

Moment Sampling in Video LLMs for Long-Form Video QA

Mustafa Chasmai^{1*} Gauri Jagatap² Gouthaman KV²
Grant Van Horn¹ Subhransu Maji¹ Andrea Fanelli²

¹University of Massachusetts Amherst ²Dolby Laboratories

{mchasmai, gvanhorn, smaji}@umass.edu {Gauri.Jagatap, Gouthaman.KV, Andrea.Fanelli}@dolby.com

Abstract

Recent advancements in video large language models (Video LLMs) have significantly advanced the field of video question answering (VideoQA). While existing methods perform well on short videos, they often struggle with long-range reasoning in longer videos. To scale Video LLMs for longer video content, frame sub-sampling (selecting frames at regular intervals) is commonly used. However, this approach is suboptimal, often leading to the loss of crucial frames or the inclusion of redundant information from multiple similar frames. Missing key frames impairs the model’s ability to answer questions accurately, while redundant frames lead the model to focus on irrelevant video segments and increase computational resource consumption. In this paper, we investigate the use of a general-purpose text-to-video moment retrieval model to guide the frame sampling process. We propose “moment sampling”, a novel, model-agnostic approach that enables the model to select the most relevant frames according to the context of the question. Specifically, we employ a lightweight moment retrieval model to prioritize frame selection. By focusing on the frames most pertinent to the given question, our method enhances long-form VideoQA performance in Video LLMs. Through extensive experiments on four long-form VideoQA datasets, using four state-of-the-art Video LLMs, we demonstrate the effectiveness of the proposed approach.

1. Introduction

Video question answering (VideoQA) is a challenging and impactful task with numerous real-world applications, including egocentric assistants for wearables, conversational agents that interpret video content, intelligent tools for long-form video editing, educational aids in lecture videos, and efficient browsing of surveillance footage, etc. These scenarios often involve complex, untrimmed videos that require multimodal reasoning across visual, textual, and

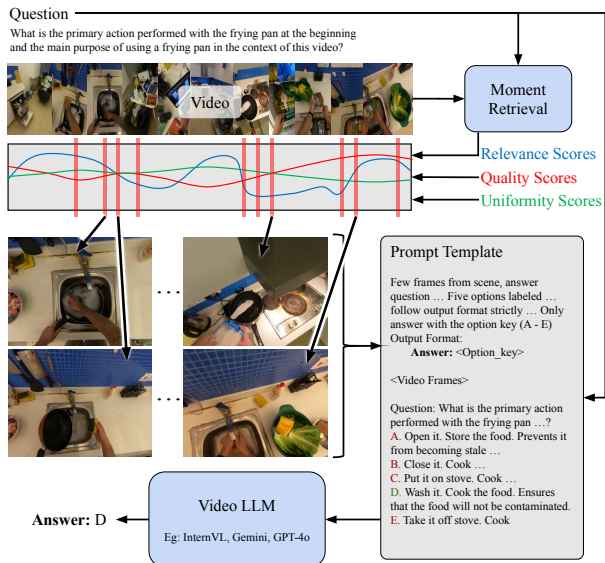


Figure 1. **Moment Sampling for VideoQA.** Given a video and a question (top), we first retrieve moments in the video that are relevant to answer the question. Retrieved moments, along with quality and uniformity scores are used to sample a few frames. These are given as input to a Video LLM to obtain the answer.

sometimes auditory information, within a shared context. To answer questions accurately, a model must identify salient objects and events, understand relationships between them, reason over time, and in many cases, apply common-sense or domain-specific knowledge, such as interpreting character interactions in a film or understanding mechanical concepts in an instructional video, etc.

Recent advancements in multimodal foundation models, particularly Video Large Language Models (VideoLLMs), have significantly elevated the capabilities of VideoQA systems. These models are designed to integrate visual [4, 5, 30, 34], auditory [25, 58], and textual [1, 12, 21, 47, 48] modalities into a unified reasoning framework. While current VideoLLMs have demonstrated strong performance on short video clips, typically ranging from a few seconds to a few minutes [10, 11, 20, 33, 51], scaling them to long-

*Work done during internship at Dolby Laboratories

form VideoQA introduces new challenges. These include maintaining temporal coherence across extended durations and managing computational constraints imposed by the length and complexity of input videos. To reduce the input size, existing approaches often resort to uniform sampling, where a fixed number of frames or segments are evenly sampled from the video (e.g., 100 frames from a 10-minute clip) [10, 33, 45]. However, this strategy is often suboptimal: it risks overlooking semantically important content while including redundant or irrelevant frames, ultimately limiting the model’s ability to perform accurate reasoning and increasing computational overhead. Addressing these limitations is especially crucial for VideoLLMs, which are often sensitive to input quality due to token and context length constraints. As such, there is a pressing need for intelligent, question-aware frame sampling strategies that dynamically select frames based on their relevance to the given question. By aligning the input with the question’s semantic and temporal requirements, such methods can enhance both the efficiency and effectiveness of VideoLLMs in long-form VideoQA, paving the way for more scalable and capable systems in real-world video understanding.

