Appendix

The appendix is structured as follows:

- The benchmark details in Section A.
- The detailed manual query identification in Section B.
- The prompt design of **DIG** in Section C.
- More details about query identification in Section D.
- More details about CAFS in Section E.
- More details about experiments in Section F.
- More efficiency analysis in Section G.

A. Benchmark Details

This section details the benchmarks used in our evaluation. A statistical overview of each dataset is provided in Table 1.

MLVU. MLVU [13] is a multi-task benchmark for long video understanding, comprising 3,102 questions across 9 categories. The dataset is partitioned into a dev set (2,593 questions) and a test set (509 questions). Tasks are categorized into three primary types: 1) holistic analysis, 2) single-detail identification, and 3) multi-detail reasoning. For our evaluation, we utilize only multiple-choice questions from the dev set and exclude open-ended questions.

LongVideoBench. LongVideoBench [11] is a question-answering benchmark featuring 3,763 web-collected videos and 6,678 human-annotated, multiple-choice questions spanning 17 fine-grained categories. The benchmark is designed to test referring reasoning by requiring models to retrieve and reason over detailed information. In our study, we utilize only the validation set of this benchmark.

VideoMME. VideoMME [4] is a multi-modal benchmark covering 30 subdomains across 6 primary visual domains. It contains 900 videos, totaling approximately 254 hours, and 2,700 question-answer pairs. The dataset includes multiple modalities (e.g., video, subtitles, audio) and splits videos by duration (short, medium, long). To focus our evaluation on long-form video understanding, we use only the medium and long duration splits. Furthermore, we leverage only the video data and corresponding questions, excluding all other modalities like subtitles.

Table 1. **Dataset Statistics.** Overview of the data statistics across LongVideoBench [11], MLVU [13] and VideoMME [4].

Dataset	Avg. Duration (s)	#QA Pairs
MLVU [13]	636.2	2174
LongVideoBench-val [11]	732.2	1337
VideoMME-short [4]	80.7	900
VideoMME-medium [4]	516.8	900
VideoMME-long [4]	2466.3	900

B. Query Identification by Human Annotator

In this section, we elaborate on the query identification process described in Section 3, detailing the methodology used to classify queries from each benchmark.

MLVU. The task structure of MLVU [13] maps directly to our proposed query definitions. Queries associated with its "holistic tasks", which necessitate a comprehensive understanding of the entire video's overarching narrative, themes or a summary of its content, are classified as global queries. Conversely, queries within its "single-detail" and "multi-detail" task categories, which inherently demand that the model focus on specific, discrete temporal segments or isolated events, are classified as localized queries. Applying this classification scheme, we identified 462 global queries and 1708 localized queries within MLVU [13].

LongVideoBench. The design of LongVideoBench [11] is centered on "referring reasoning." This evaluation paradigm is explicitly designed to test a model's capacity to ground its reasoning in specific, fine-grained visual information. By their very nature, such queries require pinpointing information within distinct temporal or spatial segments rather than assessing the video as a whole. Consequently, all queries within this benchmark correspond directly to our definition of localized queries.

VideoMME. VideoMME [4] lacks an intrinsic task classification that aligns with our global-versus-localized classification. To address this gap, we implemented a rigorous manual annotation process. We established a standardized protocol wherein human annotators were provided with detailed instructions and precise criteria (as illustrated in Figure 1) to distinguish between the two query types. To ensure the reliability of these labels and mitigate subjective bias, the final classification for each query was determined by a majority vote consensus. This meticulous annotation procedure resulted in the identification of 479 global queries and $2,221$ localized queries.

C. Prompt Design

Prompt engineering is a cornerstone of harnessing the sophisticated reasoning capabilities of LLMs and LMMs. For our **DIG** framework, we designed a series of specialized prompts to guide the models through our multi-stage video question-answering pipeline. This section details the design and rationale for the three core prompts: (1) Query Identification, (2) Reward Assignment, and (3) Direct Inference.

Query identification. The initial and most critical step in our framework is to determine the type of the user's query. This classification dictates the subsequent processing strategy. As illustrated in Figure 1, the prompt leverages a Chain-of-Thought (CoT) strategy [10] to deconstruct

Query Identification Prompt

You are a helpful assistant in a video-based question-answering process.

Core Task & Definitions

You will classify the given query into one of two categories:

1. **Global Query (isGlobal: true):** The query requires going through and understanding the entire video content.
2. **Localized Query (isGlobal: false):** The query that can be fully answered by extracting and analyzing several specific segments within the video.

Instructions for Analysis and Response

In your analysis, please follow this structured reasoning process to classify the query:

Step 1. Understand the Query: First, read the query to understand its general meaning and core intent.

Step 2. Infer Video Style (Hypothetically): Based on the query's phrasing, make a reasonable inference about the style of the video (e.g., is it a narrative film, an educational lesson, a documentary, etc.)?

Step 3. Identify Referents: Analyze if the query has specific referents. A referent is an entity (person, object), action, event, or even specific piece of information, depending on the type of video you inferred. For instance, in 'What does Professor Smith write about quantum physics?', the referent is 'Professor Smith' and 'quantum physics' since the video style is likely a lesson.

Step 4. Evaluate Referents in Context: Based on the results from step 3 and the criteria below, determine whether the query is Global or Localized.

(i) **The query is Global** if it meets either condition:
   1. Lacks a specific referent. The examples include: Summary-based: "primary focus," "in summary," "what is the video about?"
   2. Has a referent, but answering still requires a holistic understanding from going through the entire video. The examples include: "what is the boy's overall role?"

(ii) **The query is Localized** if it has specific referents, and the answer can be found by focusing on specific, related segments where it appears. Here are some examples:
   - Entity-based: "the person in the red shirt," "the black dog," "Professor Smith," "the little girl."
   - Action/Event-based: "what is [X] doing," "how does [X] build,"
   - Temporal/Sequential: "at the beginning," "after the explosion,"

Please provide your answer in the following format: {`"analysis_step1": str, "analysis_step2": str, "analysis_step3": str, "analysis_step4": str, "isGlobal": true/false`}

User Query: <Question>

Figure 1. **Query Identification Prompt.** The LLM is first provided with the task definition, followed by an application of the chain-of-thought [10] technique to arrive at a judgment.

the classification task into a series of explicit, verifiable reasoning steps. The model is instructed to first analyze the query's intent, then hypothesize the video's genre (e.g., narrative, instructional), identify specific referents (entities, actions, or concepts), and finally synthesize this information to classify the query as either global or localized. This structured approach ensures a robust and transparent classification.

