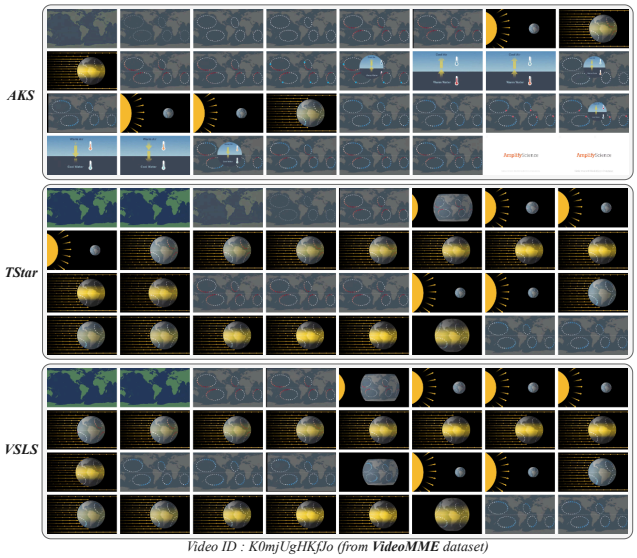

Figure 3. **Case Study 2: 'Visual Echoes' in a Scientific Animation Video.** This example (Video ID: K0mjUgHKfJo) shows that even in educational content with distinct conceptual phases, all three selectors produce highly repetitive frames of the Earth's rotation and its position relative to the sun.

D. Additional Details on Experimental Setup

D.1. Dataset Subsets for Fair Evaluation

To ensure maximum transparency, reproducibility, and fairness, our experiments were conducted on consistent, curated subsets of the official LONGVIDEOBENCH and VIDEOMME benchmarks. Our initial setup involved baselines like *VSLS**, whose original implementation filtered out videos that were unreadable or caused processing errors. This resulted in a clean, stable subset of 646 samples for LONGVIDEOBENCH and 2657 samples for VIDEOMME. To maintain a strict apples-to-apples comparison, we adopted these exact same subsets for **all** methods reported in our paper, including our main baseline *AKS** and our proposed ***AFP + Graph*** framework. This approach ensures that all performance differences are attributable solely to the methods themselves, not variations in data quality.

Table 2. **Generalizability study on the *T** selector.** This table shows the complete component-wise ablation results of applying our framework to the output of *T** on both datasets. **Bold** indicates the best performance among the four matched-budget strategies for each accuracy column.

Method	Avg. Fr.	Long(%)	Med.(%)	Short(%)
Evaluation from Top 8 Keyframes				
Model: GPT-4o				
*T** Baseline	8.0	44.3	46.2	66.0
Uniform (Matched)	2.2	45.5	48.8	64.0
*T** (Top-N, Matched)	2.2	46.4	**53.1**	64.0
AFP only (Ours)	2.2	**47.3**	48.8	68.0
AFP + Graph (Ours)	**2.2**	46.4	52.7	**80.0**
Model: Qwen2.5-VL-7B-Instruct				
*T** Baseline	8.0	42.0	47.7	54.0
Uniform (Matched)	2.2	35.7	38.5	**70.0**
*T** (Top-N, Matched)	2.2	**43.8**	**48.1**	56.0
AFP only (Ours)	2.2	38.7	41.9	66.0
AFP + Graph (Ours)	**2.2**	42.0	45.4	66.0
Model: LLaVA-Video-7B-Qwen2				
*T** Baseline	8.0	42.0	47.7	44.0
Uniform (Matched)	2.2	40.8	46.5	48.0
*T** (Top-N, Matched)	2.2	39.3	45.4	42.0
AFP only (Ours)	2.2	39.6	49.2	52.0
AFP + Graph (Ours)	**2.2**	**45.5**	**51.5**	**62.0**
Evaluation from Top 32 Keyframes				
Model: GPT-4o				
*T** Baseline	32.0	53.1	48.8	74.3
Uniform (Matched)	4.2	46.4	48.8	62.0
*T** (Top-N, Matched)	4.2	**49.1**	**55.0**	62.0
AFP only (Ours)	4.2	45.8	48.8	70.0
AFP + Graph (Ours)	**4.2**	48.5	53.1	**80.0**
Model: Qwen2.5-VL-7B-Instruct				
*T** Baseline	32.0	38.7	41.9	40.0
Uniform (Matched)	4.2	38.1	43.5	64.0
*T** (Top-N, Matched)	4.2	41.4	**48.5**	**68.0**
AFP only (Ours)	4.2	41.1	41.2	62.0
AFP + Graph (Ours)	**4.2**	**42.6**	43.1	**68.0**
Model: LLaVA-Video-7B-Qwen2				
*T** Baseline	32.0	40.2	44.2	50.0
Uniform (Matched)	4.2	38.7	46.9	48.0
*T** (Top-N, Matched)	4.2	**42.3**	46.9	48.0
AFP only (Ours)	4.2	41.1	49.2	44.0
AFP + Graph (Ours)	**4.2**	42.0	**52.7**	**62.0**

(a) Results on the LONGVIDEOBENCH dataset.

