BOLT: Boost Large Vision-Language Model Without Training for Long-form Video Understanding

Shuming Liu Chen Zhao* Tianqi Xu Bernard Ghanem
King Abdullah University of Science and Technology (KAUST)

Abstract

*Large video-language models (VLMs) have demonstrated promising progress in various video understanding tasks. However, their effectiveness in long-form video analysis is constrained by limited context windows. Traditional approaches, such as uniform frame sampling, often inevitably allocate resources to irrelevant content, diminishing their effectiveness in real-world scenarios. In this paper, we introduce **BOLT**, a method to **BO**ost **L**arge VLMs without additional **T**raining through a comprehensive study of frame selection strategies. First, to enable a more realistic evaluation of VLMs in long-form video understanding, we propose a multi-source retrieval evaluation setting. Our findings reveal that uniform sampling performs poorly in noisy contexts, underscoring the importance of selecting the right frames. Second, we explore several frame selection strategies based on query-frame similarity and analyze their effectiveness at inference time. Our results show that **inverse transform sampling** yields the most significant performance improvement, increasing accuracy on the Video-MME benchmark from 53.8% to 56.1% and MLVU benchmark from 58.9% to 63.4%. Our code is available at* `https://github.com/sming256/BOLT`.

1. Introduction

Recently, large Video-Language Models (VLMs) have emerged as powerful tools for video understanding [11, 17, 18, 20, 21, 27, 41]. These models leverage the synergy between visual and linguistic modalities through large language models, enabling them to capture complex relationships and contextual cues over time. They have demonstrated remarkable advancements in tasks such as video captioning [4, 33, 37] and video question answering [12, 21, 23]. However, large VLMs face significant challenges in processing long videos containing thousands of frames due to the sheer volume of video tokens. For instance, LLaVA-OneVision [13] utilizes 196 tokens per frame, meaning that without reduction, an hour-long video at 30 FPS would require 20 million tokens. Given the limited context length of the underlying language models and computational constraints [28, 39], **video token reduction** has become essential for enabling large VLMs to efficiently handle video understanding tasks.

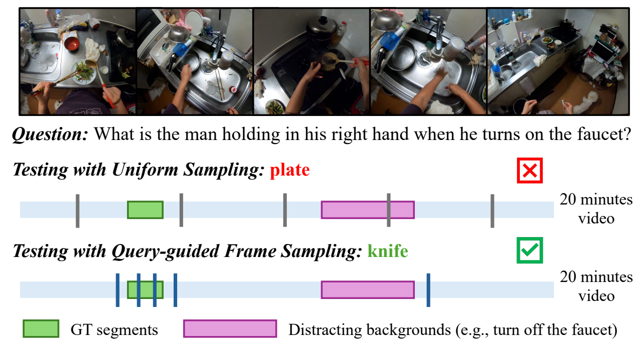

Figure 1. Given a long video, conventional VLMs adopt uniform sampling during inference, which may divert VLM's focus due to the distracting backgrounds. In contrast, our query-guided frame sampling can identify and prioritize question-relevant frames, boosting the VLM's performance without requiring fine-tuning.

To reduce video tokens, previous works have primarily explored strategies from two perspectives. The first is *spatial compression*, which reduces the number of tokens per frame, as employed by methods such as MovieChat [30] and SlowFast-LLaVA [36]. This paper focuses on the second perspective, *temporal selection*, which involves selecting a subset of frames from raw videos. Most prior works have relied on uniform frame sampling, in which a fixed number of frames are evenly selected from the video [1, 14, 38]. Although simple and efficient, this approach often fails in complex real-world scenarios because it assumes that all frames contribute equally to content understanding, which may hold true for short, query-trimmed videos but not for long, unedited ones. For example, instructional or surveillance videos usually contain lengthy segments irrelevant to the given query. Sampling frames from these segments not only consumes the token budget but may also distract the VLM from pertinent information, thereby impairing its performance [14, 29].

*Corresponding author.

Some recent methods, such as Goldfish [2], LongVU [28], and Frame-Voyager [39], have started to explore adaptive temporal selection strategies by using visual-text similarity to identify relevant frames. However, these methods require additional training on existing VLMs, demanding substantial computational resources and extensive fine-tuning data. For instance, Frame-Voyager [39] requires 32 H100 GPUs for training. This raises a crucial question: *Can we boost the performance of existing large VLMs without additional training by designing better temporal selection methods?* This paper provides an affirmative answer to this question through two key explorations: how *selecting the right frames* matters and how to *select the frames right*.

First, we investigate the importance of *selecting the right frames* for long-form video understanding using two newly designed evaluation setups. (1) We utilize the MultiHop-EgoQA dataset [5], which provides question-related segments, to examine the performance of large VLMs with and without access to ground-truth video segments. (2) To further study the impact of frame selection in real-world scenarios containing diverse background clips in longer videos, we propose a multi-source retrieval evaluation setting by extending the videos via internal retrieval in the existing dataset. The results show that selecting irrelevant frames, *i.e.*, query-unrelated ones, significantly degrades the performance on VQA tasks — a common issue encountered with uniform sampling.

