LVBench: An Extreme Long Video Understanding Benchmark

Weihan Wang[1] Zehai He[2] Wenyi Hong[2] Yean Cheng[1] Xiaohan Zhang[1]
Ji Qi[1] Ming Ding[1] Xiaotao Gu[1*] Shiyu Huang[1*] Bin Xu[2] Yuxiao Dong[2*] Jie Tang[2]
[1]Zhipu AI [2]Tsinghua University
weihan.wang@aminer.cn, xiaotao.gu@aminer.cn,
shiyu.huang@zhipuai.cn, yuxiaod@tsinghua.edu.cn

Abstract

Recent progress in multimodal large language models has markedly enhanced the understanding of short videos (typically under one minute), and several evaluation datasets have emerged accordingly. However, these advancements fall short of meeting the demands of real-world applications such as embodied intelligence for long-term decision-making, in-depth movie reviews and discussions, and live sports commentary, all of which require comprehension of long videos spanning several hours. To address this gap, we introduce LVBench, a benchmark specifically designed for long video understanding. Our dataset comprises publicly sourced videos and encompasses a diverse set of tasks aimed at long video comprehension and information extraction. LVBench is designed to challenge multimodal models to demonstrate long-term memory and extended comprehension capabilities. Our extensive evaluations reveal that current multimodal models still underperform on these demanding long video understanding tasks. Through LVBench, we aim to spur the development of more advanced models capable of tackling the complexities of long video comprehension.

1. Introduction

The rapid development of Multi-modal Large Language Models (MLLMs) has led to significant advances in vision-language understanding [26, 37, 38], driven by breakthroughs in both large language models [3, 8, 25] and visual encoders [30, 35, 46]. These models have demonstrated impressive capabilities across various video understanding tasks, from fundamental tasks like action recognition and object tracking, to more complex tasks such as temporal event localization, dense video captioning, and abstractive video summarization. However, while existing MLLMs excel at processing short video clips, they face substantial challenges when dealing with longer temporal sequences - a crucial limitation that hinders their practical applications. Despite numerous video understanding benchmarks being proposed, most primarily focus on evaluating spatial understanding in static images or short video clips, overlooking the critical aspect of temporal understanding in long-form videos. This gap is particularly significant given that real-world applications often require comprehension of extended temporal contexts. The scarcity of long video understanding benchmarks can be attributed to the challenges in data collection and the complexity of annotation. To address these limitations, we introduce LVBench, a comprehensive benchmark designed specifically for evaluating MLLMs' capabilities in long video understanding. Our benchmark distinguishes itself from existing datasets through several key features:

- We establish a systematic framework of six core temporal understanding capabilities, which can be flexibly combined to create complex, multi-faceted evaluation tasks. This design enables a thorough assessment of models' ability to process and reason about extended temporal sequences.
- We curate a diverse collection of long-form videos from various domains, with an average duration approximately four times longer than existing benchmarks. The videos span multiple categories (as illustrated in Figure 1), providing a robust foundation for evaluating temporal understanding across different contexts.
- Through a rigorous combination of expert human annotation and quality control processes, we ensure high-quality ground truth annotations, establishing a reliable benchmark for assessing long video understanding capabilities.

2. Related Works

2.1. Multi-modal Large Language Models

Building upon the achievements in LLMs, the field has shifted towards MLLMs to enhance multi-modal understanding capabilities [2, 11, 26, 32, 38]. Early advance-

*Corresponding authors

Figure 1. Examples from LVBench. LVBench covers problems involving various temporal and spatial dimensions.

ments in this area include models like Flamingo [2], which introduced a novel approach by adding cross-attention layers to a frozen language model, enabling efficient vision-language alignment without full model retraining. Recent works have made significant progress in video understanding through various architectural innovations. Video-LLaMA [48] employs a video transformer with Q-Former for effective temporal modeling. Models like InternVL2 [6], LLaVA-Next [50], and CogVLM2 [12] unify high-resolution images and videos as multi-frame sequences, enabling seamless processing of both modalities. Qwen2-VL [37] introduces 3D-RoPE to achieve strong temporal extrapolation capabilities for long video understanding. To handle extended temporal contexts, MovieChat [34] proposes a dual-memory architecture with short-term and long-term memory modules that can efficiently process videos over 10,000 frames. LWM[20] with ring attention has pushed context lengths to the million-token scale, providing a promising foundation for long video understanding. PLLaVA[43] introduced a resource-

Table 1. Comparison of different datasets. **Open-domain** represents whether the video source is diversified. **Multi-type** represents whether the types of questions are greater than 2 categories.