Method	Avg. Fr.	Long(%)	Med.(%)	Short(%)
Evaluation from Top 8 Keyframes				
Model: GPT-4o				
*T** Baseline	8.0	51.3	51.5	55.2
Uniform (Matched)	2.1	50.1	54.2	58.6
*T** (Top-N, Matched)	2.1	49.9	51.7	56.2
AFP only (Ours)	2.1	52.2	51.1	56.7
AFP + Graph (Ours)	**2.1**	**52.3**	**55.3**	**66.1**
Model: Qwen2.5-VL-7B-Instruct				
*T** Baseline	8.0	**38.2**	38.3	43.1
Uniform (Matched)	2.1	37.7	39.4	42.5
*T** (Top-N, Matched)	2.1	36.1	38.5	41.4
AFP only (Ours)	2.1	38.0	39.3	41.5
AFP + Graph (Ours)	**2.1**	36.2	**40.4**	**49.5**
Model: LLaVA-Video-7B-Qwen2				
*T** Baseline	8.0	40.4	41.3	41.6
Uniform (Matched)	2.1	38.8	41.6	39.2
*T** (Top-N, Matched)	2.1	39.3	41.0	38.8
AFP only (Ours)	2.1	37.8	40.4	40.0
AFP + Graph (Ours)	**2.1**	**42.3**	**46.2**	**56.1**
Evaluation from Top 32 Keyframes				
Model: GPT-4o				
*T** Baseline	32.0	52.6	54.4	57.5
Uniform (Matched)	4.1	52.4	55.2	57.7
*T** (Top-N, Matched)	4.1	49.8	49.4	54.6
AFP only (Ours)	4.1	51.8	53.6	55.7
AFP + Graph (Ours)	**4.1**	**52.6**	**55.3**	**63.1**
Model: Qwen2.5-VL-7B-Instruct				
*T** Baseline	32.0	36.9	40.2	42.2
Uniform (Matched)	4.1	**39.9**	**43.1**	43.5
*T** (Top-N, Matched)	4.1	37.9	40.1	43.4
AFP only (Ours)	4.1	39.6	40.4	44.6
AFP + Graph (Ours)	**4.1**	39.4	40.7	**52.3**
Model: LLaVA-Video-7B-Qwen2				
*T** Baseline	32.0	38.2	42.2	41.6
Uniform (Matched)	4.1	38.7	42.9	41.3
*T** (Top-N, Matched)	4.1	40.2	39.3	40.3
AFP only (Ours)	4.1	38.4	41.7	39.3
AFP + Graph (Ours)	**4.1**	**41.7**	**47.1**	**55.0**

(b) Results on the VIDEOMME dataset.

E. Note on the *AKS** Baseline Adaptation

In our main experiments, we use *AKS** as our primary sophisticated baseline. It is important to clarify the distinction between the original *AKS* [4] framework and our adapted *AKS** version, and to justify this choice.

Rationale for Adaptation. Unlike other selectors such as *T** and *VSLS* that employ a direct Top-K selection, the original *AKS* utilizes a more sophisticated **dynamic sampling algorithm**. This algorithm recursively allocates frame budgets to temporal segments based on score distributions. While innovative, this dynamic approach does not guarantee a fixed number of output frames (e.g., it may return 6 frames when a budget of 8 is requested), making rigorous, fixed-budget comparisons challenging. To ensure a fair and reproducible evaluation protocol, we therefore modify its final stage to a standard Top-K selection. Crucially, this choice is empirically supported by the ablation studies within the original *AKS* paper itself. Their experiments show that a standard Top-K selection (denoted 'TOP') is the best performing alternative, achieving results highly competitive with their full adaptive method ('ADA') (e.g., 62.4% vs. 62.7% on LONG

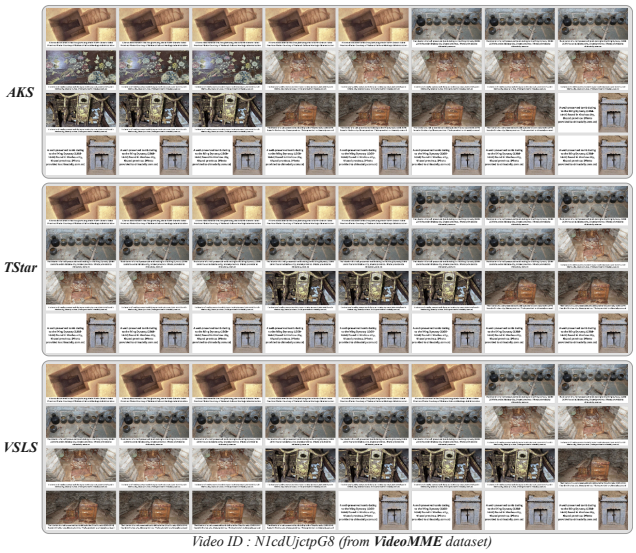

Figure 4. **Case Study 3: 'Visual Echoes' in a Real-World Video.** This example (Video ID: N1cdUjctpG8) demonstrates the prevalence of redundancy in complex, real-world footage. Despite challenging lighting and rich details, all selectors still select numerous near-identical shots of the same artifacts.

VIDEOBENCH). This confirms that our *AKS** adaptation, based on Top-K selection, serves as a powerful and faithful representation of the original method's capabilities for the purpose of a strong baseline comparison.