Then, we explore how to *select the frames right* by introducing several query-aware frame selection approaches, including top-K selection, watershed grouping, and inverse transform sampling. All these approaches are applied only at test time, without any retraining of the VLMs. We find that inverse transform sampling consistently improves VLM performance, particularly in the multi-source retrieval evaluation scenario. Using this frame selection strategy, we can **BO**ost the performance of existing **L**arge VLMs without additional **T**raining. We thus name our method **BOLT**.

Our main contributions can be summarized as:

- We demonstrate that selecting the right frames significantly impacts large VLMs' test-time performance for long-form video understanding. We validate this through experiments on the MultiHop-EgoQA dataset [5], and by introducing multi-source retrieval evaluation setup, which better reflects real-world video understanding challenges.
- We explore various training-free frame selection strategies and propose leveraging inverse transform sampling to select query-relevant frames to improve VQA performance without requiring model retraining.
- Our method achieves consistent improvements on five benchmarks, *e.g.*, from 53.8% to 56.1% on the Video-MME [9], and from 58.9% to 63.4% on MLVU [46].

2. Related Works

2.1. Large Vision-Language Models for Video

Transformer-based large language models (LLMs) have revolutionized natural language processing [8, 24]. Researchers have extended LLMs to handle multiple modalities by incorporating multimodal inputs, typically visual content such as images and videos [13, 47] to make large vision-language models (VLMs). These models leverage large-scale pre-training on extensive data to learn rich representations that capture intricate relationships between visual and textual modalities. Building upon these, large VLMs have significantly advanced various video understanding tasks such as video captioning [4, 33, 37], video question answering [12, 21, 23], and temporal reasoning [25, 34]. Recent surveys [16, 31] comprehensively review the architectures, training strategies, and applications of VLMs in video understanding, highlighting both their capabilities and the challenges they address.

Despite these successes, large VLMs encounter significant challenges when processing videos due to the substantial amount of video frames, the limited context length of language models [32], and the "lost in the middle" problem [19]. These constraints necessitate efficient frame selection strategies to identify and retain essential information from videos without overwhelming the model. Thus, developing effective frame selection methods is critical to further enhance VLM performance in video understanding tasks.

2.2. Video Token Reduction in Large VLMs

Reducing video tokens is essential for making large VLMs computationally feasible. The most common and straightforward approach is uniform frame sampling, which selects frames without considering their individual importance to the task. Unfortunately, this method may omit crucial frames needed for understanding complex or long videos, thereby limiting reasoning performance. Several strategies have been proposed in the literature to make token reduction more adaptive and task-specific.

Token Merging. One approach is to merge information within or across frames to form a compact yet comprehensive representation. MovieChat [30] implements short-term and long-term memory mechanisms to integrate and compress frames from a long video, ensuring minimal information loss. LLaMA-VID [15] encodes each frame into separate context and content tokens, whereas LongVU [28] leverages cross-modal queries and inter-frame dependencies to reduce redundancy while preserving visual details. However, token merging approaches risk oversimplifying fine-grained information, potentially preventing the model from capturing subtle differences between frames.

Query-Based Frame Sampling. To overcome limitations associated with uniform sampling, recent similarity-based methods select frames based on their relevance to specific queries or tasks. Goldfish [2] addresses video QA by segmenting long videos into shorter clips and retrieving relevant clips using cosine similarity between query embeddings and captions. Other approaches learn optimal frame selection strategies through additional training. For instance, Frame-Voyager [39] learns informative frame combinations by training on a dataset that ranks frames based on prediction losses from a VLM [3]. Although potentially more effective, such learnable sampling methods require additional training and substantial computational resources, limiting their practicality.

Unlike the above-mentioned methods, in this work, we propose a training-free frame selection strategy to enhance the performance of existing large VLMs at inference time.

3. Revisiting Large VLMs

Consider a video V with T frames $\{f_1, f_2, \ldots, f_T\}$ and an associated query Q (system prompts omitted for brevity). The VQA task aims to generate an answer A that accurately addresses the query based on the content of V. The response format depends on the query type, encompassing both open-ended and multiple-choice answers.

State-of-the-art VLMs typically address this task using the following pipeline: First, due to computational constraints and the video's high frame rate, a subset of N frames $\{f'_1, f'_2, \ldots, f'_N\}$ is selected, where $N \ll T$. In common implementations, N is often set to 16 or 32 frames. Next, each selected frame f'_i is processed through a vision encoder E_v to obtain visual embeddings: $v_i = E_v(f'_i)$, where $v_i \in \mathbb{R}^{M\times C}$. Here, C denotes the embedding dimension, and M represents the number of tokens per frame.

Following visual encoding, a projection module maps visual embeddings v_i into the language embedding space, ensuring dimensional compatibility with textual tokens. For simplicity, we omit the projector from subsequent notation. The visual embeddings are concatenated with the question's textual embeddings $E_t(Q)$, forming an input sequence processed by an LLM to generate the answer auto-regressively:

$$A = \text{LLM}([v_1; v_2; \ldots; v_N; E_t(Q)]). \tag{1}$$

Note that the context-length limitation L of the above language model, *eg.,* 16K, constrains the total number of tokens, including visual and textual embeddings:

$$\sum_{i=1}^{N} \text{len}(v_i) + \text{len}(E_t(Q)) \leq L. \tag{2}$$

Unlike image understanding tasks that process a single frame, this constraint makes frame selection crucial for effective video understanding, as the model must select frames judiciously within the limited context window.

