CG-BENCH: CLUE-GROUNDED QUESTION ANSWERING BENCHMARK FOR LONG VIDEO UNDERSTANDING

Guo Chen[1,*], **Yicheng Liu**[1,*], **Yifei Huang**[2,3,*], **Yuping He**[1], **Baoqi Pei**[2,4], **Jilan Xu**[2,5]
Yali Wang[2], **Tong Lu**[1], **Limin Wang**[1,2,†]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Shanghai Artificial Intelligence Laboratory [3]The University of Tokyo
[4]Zhejiang University, [5]Fudan University
chenguo1177@gmail.com

ABSTRACT

The existing video understanding benchmarks for multimodal large language models (MLLMs) mainly focus on short videos. The few benchmarks for long video understanding often rely on multiple-choice questions (MCQs). Due to the limitations of MCQ evaluations and the advanced reasoning abilities of MLLMs, models can often answer correctly by combining short video insights with elimination, without truly understanding the content. To bridge this gap, we introduce CG-Bench, a benchmark for clue-grounded question answering in long videos. CG-Bench emphasizes the model's ability to retrieve relevant clues, enhancing evaluation credibility. It includes 1,219 manually curated videos organized into 14 primary, 171 secondary, and 638 tertiary categories, making it the largest benchmark for long video analysis. The dataset features 12,129 QA pairs in three question types: perception, reasoning, and hallucination. To address the limitations of MCQ-based evaluation, we developed two novel clue-based methods: clue-grounded white box and black box evaluations, assessing whether models generate answers based on accurate video understanding. We evaluated multiple closed-source and open-source MLLMs on CG-Bench. The results show that current models struggle significantly with long videos compared to short ones, and there is a notable gap between open-source and commercial models. We hope CG-Bench will drive the development of more reliable and capable MLLMs for long video comprehension. All annotations and video data are available at https://cg-bench.github.io/leaderboard/.

1 INTRODUCTION

Recently, video understanding has made significant progress with the advent of multimodal large language models (MLLMs). To evaluate these models, many recent efforts have been made to create video understanding benchmarks (Li et al., 2023b; Mangalam et al., 2024; Liu et al., 2024e), providing assessments of model comprehension capabilities and clues for future improvement.

Since early benchmarks only focus on short video clips, recent works have started to create benchmarks (Fu et al., 2024a; Wu et al., 2024b; Zhou et al., 2024; Huang et al., 2024) for longer videos ($\geq$ 10 minutes). However, these works employ multiple-choice questions (MCQ), where the difficulty level is heavily influenced by the configuration of negative options. In such scenarios, models (Chen et al., 2023; Li et al., 2024; Zhang et al., 2024b; Lin et al., 2024) tend to focus on only general video knowledge and use elimination to avoid selecting the negative options. As a result, the models can achieve correct answers without genuinely engaging with the relevant video content, leading to a lack of trustworthiness. One illustration can be found in question 2 of Figure 1, the option 'A' can be easily eliminated based purely on textual information. Recently, the NExT-GQA (Xiao et al., 2024) benchmark tries to address the problem of credible models by incorporating temporal grounding into MCQ. However, NExT-GQA is limited to the NextQA (Xiao et al., 2021) dataset, which lacks di-

*Equal contribution. † Corresponding author.

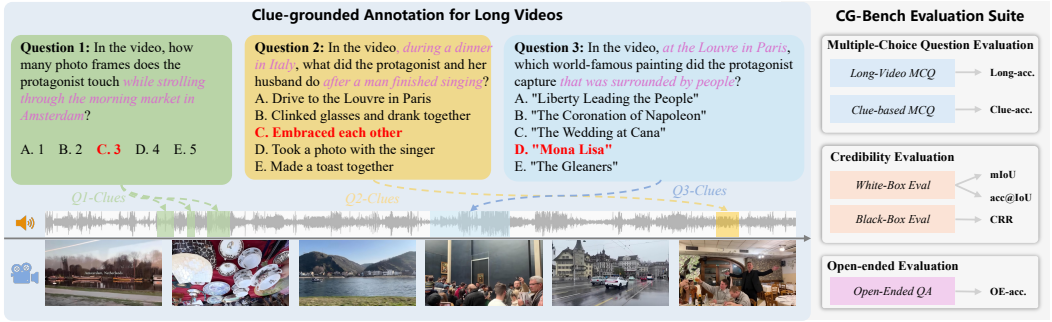


Figure 1: *Left:* examples of CG-Bench's clue-grounded annotation. To correctly answer the questions, models need to ground their reasoning into the correct clue. *Right:* CG-Bench provides an evaluation suite with two novel credibility evaluation criteria while supporting both MCQ and open-ended evaluations.

versity and primarily consists of short videos. A comprehensive benchmark for credibly evaluating *generalist* MLLMs for long video understanding, is still missing in the research community.

To make up this gap, we introduce **CG-Bench**, illustrated in Figure 1, a novel benchmark designed to evaluate clue-grounded question answering in long videos. In contrast to traditional benchmarks that focus primarily on the accuracy of question answering, **CG-Bench** goes a step further by evaluating whether the model bases its answers on relevant clues within the video. **CG-Bench** designs two novel clue-based evaluation methods to provide more reliable model performance assessments. 1) *clue-grouded white box evaluation* requires the model to directly provide the clue interval corresponding to the question while selecting the correct answer. 2) *clue-grouded black box evaluation* requires the model to align the accuracy of video-level and clue-level MCQ. Furthermore, we propose a novel heuristic method, aided by human-annotated clues, for open-ended QA evaluation, to effectively balance the cost and performance.

CG-Bench features 1,219 meticulously curated videos and 12,129 human-annotated question-answer-clue (QAC) triplets, establishing it as the largest and held-out VideoQA and question grounding benchmark for long videos. It employs a highly detailed manual classification system, organizing each video into 14 primary categories, 171 secondary categories, and 638 tertiary categories. The benchmark includes three main question types: perception, reasoning, and hallucination. Perception questions are further divided into 10 subcategories, such as object and attribute recognition, while reasoning questions are categorized into 12 subcategories, including relation reasoning, etc.

