THREADING KEYFRAME WITH NARRATIVES: MLLMS AS STRONG LONG VIDEO COMPREHENDERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Employing Multimodal Large Language Models (MLLMs) for long video understanding remains a challenging problem due to the dilemma between the substantial number of video frames (i.e., visual tokens) versus the limited context length of language models. Traditional uniform sampling often leads to selection of irrelevant content, while post-training MLLMs on thousands of frames imposes a substantial computational burden. In this paper, we propose ***Nar**rating **K**ey**F**rames **C**apturing* (**Nar-KFC**), a plug-and-play module to facilitate effective and efficient long video understanding. Nar-KFC generally involves two collaborative steps. First, we formulate the *keyframe* selection process as an integer quadratic programming problem, jointly optimizing query-relevance and frame-diversity. To avoid its computational complexity, a customized greedy search strategy is designed as an efficient alternative. Second, to mitigate the temporal discontinuity caused by sparse keyframe sampling, we further introduce interleaved textual *narratives* generated from non-keyframes using off-the-shelf captioners. These narratives are inserted between keyframes based on their true temporal order, forming a coherent and compact representation. Nar-KFC thus serves as a temporal- and content-aware compression strategy that complements visual and textual modalities. Experimental results on multiple long-video benchmarks demonstrate that Nar-KFC significantly improves the performance of popular MLLMs. Code will be made publicly available.

1 INTRODUCTION

Building upon the success of revolutionary Large Language Model (LLMs) (Touvron et al., 2023; Team et al., 2024), recent advances in Multimodal Large Language Models (Liu et al., 2023; Li et al., 2024b; Wang et al., 2024b; Chen et al., 2024c; Tong et al., 2024; Lin et al., 2024b) have significantly improved open-world visual understanding. Moving beyond static images, a natural extension of MLLMs is their application to video understanding. Existing studies have validated their effectiveness in comprehending short videos (∼10 s) (Yang et al., 2022; Kim et al., 2024; Yao et al., 2024a). However, when scaling MLLMs to long videos (Fu et al., 2025; Wu et al., 2024b; Chandrasegaran et al., 2024; Zhou et al., 2025) (e.g., hours), several critical challenges emerge.

The primary challenge stems from the inherent context limitation of MLLMs, which cannot accommodate the vast volume of visual tokens generated from the whole video. A prominent solution is to extend the context window of language models and fine-tune them on carefully collected long videos. Current video-oriented LLMs, known as VideoLLMs (Lin et al., 2024a; Jin et al., 2024; Song et al., 2024; Xu et al., 2024a; Chen et al., 2024b; Zohar et al., 2024; Shu et al., 2025; Cheng et al., 2025a; Wang et al., 2025a), typically undergo post-training on existing LLMs/MLLMs through: 1) employing a relatively large stride uniform sampling scheme, and 2) incorporating token-level merging or compression techniques to enable broader temporal coverage. However, uniform sampling often fails to preserve key moments relevant to specific instructions, while feeding an excessive number of frames as input introduces redundancy, leading to substantial computational overhead. An alternative solution follows a training-free paradigm (Zhang et al., 2024a; Kahatapitiya et al., 2024; Wang et al., 2024d; 2025b; Park et al., 2024; Ma et al., 2025), where raw videos are first converted into sequential captions, which are subsequently processed using the long-range reasoning abilities of LLMs (Achiam et al., 2023). Compared to direct video frame encoding, textual captions inherently require far fewer tokens, allowing efficient inference in a single forward pass. Nonetheless, the

translation from video frame to caption inevitably results in critical information loss (e.g., important visual features), potentially leading to hallucinated answers caused by the LLM bias.

Regarding the aforementioned paradigms, e.g., training a VideoLLM or reasoning with LLMs on textual captions, are current MLLMs fully equipped to comprehend long videos despite their limited context length? Instead of relying on uniform sampling, recent studies have focused on learning to select query-relevant keyframes (Yu et al., 2023; Hu et al., 2025; Yao et al., 2025) to facilitate inference with MLLMs. Due to the temporal redundancy among adjacent frames, trival similarity-based keyframe selection tends to retrieve frames located within narrow time windows, thereby compromising accuracy. To this end, adaptive keyframe sampling (Tang et al., 2025), inverse transform sampling (Liu et al., 2025b), DPP sampling (Sun et al., 2025) have been proposed to promote content diversity to mitigate the concentration of keyframes. Despite a decent boost over existing MLLMs, these methods largely depend on handcrafted or heuristic strategies with limited theoretical formulations, and empirically, the retrieved frames can be temporally distant, especially in long videos. Consequently, the keyframe selection process can introduce temporal discontinuities into the input provided to the MLLM, ultimately hindering its holistic understanding of video content.

In this paper, we propose ***Nar-KFC*** (**Nar**rating **K**ey**F**rames **C**apturing), a training-free framework for long video understanding with MLLMs. Unlike previous approaches, Nar-KFC jointly considers *query-relevance*, *frame-diversity* and *temporal-continuity* through two collaborative stages. The first stage **KFC** selects keyframes by considering both query relevance and frame diversity, so as to resolve the issues of critical information loss from uniform sampling and the too-narrow focus using just query-relevance. We consider keyframe selection as a graph problem, where each node is a frame and the edge weight (score) between nodes combines query-relevant similarities and frame-to-frame dissimilarities (frame-diversity). The optimal keyframes are obtained by finding the subgraph with largest total edge weight, which can be formulated as an integer quadratic programming (IQP) problem. However, since IQP is NP-hard with exponential complexity, finding exact solutions is infeasible in practice. To overcome this, we introduce a robust and efficient greedy search (GS) strategy, which, with proper preprocessing of the score matrix, achieves near-optimal performance with significantly reduced computational complexity.

The second stage **Nar-KFC** addresses the problem of temporal discontinuities caused when selecting keyframes at uneven timestamps. Specifically, Nar-KFC works by threading keyframes (visual tokens) with *non-keyframe narratives* (text tokens), generated by captioning the intermediate, unselected frames in between, aiming to reconstruct the video as a continuous and coherent sequence in both textual and visual modalities. A narrative interval is further applied to control the total number of captions and to reduce the similarity between neighboring descriptions. Leveraging only a lightweight 2B captioning model, e.g., Qwen2-VL-2B (Wang et al., 2024b), Nar-KFC demonstrates significant improvements over existing MLLMs. In summary, the contributions of this paper are three-fold:

- Jointly considering query-relevance and frame-diversity, we formulate the keyframe capturing process (KFC) in long videos as a subgraph selection problem, implemented as an integer quadratic programming problem. We introduce a customized greedy search algorithm to solve this problem with significantly reduced and practical time complexity.
- We propose Nar-KFC, which threads the optimized keyframes with non-keyframe narratives. By interleaving the two modalities in a temporally continuous manner, Nar-KFC constructs coherent and compact video representations, enabling a broader video coverage under the constraint of frame length limitations in current MLLMs.
- Our KFC and Nar-KFC are generally compatible with many MLLMs, achieving consistent improvements across four mainstream MLLMs on multiple long-video benchmarks.

2 Related Work

Transformer-based LLMs have revolutionized the field of natural language processing (Brown et al., 2020; OpenAI, 2023; Grattafiori et al., 2024; Achiam et al., 2023). By incorporating multimodal inputs such as images and videos (Li et al., 2024b; Zhu et al., 2023) with a vision encoder, e.g., ViT (Dosovitskiy et al., 2020), researchers further extend powerful LLMs to multimodal large language models (MLLMs) for open-world visual understanding (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023). Despite similar advancements of MLLMs on various video understanding tasks including video captioning (Chen et al., 2024a; Yang et al., 2023; Wu et al., 2024a), video question

answering (Maaz et al., 2023; Li et al., 2023b; Min et al., 2024), and temporal reasoning (Qian et al., 2024), significant challenges emerge when scaling to long videos due to the substantial amount of video frames not fitting in the limited context length of LLMs (Wu et al., 2024b).

Recent studies have explored methods to extend the context length of LLMs (Wan et al., 2024; Xiong et al., 2024), or introduced various token-level merging and compression techniques (Song et al., 2024; Shen et al., 2024; Li et al., 2024d; Wang et al., 2024c; Shu et al., 2025) to accommodate more frames as input. However, these approaches typically require additional fine-tuning of existing language models, which increases computational complexity and introduces the risk of hallucinations (Liu et al., 2024c). Given that textual tokens are significantly fewer than visual frames, another line of research first converts all video frames into textual descriptions, which are then used for long video inference, either by summarizing them (Zhang et al., 2024a; Park et al., 2024) or identifying central frames based on textual similarity via agents (Wang et al., 2024d; 2025b; Ma et al., 2025; Ye et al., 2025; Liu et al., 2025a). Nonetheless, the converting process inevitably leads to critical information loss, thereby compromising performance. Other studies, while maintaining the number of input frames, adopt alternative sampling strategies instead of default uniform sampling to obtain higher-quality frames for input. In general, query relevance is the primary criterion for selecting frames that are semantically closest to the query (Yu et al., 2023; Lin et al., 2024b; Wang et al., 2024d;a; Suo et al., 2025). Methods such as AKS (Tang et al., 2025), BOLT (Liu et al., 2025b), Frame-Voyager (Yu et al., 2025) further propose adaptive sampling, inverse transform sampling, and optimal frame combination sampling to identify keyframes that are both query-relevant and temporally distinctive. Nevertheless, the methods often rely on manually designed heuristics without principled theoretical guidance, and the selected keyframes are often undistributed and distant over long intervals, especially in hours-long videos (e.g., 3600 frames per hour at 1 fps). This temporal sparsity weakens the relationships between frames and can cause confusion in MLLM inference.

In contrast to previous works, we formulate long video keyframe selection as a graph-based optimization problem with a clearly defined objective, and further leverage the efficiency of textual descriptions. Our approach jointly considers query relevance, content diversity, and temporal continuity, aiming to construct optimal combinations of keyframes with interleaved narratives, under the constraints of MLLM context length.

3 METHOD

3.1 KFC: KEYFRAME CAPTURING

Uniform sampling is commonly used in *short* video understanding for consistent temporal structure. However, for *long* videos, it often misses important information with limited input. While recent works emphasize selecting query-relevant frames for long video QA, they tend to overlook the problem of narrow focus due to the high similarity between adjacent frames. To address this, we first propose a keyframe capturing method that simultaneously considers query-relevance and frame-diversity, modeling the selection process as subgraph selection problem.

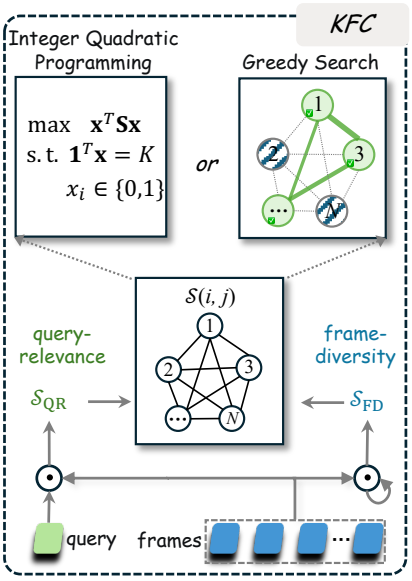

Figure 1: Illustration of keyframe capturing (KFC). $\mathcal{S}_{\text{QR}}$ and $\mathcal{S}_{\text{FD}}$ scores are computed via inner dot production.