Table 1. **Demonstration of the importance of selecting the right frames** (*i.e.*, GT segments) on MultiHop-EgoQA dataset. All models use 8 frames as input. We report the average sentence similarity as the evaluation metric.

Frame Selection	LLaVA-NeXT-Video-7B	InternVL2-8B
Uniform	61.7	71.8
GT segments	**62.8** (+1.1)	**72.5** (+0.7)
w.o. GT segments	59.2 (-2.5)	69.1 (-2.7)

4. BOLT

In this section, we first perform an empirical analysis to validate the importance of frame selection. This analysis incorporates annotated ground-truth segments from untrimmed videos and a challenging multi-source retrieval condition. Subsequently, we compare four different frame selection strategies, finding that inverse transform sampling can effectively prioritize question-relevant frames.

4.1. Impact of Ground Truth Segments

In VQA tasks, particularly those involving long-form video understanding, many queries pertain to specific temporal segments within hour-long untrimmed videos. We hypothesize that effectively retrieving query-relevant segments significantly reduces the complexity of reasoning for VLMs. To validate this hypothesis, we leverage existing VQA benchmarks that provide temporal interval annotations.

Specifically, we utilize MultiHop-EgoQA [5], a recently introduced long-form egocentric benchmark. This dataset provides temporal annotations of relevant intervals for each video-question pair, such as $\{(t_{s1}, t_{e1}), (t_{s2}, t_{e2}), \ldots\}$. To examine the significance of frame selection, we conduct experiments under three conditions: (1) *Uniform sampling (baseline)*: Frames are uniformly sampled from the entire video. (2) *Ground-truth segment-only sampling*: Frames are uniformly sampled only from question-relevant ground-truth segments. (3) *Complement-set sampling*: Frames are uniformly sampled only from segments outside the ground-truth intervals. The results are summarized in Table 1.

We evaluate pretrained models InternVL2-8B [6] and LLaVA-NeXT-Video-7B [44] using 8 sampled frames, reporting average sentence similarity as the evaluation metric (higher is better). Experimental results show that, compared with uniform sampling, excluding irrelevant context enables VLMs to better focus on informative visual features for question answering. Conversely, when only irrelevant frames are provided, the generated predictions exhibit lower similarity to ground-truth answers.

4.2. Multi-Source Retrieval Evaluation

The above preliminary analysis demonstrates the significance of frame selection in VQA tasks by directly evaluating the models' capability on ground-truth segments. How-

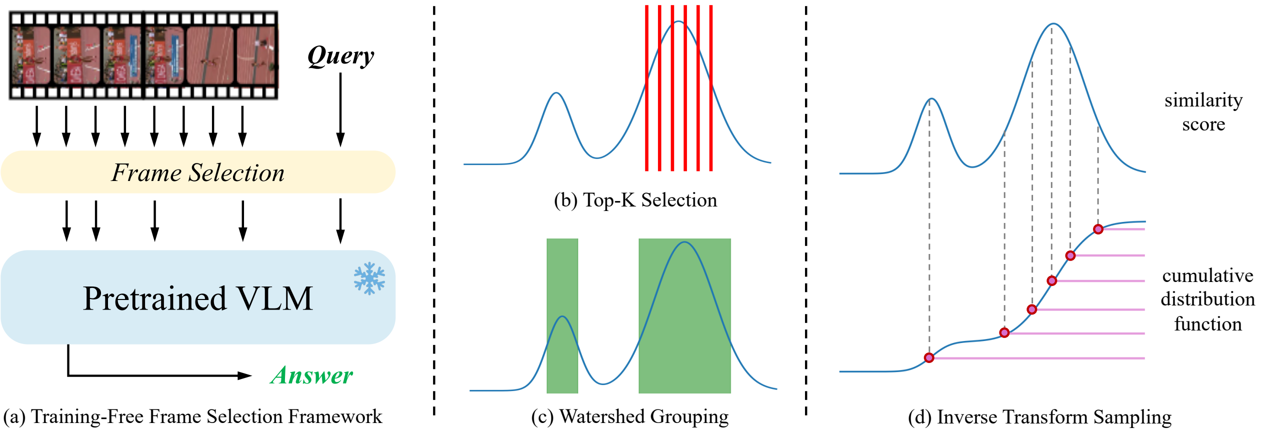

Figure 2. **(a) Proposed training-free frame selection framework.** Without exhaustive fine-tuning of VLMs, their performance can be improved by selecting query-related frames. **(b) Top-K selection** often yields redundant frames with limited diversity. **(c) Watershed grouping** may overlook broader temporal context. **(d) Inverse transform sampling** can prioritize relevant frames while preserving selection diversity, achieving stronger performance on benchmarks, particularly for videos with noisy backgrounds.

Table 2. **Multi-source retrieval evaluation on Video-MME.** K is set to 4 in this experiment. Using the 'right' frames (GT segments) significantly outperforms uniformly sampling frames.