Moment retrieval has emerged as another task in video understanding, focused on identifying and localizing the most relevant temporal segment within an untrimmed video that aligns with a given natural language query [28, 29, 35, 38, 39]. This task demands fine-grained temporal grounding, often across lengthy and complex video content. Recent advances in vision-language modeling have significantly enhanced the ability to align textual queries with specific video segments, enabling more accurate and context-aware moment retrieval. Given these strengths, moment retrieval models offer a compelling foundation for guiding question-aware frame sampling in long-form VideoQA.

In this paper, we systematically explore how recent advances in moment retrieval can be leveraged to improve long-form VideoQA in VideoLLMs. To this end, we propose a model-agnostic approach called *moment sampling*, which intelligently guides frame selection based on moment retrieval cues. Our method is compatible with any VideoLLM and does not require additional retraining or fine-tuning. An overview of the proposed framework is illustrated in Fig. 1. We evaluate the effectiveness of *moment sampling* across a diverse set of VideoLLMs, including proprietary models such as GPT-4o [40] and Gemini 1.5 Pro [46], as well as the open-source models such as VideoLLaVA [33] and InternVL2 [10]. Experiments are conducted on four public datasets: EgoSchema [36], CinePile [45], NextQA [54], and IntentQA [31].

To the best of our knowledge, this is the first work to leverage advancements in moment retrieval for improving long-form VideoQA within the context of VideoLLMs. Our experiments demonstrate that the proposed Moment Sam-

pling strategy significantly enhances frame sampling efficiency and consistently improves performance across all evaluated VideoLLMs and benchmark datasets.

2. Related Work

Long-form Video Understanding: Video understanding showed significant advancements in the literature encompasses a wide range of tasks, such as action recognition [3, 18, 24, 27, 53, 57, 60], video captioning and question-answering [8, 9, 13, 26, 55], summarization [2, 17, 61], and retrieval [16, 29, 39, 56]. These methods often works well in the case of short-clips spanning a few seconds to minutes. Understanding long-form videos presents unique challenges that remain relatively underexplored. Unlike short clips, long-form videos require sophisticated methods for extended temporal reasoning, modeling event sparsity, and efficiently retrieving relevant information across lengthy sequences. Existing methods for long-form video understanding can be broadly categorized into two approaches: hierarchical strategies like Video Re-Cap [19], which focus on summarizing and captioning long videos, and sequence-based models such as Mamba [14] and state-space architectures [15], designed to process and integrate extended temporal contexts [32, 42].

Long-form Video Question Answering: VideoQA has emerged as a crucial benchmark for evaluating fine-grained video understanding, particularly in long-form scenarios. Traditional supervised methods train multimodal models on video-question-answer triplets [6, 23, 41]. For example, FlippedVQA [23] not only trains models to predict answers but also to generate corresponding questions and reconstruct video content, requiring minimal fine-tuning of adapter layers. MC-ViT [6] introduces a transformer architecture that scales effectively for long-context video understanding, while LongViViT [41] employs enhanced contrastive objectives for optimizing long-form VideoQA training. However, these approaches often rely on domain-specific models, and domain shifts, such as those between movie and egocentric videos, pose significant challenges for generalization across datasets.

VideoLLMs for Long-form VideoQA: To address the limitations of supervised and domain-specific approaches, zero-shot and prompt-based methods utilizing VideoLLMs have gained significant attention. These models can be broadly classified into two categories: those that first generate textual narrations through captioning or summarization models, and those that directly process frame-level video input. The former generates dense captions or summaries, which are subsequently queried by a pre-trained LLM to answer questions. For instance, VideoAgent [50] uses an LLM as a central reasoning module, iteratively gathering relevant content to answer questions, while

VideoTree [52] improves efficiency by scoring frame relevance and constructing an adaptive tree-based textual representation. While these methods offer scalable inference and generalization, they rely heavily on the quality of captioning models, which often miss crucial visual cues necessary for precise answers.

In contrast, VideoLLMs that operate directly on frame-level video input offer a promising alternative. These models typically down-sample long videos by selecting frames at regular intervals or segmenting them into shorter clips of just a few seconds. Proprietary models like GPT-4o [40] and Gemini 1.5 Pro [46], and open-source models such as Video-LLaVA [33], InternVL2 [10], and Video-LLaMA [11], fall into this category. However, uniform frame sampling is inherently limited, as it may miss semantically important moments while including redundant or irrelevant content, which compromises the model’s accuracy and increases processing overhead. These challenges highlight the need for more intelligent and adaptive frame selection strategies. Recent efforts, such as LVNet [43], have made progress in this direction by developing advanced frame-selection methods. This paper follows this direction by proposing a “moment sampling” technique for frame selection in VideoLLMs, drawing inspiration from advancements in the moment retrieval literature. We benchmark our approach against the above-mentioned VideoQA paradigms in VideoLLMs: 1) caption-based methods that generate textual descriptions of the video followed by QA using LLMs, 2) direct frame-based methods that input sampled frames along with the text query. Our experiments demonstrate the superior effectiveness of the proposed approach.

