# Towards Event-oriented Long Video Understanding

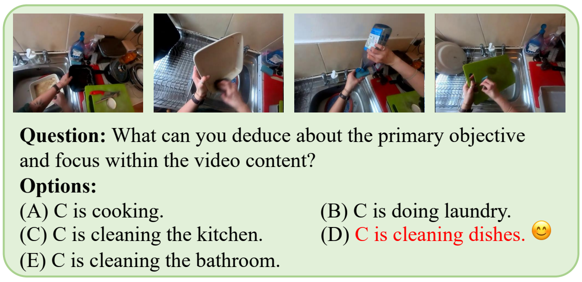
Anonymous ACL submission

## Abstract

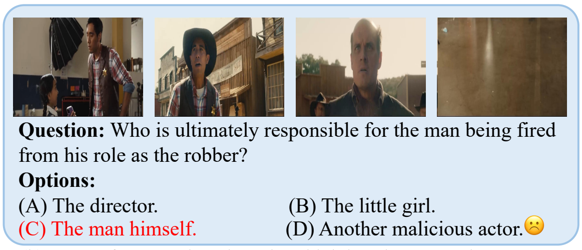
With the rapid development of video Multimodal Large Language Models (MLLMs), a surge of evaluation datasets is proposed to evaluate their video understanding capability. However, due to the lack of rich events in the videos, these datasets may suffer from the short-cut bias that the answers can be easily deduced by a few frames, without watching the entire video. To address this issue, we construct an event-oriented long video understanding benchmark, ***Event-Bench***, building upon existing datasets and human annotations. The benchmark includes six event-related tasks and a total of 2,190 test instances to comprehensively evaluate the capability to understand video events. Additionally, we propose ***Video Instruction Merging (VIM)***, a low-cost method to enhance video MLLMs by using merged event-intensive video instructions, aiming to overcome the scarcity of human-annotated, event-intensive data. Extensive experiments show that the best-performing GPT-4o achieves an overall accuracy of 53.33, significantly outperforming the best open-source model by 15.62. Leveraging the effective instruction synthesis method and model architecture, our VIM outperforms both state-of-the-art open-source video MLLMs and GPT-4V on Event-Bench. All the code, data, and models will be publicly available.

## 1 Introduction

Video understanding stands as the key capability of AI models to perceive the visual world like humans. It requires models to recognize the features and changes in regions or objects, and to understand the overall context and storyline throughout the video. Building upon Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Zhao et al., 2023), current Video Multimodal Large Language Models (Video MLLMs) (Tang et al., 2023; Zhang et al., 2023; Maaz et al., 2023) exhibit surprising video understanding capabilities. Concurrently, a surge of benchmarks are proposed to evaluate their performance in different video understanding scenes, *e.g.,* contextual reasoning (Mangalam et al., 2023) and situated reasoning (Wu et al., 2021).




(a) A case from EgoSchema, which involves a single event.




(b) A case from our benchmark, which involves complex events.

Figure 1: The comparison of two representative examples from existing benchmarks and our Event-Bench.

Despite these advancements, recent work has found that these datasets may suffer from the short-cut bias (Lei et al., 2023). It refers to the fact that the answers to part of the questions could be accurately deduced without fully reading the video, which would affect the evaluation reliability. As shown in Figure 1 (a), although the video lasts for 3 minutes, it simply describes the behavior of cleaning dishes. Therefore, questions related to the video can be easily answered by viewing just a single frame. Essentially, the cause of the short-cut bias is the *lack of rich events* in the video. Events are the high-level semantic concepts that humans perceive when observing a video (Lavee et al., 2009) (*e.g.,* the moment a player makes a shot in a soccer match), which are crucial to represent the unique and dynamic insights that differentiate various videos. Since the necessity of event-oriented video understanding might be neglected in existing datasets, their annotated test instances may fail to accurately estimate human-like video understanding capability.

In light of this, we present an event-oriented long video understanding benchmark, namely ***Event-Bench***. It focuses on comprehensively evaluating video MLLMs from three levels of event understanding capabilities, *i.e.,* atomic, composite, and overall understanding, totally consisting of six event-related tasks. To construct it, we design a low-cost automatic pipeline to meticulously collect unbiased test instances corresponding to the above tasks from existing datasets, then unify their format and filter low-quality ones. Additionally, we also manually craft multiple test instances based on the event-intensive long videos from YouTube, to improve the coverage of our benchmark on complex real-world scenarios. Totally, Event-Bench contains 2,190 samples. As shown in Table 1, our benchmark distinguishes itself with longer time scopes and an event-oriented focus.

To elicit the capability of human-like video understanding, it is necessary to utilize massive event-intensive video instruction for training video MLLMs (Chen et al., 2024c). However, it is costly to annotate sufficient high-quality video instructions with rich events. To solve it, we aim to make use of existing image instructions and simple video instructions, to compose more complex training data. Concretely, we first employ an adaptive model architecture to handle both image and video inputs, enabling us to add high-quality image instructions for training. Second, we propose to merge several similar video instructions from existing datasets into a new one, which contains all the events from them and are also longer and more complex. We conduct extensive experiments on our benchmark, and the results show that our method can perform better than all open-source models of comparable parameter scales, even outperforming GPT-4V on average (*i.e.,* 41.64 VS. 32.65).

