TempCompass: Do Video LLMs Really Understand Videos?

Yuanxin Liu[§*], Shicheng Li[§*], Yi Liu[§], Yuxiang Wang[§], Shuhuai Ren[§], Lei Li[†], Sishuo Chen[¶], Xu Sun[§], Lu Hou[‡]
[§] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[¶] Center for Data Science, Peking University
[†] The University of Hong Kong [‡] Huawei Noah's Ark Lab
{liuyuanxin, yuxiangwang, shuhuai_ren}@stu.pku.edu.cn nlp.lilei@gmail.com
{lisc99, imliuyi, chensishuo, xusun}@pku.edu.cn houlu3@huawei.com

Abstract

Recently, there is a surge in interest surrounding video large language models (Video LLMs). However, existing benchmarks fail to provide a comprehensive feedback on the temporal perception ability of Video LLMs. On the one hand, most of them are unable to distinguish between different temporal aspects (e.g., speed, direction) and thus cannot reflect the nuanced performance on these specific aspects. On the other hand, they are limited in the diversity of task formats (e.g., only multi-choice QA), which hinders the understanding of how temporal perception performance may vary across different types of tasks. Motivated by these two problems, we propose the **TempCompass** benchmark, which introduces a diversity of temporal aspects and task formats. To collect high-quality test data, we devise two novel strategies: (1) In video collection, we construct conflicting videos that share the same static content but differ in a specific temporal aspect, which prevents Video LLMs from leveraging single-frame bias or language priors. (2) To collect the task instructions, we propose a paradigm where humans first annotate meta-information for a video and then an LLM generates the instruction. We also design an LLM-based approach to automatically and accurately evaluate the responses from Video LLMs. Based on TempCompass, we comprehensively evaluate 9 state-of-the-art (SOTA) Video LLMs and 3 Image LLMs, and reveal the discerning fact that these models exhibit notably poor temporal perception ability. The data and evaluation code are available at `https://github.com/llyx97/TempCompass`.

1 Introduction

The development of video understanding systems has long been a popular topic in artificial intelligence research. Inspired by the unprecedented progress of large language models (LLMs), a line of initial efforts (Li et al., 2023c; Zhang et al., 2023; Maaz et al., 2023; Luo et al., 2023; Ren et al., 2023b) have been devoted to build LLMs with video understanding ability. These Video LLMs can serve as versatile multi-modal solvers for video and language tasks, demonstrating strong potential across various real-world applications.

With the rapid development of Video LLMs, a compelling question arises: "Do Video LLMs really understand the temporal dynamics of videos" Despite the importance of this question, current benchmarks fail to provide a satisfactory answer. **Firstly**, a majority of them neglect differentiating between various temporal aspects (e.g., type of action, speed and direction), thereby failing to offer a comprehensive view to diagnose the temporal perception ability. **Secondly**, while some Video LLM benchmarks (Chen et al., 2023; Li et al., 2023d) have categorized various temporal aspects, they are restricted in task format variety (e.g., only multi-choice QA). Consequently, they are not optimally suited for assessing Video LLMs, which are expected to generalize across diverse tasks and instruction formats.

In response to the above issues, this work proposes the **TempCompass**, a benchmark to comprehensively evaluate the temporal perception ability of Video LLMs. TempCompass introduces five basic temporal aspects (*Action, Speed, Direction, Attribute Change* and *Event Order*) and ten fine-grained sub-aspects, as shown in Figure 1. Additionally, TempCompass involves four distinct types of task formats (*Multi-Choice QA, Yes/No QA, Caption Matching* and *Caption Generation*), as shown in Figure 2, which allows us to investigate how the temporal perception ability of Video LLMs varies across different task formats.

The videos in TempCompass are originated from the ShutterStock[1] platform. These open-domain

[*]Equal contribution.

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

Findings of the Association for Computational Linguistics: ACL 2024, pages 8731–8772
August 11-16, 2024 ©2024 Association for Computational Linguistics

videos cover a diverse variety of contents, ranging from human activities to natural scenarios, among others. To prevent Video LLMs from leveraging single-frame bias or language priors to complete the tasks, we construct *conflicting video* pairs/triplets, within which the videos share the same static content but differ from each other in a specific temporal aspect. Given the collected videos, we derive 7,540 task instructions for the four types of tasks, using a collaboration of human annotated meta-information and LLM generation.