Moment Retrieval: This video understanding task requires models to identify and localize segments of a video that are most relevant to a given text query. Early approaches tackled this alongside highlight generation [35]. A major shift occurred with Moment-DETR [29], which framed moment retrieval as a temporal object detection problem, treating relevant moments as temporal “objects” that match the query. Inspired by the DETR [7] framework for object detection, Moment-DETR introduced a query-guided detection architecture and also proposed the QVHighlights dataset to advance research in this area. Building on this foundation, several subsequent works have extended the DETR-based architecture for improved performance. QD-DETR [39] incorporates cross-attention between video and textual features to more effectively guide moment detection. BAM-DETR [28] focuses on predicting boundary-oriented segments instead of center-aligned ones, addressing the inherent ambiguity in moment centers. CG-DETR [38] further refines the approach by modeling fine-grained correlations between question words and video clips. In this paper, we extend the application of moment retrieval to long-form VideoQA within the con-

text of VideoLLMs. We leverage QD-DETR, one of the best-performing moment retrieval models, for all our experiments, integrating its outputs into a query-focused frame sampling strategy that enhances VideoLLM performance on long-form video question answering tasks.

3. Methodology

3.1. Moment Sampling

We propose a query-focused frame sampling strategy for VideoLLMs, termed “Moment sampling”, which leverages moment retrieval models as an alternative to the uniform sampling commonly used in prior work. Specifically, we employ the QD-DETR model [39], a state-of-the-art moment retrieval framework trained on the QVHighlights dataset [29]. Given a video and its corresponding question for the VideoQA task, QD-DETR predicts a set of temporal segments, referred to as “moments”, along with associated relevance scores that indicate how relevant each moment is to the given query. These moment-level predictions form the foundation of our frame sampling strategy.

We begin by extracting moments using QD-DETR and converting their relevance scores into frame-level relevance scores. Since the predicted moments have hard boundaries, the initial relevance scores resemble a step function across the timeline. To encourage temporal smoothness, we apply Gaussian smoothing over these scores. If a frame belongs to multiple overlapping moments, its relevance score is the cumulative sum of the corresponding relevance values. However, relying solely on raw moment predictions can introduce artifacts or lead to redundant frame selection. To overcome these limitations and enhance the informativeness and diversity of sampled frames, we introduce the following additional refinements:

Quality scores: Some frames may suffer from visual degradation due to artifacts like motion blur. To down-weight such frames, we compute the variance of the Laplacian for each frame as a blur detection metric. This score is calibrated using an appropriate exponent and incorporated as a weighted penalty into the final relevance score.

Uniformity scores: Since some detected moments may be short or clustered closely in time, we introduce a *uniformity score* to encourage temporal dispersion of sampled frames. This score is computed as the sum of squared differences between a candidate frame’s timestamp and those of already selected frames, prioritizing frames that are temporally distant.

Frame clustering: In videos with repetitive or static scenes, visual redundancy can persist across non-adjacent frames. To mitigate this, we extract frame-level visual features and cluster them into a predefined number of groups. During selection, we ensure that only one frame is sampled from each cluster to maximize visual diversity.

With the per-frame scores computed, we perform greedy sampling to select the final frame set. At each step, we select the frame with the highest combined score. After each frame is selected, the uniformity scores are updated to reflect its timestamp, while the relevance, quality, and clustering scores remain fixed. This process continues until the desired number of frames is sampled. The overall sampling pipeline is illustrated in Fig. 2.

In addition to improving the performance of VideoQA tasks, our sampling strategy provides an interpretable mechanism for temporally grounding predictions made by otherwise black-box VideoLLMs. By visualizing relevance scores and sampled frames, we offer a transparent explanation of the model’s reasoning process, potentially increasing user trust and model accountability.

3.2. VideoLLM Setup and Evaluation

We experiment with a diverse set of VideoLLMs, including both proprietary models, such as GPT-4o [40] and Gemini 1.5 Pro [46], and open-source alternatives like VideoLLaVA [33] and InternVL2 [10]. These models typically accept a sequence of video frames alongside a paired text prompt that contains the question to be answered.

We choose to focus on multiple-choice questions instead of open-ended ones for several compelling reasons. First, the multiple-choice format enables objective and consistent evaluation across models, minimizing the ambiguity and variability often associated with free-form responses. Second, it provides a clearer and more reliable signal of model accuracy, particularly in zero-shot settings where minor variations in phrasing can significantly affect responses. Lastly, this format simplifies downstream analysis and facilitates meaningful comparisons across different frame sampling strategies and model types.

To ensure fair and standardized evaluation, we adopt a consistent zero-shot prompting strategy. Each prompt begins with an instruction describing the task, answering a question based on the video content, followed by the question itself and a list of answer options (A to E). We explicitly instruct the model to return only the letter corresponding to the selected answer.

While most models generally comply with this output format, we observe occasional deviations where responses do not exactly match one of the provided options. In such cases, we apply a fallback mechanism: we compute the longest common subsequence (LCS) between the model’s answer and each option, selecting the one with the highest overlap.