Dataset	Num QA	Avg sec	Open-domain	Multi-type	Annotation
TGIF-QA [14]	165,165	3	✓	✗	Auto
MSVD-QA [42]	13,157	10	✓	✗	Auto
MSRVTT-QA [42]	72,821	15	✓	✗	Auto
MVBench [17]	4,000	16	✓	✓	Auto
Perception Test [29]	44,000	23	✗	✓	Auto&Manual
NExT-QA [41]	52,044	44	✓	✗	Manual
How2QA [18]	44,007	60	✓	✓	Manual
ActivityNet-QA [45]	800	111	✗	✗	Manual
CinePile [31]	**303,828**	160	✗	✓	Auto&Manual
EgoSchema [24]	5,000	180	✗	✓	Auto&Manual
MovieQA [36]	6,462	203	✗	✓	Manual
LongVideoBench [40]	6,678	473	✓	✓	Manual
MovieChat-1K [34]	13,000	564	✗	✓	Manual
MoVQA [49]	21,953	992	✗	✓	Manual
Video-MME [9]	2,700	1018	✓	✓	Manual
LVBench(Ours)	1,549	**4,101**	✓	✓	Manual

efficient method for adapting image-language pre-trained models to dense video understanding through a novel feature pooling strategy. Despite these significant architectural advances, our experiments indicate that current video understanding models still fall short on tasks requiring long-range comprehension, highlighting an urgent need for the development of models specifically tailored for long video understanding.

2.2. Benchmarks for MLLM

Recent advancements in vision-language benchmarks have largely focused on images and short videos, as seen in datasets like MMBench [22], SEED-Bench-2 [15], TGIF-QA [14] and MVBench [17]. For long video understanding, previous benchmarks such as Perception Test [29] have explored multi-modal video perception and reasoning but often with shorter video clips and limited temporal context. Datasets like How2QA [18] and ActivityNet-QA [45] are domain-specific and do not adequately capture the complexity of long-term video understanding. EgoSchema [24] and MovieQA [36] provide insights into narrative and thematic understanding but are constrained by shorter video durations and limited granularity. While LongVideoBench [40], MovieChat [34], MoVQA [49], and Video-MME [9] utilize longer videos to test models, their average duration is still limited to around 10 minutes. In contrast, LVBench features significantly longer video segments averaging 4101 seconds, pushing the boundaries of long-term video understanding with comprehensive tasks and detailed annotations. A detailed comparison between LVBench and existing video benchmarks is presented in Table 1.

3. LVBench

In this chapter, we primarily discuss the construction of the original dataset for LVBench and the generation and optimization of the question-answer pairs.

3.1. Dataset Collection

We define long videos as those having a minimum duration of 30 minutes and containing rich, dynamic visual information with multiple events and scene transitions. To construct a comprehensive benchmark for long video understanding, we curated a diverse collection of videos from YouTube through the following systematic process.

Video Collection. Starting with YouTube's search and recommendation algorithms, we gathered an initial pool of 500 videos using carefully selected keywords across different domains. Our search terms covered six primary categories: Sports Competitions (e.g., basketball matches and Olympic games), Documentary Films (e.g., nature documentaries and historical records), Event Records (e.g., ceremonies and festivals), Lifestyle and Daily Activities (e.g., cooking tutorials and travel vlogs), TV Shows and Drama Series (e.g., sitcoms and reality shows), and Cartoon Videos (e.g., animated movies and cartoon series).