Reward assignment. To generate fine-grained feedback for optimizing our video refinement process, we utilize an LMM to assign relevance scores to sampled frames. The prompt, shown in Figure 2, presents the LMM with the user's question, a specific video frame, and associated metadata (video duration and frame timestamp). The model is tasked with a two-part CoT process: first, to provide a qualitative description of the frame's content, focusing on elements pertinent to the query, and second, to assign a quantitative reward score from 0 to 100. The reward criteria are carefully defined to capture not only the frame's direct usefulness but also its contextual value, that is, whether the frame suggests that temporally adjacent segments contain the necessary information.

Direct inference. For final evaluation, we use a direct inference prompt, exemplified in Figure 3. This prompt is designed for a standard multiple-choice question-answering format. It presents the LMM with the question and a set of candidate options (A, B, C, D). Additionally, the prompt instructs the model to return only the letter corresponding to the best answer.

Reward Assignment Prompt

You are a reward model for a video-based question-answering system.

Task
You will receive a question and a sampled video frame. Your task is to evaluate the relevance of this frame for answering the question. Please assign a reward score that indicates how useful or informative the provided frame is in the context of the given question.

Instructions for Analysis and Response
In your analysis, please perform the following steps to finish your evaluation:

1. Describe the visual content of the sampled frame, focusing on elements relevant to the question, if such elements are present.
2. Assign a relevance reward between 0 and 100 based on: (1) The sampled frame's direct usefulness in answering the question (2) Whether the frame suggests that adjacent frames might provide additional information that help answer the question more effectively.

Please provide your answer in the following format: {`"description": str, "reward": int`}.

User Input
Video Duration: <Duration> seconds; Sampled Frame Timestamp: <Timestamp> seconds; Question: <Question>

Figure 2. **Reward Assignment Prompt.** The LMM is first presented with the task definition and associated metadata. Then, the chain-of-thought reasoning technique [10] is applied to assign the reward for the input frame.

Inference Prompt

Question: What is the video mainly about?
A. Planes invented by the Wright Brothers.
B. The structural difference between the planes created by Whitehead and planes created by the Wright Brothers.
C. Who invented the first plane.
D. How Whitehead and the Wright Brothers cooperated to invent the first motorized flight.
Please select the best answer from the options provided and directly provide the letter representing your choice without giving any explanation.

Figure 3. **Prompt Template Example.** Example of the prompt template used by LMMs to perform direct inference.

D. More Details about Query Identification

In this section, we evaluate the capability of contemporary LLMs to distinguish between global and localized queries. We assess the alignment between LLM predictions and human annotations by computing classification accuracy across three benchmarks: MLVU [13], LongVideoBench [11], and VideoMME [4]. The ground truth labels for these query types are derived from human annotations, as detailed in Section B.

LLMs exhibit strong alignment with human annotation. As presented in Table 2, nearly all evaluated LLMs achieve an overall classification accuracy exceeding 80%. This indicates that off-the-shelf LLMs possess sufficiently robust reasoning capabilities to effectively differentiate between localized and global queries without extensive fine-tuning when given proper prompt.

Localized queries are more readily identifiable. Table 2 further reveals that accuracy on localized queries consistently surpasses that of global queries. While GQ accuracy is comparatively lower, this has a negligible impact on final model performance; it primarily incurs a minor computational overhead. This because as established previously, performance differences between query-aware frame selection and uniform sampling are minimal for global queries. In addition, the critical metric is LQ accuracy that may influence the final performance. On this metric, almost all LLMs achieve an accuracy greater than 90%, ensuring the final performance is good. And to make a tradeoff between compute cost and final model performance, we choose to use Qwen3-Next-80B-A3B-Instruct [9] in our main experiments.

Table 2. Accuracy (%) of different LLMs in identifying localized queries (LQ) and global queries (GQ) across multiple benchmarks.

LLM	MLVU [13]			LongVideoBench [11]			VideoMME [4]		
	LQ	GQ	Overall	LQ	GQ	Overall	LQ	GQ	Overall
Qwen3-Next-80B-A3B-Instruct [9]	87.02	38.26	78.52	97.53	N/A	97.53	89.13	65.76	83.90
Llama-3.1-8B-Instruct [8]	93.65	24.01	81.50	98.20	N/A	98.20	96.99	34.24	82.95
GPT-OSS-20B [6]	82.00	74.93	80.77	93.04	N/A	93.04	89.20	69.97	84.90
DeepSeek-R1-Distill-Qwen-32B [3]	93.03	26.38	81.42	99.18	N/A	99.18	97.21	52.85	87.28

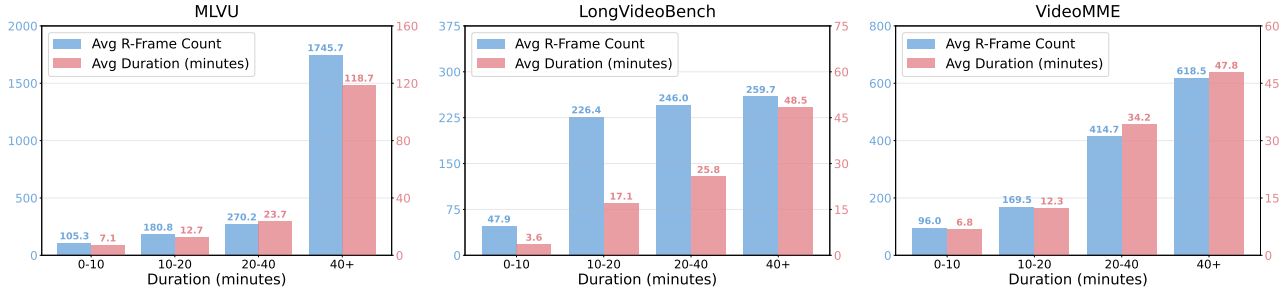

Figure 4. Correlation between video duration and the number of *r-frames* selected by the CAFS method across different benchmarks.

E. More Details about CAFS

This section provides a detailed examination of the CAFS method. Section E.1 formally specifies the algorithm of CAFS, while Section E.2 presents a statistical analysis of its output characteristics based on practical application.

E.1. Detailed Algorithm of CAFS

Algorithm 1 details our CAFS method. The process is structured into three sequential stages, taking a frame-to-frame distance sequence $d = [d_1, \ldots, d_{M-1}]$ and their corresponding original frame indices $I = [I_1, \ldots, I_M]$ as input, to produce a final set of *r-frame* indices, `r_idx`.

Initial peak detection. First, we identify all potential content boundaries. It iterates through the distance sequence, identifying any point d_i that is a local maximum, defined as being greater than its two immediate neighbors ($d_{i-1} < d_i < d_{i+1}$). The indices i of all such local maxima are collected into an initial `peaks` set.