4. Implementation Details

All moment retrieval experiments were conducted on a single NVIDIA A10 GPU with 24GB of RAM. The open-source VideoLLMs, VideoLLaVA[33] and InternVL2

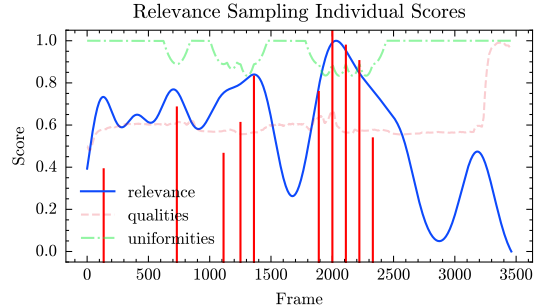


Figure 2. **An Example of Moment Sampling using various scores:** The blue, orange, and green curves represent the relevance, quality, and uniformity scores, respectively, while the red vertical lines indicate the frames selected for sampling. The height of each red line reflects the sampling order in the process. In this example, although one moment has a significantly higher relevance score, the sampling strategy, guided by quality and uniformity scores, ensures that frames are also selected from less relevant moments. This promotes temporal diversity and robustness in the sampled set.

(8B) [10], were also run under the same GPU setup. For the proprietary models, GPT-4o from OpenAI and Gemini 1.5 Pro from Google, we only had access via their respective APIs, and thus all related inference was performed on CPU. The frame relevance sampling process was likewise executed on CPU.

We experimented with different weight combinations for the scoring components (see Fig. 2), and we observed strong performance with quality scores weighted with 0.5 and uniformity scores as 2. For quality assessment, we estimate the blur level of each frame using the variance of its Laplacian. These scores are calibrated using an exponent and combined with the other scores. To reduce redundancy, we cluster frame-level CLIP [44] visual features into 30 clusters using K-means and enforce diversity by sampling at most one frame from each cluster. Relevance scores are computed as a sum of Gaussian functions centered at the predicted moment timestamps. Each Gaussian has a standard deviation equal to half the duration of the corresponding moment and is weighted by the moment’s relevance score. The final per-frame score is a weighted average of all normalized components (relevance, quality, and uniformity), and frames are selected greedily based on these scores.

5. Experiments and Results

5.1. Datasets

We conduct experiments on four publicly available long-form VideoQA datasets, each representing a distinct domain of video content. More details of the datasets are as follows:

EgoSchema [36] comprises egocentric videos of everyday tasks, recorded with wearable headsets as part of the

Table 1. **Video LLMs for Long-Form VideoQA:** For each Video LLM, we report performance with and without moment sampling. For each dataset, the best performance across all models is underlined. Metric: MCQ answering accuracy (Higher is better).

Video LLM	Moment Sampling (Ours)	EgoSchema subset	AVG	CinePile			NextQA			IntentQA full	
				CRD	NPA	TEMP	AVG	Temp	Cau		Des
InternVL2 (8B)	X	51.4	31.6	32.1	32.9	26.6	77.7	75.3	79.4	76.6	82.4
	✓	52.0	33.1	33.4	35.2	30.0	78.8	76.9	79.7	79.2	82.7
Gemini-1.5-pro	X	66.4	31.7	38.1	22.8	28.7	72.8	67.0	75.9	74.0	68.1
	✓	67.8	33.8	38.4	25.3	28.7	74.0	71.4	75.6	74.0	70.0
GPT-4o-mini	X	56.2	33.1	37.1	35.4	27.0	68.6	64.3	69.8	74.0	67.7
	✓	57.6	34.7	39.1	36.7	24.5	69.5	63.7	71.4	75.3	70.2
GPT-4o	X	72.0	35.2	40.7	30.4	<u>30.8</u>	78.1	73.6	81.4	75.3	75.8
	✓	73.6	35.5	41.1	32.9	<u>30.8</u>	77.4	72.5	80.4	76.6	77.3

Ego4D dataset. Each 3-minute video is paired with a single question, totaling 5K videos, 500 of which have publicly available answers. We conduct experiments on this answer-available subset. Videos are accompanied by manually annotated narrations, and the question-answer pairs are generated using LLMs based on these narrations.

CinePile [45] consists of third-person video clips sourced from movies. The test set includes approximately 200 videos, each with an average duration of 2 minutes and 40 seconds, accompanied by around 5,000 question-answer pairs. In addition to video content, the dataset also provides subtitles and visual descriptions authored by the movie creators. The question-answer pairs are generated using large language models (LLMs), with human oversight to ensure quality. Questions are categorized to evaluate distinct reasoning capabilities, including Character or Relationship Dynamics (CRD), Narrative or Plot Analysis (NPA), and Temporal Reasoning (TEMP). A subset of particularly challenging questions is further separated into a “hard” split. In our experiments, we focus exclusively on the hard split, and do not provide models with access to subtitles.

NextQA [54] consists of 5,440 videos averaging 44 seconds in length, with around 52K questions split across training, validation, and test sets. As we focus on zero-shot VideoQA, we use only the validation set, which includes 570 videos. While the dataset features both open-ended and multiple-choice questions, our experiments are conducted with the multiple-choice format. The questions are categorized into three types, Temporal (Tem.), Causal (Cau.), and Descriptive (Des.), each designed to evaluate different reasoning abilities of VideoQA models.

IntentQA [31] contains around 4K videos and 16K multiple-choice questions. These questions focus on reasoning about video intent. They use NextQA [54] as the source dataset, identify questions based on intent and include some additional manual annotations. The dataset is

designed so that similar actions can imply different intents depending on the context. For our experiments, we use the test set, which includes 567 videos.