Due to the diverse task formats in TempCompass and the free-form nature of Video LLM responses, it is non-trivial to automatically evaluate the performance of Video LLMs. To address this challenge, we resort to the language understanding ability of LLMs for evaluation. For each type of task, we use tailored evaluation prompts for ChatGPT (`gpt3.5-turbo`) to assess whether the Video LLM response is correct. To balance the cost and accuracy of evaluation, we also adopt some rule-based assessment methods, which are implemented prior to utilizing ChatGPT.

Based on our TempCompass benchmark, we evaluate 12 SOTA multi-modal LLMs (MLLMs), including 9 Video LLMs and 3 Image LLMs. The evaluation results reveal that the Video LLMs demonstrate a deficiency in temporal perception skills, failing to surpass their Image LLMs counterparts. We also find that the temporal perception ability of MLLMs indeed varies a lot across different task formats, which emphasizes the need to incorporate diverse task formats in the assessment process.

The main contributions of this work are summarized as follows: (1) We present a benchmark with diverse temporal aspects and task formats to comprehensively evaluate the temporal perception ability of Video LLMs. (2) We introduce conflicting videos that prevent Video LLMs from exploiting singe-frame bias or language priors. (3) We combine rule-based and LLM-based methods to efficiently and accurately evaluate the responses from Video LLMs. (4) Our empirical results reveal the weak temporal perception ability of SOTA Video LLMs.

2 Related Work

2.1 Multi-Modal Large Language Models

Following the success of pure-text LLMs (Brown et al., 2020; OpenAI, 2022; Touvron et al., 2023a,b; Taori et al., 2023; Chiang et al., 2023), numerous recent efforts have been made to build multi-modal LLMs (MLLMs). To enable LLMs to comprehend visual context, two categories of paradigms have emerged and evolved. The *Pipeline* paradigm (Shen et al., 2023; Sur'is et al., 2023; Wu et al., 2023; Yang et al., 2023) leverages off-the-shelf vision expert models to extract visual information in the form of texts, which are then fed to LLMs to perform the downstream vision tasks. The *End-to-End* paradigm integrates vision encoders and LLM in an end-to-end trainable manner. The outputs from vision encoders are mapped to the LLM embedding space, using linear projectors (Liu et al., 2023b,a; Zhu et al., 2023b), attention-based projections (Li et al., 2023b; Ye et al., 2023; Dai et al., 2023; Bai et al., 2023b) or mixed projections (Lin et al., 2023b; Gao et al., 2024). Recent Video LLMs (Su et al., 2023; Li et al., 2023c; Zhang et al., 2023; Lin et al., 2023a; Li et al., 2023d; Maaz et al., 2023; Luo et al., 2023; Li et al., 2023g; Jin et al., 2023) primarily follow the *End-to-End* paradigm, with optional temporal modules to model the temporal information across frames.

Benchmark	Temporal Diversity	Task Diversity	Open Domain
Video Understanding Benchmarks			
MSVD-QA (Xu et al., 2017)	✗	✗	✓
MSRVTT-QA (Xu et al., 2017)	✗	✗	✓
TGIF-QA (Jang et al., 2017)	✗	✗	✓
SSv2 (Goyal et al., 2017)	✗	✗	✗
SSv2-label (Lei et al., 2022)	✗	✗	✗
CLEVRER (Yi et al., 2020)	✗	✗	✗
ActivityNet-QA (Yu et al., 2019)	✗	✗	✗
NEXT-QA (Xiao et al., 2021)	✗	✓	✗
ViLMA (Kesen et al., 2024)	✓	✗	✓
Perception Test (Puatruaucean et al., 2023)	✓	✓	✗
VITATECS (Li et al., 2023e)	✓	✗	✓
Video LLM Benchmarks			
SEEDBench (Li et al., 2023a)	✗	✗	✗
Video-Bench (Ning et al., 2023)	✗	✗	✓
VLM-Eval (Li et al., 2023f)	✗	✓	✓
AutoEval-Video (Chen et al., 2023)	✓	✗	✓
MVBench (Li et al., 2023d)	✓	✗	✓
TempCompass (Ours)	✓	✓	✓

Table 1: Comparison with related benchmarks. The rightmost three columns represent, respectively, whether the benchmark assesses performance across diverse temporal aspects, task formats, and includes open-domain videos. The detailed temporal aspects and task formats are described in Appendix A.6.

