# LONGVIDEOBENCH: A Benchmark for Long-context Interleaved Video-Language Understanding

**Haoning Wu   Dongxu Li   Bei Chen   Junnan Li**

https://longvideobench.github.io.

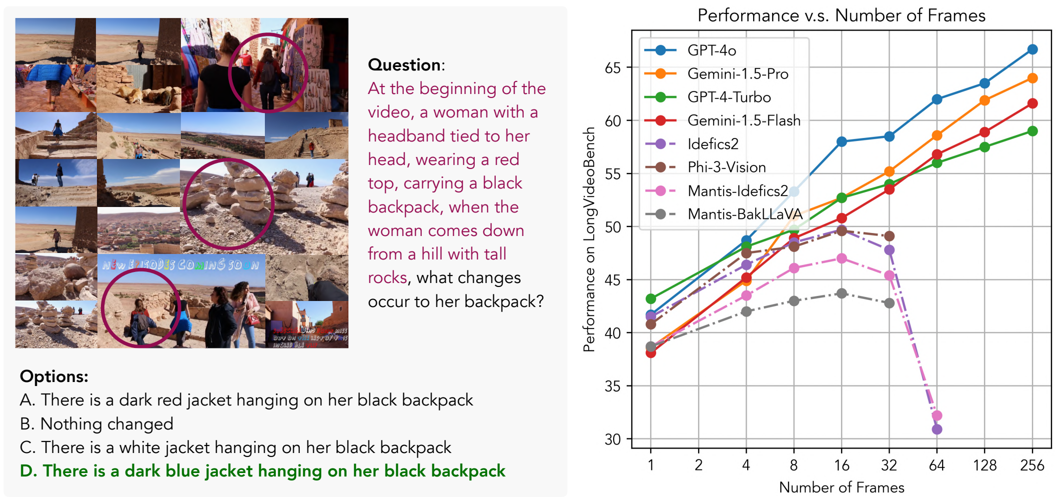



Figure 1: **(Left)** LONGVIDEOBENCH features *referring reasoning* questions, with a *referring query* that references particular video contexts (*i.e. referred context*) to answer questions about. **(Right)** Proprietary models perform better with more frames while open-source models cannot properly scale.

## Abstract

Large multimodal models (LMMs) are processing increasingly longer and richer inputs. Albeit the progress, few public benchmark is available to measure such development. To mitigate this gap, we introduce **LONGVIDEOBENCH**, a question-answering benchmark that features video-language interleaved inputs up to an hour long. Our benchmark includes 3,763 varying-length web-collected videos with their subtitles across diverse themes, designed to comprehensively evaluate LMMs on long-term multimodal understanding. To achieve this, we interpret the primary challenge as to accurately *retrieve* and *reason over* detailed multimodal information from long inputs. As such, we formulate a novel video question-answering task termed *referring reasoning*. Specifically, as part of the question, it contains a *referring query* that references related video contexts, called *referred context*. The model is then required to reason over relevant video details from the referred context. Following the paradigm of referring reasoning, we curate 6,678 human-annotated multiple-choice questions in 17 fine-grained categories, establishing one of the most comprehensive benchmarks for long-form video understanding. Evaluations suggest that the LONGVIDEOBENCH presents significant challenges even for the most advanced proprietary models (*e.g.* GPT-4o, Gemini-1.5-Pro, GPT-4-Turbo), while their open-source counterparts show an even larger performance gap. In addition, our results indicate that model performance on the benchmark improves only when they are capable of processing more frames, positioning LONGVIDEOBENCH as a valuable benchmark for evaluating future-generation long-context LMMs.

Preprint. Under review.

# 1 Introduction

Recent foundation models are processing inputs of longer contexts, with a growth from 2K tokens as in LLaMA [Touvron et al., 2023], to 128K as in GPT-4 [OpenAI, 2024a] and further into millions in models like Gemini-1.5-Pro [Team, 2024]. To measure such development, most benchmarks focus on text-only inputs [Hsieh et al., 2024, Wang et al., 2024a, gkamradt, 2024], while those for long multimodal context remain lacking. In this regard, the task of understanding long-duration videos, such as those extending up to hours, is considered a promising testbed. However, existing video benchmarks exhibit strong single-frame bias. Namely, their results do not improve even models can process more frames. This longstanding issue has continued to be a pain in the neck for video understanding, making evaluation of long-context multimodal inputs a significant challenge.

To address this challenge, this work introduces **LONGVIDEOBENCH**, a video understanding benchmark that measures the progress of LMMs in processing hour-long subtitled videos. In contrary to findings from previous benchmarks, we observe consistent performance improvements when an LMM is capable of processing a larger number of frames (Fig. 1 (b)). To achieve this, we begin by identifying two capabilities essential for long-context multimodal understanding. First, akin to the needle in a haystack (NIAH) evaluation for text LLMs [gkamradt, 2024], effective LMMs must be adept at perceiving specific multimodal details in response to user queries, a task that becomes harder with longer input lengths. Second, in addition to recalling specified elements, the model must be able to relate them and reason about them coherently and contextually. This challenges the model to interpret and integrate large volumes of multimodal information meaningfully.

To effectively evaluate these abilities, we design ***referring reasoning*** (Fig. 1 (a)) as the foundation task for our benchmark. In particular, this task initially introduces a *referring query*. It references particular video contexts, which are termed the *referred context*. Subsequently, the model is presented with a question related to this referred context. This question tests the model's multimodal understanding capabilities, such as visual perception and relational reasoning. To achieve good performance in referring reasoning, models have to interpret the referring query and accurately recall the referred context from the long-context inputs. In addition, they need to perform complex multimodal reasoning. These challenges are closely aligned with the required capabilities as outlined previously.

Following the task of referring reasoning, the LONGVIDEOBENCH contains 6,678 multiple-choice questions on 3,763 videos. These videos are diverse in their themes, including movies, news, life and knowledge, covering 4 progressive duration groups: *8-15 seconds*, *15-60 seconds*, *3-10 minutes*, and *15-60 minutes*, making LONGVIDEOBENCH widely relevant for real-world video applications. Videos are also accompanied with original or transcribed subtitles, which challenges the model to understand long-context interleaved multimodal inputs.

We incorporate *perception* and *relation* questions in the benchmark. Specifically, perception questions require the model to perceive visually on an individual referred video scene, such as to recognize objects, attributes and events. In contrast, relation questions require the model to associate multiple scenes within the referred context, and answer questions about their temporal ordering, attribute change or to track referred objects. These questions are further divided into 17 fine-grained categories, with human-annotated choices, covering a wide range of video understanding tasks.