Our main contributions are listed as follows:

(1) We propose an event-oriented long video benchmark, Event-Bench, to evaluate the human-like video understanding capability;

(2) We devise VIM, a low-cost method to improve video MLLMs using merged event-intensive video and high-quality image instructions;

(3) Experiment results show the comprehensive

| Benchmark | Time Scope (s) | Open Domain | Complex Reasoning | Event Oriented |
|---|---|---|---|---|
| MSVD-QA | 0∼60 | ✓ | ✗ | ✗ |
| MSRVTT-QA | 10∼30 | ✓ | ✗ | ✗ |
| TGIF-QA | - | ✓ | ✗ | ✗ |
| ActivityNet-QA | 0∼975 | ✗ | ✗ | ✗ |
| NeXT-QA | 5∼180 | ✓ | ✗ | ✗ |
| STAR | 2∼195 | ✓ | ✓ | ✗ |
| CLEVRER | 5 | ✗ | ✓ | ✗ |
| EgoSchema | 180 | ✗ | ✓ | ✗ |
| MVBench | 5∼40 | ✓ | ✓ | ✗ |
| TempCompass | 0∼35 | ✓ | ✗ | ✗ |
| MovieChat | 401∼602 | ✓ | ✗ | ✗ |
| VIM | 2∼1088 | ✓ | ✓ | ✓ |

Table 1: Comparing our Event-Bench with existing video benchmarks. Event-Bench stands out due to the longer time scope and event-oriented design. The details are in the Appendix.

evaluation capability of Event-Bench for video MLLMs and the effectiveness of VIM.

## 2 Related Work

### 2.1 Video Multimodal Large Language Model

Building upon the Large Language Model (LLM), Multi-modal Large Language Models (MLLMs) have recently obtained notable progress. Among them, Video MLLMs exhibit surprising performance on various tasks (Zhang et al., 2023; Maaz et al., 2023; Ren et al., 2023). Typically, a Video MLLM consists of a video encoder (or image encoder), a LLM, and a connector to bridge these two components (Zhang et al., 2023; Li et al., 2023b; Maaz et al., 2023). Based on this type of architecture, the following works explore several ways to enhance the Video MLLMs, *e.g.,* utilizing a more powerful video encoder (Lin et al., 2023), supporting long context video (Song et al., 2023; Wang et al., 2024), and fine-tuning with large-scale instructions (Li et al., 2023c). In this work, we aim to synthesize video instructions with more complex events and explore scalable model architecture.

### 2.2 Video Understanding Benchmark

Previous works propose benchmarks to evaluate various reasoning abilities in videos, including temporal reasoning (Xiao et al., 2021), situated reasoning (Wu et al., 2021), compositional reasoning (Grunde-McLaughlin et al., 2021), *etc.* However, most videos in these benchmarks are short clips and lack diversity. With the development of Video MLLMs, several works collect diverse videos to evaluate these models comprehensively (Ning et al., 2023; Chen et al., 2023), but most videos in these benchmarks are no more than 1 minute. Following works like Egoschema (Mangalam et al., 2023) and MovieChat (Song et al., 2023) collect long videos and create questions based on them. Despite this, the videos and questions in these benchmarks either do not involve complex reasoning in the event or are not open-domain. Therefore, we present an event-oriented long video understanding benchmark with diverse videos to comprehensively evaluate the model's ability to understand complex event narratives.

## 3 Event-oriented Benchmark

We propose Event-Bench, an event-oriented long video understanding benchmark for evaluating existing video MLLMs. It consists of massive videos, each paired with multi-choice questions from various event-related sub-tasks. Thus, we first establish a hierarchical task taxonomy for our benchmark and collect the data according to it.

### 3.1 Hierarchical Task Taxonomy

We organize our benchmark into three categories according to the number of events in a video, each of which comprises several sub-tasks.

**Atomic Events Understanding.** This task aims to evaluate the model's understanding of an atomic event (*e.g.,* an action of a human or object) in the video, which is one of the most basic video understanding capabilities.

• *Event Description.* For this sub-task, we collect question-answering pairs to evaluate whether the model can accurately recognize and describe a specific atomic event in the video, *e.g.,* "*What did the person do with the towel?*"

**Composite Events Understanding.** It focuses on understanding the relation between two atomic events in a video, from the following two aspects.

• *Temporal Reasoning.* We collect question-answer pairs that require to perform reasoning based on the understanding of the temporal order for two events in the video, *e.g.,* "What did the man do after putting down the towel".

• *Causal Reasoning.* This sub-task focuses on the casual relation between two events in the video, especially for explaining the reason why an event happened, *e.g.,* "Why did the man open the box".

**Overall Understanding.** It requires understanding the relations across all events in the video, to

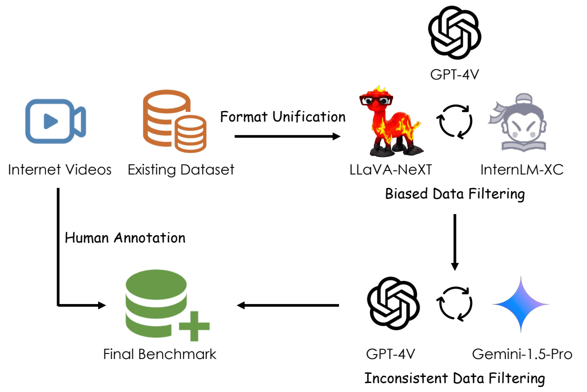



Figure 2: The data in Event-Bench are sourced from existing datasets or human annotations, involving three stages: format unification, biased data filtering, and inconsistent data filtering.

| Atomic | Composite | | Overall | | | Total |
|---|---|---|---|---|---|---|
| ED | TR | CR | CIR | CU | ER | |
| 468 | 400 | 400 | 227 | 395 | 300 | 2190 |

Table 2: The statistic of Event-Bench. Each header is the abbreviation of the corresponding sub-tasks.

capture the high-level overall information from it. We design the following three sub-tasks:

• *Contextual Reasoning.* This sub-task aims to perform reasoning based on the overall context in the video, where the model needs to summarize the content from a series of events, *e.g.,* "Describe the overarching process is conducting in the lab".