Quality Filtering. Our annotators carefully screened these videos using five well-defined criteria:

- **Clear Protagonist Presence**: Videos must feature identifiable main characters (human or virtual) who appear consistently and interact meaningfully with their environment. For first-person perspective videos, the narrator's actions and decisions should be clearly observable.
- **Structural Coherence**: Videos should maintain a well-defined narrative structure with clear beginning, development, and conclusion phases. We define "coherent logical flow" as the presence of causally connected events that progress naturally through time.
- **Event Density**: Videos must contain multiple distinguishable events (at least one significant event every 5 minutes on average) arranged in chronological order. We define events as meaningful actions, interactions, or state changes that contribute to the video's narrative.
- **Visual Clarity**: The visual content should be of high quality (minimum 720p resolution) with stable camera work and clear scene compositions. We exclude videos with excessive rapid cuts, shaky footage, or poor lighting conditions.
- **Modality Independence**: While audio may enhance understanding, all critical information should be conveyed through visual channels to ensure fair evaluation of vision-language models.

After applying these filtering criteria, we retained 103 high-quality videos totaling 117 hours of content.

3.2. Task Definition and Question Types

To systematically evaluate models' capabilities in understanding long videos, we define a comprehensive taxonomy of six core skills essential for video comprehension. Each skill category is carefully designed to test specific aspects of temporal and semantic understanding while allowing for compositional combinations that assess more complex reasoning abilities.

Temporal Grounding (TG). These questions assess the model's ability to locate and understand events within the video's timeline, such as identifying specific moments (*"What happened at 29:30?"*), determining event durations, and recognizing event sequences. This capability is fundamental for understanding the temporal dynamics in long videos.

Summarization (Sum). Tests the ability to synthesize and abstract information across the entire video content. Questions in this category require models to generate concise overviews, identify key developments, and understand the video's central themes, demonstrating a comprehensive understanding of long-form content.

Reasoning (Rea). Evaluates higher-order cognitive skills including causal understanding (*"Why did the experiment fail?"*), emotional comprehension, intention interpretation, and future prediction. This category specifically examines a model's ability to make complex inferences beyond surface-level observations.

Entity Recognition (ER). Focuses on identifying and tracking key entities (people, objects, places) throughout the video. This includes recognizing distinct entities, understanding their relationships, tracking their actions, and connecting them to relevant events - skills crucial for following long-form narratives.

Event Understanding (EU). Tests comprehension of high-level semantic concepts by requiring models to classify video types, detect significant events and understand scene transitions. This capability is essential for grasping the overall structure and flow of long videos.

Key Information Retrieval (KIR). Evaluates attention to specific details by requiring extraction and comprehension of visual text, numerical data, and other precise information presented in the video, such as *"What revenue growth did the firm report at the conference?"*

Figure 2 provides the distribution of each question type. As shown in the figure, Entity Recognition (ER) and Event Understanding (EU) account for a larger proportion of the questions. This imbalance naturally arises from the inherent structure of video content, where almost all questions inevitably involve identifying or tracking entities and understanding events or actions. Consider the question *"Why did Player number 4 in white push down Player number 17 in purple during the match?"*. While this question primarily tests reasoning capabilities (understanding the motivation behind an action), it inherently requires entity recognition (identifying and distinguishing the two players by their numbers and jersey colors) and event understanding (recognizing the pushing action within the context of the match).

3.3. Question-Answer Pair Generation

The annotation of long videos presents unique challenges compared to short-form content, requiring careful consideration of temporal dynamics and comprehensive coverage. We developed a systematic annotation protocol consisting of three key stages to ensure high-quality, diverse, and challenging question-answer pairs.

Video Analysis. In Stage 1 (Video Analysis), annotators first watch each video in its entirety. During this initial viewing, they mark significant events, scene transitions, and key entities, while also identifying potential temporal dependencies and causal relationships that could form the basis for meaningful questions.