Our contributions are summarized in three-fold:

1. We introduce LONGVIDEOBENCH (Tab. 1), a multi-choice question-answering benchmark for long-context multimodal video understanding. Our benchmark consists of 6,678 human-crafted comprehensive questions posed on vary-length videos up to an hour long on diverse themes, widely relevant for video understanding applications in the wild.
2. We propose the task of *referring reasoning* to effectively address the longstanding issue of single frame bias in video understanding metrics. As a result, models have to be capable of processing effectively more frames, longer multimodal inputs to improve performance. This requirement distinguishes LONGVIDEOBENCH from existing video benchmarks;
3. We evaluate comprehensively the proprietary and open-source models to understand their long-context multimodal modeling capabilities. Our results demonstrate significant challenges posed by LONGVIDEOBENCH. In addition, the evaluation results show intriguing insights into deficiencies of existing models, thereby offering valuable directions for future research on multimodal long-context understanding.

Table 1: The LONGVIDEOBENCH and popular benchmarks for video LMMs. The *(HT)* denotes the benchmarks split test sets with hidden answers to avoid contamination.

| **Benchmark** | *Labels* | *#Eval Videos* | *#Eval QAs* | *Avg Duration (s)* | *Theme Category* | *Interleaved?* |
|---|---|---|---|---|---|---|
| MSVD-QA [Xu et al., 2017] | Auto | 520 | 13,157 | 10 | Everyday Life | ✗ |
| MSRVTT-QA [Xu et al., 2017] | Auto | 2,990 | 72,821 | 15 | Everyday Life | ✗ |
| ActivityNet-QA Yu et al. [2019] | Human | 800 | 8,000 | 180 | Everyday Life | ✗ |
| NeXT-QA [Xiao et al., 2021] | Human | 1,000 | 8,564 | 44 | Everyday Life | ✗ |
| MVBench [Wang et al., 2023] | Auto | 4,000 | 4,000 | 16 | Life, Human Action, Movie | ✗ |
| EgoSchema [Mangalam et al., 2023] | Auto | 5,031 | 5,031*(HT)* | 180 | Life, Human Action | ✗ |
| MovieChat-1K [Song et al., 2023] | Human | 130 | 1,950 | 500 | Movie | ✗ |
| **LONGVIDEOBENCH (ours)** | Human | 3,763 | 6,678*(HT)* | 473 | Life, Movie, Knowledge, News | ✓ |

Table 2: Definition of 17 categories of *referring reasoning* questions in the LONGVIDEOBENCH.

| **Level** | **Task** | Type of *referring query* (**Q**) | Type of Target Answer | **Code** | # |
|---|---|---|---|---|---|
| **Perception** (L1, 3204) | SCENE-REFERRED EVENT | a scene | an event that happens in **Q** | S2E | 410 |
| | SCENE-REFERRED OBJECT EXISTENCE | a scene | an object that exists in **Q** | S2O | 403 |
| | SCENE-REFERRED OBJECT ATTRIBUTE | a scene$^{q_1}$+an object$^{q_2}$ | an attribute of $q_2$ in $q_1$ | S2A | 403 |
| | EVENT-REFERRED OBJECT | an event | an object that participates **Q** | E2O | 393 |
| | OBJECT-REFERRED EVENT | an object | an event while **Q** appears | O2E | 401 |
| | TEXT-REFERRED EVENT | a subtitle | an event concurrent with **Q** | T2E | 398 |
| | TEXT-REFERRED OBJECT EXISTENCE | a subtitle | an object that exists while **Q** | T2O | 387 |
| | TEXT-REFERRED OBJECT ATTRIBUTE | a subtitle$^{q_1}$+an object$^{q_2}$ | an attribute of $q_2$ while $q_1$ | T2A | 402 |
| **Relation** (L2, 3474) | EVENT BEFORE/AFTER EVENT | an event | an event that happens before/after **Q** | E3E | 406 |
| | OBJECT BEFORE/AFTER OBJECT | an object | an object that appears before/after **Q** | O3O | 394 |
| | SEQUENCE OF SCENES | multiple scenes | the sequential order among **Q** | SSS | 398 |
| | SCENE-REFERRED OBJECT TRACKING | a scene$^{q_1}$+an object$^{q_2}$ | another scene that $q_2$ appears | SOS | 381 |
| | SCENE-REFERRED OBJECT ATTRIBUTE CHANGE | two scenes$^{q_1,q_2}$+an object$^{q_3}$ | attribute change of $q_3$ from $q_1$ to $q_2$ | SAA | 375 |
| | EVENT BEFORE/AFTER TEXT | a subtitle | an event that happens before/after **Q** | T3E | 401 |
| | OBJECT BEFORE/AFTER TEXT | a subtitle | an object that appears before/after **Q** | T3O | 391 |
| | TEXT-REFERRED OBJECT TRACKING | a scene$^{q_1}$, an object$^{q_2}$ | subtitle at $q_2$'s appearance other than $q_1$ | TOS | 380 |
| | TEXT-REFERRED OBJECT ATTRIBUTE CHANGE | two subtitles$^{q_1,q_2}$+an object$^{q_3}$ | attribute change of $q_3$ from $q_1$ to $q_2$ | TAA | 348 |

## 2 The *Referring Reasoning* Task

In this section, we first identify the primary challenges for multimodal long-context understanding. To reflect these challenges, we further define *referring reasoning*, the foundational task for LONGVIDEOBENCH. We introduce its general task scheme and specific categories as follows.

**Challenges for the LONGVIDEOBENCH.** Similar to challenges identified in text-only long-context benchmarks [gkamradt, 2024, Hsieh et al., 2024], the LONGVIDEOBENCH designs question-answering tasks to reflect the following two major difficulties in understanding long videos:

First, **retrieving details** from long videos. Existing studies [gkamradt, 2024, Team, 2024] notice that LLMs or LMMs often struggle to extract specific details from long sequences. To accurately assess this capability in the domain of long videos, the tasks in LONGVIDEOBENCH demand a focus on granular details such as *objects, events*, or *attributes*, rather than a summary or topic overview.

Second, **reasoning contextual relations** in long videos. According to Hsieh et al. [2024], beyond mere retrieval, it is significantly challenging for models to reason about the relationships among extensive inputs. Questions in LONGVIDEOBENCH are therefore designed to compel LMMs to analyze the interconnections among diverse content within a long video to derive the correct answer.

**General Scheme for *Referring Reasoning*.** To effectively measure model performance against aforementioned challenges, we establish the *referring reasoning* task as the fundamental paradigm for LONGVIDEOBENCH. Each question begins by describing a *referring query*, pinpointing one or multiple moments from the video. These video moments, composed of frames and subtitles, are denoted as *referred context*. A specific question body follows the referring query, which requires the model to reason over the referred context to deduct the answer. We employ the multiple-choice question format, where several distracting options are provided alongside the correct answer option.