We evaluate a range of closed-source and open-source MLLMs using this benchmark. The commercial models, GPT-4o (OpenAI, 2024) and Gemini-1.5 Pro (Anil et al., 2023) achieve scores of 53.9 and 43.4, respectively, with 128 frames for long-video multiple-choice questions. The leading open-source MLLM, Qwen2-VL-72B (Wang et al., 2024a), scores 51.4 under the same conditions, indicating its initial benchmarking against GPT-4o. However, our credibility assessments and open-ended evaluations reveal a significant drop in accuracy for existing MLLMs, with scores decreasing from 53.9 to 21.7. This underscores the considerable room for improvement in current MLLMs for long video understanding. We hope this benchmark can become a vital tool for advancing research and development of more reliable and capable MLLMs.

2 Related Work

Multimodal Large Language Models (MLLMs) have rapidly gained popularity due to their proficiency in integrating visual and textual information (Liu et al., 2024a; 2023; Chen et al., 2023; Wang et al., 2022; 2024c). Recent advancements, such as LLaVA-Next-Video (Zhang et al., 2024b), LLaVA-OneVision (Li et al., 2024), and InternVL2 (Chen et al., 2024d), focus on enhancing MLLMs by integrating LLM backbones with visual encoders and specialized adapters, or creating higher-quality multimodal instruction data. This results in improved performance across tasks that involve both text and images.

Another area of focus is multimodal video understanding. Most models (Chen et al., 2024d; Li et al., 2023a; Maaz et al., 2023; Pei et al., 2024; Huang et al., 2018; 2020) are optimized for short videos, typically a few seconds or at most a few minutes, without exploring their visual understanding

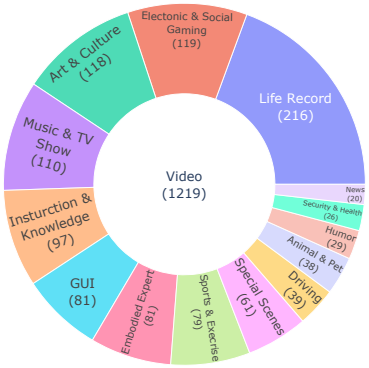


Figure 2: Distribution of video root categories, displaying the number of videos within each category.

Figure 3: Distribution of question root types, illustrating the frequency of different question types.

with longer context. In response, researchers have explored methods such as compressing video frames into fewer visual tokens to allow for the handling of longer videos, as seen in models like LLaMA-Vid (Li et al., 2023c), LVChat (Wang et al., 2024d), MovieChat (Song et al., 2024), MA-LMM (He et al., 2024) and Oryx (Liu et al., 2024f). In addition, LongVA (Zhang et al., 2024a) and LongViLA (Xue et al., 2024) explore the system-level optimization for long-context MLLMs which can natively support long video understanding. Despite the continuous proposal of various MLLMs, their real-world performance in long video understanding is still under explored.

MLLM Benchmarks. The development of benchmarks is becoming increasingly essential, especially for evaluating the MLLM performance in video understanding tasks. As the field develops, various benchmarks have been established to assess MLLMs across different modalities and video lengths. Previous efforts primarily focused on short videos, with traditional specialized VideoQA datasets like TVQA (Lei et al., 2018), NextQA (Xiao et al., 2021), and benchmarks for MLLM like VideoBench (Ning et al., 2023), MVBench (Li et al., 2023b) and EgoSchema (Mangalam et al., 2024). MVBench provides a comprehensive framework for evaluating general temporal understanding capabilities through question-answering on short clips, while EgoSchema focuses on egocentric video understanding with multi-choice questions. The videos in these benchmarks typically range from a few seconds to several tens of seconds, making them similar to image benchmarks and thus hindering the development of general video LLMs.

Recently, several works such as VideoMME (Fu et al., 2024a), CinePile (Rawal et al., 2024), MLVU (Zhou et al., 2024), LongVideoBench (Wu et al., 2024b), MoVQA (Zhang et al., 2023b), HourVideo (Chandrasegaran et al., 2024), and LVBench (Wang et al., 2024b), have introduced long video benchmarks to evaluate MLLMs. VideoMME constructs a diverse video MCQ dataset, incorporating multimodal evaluations with visuals, subtitles, and audio. MLVU designs a range of tasks that focus on granular detail understanding to assess long video comprehension capabilities. However, a common limitation of these benchmarks is their reliance on MCQs, where the difficulty is heavily influenced by the construction of negative options. This allows MLLMs to often eliminate incorrect answers using sparse frames and common sense reasoning, which can inflate performances. With our clue interval annotation, CG-Bench enhances the evaluation quality of MLLMs in long video understanding by introducing new evaluation mechanisms on credibility.

3 CG-Bench

3.1 Dataset Construction

The dataset construction process of CG-Bench consists of three steps: video collection, question-answering-clue annotation, and quality review iteration. We provide details as follows.

Video Collection. To avoid using videos that have been used for pre-training by existing MLLMs, we manually collect videos from the internet and provide new annotations on them. To facilitate the collection of raw videos from the Internet, we define 14 root domains as listed in Figure 2. During the collection process, we manually assign a brief tag (4-8 words) to categorize the content of each video. This supplementary tagging helps to ensure the diversity of the videos. We define

a video to be long if it exceeds 10 minutes in duration. Accordingly, we collected videos longer than 10 minutes while considering the distribution of video duration. Furthermore, we retain the accompanying subtitles and audio to provide multimodal information. We carefully review and filter the videos manually for 7 rounds. More details about the video collection can be found in the supplementary material.