Question Generation. Stage 2 (Question Generation) focuses on creating questions that thoroughly probe the video content. The number of questions scales with video duration, averaging 24 questions per hour. Questions are strategically distributed across the entire video duration to ensure comprehensive temporal coverage. Each video must

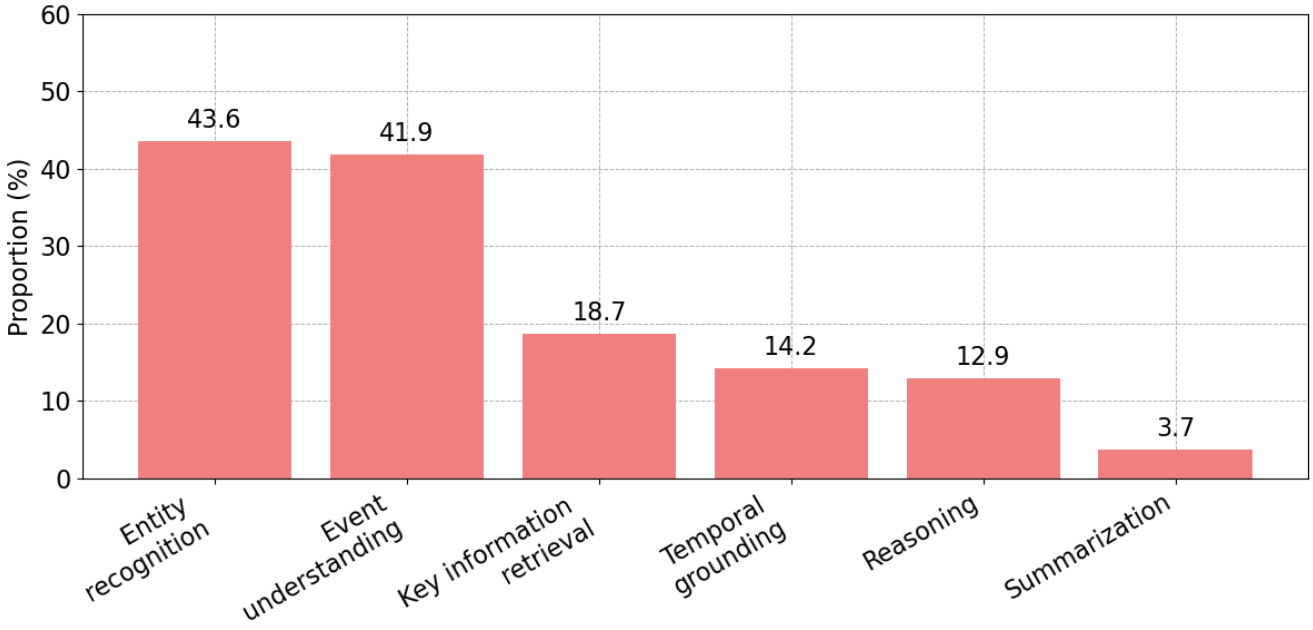

Figure 2. The proportion of different core capabilities.

include questions testing all six core capabilities. In constructing these questions, we emphasize specificity to ensure unique and unambiguous references to video content. For example, when questioning an argument scene, we prefer precise formulations like *"When did A and B start arguing?"* or *"How did the person in red's expression change during the hallway argument?"* over vague alternatives such as *"Why did they start arguing?"* or *"Who are the people arguing?"*. This specificity principle ensures that questions have clear, uniquely identifiable answers within the video context.

Answer Construction. Stage 3 (Answer Construction) focuses on both creating answer choices and annotating temporal information needed for answering each question. For each question, annotators create four answer choices - one correct answer and three distractors. The correct answer precisely addresses the question, while distractors are carefully designed to share similar format and length with the correct answer. These distractors incorporate plausible but incorrect information from the video and test different aspects of understanding.

Importantly, during this stage, annotators also manually identify and mark what we term the **"clue duration"** - the minimum video segment needed to accurately answer each question. This temporal annotation involves carefully reviewing the video content to determine the exact start and end timestamps of the critical segments containing the information required for answering the question. For instance, a question about causal relationships might require a longer clue duration spanning multiple related events, while a question about a specific visual detail might have a shorter clue duration focused on a single scene.

3.4. Data Quality Control

During the annotation process, we observed that annotators had a tendency to label most questions as temporal grounding, i.e., specifying a time range to limit the referent of the question. This practice may inadvertently reduce the difficulty of the questions and unfairly disadvantage video-understanding models that lack the ability to perceive the temporal dimension. To address this issue, we instructed annotators to minimize the number of temporal questions while still ensuring the uniqueness of the referents, effectively converting temporal grounding questions into other question types.