2.2 Temporal Perception Evaluation

Temporal perception is a fundamental distinction between video-centered and image-centered applications. Prior to the age of LLMs, a lot of studies (Goyal et al., 2017; Yi et al., 2020; Yu et al., 2019;

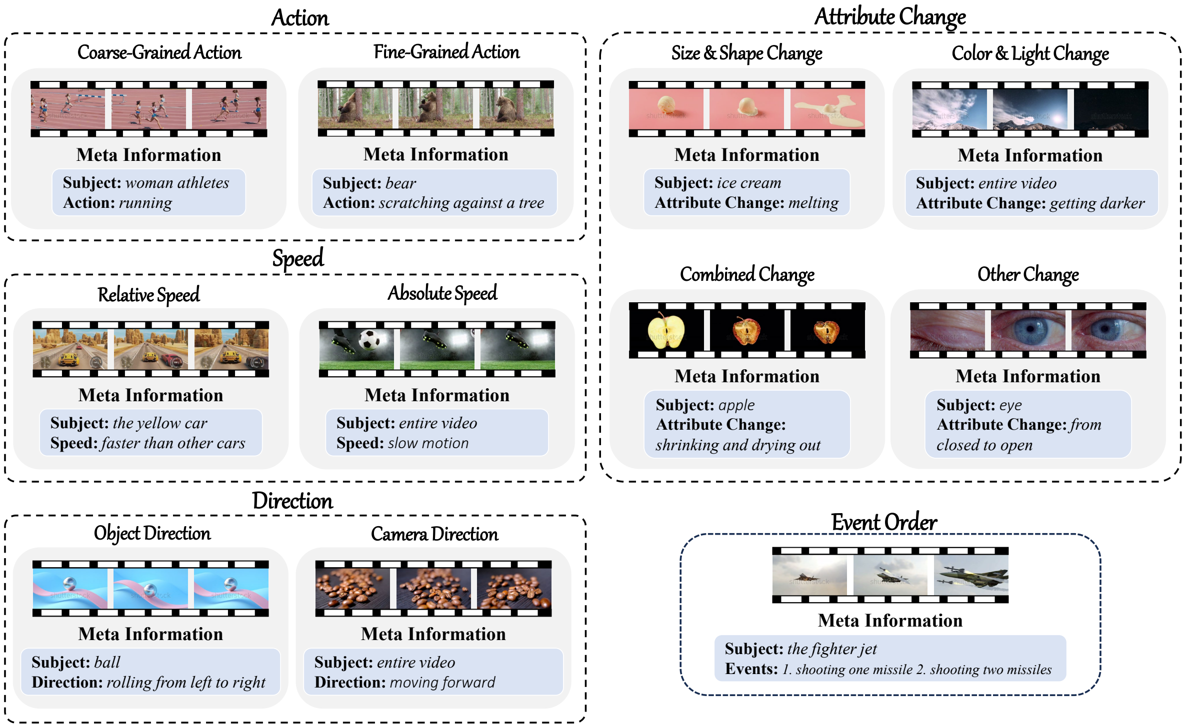


Figure 1: Illustration of the temporal aspects (Section 3.1.1) and meta-information (Section 3.2.2).

Bagad et al., 2023; Buch et al., 2022; Hendricks et al., 2018; Sevilla-Lara et al., 2019; Jang et al., 2017; Ren et al., 2023a; Xiao et al., 2021) have been conducted to evaluate the temporal perception performance of video-language models. However, most of these works neglect the distinction between various temporal aspects. To tackle this issue, the Perception Test (Puatruaucean et al., 2023), VITATECS (Li et al., 2023e) and ViLMA (Kesen et al., 2024) introduce a diversity of fine-grained temporal aspects, thereby enabling a more comprehensive and nuanced evaluation of the temporal perception capability. However, VITATECS and ViLMA are limited in the diversity of task formats and Perception Test is constrained to indoor videos, making them less ideal to evaluate Video LLMs.

2.3 MLLM Benchmarks

With the advent of MLLMs, there is an increasing number of MLLMs benchmarks. A majority of them (Fu et al., 2023; Liu et al., 2023c; Yu et al., 2023; Bai et al., 2023c; Xu et al., 2023) are specifically designed for Image LLMs. Recently, some tailored benchmarks have also been proposed for Video LLMs. However, among these Video LLM benchmarks, SEEDBench (Li et al., 2023a), VLM-Eval (Li et al., 2023f) and Video-Bench (Ning et al., 2023) fall short in discerning between various temporal aspects. AutoEval-Video (Chen et al., 2023) and MVBench (Li et al., 2023d) define and incorporate a range of temporal aspects while lacking diverse task formats.