Preliminaries. General video understanding tasks, e.g., video summarization and grounding (Liu et al., 2024b; Xiao et al., 2024) and long-video QA, can be similarly formulated as $(V, q) \rightarrow \text{Answer}$, where $V = \{f_i\}_{i=1}^{N}$ represents a video with N frames, f_i is the i-th frame, and q is the query. Considering an MLLM model as a neural function $\mathcal{M}(\cdot)$ with its limited contextual perceiving length, the normal video QA process reasoned by an MLLM model can be formulated as $\mathcal{M}(\{f_i\}_{i=1}^{K}, q) \rightarrow \text{Answer}, 1 \leq K \ll N$, meaning that only K frames are captured for representing video V. We next consider two criteria for selecting the K frames, query-relevance and frame-diversity.

Query-relevance. Since different questions can be asked on the single video, it is crucial to identify frames that correspond to a specific query first. Here, a standard two-stream vision-language model

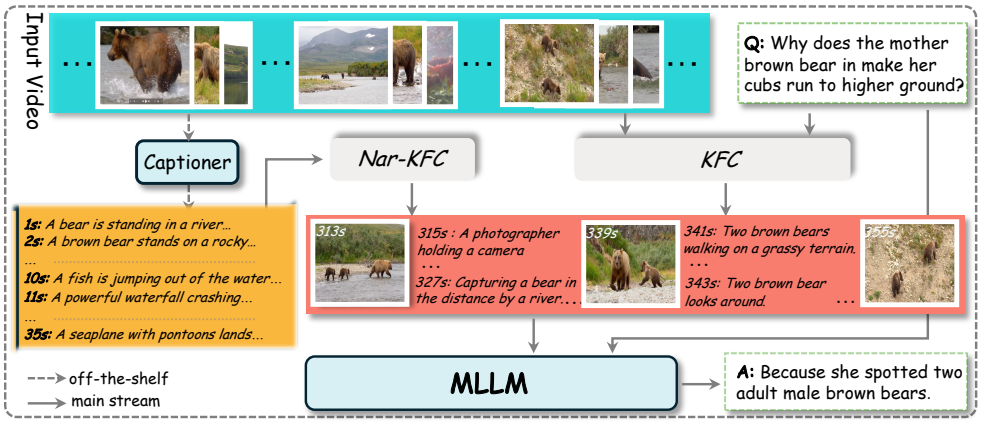

Figure 2: Illustration of Nar-KFC. We represent long videos by threading KFC-optimized keyframes with temporally interleaved narratives, where the narratives are generated frame-wise by an off-the-shelf captioner. Nar-KFC constructs a continuous representation to facilitate MLLM inference.

(VLM), e.g., CLIP (Radford et al., 2021), is used to extract embeddings $\{\mathbf{f}_i\}_{i=1}^N$ and $\mathbf{q}$ for the frames and the query, respectively. After standard normalization of all embeddings, the query-relevance score $\mathcal{S}_{\text{QR}}$ is computed as the cosine similarity between the two, $\mathcal{S}_{\text{QR}}(i) = \text{sim}(\mathbf{f}_i, \mathbf{q})$.

Frame-diversity. To avoid retrieving query-relevant only frames that are narrowly located in a small time range, we explicitly encourage diversified content when choosing the K frames. In particular, we use the *inverse* of cosine similarity between every pair of frame embeddings (normalized) to represent the diversity score. The function $\exp(\cdot)$ is applied to constrain the score between 0 and 1, formulated as $\mathcal{S}_{\text{FD}}(i,j) = \exp(-\text{sim}(\mathbf{f}_i, \mathbf{f}_j))$.

Objective. The final score combines $\mathcal{S}_{\text{QR}}$ and $\mathcal{S}_{\text{FD}}$ to jointly identify keyframes that are both query-relevant and diversified for KFC,

$$\mathcal{S}(i,j) = \mathcal{S}_{\text{QR}}(i) + \mathcal{S}_{\text{FD}}(i,j) = \text{sim}(\mathbf{f}_i, \mathbf{q}) + \exp(-\text{sim}(\mathbf{f}_i, \mathbf{f}_j)). \tag{1}$$

Next, as illustrated in Fig.1, we construct a graph where each node is a frame, and the edge weight between node pair (i,j) is $\mathcal{S}(i,j)$. The selection of K keyframes can then be cast as a subgraph selection problem with the original objective as follows: *given N nodes (frames), construct a subgraph by selecting K nodes (keyframes) so as to maximize the total edge weight of the subgraph.* Mathematically, this objective can be expressed as the optimization problem:

$$\max_{Y \subset \{1,\cdots,N\}, |Y|=K} \sum_{(i,j)\in\mathcal{I}} \mathcal{S}(i,j), \tag{2}$$

where $Y = \{y_1, \cdots, y_K\}$ is the index set of the K keyframes and $\mathcal{I}$ denotes all pairs (i,j).

3.1.1 Theoretical optimum: Integer Quadratic Programming

Our objective closely resembles the classic Knapsack problem (Salkin & De Kluyver, 1975), which can be commonly solved by dynamic programming or integer linear programming. The problem in (2) can be rewritten equivalently as an **integer quadratic programming (IQP)** problem,

$$\max_{\mathbf{x}} \ \mathbf{x}^T\mathbf{S}\mathbf{x} \quad \text{s.t. } \mathbf{1}^T\mathbf{x} = K,\ x_i \in \{0,1\}, \tag{3}$$

where $x_i = 1$ indicates that the i-th frame is selected, $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T$, and $\mathbf{S} \in \mathbb{R}^{N\times N}$ is the score matrix with $\mathbf{S}_{i,j} = \mathcal{S}(i,j)$ for $i < j$, and $\mathbf{S}_{i,j} = 0$ otherwise. Here, only the upper triangle of $\mathbf{S}$ is considered. A discussion of symmetrical $\mathbf{S}$ is detailed in Appendix §E.1. The search space is $C(N,K)$, and the time complexity of solving IQP is exponential regardless of whether the objective is convex or non-convex, making it impractical to get exact solutions in real cases. Modern optimization tools, e.g., CPLEX (Bliek1ú et al., 2014), typically address this by relaxing the binary constraint and allowing $x_i \in [0,1]$, converting the problem into a continuous optimization task. Solutions can then be obtained using methods like interior-point or Lagrange multiplier methods,

with a complexity of $\mathcal{O}(N^3)$. Subsequently, the Branch & Bound algorithm (Morrison et al., 2016) is applied to prune the search space and retrieve optimal integer solutions of x_i, but the worst-case time complexity remains exponential.

3.1.2 PRACTICALLY FEASIBLE APPROACH: GREEDY SEARCH

Solving the IQP optimally is computationally intractable for large N, e.g., long videos with thousands of frames. To search keyframes within practical latency constraints, we propose an efficient **greedy search (GS)** strategy that yields robust and near-optimal effects to the IQP solution. We first pre-process the score matrix to reduce noise across adjacent columns/rows, and shrinks the problem size for greater computational efficiency. Specifically, we apply singular value decomposition (SVD) to the score matrix $\mathbf{S}$, retaining the top r singular values to construct a low-rank approximation $\mathbf{S}_r \in \mathbb{R}^{N\times N}$. This matrix is then uniformly downsampled to $\mathbf{S}_{rd} \in \mathbb{R}^{\frac{N}{d}\times\frac{N}{d}}$ with a downsampling ratio d. The GS algorithm begins by selecting the most query-relevant frame as the starting point. It then iteratively adds the frame with the highest *cumulative* score relative to the already selected frames. In the final refinement step, the algorithm examines the k-nearest neighbors of each selected frame y_i, replacing y_i with a neighboring frame if it yields a higher cumulative score based on $\mathcal{S}_r$. A summary of the algorithm is provided in Alg. 1, and its overall time complexity is $\mathcal{O}(NK)$.

Algorithm 1: Practically Feasible Approach with Greedy Search

Input: Query-relevant score $\mathcal{S}_{\text{QR}}$, score matrix $\mathbf{S}$, number of retained singular values r, downsample ratio d, number of frames N, neighbor window k.
Output: Indices of selected K frames set $Y = \{y_1, y_2, \cdots, y_K\}$

$\mathbf{S}_r \leftarrow \text{LowRank}(\mathbf{S})$; $\mathbf{S}_{rd} \leftarrow \text{Downsample}(\mathbf{S}_r, d)$; `// Decompose and downsample` $\mathbf{S}$
$y_1 = \text{argmax}_i\, \mathcal{S}_{\text{QR}}(i)$; $Y \leftarrow \{y_1\}$ `// Initialize with most query-relevant frame`
for $i \leftarrow 2$ **to** K **do**
  for $j \leftarrow 1$ **to** N **do**
    $y_i = \text{argmax}_j \sum_{y\in Y} S_{rd}(y, y_j)$ `// Select frame with highest sum`
    $Y \leftarrow Y \cup y_i$
for $i \leftarrow 1$ **to** K **do**
  $y_i = \text{Refine}(y_i, k|\mathbf{S}_r)$; `// Refine selection within` k`-nearest neighbors`
return $Y = \text{sorted}\{y_1, y_2, \cdots, y_K\}$;

3.2 NAR-KFC: THREADING KEYFRAME WITH NARRATIVES

Keyframes captured by KFC significantly enhance the performance of MLLMs compared to the default uniform inference mechanism. However, it overlooks the *temporal-continuity* in frame sequences. Due to the severely uneven distribution of selected frames, temporal relationships become weak, often leading to confusion during inference.

To this end, we propose **Nar-KFC**, which threads keyframes with text narratives to construct a continuous and coherent input in an interleaved form. Specifically, we first use a lightweight off-the-shelf captioner, e.g., Qwen2-VL-2B, to generate captions $\{c_i\}_{i=1}^{N}$ for non-keyframes using a simple prompt as "`<USER> Describe this video frame in no more than 15 words.`" Given the unevenly distributed keyframes $\{f_{y_i}\}_{i=1}^{K}$ from KFC, we insert *captions from non-keyframes* between the keyframes, arranging them according to their true temporal order. Each y_i denotes the timestamp, and a uniform interval $\triangle$ is set between captions to control the total number of inserted narratives. The overall long video inference to a MLLM model $\mathcal{M}$ is formulated as:

$$\mathcal{M}\big(\{f_{y_1}, c_{y_1+\triangle}, \cdots, c_{y_2-\triangle}, f_{y_2}, c_{y_2+\triangle}, \cdots, c_{y_K-\triangle}, f_{y_K}\}, q\big) \rightarrow \text{Answer}. \tag{4}$$

Viability of Nar-KFC. MLLMs are typically trained via instruction tuning on both visual and textual modalities, making them well-suited to process our interleaved inputs of keyframes and narratives.
Rationality of Nar-KFC. The approach provides a temporally continous input that helps MLLMs "narrate" the story between keyframes. From another perspective, Nar-KFC can be seen as a form of compression, retaining only the most informative keyframes, while representing less critical segments with brief textual descriptions. This complementary two-stream mechanism is analogous to method like Two-Stream (Simonyan & Zisserman, 2014), which combines RGB frames with

optical flow. Also, it shares conceptual similarities with SlowFast (Feichtenhofer et al., 2019) and SlowFast-LLaVA (Xu et al., 2024b), where the caption stream serves as *a fast branch* traversing a broader temporal range (as in the low frame rate of the slow branch in SlowFast). These mechanisms together help explain the effectiveness of Nar-KFC in (long) video understanding.