Upon constructing all the questions, we discovered that for certain questions, a language model could generate answers without any visual input. As highlighted in MM-star [5], many multimodal benchmarks can be effectively solved using pure text input alone. To mitigate this issue, we employed a straightforward yet effective approach. We utilized two powerful large language models, GLM-4 [8] and GPT-4 [1], to independently generate answers for all the questions. In cases where the outputs from both models were identical and matched the ground truth answer, we removed that particular data sample from the dataset. This filtering process successfully eliminated the majority of questions that did not rely on video content for answering. Following this filtering step, we obtained a refined set of 1,549 question-answer pairs.

Table 2. LVBench evaluation results regarding each core long video understanding capability. The highest score is highlighted with green, and the second highest is highlighted with purple. All the numbers are presented in % and the full score is 100%.

Model	LLM	ER	EU	KIR	TG	Rea	Sum	Overall
Non-Native Long Video Support Models								
TimeChat [33]	LLaMA2-7B	21.9	21.7	25.9	22.7	25.0	24.1	22.3
Video-ChatGPT [23]	Vicuna-1.5-13B	22.9	22.6	22.7	25.5	23.4	24.1	23.1
PLLaVA [43]	Yi-34B	25.0	24.9	26.2	21.4	30.0	25.9	26.1
LLaVA-OneVision [16]	LLaMA3-70B	25.0	26.9	29.2	30.9	25.4	31.0	26.9
CogVLM2-Video [12]	LLaMA3-8B	28.3	26.9	31.0	25.1	25.5	38.9	28.1
LLaVA-NeXT [50]	Yi-34B	30.1	31.2	34.1	31.4	35.0	27.6	32.2
InternVL2-40B [6]	Nous-Hermes-2-Yi-34B	37.4	39.7	43.4	31.4	42.5	41.4	39.6
mPLUG-Owl3 [44]	Qwen2-7B	46.0	41.6	42.4	41.1	47.5	40.4	43.5
GLM-4.1V-Thinking [13]	GLM-4-9B	42.8	43.5	50.0	39.1	46.0	25.9	44.3
VideoLLaMA3-7B [47]	Qwen2.5-7B	45.8	42.4	47.8	35.9	45.8	36.2	45.3
GLM4V-Plus [12]	GLM-4	46.2	47.8	54.1	42.7	46.5	37.9	48.7
GPT-4o-20241120 [26]	GPT-4o	48.9	49.5	48.1	40.9	50.3	50.0	48.9
GPT-4.1 [27]	GPT-4.1	-	-	-	-	-	-	60.1
Native Long Video Support Models								
MovieChat [34]	Vicuna-7B	21.3	23.1	25.9	22.3	24.0	17.2	22.5
LLaMA-VID [19]	Vicuna-13B	25.4	21.7	23.4	26.4	26.5	17.2	23.9
LWM [20]	LLaMA2-7B	24.7	24.8	26.5	28.6	30.5	22.4	25.5
Gemini-1.5-Pro [32]	Gemini 1.5 Pro	32.1	30.9	39.3	31.8	27.0	32.8	33.1
Kangaroo [21]	LLaMA3-8B	38.6	37.9	29.6	35.0	41.3	36.2	38.3
Qwen2-VL-72B [37]	Qwen2-72B	38.0	41.1	38.3	41.4	46.5	46.6	41.3
Qwen2.5-VL-72B [4]	Qwen2.5-72B	44.2	40.9	55.6	37.7	45.2	34.5	44.0
Gemini-2.0-Flash [7]	Gemini-2.0-Flash	47.4	48.5	56.8	39.3	44.4	41.4	48.6
AdaReTaKe [39]	Qwen2.5-72B	53.0	50.7	62.2	45.5	54.7	37.9	53.3
Gemini-2.5-Flash [7]	Gemini-2.5-Flash	55.2	55.5	63.8	52.7	55.5	44.8	56.7
MR.Video [28]	Gemini-2.0-Flash	59.8	57.4	71.4	58.8	57.7	50.0	60.8
Seed1.5-VL [10]	Seed1.5	64.3	64.0	64.7	52.3	65.0	51.7	64.0
Seed1.5-VL-Thinking [10]	Seed1.5	65.4	63.4	68.0	53.6	63.7	46.6	64.6
Gemini-2.5-Pro [7]	Gemini-2.5-Pro	64.5	67.5	72.8	65.9	66.5	58.6	67.4