Topographic prominence filtering. Second, we prune the `peaks` set, retaining only the most significant content transitions. For each peak $j \in$ `peaks`, it calculates its "prominence" by finding the lowest base levels to its left (l_{min}) and right (r_{min}). The prominence is then defined as the peak's height d_j minus the higher of its two bases (`prominence` $= d_j - \max(l_{\text{min}}, r_{\text{min}})$). This metric quantifies how much a peak "stands out" from the surrounding distance signal. Only peaks whose `prominence` exceeds a threshold (e.g., 0.1) are added to the `filtered_peaks` set, effectively discarding minor, localized fluctuations.

R-Frame selection. Finally, we generate the output by identifying frames that best represent the stable content *between* these significant transitions. The algorithm iterates through consecutive pairs of prominent peaks (p_1, p_2) from the filtered set. For each pair, it calculates the temporal midpoint using their associated original frame indices from I: `midpoint` $= (I_{p_1} + I_{p_2})/2$. These midpoints, which correspond to the center of the most stable segments, are aggregated into the final `r_idx` set.

E.2. More Results of CAFS

To further analyze the performance of CAFS on specific examples, we conduct a evaluation about the relationship between the number of *r-frames* and video duration.

Non-Linear information scaling in videos. Figure 4 reveals that the *r-frame* count does not scale linearly with video duration. This non-linearity is prominent in LongVideoBench [11]: videos in the $0-10$ minute bracket average 47.9 *r-frames*, whereas those in the $10-20$ minute bracket average 226.4. This finding exposes a fundamental limitation of fixed-rate sampling strategies (e.g., N frames/video or M frames/sec). Such approaches implicitly assume a uniform information distribution, leading to a suboptimal trade-off: sparse sampling risks information loss, while dense sampling incurs high temporal redundancy. CAFS bypasses this limitation by dynamically adapting its selection to the video's content density.

High context compression efficiency. CAFS effectively condenses prolonged video-level context into a sparse, salient set of *r-frames*. For instance, on MLVU [13], videos in the $10-20$ minute bracket (12.7 min avg.) are reduced to just 180.8 *r-frames* on average. This represents a sparse sampling interval of approximately one *r-frame* every 4.22 seconds, demonstrating CAFS's capability to efficiently distill essential information from extended video sequences.

Algorithm 1: Content-Adaptive Frame Selection

Input: Distance sequence $d = [d_1, d_2, \ldots, d_{M-1}]$,
Frame indices $I = [I_1, I_2, \ldots, I_M]$
Output: Selected r-frame indices `r_idx`

```
peaks ← ∅;
for i = 2 to M − 2 do
    // A peak is a point higher than its neighbors
    if d_{i−1} < d_i and d_i > d_{i+1} then
        peaks.add(i);
// Calculate the topographic prominence for each peak ↓
filtered_peaks ← ∅
for j ∈ peaks do
    // Find the lowest point to the left of the peak
    l_min ← d_j;
    k ← j − 1;
    while k ≥ 1 and d_k ≤ d_j do
        l_min ← min(l_min, d_k);
        k ← k − 1;
    // Find the lowest point to the right of the peak
    r_min ← d_j;
    m ← j + 1;
    while m ≤ M − 1 and d_m ≤ d_j do
        r_min ← min(r_min, d_m);
        m ← m + 1;
    prominence ← d_j − max(l_min, r_min) if
     prominence > 0.1 then
        filtered_peaks.add(j)
// Calculate midpoints between consecutive prominent peaks ↓
peaks ← filtered_peaks
r_idx ← ∅
for i = 0 to len(peaks) − 2 do
    p_1 ← peaks[i];
    p_2 ← peaks[i + 1];
    midpoint ← (I_{p_1} + I_{p_2})/2;
    r_idx.add(midpoint);
return r_idx
```

F. More Details about Experiment

F.1. Detailed Experiment Settings

Baseline setup. For AKS [7], we adhered to the default configuration: candidate frames were sampled at 1 fps, and frame-question similarity was computed via BLIP [5]. Based on the algorithm's selection logic, we evaluated frame budgets of $\{32, 64, 128, 192, 256\}$. We excluded budgets of 8 and 16 as the algorithm occasionally yielded null returns at these low settings. For Q-Frame [12], we employed the default "fixed frame count" strategy. Since this method limits the initial candidate pool to 128 frames, our evaluation was restricted to budgets of $\{8, 16, 32, 64, 128\}$.

Table 3. **Performance comparison between different frame selection methods.** Base LMM is Qwen3-VL-8B [1]. **Bold** indicates best performance, while Red Box denote results inferior to uniform sampling.

Method	#Frames	MLVU [13]	LVB [11]	VideoMME [4]	
				Medium	Long
UNI	8	53.4	50.3	48.4	49.2
DIG (*Ours*)	8	**58.2**	**54.9**	**53.1**	49.8
UNI	16	53.9	50.9	51.3	48.0
DIG (*Ours*)	16	**58.9**	**53.9**	**52.9**	**49.9**
UNI	32	53.7	51.0	49.4	48.6
AKS [7]	32	57.3	**54.4**	52.3	**50.1**
DIG (*Ours*)	32	**58.7**	53.2	**53.9**	49.4
UNI	64	54.7	51.2	49.9	48.9
AKS [7]	64	56.3	52.7	50.9	**51.6**
DIG (*Ours*)	64	**59.6**	**54.8**	**54.7**	49.0
UNI	128	57.2	54.4	55.1	**51.3**
AKS [7]	128	58.9	53.0	54.4	51.1
DIG (*Ours*)	128	**64.4**	**58.3**	**58.4**	51.1
UNI	192	58.9	57.1	57.6	51.6
AKS [7]	192	61.1	54.5	**61.0**	**53.8**
DIG (*Ours*)	192	**66.8**	**60.4**	58.2	50.7
UNI	256	60.4	57.6	57.8	53.4
AKS [7]	256	63.8	55.6	59.2	53.2
DIG (*Ours*)	256	**69.0**	**61.2**	**61.6**	**53.8**
UNI	512	65.4	60.2	61.4	55.0
AKS [7]	512	65.5	57.6	60.7	56.0
DIG (*Ours*)	512	**71.7**	**63.8**	**65.6**	**56.4**
UNI	768	67.5	60.9	64.3	56.6
AKS [7]	768	65.3	58.3	62.3	57.3
DIG (*Ours*)	768	**72.2**	**64.6**	**67.8**	**59.0**

F.2. Extended Experiments with DIG

To investigate the scalability of **DIG** in ultra-long context scenarios, we extended our experiments using Qwen3-VL-8B [1], an open-source LMM distinguished for its robust long-context processing capability. We test **DIG** against the uniform sampling baseline and AKS [7].

Experiment settings. For **DIG**, the query identification and CAFS configurations align with Section 5, with the exception that we employ Qwen3-VL-8B [1] as the unified backbone for both reward assignment and final inference. Similarly, AKS [7] setup mirrors Section 5 but utilizes Qwen3-VL-8B [1] as the base model. To rigorously test performance across varying context lengths, we scaled input frame counts from 8 to 768, with each frame encoded into approximately 150 tokens. Results are in Table 1, 2.