5.2. Quantitative Results

Table 1 presents the quantitative comparison of various VideoLLMs with and without the proposed moment sampling strategy. Across all models and datasets, we observe a consistent performance improvement when using moment sampling over traditional uniform frame sampling. These results highlight the effectiveness of the query-focused approach in identifying semantically relevant frames. Since all evaluated datasets involve long-form videos, where crucial information is sparse and unevenly distributed, moment sampling significantly enhances performance by prioritizing frames most aligned with the question context, thereby improving both accuracy and efficiency.

Analyzing model performance across datasets, we can see that GPT-4o consistently outperforms others on EgoSchema and CinePile, while InternVL2 shows stronger results on NextQA and IntentQA. One possible explanation lies in the nature of the question-answer annotations: NextQA and IntentQA are manually curated by human annotators, often requiring fine-grained reasoning grounded in commonsense and real-world understanding, where InternVL2 may excel due to its training data or architecture. In contrast, EgoSchema and CinePile include question-answer pairs that are either fully or partially generated by LLMs, potentially aligning better with the style and structure of GPT-4o’s own pretraining or decoding behavior. Additionally, domain differences, such as egocentric vs. third-person perspectives and cinematic vs. daily activity, might contribute to the variation in performance across models.

Another notable observation from Table 1 is that InternVL2 consistently improves across all datasets, whereas other models exhibit occasional minor drops in performance. A likely explanation is InternVL2’s greater reliance

Table 2. **Comparison with captioning-based models:** Contrasting performance of traditional Long VideoQA pipelines (1) models that do video captioning followed by text based QA, (2) VideoLLM models that use the raw video and text question as input to produce answer, (3) frame-based Moment Sampling (MS) followed by VideoLLM based answering.

Method	LLM Backbone	EgoSchema subset	CinePile				NextQA			IntentQA full	
			AVG	CRD	NPA	TEMP	AVG	Temp	Cau		Des
Answering from Captions											
LLoVi [59]		57.6	-	-	-	-	67.7	61.0	69.5	75.6	64.0
VideoAgent [50]		60.2	-	-	-	-	71.3	64.5	72.7	81.1	-
Narration + LLM	Llama3	53.0	28.8	31.5	29.1	27.2	-	-	-	-	-
VideoTree [52]	GPT-4	66.2	-	-	-	-	73.5	67.0	75.2	81.3	66.9
LVNet [43]	GPT-4o	66.0	-	-	-	-	72.9	65.5	75.0	81.5	71.1
Answering from Video											
Flipped-VQA [23]		44.7	32.5	36.2	35.6	23.8	72.0	69.2	72.7	75.8	-
MC-ViT-L [41]		56.8	-	-	-	-	65.0	-	-	-	-
LongViViT [41]		62.6	-	-	-	-	-	-	-	-	-
VideoLLaVA [33]	LLaVA (7B)	20.6	19.3	18.9	16.5	23.2	20.5	17.6	20.6	27.3	23.3
VideoLLaMA2 [11]	LLaMA2	42.2	-	-	-	-	-	-	-	-	-
Tarsier [49]	Tarsier	68.6	-	-	-	-	-	-	-	-	79.2
LangRepo [22]	Mixtral (12B)	66.2	-	-	-	-	60.9	51.4	64.4	69.1	59.1
MoReVQA [37]	PAL-3 (5B)	51.7	-	-	-	-	69.2	64.6	70.2	-	-
GPT-4o with MS (Ours)	GPT-4o	73.6	35.5	41.1	32.9	30.8	77.4	72.5	80.4	76.6	77.3

on visual content. This indicates that models with stronger visual grounding tend to benefit more from query-focused frame selection strategies like moment sampling.

Comparison with captioning-based models: As discussed in Sec. 2, a common alternative for handling long-form videos involves first generating textual narrations using off-the-shelf captioning or summarization models, followed by querying an LLM to answer questions based solely on this generated text. While such captioning-based methods offer scalability and reduce video processing overhead, they suffer from a critical limitation: they depend entirely on the coverage and quality of the captions. In practice, these captions often miss subtle visual cues or context-specific details that are essential for accurate question answering, particularly in complex, long-form videos.

Table 2 compare the performance of the best-performing captioning-based models with our strongest model, GPT-4o equipped with moment sampling (GPT-4o + MS) from Table 1. The results clearly show that GPT-4o + MS consistently outperforms captioning-based approaches. This highlights the importance of preserving visual context in VideoLLMs and demonstrates that directly feeding semantically relevant frames, selected via the query-focused moment sampling strategy, enables more precise and context-aware reasoning. Moreover, this approach offers a more faithful representation of the video’s visual semantics, allowing the model to interpret and align information more effectively with the question. Especially in long-form videos, where key information is sparse and not evenly distributed, relying

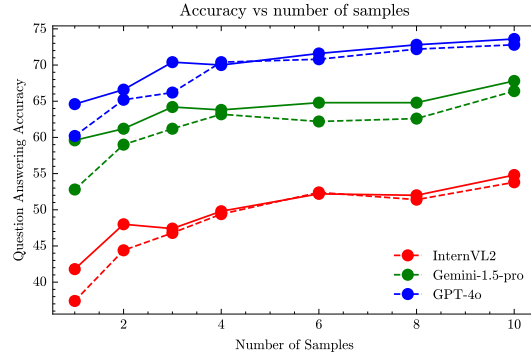


Figure 3. **Frame Sample Efficiency.** Accuracy for EgoSchema as the number of sampled frames is increased with tradition uniform sampling (dashed line) and moment sampling (solid line).

on visual inputs rather than compressed textual summaries proves significantly more effective for VideoQA tasks.