Validity of the 'Visual Echoes' Motivation. This modification does not weaken our core claim regarding the prevalence of 'visual echoes'. The problem of frame redundancy is not an artifact of simple Top-K selection. As visually demonstrated in our qualitative examples (Figure 3, Figure 4 and Section C), 'visual echoes' are a fundamental challenge rooted in the temporal nature of video. The original, more complex dynamic sampling of *AKS* also suffers from this issue because if a temporal segment with high visual similarity receives a high score, the algorithm will still allocate budget to select multiple, similar frames from within that segment. This reinforces our central argument: regardless of the final selection strategy, the 'visual echoes' problem persists, universally motivating the need for a dedicated refinement layer like our ***AFP*** framework.

F. Further Implementation Details

This section provides in-depth details about our implementation to ensure full reproducibility of our work.

F.1. *AFP* Algorithm Details

Our ***Adaptive Frame-Pruning (AFP)*** algorithm is implemented in Python utilizing the scikit-learn [3] library. The following implementation specifics directly complement the Methodology section of the main paper.

Clustering Parameters. For the core clustering step, we employ the `AgglomerativeClustering` class from scikit-learn. The `linkage` parameter is set to '`average`', which means the distance between two clusters is defined as the average of the distances between all pairs of samples, with one sample from each cluster.

Small Cluster Refinement. To enhance the robustness of our clustering, we implement a refinement step (`refine_clusters` in our script). After the initial clustering, any cluster containing fewer than two frames is considered unstable. Such singleton clusters are merged into their nearest neighboring cluster, where proximity is determined by the average visual cosine distance between the frames of the two clusters. This prevents trivial clusters and ensures a more meaningful grouping.

Adaptive Threshold Calculation. As mentioned in the main paper, the `distance_threshold` for clustering is determined adaptively. Specifically, after calculating all pairwise visual cosine distances, we fit a Gaussian Kernel Density Estimator (KDE) to their distribution. The threshold is then calculated as $\tau = p + 0.15$, where p is the distance value corresponding to the peak of the density function. The constant offset of 0.15 is an empirically chosen value designed to prevent the clustering from being overly conservative (i.e., creating too many small clusters) and to encourage the merging of closely related 'visual echoes'.

F.2. Prompt Structure and MLLM Inference

The structure of the prompts sent to the MLLM is critical for achieving consistent and reproducible results. In this section, we detail the two primary prompt templates used in our framework: one for our versatile semantic graph generation (Case 2), and one for the final downstream Video-QA task.

F.3. Prompt Template for Semantic Graph

Our framework generates a concise, structured block of text, termed the **Textual Semantic Graph**, which is programmatically injected into the final prompt for the downstream MLLM. This graph provides high-level semantic context, summarizing the key entities and their inter-relationships as inferred from the query. Figure 5 illustrates the structural template of this generated text block.

Template for Textualized Semantic Graph

[Here is an additional semantic graph context for this Video-QA]
Objects in video context: <Objects>
Relationships between objects: <Relations>

Figure 5. **The structural template for the textualized semantic graph.** This block is programmatically generated based on the query and inserted into the final downstream prompt. It serves to provide the MLLM with high-level, pre-digested semantic context to guide its reasoning over the visual frames.

Prompt for Semantic Graph Generation. For our low-cost, universally applicable semantic graph generation strategy (Case 2), we employ a sophisticated prompt designed to elicit rich, structured information from the MLLM based solely on textual input. As detailed in Figure 6, the prompt is structured using several advanced prompting techniques:

- **Persona Prompting**: We assign the MLLM the role of an "expert visual scene analyst" to activate its advanced reasoning capabilities.
- **Chain-of-Thought Instructions**: The prompt provides a clear, step-by-step thinking process (Key Entity Identification, Contextual Cue Identification, Relationship Extraction) for the MLLM to follow.
- **Open-ended yet Structured Relation Extraction**: Crucially, our prompt encourages flexibility. It explicitly defines two categories of relations—Concrete (e.g., verbs, prepositions) and Abstract Logical (e.g., 'spatial', 'causal')—and provides clear examples and definitions for each. This "dual-track" design allows for the capture of a much richer and more nuanced set of semantic connections while maintaining a structured output.
- **Few-Shot Example**: A comprehensive example is provided to demonstrate the desired input-output format and the expected quality of the extracted graph.

The full prompt is shown in Figure 6.

Prompt for Downstream Video-QA. For the final question-answering stage, we utilize a clean, direct, and robust prompt template, detailed in Figure 7. This prompt is designed for maximum compatibility across different MLLMs. It clearly presents the visual evidence (via '<image>' placeholders), followed by the optional textual semantic graph, and finally the question and multiple-choice options. The instructions are direct and unambiguous, strictly constraining the model to output a single uppercase letter corresponding to its chosen answer. This minimalist design minimizes prompt engineering sensitivity and ensures that performance differences are attributable to the quality of the visual and semantic inputs, rather than variations in prompt interpretation.

G. Hyperparameter and Component Analysis

G.1. Ablation Study on Prompt Formulation

The formulation of the prompt, particularly how the semantic graph is textualized and how instructions are given to the MLLM, can significantly impact performance. To determine the optimal prompt structure for our main experiments, we conducted an ablation study on several variants. All experiments were performed on the LongVideoBench dataset with the GPT-4o model, starting from Top-8 keyframes.