DIG delivers consistent performance gains. As evidenced in Table 3, **DIG** yields substantial improvements across nearly all frame configurations. Notably, with 256 input frames, DIG achieves an 8.6% performance boost on MLVU [13] compared to uniform sampling. Crucially, DIG maintains robustness even at the extreme scale of 768 frames, surpassing the baseline by 4.7% on MLVU [13], 3.7% on LongVideoBench [11], and 3.5% on VideoMME-

Table 4. **Performance comparison between different frame selection methods on MLVU.** Base LMMs are Qwen2.5-VL-32B [2] (left) and Qwen2.5-VL-7B [2] (right). **Bold** indicates best performance. The tasks of MLVU [13] are PlotQA (PQA), NeedleQA (NQA), Action Count (AC), Action Order (AO), Ego Reasoning (ER), Anomaly Recognition (AR), Topic Reasoning (TR).

Method	#Frames	MLVU [13]						
		PQA	NQA	AC	AO	ER	AR	TR
UNI	8	55.8	58.6	18.5	51.4	50.6	66.5	**85.6**
Q-Frame [12]	8	51.4	63.9	18.4	**60.2**	50.3	**70.5**	76.8
DIG (*Ours*)	8	**62.3**	**73.0**	**27.2**	58.3	**56.2**	66.0	84.0
UNI	16	59.4	63.9	18.0	54.8	52.8	69.5	86.3
Q-Frame [12]	16	56.4	64.8	19.9	59.5	51.4	**70.5**	77.7
DIG (*Ours*)	16	**67.9**	**78.0**	**35.0**	**66.8**	**57.1**	69.0	**86.7**
UNI	32	61.8	67.9	18.5	58.7	57.4	**76.0**	86.7
AKS [7]	32	66.8	73.0	40.3	56.0	**59.9**	74.5	**90.1**
Q-Frame [12]	32	61.4	67.9	18.9	63.8	53.1	71.5	83.3
DIG (*Ours*)	32	**72.4**	**79.2**	**48.1**	**75.7**	59.1	74.0	87.5
UNI	64	68.5	72.1	25.7	61.4	61.1	**80.0**	86.7
AKS [7]	64	73.8	76.6	40.8	58.3	63.1	75.0	88.2
Q-Frame [12]	64	68.5	73.2	21.8	66.0	59.7	77.5	85.9
DIG (*Ours*)	64	**75.9**	**81.1**	**49.5**	**78.4**	**66.5**	78.5	**89.7**
UNI	128	73.5	76.3	30.6	68.7	64.2	79.0	89.4
AKS [7]	128	78.3	**80.3**	42.2	61.8	**69.0**	77.0	87.8
Q-Frame [12]	128	73.1	76.1	30.1	69.1	64.5	**79.5**	88.6
DIG (*Ours*)	128	**79.8**	80.0	**52.4**	**79.2**	65.6	78.5	**89.7**
UNI	192	75.0	78.0	36.9	69.9	64.5	78.0	90.9
AKS [7]	192	77.6	**81.4**	47.1	63.3	**68.2**	77.0	89.4
DIG (*Ours*)	192	**82.6**	**81.4**	**53.9**	**80.7**	65.3	**79.0**	**91.6**

Model	#Frames	MLVU [13]						
		PQA	NQA	AC	AO	ER	AR	TR
UNI	8	52.1	62.5	19.4	44.0	48.6	66.5	**82.9**
Q-Frame [12]	8	50.6	67.0	19.9	**48.6**	49.7	**68.0**	73.8
DIG (*Ours*)	8	**57.1**	**73.2**	**31.6**	**48.6**	**51.1**	66.5	**82.9**
UNI	16	56.0	63.4	19.9	42.5	54.8	70.0	84.4
Q-Frame [12]	16	55.5	67.6	20.4	50.6	49.7	**70.5**	78.7
DIG (*Ours*)	16	**66.4**	**79.7**	**36.9**	**51.7**	**55.4**	68.5	**85.2**
UNI	32	59.7	69.0	22.8	51.4	54.0	**74.5**	84.4
AKS [7]	32	69.6	76.3	**42.2**	50.2	56.5	72.0	85.2
Q-Frame [12]	32	60.1	67.9	23.9	54.1	54.3	70.0	83.7
DIG (*Ours*)	32	**70.3**	**80.6**	**42.2**	**54.4**	**59.1**	73.0	**87.5**
UNI	64	64.0	74.4	26.2	51.7	59.9	**76.0**	87.1
AKS [7]	64	69.9	80.8	41.3	54.1	60.5	69.5	84.8
Q-Frame [12]	64	63.8	73.5	26.2	56.0	58.2	72.0	85.9
DIG (*Ours*)	64	**75.3**	**82.8**	**46.6**	**60.2**	**62.2**	75.5	**87.8**
UNI	128	71.4	79.2	34.5	58.7	61.4	73.0	86.7
AKS [7]	128	71.6	**83.7**	**48.5**	57.1	60.8	72.0	84.0
Q-Frame [12]	128	71.4	79.2	34.5	58.7	61.4	**73.0**	86.7
DIG (*Ours*)	128	**78.3**	82.3	45.6	**62.5**	**63.6**	72.0	**87.8**
UNI	192	72.0	80.8	40.3	61.4	63.6	**73.0**	**87.5**
AKS [7]	192	74.2	83.1	46.1	58.7	63.6	**73.0**	85.6
DIG (*Ours*)	192	**78.7**	**84.5**	**47.1**	**63.3**	**65.3**	72.0	**87.5**
UNI	256	73.5	80.0	41.3	61.4	61.1	73.0	**89.0**
AKS [7]	256	75.3	84.2	47.3	59.5	**66.2**	**75.5**	87.8
DIG (*Ours*)	256	**78.1**	**84.5**	**49.0**	**62.2**	65.1	73.0	**89.0**

Medium [4]. In contrast, while AKS [7] remains competitive at lower frame counts ($\leq$ 64), it exhibits marked performance degradation as the context length increases, frequently falling below the uniform sampling baseline. Given that practical video understanding tasks necessitate maximizing input frames to capture comprehensive temporal details, AKS [7] demonstrates limited utility for real-world applications. Conversely, **DIG** exhibits superior scalability, effectively delivering sustained performance gains.

F.3. Detailed Experiment Results & More Analysis

We present detailed performance breakdowns corresponding to the benchmarks discussed in Section 5. Comprehensive quantitative results are in Tables 4, 6, and 5.