Frame Selection Strategy	LLaVA-OneVision-7B
GT segments (16 frames)	56.85
Uniform (64 frames)	52.11 (-4.74)
Uniform (16 frames)	47.78 (-9.07)
Blind test	40.85

ever, extending this analysis to general VQA benchmarks, such as Video-MME [9], is constrained by the absence of temporal segment annotations. To overcome this limitation and assess frame selection efficacy under more challenging conditions, we introduce a novel multi-source retrieval evaluation setting.

The proposed multi-source retrieval evaluation employs a systematic construction procedure. First, K distinct videos are sampled from the benchmark and concatenated along their temporal dimension, with brief black-frame transitions inserted between adjacent videos. During video selection, we ensure that the sampled videos have similar durations yet differ significantly in their visual content, such as different categories or domains. This concatenation generates a pseudo-long-form video that effectively simulates real-world continuous video content. Next, we randomly select a question Q from the associated question sets of these videos and evaluate the model's response accuracy using the concatenated video, deviating from the conventional trimmed-video evaluation paradigm.

Our proposed evaluation framework offers distinct advantages over standard evaluation protocols. Primarily, it better approximates real-world scenarios in which models must identify relevant content within extended video sequences, such as lengthy instructional or surveillance footage. Furthermore, it imposes more rigorous reasoning demands on VLMs, providing a more comprehensive assessment of their capabilities. We implement this evaluation framework on the Video-MME dataset, and the preliminary results are presented in Table 2.

The experimental results reveal several critical insights. Most notably, uniform sampling exhibits substantial performance degradation under proposed evaluation framework, indicating its fundamental inadequacy in handling long video content containing substantial query-irrelevant segments. Despite increasing the input frame count to 64 frames, accuracy still exhibits a significant decline from 56.85% to 52.11%. With a fixed frame count of 16 frames, performance further deteriorates to 47.78%, representing a notable 9.07% reduction compared to the baseline. This result is even close to the blind test result, in which no visual information is provided. These findings underscore the importance of intelligent frame selection strategies for VQA tasks, particularly when processing long video sequences.

4.3. Selecting the Frames Right

Based on the above analysis, we emphasize that it is essential for VLMs to select frames relevant to the query. Next, we discuss how to select those frames effectively, and introduce several frame selection strategies designed to improve the identification and prioritization of question-relevant frames without requiring model fine-tuning.

Uniform Sampling. We first revisit conventional uniform sampling, which selects frames at fixed intervals: $f'_i = f_{\lfloor i \times T/N \rfloor}$, where $i \in \{1, \ldots, N\}$. Although uniform sampling is computationally efficient, it neglects the query context by implicitly assuming a uniform distribution of information throughout the video. Since uniform sampling remains widely used during both training and testing in current state-of-the-art VLMs, we adopt this as our baseline.

Top-K Selection. To identify query-relevant frames, an intuitive approach is to leverage text-video similarity and select frames with high similarity scores. A commonly used method is top-K sampling, which operates as follows:

First, the question Q and each frame f_i are encoded into embeddings using a pretrained contrastive lightweight multimodal network, such as CLIP [26] or SigLIP [40]. The similarity between the visual embedding of a frame $E_v(f_i)$ and the textual embedding of the question $E_t(Q)$ can then be computed, typically using cosine similarity:

$$s(f_i) = \cos(E_t(Q), E_v(f_i)). \tag{3}$$

Next, we select the top K frames with the highest similarity scores. These query-relevant frames are retained and subsequently processed by the VLM. Formally, the selection process is defined as:

$$\mathcal{F}_{\text{top-}K} = \{f_i \mid s(f_i) \in \text{Top-}K(\{s(f_j)\}_{j=1}^{N})\}. \tag{4}$$

Although effective for direct visual matching, top-K sampling exhibits notable limitations. As shown in Figure 2(b), when applied to long video sequences, top-K sampling often yields redundant frames with high visual similarity, potentially missing diverse contextual information crucial for complex queries. Moreover, this strategy may inadequately address temporal dependencies and causal relationships, which require analyzing scene transitions and broader contextual continuity.

Watershed Grouping. The watershed algorithm, originally developed for image segmentation [10], can be effectively adapted for temporal clustering in video understanding tasks. This approach conceptualizes similarity scores as a topographical landscape, where local extrema define distinct temporal segments.

Formally, given the frame-query similarity scores $s(f_i)$, we first apply a smoothing operation, such as sliding-window average pooling, to mitigate outliers and obtain refined scores: $s_i^r = \text{smooth}(s(f_i))$. Subsequently, the watershed algorithm computes the mean similarity score across the temporal sequence: $\overline{s} = \text{mean}\{s_1^r, s_2^r, \ldots, s_T^r\}$. This mean score $\overline{s}$ serves as a threshold for segmenting frames into distinct peak and valley regions, based on whether $s_i^r > \overline{s}$. Finally, we select representative frames exclusively from the top K highest-scoring peak regions.

The watershed-based approach is particularly effective for long videos, where traditional top-K sampling tends to yield redundant frames. By clustering temporally adjacent frames, this method facilitates the extraction of a more diverse subset of frames. Nevertheless, due to its grouping mechanism, this strategy can overlook important temporal context and still retain temporal redundancy, as illustrated in Figure 2(c). This limitation poses challenges, especially when addressing queries involving causal or explanatory reasoning (*e.g.*, *why* or *how*).