**Two Levels: Perception *and* Relation.** We divide *referring reasoning* questions into two levels. In **(L1) Perception**, the referring query references a single moment of the video. Then, a question body is posed to ask about the visual perception of a specific concept in the referred moment, such as object, action, or event. (L1) questions mainly challenge models on locating the referred context from the long inputs and understand its visual information. In **(L2) Relation**, the referred context spans across multiple moments of the video. These moments are either related with a specific sequential

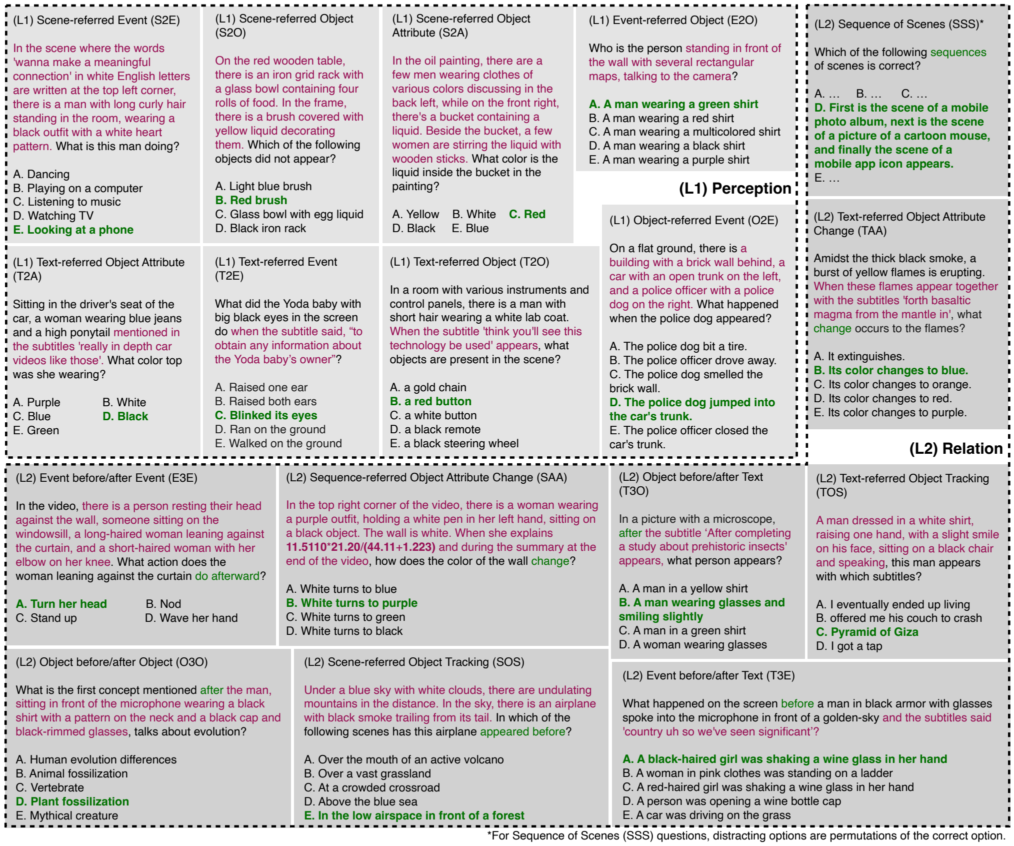



Figure 2: Examples of 17 categories of *referring reasoning* questions in the LongVideoBench.

order (before/after/concurrent) or containing the same concept (*e.g.* the same object appears in these moments). The question is then posed regarding the relations of the moments, and answering these questions require models to not only locate the referred moments, but further reason over their relations. This makes (L2) questions in general more challenging than (L1) questions.

**17 Finer-grained Question Categories.** We further subdivide the two levels of questions into 17 finer-grained categories, dividing based on the type of referring query and the type of target answer. As listed in Tab. 2, given interleaved multimodal inputs, the referring query could either be describing a scene, an event, or an object from the video frames, or be narrating a sentence or a phrase from the text subtitles. The target answer typically is about a visual concept (an event, object, or attribute) from one of the referred moments, with two exceptions: the Sequence of Scenes (SSS) category requires to answer the correct sequential order of multiple ($> 3$) scenes in the video, and the Text-referred Object Tracking (TOS) requires to answer the specific subtitle while a given object appears.

## 3 Dataset Construction

In this section, we discuss the dataset construction for the LongVideoBench. We first define the category and duration groups of videos (Sec. 3.1), then we introduce the process of collecting and creating interleaved video-subtitle data (Sec. 3.2), lastly we elaborate on the human annotation process to collect high-quality referring questions and answers for LongVideoBench (Sec. 3.3).

### 3.1 Groups of Videos

**Progressive Duration Groups.** In LongVideoBench, we aim to not only evaluate LMMs on ultra-long videos, but analyze how their ability changes from short videos (*about 10s*) to long (*hour-long*). In light of this, we propose to collect videos in four progressive duration groups, as listed in Tab. 3.1. The first two groups contain shorter videos of length *(8s, 15s]* and *(15s, 60s]*, whereas the latter two duration groups contain long videos of length *(180s, 600s]* and *(900s, 3600s]*. The four groups not

Table 3: Statistics of videos in LongVideoBench, by duration groups and video layouts.

| **Duration Group** | *(8s, 15s]* | | | | *(15s, 60s]* | | | | *(180s, 600s]* | | *(900s, 3600s]* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Source Platform** | Landscape | | Portrait | | Landscape | | Portrait | | Landscape | | Landscape | |
| *Statistics* | *Duration* | *#Videos* | *Duration* | *#Videos* | *Duration* | *#Videos* | *Duration* | *#Videos* | *Duration* | *#Videos* | *Duration* | *#Videos* |
| | 11.06 | 546 | 11.93 | 338 | 33.88 | 551 | 38.59 | 374 | 389 | 986 | 1408 | 966 |

Table 4: Statistics of videos in LongVideoBench, by category groups (*in two-letter codes, as defined in Sec. 3.1*) and video layouts (*LS*: Landscape; *PT*: Portrait.)

| **Category Group** | Movie Recaps (*MR*) | | *Everyday Life* | | | | | | News Program (*NP*) | Knowledge | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *LT* | | *LV* | | *LC* | | | *KA* | *KH* | *KG* | *KS* | *KC* |
| Source Platform | *LS* | *PT* | *LS* | *PT* | *LS* | *PT* | *LS* | *PT* | *LS* | *LS* | *LS* | *LS* | *LS* | *LS* |
| *#Channels* | 18 | 4 | 10 | 6 | 7 | 5 | 8 | 5 | 12 | 5 | 8 | 7 | 10 | 7 |
| *#Downloaded Videos* | 7679 | 1106 | 2230 | 2532 | 1009 | 1891 | 2731 | 4706 | 24002 | 2010 | 2900 | 1280 | 1350 | 335 |
| *#Annotated Videos* | 352 | 160 | 343 | 173 | 338 | 179 | 336 | 203 | 329 | 327 | 336 | 330 | 200 | 160 |

only cover the duration ranges of existing video understanding benchmarks, but also provide a unique hour-long subset to further expand the video length beyond existing benchmarks.