Uniform sampling suffices for global queries. For global queries, specifically Anomaly Recognition and Topic Reasoning tasks within MLVU [13], all evaluated methods perform comparably to uniform sampling, regardless of the input frame count. This observation reaffirms our previous assertion: uniform sampling is the preferred strategy for global queries, as it achieves sufficient performance while maintaining high efficiency.

Inference for localized queries operates in two distinct stages: query-aware frame selection and subsequent reasoning based on the retrieved content. Without the initial selection stage, evaluating the model's fundamental performance is challenging, as errors may stem from information-deficient inputs rather than inherent model limitations. By incorporating this stage to ensure the input contains relevant information, we can decouple data retrieval issues from reasoning capabilities. This allows for a more accurate assessment of the model's intrinsic proficiency across different tasks, yielding deeper insights.

Query-aware selection uncovers intrinsic visual perception capabilities. As shown in Table 4 and 5, our method significantly and consistently outperforms uniform sampling on localized perception tasks (e.g., PlotQA, NeedleQA, and L1-Perception). Notably, these tasks primarily evaluate fundamental visual perception capabilities. Our findings suggest that LMMs are intrinsically capable of solving such tasks, provided the query-relevant information is effectively supplied. This explains the substantial performance gap: while uniform sampling often introduces significant noise by including irrelevant content, query-aware selection ensures the model is conditioned on relevant frames.

Temporal reasoning remains a fundamental bottleneck. Conversely, regarding tasks requiring temporal logic (e.g., Action Order and L2-Relation), performance remains stagnant across all methods. Even when provided with query-relevant visual information, model performance does not improve. This underscores a critical limitation: current LMMs struggle with temporal reasoning and sequencing, a deficiency that persists independently of the quality of visual information retrieval.

Table 5. **Performance Comparison between Different Frame Selection Methods on LongVideoBench.** Base LMMs are Qwen2.5-VL-7B [2](top) and Qwen2.5-VL-32B [2](bottom). **Bold** indicates best performance.

Model	#Frames	LongVideoBench [11]																		
		L1-Perception									L2-Relation									
		S2E	S2A	O2E	T2O	S2O	T2E	E2O	T2A	Avg	TOS	E3E	SAA	O3O	T3O	T3E	TAA	SSS	SOS	Avg
UNI	8	57.0	51.1	**62.8**	56.6	45.8	58.5	56.9	49.4	54.4	**38.4**	**62.8**	47.2	45.5	47.3	47.9	**46.3**	**34.0**	64.2	48.3
Q-Frame [12]	8	67.7	**73.9**	60.9	**57.9**	55.6	**64.6**	**63.1**	55.7	**62.7**	31.5	62.8	**52.8**	47.0	39.2	**48.0**	45.1	28.9	65.4	46.8
DIG (*Ours*)	8	**69.9**	68.2	62.1	50.0	**55.6**	61.5	61.5	**62.0**	61.8	37.0	61.7	50.0	**48.5**	**54.1**	43.8	45.1	29.9	**67.9**	**48.6**
UNI	16	65.6	64.8	62.8	53.9	48.6	61.5	61.5	51.9	59.0	**37.0**	**62.8**	50.0	47.0	**56.8**	45.2	**51.9**	36.1	63.0	49.3
Q-Frame [12]	16	66.7	70.5	**65.5**	63.2	**61.1**	64.6	**69.2**	63.3	**65.6**	34.3	59.6	**56.9**	**54.6**	43.2	49.3	50.0	**38.1**	65.4	50.1
DIG (*Ours*)	16	**72.0**	**71.6**	59.8	**65.8**	54.2	**69.2**	**69.2**	**63.3**	65.4	**37.0**	57.4	52.8	43.9	**56.8**	**52.1**	43.9	36.1	**74.1**	**50.4**
UNI	32	67.7	58.0	61.7	56.6	62.5	**67.7**	67.7	51.9	61.9	37.0	61.7	55.6	**56.1**	**55.4**	49.3	**52.4**	**40.2**	66.7	52.7
AKS [7]	32	65.6	77.3	**67.8**	63.2	**63.9**	63.1	63.1	**64.6**	66.1	37.0	**67.0**	58.3	**56.1**	45.9	**53.4**	48.8	38.1	**74.1**	**53.2**
Q-Frame [12]	32	64.5	69.3	60.9	59.2	61.1	61.5	**69.2**	60.8	63.3	32.9	59.6	**62.5**	50.0	50.0	45.2	46.3	39.2	66.7	50.3
DIG (*Ours*)	32	**72.0**	**78.4**	63.2	**68.4**	62.5	**67.7**	67.7	62.0	**68.0**	**41.1**	62.8	52.8	51.5	**55.4**	50.7	48.8	36.1	70.4	52.1
UNI	64	**73.1**	67.0	62.8	59.2	56.9	63.1	66.2	62.0	64.6	34.2	62.8	58.3	59.1	**60.8**	47.9	51.2	**43.3**	67.9	53.9
AKS [7]	64	69.9	77.3	**71.3**	65.8	59.7	60.0	64.6	64.6	67.2	**41.1**	**70.2**	**61.1**	56.1	47.3	49.3	47.6	40.2	**76.5**	54.5
Q-Frame [12]	64	63.4	70.5	63.2	57.9	**61.1**	63.1	64.6	65.8	63.8	34.3	67.0	58.3	53.0	52.7	46.6	**51.2**	37.1	66.7	52.0
DIG (*Ours*)	64	69.9	**78.4**	66.7	**69.7**	55.6	**67.7**	**72.3**	**68.4**	**68.8**	38.4	66.0	58.3	**62.1**	59.5	**53.4**	46.3	42.3	69.1	**54.9**
UNI	128	71.0	67.0	67.8	64.5	**61.1**	67.7	72.3	**73.4**	68.2	**38.4**	**68.1**	56.9	59.1	**56.8**	50.7	**54.9**	46.4	74.1	**56.3**
AKS [7]	128	69.9	72.7	66.7	64.5	**61.1**	60.0	66.2	64.6	66.1	37.0	**68.1**	**63.9**	56.1	50.0	50.7	48.8	**47.4**	**76.5**	55.6
Q-Frame [12]	128	66.7	67.1	**69.0**	60.5	59.7	64.6	67.7	64.6	65.1	**38.4**	64.9	58.3	54.6	**56.8**	49.3	51.2	45.4	75.3	55.1
DIG (*Ours*)	128	**72.0**	**79.5**	66.7	**71.1**	**61.1**	**69.2**	**75.4**	70.9	**70.9**	**38.4**	**68.1**	61.1	**65.2**	**56.8**	**57.5**	47.6	45.4	67.9	**56.3**
UNI	192	**74.2**	72.7	66.0	**68.4**	59.7	67.7	70.8	63.3	68.2	**35.6**	66.0	59.7	59.1	**58.1**	58.9	**56.1**	46.4	67.9	56.5
AKS [7]	192	69.9	76.1	**67.8**	63.2	**61.1**	63.1	67.7	65.8	67.2	**35.6**	**68.1**	**62.5**	56.1	56.8	52.1	48.8	46.4	**72.8**	55.6
DIG (*Ours*)	192	**74.2**	**84.1**	66.7	67.1	59.7	**69.2**	**73.8**	**78.5**	**72.0**	**35.6**	**68.1**	61.1	**66.7**	56.8	**61.6**	48.8	**49.5**	70.4	**57.6**
UNI	256	69.9	73.9	69.1	63.2	59.7	63.1	72.3	68.4	66.7	**37.0**	69.1	**62.5**	60.6	55.4	**56.2**	**54.9**	46.4	69.1	**56.9**
AKS [7]	256	68.8	72.7	70.1	65.8	61.1	66.2	63.1	65.8	67.0	35.6	69.1	**62.5**	56.1	**56.8**	52.1	48.8	43.3	**71.6**	55.2
DIG (*Ours*)	256	**76.3**	**80.7**	**71.3**	**72.4**	**65.3**	**69.2**	**75.4**	**74.7**	**73.4**	**37.0**	**71.3**	59.7	**68.2**	**56.8**	**56.2**	45.1	**49.5**	67.9	**56.9**

