ALLVB: All-in-One Long Video Understanding Benchmark

Xichen Tan[1], **Yuanjing Luo**[1], **Yunfan Ye**[2], **Fang Liu**[2*], **Zhiping Cai**[1]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, China
[2]School of Design, Hunan University, Changsha, China
tanxc23@nudt.edu.cn

Abstract

From image to video understanding, the capabilities of Multimodal LLMs (MLLMs) are increasingly powerful. However, most existing video understanding benchmarks are relatively short, which makes them inadequate for effectively evaluating the long-sequence modeling capabilities of MLLMs. This highlights the urgent need for a comprehensive and integrated long video understanding benchmark to assess the ability of MLLMs thoroughly. To this end, we propose ALLVB (ALL-in-One Long Video Understanding Benchmark). ALLVB's main contributions include: 1) It integrates 9 major video understanding tasks. These tasks are converted into video QA formats, allowing a single benchmark to evaluate 9 different video understanding capabilities of MLLMs, highlighting the versatility, comprehensiveness, and challenging nature of ALLVB. 2) A fully automated annotation pipeline using GPT-4o is designed, requiring only human quality control, which facilitates the maintenance and expansion of the benchmark. 3) It contains 1,376 videos across 16 categories, averaging nearly 2 hours each, with a total of 252k QAs. To the best of our knowledge, it is the largest long video understanding benchmark in terms of the number of videos, average duration, and number of QAs. We have tested various mainstream MLLMs on ALLVB, and the results indicate that even the most advanced commercial models have significant room for improvement. This reflects the benchmark's challenging nature and demonstrates the substantial potential for development in long video understanding.

Datasets —
https://huggingface.co/datasets/ALLVB/ALLVB

Introduction

The field of Large Language Models (LLMs) is currently a highly popular research area, rapidly evolving from text-focused to multi-modal, incorporating image and video inputs. To objectively assess the performance of these models, various Q&A benchmarks are continually being proposed.

For pure text benchmarks, notable examples include MMLU (Hendrycks et al. 2020), which evaluates LLMs across 57 tasks in various academic fields, and AGIEval (Zhong et al. 2023), which focuses on performance in standardized exams like GRE, GMAT, and China's Gaokao. For image-text multi-modal benchmarks, ScienceQA (Lu et al. 2022) includes Q&As from elementary and middle school curricula, while MMMU (Yue et al. 2024) features questions from six core academic subjects at the university level.

Regarding video-text benchmarks, current benchmarks mainly focused on short videos, such as MSVD-QA (Xu et al. 2017), MSRVTT-QA (Xu et al. 2017), TGIF-QA (Jang et al. 2017), ActivityNet-QA (Yu et al. 2019), and BDIQA (Mao et al. 2024). These benchmarks typically involve videos averaging under 10 seconds, except for ActivityNet-QA, which averages 180 seconds.

When it comes to benchmarks for long videos, several contemporary works are noteworthy, including MLVU (Zhou et al. 2024), Video-MME (Fu et al. 2024), and LVBench (Wang et al. 2024). They rely on manual Q&As annotations, limiting scalability due to labor costs, and resulting in shorter videos and fewer Q&As per video. Moreover, the lack of a unified standard in designing video comprehension questions poses challenges. To address these issues, we develop an automated pipeline for a more efficient and scientifically robust benchmark.

Since existing LLMs do not yet natively support video modality input, the current method involves generating textual descriptions of video content for the models. Given the significant challenges of describing hour-long videos, we have opted to use existing text descriptions closely related to the video content, with movie scripts being the most common example.

First, we collect a large number of movie scripts from open-source websites, then filter out duplicates and unsuitable scripts, ultimately obtaining 1,376 movie scripts. These scripts are used as input to GPT-4o (OpenAI 2024) for generating video-related Q&As. To ensure that the LLM can capture the details within the movie and considering its limitations in handling ultra-long contexts, we use a two-stage segmentation method to divide the script into different plots and sub-plots, and then construct overarching Q&As for the entire script and detailed Q&As for the sub-plots. This method ensures the correctness of the Q&As construction.

To design the Q&As as objectively as possible and to enhance the benchmark's versatility, we select 9 major ex-

[*]Corresponding author.