Question-Answer-Clues Annotation. After collecting the raw video data, we annotate it with high-quality question-answer-clue (QAC) triplets. To ensure question diversity, we establish a taxonomy with three types: Perception, Reasoning, and Hallucination. As shown in Figure 3, Perception and Reasoning questions are further divided into 10 and 14 subcategories, respectively, while Hallucination questions combine elements of both. Annotators are instructed to include negative options to create a multiple-choice QA format, facilitating straightforward and cost-effective assessments. To minimize expression loss, annotators use their native language during the annotation process. Each video is annotated with 6 to 15 QAC triplets, depending on its duration. To ensure consistency in QAC triplets, we standardized the annotation process by first annotating the QA pairs and then identifying the clues. Annotators must watch the entire video, select a question type from the predefined categories, and then annotate a new question and its corresponding answer. Next, they select one or more intervals from the video to form a QAC triplet. Since the actual clue intervals often consist of multiple short moments, annotating each fragment is costly. Therefore, annotators are required to mark intervals that cover these short moments while ensuring the completeness of each event.

Review Iteration. To ensure the difficulty and quality of the dataset, we conduct a repetitive review and iteration process to enhance annotation quality. We reject annotations that do not meet our quality standards and request annotators to revise them. Our quality requirements for annotations and the measures taken to ensure them are as follows: 1) *The rationality of the question, options, and answer*: we conduct manual reviews; 2) *The video dependency of the question, options, and answer*: we input questions and options into GPT-4 and filter out QA pairs that can be answered solely based on pure text; 3) *The difficulty of negative options in multiple-choice questions*: we input the video, questions and options into MLLMs and filter out QA pairs that can be answered using only sparse frames and small models; 4) *The positional diversity of clue intervals:* We monitor the distribution of clue duration and position and provide timely guidance to annotators.

3.2 Dataset Statistics & Comparisons

We present the detailed statistics of our dataset to provide a more comprehensive understanding, including meta-information, QAC triplets, qualitative analysis, and comparison to previous works.

3.2.1 Dataset Statistics

Video Meta. Our dataset comprises a total of 1219 videos with multimodal information, including vision, audio, and subtitles. The duration of the videos varies between 10 and 80 minutes, with a distribution illustrated in Figure 4. Notably, videos that last between 20 and 30 minutes are the most prevalent. This selection process is manual, based on content relevance, which mirrors real-world duration distributions and highlights a long-tail effect for longer videos. As illustrated in Figure 2, each video is classified using a three-tiered tagging system that succinctly encapsulates its content and assigns it to fundamental categories. The primary classification is augmented by a secondary layer of 171 tags and a tertiary layer consisting of 638 tags. This multi-level tagging mechanism guarantees a broad diversity of data content. For a more detailed classification of tags, please consult the supplementary materials.

QAC Annotation. CG-Bench includes 12,129 annotations consisting of questions, answers, and clues. Table 1 presents the sentence lengths and totals for the annotated questions and answers, highlighting the linguistic diversity within our dataset. Each QAC triplet is annotated with 4 to 7 negative samples, resulting in an approximately uniform distribution with ratios of options A to H of 12.4%, 14.7%, 12.1%, 14.8%, 15.1%, 16.1%, 11.6%, and 3.1%. There are a total of 14,362 clue intervals across all QAC triplets, with an average duration of 19.24 seconds each. Additionally, we conduct a further analysis of the positions of clue intervals within the video. Figure 5 illustrates the frequency with which each normalized timestamp is represented by intervals. This demonstrates the unbiased nature of our interval annotations and highlights the diversity of our QA content in temporal position.

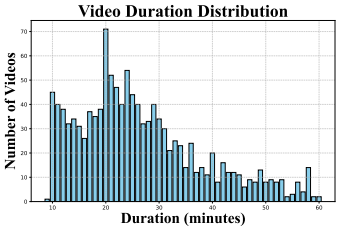


Figure 4: Video duration distribution, showing the number of videos for different duration intervals.

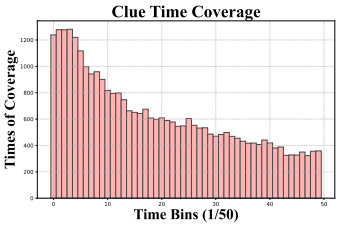


Figure 5: Clue time coverage, illustrating the frequency of clues across different time bins.

Annotation Statistics	
#QAC Triplets	12129
#Avg/QAC per video	9.95
#Avg/Option per QAC	6.96
#Avg/Clue per QAC	1.18
#Avg/Words of Questions	20.07
#Avg/Words of Options	22.88
#Avg/Duration of Clues	19.24

Table 1: Annotation statistics, detailing the number of QAC triplets, questions, options, and clues.

Table 2: Comparison of benchmarks across key aspects: number of videos (#Video), average duration (#Duration), number of QA pairs (#QA Pairs), number of clues (#Clue), annotation method (M/A for manual/automatic), Open-Domain (OD), Open-Ended (OE), Multi-modal (MME), and Credibility (CE) Evaluation.