4. Experiments

In this chapter, we report the experimental results of different video understanding models on LVBench and also compare them with human performance.

4.1. Settings

We evaluated the performance of 13 models that support multi images or short video input: TimeChat [33], Video-ChatGPT [23], PLLaVA [43], LLaVA-OneVision [16], CogVLM2-Video [12], LLaVA-NeXT [50], InternVL2-40B [6], mPLUG-Owl3 [44], GLM-4.1V-Thinking [13], VideoLLaMA3-7B [47], GLM4V-Plus [12], GPT-4o [26] and GPT-4.1 [27]. To adapt these models for long video inputs, we sample a fixed number of frames from the original video, such as 32 or 96 frames, to maintain consistency with the model's training sequence length. Additionally, we assessed 13 models that natively support long videos: LLaMA-VID [19], MovieChat [34], LWM [20], Gemini 1.5 Pro [32], Kangaroo [21], Qwen2-VL-72B [37], Qwen2.5-VL-72B [4], Gemini-2.0-Flash [7], AdaReTaKe [39], Gemini-2.5-Flash [7], MR.Video [28], Seed1.5-VL [10] and Gemini-2.5-Pro [7]. We processed the videos at a rate of 1 frame per second and fed them into the models, only performing downsampling when the video's length exceeded the model's maximum processing capability. It is worth noting that although Gemini 1.5 Pro

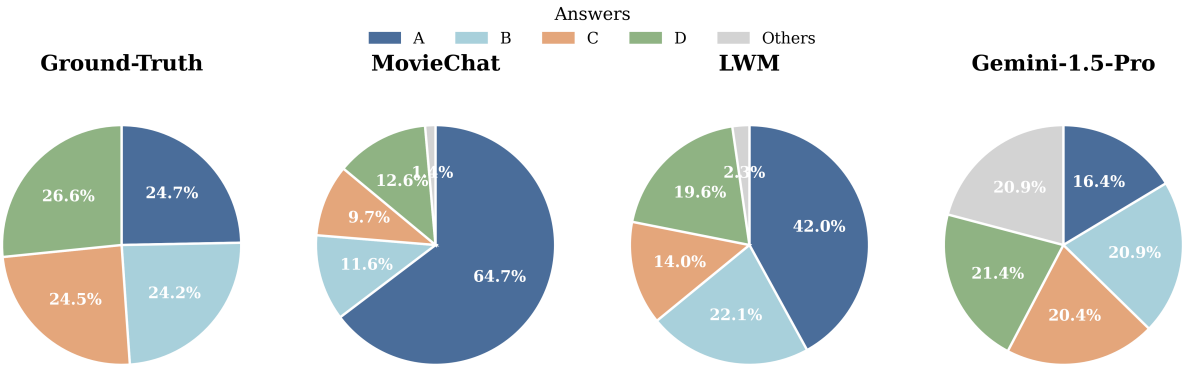

Figure 3. Failure analysis of native long-video models via answer distribution.

Table 3. LVBench evaluation results across different video categories.

Model	Sports	Documentary	Event Record	Lifestyle	TV Show	Cartoon	Overall
Human	96.3	89.8	87.4	98.4	97.2	95.8	94.4
Seed1.5-VL	63.3	67.5	60.3	66.9	60.4	65.0	64.0
Gemini-2.5-Pro	71.3	54.3	72.1	68.5	69.2	66.1	67.4

can handle videos up to 10 hours long, its publicly available interface is limited to processing videos of up to 1 hour in length. For each question, we provided the following prompt as input to the models:

> *Question (A) Option1 (B) Option2 (C) Option3 (D) Option4. Please select the best answer from the options above and directly provide the letter representing your choice without giving any explanation.*

After obtaining the model responses, we first attempted to extract the answers using regular expression matching. For questions where the matching process was unsuccessful, we employed a LLM to extract the answers from the responses.