Model	#Frames	LongVideoBench [11]																		
		L1-Perception									L2-Relation									
		S2E	S2A	O2E	T2O	S2O	T2E	E2O	T2A	Avg	TOS	E3E	SAA	O3O	T3O	T3E	TAA	SSS	SOS	Avg
UNI	8	62.4	61.4	**65.5**	54.0	51.4	58.5	58.5	51.9	58.2	30.1	**63.8**	55.6	42.4	47.3	**49.3**	45.1	**46.4**	58.0	49.2
Q-Frame [12]	8	62.4	**81.8**	62.1	55.3	**55.6**	**61.5**	**66.2**	53.2	**62.6**	30.1	57.5	56.9	42.4	43.2	46.6	41.5	36.1	59.3	46.1
DIG (*Ours*)	8	**63.4**	75.0	60.9	**59.2**	48.6	60.0	66.2	**60.8**	61.8	**37.0**	60.6	**61.1**	**50.0**	**52.7**	**49.3**	**47.6**	44.3	**65.4**	**52.0**
UNI	16	57.0	71.6	57.5	52.6	48.6	63.1	63.1	45.6	57.4	**37.0**	58.5	58.3	53.0	54.1	50.7	**53.7**	**42.3**	66.7	**52.7**
Q-Frame [12]	16	**69.9**	77.3	**64.4**	**64.5**	58.3	64.6	63.1	62.0	**65.9**	31.5	**62.8**	58.3	51.5	46.0	43.8	46.3	40.2	54.3	49.5
DIG (*Ours*)	16	66.7	**80.7**	63.2	55.3	**62.5**	**66.2**	**64.6**	**68.4**	**65.9**	34.3	56.4	**63.9**	**54.5**	**56.8**	**54.8**	46.3	39.2	**67.9**	52.7
UNI	32	68.8	68.2	64.4	57.9	54.2	63.1	61.5	53.2	61.8	32.9	66.0	56.9	53.0	52.7	**58.9**	50.0	47.4	70.4	**54.5**
AKS [7]	32	65.6	**77.3**	**69.0**	**61.8**	**65.3**	61.5	66.2	**64.6**	**66.7**	34.3	69.2	59.7	50.0	41.9	48.0	**51.2**	44.3	**72.8**	52.8
Q-Frame [12]	32	69.9	**77.3**	63.2	59.2	**65.3**	61.5	60.0	54.4	61.4	**37.0**	59.6	54.2	48.5	48.6	48.0	**51.2**	50.5	60.5	51.3
DIG (*Ours*)	32	**71.0**	75.0	65.5	57.9	**65.3**	**72.3**	**67.7**	60.8	66.9	35.6	**70.2**	**66.7**	**62.1**	**56.8**	56.2	43.9	**51.5**	71.6	**57.2**
UNI	64	**69.9**	65.9	64.4	61.8	58.3	60.0	70.8	57.0	63.7	31.5	67.0	58.3	51.5	55.4	**54.8**	51.2	51.5	69.1	54.9
AKS [7]	64	67.7	76.1	**66.7**	60.5	**69.4**	61.5	**72.3**	**67.1**	67.8	**37.0**	**73.4**	56.9	53.0	47.3	52.1	48.8	52.6	**75.3**	**55.8**
Q-Frame [12]	64	63.4	77.3	**66.7**	56.6	66.7	64.6	67.7	65.8	64.0	35.6	63.8	61.1	47.0	46.0	52.1	**52.4**	53.6	67.9	53.8
DIG (*Ours*)	64	**69.9**	**79.5**	65.5	**64.5**	65.3	**66.2**	**72.3**	**67.1**	**68.8**	34.3	68.1	**73.6**	**62.1**	**59.5**	**54.8**	46.3	**57.7**	72.8	**55.8**
UNI	128	71.0	70.5	62.1	61.8	68.1	63.1	67.7	60.8	65.7	35.6	70.2	62.5	51.5	55.4	57.5	**53.7**	**60.8**	71.6	58.3
AKS [7]	128	69.9	72.7	**67.8**	63.2	65.3	61.5	70.8	**63.3**	67.0	**37.0**	**74.5**	58.3	56.1	51.4	53.4	52.4	53.6	**76.5**	58.6
Q-Frame [12]	128	65.6	69.3	60.9	59.2	65.3	61.5	67.7	59.5	63.7	35.6	71.3	61.1	48.5	51.4	53.4	52.4	**60.8**	70.4	56.9
DIG (*Ours*)	128	**76.3**	**83.0**	65.5	**65.8**	**70.8**	**67.7**	**80.0**	**63.3**	**71.6**	34.3	72.3	**68.1**	**68.2**	**66.2**	**57.5**	45.1	55.7	74.1	**60.2**
UNI	192	72.0	73.9	66.7	67.1	66.7	**67.7**	72.3	63.3	68.8	34.3	**76.6**	58.3	62.1	58.1	**60.3**	**52.4**	**56.7**	71.6	59.4
AKS [7]	192	71.0	79.5	**67.8**	67.1	66.7	60.0	69.2	62.0	68.3	35.6	74.5	58.3	57.6	54.1	50.7	48.8	54.6	**76.5**	57.3
DIG (*Ours*)	192	**73.1**	**84.1**	**67.8**	**68.4**	**68.1**	64.6	**76.9**	**65.8**	**71.1**	**38.4**	74.5	**70.8**	**69.7**	**71.6**	57.5	43.9	**56.7**	75.3	**60.7**

G. More Efficiency Analysis of DIG

G.1. Detailed Runtime Profiling

We evaluate the computational efficiency of **DIG** compared to distinct baselines, AKS [7] and Q-Frame [12]. The total runtime of each method can be divided into two stages:

- *Key Frame Selection*, where the method identifies optimal indices from raw video.
- *Inference*, where the LMM processes the selected frames to generate a response.