Table 1 compares TempCompass with representative video understanding and Video LLM benchmarks. We can see that TempCompass stands out by emphasizing diverse temporal aspects, task formats and open-domain videos.

3 TempCompass Benchmark

TempCompass is a dataset of videos and task instructions intended to test the temporal perception ability of Video LLMs. This section will introduce the temporal aspects, task formats and static contents included in TempCompass (Section 3.1), how to collect the videos and task instructions (Section 3.2) and how to automatically evaluate Video LLMs on TempCompass (Section 3.4).

3.1 Benchmark Structure

3.1.1 Temporal Aspects

In contrast to images that only contain static visual information, videos convey dynamic visual information over time, i.e., temporal information. As shown in Figure 1, we identify five basic aspects of temporal information in TempCompass:

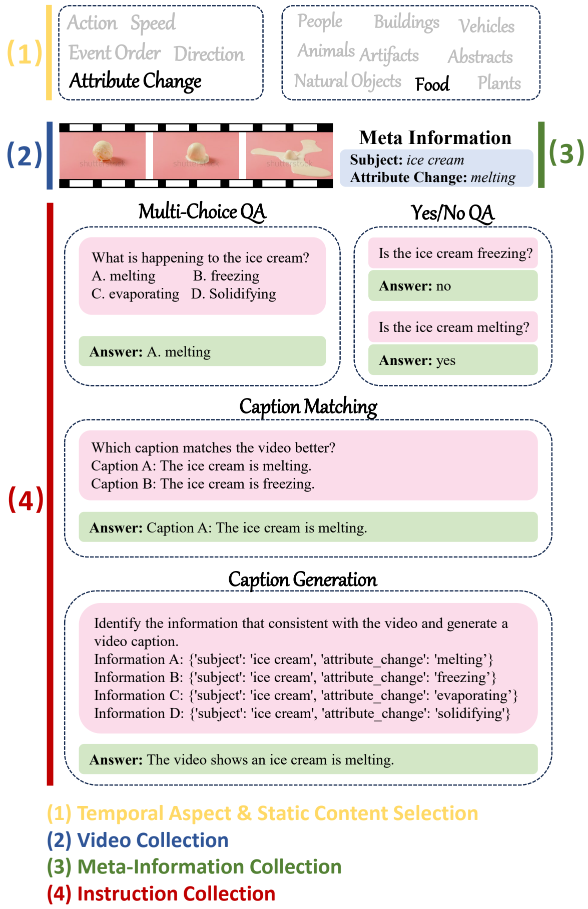


Figure 2: Illustration of the four types of task formats and the data collection steps.

Action. This aspect assesses the ability to distinguish between different types of actions, which is a common task for video understanding models. We further divide this aspect into *Coarse-Grained Action* and *Fine-Grained Action*. The former involves a broader set of activities or movements while the latter is about more specific and detailed actions.

Speed. This aspect delves into the capacity to discern variations in speed, which is further categorized into two components. *Absolute Speed* focuses on the speed of a specific object or the pace of an entire video while *Relative Speed* compares the speed of different objects.

Direction. This aspect emphasizes the perception of movement direction. Under this aspect, we separately consider the direction of objects (*Object Direction*) and the direction of camera (*Camera Direction*).

Attribute Change. This aspect centers on how the attribute of objects or the entire video change over time. Attribute change encompasses four sub-aspects, including *Size & Shape*, *Color & Light Change*, *Combined Change* and *Other Change*.

Event Order. This aspect focuses on the chronological order that different events happen in a video.

3.1.2 Task Formats

Having established the definition of different aspects of temporal information, we now deal with the question of "how to examine whether a Video LLM understands a specific temporal information?" As illustrated in Figure 2, for a specific temporal information in the given video, we test the temporal perception ability of Video LLMs using four types of tasks: **(1)** *Multi-Choice QA* asks the model to select the correct answer from multiple candidate choices. **(2)** *Yes/No QA* involves the model determining whether a statement is correct based on the video. **(3)** *Caption Matching* requires the model to distinguish between two video captions, one of which is consistent with the video while the other is inconsistent with the video in the temporal aspect of interest. **(4)** In the task of *Caption Generation*, several pieces of information about the given temporal aspect are presented to the model, which is then asked to select the correct one and generate a video caption accordingly. Such a constrained form of captioning makes it easier to automatically evaluate the correctness of the generated caption (see Section 3.4 for details).