Benchmark	**#Video**	**#Dur.(s)**	**#QA Pairs**	**#Clue**	**Anno.**	**OD**	**OE**	**MME**	**CE**
Question-Clue Grounding									
NextGQA (Xiao et al., 2024)	1,000	39.5	-	10,531	M	✗	-	-	-
Ego4D-NLQ$_{\texttt{val}}$ (Grauman et al., 2022)	415	499.7	-	4,554	M	✗	-	-	-
Ego4D-NLQ$_{\texttt{test}}$ (Grauman et al., 2022)	333	493.7	-	4,005	M	✗	-	-	-
MultiHop-EgoQA$_{\texttt{test}}$ (Chen et al., 2024c)	360	-	-	1,080	A&M	✗	-	-	-
E.T. Bench$_{\texttt{test}}$ (Liu et al., 2024d)	-	129.3	-	2,011	M	✓	-	-	-
RexTime$_{\texttt{test}}$ (Chen et al., 2024a)	-	141.1	-	2,143	A&M	✗	-	-	-
CG-Bench-QG	1,219	1624.4	-	14,362	M	✓	-	-	-
Short-Video QA									
TVQA (Lei et al., 2018)	2,179	11.2	15,253	15,253	M	✗	✗	✗	✗
STAR (Wu et al., 2024a)	914	11.9	7,098	7,098	A	✗	✗	✗	✗
NextQA (Xiao et al., 2021)	1,000	44.0	8,564	✗	A	✗	✓	✗	✗
EgoSchema (Mangalam et al., 2024)	5,063	180.0	5,063	✗	A&M	✗	✗	✗	✗
TempCompass (Liu et al., 2024e)	410	11.4	7,540	✗	A&M	✗	✗	✗	✗
RexTime$_{\texttt{test}}$ (Chen et al., 2024a)	-	141.1	-	2,143	A&M	✗	✗	✗	✓
MVBench (Li et al., 2023b)	3,641	16.0	4,000	✗	A&M	✗	✗	✗	✗
MMBench-Video (Fang et al., 2024)	600	165.4	1,998	✗	M	✓	✓	✗	✗
CG-Bench-Clue	12,129	22.8	12,129	-	M	✓	-	✓	-
Long-Video QA									
EgoTimeQA$_{\texttt{test}}$ (Di & Xie, 2024)	148	492	500	✗	A	✗	✗	✗	✗
MovieChat-1K (Song et al., 2024)	130	500.0	1,950	✗	M	✗	✗	✗	✗
Video-MME (Fu et al., 2024a)	900	1017.9	2,700	✗	M	✓	✗	✓	✗
LongVideoBench (Wu et al., 2024b)	966	1408.0	6,678	✗	M	✓	✗	✗	✗
MLVU (Zhou et al., 2024)	757	720.0	2,593	✗	M	✗	✗	✗	✗
CG-Bench	1,219	1624.4	12,129	14,362	M	✓	✓	✓	✓

3.2.2 Comparison with Previous Benchmarks

CG-Bench is characterized by its diverse features, allowing it to be compared with three distinct types of benchmarks, as depicted in the three sections of Table 2: Question Clue Grounding, Short-Video QA, and Long-Video QA benchmarks. For the question clue grounding benchmarks, NextGQA (Xiao et al., 2024), Ego4D-NLQ (Grauman et al., 2022), MultiHop-EgoQA (Chen et al., 2024c), E.T. Bench (Liu et al., 2024d), and RexTime (Chen et al., 2024a) are primarily centered around action and egocentric domains. Their videos are sampled from academic datasets. In comparison, the question clue grounding part of CG-Bench, CG-Bench-QG, stands out with the highest number of videos and the longest average length, the diversity of which fosters a broad spectrum of question-grounding queries.

Furthermore, we transform QAC triplets to our novel Short-Video QA benchmark, termed CG-Bench-Clue. When contrasted with prior short video benchmarks such as TempCompass (Liu et al., 2024e), MVBench (Li et al., 2023b) and MMBench-Video (Fang et al., 2024), our CG-Bench-Clue emerges as the ***largest***, ***held-out***, ***open-domain*** and ***multimodal*** Short-Video QA benchmark.

As for the Long-Video QA benchmark, CG-Bench excels in the number of videos, length, quantity of questions, and annotation quality. Owing to our clue interval annotations, CG-Bench further facilitates reliable evaluations for long videos and open-ended evaluations with clue assistance, a feature that sets it apart from existing long video benchmarks like Video-MME (Fu et al., 2024a) and MLVU (Zhou et al., 2024).

3.3 EVALUATION

In this section, we describe the evaluation tasks of our CG-Bench which include traditional MCQ, the unique credibility evaluation, and clue-aided open-ended QA evaluation.

3.3.1 MULTIPLE-CHOICE QUESTION EVALUATION

We assess the accuracy of MCQ in two settings: **Long-Video MCQ** and **Clue-based MCQ**. In the Long-Video MCQ setting, the model receives the entire video as input and is required to select the correct answer based on the video, the question, and the candidate options. For the Clue-based MCQ setting, the model is given only the video within the annotated clue interval as input. The model has access only to the clue clip, the question, and the candidate options. It does not have access to the original long video. Since a single QA may correspond to multiple clues, we merge these clues and treat the combined clue as a single, cohesive clue segment.

3.3.2 CREDIBILITY EVALUATION

The ability of a model to identify relevant clues related to questions is a crucial factor in determining its reliability. Therefore, we define a model's reliability based on its proficiency in locating accurate clues when addressing problems. To achieve this, we introduce two clue-grounded mechanisms for credibility assessment: white-box evaluation and black-box evaluation.

White-Box Evaluation requires the model to directly output the intervals of clues that can accurately answer the question. This task is similar to video temporal grounding (Lei et al., 2021; Huang et al., 2023). Therefore, we use tIoU (Temporal Intersection over Union) as the evaluation metric. Since each question may correspond to multiple intervals of clues, we allow the model to predict multiple possible intervals. Given a set of prediction $\mathcal{P}$ and ground truths $\mathcal{G}$, the tIoU is defined as:

$$\text{tIoU} = \frac{\sum_{i \in \mathcal{G}, j \in \mathcal{P}} \max(0, \min(b_i, d_j) - \max(a_i, c_j))}{\sum_{i \in \mathcal{G}} (b_i - a_i) + \sum_{j \in \mathcal{P}} (d_j - c_j) - \sum_{i \in \mathcal{G}, j \in \mathcal{P}} \max(0, \min(b_i, d_j) - \max(a_i, c_j))} \times 100\%, \quad (1)$$

where a_i, b_i are the start and end timestamps of the i-th ground truth interval of $\mathcal{G}$. c_j, d_j are the start and end timestamps of the j-th predicted interval of $\mathcal{P}$. We calculate the mean IoU (**mIoU**) by averaging the tIoU scores obtained by the model across all question queries. To further improve the robustness of question grounding evaluation, we introduce the **rec.@IoU** metric. This metric measures the probability of successfully recalling clue intervals at various IoU thresholds.