All experiments were conducted on a node equipped with 8 NVIDIA A100 GPUs. To provide a comprehensive analysis, we report the standard LMM inference latency across varying input frame counts in Table 7 and detail the selection overhead introduced by specific methods in Table 8.

DIG achieves a favorable efficiency-performance trade-off. As evidenced in Table 8, **DIG** offers a significant efficiency advantage over AKS [7], reducing computational overhead by an order of magnitude while maintaining superior downstream performance (see Section 5). While **DIG** incurs a marginal increase in processing time compared to Q-Frame [12], this cost is justified by substantial robustness gains; specifically, Q-Frame [12] fails to outperform uniform sampling as frame counts exceed 32, whereas **DIG** consistently surpasses baselines across all settings. Furthermore, comparing the selection overhead (Ta-

Table 6. **Performance comparison between different frame selection methods on VideoMME.** Base LMMs are Qwen2.5-VL-7B [2](left) and Qwen2.5-VL-32B [2](right). **Bold** indicates best performance. The tasks are Object Reasoning (ORA), Object Recognition (ORC), Action Reasoning (ARA), Information Synopsis (INS), Counting Problem (COP), Temporal Reasoning (TER), Temporal Perception (TEP), Spatial Perception (SPP), Spatial Reasoning (SPR), OCR, Attribute Perception (ATP), Action Recognition (ACR).

Model	#Frames	VideoMME [4]											
		ORA	ORC	ARA	INS	COP	TER	TEP	SPR	SPP	OCR	ATP	ACR
UNI	8	49.5	54.5	49.5	67.8	36.2	**40.7**	47.3	58.9	63.0	48.9	62.2	53.4
Q-Frame [12]	8	**50.5**	50.1	50.8	64.8	36.9	36.0	43.3	62.1	33.3	44.0	57.0	52.2
DIG (*Ours*)	8	47.6	**60.2**	**52.3**	**70.0**	**39.6**	**40.7**	**56.4**	**64.3**	**66.7**	**55.4**	**65.3**	**55.9**
UNI	16	53.5	58.8	51.2	**74.6**	37.7	**43.5**	**63.6**	64.3	72.2	54.7	67.1	**56.9**
Q-Frame [12]	16	52.4	52.2	52.2	66.0	36.8	36.1	56.9	62.1	45.9	48.8	59.0	49.0
DIG (*Ours*)	16	**57.9**	**61.0**	**54.4**	**74.6**	**41.4**	**43.5**	56.4	**66.1**	**75.9**	**59.0**	**71.2**	56.2
UNI	32	57.5	63.6	55.8	76.5	42.5	43.5	**67.3**	**76.8**	**72.2**	62.6	**74.8**	59.7
AKS [7]	32	56.8	66.7	51.9	**80.2**	**42.9**	**49.2**	60.0	**76.8**	**72.2**	66.2	**74.8**	59.7
Q-Frame [12]	32	55.3	56.6	54.2	68.9	38.3	37.2	59.6	65.5	50.1	50.1	64.0	50.6
DIG (*Ours*)	32	**59.5**	**67.2**	**56.1**	75.9	40.3	46.9	52.7	**76.8**	**72.2**	**69.8**	73.4	**61.0**
UNI	64	58.1	66.7	54.0	76.8	41.4	43.5	69.1	**76.8**	68.5	67.6	**74.8**	**63.9**
AKS [7]	64	58.4	67.5	56.1	**79.3**	44.4	**52.0**	**72.7**	**76.8**	72.2	72.7	**74.8**	61.0
Q-Frame [12]	64	56.0	62.4	56.0	72.6	**46.5**	45.2	56.8	69.1	54.2	51.3	72.0	51.2
DIG (*Ours*)	64	**60.6**	**68.4**	**57.5**	78.0	46.3	49.2	58.2	73.2	**75.9**	**73.4**	73.4	63.3
UNI	128	61.2	71.2	58.9	79.9	**45.1**	**57.1**	70.9	76.8	**68.5**	71.2	76.6	66.1
AKS [7]	128	61.0	68.4	**59.3**	80.2	44.8	**57.1**	**76.4**	75.0	**68.5**	**77.0**	77.9	61.3
Q-Frame [12]	128	58.7	64.5	55.6	78.0	38.9	55.6	64.9	72.4	54.3	57.4	67.0	59.3
DIG (*Ours*)	128	**61.7**	**72.3**	57.2	**80.5**	**45.1**	52.5	58.2	**78.6**	**68.5**	71.2	**78.8**	**66.8**
UNI	192	62.3	72.0	**60.7**	79.9	**48.5**	54.2	72.7	76.8	70.4	71.9	77.5	**68.1**
AKS [7]	192	60.8	69.2	57.5	79.3	45.5	**57.6**	**81.8**	76.8	**72.2**	**76.3**	77.9	63.9
DIG (*Ours*)	192	**64.5**	**73.7**	60.0	**80.2**	46.3	54.2	65.5	**78.6**	68.5	74.1	**79.3**	65.8
UNI	256	62.1	71.5	59.6	**82.4**	**46.6**	**57.6**	76.4	75.0	66.7	71.9	77.9	**68.1**
AKS [7]	256	61.0	70.9	57.9	79.3	44.8	57.1	**83.6**	75.0	**70.4**	**74.1**	77.9	64.5
DIG (*Ours*)	256	**63.0**	**72.0**	**62.1**	**82.4**	46.3	54.8	67.3	**78.6**	66.7	73.4	**80.2**	67.4