**Category Groups.** Existing LMM benchmarks for long videos typically focus on a specific category of videos, *e.g.* egocentric videos [Mangalam et al., 2023], or movies [Song et al., 2023, Zhang et al., 2023a]. In comparison, LongVideoBench is a more comprehensive benchmark that covers diverse categories of contents. The videos in LongVideoBench are collected from 99 different channels for landscape videos and 20 channels for portrait videos, in the 10 following categories: Movie Recaps (*MR*); three life-related categories: Travel Guides (*LT*), Life Vlogs (*LV*), Cooking/Recipes (*LC*); News Programs (*NP*); and five knowledge-related categories: Art (*KA*), History (*KH*), Geography (*KG*), STEM (*KS*), Computer Science (*KC*). As listed in Tab. 4, LongVideoBench includes a sufficient number of videos from all 10 category groups, spanning over a diverse distribution.

### 3.2 Video and Subtitle Collection

The video and subtitle collection process is illustrated in Fig. 3. First, all videos with at least 720P resolution from the 119 channels are downloaded. After downloading the videos, for the source platforms that provide transcribed subtitles, we remove the videos without transcribed subtitles or with non-English subtitles. For those videos without provided transcribed subtitles, we employ Whisper-V3-Large [OpenAI, 2024b] to generate subtitles for them. These videos are further sampled to cover different topics uniformly. Finally, we evaluate their video quality via Q-Align [Wu et al., 2024] and remove especially low-quality videos to ensure that all videos have scores $> 0.25$ (in range $[0, 1]$). Remaining videos are further manually filtered by annotators (in Sec. 3.3) to the final 3,763 videos.

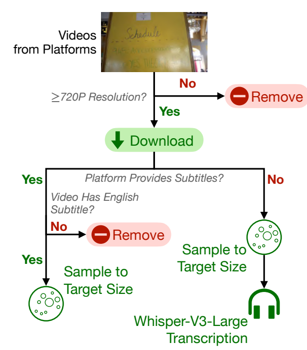



Figure 3: Video collection for LongVideoBench, ensuring all videos have subtitles.

Subtitles are important for multimodal video understanding, as they provide vital text information from human speech and reduce ambiguity from pure visual scenes. Aligning with the way humans watch videos with subtitles, in LongVideoBench, we require LMMs to receive the text subtitles simultaneously with concurrent frames. To achieve this, we define the **interleaved multimodal input** format to feed videos and subtitles together into LMMs as temporally-aligned multimodal sequences. Specifically, a chunk of subtitle will be inserted in-between the two frames before and after the mid-timestamp of the subtitle.

### 3.3 Annotating Questions and Answers

We conduct the annotation process in a well-controlled lab environment with experienced annotators. Before annotation, we conduct a special training to all annotators for them to understand the requirements of each specific question category. During the annotation process, the subtitles are appended at the bottom of the video with aligned timestamps, and displayed to annotators. The annotator is required to watch the full video before starting the annotation, and is allowed to drag back to to any specific timestamps after full watching. We collect one question per video for videos longer than 60 seconds, two questions for videos in *(180s, 600s]*, three questions for *(900s, 3600s]*. The annotator also needs to provide 3-4 distracting answer options that are relevant to the question and the video.

We further introduce two additional annotation requirements to ensure high-quality *referred reasoning* questions: 1) We explicitly require annotators to include and highlight the referred query in all questions[1]; 2) To ensure that the referred context uniformly span over the video, we ask annotators to explicitly label the frame index for all referred moments in each question. This additional requirement further facilitates our in-depth study of long-context understanding abilities for LMMs with respect to the relative token-wise distance between the question and the referred context.

To control the annotation quality, each video is passed through three annotators: 1) The primary annotator, whose duty is to provide annotations and filter out videos that are not available for annotation (*e.g.* still frames, incomplete subtitles); 2) The examiner, who examines whether the annotated question is in the correct question category, and whether the annotation requirements are all met; 3) The reviser, to revise the annotations labeled as incorrect by examiners. The examiner and reviser have identified 20% of annotations to be problematic and revised them, which significantly improved the quality of the LongVideoBench.

As we require all questions to include the question body itself as well as a referring query, the average question length is as long as **43.53** words, ensuring that the referred context is clearly depicted in the question without introducing ambiguity. The average length of an answer is 8.28 words.

# 4 Evaluation of LongVideoBench

## 4.1 Models and Evaluation Strategies

**Participating LMMs.** We include in total 22 LMMs for evaluation. The main participants are long-context LMMs, including four proprietary models: GPT-4o (`gpt-4o-0513`), Gemini-1.5-Pro (`gemini-1.5-pro-0514`), GPT-4-Turbo (`gpt-4-turbo-0409`), and Gemini-1.5-Flash (`gemini-1.5-flash-0514`), and four state-of-the-art open-sourced long-context LMMs: Phi-3-Vision-Instruct (*128K*), Idefics2 (*32K*), Mantis-Idefics2 (*32K*), and Mantis-BakLLaVA (*32K*). All these models above support interleaved video-language inputs. We also evaluate 9 representative video-specific LMMs, and 6 image LMMs that support $\geq$8 images.

**Validation and Test Subsets.** We split the LongVideoBench into two subsets, the *validation set* (752 videos, 1337 MCQs), and the *test set* (3011 videos, 5341 MCQs). We use the *validation set* to analyze the performance of LMMs under different settings. Afterwards, we pick the optimal setting for each LMM to report their performance on test set leaderboard.

## 4.2 Main Results

In Tab. 5 and Tab. 6, we analyze the performance of six long-context LMMs under different settings on the val set of LongVideoBench. Our evaluation brings several important findings, as follows:

**1)** ***LMMs have to understand long inputs for better results.*** As shown in Tab. 5 (a), (b), (c) and (d), all four proprietary models, especially more advanced GPT-4o and Gemini-Pro, have shown significant improvements while increasing their input length, in particular for long videos. For videos longer than 180 seconds, GPT-4o and Gemini-1.5-Pro can improve more than **10%** by increasing input length from 16 to 256 frames. In contrast, on EgoSchema, Gemini-1.5-Pro only improves 2.5% from 16 to 150 frames. This validates the effectiveness of LongVideoBench as a longstanding challenging benchmark for models to evaluate their long-context multimodal understanding abilities.