Inverse Transform Sampling. Analysis of top-K sampling and watershed grouping methods reveals two critical requirements: the selection of query-relevant frames and the preservation of frame diversity. Additionally, contextual information surrounding query-relevant segments should be retained. To meet these requirements without additional model fine-tuning, we propose an adaptive frame selection method based on *inverse transform sampling* (ITS).

ITS is a probabilistic sampling technique that generates samples from a specified probability distribution. When applied to video understanding, ITS leverages frame-query similarity scores to construct a query-guided probability distribution over frames, where the probability of selecting a frame f_i is proportional to its similarity score s_i. Specifically, we first normalize the similarity scores s_i across the entire video sequence to the range [0, 1] and refine the normalized score as follows:

$$s_i^r = \left(\frac{s_i - \min(s)}{\max(s) - \min(s)}\right)^{\alpha}, \tag{5}$$

where α controls the sharpness of the distribution. Smaller values of α produce a more uniform distribution, increasing frame diversity, while larger values yield a sharper distribution, emphasizing query-relevant frames.

Next, we construct the cumulative distribution function, $F(i)$, based on the refined scores:

$$F(i) = \frac{\sum_{j=1}^{i} s_j^r}{\sum_{j=1}^{N} s_j^r}. \tag{6}$$

Finally, we sample N frames using uniform sampling over the inverse transform of this cumulative distribution function:

$$f_i' = f_{\arg\min_k\{F(k) \geq i/N\}}, \quad i \in \{1, \ldots, N\}, \tag{7}$$

where f_i' represents the i-th sampled frame.

By employing the above inverse transform sampling, frames with higher relevance to the query have an increased probability of being sampled, while frames with lower relevance retain a non-zero probability. This ensures an effective balance between capturing essential query-related visual content and preserving sufficient contextual and temporal diversity. Additionally, by controlling the distribution sharpness through the parameter α, ITS provides flexible adjustment of the relevance-diversity trade-off. Empirically, we find that setting α between 2.5 and 3.5 yields satisfactory improvements.

Importantly, ITS requires no additional training and can be seamlessly integrated into existing VLMs during inference. Our experimental results demonstrate that ITS consistently outperforms alternative approaches, achieving a 2.3% improvement over uniform sampling on the Video-MME benchmark and a 4.5% improvement on the MLVU benchmark.

Table 3. **Comparison of our proposed BOLT with state-of-the-art models on standard benchmarks.** Inverse transform sampling is used. Our training-free approach BOLT enables off-the-shelf VLM LLaVA-OneVision to focus on query-related frames, thereby boosting the VQA performance in various settings. It also outperforms other large VLMs in the literature with comparable model scales.

Model	**LLM Size**	**Training Free**	**Video-MME** (w.o. sub.)				**EgoSchema**		**LongVideo Bench**	**MLVU**	**NextQA**
			Overall	Short	Medium	Long	Full	Subset			
Video-LLaVA [17]	7B	✗	39.9	45.3	38.0	36.2	-	-	39.1	47.3	-
VideoChat2 [14]	7B	✗	39.5	48.3	37.0	33.2	54.4	63.6	39.3	44.5	-
ShareGPT4Video [4]	8B	✗	39.9	48.3	36.3	35.0	-	-	41.8	46.4	-
Chat-UniVi-V1.5 [11]	7B	✗	40.6	45.7	40.3	35.8	-	-	-	-	-
VideoLLaMA2 [7]	7B	✗	47.9	56.0	45.4	42.1	-	53.1	-	-	-
Frame-Voyager [39]	8B	✗	57.5	67.3	56.3	48.9	-	-	-	65.6	73.9
LongVU [28]	7B	✗	60.9	64.7	58.2	59.5	-	-	-	65.4	-
LLaVA-OneVision$^{8\text{ frame}}$ [13]	7B		53.8	63.6	52.0	45.7	59.17	62.0	54.2	58.9	77.4
LLaVA-OneVision$^{8\text{ frame}}$+BOLT	7B	✓	**56.1**	**66.8**	**54.2**	**47.3**	**59.23**	**62.2**	**55.6**	**63.4**	**77.4**
LLaVA-OneVision$^{16\text{ frame}}$ [13]	7B		56.9	68.3	54.0	**48.2**	59.49	61.4	55.7	61.2	78.1
LLaVA-OneVision$^{16\text{ frame}}$+BOLT	7B	✓	**57.8**	**69.2**	**56.8**	47.3	**59.86**	**61.8**	**57.0**	**65.8**	**78.3**
LLaVA-OneVision$^{32\text{ frame}}$ [13]	7B		58.5	**70.3**	56.6	48.8	60.36	62.2	56.4	63.2	79.4
LLaVA-OneVision$^{32\text{ frame}}$+BOLT	7B	✓	**59.9**	70.1	**60.0**	**49.6**	**60.66**	**64.0**	**59.6**	**66.8**	**79.5**

5. Experiments

5.1. Experimental Settings

Datasets. We evaluate our approach on five widely used VQA benchmarks: Video-MME [9], EgoSchema [22], LongVideoBench [32], MLVU [46], and NextQA [35]. Among them, the NextQA and EgoSchema datasets contain short videos averaging less than 2 minutes in length, focusing primarily on short-form question answering. In contrast, Video-MME, MLVU, and LongVideoBench include longer videos, ranging from a few minutes to around two hours, thus evaluating models' capabilities in long-form video understanding. Notably, Video-MME consists of 900 videos with 2,700 questions and categorizes videos into short, medium, and long subsets, which we specifically leverage in our ablation studies. Importantly, we evaluate Video-MME without subtitles.