• *Episodic Reasoning.* For a video, we also consider its contained episodes (*i.e.,* stories) about the characters and objects across all the events, where the model need to characterize high-level semantics to answer complex questions, *e.g.,* "What led to Bean deciding to quickly leave the restaurant".

• *Counter-intuitive Reasoning.* For this sub-task, the videos involve counter-intuitive elements (*e.g.,* magical spells), and the model needs to identify the abnormal details to answer corresponding questions, *e.g.,* "Why the video is magical".

### 3.2 Data Construction

Our benchmark consists of data collected from existing datasets and newly human-annotated internet videos. The overall construction process is illustrated in Figure 2.

#### 3.2.1 Construction Based on Existing Datasets

As there exist multiple open-source VideoQA datasets, we aim to collect useful instances from them to compose our event-oriented benchmark.

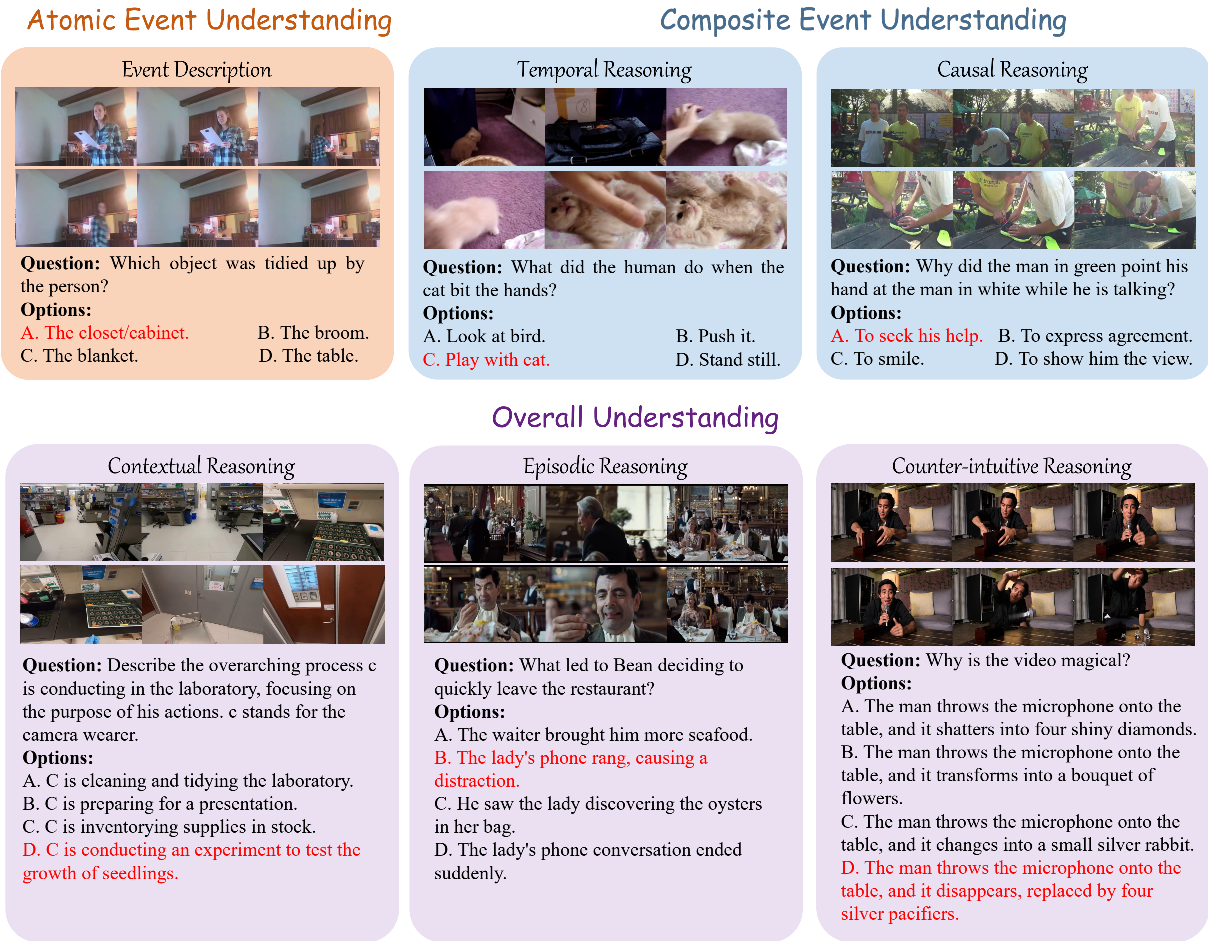



Figure 3: Overview of our Event-Bench. Our benchmark includes six sub-tasks across three event understanding abilities: atomic event understanding, composite event understanding, and overall understanding. The ground-truth answer is highlighted in red.

Specifically, we select the instances from four datasets, *i.e.,* STAR (Wu et al., 2021), NeXT-QA (Xiao et al., 2021), EgoSchema (Mangalam et al., 2023), and FunQA (Xie et al., 2023), owing to their diverse domains and rich annotations. However, after human review, we find three key issues in these instances: (1) different data formats and evaluation settings; (2) biased short-cut questions requiring no video understanding; (3) inconsistency between the answers and the video content. To address them, we develop the corresponding three-stage pipeline to preprocess the data.

**Format Unification.** We first convert all open-ended questions into multi-choice questions using GPT-4, where the prompt is *"Please change this task into a 4-way multi-choice question based on their descriptions"*. The generated questions are further examined and revised by human annotators.

**Biased Data Filtering.** Inspired by existing work (Chen et al., 2024b), we filter the short-cut questions that can be answered by only a single frame of the video, which are biased test data for evaluating video understanding capability. Concretely, we employ three Image-based MLLMs (*i.e.,* GPT-4V (OpenAI, 2023), LLaVA-NeXT-34B (Liu et al., 2024a), InternLM-XComposer2-4kHD (Dong et al., 2024)) on collected data and remove those can be accurately answered using only one frame. Such a way can leverage the short-cut bias to identify and remove the biased data.