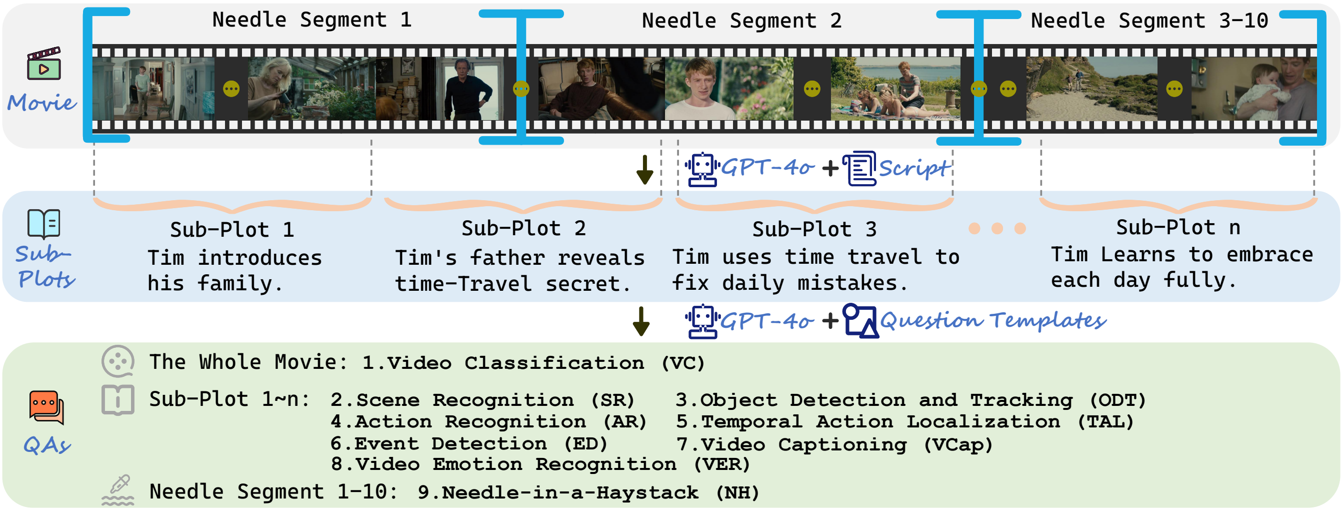
Needle Segment 1
Needle Segment 2
Needle Segment 3-10
Movie
GPT-4o + Script
Sub-Plots
Sub-Plot 1
Tim introduces his family.
Sub-Plot 2
Tim's father reveals time-Travel secret.
Sub-Plot 3
Tim uses time travel to fix daily mistakes.
Sub-Plot n
Tim Learns to embrace each day fully.
GPT-4o + Question Templates
QAs
The Whole Movie: 1.Video Classification (VC)
Sub-Plot 1~n: 2.Scene Recognition (SR) 3.Object Detection and Tracking (ODT)
4.Action Recognition (AR) 5.Temporal Action Localization (TAL)
6.Event Detection (ED) 7.Video Captioning (VCap)
8.Video Emotion Recognition (VER)
Needle Segment 1-10: 9.Needle-in-a-Haystack (NH)

Figure 1: The construction pipeline of ALLVB. Utilizing the powerful processing capabilities of GPT-4o, we first segment the movie into different sub-plots based on the corresponding script content. We then create Q&As for the entire video, each sub-plot, and evenly divided needle segments using 91 question templates. Note that needle segments do not correspond to the sub-plot segments.

isting video understanding tasks and expand them into 91 sub-tasks, designing corresponding question templates for each. These question templates comprehensively assess the MLLMs' abilities in summarization, information extraction, temporal reasoning, and more. GPT-4o then generates the final Q&As using these 91 templates and the video content.

To ensure the correctness of the generated Q&As, we implement strict quality control measures for ALLVB. First, during the plot segmentation phase, scripts are used to verify the continuity and completeness of the generated plots. In the Q&As construction phase, scripts ensure that the format of questions and answers complies with the designed rules. Finally, the paper's authors and recruited volunteers conduct a three-stage manual review, with details provided in the quality control section below.

The above provides an overview of the automated pipeline used to construct ALLVB. This pipeline distinguishes ALLVB from existing long video understanding benchmarks in several key aspects:

1. We design 91 question templates for 9 types of video understanding tasks, effectively integrating existing tasks into a comprehensive video Q&A framework. This approach thoroughly evaluates the MLLMs' ability to understand long videos, providing substantial practical value.

2. The benchmark leverages external text information and GPT-4o's powerful processing capabilities, utilizing a custom-designed, fully automated pipeline with rigorous quality control. This ensures its correctness, excellent scalability, and simplifies future maintenance.

3. ALLVB includes 1,376 long videos, averaging nearly two hours each, with a total of 252k Q&As, or 183 Q&As per video. In the field of long video understanding benchmarks, its average length is nearly 2× longer than the second longest benchmark, and its total number of Q&As is 11.5× greater than the second largest benchmark. From these two dimensions, ALLVB is the most comprehensive and largest long video understanding benchmark to date.

Related Work

Multi-model LLMs

Following the introduction of ChatGPT (OpenAI 2022), considerable effort (Yang et al. 2024) has been made to combine vision encoders with pre-trained LLMs to create MLLMs, enabling support for both textual and visual inputs. Image-based MLLMs such as Otter-I (Li et al. 2023a) and LLaVA-1.6 (Liu et al. 2024) use MPT (Team 2023) and Vicuna (Chiang et al. 2023) as base models, respectively, and incorporate CLIP's ViT-L/14 (Radford et al. 2021) for processing multiple image inputs. Video-based MLLMs, often built on Vicuna or LLaMA (Touvron et al. 2023), also utilize ViT (Dosovitskiy et al. 2020) encoders to handle video data, where each adopts different frame processing techniques: mPLUG-Owl-V (Ye et al. 2023) uses a set of learnable tokens to summarize visual information, MovieChat (Song et al. 2024) employs methods from ToMe (Bolya et al. 2022) to merge similar tokens between adjacent frames, and LLaMA-VID (Li, Wang, and Jia 2023) directly applies average pooling to reduce the number of image tokens.