4 Experiments

4.1 Experiment Settings

Evaluation Benchmarks. We evaluate our methods on several widely-used long-video question-answering benchmarks:: 1) *Video-MME* (Fu et al., 2025), consisting of 2,700 human-annotated QA pairs, with an average video duration of 17 min; 2) *LongVideoBench* (Wu et al., 2024b) validation set (denoted as LVB), which contains 1,337 QA pairs with average duration of 12 min; 3) *MLVU* (Zhou et al., 2025), where we use the multiple-choice task (M-avg), comprising 2,593 questions across 9 categories, with an average duration of 12 min. We provide more results of on relatively short EgoSchema (3 min) (Mangalam et al., 2023) and NExTQA (44 sec) (Xiao et al., 2021) benchmarks in Appendix §D.3. Furthermore, we evaluate open-ended generation performance on MMBench-Video (Fang et al., 2024) and MLVU-OpenEnded (Zhou et al., 2025) (G-avg), to verify the fine-grained capabilities of our methods.

Evaluation Models. We consider multiple advanced MLLMs, including InternVL2 (Chen et al., 2024c), Qwen2.5-VL (Bai et al., 2025), LLaVA-OneVision (Li et al., 2024b), LLaVA-Video (Zhang et al., 2024d), and InternVL3 (Zhu et al., 2025), to verify the effectiveness of our method. In Appendix Tab. 8, we further report performance with very recent Qwen3-VL (Team, 2025) model. We re-implement baseline results (uniform sampling) of these MLLMs using VLMEvalKit (Duan et al., 2024), which may yield slight differences compared to other public toolkits, e.g., LMMs-Eval (Li et al., 2024a).

Implementation Details. We use CLIP-ViT-L-336px (Radford et al., 2021) to extract query and video frame embeddings. Candidate frames are sampled from raw videos at 1 fps. For solving the IQP, we limit the maximum search nodes to 40k in CPLEX (Bliek1ú et al., 2014). In our customized greedy search algorithm, we empirically retain the top $\frac{N}{4}$ singular values to form the low-rank approximation of the score matrix $\mathbf{S}$ and further downsample it to a fixed resolution of 128×128 following previous work (Yu et al., 2025; Sun et al., 2025). The refinement window size k is set to 2 (see ablations of hyperparameters in Appendix §E.3). Unless otherwise stated, all ablations are conducted using the InternVL2 model on Video-MME. Experiments are run on 8 A100 GPUs.

4.2 Benchmark Results

Comparisons with State-of-the-Arts. We conduct comprehensive comparisons between our approach and several recent MLLMs and VideoLLMs in Tab. 1. Earlier works, e.g., Video-LLaVA (Lin et al., 2024a), Chat-UniVi-V1.5 (Jin et al., 2024), VideoLLaMA2 (Cheng et al., 2024), *etc*, are fully included in Appendix §D.1. Our methods, KFC and Nar-KFC, deliver consistent and significant gain over five baselines across three long-video benchmarks. On *Video-MME* (no sub.), Nar-KFC outperforms five MLLM baselines by 4.38% in average. Using the strongest baseline, i.e., InternVL3, Nar-KFC achieves state-of-the-art performance (63.8%), surpassing previous VideoLLMs - even those using larger LLMs (e.g., VILA-34B, 58.3%) or more frames (e.g., Video-XL$^{256\text{frm}}$, 55.5%). Incorporating larger numbers of frames may introduce noise and irrelevant information, which can be well addressed by our keyframe capturing and narrating strategies. On *LVB*, our method also achieves notable performance improvements, e.g., 52.3% *vs.* 53.9% with InternVL2 and 52.7% *vs.* 55.3% with Qwen2.5-VL, although the overall gain is partly offset by videos shorter than 1 min, demonstrating clear advantages in long video understanding. On *MLVU*, our KFC-only strategy (without narrations) yields an average improvement of over 6% across five MLLMs. The use of query-relevant and diverse keyframes significantly boosts performance on Needle-in-a-haystack (Zhang et al., 2024b) and counting questions. Furthermore, appending narratives provides additional and robust gains by preserving temporal continuity. Detailed analysis are further presented in Appendix §D.4.

Comparisons with varying number of keyframes. In Fig. 3, we compare KFC and Nar-KFC against uniform sampling with varying frames across three benchmarks and three models. Due to Qwen2.5-VL's dynamic resolution mechanism (Dehghani et al., 2023), increasing keyframes often leads to memory overflow, so its results are omitted. Notably, Nar-KFC shows substantial gains when the number of keyframes is limited (e.g., 4 or 8), due to its ability to provide broad video coverage via interleaved textual narratives. As the number of keyframes increases, the performance

Table 1: Comparisons with previous VideoLLMs on three common long-video benchmarks: Video-MME, LVB, and MLVU. All methods are evaluated using 8 frames. For Video-MME, we report performance with two standard settings: without subtitles (no sub.) and with subtitles (sub.). LVB denotes the LongVideoBench set. Methods that use significantly more frames and larger-sized LLM are marked in gray. The reported results are accuracy percentage.

Model	Size	Video-MME$_{\text{(no sub. / sub.)}}$				LVB	MLVU
		Short	Medium	Long	Overall$_{\sim 17m}$	$_{\sim 12m}$	$_{\sim 12m}$
VILA (Lin et al., 2024b)	8B	57.8 / 61.6	44.3 / 46.2	40.3 / 42.1	47.5 / 50.0	-	46.3
LLaVA-NeXT-QW2 (Liu et al., 2024a)	7B	58.0 / -	47.0 / -	43.4 / -	49.5 / -	-	-
MiniCPM-V2.6 (Yao et al., 2024b)	7B	61.1 / 63.8	50.3 / 50.2	46.4 / 45.4	52.6 / 53.1	51.2	55.4
LongVU (Shen et al., 2024)	7B	64.7 / -	58.2 / -	59.5 / -	60.6 / -	-	65.4
BOLT (Liu et al., 2025b)	7B	66.8 / -	54.2 / -	47.3 / -	56.1 / -	55.6	63.4
Frame-Voyager (Yu et al., 2025)	8B	67.3 / -	56.3 / -	48.9 / -	57.5 / -	-	65.6
LongVILA^{256frm} (Chen et al., 2024b)	8B	61.8 / -	49.7 / -	39.7 / -	50.5 / -	-	-
Video-XL^{256frm} (Shu et al., 2025)	7B	64.0 / 67.4	53.2 / 60.7	49.2 / 54.9	55.5 / 61.0	50.7	64.9
LLaVA-NeXT-Video (Zhang et al., 2024c)	34B	61.7 / 65.1	50.1 / 52.2	44.3 / 47.2	52.0 / 54.9	50.5	58.8
VILA (Lin et al., 2024b)	34B	70.3 / 73.1	58.3 / 62.7	51.2 / 55.7	58.3 / 61.6	-	57.8
InternVL2 (Chen et al., 2024c)	8B	62.1 / 63.9	48.2 / 48.7	45.2 / 44.9	51.9 / 52.5	52.3	54.3
+ KFC	8B	64.5 / 65.4	50.0 / 52.3	46.5 / 47.3	53.5 / 55.0	53.3	62.2
+ Nar-KFC	8B	67.2 / 67.7	54.7 / 57.9	47.1 / 48.9	**56.3 / 58.1**	**53.9**	**64.4**
Qwen2.5-VL (Bai et al., 2025)	7B	65.9 / 66.4	54.4 / 54.3	45.8 / 46.9	55.4 / 55.9	52.7	55.8
+ KFC	7B	68.8 / 70.7	52.6 / 54.9	49.3 / 51.4	56.9 / **59.0**	54.3	62.6
+ Nar-KFC	7B	70.1 / 71.0	54.4 / 55.2	49.0 / 49.4	**57.9** / 58.6	**55.3**	**64.4**
LLaVA-OneVision (Li et al., 2024b)	7B	65.2 / 67.1	51.7 / 54.4	45.1 / 46.1	53.3 / 55.9	54.5	58.5
+ KFC	7B	66.4 / 69.1	52.9 / 56.8	46.8 / 48.8	55.4 / 58.2	55.6	65.0
+ Nar-KFC	7B	67.2 / 68.6	57.1 / 59.8	49.1 / 51.0	**57.8 / 59.8**	**56.5**	**66.2**
LLaVA-Video (Zhang et al., 2024d)	7B	67.2 / 69.4	53.2 / 53.4	47.2 / 47.3	55.9 / 56.7	54.2	60.5
+ KFC	7B	68.3 / 70.0	55.1 / 57.4	49.4 / 51.6	57.6 / 59.7	56.5	66.9
+ Nar-KFC	7B	71.2 / 72.7	61.4 / 62.3	52.0 / 53.9	**61.6 / 63.0**	**57.7**	**67.7**
InternVL3 (Zhu et al., 2025)	8B	68.7 / 70.9	58.3 / 58.2	50.0 / 50.9	59.0 / 60.0	53.6	60.9
+ KFC	8B	70.9 / 71.9	60.6 / 60.1	50.9 / 51.8	60.8 / 61.4	54.5	67.5
+ Nar-KFC	8B	72.9 / 73.9	62.9 / 62.7	55.7 / 55.8	**63.8 / 64.1**	**54.8**	**68.4**

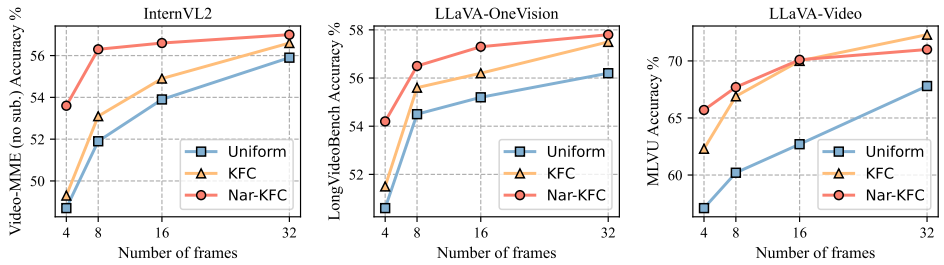

Figure 3: Accuracies (%) of uniform sampling, KFC, and Nar-KFC versus numbers of keyframes.

gap between uniform sampling and our methods narrows. This can be attributed to: 1) uniform sampling is more likely to capture key moments when more frames are used; and 2) many video QA questions typically only require a few number of frames to accurately answer in current benchmarks. Interestingly, on MLVU, KFC alone outperforms Nar-KFC with 32 keyframes, suggesting that when sufficient keyframes are present, the added benefit of narratives diminishes. These results underscore the strength of KFC in selecting informative keyframes while demonstrating that narratives are particularly valuable when MLLMs have limited context capacity. We further scale Nar-KFC to 72B models and compare them with proprietary models and SOTA VideoLLMs in Appendix §D.2.

Improvements on open-ended generation tasks. In Tab. 2), we show Nar-KFC consistently improves performance on open-ended generation tasks that require fine-grained reasoning, demonstrating that even with global-level frame selection and captioning, our methods can still enhance fine-grained tasks. Interestingly, we find that KFC shows decreased performance on the MLVU-OpenEnded summary task, likely because uniformly sampled frames cover the entire video range, whereas KFC-selected frames may be more concentrated. Our Nar-KFC addresses this issue by providing more comprehensive video information.

4.3 Ablation and Analysis

KFC and Nar-KFC ablations. We report the ablation results of KFC and Nar-KFC components on the Video-MME (sub.) and MLVU benchmarks in Tab. 3. Simply inserting narratives between

Table 2: Improvements on open-ended generation tasks. All results are reported based on 8-frame evaluation. For MMBench-Video, `GPT-4-1106` is used as the judge model, while for MLVU-OpenEnded, `GPT-4-0125` serves as the default judge model, following the official implementations.