**Inconsistent Data Filtering.** Finally, given the video and question from an instance, we utilize two powerful MLLMs, *i.e.,* GPT-4V and Gemini-1.5-Pro[1] to produce the answers. If their answers are the same but different from the human-annotated one, we regard the instance as an inconsistent sample and filter it out.

[1]We sample 16 frames for GPT-4V and 1fps for Gemini-Pro-1.5 as the representation of the video.

#### 3.2.2 Annotation Based on Internet Videos

Although the processed instances from existing datasets are diverse and high-quality, we find that their videos generally contain relatively fewer events and their questions mostly neglect the episodic reasoning capability, which is important for testing the understanding capability of the overall video storyline. Therefore, we collect multiple videos from YouTube, whose storylines contain rich body language information, and then annotate questions and answers for the episodic reasoning task. Considering the complexity of the episodic reasoning task, we decompose its annotation process into three stages to simplify it: caption annotation, question generation, and answer check.

**Caption Annotations.** We ask human annotators to write the captions for every 30 seconds of a video. To ensure the quality, we first utilize Gemini-Pro-1.5 and GPT-4 to synthesize 10 questions per video, and ask human annotators to answer the questions by writing detailed captions. Note that the synthetic questions may contain errors, yet can still guide the whole annotation process to control the quality.

**Question Generation.** To reduce the human annotation cost, we utilize GPT-4 to generate the question-answer pairs for the episodic reasoning task, according to the annotated captions. We utilize the following prompt with detailed guidelines (in Appendix) to guarantee their consistency with the captions: *"Based on the following descriptions, please ask 10 diverse questions about the plot and events of the video. While executing this task, please adhere to the following guidelines: ..."*

**Answer Check.** We ask human annotators to answer the generated question without giving the corresponding answer generated by GPT-4, and then compare their answers for checking. If they are the same, we add them to our benchmark. Otherwise, we invite more human annotators to check the question and vote on the final answer. Note that we also ask the human annotators to select the time interval in the video that corresponds to the question-related event, which is also used to estimate the annotation reliability.

### 3.3 Data Statistics

Our benchmark comprises a total of 2,190 video question-answer pairs on 6 tasks corresponding to different event understanding abilities, where each task has 172∼400 test samples for evaluation.

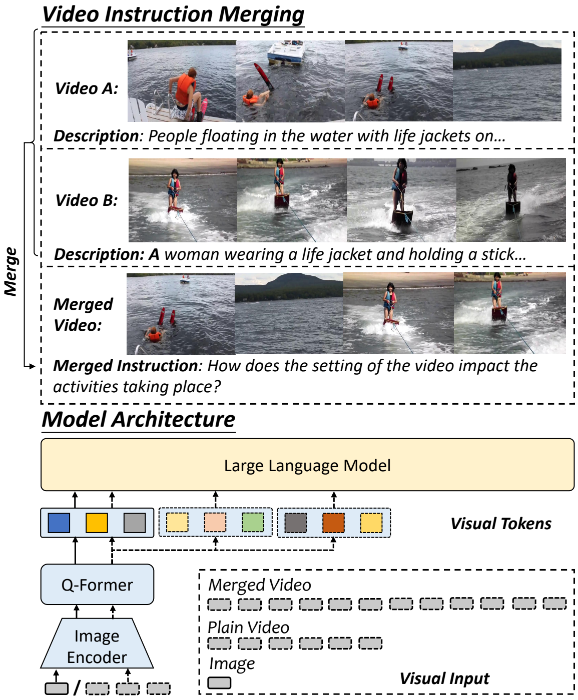



Figure 4: Overview of our method. We devise an instruction merging strategy to obtain instructions with more events based on existing data, and employ an adaptive model architecture supporting both image and video as the input.

Owing to the hierarchical task taxonomy, we can freely estimate the capability of models at different levels. Besides, as the benchmark is built based on diverse data sources, its contained videos can well cover the diverse domains in the real world and own varying lengths. These characteristics enable our benchmark to provide a comprehensive evaluation of existing video MLLMs. We show the cases in our benchmark in Figure 3.

## 4 Methodology

In this section, we introduce Video Instruction Merging (VIM) to enhance the performance of video MLLMs on event-oriented long video understanding tasks. Previous approaches primarily utilize video instruction tuning (Li et al., 2023b; Maaz et al., 2023; Zhang et al., 2023), which typically require extensive human effort to annotate massive video instructions. To address this, our proposed VIM integrates several similar video instructions from existing datasets into a new event-intensive one as additional training data. We also adopt a scalable visual processor in our video MLLM that interprets video as sequences of images, thereby handling both image and video inputs. This architecture allows us to combine existing high-quality image instructions with the newly created merged video instructions for training. The overall architecture of our approach is illustrated in Figure 4.

### 4.1 Video Instruction Merging

Existing video instruction datasets suffer from the issues of lacking rich events (Heilbron et al., 2015), *e.g.,* 1.41 on average for Video-ChatGPT-100K (Maaz et al., 2023). Thus, inspired by the mix-up strategy (Zhang et al., 2018), we propose to merge several simple video instructions to obtain a complex one with more events. Concretely, for each video and its corresponding instruction, we first find its most similar ones and then merge them into a new sample.

**Similar Video Selection.** We select the most similar video instructions to merge, to ensure the coherence of the synthetic new one. Specifically, we concatenate the input question and answer into one sentence $[q_i; a_i]$, and convert it into the text embedding $\mathbf{h}_i$ via state-of-the-art BGE model (Chen et al., 2024a). Then, the embedding is regarded as the semantic representation of the whole instruction, and we compute its cosine similarity to other instructions for selecting the $k-1$ nearest ones:

$$\text{Cos}(i, j) = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{|\mathbf{h}_i| * |\mathbf{h}_j|}. \quad (1)$$

In this way, we can divide the entire video instruction dataset $\mathcal{D}$ into $|\mathcal{D}|/k$ subsets.