Additionally, we propose **acc.@IoU** that evaluates both MCQ accuracy and clue-grounding ability. A response is considered correct only if the selected answer is accurate and the tIoU exceeds ($>$) a predefined threshold τ. Since locating short-duration clues in the long videos in CG-Bench is inherently challenging, we set the default τ to be 0 for the more obvious comparison on ablation studies. Setting $\tau = 0$ ensures that acc.@IoU requires the model to select the correct option and produce a time interval that overlaps at least slightly (tIoU > 0) with the annotated clue interval, rather than reducing to naive MCQ accuracy. We calculate the **rec.@IoU** and **acc.@IoU** at IoU thresholds of 0.1, 0.2, 0.3, 0.4, and 0.5 to determine the final result.

Black-Box Evaluation aims to evaluate the model's ability to seek out clues implicitly. Understanding long videos involves the retrieval of clues distributed across various spatiotemporal locations within the entire video. Therefore, an effective model for long videos should naturally focus on capturing human-annotated clue intervals in its hidden states. However, beyond the explicitly annotated clue intervals, there are likely hidden clues scattered throughout the video that can also help to determine the correct answer. Thus, a model with access to the full video should yield higher accuracy compared to solely relying on the clue interval. In other words, the accuracy of Long-Video MCQ (**long-acc.**) should be greater than or equal to the accuracy of Clue-based MCQ (**clue-acc.**).

With this insight, for the black box evaluation, we define a new metric called Clue Recovery Rate (**CRR**). This metric evaluates the model's robustness to context dilution, *i.e.*, how stable a model can find related clues from long but diluted video context. CRR is calculated by:

$$\text{CRR} = \frac{\min(\textbf{long-acc.}, \textbf{clue-acc.})}{\textbf{clue-acc.}} \times 100\%, \quad (2)$$

A CRR of less than 100% suggests that the MLLM's ability to retrieve short clues from long video representations is not optimal.

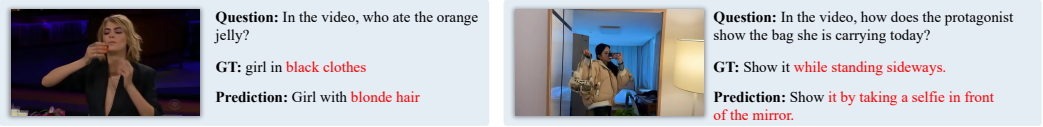


Figure 6: Two examples illustrating the ambiguity challenge of using LLMs for open-ended evaluation. While in different expressions, GT and prediction should both be treated as correct answers.

3.3.3 CLUE-AIDED OPEN-ENDED QA EVALUATION

CG-Bench supports open-ended QA evaluation for a more comprehensive assessment. Previous works like MM-Vet (Yu et al., 2023) and MMBench-Video (Fang et al., 2024) have used LLMs to evaluate open-ended QA for images and short videos. However, long videos contain more complex information, leading to ambiguous user-generated questions. This can result in discrepancies between LLM-evaluated scores and the actual QA ability of models, as shown in Figure 6.

To address this, we use a low-hallucination MLLM to evaluate the similarity between text output and visual information. We selected GPT-4o (OpenAI, 2024) as the multimodal evaluator due to its high ranking in benchmarks like OpenCompass (Contributors, 2023) and the Lmsys leaderboard (Chiang et al., 2024), and its relatively low hallucination rate compared to other MLLMs. Since using GPT-4o directly for multimodal judging can still introduce hallucination errors and incur high API costs, we propose a heuristic evaluation method to mitigate biases and reduce costs. First, GPT-4o determines if the output can be evaluated based solely on the text answer. It outputs `yes` or `no`; if not, it requests visual cues by stating, "I need visual clues." This prompts the inclusion of supplementary visual data to aid GPT-4o in its evaluation. By using pre-annotated time intervals with question clues, we sample frames as visual aids, further reducing hallucination errors and costs. We analyze this evaluation method in Sec 4.3, with more details available in the supplementary materials.

4 EXPERIMENTS

In this section, we evaluate a wide range of MLLMs using CG-Bench. We first introduce the evaluation setup, followed by quantitative results for both closed-source and open-source models. Finally, we analyze some key factors in the evaluation.

4.1 SETTINGS

We first briefly describe the settings used in our experiments. The supplementary material provides more detailed settings.

Models. We evaluate the performance of three mainstream commercial models on our CG-Bench: GPT4o (OpenAI, 2024), Gemini-1.5 (Anil et al., 2023), and Claude-3.5, including their different versions. Also, we assess the representative open-source video models such as LLaVA-OneVision (Li et al., 2024), Qwen2-VL (Wang et al., 2024a) and InternVL2 (Chen et al., 2024d), among others.

Frame Sampling. For long video understanding, the frame sampling strategy significantly impacts evaluation results. For open-source MLLMs, we make the best use of our computational resources to use as many frames as possible. For closed-source MLLMs, since the local computational resource is no longer a bottleneck, we can use even more frames. We uniformly sample 128 frames for Long-video MCQ, and use 32 frames as the for Clue-based MCQ.

Modality. We also explore other modalities: subtitles and audio. For subtitles, we employ a uniform sampling method. If the timestamp of a sampled frame falls within the time interval of a subtitle, that subtitle will be included in the analysis. Each subtitle is considered only once to avoid redundancy.

Prompt. For MCQ tasks, the model is prompted to provide the uppercase letter corresponding to the correct option. In Open-Ended QA tasks, the model responds freely based on the questions. For the Clue Grounding task, we append the timestamps of each frame and subtitle to enhance the model's time-awareness, requiring it to return nested lists in the format `[[s1, e1], [s2, e2], ...]`. For open-ended evaluation, we require the model to assess the correctness between the predictions and the ground truth and respond with `yes` or `no`.

Table 3: Performance of various open-source and closed-source MLLMs on CG-Bench. We provide human evaluation for showing annotation agreements and the difficulty of our benchmark.