4.2. Performance across Core Capabilities

To comprehensively evaluate the performance of various long video understanding models across core capabilities, we conducted extensive experiments on the LVBench dataset, testing multiple representative models, including both non-native and native long video support models. The experimental results are presented in Table 2.

Overall, Gemini-2.5-Pro demonstrated state-of-the-art (SOTA) performance, achieving a top score of 67.4 by outperforming all other models in 5 out of 6 tasks (EU, KIR, TG, Rea, and Sum). Seed1.5-VL-Thinking secured the second position with a score of 64.6. Our analysis reveals two key findings. First, a significant performance gap persists between proprietary and open-source models. The top-performing open-source models, VideoLLaMA3-7B (non-native long-video) and Qwen2.5-VL-72B (native long-video), scored 45.3 and 44.0, respectively, lagging behind Gemini-2.5-Pro by over 20% in both individual tasks and the overall score. Second, we observed that some models without native long-video support still achieved competitive results against their native counterparts, though their performance was not comparable to SOTA models. Conversely, several models designed for native long-video processing, including MovieChat, LLaMA-VID, LWM, and Gemini-1.5-Pro, exhibited unexpectedly low scores.

4.3. Analysis of Failure Modes

To understand why some native long video support models perform poorly on LVBench, we evaluated the distribution of answers generated by different models on LVBench and observed that existing long video understanding models struggle with precisely following instructions. For example, despite explicitly constraining the output in the prompt to be one of four provided answer choices, Gemini-1.5-Pro generated responses outside of the specified options 20.9% of the time, such as *"None of the above options are correct"* or *"I cannot answer this question"*. This occurred even though manual validation confirmed that the questions were indeed answerable from the given choices. MovieChat and LWM exhibited a strong bias towards selecting option A, regardless of the question.

We hypothesize that although these models have the ability to process long videos, they may lack instruction data for such long videos, which leads to them being unable to

effectively follow instructions, in turn affecting their performance.

4.4. Performance across Video Categories

We conducted a comprehensive evaluation across various video categories. We selected two leading models, Seed1.5-VL and Gemini-2.5-Pro, for detailed analysis and compared their results with human performance.

As shown in Table 3, humans achieve a very high accuracy of 94.4% on average, setting a strong benchmark across all categories. In contrast, the overall performance of Gemini-2.5-Pro and Seed1.5-VL was considerably lower, at 67.4% and 64.0%, respectively. This highlights a significant performance gap between current multimodal models and humans in long-video understanding, indicating substantial room for improvement.

Further analysis of the results by category reveals distinct model strengths. Gemini-2.5-Pro performed best on Event Record videos, reaching an accuracy of 72.1%, but struggled with Documentary videos, where its score dropped to 54.3%. Conversely, Seed1.5-VL demonstrated more consistent performance across categories, with all scores falling within a narrow range between 60.3% (Event Records) and 67.5% (Documentaries).

4.5. Ablation on Frame Density

To understand the relationship between temporal information density and model performance, we conduct an ablation study on the number of input frames. We evaluate three state-of-the-art models—Seed1.5-VL, Gemini-2.5-Pro, and Qwen2.5-VL-72B—with varying frame inputs: 0, 1, 4, 8, 50, and a dense sampling of 1 frame per second (FPS). The results are presented in Figure 4.

As depicted in the figure, all models exhibit a clear trend of improved performance with an increased number of input frames. A modest performance gain is observed as the frame count increases from 1 to 8. This suggests that with sparse frames, models can grasp the general context or primary event of the video but may fail to capture the transient, crucial visual cues necessary for accurate reasoning. The most significant finding is the substantial performance leap from 50 frames to the 1 FPS setting. This large gap underscores the necessity of dense visual input for resolving the complex, long-range temporal dependencies inherent in our benchmark tasks.