G.1.1. Prompt Component Variants

We tested combinations of two `graph_context` formats and two `system_prompt` formats:

Graph Context Formats.

- **Concise Triplet (G1):** Our final choice. This format is simple and structured, directly presenting nodes and raw triplet relationships (e.g., `(object1, relation, object2)`).
- **Verbose Natural Language (G2/G3):** These formats attempt to convert the graph into more human-like sentences (e.g., "object1 appears with object2").

System Prompt Formats.

- **Direct Prompt (P1):** Our final choice. This is a concise, direct instruction for the QA task.
- **Instructional Prompt (P2):** This is a more verbose prompt that assigns an "expert" persona to the MLLM and provides detailed step-by-step guidelines.

G.1.2. Results and Analysis

The results of our prompt ablation are presented in Table 3.

A "Less is More" Phenomenon in Prompting. Interestingly, this ablation study reveals a microcosm of our paper's central 'Less is More' theme, but applied to the prompt engineering space. As shown in Table 3, while more verbose and complex combinations like (G1, P2) or (G3, P2) can achieve marginally higher peak accuracy on specific video lengths (e.g., Short videos), this gain is inconsistent and comes at the cost of stability and increased token counts. The simpler, more direct (G1, P1) format demonstrates a superior balance across all metrics. This suggests that MLLMs can suffer from a form of **"instructional noise"** or "prompt dilution," where overly elaborate instructions can sometimes obscure the core task, mirroring how excessive visual frames can cause "context dilution." Based on this finding, our analysis led us to select the combination of **Concise Triplet graph (G1)** and **Direct Prompt (P1)** for all main experiments, guided by the following key reasons:

- **Robust and Stable Performance:** While other combinations, such as (G1, P2), achieve marginally higher accuracy on certain video lengths, the (G1, P1) combination demonstrated the most stable and consistent high performance across all our preliminary and main experiments. As noted in our experimental logs, other verbose combinations frequently yielded unexpectedly low results in some runs, indicating a lack of robustness. The (G1, P1) setting, in contrast, reliably produced strong results.
- **Token Efficiency:** The G1 and P1 formats are significantly more concise than their verbose counterparts. This results in a lower token count for each query, which directly aligns with our paper's core objective of maximizing token efficiency.
- **Simplicity and Generalizability:** The direct, structured format of (G1, P1) provides clear instructions to the MLLM without excessive "prompt engineering." We believe this simpler format is more likely to generalize well across different MLLMs, as it relies on fundamental instruction-following capabilities rather than sensitivity to nuanced persona-based instructions.

The prompt template for LLM-based semantic graph generation

Here is a multiple-choice question based on a video. You are an expert visual scene analyst.
Your task is to deconstruct the question and multiple-choice options to identify key entities and their relationships.

Input:
- Question: {question_text}
- Options: {options_text}

Instructions:
1. Key Entity Identification: From the Question and Options, extract 3-5 core entities (e.g., "person", "car") that are central to the query.
2. Contextual Cue Identification: Extract 2-4 additional scene elements that provide context or help locate the key entities.
3. Relationship Extraction: Infer all plausible relationships between the extracted entities.
   A relationship MUST be a triplet `(Entity1; Relation; Entity2)`. The "Relation" can be one of two types:

   A) Concrete Relation: A concise verb or prepositional phrase describing a specific, physical action or state.
      Examples: `wearing`, `holding`, `near`, `on top of`, `looking at`.
   B) Abstract Logical Relation: A word describing a fundamental logical connection between entities.

You are encouraged to infer such relations when applicable.
Examples & Definitions:
   a.`spatial`: Describes co-occurrence or positional arrangement (e.g., two objects in the same place).
   b.`attribute`: Describes a property of an entity (e.g., a person's clothing color).
   c.`temporal`: Describes a sequence in time (e.g., one event happens after another).
   d.`causal`: Describes a cause-and-effect link.
   e.`functional`: Describes the purpose or use of an object (e.g., a key is for a door).
Condition: Both entities in a triplet must be from the entities you extracted in steps 1 and 2.

Output Rules:
1. Provide your response in three separate lines starting with the exact prefixes: `Key Objects:`, `Cue Objects:`, and `Rel:`.
2. In the `Key Objects:` and `Cue Objects:` lines, separate items with a comma.
3. In the `Rel:` line, separate each complete triplet `(...)` with a comma. Inside each triplet, separate the three elements with a semicolon.
4. Never use markdown or natural language explanations in your output.

Example:
Input:
- Question: In a kitchen, after the woman in a red shirt opens the fridge with a key, what does she take out?
- Options: A) A bottle of milk, B) A green apple

Response:
Key Objects: woman, red shirt, fridge, key, milk, green apple
Cue Objects: kitchen
Rel: (woman; attribute; red shirt); (woman; in; kitchen); (key; functional; fridge); (woman; opens; fridge); (fridge; temporal; milk)

Figure 6. **The full prompt template for LLM-based semantic graph generation.** This prompt leverages persona, chain-of-thought, and open-ended relation extraction to guide the LLM in deconstructing the query.

The prompt template for downstream Video-QA

Select the best answer to the following multiple-choice question based on the video.
<image>
<image>
...
[Here is an additional semantic graph context for this Video-QA]
Objects in video context: <Objects>
Relationships between objects: <Relations>
Question: <Question>
Options: <Options>
Answer with the option's letter from the given choices directly.
Your response format should be strictly an upper case letter A,B,C,D or E.