Model	MMBench-Video			MLVU-OpenEnded		
	Perception	Reasoning	Overall	Sub_scene	Summary	G-Avg
InternVL3-8B (Zhu et al., 2025)	1.54	1.61	1.57	5.47	**4.40**	4.92
+ KFC	1.56	1.58	1.58	**5.73**	4.23	4.95
+ Nar-KFC	**1.76**	**1.78**	**1.78**	5.69	4.39	**5.02**
Qwen3-VL-8B (Team, 2025)	1.62	1.65	1.64	6.17	**6.01**	6.09
+ KFC	1.65	1.75	1.69	**6.32**	5.71	6.00
+ Nar-KFC	**1.75**	**1.76**	**1.76**	6.28	5.97	**6.12**

Table 3: Main component ablation results in Nar-KFC. "S, M, L" refer to short, medium, and long video categories in the Video-MME (sub.) benchmark.

Strategy	Video-MME				MLVU	Time
	S	M	L	Overall		
Uniform	63.9	48.7	44.9	52.5	54.3	$\mathcal{O}(1)$
+ **Nar**ratives	66.1	54.9	45.2	55.4	59.4	$\mathcal{O}(N)$
KFC (IQP)	65.9	52.9	46.4	55.1	62.0	$\mathcal{O}(2^N)$
KFC (GS)	65.4	52.3	47.3	55.0	62.2	$\mathcal{O}(NK)$
w/o $\mathcal{S}_{\text{QR}}$	62.3	47.8	45.3	51.8	57.3	$\mathcal{O}(NK)$
w/o $\mathcal{S}_{\text{FD}}$	63.6	49.4	44.6	52.5	60.9	$\mathcal{O}(NK)$
Nar-KFC	**67.7**	**57.9**	**48.9**	**58.1**	**64.4**	$\mathcal{O}(NK)$

Table 4: Effects of including pre-processing and refinement stages in the KFC Greedy Search (GS) method. V-MME denotes the overall Video-MME (sub). Line (ii′) indicates Downsampling *without* LowRank. The final KFC (GS) strategy integrates all components from (i) to (iv).

Ex#	Strategy	V-MME	MLVU
	Vanilla GS	52.3	60.4
(i)	+ Initialization	53.3	61.0
(ii)	+ LowRank	53.7	61.8
(ii′)	+ Downsample	53.9	61.6
(iii)	+ LowRank + Downsample	54.7	**62.2**
(iv)	+ Refinement (**KFC**)	**55.0**	**62.2**

uniformly sampled frames yields improvements of 2.9% on Video-MME and 5.1% on MLVU, indicating that adding narrative context, despite with frames not being query-specific, can effectively boost overall video understanding. To retrieve query-relevant and diverse keyframes, our Greedy Search (GS) strategy achieves results comparable to the optimal Integer Quadratic Programming (IQP) method (55.0% *vs.* 55.1% on Video-MME and 62.2% *vs.* 62.0% on MLVU), while being significantly more efficient with $\mathcal{O}(NK)$ complexity. Details of our IQP implementation and comparisons with GS are provided in Appendix §E.2. Further ablations show that removing the query-relevance score $\mathcal{S}_{\text{QR}}$ leads to a 3.2% drop on Video-MME and 4.9% on MLVU with greedy search. This emphasizes that retrieving query-relevant frames is critical in long videoQA. Meanwhile, incorporating frame diversity $\mathcal{S}_{\text{FD}}$ further stabilizes and enhances performance across benchmarks. When threading all keyframes with interleaved narratives, Nar-KFC achieves the best overall results on all metrics, underscoring its solid effectiveness in representing long video contents.

Component analysis of greedy search (GS). Starting from the vanilla GS, which iteratively selects the frame with the highest cumulative score relative to the already selected frames, we progressively incorporate several techniques (Tab. 4) to enhance its effectiveness to a near-optimal solution: (i) initialization with the frame most relevant to the query brings a modest yet consistent gain (from 52.3%→53.3% on Video-MME, and 60.4%→61.0% on MLVU); (ii and iii) applying low-rank denoising and downsampling further improves performance by producing a more compact and less noisy score matrix $\mathbf{S}$; and (iv) adding the final refinement step, KFC (GS) achieves the best results of 55.0% on Video-MME and 62.2% on MLVU. This highlights the cumulative benefit of combining compact frame representations, reduced redundancy, and an iterative selection mechanism.

Comparisons with other keyframe selection methods. We compare KFC with several keyframe extraction baselines in Tab. 5, all utilizing the InternVL2 backbone and 8 frames. Details are in Appendix §E.4. Methods that apply top-K frame-query matching using SigLIP (Zhai et al., 2023), or BLIP-2 (Li et al., 2023a) embeddings perform worse than uniform sampling, possibly due to keyframes being concentrated within a narrow temporal window. For those localize-then-answer methods, i.e., TempGQA (Xiao et al., 2024) and SeViLA (Yu et al., 2023), performance heavily depends on the quality of segment localization, which can be unreliable. Recent approaches including DPP (Sun et al., 2025), AKS (Tang et al., 2025), and BOLT (Liu et al., 2025b) generally yield better results by incorporating frame diversity. However, these methods rely on handcrafted and heuristic sampling strategies, lacking

Table 5: Comparisons with different frame selection methods on Video-MME.

	V-MME$_{\text{(no sub./sub.)}}$
InternVL2	51.9 / 52.5
+ CLIP (top-K)	47.7 / 50.0
+ SigLIP (top-K)	47.3 / 51.0
+ BLIP-2 (top-K)	47.8 / 50.9
+ TempGQA	50.4 / 51.1
+ SeViLA	52.2 / 53.7
+ DPP	52.2 / 53.5
+ AKS	52.8 / 53.9
+ BOLT	53.3 / -
+ KFC (Ours)	**53.5 / 55.0**

Table 6: Analysis of video input components on Video-MME (no sub). Superscript numbers indicate the quantity. Average time and tokens per video are reported.

Components	V-MME	Latency (s)	TFLOPs ↓	Token#
Narratives^{210}	51.1	0.98	109.6	4,725
Frames^{8} (uniform)	51.9	1.03	146.3	6,280
Frames^{8} (KFC)	53.5	1.31	146.3	6,280
Interleave^{8+210} (Nar-KFC)	56.3	2.13	202.6	11,005

Table 7: Temporal structure analysis between narratives and keyframes on Video-MME (no sub) benchmark.

Temporal Structure	V-MME
{Narrative}→{Keyframe}→{Query}	55.5
{Keyframe}→{Narrative}→{Query}	55.3
Interleave (Nar-KFC)→{Query}	56.3

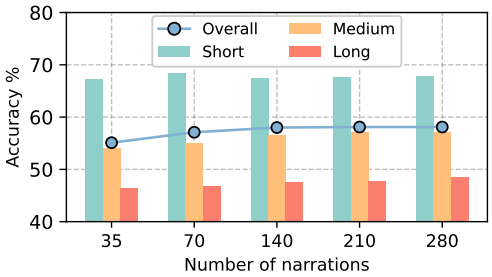

Figure 4: Effect of the total number of inserted narratives, corresponding to the narrative interval △, across videos of different lengths.

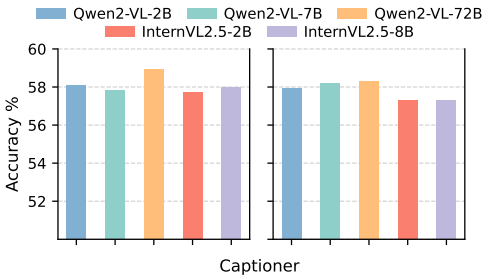

Figure 5: Impact of different captioners for generating narratives. Video-MME (sub.) results are for InternVL2-8B (left) and Qwen2-VL-7B (right).

a principled and generalized frame selection guidance. In comparison, our proposed KFC consistently outperforms all baselines, demonstrating clear superiority in subset frame selection.

Effect of narrative quantity. Due to the varying length of videos, we do not directly ablate the effect on a fixed interval value △. Instead, we control the total number of narratives appended, as shown Fig. 4 on Video-MME (sub.). Narratives are incrementally added across 7 intervals between 8 keyframes. The overall accuracy improves steadily from 55.1% to 58.1% as more narratives are available, with more performance gains on medium and long videos. However, since adjacent frames often contain similar visual information, adding more narratives results in diminishing returns due to redundant descriptions. We thus use 210 narratives as the default.

Effect of narrative quality. Fig. 5 presents the impact of different captioners on the quality of generated narratives and the resulting performance of Nar-KFC on the Video-MME (sub.). We evaluate five MLLMs of varying sizes and sources as captioners. Narratives extracted from the largest captioner, Qwen2-VL-72B, achieves the best accuracy, i.e., 58.9% on InternVL2-8B and 58.3% on Qwen2-VL-7B, highlighting the benefit of higher-quality narratives. Nevertheless, the overall performance gap across all captioners is small (less than 1%). This suggests that keyframes play a dominant role in long video understanding, while captions serve as auxiliary and supportive context. We thus use the lightweight Qwen2-VL-2B as the default captioner for other benchmarks.

Efficiency and effectiveness between narratives and keyframes. We decompose Nar-KFC into standalone narratives and frames in Tab. 6. Although translated from 210 frames, pure narratives perform worse than even 8 uniformly sampled frames (51.1% *vs.* 51.9%), which reflects that substantial information is lost during the frame-to-caption conversion. Nevertheless, narratives exhibit advantages with the shortest latency (0.98s) and the fewest tokens (4,725 per video). Combining narratives with KFC-selected keyframes (Nar-KFC) achieves both the best accuracy and also maintains reasonable efficiency. We discuss detailed computational overhead in Appendix §D.5. In addition, Tab. 7 investigates the temporal structure between narratives and keyframes. Placing all keyframes either before or after the narratives degrades the performance by 0.8% and 1.0%, likely due to disrupted temporal sequences. In contrast, interleaving narratives and frames, as in Nar-KFC, yields superior results. These findings further validate our primary goal: constructing temporally continuous representations for long video understanding.

4.4 Qualitative Results

Fig. 6 presents two qualitative examples of our method. In the first example (left), our KFC effectively identifies frames that are both query-relevant and content-diverse, resulting in the correct answer. In the second example (right), we demonstrate that Nar-KFC substantially improves reasoning in a complete relay race scenario by threading temporally interleaved keyframes with coherent narratives.

Figure 6: Qualitative results. (left) Comparison of frames selected by uniform sampling, top-K sampling, and our KFC. (right) Key narratives generated by Nar-KFC that lead to the correct answer. Zoom in for details.

This enables accurate inference of the final winner, whereas KFC fails due to limited number of frames. More examples can be found in Appendix §F.

5 Conclusion

In this paper, we propose a keyframe capturing strategy (KFC) and a narrating keyframe method (Nar-KFC) to boost existing MLLMs for long video understanding, under the constraint of limited context length in language models. Our approach constructs long video representations that are query-relevant, content-diverse, and temporally continuous, all achieved in a training-free manner. This significantly improves the performance of current MLLMs on widely-used long video benchmarks. Our findings strongly validate the potential of MLLMs as effective long video comprehenders.

References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems (NeurIPS), 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 1(2):3, 2023.

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

Christian Bliek1ú, Pierre Bonami, and Andrea Lodi. Solving mixed-integer quadratic programming problems with ibm-cplex: a progress report. In Proceedings of the twenty-sixth RAMP symposium, pp. 16–17, 2014.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems (NeurIPS), 33:1877–1901, 2020.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. Advances in Neural Information Processing Systems (NeurIPS), 37:53168–53197, 2024.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. Advances in Neural Information Processing Systems (NeurIPS), 37:19472–19495, 2024a.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. arXiv preprint arXiv:2408.10188, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24185–24198, 2024c.

Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. Scaling video-language models to 10k frames via hierarchical differential distillation. arXiv preprint arXiv:2504.02438, 2025a.

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, 2025b.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476, 2024.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. Advances in Neural Information Processing Systems (NeurIPS), 36:2252–2274, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR), 2020.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In Proceedings of the 32nd ACM International Conference on Multimedia (MM), pp. 11198–11201, 2024.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. NeurIPS, 37:89098–89124, 2024.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp. 6202–6211, 2019.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 24108–24118, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 13702–13712, 2025.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13700–13710, 2024.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. In Workshop on Video-Language Models@ NeurIPS 2024, 2024.

Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. IEEE Access, 2024.

Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024a. URL https://github.com/EvolvingLMMs-Lab/lmms-eval.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning (ICML), pp. 19730–19742. PMLR, 2023a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355, 2023b.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. arXiv preprint arXiv:2501.00574, 2024c.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In European Conference on Computer Vision (ECCV), pp. 323–340. Springer, 2024d.

Jianxin Liang, Xiaojun Meng, Yueqian Wang, Chang Liu, Qun Liu, and Dongyan Zhao. End-to-end video question answering with frame scoring mechanisms and adaptive sampling. arXiv preprint arXiv:2407.15047, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5971–5984, 2024a.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 26689–26699, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems (NeurIPS), 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Huabin Liu, Xiao Ma, Cheng Zhong, Yang Zhang, and Weiyao Lin. Timecraft: Navigate weakly-supervised temporal grounded video question answering via bi-directional reasoning. In European Conference on Computer Vision (ECCV), pp. 92–107. Springer, 2024b.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics (TACL), 12:157–173, 2024c.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In European Conference on Computer Vision (ECCV), pp. 1–18. Springer, 2024d.

Ruyang Liu, Shangkun Sun, Haoran Tang, Wei Gao, and Ge Li. Flow4agent: Long-form video understanding via motion prior from optical flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 23817–23827, 2025a.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 3318–3327, 2025b.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? arXiv preprint arXiv:2403.00476, 2024e.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 18936–18946, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems (NeurIPS), 36:46212–46244, 2023.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13235–13245, 2024.

David R Morrison, Sheldon H Jacobson, Jason J Sauppe, and Edward C Sewell. Branch-and-bound algorithms: A survey of recent advances in searching, branching, and pruning. Discrete Optimization, 19:79–102, 2016.

OpenAI. Chatgpt: Optimizing language models for dialogue, 2023. URL https://openai.com/blog/chatgpt.

Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. In Workshop on Video-Language Models@ NeurIPS 2024, 2024.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. In International Conference on Machine Learning (ICML), pp. 41340–41356. PMLR, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning (ICML), pp. 8748–8763. PmLR, 2021.

Harvey M Salkin and Cornelis A De Kluyver. The knapsack problem: a survey. Naval Research Logistics Quarterly, 22(1):127–144, 1975.

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

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 26160–26169, 2025.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems (NeurIPS), 27, 2014.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18221–18232, 2024.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp3: A training-free approach for list-wise frame selection in video-llms. arXiv preprint arXiv:2501.02885, 2025.

Yucheng Suo, Fan Ma, Linchao Zhu, Tianyi Wang, Fengyun Rao, and Yi Yang. From trial to triumph: Advancing long video understanding via visual context sample scaling and self-reward alignment. arXiv preprint arXiv:2503.20472, 2025.

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 (CVPR), pp. 29118–29128, 2025.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295, 2024.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems (NeurIPS), 37: 87310–87356, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. Efficient large language models: A survey. Transactions on Machine Learning Research (TMLR), 2024.

Haibo Wang, Chenghang Lai, Yixuan Sun, and Weifeng Ge. Weakly supervised gaussian contrastive grounding with large multimodal models for video question answering. In Proceedings of the 32nd ACM International Conference on Multimedia (MM), pp. 5289–5298, 2024a.

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

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Retake: Reducing temporal and knowledge redundancy for long video understanding. arXiv preprint arXiv:2412.20504, 2024c.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. arXiv preprint arXiv:2503.12559, 2025a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In European Conference on Computer Vision (ECCV), pp. 58–76. Springer, 2024d.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191, 2022.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 3272–3283, 2025b.

Hao Wu, Huabin Liu, Yu Qiao, and Xiao Sun. Dibs: Enhancing dense video captioning with unlabeled videos via pseudo boundary enrichment and online refinement. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 18699–18708, 2024a.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems (NeurIPS), 37:28828–28857, 2024b.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 9777–9786, 2021.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13204–13214, 2024.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (NAACL), pp. 4643–4663, 2024.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. arXiv preprint arXiv:2404.16994, 2024a.

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. arXiv preprint arXiv:2407.15841, 2024b.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. Advances in Neural Information Processing Systems (NeurIPS), 35:124–141, 2022.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10714–10726, 2023.

Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for mllms. Advances in Neural Information Processing Systems (NeurIPS), 37:33108–33140, 2024a.

Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. Generative frame sampler for long video understanding. arXiv preprint arXiv:2503.09146, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024b.

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 Computer Vision and Pattern Recognition Conference (CVPR), pp. 8579–8591, 2025.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. Advances in Neural Information Processing Systems (NeurIPS), 36: 76749–76771, 2023.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models.(2025). In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR), pp. 24–28, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp. 11975–11986, 2023.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106, 2025.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 21715–21737, 2024a.

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

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024c. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713, 2024d.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 13691–13701, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023.

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

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. arXiv preprint arXiv:2412.10360, 2024.

APPENDIX

A Limitations and Future Work

We discuss limitations and possible extensions of Nar-KFC. Despite current MLLMs being able to process our interleaved inputs of keyframes and narratives, thanks to their instruction tuning step, they are not trained with such input formats. This may weaken their ability to fully understand the structure and relationships within our specialized long video representations. A valuable future direction is to incorporate keyframe selection and narrative interleaving into the training of MLLMs, thereby aligning training and testing procedures for improved long video understanding. Furthermore, our method relies mainly on interleaving visual information with narrations and does not incorporate additional modalities such as audio or subtitles. Exploring these modalities in future work may further improve multi-modal long video understanding.

B The Use of Large Language Models (LLMs)

In this paper, we exclusively utilize advanced LLMs to refine and polish the manuscript. Our prompts to the LLMs include requests such as: "`Please help me polish this academic writing paragraph. It should be concise, fluent, logical, and in line with academic standards.`" LLMs are not employed for any purposes beyond writing improvement.

C Broader Impacts

Effective and efficient long video understanding is a critical task, especially as Internet video streams often last tens of minutes or even hours. We expect that the proposed keyframe selection and narration methods will benefit society by enabling MLLMs to comprehend long videos more accurately and efficiently. However, it is essential to ensure that the narratives generated by specific models remain free from harmful or unrelated content.

D Main Results Supplementary

We provide supplementary results to the main experiments: Sec. D.1 covers earlier works. Sec. D.2 scales Nar-KFC to 72B models and compares its performance with proprietary models and VideoLLMs capable of reasoning over thousands of frames. Sec. D.3 presents the performance of KFC and Nar-KFC on additional EgoSchema and NExTQA benchmarks, and Sec. D.4 provides a detailed analysis on the MLVU benchmark. Finally, Sec. D.5 discusses the detailed computational overhead introduced by Nar-KFC.

D.1 Comprehensive Comparisons with Previous Methods.

VideoLLMs for video understanding have become a popular research area in recent years. However, directly applying previous VideoLLMs to long videos, such as Video-MME, LongVideoBench, and MLVU, often leads to unsatisfactory performance. To provide a more comprehensive comparison, as an extension to the main paper in Tab. 1, we also include the performance of earlier works, such as Video-LLaVA (Lin et al., 2024a), Qwen-VL-Chat (Bai et al., 2023), ST-LLM (Liu et al., 2024d), VideoChat2 (Li et al., 2023b), ShareGPT4Video (Chen et al., 2024a), Chat-UniVi-V1.5 (Jin et al., 2024), and VideoLLaMA2 (Cheng et al., 2024), in Tab. 8.

D.2 Scaling Nar-KFC to 72B models

To further evaluate the ability of Nar-KFC to enhance SOTA performance, we scale our Nar-KFC framework to two advanced models: LLaVA-OneVision-72B (32 frames) and LLaVA-Video-72B-Qwen2 (64 frames). We also compare our results with those of SOTA proprietary models and recent works, as shown in Tab. 9 and Tab. 10, with our results highlighted in bold. Extensive experiments demonstrate that the Nar-KFC framework enables 72B models to achieve competitive performance on

Table 8: Comprehensive comparisons with previous VideoLLMs/MLLMs on three common long-video benchmarks: Video-MME, LVB, and MLVU. The reported results are accuracy percentage.