**2)** ***Open-source models lag significantly behind.*** Different from proprietary models, open-source LMMs are unable to improve their results by inputting more than 16 frames. Idefics2 and Mantis-Idefics2, as shown in Tab. 5 (e) and (g), even face a severe degradation on accuracy with 64 input frames, before they have reached their context length limits.

**3)** ***Longer videos are more challenging.*** As in Tab. 5, all six models show the lowest accuracy on the longest *(900,3600]* group, followed by the *(180,600]* group, and then the shorter-video groups. These results pose LongVideoBench as a meaningful and challenging benchmark for LMMs to test their video understanding abilities.

[1]Except Sequence of Scenes questions, where it is *implicitly* mentioned in all candidate choices.

Table 5: Validation set results categorized by duration groups, *w.r.t.* `max_frames` (capped at 1 *fps*). While `max_frames` is already more than the max duration of a group, the results will not change when we set a larger `max_frames`. Respective settings are labeled as "s.a." (same as above).

| Model | `max_frames` | Duration Group (unit: second) *(8,15]* | *(15,60]* | *(180,600]* | *(900,3600]* | *all* |
|---|---|---|---|---|---|---|
| (a) GPT-4o | 1 | 52.9 | 50.6 | 40.8 | 36.0 | 41.7 |
| | 4 | 63.5 | 64.3 | 47.2 | 40.3 | 48.7 |
| | 8 | 69.7 | 67.3 | 49.4 | 47.1 | 53.3 |
| | 16 | **71.4** | 73.7 | 53.8 | 52.2 | 58.0 |
| | 32 | s.a. | 73.5 | 57.3 | 50.5 | 58.5 |
| | 64 | s.a. | **76.7** | 61.4 | 55.8 | 62.0 |
| | 128 | s.a. | s.a. | 64.2 | 56.5 | 63.5 |
| | 256 | s.a. | s.a. | **69.1** | **60.9** | **66.7** |
| (b) Gemini-1.5-Pro | 1 | 46.6 | 45.2 | 35.7 | 35.8 | 38.6 |
| | 4 | 59.6 | 62.9 | 37.7 | 39.0 | 44.9 |
| | 8 | 62.4 | 68.0 | 44.9 | 46.0 | 51.0 |
| | 16 | **67.4** | 69.6 | 50.3 | 44.0 | 52.7 |
| | 32 | s.a. | 74.3 | 51.2 | 48.0 | 55.2 |
| | 64 | s.a. | **75.1** | 59.3 | 50.9 | 58.6 |
| | 128 | s.a. | s.a. | 64.9 | 54.0 | 61.9 |
| | 256 | s.a. | s.a. | **65.3** | **58.6** | **64.0** |
| (c) GPT-4-Turbo | 1 | 49.2 | 48.3 | 43.7 | 39.2 | 43.2 |
| | 4 | 57.1 | 57.0 | 46.6 | 43.8 | 48.1 |
| | 8 | 59.8 | 62.8 | 50.7 | 41.5 | 49.7 |
| | 16 | **65.2** | 67.9 | 51.7 | 44.5 | 52.7 |
| | 32 | s.a. | 66.9 | 53.1 | 47.5 | 54.0 |
| | 64 | s.a. | **68.2** | 59.0 | 47.0 | 56.0 |
| | 128 | s.a. | s.a. | 60.3 | 49.3 | 57.5 |
| | 256 | s.a. | s.a. | **62.4** | **50.5** | **59.0** |
| (d) Gemini-1.5-Flash | 1 | 48.6 | 42.9 | 35.1 | 35.4 | 38.1 |
| | 4 | 53.3 | 64.5 | 40.0 | 40.4 | 45.2 |
| | 8 | 62.5 | 65.3 | 45.8 | 41.8 | 48.9 |
| | 16 | **68.3** | 66.9 | 49.0 | 43.9 | 50.8 |
| | 32 | s.a. | 74.1 | 50.0 | 44.5 | 53.5 |
| | 64 | s.a. | 76.2 | 54.4 | 48.6 | 56.8 |
| | 128 | s.a. | s.a. | 56.9 | 51.7 | 58.9 |
| | 256 | s.a. | s.a. | **62.6** | **54.0** | **61.6** |

| Model | `max_frames` | Duration Group (unit: second) *(8,15]* | *(15,60]* | *(180,600]* | *(900,3600]* | *all* |
|---|---|---|---|---|---|---|
| (e) Idefics2 | 1 | 48.6 | 48.8 | 39.3 | 38.5 | 41.5 |
| | 4 | 62.4 | 58.1 | 41.3 | 41.3 | 46.4 |
| | 8 | 59.3 | 63.4 | 46.8 | 41.7 | 48.5 |
| | 16 | **59.8** | **65.7** | **47.8** | **42.7** | **49.7** |
| | 32 | s.a. | 64.5 | 44.0 | 41.5 | 47.8 |
| | 64 | s.a. | 52.3 | 22.1 | 21.2 | 30.9 |
| (f) Phi-3-Vision | 1 | 49.2 | 46.5 | 39.3 | 37.4 | 40.8 |
| | 4 | 56.6 | 57.5 | 44.4 | 43.6 | 47.5 |
| | 8 | 60.8 | 62.2 | 42.5 | 43.6 | 48.1 |
| | 16 | **59.3** | 61.6 | **46.8** | **44.7** | **49.6** |
| | 32 | s.a. | **66.3** | 46.6 | 42.3 | 49.1 |
| | 64 | – Context Length Exceeded – | | | | |

| Model | `max_frames` | Duration Group (unit: second) *(8,15]* | *(15,60]* | *(180,600]* | *(900,3600]* | *all* |
|---|---|---|---|---|---|---|
| (g) Mantis-Idefics2 | 1 | 48.1 | 44.2 | 35.4 | 36.2 | 38.7 |
| | 4 | 53.4 | 51.2 | 42.5 | 38.7 | 43.5 |
| | 8 | **57.7** | **57.0** | 45.4 | 39.5 | 46.1 |
| | 16 | 56.6 | 55.8 | **45.6** | **42.2** | **47.0** |
| | 32 | s.a. | 55.8 | 42.7 | 40.4 | 45.4 |
| | 64 | s.a. | 48.0 | 24.7 | 24.9 | 30.2 |
| (h) Mantis-BakLLaVA | 1 | 48.1 | 44.2 | 35.4 | 36.2 | 38.7 |
| | 4 | **57.7** | 50.0 | 38.8 | 36.5 | 42.0 |
| | 8 | 54.0 | 55.8 | 39.8 | 37.8 | 43.0 |
| | 16 | 53.4 | **57.6** | **40.3** | **38.7** | **43.7** |
| | 32 | s.a. | 54.7 | 39.8 | 37.8 | 42.8 |
| | 64 | – Context Length Exceeded – | | | | |