3.1.3 Static Contents

We define nine categories of static contents: *people, animals, plants, food, natural objects, vehicles, artifacts, buildings, abstract* (please refer to Appendix A.1 for detailed descriptions). Each video in TempCompass is classified into one or multiple categories, depending on the static visual content.

3.2 Data Collection

Each data example in TempCompass contains four components: video, meta-information, static content categories and task instructions. As shown in Figure 2, we collect these components in four steps. **(1)** We first select a set of temporal aspects and static content categories, based on which we then **(2)** collect a video together with **(3)** annotated meta-information. **(4)** Following this, we employ ChatGPT (`gpt3.5-turbo`) (OpenAI, 2022), an LLM, to generate task instructions according to the meta-information. Next, we will describe how to collect the three components in detail.

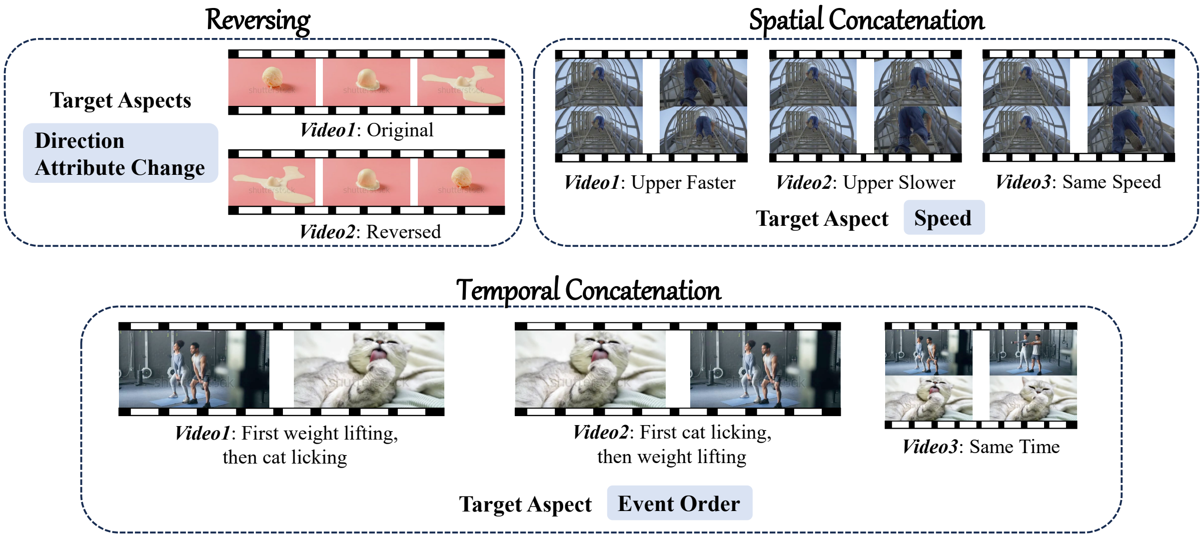


Figure 3: Illustration of conflicting video pairs/triplets for different temporal aspects.

3.2.1 Video Collection

We collect raw videos from the ShutterStock platform. To enhance video diversity, we carefully control the static content distribution, guaranteeing that each category contains an adequate number of video samples. (Figure 4(b) shows the distribution). At the same time, we ensure that the videos are not included in WebVid (Bain et al., 2021), a dataset widely used in pre-training video-language models.

In the literature, it has been shown that video understanding models may utilize language priors or single-frame bias as shortcuts to obtain the correct answer, without truly understanding the temporal content of a video (Huang et al., 2018; Buch et al., 2022; Sevilla-Lara et al., 2019; Lei et al., 2022; Girdhar and Ramanan, 2019). Language priors is the prior knowledge learned from language modeling (e.g., an ice cream is more likely to be melting instead of freezing). Single-frame bias refers to the reliance on static visual cues in a single frame, which strongly correlates with the correct answer (e.g., inferring the moving direction of a vehicle from its orientation in a single frame).