Model	Size	Video-MME$_{(\text{no sub. / sub.})}$				LVB	MLVU
		Short	Medium	Long	Overall$_{\sim 17m}$	$\sim 12m$	$\sim 12m$
Video-LLaVA (Lin et al., 2024a)	7B	45.3 / 46.1	38.0 / 40.7	36.2 / 38.1	39.9 / 41.6	39.1	47.3
Qwen-VL-Chat (Bai et al., 2023)	7B	46.9 / 47.3	38.7 / 40.4	37.8 / 37.9	41.1 / 41.9	-	-
ST-LLM (Liu et al., 2024d)	7B	45.7 / 48.4	36.8 / 41.4	31.3 / 36.9	37.9 / 42.3	-	-
VideoChat2 (Li et al., 2023b)	7B	48.3 / 52.8	37.0 / 39.4	33.2 / 39.2	39.5 / 43.8	39.3	44.5
ShareGPT4Video (Chen et al., 2024a)	8B	48.3 / -	36.3 / -	35.0 / -	39.9 / -	41.8	46.4
Chat-UniVi-V1.5 (Jin et al., 2024)	7B	45.7 / 51.2	40.3 / 44.6	35.8 / 41.8	40.6 / 45.9	-	-
VideoLLaMA2 (Cheng et al., 2024)	7B	56.0 / -	45.4 / -	42.1 / -	47.9 / -	-	-
VILA (Lin et al., 2024b)	8B	57.8 / 61.6	44.3 / 46.2	40.3 / 42.1	47.5 / 50.0	-	46.3
LLaVA-NeXT-QW2 (Liu et al., 2024a)	7B	58.0 / -	47.0 / -	43.4 / -	49.5 / -	-	-
MiniCPM-V2.6 (Yao et al., 2024b)	7B	61.1 / 63.8	50.3 / 50.2	46.4 / 45.4	52.6 / 53.1	51.2	55.4
LongVU (Shen et al., 2024)	7B	64.7 / -	58.2 / -	59.5 / -	60.6 / -	-	65.4
Frame-Voyager (Yu et al., 2025)	8B	67.3 / -	56.3 / -	48.9 / -	57.5 / -	-	65.6
LongVILA$^{256\text{frm}}$ (Chen et al., 2024b)	8B	61.8 / -	49.7 / -	39.7 / -	50.5 / -	-	-
Video-XL$^{256\text{frm}}$ (Shu et al., 2025)	7B	64.0 / 67.4	53.2 / 60.7	49.2 / 54.9	55.5 / 61.0	50.7	64.9
VILA (Lin et al., 2024b)	34B	70.3 / 73.1	58.3 / 62.7	51.2 / 55.7	58.3 / 61.6	-	57.8
InternVL2 (Chen et al., 2024c)	8B	62.1 / 63.9	48.2 / 48.7	45.2 / 44.9	51.9 / 52.5	52.3	54.3
+ KFC	8B	64.3 / 65.4	49.6 / 52.3	46.1 / 47.3	53.1 / 55.0	53.3	62.2
+ Nar-KFC	8B	67.2 / 67.7	54.7 / 57.9	47.1 / 48.9	**56.3 / 58.1**	**53.9**	**64.4**
Qwen2-VL (Wang et al., 2024b)	7B	65.7 / 66.9	52.8 / 53.0	46.7 / 48.6	55.0 / 56.1	53.4	59.6
+ KFC	7B	68.2 / 69.7	53.3 / 54.9	48.4 / 50.2	**56.7 / 58.3**	**54.6**	65.9
+ Nar-KFC	7B	68.8 / 69.3	53.4 / 55.3	48.0 / 49.0	**56.7** / 57.9	53.6	**68.5**
Qwen2.5-VL (Bai et al., 2025)	7B	65.9 / 66.4	54.4 / 54.3	45.8 / 46.9	55.4 / 55.9	52.7	55.8
+ KFC	7B	68.8 / 70.7	52.6 / 54.9	49.3 / 51.4	56.9 / **59.0**	54.3	62.6
+ Nar-KFC	7B	70.1 / 71.0	54.4 / 55.2	49.0 / 49.4	**57.9** / 58.6	**55.3**	**64.4**
LLaVA-OneVision (Li et al., 2024b)	7B	65.2 / 67.1	51.7 / 54.4	45.1 / 46.1	53.3 / 55.9	54.5	58.5
+ KFC	7B	66.4 / 69.1	52.9 / 56.8	46.8 / 48.8	55.4 / 58.2	55.6	65.0
+ Nar-KFC	7B	67.2 / 68.6	57.1 / 59.8	49.1 / 51.0	**57.8 / 59.8**	**56.5**	**66.2**
LLaVA-Video (Zhang et al., 2024d)	7B	67.2 / 69.4	53.2 / 53.4	47.2 / 47.3	55.9 / 56.7	54.2	60.5
+ KFC	7B	68.3 / 70.0	55.1 / 57.4	49.4 / 51.6	57.6 / 59.7	56.5	66.9
+ Nar-KFC	7B	71.2 / 72.7	61.4 / 62.3	52.0 / 53.9	**61.6 / 63.0**	**57.7**	**67.7**
InternVL3 (Zhu et al., 2025)	8B	68.7 / 70.9	58.3 / 58.2	50.0 / 50.9	59.0 / 60.0	53.6	60.9
+ KFC	8B	70.9 / 71.9	60.6 / 60.1	50.9 / 51.8	60.8 / 61.4	54.5	67.5
+ Nar-KFC	8B	72.9 / 73.9	62.9 / 62.7	55.7 / 55.8	**63.8 / 64.1**	**54.8**	**68.4**
Qwen3-VL (Team, 2025)	8B	68.4 / 70.7	55.4 / 55.3	50.1 / 52.0	58.0 / 59.1	54.7	49.5
+ KFC	8B	68.4 / 71.9	57.3 / 57.0	50.8 / 50.7	58.9 / 59.9	55.8	63.0
+ Nar-KFC	8B	70.4 / 72.9	60.1 / 59.7	52.7 / 52.4	**61.1 / 61.7**	**56.2**	**65.8**

Table 9: Scaling to 72B models on the Video-MME benchmark. Results from our Nar-KFC method are in bold.

Model	Frames	Video-MME (no sub.)			
		Short	Medium	Long	Overall
LLaVA-OneVision-72B	32	76.7	62.2	60.0	66.3
+ **Nar-KFC**	32	**77.5**	**68.6**	**61.9**	**69.6**
LLaVA-Video-72B	64	81.7	67.9	61.8	70.4
+ **Nar-KFC**	64	**82.0**	**68.9**	**63.6**	**71.5**
VideoChat-Flash@448-7B (Li et al., 2024c)	N/A	-	-	-	65.3
LLaVA-OneVision-72B + T* (Ye et al., 2025)	32	77.5	66.6	61.0	68.3
VILAMP-7B (Cheng et al., 2025a)	1 fps	-	-	-	67.5
Aria-8x3.5B	256	76.9	67.0	58.8	67.6
GPT-4o (0615)	384	80.0	70.3	65.3	71.9
Qwen2-VL-72B (Wang et al., 2024b)	768	80.1	71.3	62.2	71.2
AdaReTake-72B (Wang et al., 2025a)	2 fps	-	-	-	73.5
Gemini-1.5-Pro (0615)	1/0.5 fps	81.7	74.3	67.4	75.0

Table 10: Scaling to 72B models on the MLVU benchmark. Results from our Nar-KFC method are shown in bold. * indicates results obtained from our own implementation.

Model	Frames	MLVU
LLaVA-OneVision-72B	32	66.4
+ **Nar-KFC**	32	**74.4**
LLaVA-Video-72B	64	74.4 (73.6*)
+ **Nar-KFC**	64	**75.0**
GPT-4o (0615)	0.5 fps	64.6
VideoLLaMA3-7B (Zhang et al., 2025)	≤180	73.0
VILAMP-7B (Cheng et al., 2025a)	1 fps	72.6
VideoChat-Flash@448-7B (Li et al., 2024c)	1 fps	74.7
AdaReTake-72B (Wang et al., 2025a)	2 fps	78.1

the Video-MME benchmark (71.5%) and leading results on MLVU (75.0%). Notably, our approach uses significantly fewer frames (32 or 64) compared to proprietary models such as Gemini-1.5-Pro and VideoLLMs that reason over thousands of frames, including VILAMP (Cheng et al., 2025a) and AdaReTake (Wang et al., 2025a). These findings underscore the potential significance of our framework, particularly under the limited context length constraints of MLLMs.

D.3 Results on more Benchmarks

Table 11: Results on EgoSchema and NExTQA benchmarks. Accuracy sign % is omitted for clarity.

Model	Frames	EgoSchema 3min	NExT-QA 0.7min
InternVideo (Wang et al., 2022)	90	32.1	49.1
LLoVi (Zhang et al., 2024a)	90	57.6	67.7
LangRepo (Kahatapitiya et al., 2024)	180	66.2	60.9
VideoAgent (Wang et al., 2024d)	8.4	60.2	71.3
LVNet (Park et al., 2024)	12	66.0	72.9
VidF4 (Liang et al., 2024)	8	-	74.1
VideoTree (Wang et al., 2025b)	63.2	66.2	73.5
InternVL2-8B (Chen et al., 2024c)		59.8	76.5
+ **KFC**	8	58.6	77.8
+ **Nar-KFC**		**64.0**	**78.1**
Qwen2-VL-7B (Wang et al., 2024b)		60.8	76.3
+ **KFC**	8	63.2	76.6
+ **Nar-KFC**		**65.8**	**77.6**

We further report performance of our KFC and Nar-KFC on two relatively shorter video benchmarks, i.e., EgoSchema (Subset) (Mangalam et al., 2023) and NExTQA (Xiao et al., 2021), in Tab. 11.

Unlike the long video datasets discussed in the main paper, our keyframe selection strategy (i.e., KFC) may underperfom compared to uniform sampling when applied to shorter videos. For example,

InternVL2-8B yields 58.6% accuracy on EgoSchema when using KFC. This performance drop is primarily due to KFC disrupting the temporal consistency of frame sequences, which is particularly important for short video understanding. Nevertheless, supplementing with non-keyframe narratives (Nar-KFC) leads to consistent performance improvements even on these shorter benchmarks. The gains are especially evident on EgoSchema, while the improvement on NExTQA is more limited, likely due to its relatively short average video length of approximately 44 sec.

Table 12: Results on TempCompasss and Video-Holmes benchmarks. Accuracy sign % is omitted for clarity.

Model	TempCompass		Video-Holmes
	Caption Matching	Overall	
Qwen2.5-VL-7B (Bai et al., 2025)	74.0	72.2	20.4
+ KFC	**74.1**	**72.2**	20.7
+ Nar-KFC	-	-	**22.9**
InternVL3-8B (Zhu et al., 2025)	80.1	74.8	33.5
+ KFC	**80.2**	**74.9**	**34.5**
+ Nar-KFC	-	-	**34.5**
Qwen3-VL-8B (Team, 2025)	79.1	**74.4**	30.9
+ KFC	**80.0**	74.3	33.7
+ Nar-KFC	-	-	**33.8**

We also evaluate our methods on extremely short video understanding benchmarks (TempCompass (Liu et al., 2024e), 10s), where narratives are not required, as well as on the more complex video reasoning benchmark, Video-Holmes (Cheng et al., 2025b), as illustrated in Tab. 12. For the relatively short TempCompass benchmark, selecting query-relevant and diversified keyframes results in considerable overlap with uniform sampling, leading to limited performance improvement. In contrast, on the more challenging Video-Holmes benchmark, our approach of carefully selecting keyframes and incorporating threaded narratives significantly enhances the MLLM's video reasoning capabilities.

D.4 Detailed Analysis on MLVU Categories

In Fig.7, we provide a detailed comparison of performance across specific categories in the MLVU benchmark as a supplement to the main paper Tab. 1. Compared to uniform sampling, the overall performance improvement introduced by KFC across all four models is primarily attributed to its superior accuracy in the **needle** and **count** categories. The *needle* task involves questions based on rare or unusual frames sourced from external videos, which are more likely to be captured by our query-relevance-based sampling strategy. In contrast, such frames are often missed by uniform sampling. A similar challenge arises in the *count* task, where correct answers rely on retrieving specific frames first in order to support accurate object/crowd/event counting.

On the other hand, our Nar-KFC approach generally achieves the best performance on **plotQA** and **topic** tasks. This advantage stems from its ability to preserve temporal continuity, which is often lacking in KFC-optimized keyframes that are temporally sparse and discontinuous. Such discontinuity hinders the model's ability to comprehend holistic video contents. For instance, KFC performs the worst on the *topic* task when inferenced with LLaVA-OneVision (c) and LLaVA-Video (d), even underperforming the uniform sampling baseline. In contrast, Nar-KFC addresses this issue through a narrative threading strategy, which maintains continuity by supplementing keyframes with coherent non-keyframe descriptions. This strategy significantly enhances the model's understanding of overall video plots and topics.

D.5 Computational Overhead

We analyze and present the detailed computational complexity (efficiency), including TFLOPs, latency, and memory usage, in Tab. 13. Note that searching the entire space of IQP would require approximately 10^{13} TFLOPs, making it impractical in real-world scenarios. Therefore, we report the computational complexity based on using 30k nodes in the IQP algorithm. Here, "search efficiency" refers to the keyframe search stage, while "overall efficiency" primarily pertains to the MLLM reasoning stage.