Frame sampling efficiency: Fig. 3 analyzes the performance of selected models as the number of sampled frames is varied. Notably, the performance gap between moment sampling and uniform sampling widens as the number of frames decreases. This demonstrates that our query-based moment sampling is significantly more sample-efficient, it can often match or even surpass the performance of uniformly sampled models while using fewer frames.

Visual Relience: VideoQA tasks are fundamentally designed to evaluate a model’s ability to reason over visual

content in conjunction with natural language. Thus, a high-quality VideoQA dataset should include questions that cannot be reliably answered without access to the video itself. To assess this property, we evaluate model performance with and without video input, as shown in Table 3.

In the no-video setting, we inform the model that the question pertains to a video that is not accessible and explicitly instruct it to attempt an answer based on its prior knowledge. Across EgoSchema, NextQA, and IntentQA, we observe substantial drops in accuracy, on the order of 25-30% points (30-50% relative), indicating that these datasets require significant visual understanding and are less susceptible to language-only biases. In contrast, CinePile exhibits a much smaller drop (less than 6.5% absolute, < 20% relative), suggesting that many questions may be partially answerable using linguistic cues alone, perhaps due to the influence of templated or narratively suggestive phrasing.

To better isolate visual reasoning in CinePile, we further analyze settings that include or exclude subtitles (Table 4). The inclusion of subtitles, which themselves often contain rich contextual information, further diminishes visual reliance. Therefore, we focus our experiments on the hard subset of CinePile without subtitles, where visual context plays a more critical role.

Moment sampling aims to enrich the quality and relevance of visual information presented to the model by prioritizing frames most aligned with the query. As such, the effectiveness of moment sampling is inherently tied to the degree to which a model depends on visual content. Models with strong visual grounding are more likely to benefit from improvements in frame selection. The performance differences observed across models in Table 1 can be partially attributed to the differences in visual reliance measured in Table 3. For example, InternVL2, which shows high sensitivity to visual input, also exhibits more consistent gains with moment sampling. On the other hand, improvements are more modest for models that may rely more heavily on language priors, such as GPT-4o.

Collectively, these findings highlight the dual role of visual reliance and frame relevance in effective VideoQA. Moment sampling not only boosts sample efficiency and accuracy but also serves as a valuable strategy for enhancing multimodal alignment in long-form video understanding.

5.3. Qualitative Results

Fig. 4 shows some qualitative results showcasing model predictions along with collages of frames selected via traditional uniform sampling and our proposed moment sampling strategy. These visualizations help illustrate how moment sampling improves frame relevance and, consequently, model performance. In the EgoSchema example, the question pertains to two key segments: the beginning of the video (washing a frying pan) and the end (using it for

Table 3. **Visual vs. Language Reliance for VideoQA.** For each dataset the performance using text alone (first row) and using vision (second row) are shown. CinePile dataset performs well even without any visual information and may not be sufficiently testing the multimodal capabilities of a VideoQA model.

Method	Vis	EgoSchema	CinePile	Next	Intent
InternVL2 (8B)	X	25.8	25.3	48.8	54.7
	✓	51.4	31.6	77.7	82.4
Gemini-1.5-pro	X	33.0	29.7	50.9	57.0
	✓	66.4	31.7	72.8	68.1
GPT-4o-mini	X	30.6	28.0	51.9	57.0
	✓	56.2	33.1	68.6	67.7
GPT-4o	X	42.2	31.1	52.8	59.3
	✓	72.0	37.2	78.1	75.8

Table 4. **Visual Reliance on CinePile.** Visual reliance for different dataset settings. Inclusion of subtitles results in smaller loss in performance in language-only mode. However, the visual modality is more relevant when subtitles are not included.

Method	Vision	Subtitles		No Subtitles	
		Full	Hard	Full	Hard
GPT-4o	X	58.8	42.1	42.0	31.1
	✓	59.7	43.7	50.4	37.2

cooking). Moment sampling effectively allocates roughly half of the selected frames to the initial activity and the remaining half to the final event, both critical to answering the question. In contrast, uniform sampling fails to capture this temporal distribution, missing important context. For CinePile, the answer requires detecting a specific visual cue, black liquid on a character’s forehead. This detail is captured in the first frame of the penultimate row in the query-based (moment sampling) collage but only partially and less clearly in the third frame of the uniform sampling grid. The presence or absence of this frame significantly impacts the model’s ability to answer correctly. In both NextQA and IntentQA, the questions are tied to specific incidents within the video. Uniform sampling often fails to include these moments, while moment sampling consistently captures them. For instance, in the NextQA collage, the first frame in the last row captures the key event; similarly, for IntentQA, the final frame in the penultimate row does. These contextually crucial frames are entirely missing in the corresponding uniform-sampled collages, likely leading to incorrect predictions.