Despite progress, MLLMs face challenges in long video comprehension, requiring expanded inferential capabilities. Models like MovieChat and LLaMA-VID, meant for long videos, have yet to demonstrate proven performance on hour-long videos. A reliable long video benchmark is needed for objective and accurate evaluation of MLLMs.

Video Understanding Benchmarks

Video understanding benchmarks assess LLMs' capabilities in video analysis through a range of tasks. Most, like MSVD-QA, MSRVTT-QA, TGIF-QA, and ActivityNet-QA, concentrate on short videos. They transform video captions into Q&As (MSVD-QA, MSRVTT-QA, and TGIF-QA) or derive Q&As from the Video Classification task (ActivityNet-QA), often using crowdsourcing, which is labor-intensive. Similarly, MovieQA (Tapaswi et al. 2016) is another movie-based short-form video benchmark that, like the others, is manually annotated and has an average length of just 203 seconds.

For long video understanding, there are some benchmarks such as MLVU, Video-MME, LVBench, and MoVQA (Zhang et al. 2023) (which also uses movies), that have manually annotated through various mechanisms: MLVU categorizes Q&As into three types: holistic LVU, single-detail LVU, and multi-detail LVU. Video-MME constructs Q&As from the perspectives of perception, reasoning, and information synopsis. LVBench builds Q&As based on six core capabilities: Temporal Grounding, Summarization, Reasoning, Entity Recognition, Event Understanding, and Key Information Retrieval. MoVQA classifies all Q&As into six types: Information Synopsis, Temporal Perception, Spatial Perception, Causal Reasoning, Hypothetical Reasoning, and External Knowledge.

We believe these approaches to constructing Q&As are reasonable, but they highlight a lack of a unified standard for constructing Q&As, and there is room for improvement in terms of video duration and size. We notice that these approaches closely resemble existing video understanding tasks. To avoid subjectively classifying Q&As, our approach, using established video understanding tasks to construct question templates and generating Q&As with GPT-4o, aims to create more objective and aligned Q&As, as compared in detail in Tab. 1.

ALLVB: All-in-One Long Video Understanding Benchmark

In this section, we will provide a detailed explanation of the benchmark construction pipeline and quality control measures and present the specific statistics of ALLVB. We will also compare these with existing benchmarks to highlight ALLVB's distinctiveness.

Benchmark Construction Pipeline

Data Collection and Cleaning. We initially scrape 2,520 scripts from Script Reader Pro[1] and SWN[2]. Through regularized detection of script names (e.g., names containing "episode" or season numbers indicating TV scripts), we identify 711 TV scripts and 1,809 movie scripts. Given that TV episodes are shorter in duration and often have plots spread across multiple episodes, we decide to temporarily exclude TV scripts from annotation.

For the remaining 1,809 movie scripts in PDF format, we first filter out abnormal files with very small sizes (<10KB), as these typically do not contain complete script content. We then use a script to extract text from PDFs. Following this, we further filter out scripts with too few text characters (<3,000 characters) and those containing an excessive number of non-text characters, such as some early movie scripts in blurry PDF formats. After these steps, we ultimately obtain 1,376 text-formatted movie scripts that meet our criteria.

Finally, we download 1,376 corresponding movies from YTS[3] and their respective subtitles from YTS and SUBDL[4].

Data Preprocessing: Plot Segmentation. Previous long video annotations involve humans watching the video and asking questions based on the content according to specified requirements. This approach limits the number of Q&As per video and often overlooks numerous details in long videos. To ensure that the Q&As generated by GPT-4o more thoroughly cover the details in the video, we instruct GPT-4o to first segment the movie script into different continuous plots based on its content. During this process, we discover that single-pass segmentation still results in plots that span considerable durations, so we adopt a two-stage segmentation strategy. In this strategy, sub-plots are further divided within the initially segmented plots, enabling finer-grained video segmentation and ensuring the correctness of the generated questions.