Figure 7. **The full MLLM prompt template used for the final Video-QA task.** It integrates image placeholders, the optional semantic graph, and the QA content into a single, direct query.

In conclusion, our chosen (G1, P1) prompt formulation represents the best trade-off between performance, stability, and token efficiency, making it the most suitable choice for our proposed method.

G.2. Analysis of Hyperparameters

This section provides the detailed data and analysis supporting our hyperparameter selection, ensuring the reproducibility of our work. Our method's core hyperparameters, the feature fusion ratio α and the distance metric weight β, were determined through a systematic sensitivity analysis. This analysis reveals a clear cost-utility trade-off, allowing our framework to be tuned to prioritize either performance (utility) or efficiency (cost). All experiments were performed on the LongVideoBench dataset using the Qwen2.5-VL-7B-Instruct model, starting from *VSLS** Top-8 keyframes.

Table 3. **Ablation study on prompt formulation. Results are averaged over multiple runs.** Our chosen combination (G1, P1) is highlighted in **bold**.

Graph Format	Prompt Format	Avg. Accuracy (%)		
		Long	Medium	Short
G1 (Ours)	**P1 (Ours)**	**47.0**	52.5	69.0
G2	P1	46.1	51.7	66.4
G1	P2	46.7	**52.9**	**70.7**
G2	P2	45.9	49.9	69.0
G3	P2	44.9	51.2	**70.7**

The results, averaged over multiple runs, are visualized in Figure 8 and detailed with exact values in Table 4.

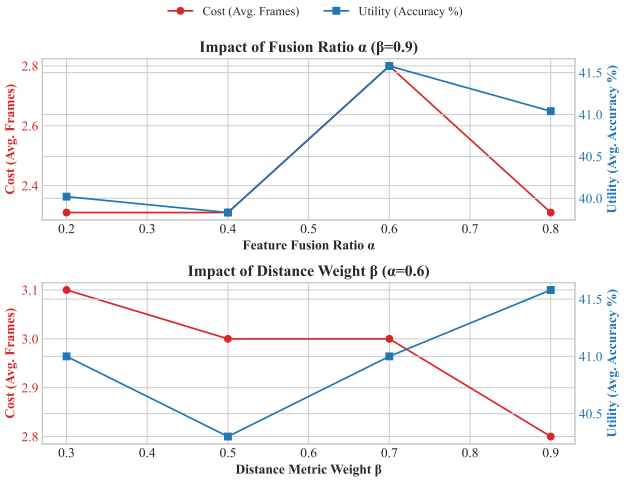

Figure 8. **Cost-Utility trade-off analysis for hyperparameters α and β. Cost** (left y-axis, red) is measured by the average number of output frames, while **Utility** (right y-axis, blue) is measured by the average QA accuracy on a subset of LONGVIDEOBENCH. The plots demonstrate that our parameters are tunable knobs: users can adjust them to prioritize either higher accuracy (e.g., lower β) or maximum efficiency (e.g., higher β). Our default choice ('$\alpha = 0.6, \beta = 0.9$') represents a strong balance.

Table 4. **Detailed sensitivity analysis of hyperparameters α and β.** Results are averaged over multiple runs on the LONG VIDEOBENCH dataset using the Qwen2.5-VL-7B-Instruct model (from Top-8). Our chosen parameters ('$\alpha = 0.6, \beta = 0.9$') are highlighted in **bold**.

Hyperparameter Setting	Avg. Accuracy (%)	Avg. Frames
Analysis of Feature Fusion Ratio α (fixing $\beta = 0.9$)		
$\alpha = 0.2$	40.02	**2.31**
$\alpha = 0.4$	39.83	**2.31**
$\alpha = 0.6$ (Ours)	**41.58**	2.80
$\alpha = 0.8$	41.04	**2.31**
Analysis of Distance Metric Weight β (fixing $\alpha = 0.6$)		
$\beta = 0.3$	41.00	3.10
$\beta = 0.5$	40.30	3.00
$\beta = 0.7$	41.00	3.00
$\beta = 0.9$ (Ours)	**41.58**	**2.80**

Analysis and Selection Strategy. As visualized in Figure 8 and detailed in Table 4, our analysis reveals a clear cost-utility trade-off. For the feature fusion ratio α (top plot), the results show that $\alpha = 0.6$ achieves the highest average accuracy of 41.58%, suggesting an optimal balance between visual and semantic features, albeit at a slightly higher frame cost. For the distance metric weight β (bottom plot), a higher value (e.g., 0.9) leads to more aggressive clustering (fewer frames, as shown by the red line) and, in this case, also yields the highest accuracy. This demonstrates that the hyperparameters of ***AFP*** are not arbitrary fixed values, but rather tunable "knobs" that control the cost-utility balance. Based on this systematic analysis, we selected the configuration of $\alpha = 0.6$ and $\beta = 0.9$ for all main experiments, as it consistently provided the best overall performance and a favorable efficiency-accuracy balance in our development set.