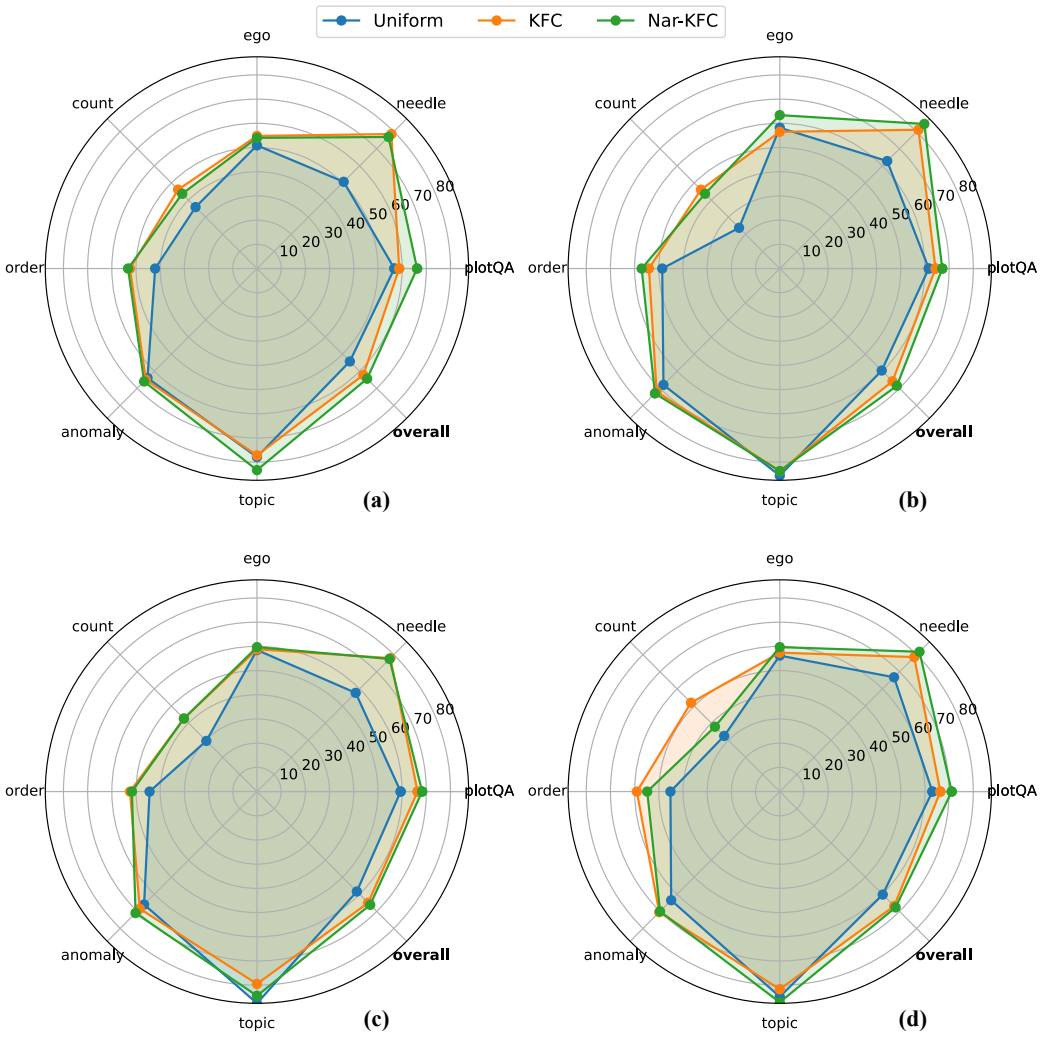

Figure 7: Performance comparison across specific categories of the MLVU benchmark. Results are shown for (a) InternVL2-8B, (b) Qwen2-VL-7B, (c) LLaVA-OneVision-7B, and (d) LLaVA-Video-7B, evaluated using three keyframe selection strategies: Uniform, KFC, and Nar-KFC.

Table 13: Computational efficiency comparison, including TFLOPs, latency, and memory usage for both the searching and overall inference stages. Results are reported using 8 frames and 210 narratives with the InternVL2-8B model.

Method	Search Efficiency			Overall Efficiency		
	TFLOPs↓	Latency (s) ↓	Memory (GB)	TFLOPs↓	Latency (s) ↓	Memory (GB)
Uniform-8	N/A	0.20	N/A	146.3	1.03	21.8
Top-k	N/A	0.24	N/A	146.3	1.07	21.8
KFC (IQP-30k)	6.9	7.18	N/A	153.2	8.01	21.8
KFC (GS)	∼0	0.48	N/A	146.3	1.31	21.8
Nar-KFC	∼0	0.60	N/A	202.6	2.13	32.2

Since we use an offline CLIP model to extract video embeddings (including query embeddings) and a Qwen2-VL-2B captioning model to generate video narratives, we also report their computational complexity in Tab. 14. The results are evaluated on an average 17-minute video (1,020 frames at 1 fps). It is important to note that these extraction processes are performed offline prior to online reasoning, which is the same as all previous keyframe selection strategies. Therefore, although the

preprocessing step is time-consuming, it impacts all keyframe selection methods equally, but does not impact the final inference complexity.

For an on-demand (long) video understanding system and suppose we are given an on-demand video, our lightweight captioner only needs to extract less than 210 narratives no matter how long the video is (since we have proved in our paper that more narrations won't bring further improvements and may exceed the context length of MLLMs). The caption extraction process requires less than 74.2 sec of latency. In practice, there are often no more than 210 frames between the first and last sampled keyframes, which can further reduce preprocessing time. The low computational cost of captioning is primarily due to our lightweight captioner, as we demonstrate that Nar-KFC's performance is not sensitive to captioner size and only a small number of frames are processed. If a latency of 74.2 sec (or less) remains a concern for on-demand video systems, our keyframe selection method, KFC-GS, can be used without the captioning stage for faster inference compared with prior frame selection methods. Overall, our approach achieves a favorable balance between accuracy and efficiency.

Table 14: Computational overhead for CLIP embedding extraction and frame captioning. Results are reported on an average 17 min video at 1 fps (1020 frames) frame sampling.

Model	Frames	TFLOPs↓	Latency (s)↓	Memory (GB)
Offline Frame Embedding & Caption Extraction				
CLIP-ViT-L-336px	1020	420.8	25.8	1.6
Qwen2-VL-2B	1020	4462.5	360.5	7.2
On-demand Video System Processing				
Qwen2-VL-2B	$\leq$210	$\leq$ 918.8	$\leq$74.2	7.2

E Additional Ablation Results

E.1 A Symmetrical Formulation of Original Objective and Analysis.

Objective Revisiting. In the main paper Sec. 3, we formulate the keyframe selection task as a *graph* problem and model it using integer quadratic programming (IQP) (3). However, the constructed score matrix (1) is asymmetric, as it only accounts for the query relevance of the i-th frame and the diversity between the i-th and j-th frames, while neglecting the query relevance of the j-th frame. This asymmetry introduces a minor discrepancy compared to the standard subgraph selection procedure. We illustrate this discrepancy with an example.

Example. Suppose we aim to retrieve 3 keyframes from 5 frames, and the optimal selection is given by $\mathbf{x} = [1, 1, 1, 0, 0]^T$, indicating that first three frames are selected. The score matrix $\mathbf{S}$ is defined as:

$$\mathbf{S}_{i,j} = \mathcal{S}(i,j) = S_{\text{QR}}(i) + S_{\text{FD}}(i,j) = \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & 0 & a_{23} & a_{24} & a_{25} \\ 0 & 0 & 0 & a_{34} & a_{35} \\ 0 & 0 & 0 & 0 & a_{45} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{5}$$

where $a_{i,j}$ denotes the score term for $i < j$ (i.e., only the upper trangular part of $\mathbf{S}$ is considered). According to (3), the maximum sum score (the total edge weight of the subgraph) should be:

$$\begin{aligned} \mathbf{x}^T\mathbf{S}\mathbf{x} &= [1,1,1,0,0] \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & 0 & a_{23} & a_{24} & a_{25} \\ 0 & 0 & 0 & a_{34} & a_{35} \\ 0 & 0 & 0 & 0 & a_{45} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} [1,1,1,0,0]^T \\ &= [1,1,1,0,0][a_{12}+a_{13}, a_{23}, 0, 0, 0]^T \\ &= a_{12} + a_{13} + a_{23} \\ &= S_{\text{QR}}(1) + S_{\text{FD}}(1,2) + S_{\text{QR}}(1) + S_{\text{FD}}(1,3) + S_{\text{QR}}(2) + S_{\text{FD}}(2,3). \end{aligned} \tag{6}$$

From this computation, we know that the query relevance of the first frame is counted twice, while that of the last selected frame (3^{rd}) is not counted at all, as there are no subsequent frames after it. This *discrepancy* shows the deviation from the standard graph-based subgraph selection formulation.

Symmetric Score Matrix. To mitigate this discrepancy and align the keyframe selection process with a standard graph problem, we reconstruct the original score matrix $\mathbf{S}$ to be symmetric by incorporating the query relevance of the j-th frame, defined as:

$$\mathbf{S}_{i,j} = S(i,j) = S_{\text{QR}}(i) + 2S_{\text{FD}}(i,j) + S_{\text{QR}}(j). \tag{7}$$

Experimental Results and Analysis. Compared with the symmetric $\mathbf{S}$ in (7), our original asymmetric matrix involves fewer terms with reducing size (only the upper triangular part is calculated), which leads to faster inference. Tab. 15 presents additional experimental results for replacing the original score matrix $\mathbf{S}$ with its symmetric counterpart. Modifying $\mathbf{S}$ to be symmetric – thus aligning the formulation with a standard graph problem – results in a 1% performance drop when using the IQP solver. This result supports the benefit of assigning higher weights to the initially selected frame at the beginning. Since the first keyframe is heuristically selected based on query relevance, this modification has negligible impact when using the GS strategy. We thus adopt the asymmetric score matrix defined in (1) for the remainder of our process.

Table 15: Impact of whether replacing score matrix to its symmetric counterpart. Results are reported on the Video-MME (sub.) benchmark using InternVL2-8B model. The search node number is 40k for solving IQP.

Setting	Strategy	Video-MME (sub.)			
		Short	Medium	Long	Overall
asymmetric $\mathbf{S}$ (1)	IQP	65.9	52.9	46.4	**55.1**
symmetric $\mathbf{S}$ (7)	IQP	66.1	50.1	46.1	54.1
asymmetric $\mathbf{S}$ (1)	GS	65.4	52.3	47.3	55.0
symmetric $\mathbf{S}$ (7)	GS	65.7	52.6	47.2	**55.1**

E.2 Integer Quadratic Programming (IQP) *vs.* Greedy Search (GS)

Table 16: Impact of expanding the IQP search space on performance and efficiency. Results are reported on the Video-MME (sub.) benchmark using InternVL2-8B model, with average computational time per video (in seconds) evaluated on a single NVIDIA A100 GPU.

Setting	Nodes#	Video-MME (sub.)				Time (s)
		Short	Medium	Long	Overall	
Uniform	-	63.9	48.7	44.9	52.5	1.03
GS	-	65.4	52.3	47.3	55.0	1.31
IQP	5k	64.2	52.6	46.7	54.5	3.91
	10k	64.4	52.6	45.8	55.0	4.81
	20k	64.3	52.3	**47.9**	54.9	6.23
	30k	65.6	52.6	46.2	54.8	8.01
	40k	**65.9**	**52.9**	46.4	**55.1**	9.26
IQP (GS init)	5k	64.3	52.0	48.0	54.7	5.22
	10k	65.1	51.9	47.3	54.5	6.12
	20k	65.1	52.3	47.5	54.9	7.54
	30k	65.3	52.3	45.7	54.4	9.32
	40k	65.8	51.4	46.0	54.4	10.57

We implement the Integer Quadratic Programming (IQP) algorithm using CPLEX and set a maximum number of search nodes to obtain the optimal set of keyframe indices within a limited time. The corresponding IQP results are reported in Tab. 16. As the search space increases from 5k to 40k nodes, performance on short videos gradually improves from 64.2% to 65.9%, which validates the effectiveness of modeling keyframe selection as an IQP problem. However, this improvement does not hold for long videos, where performance becomes unstable as the search space expands. We speculate that this is because even 40k nodes are still insufficient to cover the full solution space for long videos. For instance, in a 15-minute video (900 frames at 1 fps), selecting 8 keyframes results

in approximately $C(900, 8) \simeq 2.5 \times 10^{18}$, i.e., roughly 2.5 quintillion possible combinations. This vast search space far exceeds what can be practically explored with a node limit of 40k, let alone for videos that span several hours.