Question Template Design. The tasks addressed by the source videos of short video benchmarks and the tasks considered when constructing Q&As for long video benchmarks are very similar to existing video understanding tasks. Therefore, we design our question templates directly based on these established video understanding tasks. We have collected 9 common types of video understanding tasks suitable for conversion into Q&As, and these 9 tasks comprehensively cover the content assessed by other short video or long video understanding benchmarks, such as reasoning, summarization, recognition, and more. The 9 tasks include: *Video Classification (VC), Scene Recognition (SR), Object Detection and Tracking (ODT), Action Recognition (AR), Temporal Action Localization (TAL), Event Detection (ED), Video Captioning (VCap), Video Emotion Recognition (VER), and Needle-in-a-Haystack (NH) task for video*. For the 9 tasks, we have GPT-4o provide diverse and generalized sub-tasks. From these, we select 11, 12, 10, 11, 7, 12, 8, 10, and 10 non-overlapping sub-tasks that comprehensively assess various capabilities of MLLMs, further expanding the scope of the original tasks

For the "Needle-in-a-Haystack" (NH) task, a common approach is to insert content unrelated to the original video and then ask questions about it. For example, Gemini 1.5 Pro (Reid et al. 2024) overlays the text ”The secret word is needle” on a single randomly sampled video frame and then asks the LLM, ”What is the secret word?” However, advanced commercial models might recognize these as tests and refuse to answer. To address this, we ask detailed ques-

[1] https://www.scriptreaderpro.com

[2] https://www.screenwritersnetwork.org/script/?page=1

[3] https://ww4.yts.nz/

[4] https://subdl.com/

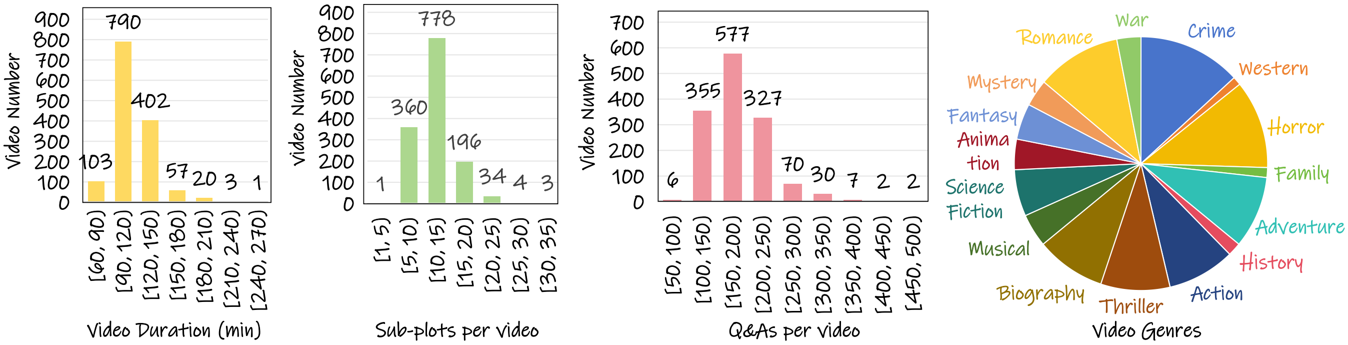

Figure 2: ALLVB Benchmark Statistics Chart. The distributions shown, from left to right, are **video duration**, **number of sub-plots** per video, **number of Q&As** per video, and **video genres**. Most videos are between 90-150 minutes in length, which is significantly longer than those in other benchmarks, highlighting the challenge of ALLVB. The majority of videos are divided into 5-20 sub-plots, resulting in most videos having 100-250 Q&As, showcasing the benchmark's comprehensiveness. Finally, our videos span 16 diverse genres, ensuring the benchmark's general applicability.

tions about the content of specific video frames, thereby avoiding this issue and preventing redundancy with the other 8 task types, such approach is more challenging and can provide a better assessment of the LLM's retrieval capabilities in long contexts. Every video is evenly divided into 10 segments, with each segment containing one sub-task that targets specific details in the frames.

Finally, we design 91 corresponding question templates based on these 91 sub-tasks.

Constructing Q&As. *1) Number of Q&As.* First, it is important to note that the Video Classification task applies to the entire video, while the Needle-in-a-Haystack task is applied to each evenly divided video segment. The remaining 7 tasks generate questions for each sub-plot of the video. To avoid excessive homogeneous questions in the benchmark, we randomly select 2 question templates from each of the 7 tasks to design questions for each sub-plot. Assuming a movie has n sub-plots, the corresponding number of Q&As would be: $11(VC) + 2 * 7 * n + 10(NH) = 21 + 14n$.

2) Questions. For the Video Classification task, we input the entire script content and question templates into GPT-4o, prompting it to design Q&As based on the requirements outlined in the prompt. For questions targeting sub-plots, we instruct GPT-4o to first describe the scene content of the sub-plot in the question, helping to locate the question within the video and reduce ambiguity. For Needle-in-a-Haystack questions, we evenly divide the video into segments and sample 11 continuous frames from a random position in each segment at a frame rate of 1 fps. GPT-4o then describes the scene corresponding to these 11 frames and generates detailed questions about the middle frame or a specific frame.

3) Answers. All Q&As are presented as multiple-choice questions, with each question offering 5 options, one correct answer, and 4 distractors. This format allows for easy calculation of accuracy during testing and eliminates the subjective judgment issues associated with open-ended answers.