Table 6: Validation set results *w.r.t.* input modalities.

| Video Frames? | Text Subtitles? | GPT-4o | Gemini-1.5-Pro | GPT-4-Turbo | Gemini-1.5-Flash | Idefics2 | Phi-3-Vision | Mantis-Idefics2 | Mantis-BakLLaVA |
|---|---|---|---|---|---|---|---|---|---|
| ✗ | ✓ | 44.6 | 43.0 | 45.2 | 39.2 | 25.6 | 40.7 | 31.7 | 31.1 |
| ✓ | ✗ | 60.6 | 62.9 | 56.0 | 60.2 | 49.4 | 49.5 | 45.8 | 43.5 |
| ✓ | ✓ | **66.7** | **63.9** | **59.0** | **61.6** | **49.7** | **49.6** | **47.0** | **43.7** |

**4)** ***Interleaved inputs are hard.*** As shown in Tab. 6, all models can improve their results by inserting subtitles to videos as inputs. However, compared to GPT-4o, open-source LMMs are still unable to effectively integrate subtitle information to facilitate video understanding and improve their accuracy on LongVideoBench, demonstrating a gap in long-context multimodal understanding.

**5)** ***Visual modality is fundamental.*** Results from Tab. 6 also demonstrate that video frames, *i.e.* visual modality, is a crucial component in the interleaved inputs, as removing them and only using the subtitles lead to much worse results for all models.

### 4.3 Leaderboard

Table 7 shows the test set results of the 6 long-context LMMs, as well as 9 representative open-source video LMMs and 6 open-source image LMMs with multi-image support. By including more LMMs for evaluation, this leaderboard raises more observations, as follows:

**6)** ***Open-source video LMMs do not show clear advantages over image LMMs.*** Under the same model architecture, LLaVA-Next-Video-M7B (*video LMM*) is less competitive than LLaVA-Next-Mistral (*image LMM*), despite being trained on additional videos. This may be due to existing video training datasets mainly consisting of short videos and summary-level tasks, leading to a decline on long-context and detailed video understanding capabilities.

Table 7: Test Set Leaderboard of the LONGVIDEOBENCH on 23 LMMs, by duration groups and question categories. We also show the validation set results ("Val Total") as a reference.