These examples further highlight the core advantage of moment sampling, its ability to identify and prioritize semantically rich frames that are closely aligned with the question, particularly in long-form videos where key events may be temporally sparse and unevenly distributed.

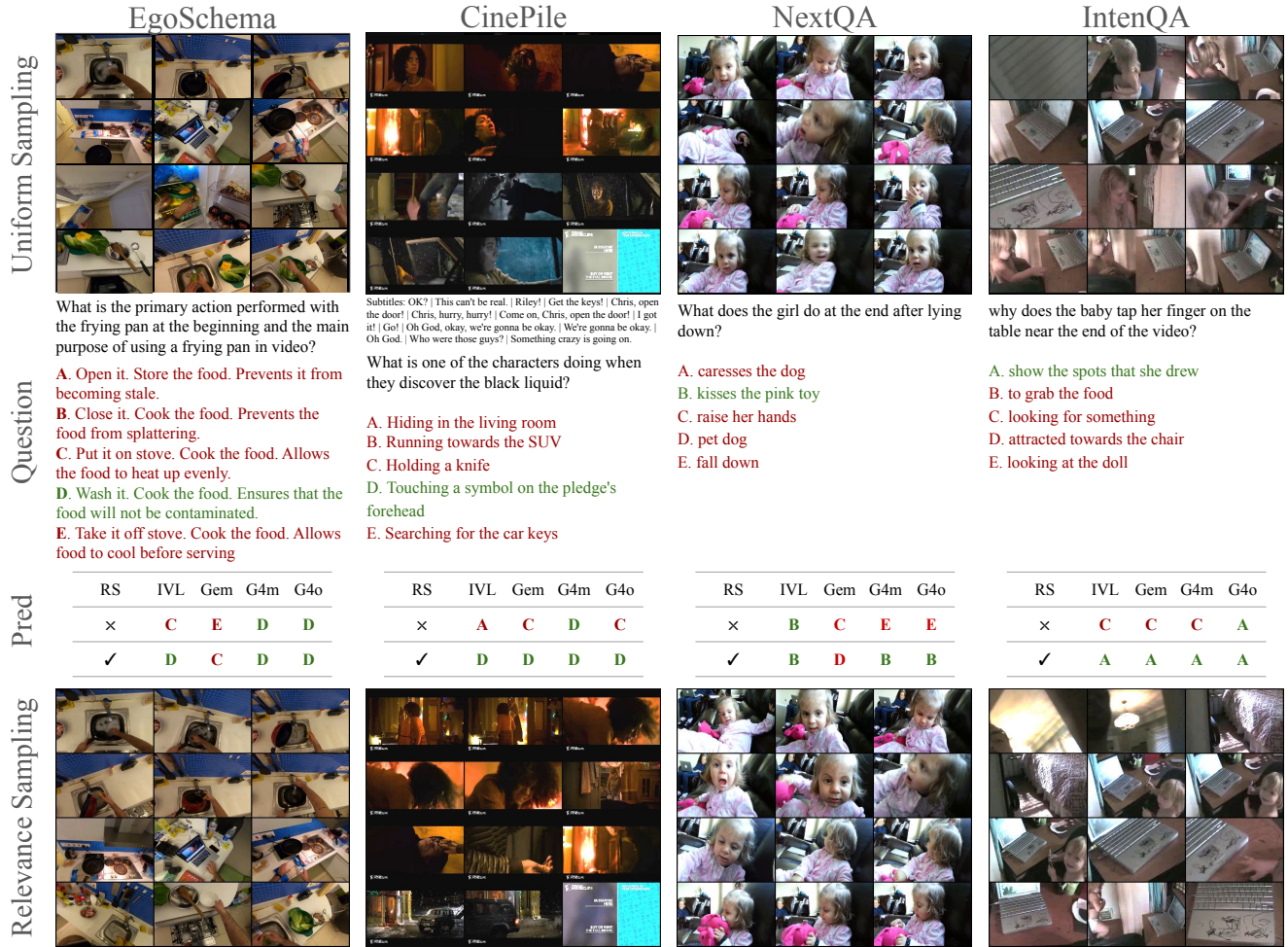


Figure 4. **Qualitative results** from the four datasets along with predictions of the VideoLLMs, with and without moment sampling. We also show collages made by uniform (top) and relevance (bottom) sampling for these questions. For each question, the correct option is in green color and incorrect options in red. Associated subtitles are shown for CinePile.

6. Conclusion and Future Work

In this work, we introduced moment sampling, a relevance-driven, query-aware frame selection strategy to address the limitations of uniform sampling in long-form VideoQA in the context of VideoLLMs. By leveraging a pre-trained moment retrieval model, our method identifies video segments that are most contextually aligned with the input question, enabling VideoLLMs to focus on informative content while avoiding redundancy. Through extensive evaluation across four long-form VideoQA datasets and four state-of-the-art VideoLLMs, we demonstrated consistent performance gains. These improvements are especially notable in datasets with higher visual reliance, highlighting the importance of intelligent frame selection in multimodal reasoning tasks. Additionally, moment sampling improves sampling efficiency and offers enhanced interpretability by explicitly grounding predictions in relevant visual evidence.