Models	**LLM**	**#F**		**MCQ**		**Cred. Eval.**				**OE**
	#param	**clue**	**long**	**clue-acc.**	**long-acc.**	**mIoU**	**rec.@IoU**	**acc.@IoU**	**CRR**	**acc.**
Random	-	-	-	14.2	14.2	0.13	0.16	0.09	100	0
Human (full-video)	-	-	-	92.2	90.3	35.5	51.2	29.8	97.9	83.7
Human (sparse frames)	-	-	128	-	59.9	-	-	-	-	-
GPT4o (text)	-	0	0	16.8	16.8	0.14	0.2	0.15	100	2.1
Open-source MLLMs										
Video-LLAVA (Lin et al., 2023)	7B	8	8	34.2	16.2	1.13	1.96	0.59	47.4	12.3
VideoLLAMA (Zhang et al., 2023a)	7B	32	32	36.8	18.4	1.21	1.87	0.84	50.0	15.8
Videochat2 (Li et al., 2023b)	7B	16	16	35.2	19.3	1.28	1.98	0.94	54.8	18.6
Qwen-VL-Chat (Bai et al., 2023)	7B	4	4	38.3	21.6	0.89	1.19	0.42	56.4	19.4
ST-LLM (Liu et al., 2024c)	7B	32	64	39.6	23.8	2.23	2.86	1.13	60.1	20.7
ShareGPT4Video (Chen et al., 2024b)	16B	16	16	41.4	26.7	1.85	2.65	1.01	64.5	22.0
Chat-UniVi-v1.5 (Jin et al., 2024)	13B	32	64	41.5	25.9	2.07	2.53	1.21	62.4	21.4
ViLA (Lin et al., 2024)	8B	14	14	41.8	28.7	1.56	2.89	1.35	68.7	24.0
GroundVQA (Liu et al., 2024d)	0.25B	-	1200	27.3	-	1.33	1.37	-	-	-
GeLM (Chen et al., 2024c)	7B	-	100	-	-	2.25	2.81	-	-	-
ET-Chat (Liu et al., 2024d)	4B	-	1fps	17.6	-	1.38	1.43	-	-	-
InternVL-Chat-v1.5 (Chen et al., 2023)	20B	10	10	42.5	28.9	2.18	2.38	1.15	68.0	23.1
MiniCPM-v2.6 (Yao et al., 2024)	8B	32	32	44.6	30.1	2.35	2.61	1.04	67.5	26.6
LongVA (Zhang et al., 2024a)	7B	32	128	42.8	28.7	2.94	3.86	1.78	67.1	25.1
LLaVA-OneVision (Li et al., 2024)	7B	16	16	43.2	31.1	1.63	1.78	1.08	72.0	25.4
Video-CCAM (Fei et al., 2024)	14B	32	96	43.6	29.7	2.63	3.48	1.83	68.1	25.3
Kangaroo (Liu et al., 2024b)	8B	32	64	45.9	30.2	2.56	2.81	1.94	65.8	24.5
VITA (Fu et al., 2024b)	8x7B	32	32	47.8	33.3	3.06	3.53	2.06	69.7	27.5
Qwen2-VL (Wang et al., 2024a)	72B	32	128	56.2	41.3	3.58	5.32	3.31	73.5	33.6
InternVL2 (Chen et al., 2024d)	78B	32	32	58.5	42.2	3.91	5.05	2.64	72.1	32.5
Closed-source MLLMs										
GPT-4o-08-06 (OpenAI, 2024)	-	32	128	58.3	45.2	5.62	8.30	4.38	77.5	39.5
GPT-4mini-08-06 (OpenAI, 2024)	-	32	128	48.3	33.4	3.75	5.18	2.21	69.2	25.4
Gemini-1.5-Pro (Anil et al., 2023)	-	32	128	50.1	37.2	3.95	5.81	2.53	74.3	29.3
Gemini-1.5-Flash (Anil et al., 2023)	-	32	128	47.0	32.3	3.67	5.44	2.45	68.7	26.3
Claude3.5-Sonnet	-	32	50	56.2	40.5	3.99	5.67	2.79	72.1	35.2

4.2 Main Results

As shown in Table 3, the closed-source MLLM GPT4o (OpenAI, 2024) leads significantly, surpassing other models across all metrics. Notably, GPT4o's long-acc. reaches 45.2%, much higher than Gemini-1.5-Pro (Anil et al., 2023), demonstrating its strong capabilities in long video understanding. Among open-source MLLMs, Qwen2-VL (Wang et al., 2024a) performs impressively, achieving results comparable to GPT4o in long-acc. and clue-acc. Other models underperform due to insufficient context support or inadequate video training. While these MLLMs perform well on MCQ tasks, they experience significant performance drops in credibility and open-ended evaluations on CG-Bench. For instance, GPT-4o's long-acc. falls from 45.2 to 4.38 in Acc@IoU and 39.5 in OE-acc. Additionally, with the same number of sampling frames, GPT-4o achieves a CRR of 77.5, whereas Gemini-1.5-Pro only reaches 74.3, indicating its weaker ability to retrieve short-term clues from long videos. Overall, current MLLMs do not perform well on CG-Bench, suggesting considerable room for improvement in their capability and credibility.

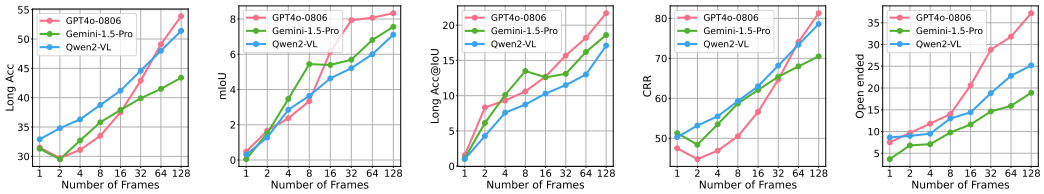
Since it is difficult to input more than 128 frames due to the hardware limitations, we alternatively conducted a human evaluation experiment under constrained visual conditions, to see how severe the "undersampling" issue is for longer video. We uniformly sampled 30 videos from CG-Bench, resulting in 296 questions. For each video, we uniformly sampled 128 frames and asked volunteers to perform an MCQ testing. The resulting accuracy was 59.85% (row 3 in Table 3). This result indicates that our dataset is indeed challenging and that it is difficult to derive solutions from a limited number of frames. It also highlights that even the most advanced models, such as GPT-4o, have ample room for improvement in long video comprehension.