| Model | Val Total | Duration Group (s) | | | | Question Category | | | | | | | | | | | | | | | | | Test Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | (L1) Perception | | | | | | | | (L2) Relation | | | | | | | | | |
| | | *(8, 15]* | *(15, 60]* | *(180, 600]* | *(900, 3600]* | S2E | S2O | S2A | E2O | O2E | T2E | T2O | T2A | E3E | O3O | SSS | SOS | SAA | T3E | T3O | TOS | TAA | |
| *Proprietary Long-context LMMs: (`max_frames` set according to Tab. 5)* | | | | | | | | | | | | | | | | | | | | | | | |
| GPT-4o (0513) | **66.7** | **71.6** | **76.8** | **66.7** | **61.6** | **76.8** | **69.8** | 70.9 | 67.3 | 72.8 | **67.2** | **65.3** | **77.2** | **62.6** | **61.3** | 44.3 | **75.6** | **62.6** | **64.0** | **66.4** | **62.1** | **66.4** | **66.7** |
| Gemini-1.5-Pro (0514) | 64.0 | 70.2 | 75.3 | 65.0 | 59.1 | 74.6 | 58.3 | **76.2** | **68.7** | **73.3** | 66.2 | 63.6 | 76.7 | 61.9 | 58.6 | **55.2** | 69.0 | 59.0 | 58.9 | 60.5 | 53.3 | 62.5 | 64.4 |
| Gemini-1.5-Flash (0514) | 61.6 | 66.1 | 73.1 | 63.1 | 57.3 | 68.5 | 64.7 | 68.0 | 64.5 | 72.5 | 63.6 | 68.0 | 76.7 | 56.5 | 61.0 | 43.1 | 67.3 | 56.2 | 57.5 | 55.0 | 55.3 | 60.7 | 62.4 |
| GPT-4-Turbo (0409) | 59.1 | 66.4 | 71.1 | 61.7 | 54.5 | 74.9 | 60.1 | 64.2 | 63.9 | 69.4 | 62.5 | 61.3 | 69.9 | 57.5 | 55.9 | 44.8 | 66.0 | 53.2 | 56.5 | 53.6 | 56.2 | 60.2 | 60.7 |
| *Open-source Long-Context LMMs: (`max_frames` set according to Tab. 5)* | | | | | | | | | | | | | | | | | | | | | | | |
| Idefics2 | 49.7 | 57.4 | 60.4 | 47.3 | 44.7 | 60.9 | 51.4 | 49.4 | 53.7 | 58.9 | 54.4 | 51.8 | 54.8 | 46.8 | 40.5 | 28.9 | 61.0 | 49.8 | 47.0 | 42.0 | 40.7 | 46.2 | 49.4 |
| Phi-3-Vision-Instruct | 49.6 | 58.3 | 59.6 | 48.4 | 45.1 | 60.3 | 52.9 | 53.4 | 51.8 | 54.1 | 52.3 | 55.3 | 53.3 | 49.4 | 47.6 | 33.6 | 59.3 | 46.2 | 44.2 | 43.2 | 38.8 | 51.5 | 49.9 |
| Mantis-Idefics2 | 47.0 | 56.1 | 61.4 | 44.6 | 42.5 | 60.3 | 51.1 | 51.2 | 53.4 | 52.9 | 51.4 | 49.5 | 57.3 | 46.2 | 45.1 | 30.2 | 53.7 | 46.5 | 44.2 | 40.1 | 30.6 | 40.2 | 47.6 |
| Mantis-BakLLaVA | 43.7 | 51.3 | 52.7 | 41.1 | 40.1 | 53.0 | 38.7 | 44.1 | 46.0 | 51.0 | 50.8 | 43.7 | 50.8 | 45.5 | 40.2 | 23.3 | 48.0 | 44.9 | 40.9 | 38.5 | 34.9 | 47.7 | 43.7 |
| *Open-source Image LMMs with Multi-Image Support: (all sample 8 frames)* | | | | | | | | | | | | | | | | | | | | | | | |
| LLaVA-Next-Mistral-7B | 49.1 | 53.4 | 57.2 | 46.9 | 42.1 | 59.0 | 46.5 | 49.4 | 49.7 | 52.2 | 52.9 | 51.1 | 51.4 | 47.4 | 45.4 | 28.2 | 56.0 | 50.8 | 38.7 | 41.6 | 31.9 | 48.1 | 47.1 |
| InstructBLIP-T5-XXL | 43.3 | 48.1 | 50.1 | 44.5 | 40.0 | 54.9 | 39.3 | 41.3 | 45.4 | 49.7 | 52.9 | 42.4 | 48.6 | 44.2 | 40.2 | 25.2 | 51.0 | 42.9 | 42.7 | 41.6 | 33.9 | 47.7 | 43.8 |
| BLIP-2-T5-XXL | 42.7 | 46.7 | 47.4 | 44.2 | 40.9 | 54.6 | 38.1 | 38.8 | 46.3 | 49.0 | 52.6 | 40.2 | 44.3 | 45.2 | 41.2 | 25.6 | 51.3 | 41.6 | 45.1 | 45.1 | 33.6 | 47.4 | 43.5 |
| LLaVA-1.5-13B | 43.4 | 49.0 | 51.1 | 41.8 | 39.6 | 54.9 | 42.6 | 40.4 | 44.8 | 49.0 | 51.1 | 43.1 | 43.0 | 45.2 | 40.9 | 29.9 | 53.3 | 44.2 | 38.7 | 35.6 | 30.0 | 46.2 | 43.1 |
| LLaVA-1.5-7B | 40.3 | 45.0 | 47.4 | 40.1 | 37.0 | 53.3 | 35.0 | 38.8 | 39.6 | 44.9 | 44.1 | 39.9 | 43.3 | 40.7 | 43.9 | 26.2 | 47.3 | 42.9 | 37.2 | 34.7 | 30.3 | 45.1 | 40.4 |
| mPLUG-Owl2 | 39.1 | 49.4 | 47.3 | 38.7 | 34.3 | 49.5 | 37.5 | 37.3 | 39.6 | 45.5 | 45.9 | 41.5 | 39.6 | 44.6 | 36.9 | 24.9 | 45.7 | 38.9 | 30.9 | 36.6 | 33.9 | 38.3 | 39.4 |
| *Open-source Video LMMs: (frame sampling set as their default settings)* | | | | | | | | | | | | | | | | | | | | | | | |
| PLLaVA-34B | 53.2 | 60.1 | 66.8 | 50.8 | 49.1 | 65.9 | 53.8 | 53.1 | 54.9 | 57.6 | 58.9 | 52.4 | 56.3 | 54.8 | 50.6 | 44.2 | 60.3 | 56.1 | 46.6 | 47.9 | 41.4 | 54.9 | 53.5 |
| LLaVA-Next-Video-34B | 50.5 | 57.6 | 61.6 | 48.7 | 45.9 | 62.1 | 50.2 | 51.2 | 50.9 | 58.5 | 59.0 | 48.2 | 48.9 | 54.8 | 49.7 | 39.2 | 58.7 | 50.8 | 46.6 | 43.8 | 36.8 | 47.2 | 50.5 |
| PLLaVA-13B | 45.6 | 52.9 | 54.3 | 42.9 | 41.2 | 57.1 | 43.5 | 41.9 | 47.3 | 53.5 | 54.4 | 46.9 | 43.7 | 47.1 | 43.6 | 27.2 | 58.0 | 44.2 | 39.6 | 40.1 | 30.9 | 47.0 | 45.1 |
| LLaVA-Next-Video-M7B | 43.5 | 50.9 | 53.1 | 42.6 | 38.9 | 54.6 | 41.7 | 47.2 | 46.3 | 52.9 | 46.8 | 46.6 | 45.8 | 44.9 | 42.1 | 24.6 | 51.3 | 40.6 | 39.0 | 40.1 | 34.5 | 39.5 | 43.5 |
| ShareGPT4Video | 39.7 | 46.9 | 50.1 | 40.0 | 38.7 | 50.2 | 37.5 | 44.4 | 44.2 | 42.7 | 43.8 | 41.2 | 45.8 | 41.7 | 42.7 | 29.9 | 50.3 | 47.2 | 38.7 | 39.7 | 29.3 | 39.8 | 41.8 |
| PLLaVA-7B | 40.2 | 45.3 | 47.3 | 38.5 | 35.2 | 52.4 | 35.3 | 40.4 | 39.3 | 46.8 | 46.5 | 39.9 | 39.3 | 41.0 | 36.3 | 26.2 | 47.7 | 41.6 | 34.1 | 30.5 | 27.7 | 38.3 | 39.2 |
| VideoChat2 (Mistral-7B) | 39.3 | 49.3 | 49.3 | 39.0 | 37.5 | 53.6 | 40.8 | 38.5 | 44.5 | 53.5 | 46.8 | 43.1 | 47.7 | 43.6 | 46.6 | 10.6 | 42.0 | 40.6 | 38.4 | 36.3 | 27.4 | 43.6 | 41.2 |
| VideoLLaVA | 39.1 | 43.1 | 44.6 | 36.4 | 34.4 | 49.5 | 29.6 | 30.6 | 40.9 | 44.9 | 43.5 | 33.8 | 40.6 | 46.5 | 38.7 | 24.3 | 40.0 | 42.9 | 35.1 | 30.5 | 23.8 | 39.5 | 37.6 |
| VideoChat2 (Vicuna 7B) | 36.0 | 38.1 | 40.5 | 33.5 | 33.6 | 44.8 | 29.0 | 27.3 | 36.9 | 41.7 | 41.7 | 34.1 | 33.1 | 37.2 | 39.6 | 22.6 | 43.0 | 30.7 | 34.1 | 33.8 | 28.3 | 37.2 | 35.1 |

**7)** ***Stronger LLM backbones are helpful.*** Compared with PLLaVA-7B, its larger variants trained with the same datasets, PLLaVA-13B and PLLaVA-34B, shows notable 5.9% and 14.3% improvements, and PLLaVA-34B ranks top among all open-source models. This observation suggests that scaling up the language model is effective for more comprehensive video understanding.

**8)** ***(L2) Relation is more challenging than (L1) Perception.*** Compared to (L1), questions in (L2) additionally require LMMs to understand the relation among multiple scenes in the video. Thus, the disparity between performance on (L1) and (L2) indicates LMMs' insufficient understanding of the temporal dynamics of videos. The most difficult category is an (L2) category, SSS (SEQUENCE OF SCENES), where the distracting options are permutations of the correct sequence order (of the scenes). All LMMs perform worst on this category of questions, further highlighting their limitation of complex temporal understanding.