G.3. Representative Frame Selection Strategy

A crucial step in our ***AFP*** algorithm is selecting a single representative frame from each generated cluster. To determine the most effective and robust approach, we conducted a comprehensive ablation study comparing three distinct strategies. The experiments were performed on the LONG VIDEOBENCH dataset with the GPT-4o model, starting from Top-8 keyframes provided by *VSLS**, and the results reported are averaged over multiple runs to ensure reliability. This analysis focuses on the '***AFP*** only' setting (without the semantic graph) to purely evaluate the quality of the selected visual information.

- **Score-based (Highest *VSLS** Score):** This strategy leverages external information by selecting the frame from each cluster that has the highest initial relevance score assigned by the upstream *VSLS** selector.
- **Centroid-based (Visual Centroid):** This strategy is self-contained. It selects the "centroid frame", the frame with the minimum average feature distance to all other frames within the same cluster, making it the most visually representative.
- **Relevance-based (Highest Query Similarity):** This strategy introduces task-specific guidance by selecting the frame with the highest CLIP similarity score to the 'Question + Options' text prompt.

As shown in Table 5, the averaged results over multiple runs confirm that the **Centroid-based strategy** holds a clear advantage. It achieves the highest overall accuracy (48.53%), outperforming both the Score-based and Relevance-based approaches. The superiority of the Centroid strategy is theoretically sound, especially in the absence of a semantic graph. By selecting the frame closest to the cluster's feature-space center, it guarantees the most visually representative frame is chosen, which is the most robust way to minimize information loss during the pruning process. This self-contained logic is arguably more generalizable than relying on external scores (from *VSLS** or CLIP), which might be noisy or biased. Based on these comprehensive findings, we adopted the **Centroid-based strategy** for all main experiments reported in this paper.

Table 5. **Ablation study on representative frame selection strategies.** The experiment was conducted in the '***AFP*** only' setting to purely evaluate the quality of the selected visual information. Results are averaged over multiple runs. The Centroid-based strategy demonstrates the best performance.

Experimental Settings				Averaged Results	
Dataset	**Model**	**Frame Source**	**Selection Strategy**	**Avg. Acc (%)**	**Avg. Frames**
LongVideoBench	GPT-4o	*VSLS** (Top-8)	Score-based	47.14	2.1
			Centroid-based	**48.53**	**2.1**
			Relevance-based	46.67	2.2

H. Robustness on Event Dense Short Videos

Our framework primarily targets long videos where sparse sampling inevitably yields severe visual redundancy. To rigorously evaluate its robustness in out of distribution scenarios, we conducted a stress test on the Video-HOLMES [1] benchmark. This dataset consists of short, event dense videos (typically 1 to 5 minutes) where a 32 frame extraction represents dense sampling with minimal inherent redundancy. We evaluated the Qwen2.5-VL-7B-Instruct model on a stratified subset of 900 videos. The results in Table 6 demonstrate that our method remains highly robust. Even with minimal redundancy to prune, the ***AFP*** module successfully filters out subtle noise, improving accuracy by 2.03% while reducing the frame count to 4.1. The integration of the semantic graph further elevates the accuracy to 30.14%, achieving a total absolute improvement of 3.47% over the *AKS** baseline. This confirms that our framework safely compresses visual information and provides a reliable semantic reasoning scaffold even in highly dynamic contexts.

Table 6. **Stress test on the Video-HOLMES benchmark.** Our method maintains robustness and improves accuracy on event dense short videos.

Model	Method	Avg. Fr.	Acc (%)	vs. Base
Qwen2.5-VL	*AKS** (Base)	32.0	26.67	/
	AFP only	4.1	28.70	+2.03
	AFP + Graph	**4.1**	**30.14**	+3.47

I. Reliability of the Semantic Graph

A potential concern regarding the use of an LLM generated semantic graph is the introduction of textual hallucinations. To quantitatively assess the reliability of our generated context, we conducted a rigorous manual audit of 100 randomly sampled graphs from the VIDEOMME dataset. The audit revealed an exceptionally high degree of fidelity, achieving an entity extraction accuracy of 98% and a relationship extraction accuracy of 96%. Qualitative analysis showed that the minimal inaccuracies observed (approximately 4%) were predominantly relevance errors (e.g., extracting a peripheral background object) rather than factual fabrications. This high reliability is deliberately engineered into our framework: by constraining the LLM to perform a structural decomposition of the textual query rather than open ended generation, we strictly limit the generative space. Furthermore, the downstream MLLM utilizes the sampled visual frames as primary evidence, inherently filtering out minor semantic noise.