We also attempt to initialize the IQP search with greedy searched results, which are highlighted in gray in Tab. 16, in hopes of better guiding the IQP solving process. Experimental results indicate that this initialization strategy does not lead to further improvements in IQP performance, likely due to the search space remaining too large to be effectively navigated. Therefore, we adopt a customized greedy search (GS) strategy as a practical and robust alternative to the IQP algorithm.

E.3 Ablations on Hyperparameters in KFC (GS)

Table 17: Impact of low-rank truncation r in our Greedy Seach (GS) algorithm.

LowRank truncation r	Video-MME (sub.)			
	Short	Meidum	Long	Overall
$N/16$	64.8	51.3	46.3	54.2
$N/8$	65.3	51.7	46.0	54.3
$N/4$	**65.4**	**52.3**	47.3	**55.0**
$N/2$	65.2	51.7	**47.6**	54.8
N (w/o SVD)	64.1	51.8	45.7	53.9

The low-rank truncation parameter r in SVD (Sec. 3.1.2) serves to compress and denoise neighboring frames in the score matrix $\mathcal{S}$. Setting r equal to the number of video frames N is equivalent to not applying the SVD technique. Our experiments in Tab. 17 demonstrate that incorporating this decomposition step facilitates frame selection and reduces the problem size. Setting $r = \frac{N}{4}$ yields the best performance, where N refers to the total number of frames in a video. Choosing a smaller value, such as $\frac{N}{16}$ or $\frac{N}{8}$, leads to excessive information loss and consequently degrades the performance.

Table 18: Impact of downsample resolution in our Greedy Seach (GS) algorithm.

Downsample Resolution	Video-MME (sub.)			
	Short	Meidum	Long	Overall
64	63.8	52.3	44.0	53.4
128	**65.4**	**52.3**	47.3	**55.0**
256	64.8	50.2	47.6	54.2
512	63.0	51.4	**47.7**	53.8

Following previous works such as Frame-Voyager (Yu et al., 2025) and MDP3 (Sun et al., 2025), we default to downsampling the frame sequence to 128 frames. Our experiments, as shown in Tab. 18, also indicate that this downsampling resolution generally yields the best performance. Similar to SVD, the downsampling operation is designed to balance the trade-off between denoising the score matrix and minimizing the information loss.

Table 19: Impact of refinement window size k in our Greedy Seach (GS) algorithm.

Window Size k	Video-MME (sub.)			
	Short	Meidum	Long	Overall
0 (w/o refine)	65.0	51.2	**47.8**	54.7
1	65.1	51.4	47.4	54.7
2	**65.4**	52.3	47.3	**55.0**
4	64.6	**53.1**	45.7	54.4
8	64.2	50.3	43.8	52.8

We analyze the impact of the neighbor window size k in the final Greedy Search (GS) refinement step. As shown in Tab. 19, setting $k = 0$ corresponds to using the GS strategy without any refinement.

Table 20: Impact of incorporating full video-level narratives. These narratives include segments that appear before the first keyframe and after the last keyframe. * indicates that only narratives *between keyframes* are utilized in Nar-KFC.

Setting	Video-MME (no sub. / sub.)			
	Short	Meidum	Long	Overall
Full-Narrative	66.3 / 66.9	56.3 / 58.0	46.7 / 47.3	**56.4** / 57.4
Nar-KFC*	67.2 / 67.7	54.7 / 57.9	47.1 / 48.9	56.3 / **58.1**

When $k = 2$, which means examining a total of four frames, two before and two after the selected keyframe, the model achieves the best overall performance. This highlights the effectiveness of the refinement step as a robust strategy to complement prior SVD and downsampling operations. However, increasing the window size further (e.g., $k = 4$ or $k = 8$) results in performance degradation. This is likely due to the disruption of holistic keyframe combinations constructed by the greedy search, as excessive frame examination may introduce noise or redundancy.

Conclusion. The core of KFC-GS is a greedy algorithm, which iteratively selects the next frame with the highest cumulative score. Although we incorporate some pre-processing (SVD, downsampling) and post-processing steps (refinement) to further enhance performance, the vanilla greedy selection (GS) is already highly effective with initialization, eg., achieving 53.3 on Video-MME and 61.0 on MLVU. These results demonstrate that KFC-GS is generally robust and capable of generalizing well across different benchmarks. In fact, we do not manually tune the hyperparameters (r, d) involved in the pre-processing techniques, as they can be empirically set within an appropriate range. Comprehensive ablations on these hyperparameters (Tab. 17, Tab. 18, Tab. 19)further demonstrate that our results are not particularly sensitive to these parameters. Re-tuning is generally unnecessary, as our approach consistently achieves improvements over multiple benchmarks with 4 different MLLMs.

E.4 Implementation Details of Frame Extraction Baselines in Tab. 5

For CLIP[1] (Radford et al., 2021), SigLIP[2] (Zhai et al., 2023), and BLIP-2[3] (Li et al., 2023a), we directly rank and select the top-K candidate keyframes based on their frame-query cosine similarity logits. For TempGQA (Xiao et al., 2024), we follow the official code[4] to first select a segment based on the question, and then uniformly sample frames from the selected segment to generate the answer. For SeViLA (Yu et al., 2023), we use its trained localizer[5] to select the K keyframes as input, while maintaining the original hyperparameter settings. As for DPP (Determinantal Point Process) selection (Sun et al., 2025), since the official code is unavailable, we reimplement the DPP algorithm by defining its kernel matrix as $\mathcal{S}(i, q)\mathcal{S}(j, q)[1 - \mathcal{S}(i, j)]$, where the first two terms represent the similarity of frames i, j to the query q, and the last term encourages frame diversity between frame i and frame j. $\mathcal{S}$ denotes the cosine similarity operation. For AKS (Tang et al., 2025), we select keryframes based on the frame scores provided in the official repository[6].

E.5 Incorporating Full Video-level Narratives.

Our default Nar-KFC configuration (see main paper Sec. 3.2) only uses narratives that appear between the first and the last keyframe, discarding those that occur at the beginning or end of the video. Here, we analyze the effect of incorporating full video-level narratives, as shown in Tab. 20, while keeping the total number of inserted narratives fixed at 210. The results suggest that including these additional narratives has minimal impact on overall video understanding. This finding further supports our primary conclusion: keyframes play a dominant role in long-form VideoQA, while narratives mainly serve as auxiliary context.

[1] https://huggingface.co/openai/clip-vit-large-patch14-336
[2] https://huggingface.co/google/siglip-so400m-patch14-384
[3] https://huggingface.co/Salesforce/blip2-opt-2.7b
[4] https://github.com/doc-doc/NExT-GQA/tree/main/code/TempGQA
[5] https://github.com/Yui010206/SeViLA?tab=readme-ov-file
[6] https://github.com/ncTimTang/AKS

F ADDITIONAL QUALITATIVE EXAMPLES

We present additional qualitative examples of our keyframe selection method (KFC) in Fig. 8, and of the narrating keyframe method (Nar-KFC) in Fig. 9. Note that the frames leading to incorrect predictions in Fig. 9 can be regarded as failure cases of KFC.

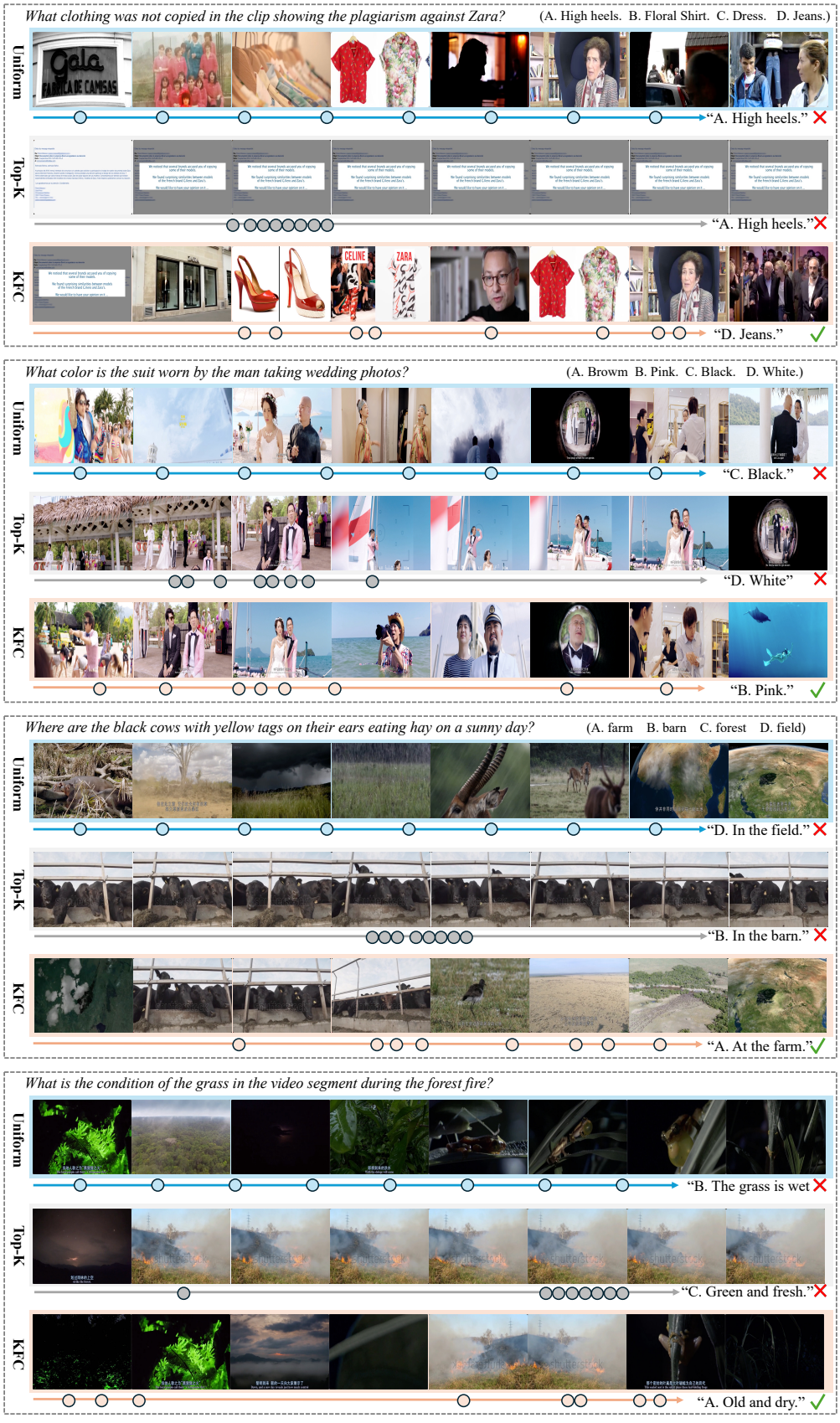

Figure 8: More qualitative examples of keyframe selection using our KFC method, compared with uniform sampling and topK sampling baselines. Zoom in for better visual details.

Figure 9: More qualitative examples of our threading keyframe methods Nar-KFC. Zoom in for details.