Furthermore, the "0 Frames" condition serves as a critical control experiment to validate the integrity of our benchmark's question-answer pairs. In this visuals-agnostic setting, models must attempt to answer based solely on linguistic cues from the question and options. The results show that all models perform near the level of random guessing. This outcome effectively demonstrates that success on our benchmark is contingent on visual understanding rather than exploitable language biases. It also validates the efficacy of our LLM-based data filtering protocol, which successfully purges questions that do not strictly require visual grounding.

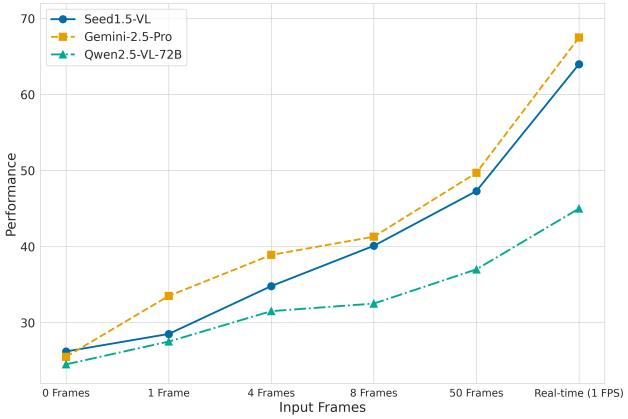

Figure 4. Impact of Frame Count on Model Performance.

5. Discussion

Conclusion. In this paper, we introduced LVBench, a benchmark designed to advance long video understanding. LVBench comprises a diverse collection of lengthy videos and a meticulously annotated question-answer dataset, presenting a robust evaluation framework for assessing multimodal models on complex video understanding tasks. Our experiments revealed that while state-of-the-art models have made strides in short video understanding, their performance on long videos falls short of human-level accuracy. By providing a challenging benchmark, we hope to stimulate the development of advanced models capable of tackling the complexities of extended video comprehension.

Limitations. A limitation of our benchmark is the exclusion of audio data. While audio can provide valuable context, we did not include it because most current models lack effective audio processing capabilities. Future work will aim to incorporate audio information to enhance the evaluation framework for multimodal understanding.

License. Our dataset is under the CC-BY-NC-SA-4.0 license. The dataset relies on YouTube as an external resource for the video content. We only provide download links rather than source videos. While there is no guarantee that the original YouTube videos will remain constant over time, we have taken measures to mitigate this issue by creating and storing copies of the videos. These copies can be provided to dataset consumers, subject to obtaining permission from the original video creators.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62425601 and 62495063, and by the New Cornerstone Science Foundation through the XPLORER PRIZE.

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. 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 2

[3] Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. 1

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 5

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2, 6

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6

[8] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021. 1, 5

[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. 3

[10] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 6

[11] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023. 1

[12] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 2, 6

[13] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. 6

[14] 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. 3

[15] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models, 2023. 3

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[17] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023. 3

[18] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 3

[19] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 6

[20] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 2, 6

[21] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024. 6

[22] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 3

[23] 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. 6

[24] 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. 3

[25] OpenAI. Gpt-4 technical report, 2023. 1

[26] OpenAI. Gpt-4o. 2024. 1, 6

[27] OpenAI. Gpt-4.1. 2025. 6

[28] Ziqi Pang and Yu-Xiong Wang. Mr. video:" mapreduce" is the principle for long video understanding. *arXiv preprint arXiv:2504.16082*, 2025. 6

[29] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023. 3

[30] 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. 1

[31] Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024. 3

[32] 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. 1, 6

[33] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2023. 6

[34] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 2, 3, 6

[35] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1

[36] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 3

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 6

[38] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[39] Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. *arXiv preprint arXiv:2503.12559*, 2025. 6

[40] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024. 3

[41] 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. 3

[42] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 3

[43] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 2, 6

[44] 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. 6

[45] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 3

[46] 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. 1

[47] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 6

[48] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2

[49] Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. Movqa: A benchmark of versatile question-answering for long-form movie understanding. *arXiv preprint arXiv:2312.04817*, 2023. 3

[50] 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. 2, 6