To ensure the scientific validity of the options, we require that all options be of similar length. The incorrect options must be related to the video content, and the correct option should be randomly distributed among the incorrect ones. We also stipulate that the answers cannot be deduced simply by examining the question content alone. Additionally, GPT-4o is required to provide reasoning when generating the answers to support the deduction process and enhance the correctness of the output.

Quality Control. *During the plot segmentation phase*, we use regularization methods to verify that the start and end positions of the main plots and sub-plots align correctly, ensure the continuity of the sub-plots, and confirm that the sub-plots comprehensively cover all the script content. Additionally, dividing the script content into different sub-plots before designing questions can ensure the correctness of the generated Q&As.

During the Q&As construction phase, we also use regularization methods to determine whether all the questions are generated completely and whether there are any errors in the question content. For instance, if a question mentions the movie title, it could lead to information leakage. If a question related to a sub-plot lacks specific scene descriptions, or if a Needle-in-a-Haystack question references the frame number, these are considered invalid. We regenerate any questions that do not meet the requirements and conduct secondary verification. We also analyze the distribution of all the correct options and use a script to randomly adjust them. As a result, the proportion of correct options being A, B, C, D, and E is 18%, 20%, 21%, 19%, and 22%, respectively, which is close to an even distribution.

Manual reviewing. Although movie scripts generally align with the movie content, discrepancies can occur in certain details, which may affect the correctness of the answers. To address this issue, all the authors of this paper, along with 12 recruited volunteers, conduct a 3-stage manual review. **The first stage** focuses on identifying potentially problematic question types. We find that tasks related to video classification, scenes, events, and emotion analysis are almost entirely correct, as the script and movie content

Benchmarks	#Videos	Avg.Len.(min)	#Q&As	# Avg. Q&As	Q&As type	Anno.	Subs.
Short Video Understanding Benchmarks							
MovieQA (Tapaswi et al. 2016)	408	3.38	14,944	36	MC	M	✓
MSVD-QA (Xu et al. 2017)	1,970	0.16	50,505	25	OE	A	✗
MSRVTT-QA (Xu et al. 2017)	10,000	0.25	243,680	24	OE	A	✗
TGIF-QA (Jang et al. 2017)	71,741	0.05	165,165	2	OE&MC	A&M	✗
TVQA (Lei et al. 2018)	21,793	1.26	152,545	7	MC	M	✓
ActivityNet-QA (Yu et al. 2019)	5,800	3	58,000	10	OE	M	✗
How2QA (Li et al. 2020)	9,035	1	44,007	5	MC	M	✗
NExT-QA (Xiao et al. 2021)	5,440	0.73	52,044	9	OE&MC	M	✗
MVBench (Li et al. 2024)	3,641	0.26	4,000	1	MC	A	✗
CinePile (Rawal et al. 2024)	9,396	2.66	303,828	32	MC	A&M	✓
EgoSchema (Mangalam, Akshulakov, and Malik 2024)	5,063	3	5,063	1	MC	A&M	✗
Long Video Understanding Benchmarks							
MoVQA (Zhang et al. 2023)	100	16.53	21,953	**219**	MC	M	✓
MLVU (Zhou et al. 2024)	1,334	12	2,593	2	OE&MC	M	✗
Video-MME (Fu et al. 2024)	900	17	2,700	3	MC	M	✓
LVBench (Wang et al. 2024)	103	68	1,549	15	MC	M	✗
ALLVB	**1,376**	**114.62**	**252,420**	183	MC	A	✓

Table 1: Comparison with other benchmarks, where the abbreviations are defined as follows: **Avg.Len.** (Average length of each video), **Avg. Q&As** (Average number of Q&As per video), **OE** (Open-Ended questions), **MC** (Multiple-Choice questions), **Anno.** (Annotation Method), **A** (Automatic Annotation), **M** (Manual Annotation), **Subs.** (Subtitles). In the realm of long video benchmarks, ALLVB leads in terms of the number of videos, average video length, and the quantity of Q&As.

typically match. The issues primarily arise with object and action recognition tasks, such as identifying the color of an object, counting objects, or recognizing actions. **In the second stage,** we filter these questions from the benchmark and manually verify them against the movie content. For questions with discrepancies, we make corrections or regenerate them to ensure the correctness of the answers. **In the third stage**, we conduct another round of manual review to ensure that no errors are found. These rigorous quality control steps ultimately ensure the high quality of ALLVB. An example of using the pipeline to fully construct Q&As from a single video is shown in Fig. 1.