4.3 Analysis

Furthermore, we perform a comprehensive analysis of the two leading closed-source MLLMs, GPT4o (OpenAI, 2024) and Gemini-1.5 Pro (Anil et al., 2023), as well as the best performing open-source MLLM, Qwen2-VL (Wang et al., 2024a), on our CG-Bench. In this analysis, we use 1000

Table 4: Impact of different prompts and modalities. Each prompt can be composed of frames (F), frame timestamps (FT), subtitles (S), subtitle timestamps (ST), and audio (A). We conduct the main experiments with GPT4o-0806 (OpenAI, 2024) while studying the audio modality with Gemini-1.5 Pro (Anil et al., 2023).

model	prompt & modality	clue-acc.	long-acc.	mIoU	Acc@IoU	CRR	OE-acc.
GPT4o	S (128 frames)	-	31.5	-	-	-	-
GPT4o	S (full-video)	-	34.3	-	-	-	-
GPT4o	F	65.8	51.8	3.39	10.7	78.7	35.4
GPT4o	F+FT	$65.3_{(-0.5)}$	$51.6_{(-0.2)}$	$5.73_{(+2.34)}$	$20.4_{(+9.7)}$	$79.0_{(-0.3)}$	$36.8_{(+1.4)}$
GPT4o	F+S	$66.7_{(+0.9)}$	$53.4_{(+1.6)}$	$3.96_{(+0.57)}$	$11.2_{(+0.5)}$	$80.1_{(+1.4)}$	$38.2_{(+2.8)}$
GPT4o	F+S+ST	$67.1_{(+1.3)}$	$54.1_{(+2.3)}$	$5.19_{(+1.80)}$	$13.2_{(+2.5)}$	$80.6_{(+1.9)}$	$38.4_{(+3.0)}$
GPT4o	F+S+FT	$67.4_{(+1.6)}$	$53.2_{(+1.4)}$	$7.80_{(+4.41)}$	$22.3_{(+11.6)}$	$78.9_{(+0.2)}$	$37.9_{(+2.5)}$
GPT4o	F+S+ST+FT	**67.5**$_{(+1.7)}$	**54.9**$_{(+3.0)}$	**9.68**$_{(+6.29)}$	**26.7**$_{(+16.0)}$	**81.3**$_{(+2.6)}$	**39.5**$_{(+4.1)}$
Gemini	F+S+ST+FT	62.1	45.1	9.16	20.7	72.6	23.2
Gemini	F+S+ST+FT+A	$62.3_{(+0.2)}$	$45.0_{(-0.1)}$	$9.10_{(-0.06)}$	$19.8_{(-0.9)}$	$72.2_{(-0.4)}$	$23.5_{(+0.3)}$



(a) Long Acc (b) mIoU (c) Acc@IoU (d) CRR (e) Open ended

Figure 7: Impact of sampling frame numbers on different metrics for GPT-4o-0806 (OpenAI, 2024), Gemini-1.5 Pro (Anil et al., 2023) and Qwen2VL-72B (Wang et al., 2024a).

QAC triplets sampled uniformly from all annotations for fast experiments. We report acc.@IoU with $\tau = 0$ for a more obvious comparison.

Impact of Prompt & Modality. As shown in Table 4, we conduct the ablation studies on the subset that contains subtitles and explore the impact of different prompts on GPT4o and the effect of the audio modality on Gemini-1.5 Pro. Our findings indicate that all prompt types (FT/S/ST), except video frames (F), provide performance benefits across most metrics. Subtitles contribute more to **long-acc.** than they do to **clue-acc.**. Additionally, the inclusion of timestamp information (FT/ST) is critical for interval prediction. Timestamps from both frames and subtitles enhance IoU-related metrics, revealing a complementary effect. When both FT and ST are added simultaneously, **mIoU** increases from 3.39 to 9.68, and **Acc@IoU** rises from 10.7 to 26.7. When S, FT, and ST are all used in the prompt, the model achieves the best performance across all metrics. In contrast, our exploration of the audio modality (A) revealed that audio does not yield significant performance gain and, in some cases, even slightly degrades the results, as shown in Table 4. Finally, we conduct experiments using only subtitles from 128 frames versus the full video. The results show that while subtitles offer useful semantic cues, their impact is significantly reduced when visual input is included. This suggests that our benchmark favors visual signals.

Impact of Frame Number. As illustrated in Figure 7, we conducted experiments to analyze the performance across various metrics as the number of frames increases. Overall, the performance of all three MLLMs gradually improves with the addition of more frames, with GPT-4o consistently outperforming the others across all metrics. For **long-acc.** and **OE acc.**, Qwen2VL achieves performance comparable to GPT-4o. However, compared with Qwen2VL, Gemini excels in terms of mIoU and Acc@IoU. Regarding **CRR**, GPT-4o demonstrates greater consistency between clue-acc. and long-acc. across more frames, indicating its superior reliability in long video understanding. For open-ended QA, Gemini's higher refusal rate results in a noticeable decline in performance.

Open-ended Evaluation Quality. To assess the stability and accuracy of various MLLMs as evaluators, we utilized four models—Gemini, Qwen2VL, Claude, and GPT-4o—each of which evaluated GPT-4o's predictions five times. Human evaluations of GPT-4o's predictions are also conducted for reference. The results, shown in Figure 8, indicate that GPT-4o has the highest stability and the smallest deviation from human-assigned scores. Furthermore, Table 5 explores the impact of different evaluation methods. When evaluators were provided only with ground truth (col. "GT") or visual information (col. "Vis"), the scoring bias (absolute difference) between human and model-

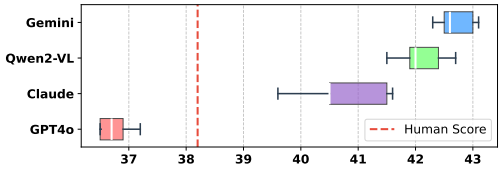


Figure 8: Comparison of using different LLMs as open-ended evaluators for GPT-4o's outputs.