**Instruction Merging.** For instructions within each similar video subset $\{v_i, q_i, a_i\}_{i=1}^{k}$, we merge them into a new one. We first temporally concatenate every video as a new one $v'$, then ask ChatGPT [2] to generate a new question $q'$ and answer $a'$ for the merged video given their original questions and answers. The process can be formulated as:

$$\begin{aligned} v' &= [v_1; v_2; \dots; v_k], \\ q', a' &= \text{ChatGPT}(p_m, q_1, \dots, a_1, \dots), \end{aligned} \quad (2)$$

where $[;;]$ is the concatenation process and $p_m$ is the prompt for ChatGPT.

[2]https://chatgpt.com/

**Prompt for Instruction Merging**

The user will give you $k$ question-answer pairs about a video. These pairs have similar semantics but are different in some details. Your task is to create a new question-answer pair based on them, which requires the tester to watch all the videos to answer. The new question should be about the similarities and differences among these videos. The question should be diverse and the corresponding answer should be as detailed as possible...

### 4.2 Adaptive Model Architecture

Our model architecture is composed of a scalable visual processor and an LLM. The scalable visual processor consists of a reusable image encoder and a cross-modal connector. For video input, we first uniformly sample $n$ frames from it, then separately feed them into the visual processor and concatenate the result visual tokens as the video representations, while image input is treated as in regular Image MLLMs. Therefore, our model can flexibly handle inputs of different sequence lengths (*e.g.,* a single image, short videos, or long videos).

In practice, we adopt EVACLIP (Fang et al., 2023) as the image encoder. For the cross-modal connector, we adopt a pre-trained Q-Former (Li et al., 2023a) to reduce the number of resulting visual tokens of input videos. The visual tokens are then concatenated with the embedding of question $q$ as the input of the LLM:

$$\text{LLM}([\mathbf{H}_{f_1}, \dots, \mathbf{H}_{f_n}; \mathbf{e}_1, \dots, \mathbf{e}_L]), \quad (3)$$

where $[\mathbf{H}_{f_1}, \cdots, \mathbf{H}_{f_n}]$ are the visual tokens and $[\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_L]$ are the text tokens. Since our model can handle both image and video inputs, we also add some high-quality image instructions to our training data, which helps the LLM better align with and understand the visual input.

## 5 Experiment

### 5.1 Experimental Setup

**Implementation Details.** We utilize EVA-CLIP (Fang et al., 2023) as the image encoder, Vicuna-v1.1 (Chiang et al., 2023) as the LLM, and initialize the Q-Former from InstructBLIP (Dai et al., 2023). We extrapolate the maximum length of Vicuna-v1.1 from 2,048 to 4,096 so that it can receive 64 frames as the input. As for the training data, we utilize 100K instructions from Video-ChatGPT (Maaz et al., 2023), 40K instructions from Something-Something-2 (Goyal et al., 2017), 34K instructions from NExT-QA (Xiao

| | Atomic | Composite | | | Overall | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| | Event Description | Temporal Reasoning | Causal Reasoning | Avg. | Counter Reasoning | Contextual Reasoning | Episodic Reasoning | Avg. | |
| *Open-Source Image MLLMs* | | | | | | | | | |
| LLaVA-NeXT (7B) | 13.68 | 14.75 | 9.75 | 12.25 | 14.98 | 9.11 | 7.30 | 9.97 | 11.59 |
| IXC2-4KHD (7B) | 26.07 | 27.50 | 32.50 | 30.00 | 9.25 | 12.15 | 17.67 | 13.23 | 22.10 |
| *Open-Source Video MLLMs* | | | | | | | | | |
| LLaMA-VID-long (7B) | 0.21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 |
| LLaMA-VID (13B) | 1.92 | 1.75 | 0.00 | 0.88 | 3.08 | 0.00 | 4.00 | 2.06 | 1.60 |
| Video-LLaVA (7B) | 12.82 | 5.50 | 0.00 | 2.75 | 6.17 | 2.78 | 7.20 | 5.05 | 5.87 |
| Video-LLaMA (7B) | 15.81 | 9.00 | 6.25 | 6.63 | 0.09 | 2.28 | 0.67 | 1.22 | 6.68 |
| Video-ChatGPT (7B)* | 9.83 | 9.50 | 15.00 | 12.25 | 14.98 | 12.66 | 10.00 | 12.37 | 11.78 |
| MovieChat (7B)* | 16.88 | 16.00 | 14.50 | 15.25 | 18.06 | 13.16 | **20.33** | 16.70 | 16.21 |
| PLLaVA (7B) | 34.62 | 40.00 | 40.50 | 40.25 | 17.62 | 15.19 | 11.00 | 14.42 | 28.17 |
| VideoChat2 (7B) | 33.76 | 37.75 | 47.75 | 42.75 | 16.74 | 15.70 | 14.67 | 15.62 | 29.41 |
| PLLaVA (13B) | 39.53 | 42.50 | 43.00 | 42.75 | **25.56** | 22.78 | 17.00 | 21.58 | 33.15 |
| ST-LLM (7B) | 47.22 | 48.75 | 59.50 | 54.13 | 9.69 | 25.32 | 16.67 | 18.66 | 37.71 |
| VIM (7B) (Ours) | **48.08** | **51.25** | **61.25** | **56.25** | 22.91 | **32.66** | 18.67 | **25.71** | **41.64** |
| *Proprietary MLLMs* | | | | | | | | | |
| GPT-4V | 29.70 | 35.00 | 40.00 | 37.50 | 36.56 | 28.35 | 27.00 | 29.93 | 32.65 |
| Gemini-1.5-Pro | 48.50 | 47.50 | 41.75 | 44.63 | 52.86 | 32.15 | **38.67** | 39.37 | 43.24 |
| GPT-4o | **54.27** | **56.75** | **58.25** | **57.5** | **63.44** | **50.13** | 37.33 | **49.24** | **53.33** |