Implementation Details. For computational efficiency, we sample video frames at 1 FPS. To compute the similarity between video frames and textual queries, we use the pretrained CLIP-L/14 model [26] to extract visual and textual embeddings. The hyperparameter α is set to 2.5 for EgoSchema, LongVideoBench, NextQA, and to 3 for Video-MME and MLVU. We utilize the LMMs-Eval library [42] to measure accuracy for multiple-choice tasks. All experiments are conducted on two A100 GPUs. Additional details can be found in the supplementary material.

5.2. Benchmark Results

Comparison with State-of-the-Art Methods on Standard Benchmarks. Table 3 presents the comparison of our approach against state-of-the-art methods. Recently, LLaVA-OneVision [13] has achieved strong performance

Table 4. **Our training-free approach BOLT boosts various VLMs.** Inverse transform sampling is used on the Video-MME benchmark. All models use 8 frames.

Model	**LLM Size**	**With BOLT**	**Video-MME** (w.o. sub.)			
			Overall	Short	Medium	Long
Video-LLaVA	7B	✗	37.6	42.7	37.1	33.0
		✓	**39.3**	**45.7**	**38.0**	**34.1**
LongVA	7B	✗	48.5	56.7	46.6	42.3
		✓	**51.6**	**59.8**	**51.0**	**43.9**
InternVL2	8B	✗	52.6	62.4	51.1	44.2
		✓	**53.3**	**65.6**	**50.4**	**43.9**
LLaVA-Video	7B	✗	56.0	67.8	53.6	46.7
		✓	**58.6**	**70.4**	**55.7**	**49.9**

on these datasets, surpassing prior methods such as VideoLLaMA2 [7] and thus establishing a robust baseline. Our training-free, query-guided inverse transform sampling (ITS) method further enhances the performance of LLaVA-OneVision. Specifically, on Video-MME, accuracy improves from 53.8% to 56.1% using 8 frames, and on MLVU, it increases from 58.9% to 63.4% with 8 frames. On NextQA, our performance remains comparable to the baseline, likely due to the brevity of the videos. Notably, our method continues to yield improvements even with 32 frames, demonstrating significant gains on long-video benchmarks, such as a 3.2% increase on LongVideoBench and a 3.6% increase on MLVU. These results highlight the importance of selecting the right frames at inference time to fully unleash the potential of existing VLMs, and our proposed inverse transform sampling consistently surpasses uniform sampling by a substantial margin.

Table 5. **Our BOLT boosts existing VLMs under multi-source retrieval evaluation.** Inverse transform sampling is used. LLaVA-OneVision-7B is utilized and the K is set to 4.

Model	Video-MME (w.o. sub.)			
	Overall	Short	Medium	Long
LLaVA-OneVision$^{64\text{ frame}}$	52.1	61.8	50.8	43.8
LLaVA-OneVision$^{64\text{ frame}}$ **+BOLT**	**54.9**	**66.1**	**54.2**	**44.4**
LLaVA-OneVision$^{16\text{ frame}}$	47.8	55.8	45.6	42.0
LLaVA-OneVision$^{16\text{ frame}}$ **+BOLT**	**53.4**	**65.1**	**51.1**	**43.3**

Results on Off-the-Shelf VLMs. As shown in Table 4, we extend the proposed approach to additional video VLMs, including Video-LLaVA [17], LongVA [43], InternVL2 [6], and LLaVA-Video [45], and evaluate its effectiveness on the Video-MME benchmark. The results clearly demonstrate that our training-free approach, BOLT, consistently boosts the performance of off-the-shelf VLMs, achieving approximately 1% to 3% accuracy improvements. For instance, the performance of LLaVA-Video notably increases from 56.0% to 58.6% with only 8 frames, verifying the strong generalization capability of inverse transform sampling.

Results on Proposed Multi-Source Retrieval Evaluation Setting. We further evaluate our method under the retrieval-based evaluation setting on Video-MME, as described in Section 4.2. This evaluation assesses the model's ability to retrieve query-relevant information from noisy video content. As shown in Tables 3 and 5, LLaVA-OneVision's performance significantly decreases from 56.9% to 47.8% when using 16 frames, demonstrating that irrelevant background frames substantially hinder comprehension. Even increasing the frame count to 64 frames (equivalent to 16 frames per candidate video) yields only partial recovery (52.1% vs. 47.8%), confirming that background noise leads to considerable distraction in pretrained VLMs.

Applying our proposed frame selection strategy consistently improves VLM accuracy. With 16 frames, performance on Video-MME increases from 47.8% to 53.4%, representing a substantial 5.6% gain. Furthermore, results indicate that BOLT effectively enhances accuracy across short, medium, and long videos. These findings highlight both the importance of selecting query-relevant frames and maintaining frame diversity to maximize model performance.