Benchmark Statistics

Through the GPT-4o automated benchmark construction pipeline, we collect a total of 1,376 high-quality movies and segment them into 15,966 sub-plots based on the script content, averaging 11.6 sub-plots per movie. Using these segmented sub-plots and entire movies, GPT-4o generates 252,420 Q&As according to 91 question templates, averaging 183 questions per movie. For more detailed data, please refer to Fig. 2. We also provide a detailed comparison with existing video understanding benchmarks in Tab. 1, and the following points are worth noting:

- In the realm of long videos, ALLVB features the highest number of videos, the longest average duration, and the most Q&As. This is made possible by our automated pipeline, which facilitates benchmark expansion and maintenance, significantly reducing labor costs.
- We select multiple-choice questions as the format to enable objective evaluation. Additionally, we provide subtitles for each video, allowing for questions that do not rely solely on video content. Many benchmarks avoid this issue by describing characters' physical features instead, which does not fully assess the model's ability to extract information from multiple inputs.

Experiments and Analysis

In this section, we test various MLLMs on ALLVB and thoroughly demonstrate the challenges of ALLVB through analysis of the experimental results.

Implementation Details

Settings. First, we divide ALLVB into a training set and a test set at a 9:1 ratio, containing 1,236 and 140 videos, respectively. The training set can be used for future pre-training or fine-tuning of other MLLMs. Then, we test various open-source and closed-source MLLMs on the test set. Among them, Otter-I (Li et al. 2023a), LLaVA-1.6 (Liu et al. 2024) and GPT4-Turbo (Achiam et al. 2023) are image MLLMs that support multiple image inputs. Otter-V (Li et al. 2023a), mPlug-Owl-V (Ye et al. 2023), LLaMA-VID (Li, Wang, and Jia 2023), VideoChat (Li et al. 2023b), VideoChat2 (Li et al. 2024), MovieChat (Song et al. 2024), TimeChat (Ren et al. 2024), GPT-4o (OpenAI 2024) and Claude 3.5 Sonnet (Anthropic 2024) are video MLLMs. Open-source models are run locally on an NVIDIA 4090, while closed-source models are accessed via official API.

To ensure fairness in testing and that each model can perform inference, all models receive 16 frames uniformly sampled across the entire video, along with the corresponding subtitles for these frames. LLM parameters are uniformly set to 7B and all tests are conducted in a 0-shot format. Due to the limitations in handling long contexts, open-source

Models	LLM	Video Understanding Tasks									Avg. Acc. (%)
		VC	SR	ODT	AR	TAL	ED	VCap	VER	NH	
Image MLLMs											
Otter-I (Li et al. 2023a)	MPT-7B	37.7	30.1	25.1	22.8	22.8	26.4	29.1	31.7	19.6	26.9
LLaVA-1.6 (Liu et al. 2024)	Vicunna-7B	68.4	51.3	39.0	43.6	43.3	46.1	56.2	57.3	32.3	48.5
GPT4-Turbo (Achiam et al. 2023)	GPT4	84.5	59.7	51.9	58.5	53.5	63.2	72.2	67.5	32.9	60.8
Video MLLMs											
Otter-V (Li et al. 2023a)	LLaMA-7B	25.6	23.6	22.3	22.8	25.9	24.4	22.7	20.0	19.6	23.1
mPlug-Owl-V (Ye et al. 2023)	LLaMA-7B	22.8	25.6	21.5	25.4	22.0	25.2	23.8	22.3	17.8	23.3
MovieChat (Song et al. 2024)	Vicunna-7B	26.2	25.6	23.0	23.8	23.0	26.7	25.1	24.6	20.9	24.4
VideoChat (Li et al. 2023b)	Vicunna-7B	35.2	31.8	26.3	29.4	29.5	30.9	31.6	34.3	23.2	30.4
VideoChat2 (Li et al. 2024)	Vicunna-7B	52.3	34.3	31.9	36.9	36.0	37.0	38.1	44.7	34.6	37.8
LLaMA-VID (Li, Wang, and Jia 2023)	Vicunna-7B	71.4	45.5	41.3	39.8	35.8	44.8	57.4	56.9	27.4	46.5
TimeChat (Ren et al. 2024)	LLaMA-2 7B	61.8	52.1	41.1	46.9	38.8	48.4	53.1	49.5	29.8	47.1
GPT-4o (OpenAI 2024)	GPT4	91.8	67.3	56.4	67.1	63.1	70.0	78.2	75.9	37.6	68.0
Claude3.5 Sonnet (Anthropic 2024)	Claude 3.5	**92.7**	**73.5**	**68.2**	**75.2**	**73.5**	**75.4**	**84.9**	**82.1**	**42.5**	**75.2**

Table 2: The test results of various MLLMs on ALLVB, including the accuracy for 9 types of video understanding tasks: **VC** (Video Classification), **SR** (Scene Recognition), **ODT** (Object Detection and Tracking), **AR** (Action Recognition), **TAL** (Temporal Action Localization), **ED** (Event Detection), **VCap** (Video Captioning), **VER** (Video Emotion Recognition), and **NH** (Needle-in-a-Haystack), as well as the **Avg. Acc.** (Average Accuracy).

models answer each Q&A in the video individually, while closed-source models answer all Q&As in a single video at once. Open-source and closed-source models each use the same prompts to ensure fairness. The specific prompt details are as follows:

- Prompts for open-source models:

  System Prompt: You are an expert at analyzing videos and their accompanying subtitles. Carefully observe the details in the video frames and their corresponding subtitles. Based on your observations, select the best option for the question provided.