Table 3: Experiment results on Event-Bench. For the Image MLLMs, we extract the frame in the middle of the video as the input. For the Video MLLMs, we uniformly sample $\{8, 16, 32\}$ frames as the input and report the best performance. *Video-ChatGPT samples 100 frames, while MovieChat samples 1fps from the video.

et al., 2021), 10K from Vript Caption (Yang, 2024), 100K visual instructions randomly sampled from LLaVA665K (Liu et al., 2023a), and 32K instructions synthesized in Section 4.1. In the training process, we freeze the image encoder and the Q-Former, only updating the parameters of the LLM. We train our model on 8 Nvidia A100 (80G) GPUs for 1 epoch and complete within 12 hours.

**Baseline Models.** We select several SOTA MLLMs as baselines. For open-source models, we select 2 Image MLLMs (LLaVA-NeXT (Liu et al., 2024a) and InternLM-XComposer2-4kHD (Dong et al., 2024)) and 7 Video MLLMs (Video-LLaMA, Video-ChatGPT (Maaz et al., 2023), MovieChat (Song et al., 2023), LLaMA-VID (Li et al., 2023d), VideoChat2 (Li et al., 2023c), Video-LLaVA (Lin et al., 2023) and ST-LLM (Liu et al., 2024b)). For proprietary models, we select GPT-4o, Gemini-1.5-Pro (Reid et al., 2024), and GPT-4V (OpenAI, 2023).

**Evaluation Protocols.** We follow the evaluation strategy proposed in MMBench (Liu et al., 2023b) to evaluate these models. Specifically, we first use regular expression to extract the options from the model's response. If successful, we use this as the prediction and compare it with the ground truth.

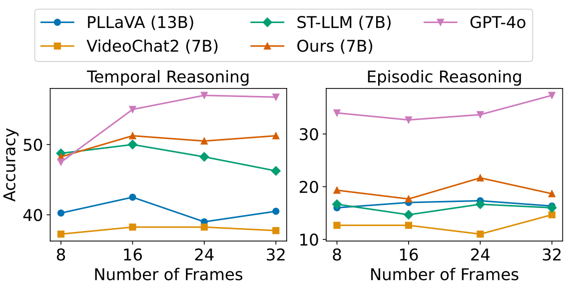



Figure 5: The relationship between the performance and the number of input frames.

Otherwise, we utilize GPT-4-turbo to judge if the prediction is correct. Besides, to ensure the consistency of models' responses on multiple choice questions, we adopt the circular evaluation strategy (Liu et al., 2023b). Specifically, we ask the models each question $N$ ($N$ is the number of choices) times and only consider the answer correct if the models provide the correct answer in every round.

### 5.2 Main Results

The performance of the models is illustrated in Table 3. We discuss the result and present the key findings from the following perspective:

**Overall Performance.** As is shown in Table 3, both Image MLLMs and Video MLLMs exhibit poor performance on these event reasoning tasks. For the Image MLLMs, LLaVA-NeXT and InternLM-XComposer2-4kHD could not achieve satisfying performance conditioned on only one frame, which proves the effectiveness of our data filtering strategies in building our benchmark. Surprisingly, most Video MLLMs even underperform these two Image MLLMs, implying their weak ability to understand complex events in the videos. From the perspective of task, we can observe that overall understanding is more challenging than composite event understanding and atomic event understanding. Especially in our newly annotated episodic reasoning task, the most powerful Gemini-1.5-Pro and GPT-4o only achieve 38.67 and 37.33.

**Comparisons of Different Models.** From the perspective of model, most open-source models obtain comparable performance as the proprietary models in the atomic and composite understanding tasks, with some models even outperforming GPT-4V (*e.g.,* ST-LLM, PLLaVA, and VideoChat2). However, the gap is enlarged in the overall understanding task, where all the open-source models lag behind the proprietary models. Among the open-source models, our model achieves the best performance across almost all the tasks. The only exception is that MovieChat achieves the best on the episodic reasoning task and PLLaVA (13B) is slightly better than ours on the counter-intuitive reasoning task. This is because MovieChat samples more frames and PLLaVA (13B) utilizes a larger LLM and more training data. However, our model still obtains the best accuracy on average.

### 5.3 Analysis

**Number of Frames.** Due to the limit of context length in LLMs, most video MLLMs sample frames from the whole video uniformly as the input. Intuitively, increasing the number of frames would help the model better understand the video, thus achieving better performance. We select the best four open-source models and one proprietary model and display the relationship between their performance and the number of input frames in Figure 5. We can observe that more input frames lead to better performance for GPT-4o. For example, the performance of GPT-4o in the temporal reasoning task is boosted from 47.50 to 56.75 when the number of input frames increases from 8 to 32. However, the open-source models do not always benefit from more input frames. Most models achieve the best performance when given 16 or 24 frames while increasing to 32 frames will lead to performance degradation. As a comparison, VIM is still boosting when the number of frames increases from 16 to 32, demonstrating its scalability.

| | Atomic | Composite | Overall | Avg. |
|---|---|---|---|---|
| Ours | **48.08** | **56.25** | **25.71** | **41.64** |
| - w/o mixup | 43.16 | 51.63 | 24.39 | 38.90 |
| - w/o image | 46.15 | 51.75 | 24.08 | 38.90 |
| - random merge | 45.94 | 54.25 | 25.38 | 40.32 |

Table 4: Ablation study of VIM on Event-Bench.