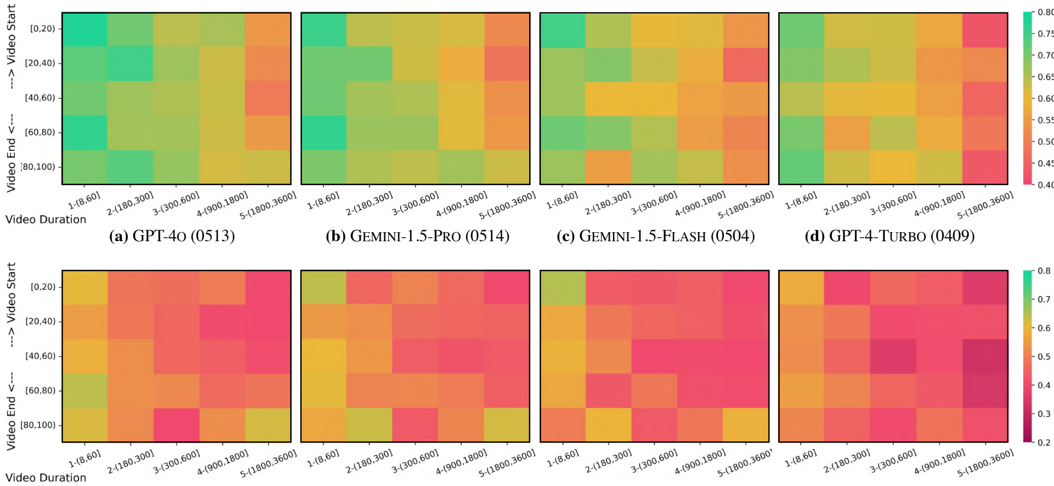
**9)** ***Results on validation and test subsets are consistent.*** This consistency demonstrates the validation set as a sufficient representation of the entire benchmark dataset, confirming the reliability of LONGVIDEOBENCH and the findings in Sec. 4.2.

### 4.4 Performance *w.r.t.* Referring Query Depth

In Fig. 4, we further analyze the performance trends of LMMs when the queried moment is located at different positions within a video. In summary, the performance of LMMs is not uniform: all models perform worse when the referred moment is closer to the beginning of the video (*i.e.* has longer distance to the question), and this trend becomes more evident as the video duration becomes longer. Additionally, we found that questions posed closer to the middle of the video, rather than the beginning or end, present a greater challenge for LMMs. These findings are consistent with respective conclusions from needle in a haystack (NIAH) [gkamradt, 2024] for long-term text understanding.

## 5 Related Works

**Video LMMs and Long-context LMMs.** Early video LMMs focus on short videos (less than one minute). These works usually build upon pre-trained video backbones [Wang et al., 2023, 2024b], temporal pooling modules [Zhang et al., 2023b,c, Xu et al., 2024] and are trained on video-specific supervised tuning datasets [Li et al., 2023a, Zhang et al., 2023c]. Several image LMMs [Li et al.,




**(a)** GPT-4o (0513) **(b)** Gemini-1.5-Pro (0514) **(c)** Gemini-1.5-Flash (0504) **(d)** GPT-4-Turbo (0409)

**(e)** Idefics2 **(f)** Phi-3-Vision **(g)** Mantis-Idefics2 **(h)** Mantis-BakLLaVA

Figure 4: Accuracy of proprietary and open-source LMMs *w.r.t.* referring query depth and video duration. All models perform worse when the referred moment is closer to video start or middle video. Please refer to Appendix Sec. B for respective visualizations on rest 15 models.

2023b, Liu et al., 2023a, Ye et al., 2023] have shown competitive performance on many traditional short-video understanding tasks [Xu et al., 2017, Yu et al., 2019].

For longer videos, recent research explores methods like compressing video frames to fewer tokens to manage hour-long content within LMMs [Li et al., 2023c], and incorporating memory banks into standard LMM architectures [Song et al., 2023, He et al., 2024, Tan et al., 2024]. Leading models, both open-source (*e.g.,* LWM [Liu et al., 2024a], Phi-3-128K [Abdin et al., 2024]) and proprietary (*e.g.,* GPT-4o [OpenAI, 2024a], Gemini-1.5-Pro [Team, 2024]), now support context lengths over 128K tokens, allowing detailed video analysis. However, robust benchmarks for long-duration video understanding are lacking, with GPT-4o assessed only on 3-minute videos [Yu et al., 2019, Mangalam et al., 2023] and Gemini-1.5-Pro on an in-house benchmark. To advance LMM capabilities in understanding longer videos, we introduce LongVideoBench, a comprehensive benchmark for evaluating LMMs across various video durations and distributions.

**Benchmarks for Video LMMs.** Traditionally, video LMMs are evaluated on classical video QA datasets like MSVD-QA, MSRVTT-QA [Xu et al., 2017], and ActivityNet-QA [Yu et al., 2019], which primarily evaluate video LMMs through global-summary questions. However, it has been demonstrated that these benchmarks are addressable by a few key frames. For a focused assessment of temporal comprehension, NeXT-QA [Xiao et al., 2021] and MVBench [Wang et al., 2023] serve to measure temporal dynamics over short clips, with average durations of *44s* and *16s*, respectively. Long-duration video understanding is targeted by benchmarks like EgoSchema [Mangalam et al., 2023], which involves multi-choice questions on *3-minute-long* egocentric videos, and MovieChat-1K [Song et al., 2023], focused on *10-minute-long* movie clips. These long-video benchmarks often limit their scope to videos on specific themes and still include a large proportion of summary questions solvable with limited frames. To address these gaps and enhance evaluation of detailed multimodal reasoning over longer videos, we introduce the LongVideoBench, a comprehensive benchmark focusing on referring reasoning questions that by-design requires dense input frames to solve, encompassing diverse video topics and varying lengths up to hour long.

## 6 Conclusion

This work introduces LongVideoBench, a comprehensive benchmark that evaluates Large Multimodal Models (LMMs) in understanding hour-long subtitled videos in diverse themes. The benchmark introduces referring reasoning questions, a novel video question-answering paradigm that addresses the longstanding issue of single frame bias in existing video understanding benchmarks. Evaluation results demonstrate that LongVideoBench presents significant challenges for both proprietary and open-source LMMs in their long-context multimodal capabilities. In addition, the benchmark results provide valuable insights on the deficiencies of existing models, making it a valuable asset to understand the current multimodal model landscape and to guide the future explorations.