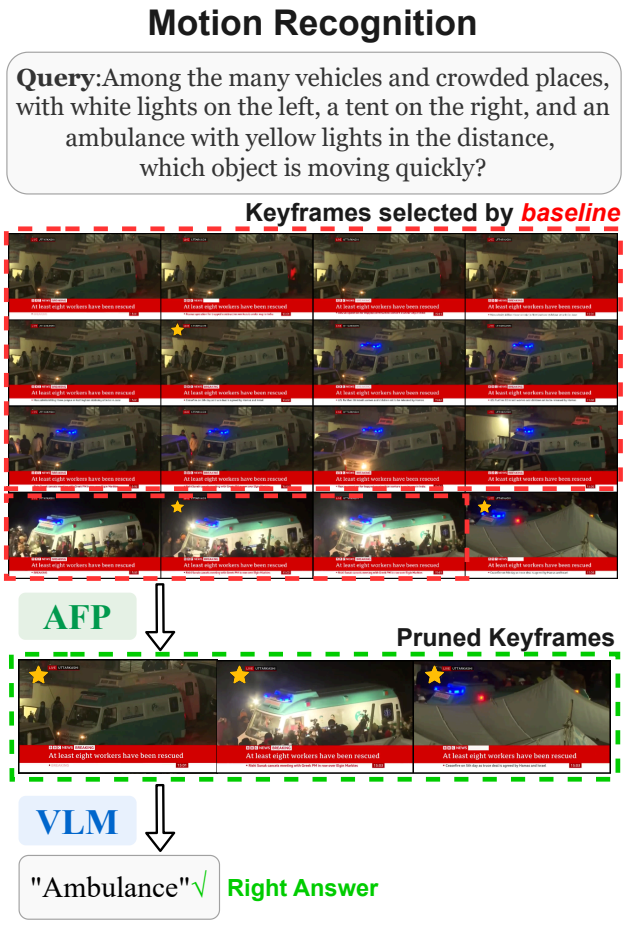

Figure 9. Qualitative example for a **motion recognition task** (Video ID: Z00vWImw1KQ). The query asks to identify a moving object. The baseline *VSLS** selects 16 frames with high redundancy. Our ***AFP*** algorithm prunes this set to just 3 frames that preserve the key motion cues, enabling the MLLM to derive the correct answer.

J. Qualitative Analysis and Limitations

J.1. Additional Qualitative Examples

To demonstrate the versatility and robustness of our ***Adaptive Frame-Pruning (AFP)*** method, this section provides qualitative examples from diverse Video-QA tasks. These cases illustrate how ***AFP*** effectively handles various forms of visual redundancy while preserving the crucial information needed to answer different types of queries. Figure 9 presents a challenging task requiring the identification of a fast-moving object based on a descriptive, distractor-rich query. The baseline *VSLS** method, aiming for comprehensive coverage, selects 16 keyframes. However, these frames are plagued by severe 'visual echoes': the first twelve frames are nearly identical, capturing a static view of an ambulance. This flood of repetitive information risks overwhelming the MLLM. Our ***AFP*** algorithm excels in such scenarios. It correctly identifies and merges the redundant static and slightly shifted shots into distinct representative frames. Consequently, the input is drastically reduced to a highly efficient set of just 3 frames. By capturing the ambulance's initial state, a subtle movement, and the surrounding context, these frames preserve the critical temporal cues needed to infer motion. This compact visual input, augmented by the textual semantic graph, provides a sufficient, non-redundant context for the MLLM to correctly identify the "Ambulance".

J.2. Qualitative Analysis of Limitations

To provide a balanced perspective on our framework, this section analyzes its inherent limitations through two representative failure cases, which also highlight avenues for future research.

Loss of Fine-Grained Information during Pruning. The first failure case (Figure 10) highlights a trade-off inherent to our global feature-based pruning. While correctly identifying the high visual similarity between frames depicting "Ramapithecus" and "Ardipithecus Ramidus", our ***AFP*** algorithm merges them into a single cluster and inadvertently discards the frame containing the correct evidence. This reveals a primary limitation: a reliance on global features can be insensitive to subtle but semantically decisive local details, such as small pieces of text. This points towards a clear avenue for future improvement by integrating more localized feature extractors, such as Optical Character Recognition (OCR) models, into the clustering process.

Ceiling Effect from Upstream Selector. The second case (Figure 11) perfectly exemplifies our framework's role as a post-processing refinement layer. Here, the query requires a deep narrative understanding that is absent even in the initial 32 frames provided by the upstream selector. Our ***AFP*** module, while correctly pruning the insufficient input, cannot recover this missing information, leading to an unavoidable failure. This case demonstrates that our framework's performance is fundamentally capped by the quality of the initial keyframe set, as its role is to prune, not to discover. This limitation highlights a broader challenge for all sparse sampling methods when dealing with questions that demand global, narrative-level reasoning.