**Training Strategy.** We study the effect of the instruction merging strategy and the benefit of adding image data in our training process. First, the result in Table 4 shows that removing the merging strategy significantly hurt the performance on all tasks. Secondly, selecting videos with similar semantics leads to better performance than random selection, which highlights that the coherence of events in a video is quite important. As for the effect of image data, we could observe that removing image instructions from our training data causes a performance decrease on all the tasks. This not only shows that image instruction could compensate for the lack of high-quality video data, but also demonstrates the compatibility and scalability of our model architecture.

## 6 Conclusion

In this work, we built an event-oriented long video understanding benchmark based on existing datasets and human annotation, namely Event-Bench. We created six event-related tasks, and collected totally 2,190 test instances in Event-Bench to comprehensively evaluate the capability of understanding events within the videos. Then, we devised an efficient training strategy to improve video MLLMs to alleviate the problems of lacking human-annotated event-intensive video instructions. We revised the model architecture to support using high-quality image-based instruction, and merged several simple video instructions into an event-intensive new one, to extend our training dataset. Extensive experiments have shown that our Event-Bench can provide a systematic comparison across the different kinds of capabilities for existing video MLLMs, and point out the major shortcomings of open-source MLLMs. Besides, our approach can outperform state-of-the-art open-source video MLLMs on average, even GPT-4V.

## 7 Limitation

First, events are not only represented by visual modality, but also by other modalities in the real world(*e.g.,* textual, audio, and speech). They convey important information in the video and complement the visual modality. As an initial exploration, we only consider the visual modality in Event-Bench. In the future, we will also add other modalities to our benchmark. Second, we only use 500K video instructions during training the Video MLLM due to the limited computation resources. However, the experimental results show that including more high-quality video instructions and image instructions has a positive impact on the model performance. In the future, we will scale the training data and model size to obtain better performance. Third, although the method we propose to merge video instructions is low-cost and effective, the quality is still lower than human annotations. In the future, we will construct more event-intensive training data through human annotation.

## References


Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *CoRR*, abs/2402.03216.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are we on the right way for evaluating large vision-language models? *CoRR*, abs/2403.20330.

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

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2023. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *CoRR*, abs/2311.14906.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.*

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k HD. *CoRR*, abs/2404.06512.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA: exploring the limits of masked visual representation learning at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19358–19369. IEEE.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5843–5851. IEEE Computer Society.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. AGQA: A benchmark for compositional spatio-temporal reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11287–11297. Computer Vision Foundation / IEEE.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer*

*Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 961–970. IEEE Computer Society.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1359–1367. IEEE Computer Society.

Gal Lavee, Ehud Rivlin, and Michael Rudzsky. 2009. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Trans. Syst. Man Cybern. Part C*, 39(5):489–504.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2023. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 487–507. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355.

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. 2023c. Mvbench: A comprehensive multi-modal video understanding benchmark. *CoRR*, abs/2311.17005.

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

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *CoRR*, abs/2311.10122.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. Mmbench: Is your multi-modal model an all-around player? *CoRR*, abs/2307.06281.

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

Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *CoRR*, abs/2306.05424.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2023. Timechat: A time-sensitive multimodal large language model for long video understanding. *CoRR*, abs/2312.02051.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2023. Moviechat: From dense token to sparse memory for long video understanding. *CoRR*, abs/2307.16449.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. 2023. Video understanding with large language models: A survey. *CoRR*, abs/2312.17432.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Yu Wang, Zeyuan Zhang, Julian J. McAuley, and Zexue He. 2024. LVCHAT: facilitating long video comprehension. *CoRR*, abs/2402.12079.

Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. 2021. STAR: A benchmark for situated reasoning in real-world videos. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2023. Funqa: Towards surprising video comprehension. *CoRR*, abs/2306.14899.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1645–1653. ACM.

Dongjie Yang. 2024. Vript. Accessed: 2024-5-17.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. CLEVRER: collision events for video representation and reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9127–9134. AAAI Press.

Hang Zhang, Xin Li, and Lidong Bing. 2023. 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*, pages 543–553. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. abs/2303.18223.

# A Appendix

### A.0.1 Data Statistics.

Our benchmark comprises a total of 2,190 video question-answer pairs on 6 tasks corresponding to different event understanding abilities, where each task has 172-400 test samples for evaluation.

### A.0.2 Ablation Study

**Number of Merged Videos.** In Section 4.1, we select $k$ samples and merge them into a new one, where a larger $k$ indicates more events happening in the new video. We experiment with $k = \{1, 2, 3, 4\}$ ($k = 1$ indicates no merge operation) and depict the corresponding performance in Figure 7. We could observe that increasing the number of events from 2 to 3 and 4 hurts performance on all the tasks, but is still better than the model trained on a single video.

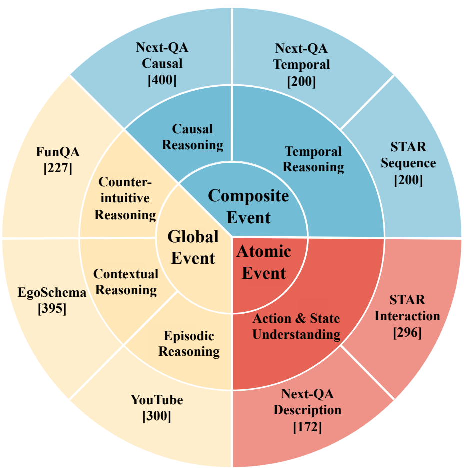



Figure 6: The dataset distribution of our benchmark.