Model	#Frames	VideoMME [4]											
		ORA	ORC	ARA	INS	COP	TER	TEP	SPR	SPP	OCR	ATP	ACR
UNI	8	**57.2**	54.8	**53.4**	**69.7**	30.1	37.8	**46.0**	65.5	54.2	47.6	**62.0**	45.6
Q-Frame [12]	8	50.5	50.1	50.8	64.8	**36.9**	36.0	43.3	62.1	33.3	44.0	57.0	**52.2**
DIG (*Ours*)	8	55.6	**58.1**	**53.4**	**69.7**	31.5	**40.9**	**46.0**	**72.4**	**62.5**	**54.9**	**62.0**	48.4
UNI	16	55.4	56.5	52.5	70.5	32.2	**47.6**	51.4	**75.9**	**62.5**	42.7	**65.0**	**50.6**
Q-Frame [12]	16	52.4	52.2	52.2	66.0	36.8	36.1	**56.9**	62.1	45.9	**48.8**	59.0	49.0
DIG (*Ours*)	16	**58.3**	**59.7**	**53.4**	**71.0**	**37.8**	47.0	56.8	**75.9**	**62.5**	45.1	**65.0**	47.3
UNI	32	55.4	55.9	54.2	76.8	35.0	40.9	**62.2**	75.9	**62.5**	47.6	66.0	**51.7**
AKS [7]	32	58.6	57.0	**57.6**	**78.0**	34.3	47.6	56.8	**79.3**	45.8	**61.0**	68.0	51.1
Q-Frame [12]	32	55.3	56.6	54.2	68.9	**38.3**	37.2	59.6	65.5	50.1	50.1	64.0	50.6
DIG (*Ours*)	32	**58.8**	**64.0**	55.5	**78.0**	35.7	**51.8**	48.7	75.9	58.3	56.1	**69.0**	50.0
UNI	64	61.8	62.9	58.0	76.8	34.3	44.5	**64.9**	**79.3**	**62.5**	58.5	69.0	**59.3**
AKS [7]	64	61.0	63.4	**59.7**	**80.1**	37.8	**54.3**	62.2	**79.3**	54.2	59.8	**73.0**	56.6
Q-Frame [12]	64	56.0	62.4	56.0	72.6	**46.5**	45.2	56.8	69.1	54.2	51.3	72.0	51.2
DIG (*Ours*)	64	**62.6**	**65.6**	57.6	76.4	32.9	51.8	59.5	75.9	58.3	**67.1**	**73.0**	58.2
UNI	128	63.4	69.9	**63.5**	81.3	42.0	54.3	59.5	**86.2**	**58.3**	68.3	73.0	57.1
AKS [7]	128	**66.0**	68.3	58.8	81.3	42.7	**57.9**	**67.6**	82.8	50.0	**69.5**	**75.0**	59.9
Q-Frame [12]	128	58.7	64.5	55.6	78.0	38.9	55.6	64.9	72.4	54.3	57.4	67.0	59.3
DIG (*Ours*)	128	65.2	**73.1**	61.8	**81.7**	**45.5**	52.4	64.9	82.8	54.2	**69.5**	**75.0**	**60.4**
UNI	192	64.7	69.9	**63.0**	81.3	**44.8**	55.5	73.0	**86.2**	**62.5**	68.3	75.0	**62.1**
AKS [7]	192	65.5	69.4	59.7	81.3	42.7	**58.5**	**75.7**	79.3	54.2	**73.2**	78.0	58.8
DIG (*Ours*)	192	**66.8**	**74.2**	62.6	**81.7**	42.0	57.3	64.9	79.3	58.3	**73.2**	**80.0**	61.0

Table 7. **Inference latency analysis.** The inference time (in minutes) of the base LMM (Qwen2.5-VL-7B [2]) across different input frame counts using standard uniform sampling.

#Frames		8	16	32	64	128	192	256
Inference Time (min)	MLVU [13]	3.2	5.0	9.3	17.6	29.1	37.3	43.4
	LongVideoBench [11]	1.4	2.2	4.3	8.3	14.0	19.9	25.6
	VideoMME [4]	3.1	4.7	8.7	15.8	26.1	36.7	46.3

Table 8. **Comparison of frame selection overhead.** The time cost (in minutes) required by different methods to process videos and select key frames. For DIG, we break down the cost into Query Identification (QI), Content-Aware Frame Selection (CAFS), Reward Assignment (RA), and Video Refinement (VR).

Method		AKS [7]	Q-Frame [12]	DIG (Ours)				
				QI	CAFS	RA	VR	Sum
Selection Time (min)	MLVU [13]	≥ 720	122.1	11.3	25.9	218.9	0.2	256.3
	LongVideoBench [11]	≥ 720	34.5	7.6	20.8	110.4	0.1	138.9
	VideoMME [4]	≥ 720	94.2	11.6	31.2	264.8	0.3	307.9

ble 8) against standard inference latency (Table 7), the additional cost remains within a reasonable range. This confirms that **DIG** effectively balances efficiency and accuracy, serving as a practical, plug-and-play module for enhanced long-form video understanding.

G.2. Efficiency Gains from Query Identification

To balance efficiency and accuracy, **DIG** employs a Query Identification module. We apply resource-intensive key frame selection only to localized queries, defaulting to efficient uniform sampling for global ones. This adaptive strategy minimizes computational cost without compromising downstream performance. Table 9 quantifies these gains by comparing our adaptive approach against the baseline that applies our specific selection universally. On LongVideoBench [11], where queries are predominantly localized, the QI module incurs a marginal overhead (3.6%) due to the additional classification step. However, on datasets with a diverse mix of query types, such as VideoMME [4] and MLVU [13], the adaptive strategy yields significant time savings (19.9% and 13.3%, respectively). This demonstrates that the QI module effectively optimizes resource allocation by bypassing unnecessary computation for global queries.

Table 9. **Impact of Query Identification on efficiency.** We compare the frame selection time (in minutes) of applying our specific selection universally (w/o. QI) versus DIG's adaptive approach (w. QI). "Percent" denotes the proportion of localized queries.

	Percent	w/o. QI	w. QI
MLVU [13]	82.8	295.7	256.3 (↓ 13.3%)
LongVideoBench [11]	97.8	134.1	138.9 (↑ 3.6%)
VideoMME [4]	77.0	384.2	307.9 (↓ 19.9%)

References

[1] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu. Qwen3-vl technical report, 2025. 5

[2] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report, 2025. 6, 7, 8

[3] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 4

[4] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, P. Chen, Y. Li, S. Lin, S. Zhao, K. Li, T. Xu, X. Zheng, E. Chen, R. Ji, and X. Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024. 1, 3, 4, 5, 6, 8

[5] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 5

[6] OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. 4

[7] X. Tang, J. Qiu, L. Xie, Y. Tian, J. Jiao, and Q. Ye. Adaptive keyframe sampling for long video understanding, 2025. 5, 6, 7, 8

[8] L. Team. The llama 3 herd of models, 2024. 4

[9] Q. Team. Qwen3 technical report, 2025. 3, 4

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 1, 2, 3

[11] H. Wu, D. Li, B. Chen, and J. Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. 1, 3, 4, 5, 7, 8

[12] S. Zhang, J. Yang, J. Yin, Z. Luo, and J. Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms, 2025. 5, 6, 7, 8

[13] J. Zhou, Y. Shu, B. Zhao, B. Wu, Z. Liang, S. Xiao, M. Qin, X. Yang, Y. Xiong, B. Zhang, T. Huang, and Z. Liu. Mlvu: Benchmarking multi-task long video understanding, 2025. 1, 3, 4, 5, 6, 8