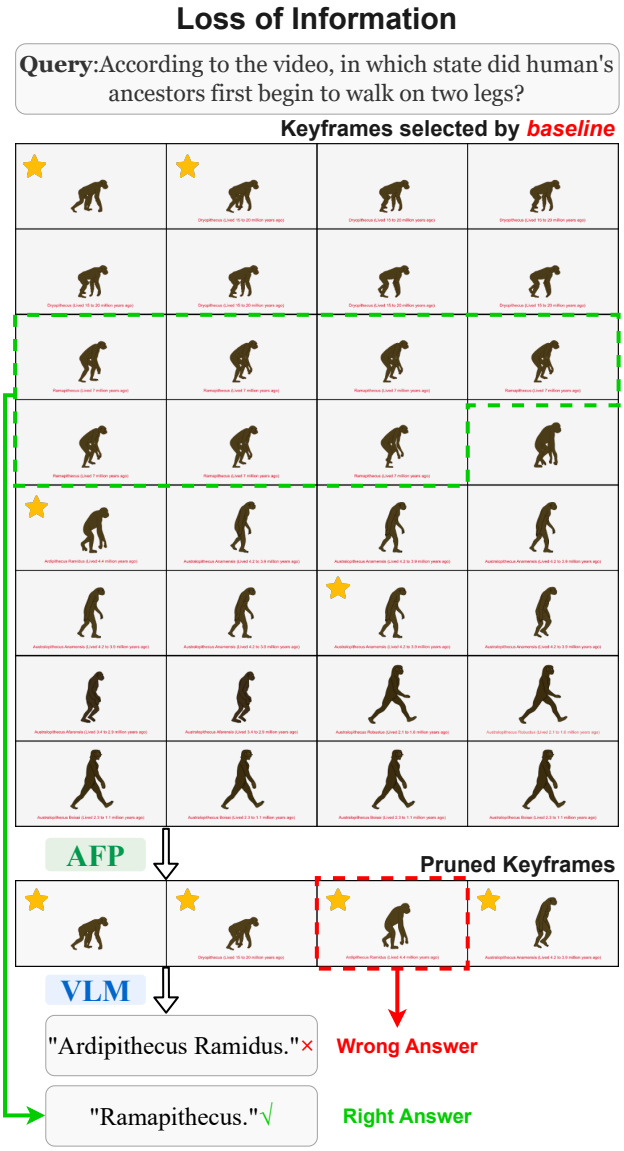

Figure 10. An analysis of a failure case caused by **information loss during pruning**. The initial keyframes from *VSLS** contain the evidence for the correct answer "Ramapithecus" (highlighted in green). However, due to high visual similarity with other frames, our ***AFP*** algorithm incorrectly prunes this crucial frame, leading the VLLM to a wrong answer based on the remaining evidence.

K. Future Directions

The preceding failure case analysis illuminates two promising avenues for future research. First, to address the loss

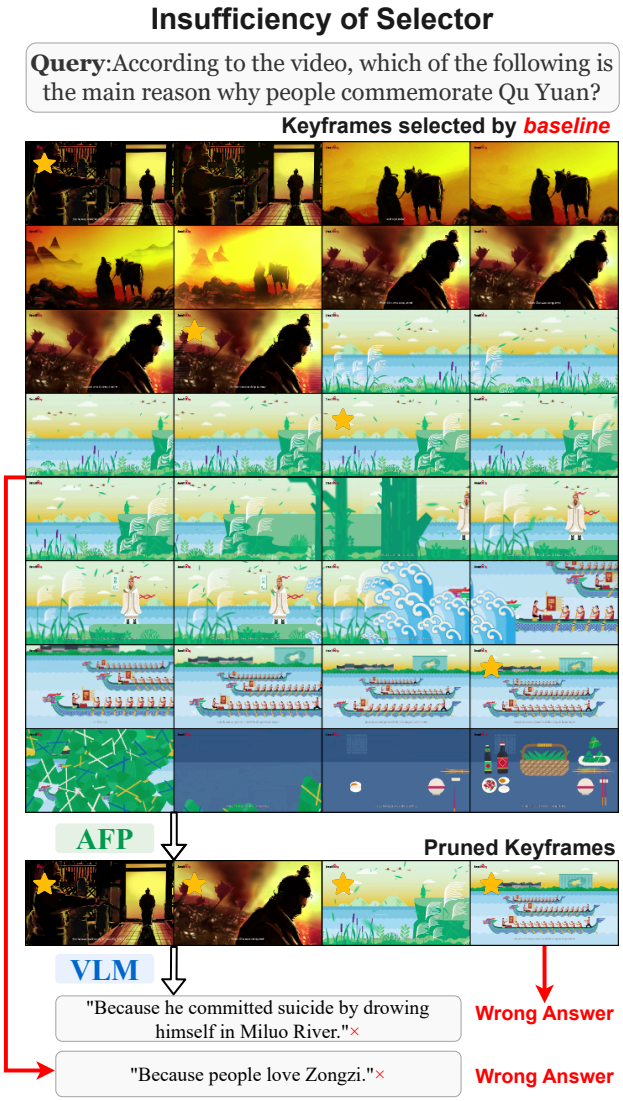

Figure 11. An analysis of a failure case caused by the **insufficiency of the upstream selector**. The question requires a global understanding of a narrative. The initial 32 frames from *VSLS** are already insufficient to answer correctly. Our ***AFP*** method, while efficiently pruning the input, cannot recover the missing information, leading to an inevitable failure by the VLLM. This illustrates the performance ceiling imposed by the initial keyframe selection.

of fine-grained details, our framework's reliance on global visual features can be overcome by developing an **adaptive feature fusion mechanism**. This system would dynamically switch to more **localized, patch-based representations** or integrate Optical Character Recognition (OCR) models to re-evaluate clusters where subtle details are semantically decisive. Ultimately, both limitations point towards a unified solution: the development of a **'Multimodal Semantic Graph'**. By encapsulating not only textual relationships but also key visual motifs, audio cues, and explicit textual evidence from OCR, this compact representation would shift the paradigm from mere frame selection towards a more holistic video comprehension. This would provide a more robust and token-efficient foundation for reasoning, paving the way for more scalable video AI.

References

[1] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. 9

[2] Weiyu Guo, Ziyang Chen, Shaoguang Wang, Jianxiang He, Yijie Xu, Jinhui Ye, Ying Sun, and Hui Xiong. Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 5

[4] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29118–29128, 2025. 4

[5] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8579–8591, 2025. 1