5.3. Ablations and Analysis

To further validate the effectiveness of our training-free approach, we conduct ablation studies and analyses from four perspectives: (1) comparisons with alternative frame selection techniques, (2) comparison of visual-text similarity measures, (3) analysis on the hyperparameter α, and (4) analysis with respect to video length. Additional experimental results and efficiency analysis are provided in the supplementary material.

Table 6. **Comparison of different training-free frame selection strategies on Video-MME dataset**. LLaVA-OneVision-7B is utilized with 16 frames.

Frame Selection	Query	Standard	Retrieval-Based
Blind Test	✗	40.85	
Uniform	✗	56.85	47.78
Uniform + Shuffle	✗	56.25 (-0.60)	46.93 (-0.85)
Random	✗	55.15 (-1.70)	46.05 (-1.73)
Top-K	✓	57.10 (+0.25)	48.45 (+0.67)
Watershed	✓	57.23 (+0.38)	49.15 (+1.37)
ITS	✓	**57.78** (+0.93)	**53.37** (+5.59)

Table 7. **Comparison of different visual-text similarity measures on Video-MME dataset**. LLaVA-OneVision-7B is utilized with 16 frames. The baseline is uniform sampling.

Baseline	Caption	SigLIP	CLIP	CLIP + SigLIP
56.85	55.47	57.00	**57.78**	**57.83**

Comparison with Alternative Frame Selection Techniques. Table 6 compares our proposed method with alternative training-free frame selection strategies. As a baseline setup, we exclude visual input entirely, relying solely on the question text (Blind Test). Surprisingly, even without visual data, the model achieves 41% accuracy on Video-MME, indicating that VLMs heavily utilize general knowledge to answer multiple-choice questions. When visual input is included, uniform and random sampling strategies yield similar performance, around 56% in standard evaluation and 47% in multi-source retrieval evaluation, suggesting limited effectiveness of traditional, query-agnostic frame selection approaches on datasets such as Video-MME, which does not require complex temporal reasoning.

However, incorporating query-visual similarity through strategies such as top-K sampling, watershed grouping, and inverse transform sampling consistently improves performance over uniform sampling. Among these methods, ITS provides robust gains across both standard and retrieval-based evaluations, achieving the best overall performance. Remarkably, in the retrieval-based evaluation, ITS significantly boosts accuracy from 47.78% to 53.35%, highlighting the importance of selecting both query-relevant frames, and underscoring the effectiveness of our proposed method.

Comparison of Visual-Text Similarity Measures. Since our approach leverages similarity between image and query embeddings for frame selection, we compare various image representation methods, including image captions and pretrained vision-language encoders, as shown in Table 7. We initially caption each frame using LLaVA-OneVision and compute frame selection based on the similarity between queries and generated captions. However, this caption-based approach underperforms due to noisy captions and in-

Table 8. **Analysis on the sharpness hyperparameter α in ITS**. LLaVA-OneVision-7B is utilized with 16 frames.

α	**Video-MME**	**MLVU**	**LongVideoBench**
Uniform	56.90	61.24	55.72
2	58.37	65.08	58.04
2.5	57.51	65.24	57.00
3	57.78	65.76	57.67
3.5	57.74	65.11	58.41

Table 9. **Impact of video number K in multi-source retrieval evaluation.** If $K = 1$, the evaluation degrades to standard VQA setting. When K is larger, the background noise is also larger.

Video Number K	**1**	**2**	**4**	**8**
Uniform	56.85	51.81	47.78	44.82
Watershed	57.23	53.70	48.67	45.89
ITS	**57.78**	**54.93**	**53.37**	**49.04**

accurate similarity scores. In contrast, directly using image encoders achieves consistent improvements over the uniform sampling baseline, with CLIP outperforming SigLIP. Furthermore, averaging the similarity scores from CLIP and SigLIP yields the best overall performance.

Analysis on the Sharpness Hyperparameter α. As described in Eq. 5, the hyperparameter α controls the sharpness of the similarity distribution. Table 8 investigates the impact of different α values. When α is set between 2 and 3.5, the VQA performance consistently surpasses uniform sampling across all three datasets by a large margin. Moreover, performance variation within this range is minor, indicating that ITS is less sensitive to the exact value of α.

Impact of Background Noise Level. To examine the effect of background noise on model performance, we conduct an analysis using multi-source retrieval evaluation setting, where target videos are concatenated with randomly selected background videos, simulating long and noisy video environments. Table 9 shows that introducing background videos degrades accuracy (frame budget 16), with uniform sampling being particularly susceptible to the increased noise. Although ITS also experiences a performance drop, it consistently outperforms uniform sampling. These results emphasize the effectiveness of query-guided frame selection in mitigating the adverse impacts of noisy contexts.

5.4. Visualization

To further illustrate the frame selection results of inverse transform sampling, we provide two examples from the Video-MME in Figure 3. As shown in the figure, ITS can effectively sample query-relevant frames with high similarity scores. At the same time, ITS also includes certain local peaks as contextual information, ensuring frame diversity and providing the VLMs with crucial context for reasoning.