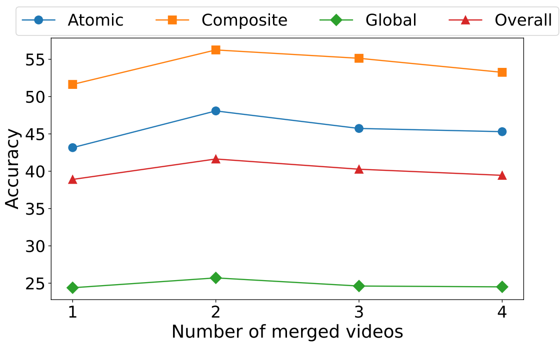



Figure 7: Performance comparison w.r.t the number of selected videos during video instruction merging.

| Benchmark | Time Scope (s) | Annotation | Open Domain | Complex Reasoning | Hierarchical Events | Multiple Scenes |
|---|---|---|---|---|---|---|
| MSVD-QA (Xu et al., 2017) | 0∼60 | Auto | ✓ | ✗ | ✗ | ✗ |
| MSRVTT-QA (Xu et al., 2017) | 10∼30 | Auto | ✓ | ✗ | ✗ | ✗ |
| TGIF-QA (Jang et al., 2017) | - | Auto+Human | ✓ | ✗ | ✗ | ✗ |
| ActivityNet-QA (Yu et al., 2019) | 0∼975 | Human | ✗ | ✗ | ✗ | ✗ |
| NeXT-QA (Xiao et al., 2021) | 5∼180 | Human | ✓ | ✗ | ✗ | ✗ |
| STAR (Wu et al., 2021) | 2∼195 | Auto | ✓ | ✓ | ✗ | ✗ |
| CLEVRER (Yi et al., 2020) | 5 | Auto | ✗ | ✓ | ✗ | ✗ |
| EgoSchema (Mangalam et al., 2023) | 180 | Auto | ✗ | ✓ | ✗ | ✗ |
| MVBench (Li et al., 2023c) | 5∼40 | Auto | ✓ | ✓ | ✗ | ✗ |
| TempCompass (Liu et al., 2024c) | 0∼35 | Auto+Human | ✓ | ✗ | ✗ | ✗ |
| MovieChat (Song et al., 2023) | 401∼602 | Human | ✓ | ✗ | ✗ | ✓ |
| Ours | 2∼1088 | Auto+Human | ✓ | ✓ | ✓ | ✓ |

Table 5: Comparison with previous video understanding benchmarks.

| | | Atomic | Composite | | Overall | | |
|---|---|---|---|---|---|---|---|
| | | Event Description | Temporal Reasoning | Causal Reasoning | Counter-intuitive Reasoning | Contextual Reasoning | Episodic Reasoning |
| 1 frame | LLaVA-NeXT (7B) | 13.68 | 14.75 | 9.75 | 14.98 | 9.11 | 7.3 |
| | IXC2-4KHD (7B) | 26.07 | 27.5 | 32.5 | 9.25 | 12.15 | 17.67 |
| 8 frame | LLaMA-VID (7B) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | LLaMA-VID (13B) | 1.92 | 1.75 | 0.00 | 3.08 | 0.00 | 4 |
| | Video-LLaVA (7B) | 12.82 | 5.5 | 0.00 | 6.17 | 2.78 | 7.2 |
| | Video-LLaMA2 (7B) | 15.81 | 9 | 6.25 | 0.09 | 2.28 | 0.67 |
| | VideoChat2 (7B) | 31.2 | 37.25 | 47.25 | 14.98 | 15.44 | 12.67 |
| | ST-LLM (7B) | 47.22 | 48.75 | 59.5 | 9.69 | 25.32 | 16.67 |
| | GPT-4V | 29.27 | 32.75 | 41.25 | 42.29 | 24.81 | 24 |
| | GPT-4o | 48.08 | 47.5 | 55.5 | 63 | 48.86 | 34 |
| 16 frame | LLaMA-VID (7B) | 0.21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | LLaMA-VID (13B) | 1.06 | 1.13 | 0.25 | 3.08 | 0.00 | 5 |
| | Video-LLaMA2 (7B) | 11.11 | 3.25 | 6 | 0.88 | 3.04 | 0.33 |
| | PLLaVA | 34.62 | 40 | 40.5 | 17.62 | 15.19 | 11 |
| | VideoChat2 (7B) | 34.19 | 38.25 | 46.25 | 17.18 | 17.22 | 12.67 |
| | ST-LLM (7B) | 47.65 | 50.00 | 56.5 | 11.45 | 26.84 | 14.67 |
| | GPT-4V | 29.7 | 35 | 40.00 | 36.56 | 28.35 | 27 |
| | GPT-4o | 52.99 | 55 | 58.25 | 63 | 49.11 | 32.67 |
| 32 frame | LLaMA-VID (7B) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | LLaMA-VID (13B) | 0.85 | 0.75 | 0.00 | 3.08 | 0.00 | 3 |
| | Video-LLaMA2 (7B) | 9.19 | 4.75 | 3.75 | 2.2 | 1.77 | 1.33 |
| | VideoChat2 (7B) | 33.76 | 37.75 | 47.75 | 16.74 | 15.7 | 14.67 |
| | ST-LLM (7B) | 46.79 | 46.25 | 55.25 | 10.13 | 26.33 | 16 |
| | GPT-4V | 23.72 | 25.75 | 33 | 40.09 | 20.51 | 20.67 |
| | GPT-4o | 54.27 | 56.75 | 58.25 | 63.44 | 50.13 | 37.33 |
| more frames | MovieChat (7B) | 16.88 | 16 | 14.5 | 18.06 | 13.16 | 20.33 |
| | Video-ChatGPT (7B) | 9.83 | 9.5 | 15 | 14.98 | 12.66 | 10.00 |
| | Gemini-1.5-Pro | 48.5 | 47.5 | 41.75 | 52.86 | 32.15 | 38.67 |

Table 6: Detailed experimental results with more frames as input.