  User Prompt:

  Frames from a video, their corresponding subtitles, and a multiple-choice question with five options have been provided. Your task is to select the best answer based on the information from the video frames and subtitles.

  Subtitles: {Insert the subtitles here}

  Question: {Insert the question here}

  Please strictly follow the format below for outputting your answers:

  Best Answer: [Correct option]

- Prompts for closed-source models:

  User Prompt: Video frames, corresponding subtitles, and {question_num} multiple-choice questions with five options each have been provided. Your task is to select the best answer based on the information from the video frames and subtitles.

  Subtitles: {Insert the subtitles here}

  Question: {Insert the question here}

  Please strictly follow the format below for outputting your answers:

  Question 1: [Correct option]

  Question 2: [Correct option]

  ...

  Question {question_num}: [Correct option]

Evaluation Metrics. We extract the answers output by each model and compare them with the correct answers to calculate the final accuracy. Despite strictly specifying the answer format in the prompts, we still encounter a variety of answer formats in the outputs. To address this, we manually create numerous regex-matching scripts tailored for different model outputs. For content without letter options, we calculated the similarity between the output text and the question options to identify the closest answer. We observe that closed-source models require significantly fewer regex rules compared to open-source models. For instance, GPT-4o only needed two rules to identify all output options, while models like mPlug-Owl-V and VideoChat required 34 and 26 regex rules, respectively, to match most output options. This suggests that closed-source models have much stronger instruction-following capabilities compared to open-source models.

Results and Analysis

We calculate the individual accuracy for each model across the 9 tasks, as well as the average accuracy. The detailed results can be found in Tab. 2. From this analysis, several key conclusions can be drawn:

- **Claude 3.5 Sonnet achieves the highest performance across all 9 tasks**, with GPT-4o consistently securing the second-best results. Although we are unable to run inference locally with a model that matches Claude's parameter size, these results highlight the superior capabilities of the Claude and GPT series in MLLMs.
- **There is a significant variation in results across different tasks.** Tasks such as VC (Video Classification), VCap (Video Captioning), and VER (Video Emotion Recognition) score noticeably higher. These tasks generally do not require extensive interaction with detailed visuals, different characters, objects, or environments, making them relatively less challenging. On the other hand, tasks like ODT (Object Detection and Tracking), TAL (Temporal Action Localization), and NH (Needle-in-a-Haystack) involve more detailed visual elements and logical interactions, such as object colors, object counts, and action sequences. These tasks are more difficult and test the model's comprehension and reasoning abilities more rigorously. The NH task, in particular, deals with

Models	Video Length Distribution (min)		
	[80, 90)	[105, 115)	[130, 140)
Otter-V	24.7	24.3 (*-0.4*)	22.5 (*-0.8*)
VideoChat2	38.5	38.0 (*-0.5*)	37.7 (*-0.3*)
LLaMA-VID	47.8	47.4 (*-0.4*)	45.9 (*-1.5*)
GPT-4o	70.3	66.9 (*-3.4*)	66.9 (*-0.0*)

Table 3: Model Accuracy at Different Video Lengths.

very fine-grained information in the video frames, leading to lower scores across all models.

- **Open-source video MLLMs do not show a significant advantage** over the image MLLMs trained on image data under the same model parameter size (7B) and input conditions. This could be partly due to the limited number of video frames used as input, but it also raises questions about whether most open-source video MLLMs have sufficient advantages in temporal modeling.
- **Even the best closed-source model, Claude 3.5 Sonnet, achieves an accuracy of only 75.2%.** One reason for this is that, due to the limitations of some open-source models, we standardize the input frame count to 16 to ensure fairness in comparison, which limits the amount of video information the models could access. Additionally, this result indicates that, whether open-source or closed-source, there is still significant room for improvement in retrieving and reasoning within ultra-long video contexts. This also highlights the challenges presented by ALLVB and its potential contributions to the future development of the video MLLMs community.

Further Discussion

Based on the analysis of the experimental results, we would like to further discuss the following two aspects:

The impact of video length. Although the videos in ALLVB are all at the hour-long level, we randomly select three video length distributions ranging from short to long to verify the impact of different video lengths on model accuracy. We then sample four baseline models to analyze their accuracy performance across these varying video lengths. The specific results are presented in Tab 3.

As the video length increases, all models exhibit varying degrees of accuracy decline. Since the input video frames are fixed at 16 frames, longer videos result in more information being missed by the models, making it more challenging to perform effective reasoning and test the models' ability to understand long contexts. We also observe that GPT-4o demonstrates better robustness with longer video durations, as its accuracy remains stable when transitioning from 105-115 minutes to 130-140 minutes. Because the overall video lengths in ALLVB are generally long, the increase in video length does not cause a particularly significant drop in accuracy. However, in general, the performance of multimodal LLMs still tends to decrease as the video length increases.