	GT	GT+Vis	Vis	**Ours**
Bias(%)↓	12.4	6.4	17.0	**1.0**
Time (s)↓	**741**	20,040	19,640	3,600
Price ($)↓	**0.05**	6.1	6	2
Trigger Rate (%)↓	0	100	100	**14**
Trigger Recall Rate (%)↑	0	100	100	**88**

Table 5: Comparison of different modes: GT-only, visual-only, GT+vision and heuristic (Ours).

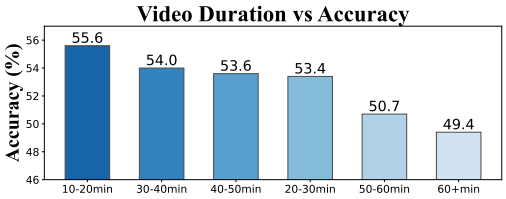


Figure 9: Long-Video-MCQ Accuracy grouped by video duration for GPT4o-0806 with 128 frames.

#Frames	Resolution	Sampling Strategy	long-acc.
128	Low	Uniform	**53.9**
50	Low	Uniform	46.7
50	Low	Keyframe	45.7
50	High	Uniform	51.0

Table 6: Impact of different frame sampling strategies on long-acc. for GPT4o-0806.

based evaluation increased. While fully leveraging visual information (col. "GT+Vis") improved evaluation accuracy, it also significantly increased the time and cost required. Our proposed heuristic evaluation method achieves the lowest evaluation bias. Additionally, we manually annotated 200 evaluation samples to determine the necessity of visual request triggers. From the bottom block in Table 5, the statistics show that our method achieved a visual request trigger rate (the probability that the model triggers "visual clues required") of 14%. The recall rate of this triggering achieves 88%. This proves that our approach effectively balances cost and performance.

Performance grouped by Video Duration. We grouped videos by duration and evaluated the **long-acc.** performance of GPT-4o-0806 using 128 frames. Figure 9 shows that the model struggles with undersampling, especially for longer videos.

Impact of Frame Sampling Strategy. We investigate how different frame sampling strategies affect performance. To expedite testing, we primarily evaluated GPT4o-0806 using 50 uniformly sampled frames, focusing on the **long-acc** metric. The experiment consists of three parts: 1) low resolution, 2) high resolution, and 3) keyframe extraction (via FFmpeg) combined with low resolution. As shown in Table 6, higher resolution offers some improvement, while keyframe extraction has no significant impact.

5 Conclusion and Future Work

In this paper, we introduce CG-Bench, a novel benchmark designed to evaluate clue-grounded question answering capabilities in long video understanding. Unlike existing benchmarks that focus on short videos or rely solely on multiple-choice questions, CG-Bench emphasizes the importance of models retrieving and grounding their answers in specific video segments, enhancing evaluation credibility. CG-Bench includes 1,219 manually curated videos organized into a detailed three-tier system, with 12,129 QA pairs covering perception, reasoning, and hallucination question types. This provides a comprehensive and diverse dataset for assessing MLLMs. We propose two clue-based evaluation methods—clue-grounded white-box and black-box evaluations—that offer novel ways to determine whether models genuinely comprehend video content or merely rely on superficial cues. Extensive experiments with various closed-source and open-source MLLMs reveal that current models significantly underperform in long video understanding compared to short videos. We hope that CG-Bench will serve as a valuable resource for the research community, driving the development of more trustworthy and capable MLLMs for long video understanding.

Acknowledgement. This work is supported by the National Key R&D Program of China (No. 2022ZD 0160900), the National Natural Science Foundation of China (No. 62372223 and U24A20330), Jiangsu Frontier Technology Research and Development Program (No. BF2024076), and Nanjing University-China Mobile Communications Group Co., Ltd. Joint Institute.

REFERENCES

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. Rextime: A benchmark suite for reasoning-across-time in videos. *arXiv preprint arXiv:2406.19392*, 2024a.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024b.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

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*, 2024d.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`, 2023.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12934–12943, 2024.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, abs/2405.21075, 2024a.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024b.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. pp. 18995–19012, 2022.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13504–13514, 2024.

Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 754–769, 2018.

Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14024–14034, 2020.

Yifei Huang, Lijin Yang, and Yoichi Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18908–18918, 2023.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22072–22086, 2024.

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

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. pp. 11846–11858, 2021.

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.

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*, 2023a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. *CoRR*, abs/2311.17005, 2023b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *CoRR*, abs/2311.17043, 2023c.

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.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 36, 2024a.

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*, 2024b.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. ST-LLM: large language models are effective temporal learners. *CoRR*, abs/2404.00308, 2024c.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Et bench: Towards open-ended event-level video-language understanding. *arXiv preprint arXiv:2409.18111*, 2024d.

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

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024f.

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

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023.

OpenAI. Hello gpt-4o. 2024. URL `https://openai.com/index/hello-gpt-4o/`.

Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, et al. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024.

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.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024b.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *CoRR*, abs/2212.03191, 2022.

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. *arXiv preprint arXiv:2403.15377*, 2024c.

Yu Wang, Zeyuan Zhang, Julian McAuley, and Zexue He. Lvchat: Facilitating long video comprehension. *arXiv preprint arXiv:2402.12079*, 2024d.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. *CoRR*, abs/2405.09711, 2024a.

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

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. pp. 9777–9786, June 2021.

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

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

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pp. 543–553. Association for Computational Linguistics, 2023a.

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*, 2023b.

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*, 2024a. URL `https://arxiv.org/abs/2406.16852`.

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

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.