Looking ahead, our work opens up several promising directions. One key avenue is end-to-end integration, where the moment retrieval and VideoQA components are trained jointly, potentially improving alignment between relevance estimation and final predictions. We also plan to explore multimodal enhancements, such as incorporating audio signals or structured video descriptions to further enrich context. Another exciting direction is temporal query decomposition, breaking complex queries into sub-questions localized to specific time segments, thereby enabling more precise and interpretable reasoning. Overall, moment sampling provides a practical foundation for scalable, efficient, and explainable long-form VideoQA in the era of VideoLLMs.

Acknowledgments and Disclosure of Funding

The research is supported in part by grant #2329927 from the National Science Foundation (USA).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. **1**
- [2] Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. Combining global and local attention with positional encoding for video summarization. In *2021 IEEE international symposium on multimedia (ISM)*, pages 226–234. IEEE, 2021. **2**
- [3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. **2**
- [4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. **1**
- [5] 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*, 2023. **1**
- [6] Ivana Balazevic, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J Hénaff. Memory consolidation enables long-context video understanding. In *Forty-first International Conference on Machine Learning*, 2024. **2**
- [7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. **3**
- [8] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023. **2**
- [9] Sihan Chen, Xingjian He, Handong Li, Xiaojie Jin, Jiashi Feng, and Jing Liu. Cosa: Concatenated sample pretrained vision-language foundation model. In *The Twelfth International Conference on Learning Representations*, 2024. **2**
- [10] Zhe Chen, Jiannan Wu, Wenhui 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*, pages 24185–24198, 2024. **1, 2, 3, 4**
- [11] 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. **1, 3, 6**
- [12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. **1**
- [13] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618, 2020. **2**
- [14] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. **2**
- [15] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. **2**
- [16] Peiyan Guan, Renjing Pei, Bin Shao, Jianzhuang Liu, Weimian Li, Jiayi Gu, Hang Xu, Songcen Xu, Youliang Yan, and Edmund Y Lam. Pidro: Parallel isomeric attention with dynamic routing for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11164–11173, 2023. **2**
- [17] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VII 13*, pages 505–520. Springer, 2014. **2**
- [18] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3148–3159, 2022. **2**
- [19] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18198–18208, 2024. **2**
- [20] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. **1**
- [21] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. **1**
- [22] 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. **6**
- [23] Dohwan Ko, Ji Soo Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo J Kim. Large language models are temporal and causal reasoners for video question answering. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. **2, 6**
- [24] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong.

- Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16020–16030, 2021. 2
- [25] Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Forty-first International Conference on Machine Learning*, 2024. 1
- [26] Weicheng Kuo, AJ Piergiovanni, Dahun Kim, xiyang luo, Benjamin Caine, Wei Li, Abhijit Ogale, Luwei Zhou, Andrew M. Dai, Zhifeng Chen, Claire Cui, and Anelia Angelova. MaMMUT: A simple architecture for joint learning for multimodal tasks. *Transactions on Machine Learning Research*, 2023. 2
- [27] Dongho Lee, Jongseo Lee, and Jinwoo Choi. Cast: cross-attention in space and time for video action recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [28] Pilhyeon Lee and Hyeran Byun. Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. *arXiv preprint arXiv:2312.00083*, 2023. 2, 3
- [29] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 2, 3
- [30] 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*, pages 19730–19742. PMLR, 2023. 1
- [31] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Inten-tqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11963–11974, 2023. 2, 5
- [32] Kunchang Li, Xinhao Li, Yi Wang, Yanan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024. 2
- [33] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 2, 3, 4, 6
- [34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1
- [35] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051, 2022. 2, 3
- [36] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4
- [37] 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*, pages 13235–13245, 2024. 6
- [38] WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023. 2, 3
- [39] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023. 2, 3
- [40] OpenAI. Gpt-4o: A language model. <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed: 2024-09-07. 2, 3, 4
- [41] Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14386–14397, 2024. 2, 6
- [42] Jinyoung Park, Hee-Seon Kim, Kangwook Ko, Minbeom Kim, and Changick Kim. Videomamba: Spatio-temporal selective state space model. *arXiv preprint arXiv:2407.08476*, 2024. 2
- [43] 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. 3, 6
- [44] 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*, pages 8748–8763. PmLR, 2021. 4
- [45] Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024. 2, 5
- [46] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2, 3, 4
- [47] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 1
- [48] 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. 1
- [49] Jiawei Wang, Liping Yuan, and Yuchen Zhang. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 6
- [50] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*, 2024. 2, 6
- [51] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *CoRR*, 2024. 1
- [52] 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. *arXiv preprint arXiv:2405.19209*, 2024. 3, 6
- [53] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 2
- [54] 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*, pages 9777–9786, 2021. 2, 5
- [55] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, pages 38728–38748. PMLR, 2023. 2
- [56] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language alignment. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [57] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3333–3343, 2022. 2
- [58] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023. 1
- [59] 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*, pages 21715–21737, 2024. 6
- [60] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 2
- [61] Wencheng Zhu, Yucheng Han, Jiwen Lu, and Jie Zhou. Relational reasoning over spatial-temporal graphs for video summarization. *IEEE Transactions on Image Processing*, 31: 3017–3031, 2022. 2