Models	Input	Avg. Acc. (%)
MovieChat	16 frames	24.4
	64 frames	26.0 (*+1.6*)
LLaMA-VID	16 frames	46.5
	64 frames	46.7 (*+0.2*)
TimeChat	16 frames	47.1
	64 frames	47.7 (*+0.6*)
GPT-4o	16 frames	68.0
	64 frames	68.5 (*+0.5*)

Table 4: Model accuracy with different numbers of input frames.

The impact of the number of input frames. Due to the limitations in context length of some open-source models, the comparison experiments in Tab. 2 uniformly set the number of input frames to 16. To assess the impact of increasing the number of input frames on model accuracy, we select four models that support longer context inputs. Since TimeChat does not support 128-frame input locally, we ultimately set the number of input frames to 64 and compare the accuracy with that of 16 frames. Detailed results are shown in Tab. 4.

As the number of input frames increases, the average accuracy of all models improves. This improvement is mainly due to the additional video information provided, which allows the models to capture more details and avoid incorrect answers caused by insufficient information. However, we also observe that with 64 frames, the accuracy only improves slightly. This is partly because 64 frames still miss a significant amount of information in hour-long videos, and partly because, although long-video MLLMs can support longer context inputs, they have not yet fully utilized this information, leaving room for further improvement.

Conclusion

In this paper, we design a comprehensive and challenging benchmark called ALLVB based on the 9 major existing video understanding tasks, combined with GPT-4o's automated long-video annotation pipeline. In the realm of long-video understanding benchmarks, ALLVB stands out with the largest number of videos, the longest average video duration, and the highest number of Q&As. Our tests on various MLLMs reveal that existing models still have significant room for improvement in long-video understanding. We hope this benchmark can serve as an objective evaluation metric for future MLLMs and contribute to the advancement of the entire video understanding community.

Limitations The videos in ALLVB are all sourced from movies, which already offer a diverse range of content. However, in the future, we hope to incorporate more types of videos, such as sports videos, documentaries, and others. Additionally, the current video Q&A content is still relatively simple compared to human capabilities. We are considering adding instructional videos, course videos, and other specialized content to assess MLLMs' understanding of professional knowledge. We will continue to maintain and expand ALLVB, striving to contribute further to the realization of true AGI.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62172155, 62402171, 62402505, 62472434), and the Science and Technology Innovation Program of Hunan Province (2022RC3061).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Claude 3.5 Sonnet. https://www.anthropic.com/claude/sonnet. Accessed: 2024-07-21.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2758–2766.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Yang, J.; and Liu, Z. 2023a. Otter: A Multi-Modal Model with In-Context Instruction Tuning. https://arxiv.org/abs/2305.03726. Accessed: 2024-07-21.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Li, Y.; Wang, C.; and Jia, J. 2023. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/. Accessed: 2024-07-21.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

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

Mao, Y.; Lin, X.; Ni, Q.; and He, L. 2024. BDIQA: A New Dataset for Video Question Answering to Explore Cognitive Reasoning through Theory of Mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 583–591.

OpenAI. 2022. ChatGPT. https://openai.com/chatgpt. Accessed: 2024-07-21.

OpenAI. 2024. GPT-4o: Generative Pre-trained Transformer 4 optimized. https://openai.com/index/hello-gpt-4o. Accessed: 2024-07-21.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rawal, R.; Saifullah, K.; Basri, R.; Jacobs, D.; Somepalli, G.; and Goldstein, T. 2024. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Ren, S.; Yao, L.; Li, S.; Sun, X.; and Hou, L. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14313–14323.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4631–4640.

Team, M. N. 2023. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs. https://www.databricks.com/blog/mpt-7b. Accessed: 2024-07-21.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Huang, S.; Xu, B.; Dong, Y.; Ding, M.; et al. 2024. LVBench: An Extreme Long Video Understanding Benchmark. *arXiv preprint arXiv:2406.08035*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, 1645–1653.

Yang, X.; Jing, H.; Zhang, Z.; Wang, J.; Niu, H.; Wang, S.; Lu, Y.; Wang, J.; Yin, D.; Liu, X.; et al. 2024. DaRec: A Disentangled Alignment Framework for Large Language Model and Recommender System. *arXiv preprint arXiv:2408.08231*.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9127–9134.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Zhang, H.; Liu, Y.; Dong, L.; Huang, Y.; Ling, Z.-H.; Wang, Y.; Wang, L.; and Qiao, Y. 2023. Movqa: A benchmark of versatile question-answering for long-form movie understanding. *arXiv preprint arXiv:2312.04817*.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Xiao, S.; Yang, X.; Xiong, Y.; Zhang, B.; Huang, T.; and Liu, Z. 2024. MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding. *arXiv preprint arXiv:2406.04264*.