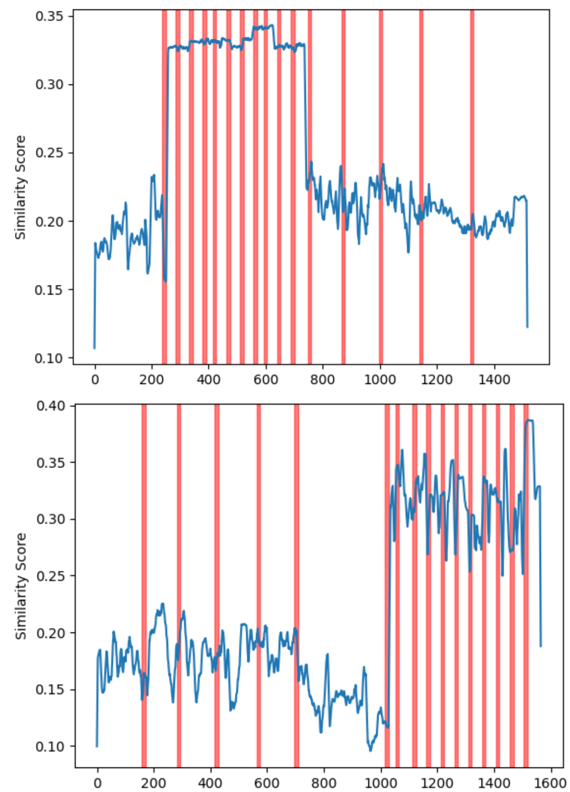

Figure 3. **Visualization of frame selection results using inverse transform sampling.** The blue curve represents the similarity scores across the entire video sequence, and the red lines indicate the selected frames.

6. Conclusion

In this paper, we introduced BOLT, a training-free method designed to enhance large VLMs for long-form video understanding. We explored various temporal frame selection techniques, focusing particularly on inverse transform sampling, to efficiently prioritize query-relevant frames, significantly improving visual question-answering performance. Our evaluations demonstrate that BOLT effectively addresses the limitations of conventional uniform sampling, providing a scalable solution to extend the capabilities of existing VLMs without requiring additional fine-tuning.

Limitations. Although BOLT significantly improves long-form video understanding, its effectiveness relies on accurate query-frame similarity measures, which may not always reflect true relevance on tasks that require reasoning. Future work could explore more robust similarity metrics that better adapt to varied video content and query contexts, including under multi-round VQA scenarios.

Acknowledgments. This work is supported by the KAUST Center of Excellence for Generative AI under award number 5940. The computational resources are provided by IBEX, which is managed by the Supercomputing Core Laboratory at KAUST.

References

[1] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024. 1

[2] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. Goldfish: Vision-language understanding of arbitrarily long videos. In *ECCV*, pages 251–267. Springer, 2024. 2, 3

[3] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. 3

[4] 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. In *NeurIPS*, pages 19472–19495, 2024. 1, 2, 6

[5] Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. *arXiv preprint arXiv:2408.14469*, 2024. 2, 3

[6] 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 *CVPR*, pages 24185–24198, 2024. 3, 7

[7] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs, 2024. 6

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. The llama 3 herd of models, 2024. 2

[9] Chaoyou Fu, Yuhan Dai, Yondong 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. *arXiv preprint arXiv:2405.21075*, 2024. 2, 4, 6

[10] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. In *TPAMI*, pages 532–550. IEEE, 1987. 5

[11] 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 *CVPR*, pages 13700–13710, 2024. 1, 6

[12] 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. 1, 2

[13] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 1, 2, 6

[14] 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*, 2023. 1, 6

[15] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models, 2023. 2

[16] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024. 2

[17] 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. *arXiv preprint arXiv:2311.10122*, 2023. 1, 6, 7

[18] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*, 2023. 1

[19] 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, 2023. 2

[20] Shuming Liu, Chen-Lin Zhang, Chen Zhao, and Bernard Ghanem. End-to-end temporal action detection with 1b parameters across 1000 frames. In *CVPR*, pages 18591–18601, 2024. 1

[21] 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. 1, 2

[22] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 6

[23] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13235–13245, 2024. 1, 2

[24] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. Gpt-4 technical report, 2024. 2

[25] 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, 2024. 2

[26] 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. 5, 6

[27] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 1

[28] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding, 2024. 1, 2, 6

[29] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding, 2024. 1

[30] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232, Seattle, WA, USA, 2024. IEEE. 1, 2

[31] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey, 2024. 2

[32] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 37:28828–28857, 2024. 2, 6

[33] 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18699–18708, 2024. 1, 2

[34] Yongliang Wu, Bozheng Li, Jiawang Cao, Wenbo Zhu, Yi Lu, Weiheng Chi, Chuyun Xie, Haolin Zheng, Ziyue Su, Jay Wu, and Xu Yang. Zero-shot long-form video understanding through screenplay, 2024. 2

[35] 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)*, pages 9777–9786, 2021. 6

[36] 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, 2024. 1

[37] 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, 2023. 1, 2

[38] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 1

[39] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-voyager: Learning to query frames for video large language models, 2024. 1, 2, 3, 6

[40] 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*, pages 11975–11986, 2023. 5

[41] 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. *arXiv preprint arXiv:2312.17235*, 2023. 1

[42] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 6

[43] 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*, 2024. 7

[44] 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, 2024. 3

[45] 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*, 2024. 7

[46] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 2, 6

[47] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 2