To mitigate the impact of such shortcuts, we construct conflicting video pairs/triplets. Within a pair/triplet, the videos have the same static content, but differ from each other in a particular temporal aspect. In this manner, the very shortcut that induces a correct answer for one video will inversely lead to an incorrect answer when applied to the conflicting counterpart. Specifically, as depicted in Figure 3, we propose three methods to construct the conflicting videos:

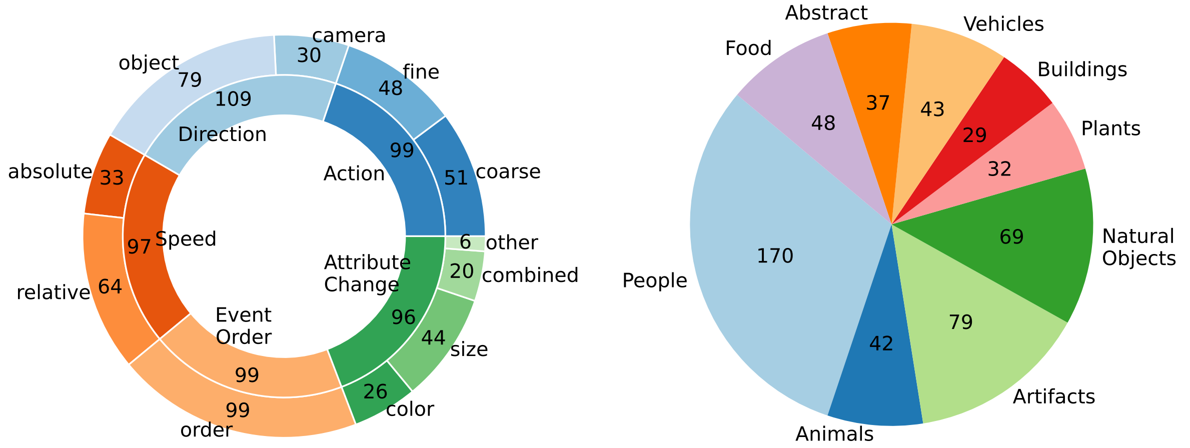
Reversing. Information of *Direction* and *Attribute Change* in a video can usually be modified by playing the video in reverse. Therefore, the conflicting video pairs for these two temporal aspects consist of an original video and its reversed counterpart.

Spatial Concatenation. For the *Speed* aspect, we first accelerate or decelerate a video. Then, we concatenate this modified video with the original one along the spatial dimension by (1) placing the faster version above or (2) placing the slower version above, creating two conflicting videos. We also construct a third video by concatenating two exactly same videos in the spatial dimension.

Temporal Concatenation. For the *Event Order* aspect, we concatenate two videos along the temporal dimension. Two conflicting videos are produced by reversing the order of the two original videos, creating two different sequences of events. Additionally, we construct a third video by spatially concatenating the two original videos, thereby presenting the two events at the same time.

3.2.2 Meta-Information Collection

Given a collected video, we convert its key information into textual format. To reduce the load of annotation, we manually annotate semi-structured meta-information. As Figure 1, 2 shows, each piece of meta-information is comprised of two parts: (1) a phrase describing the subject and (2) another phrase describing the information related to the temporal aspect of interest.



(a) Temporal Aspects.

(b) Static Content Categories.

Figure 4: Distribution of videos over temporal aspects and static content categories.

3.2.3 Instruction Collection

With the annotated meta-information, we obtain the task instructions via a process with interleaved automatic generation and manual refinement. Specifically, we first employ ChatGPT to automatically generate *Multi-Choice QA* instructions based on the meta-information. Then, these instructions are checked and rectified by humans. Subsequently, we prompt ChatGPT to generate *Yes/No QA*, *Caption Matching* and *Caption Generation* instructions, based on the manually rectified *Multi-Choice QA* instructions. These instructions are also further checked and rectified by humans. More details of instruction collection and the prompts for instruction generation are shown in Appendix A.2.

3.2.4 Data Statistics

We collect a total of 410 videos and 500 pieces of meta-information (a video may be annotated with multiple pieces of meta-information). Figure 4 depicts the video statistics, revealing an even distribution across basic temporal aspects, with roughly 100 videos representing each aspect. The nine content categories are also well covered by our collected videos. These data distributions demonstrate the diversity of TempCompass in terms of both temporal aspects and static visual contents.

Given a piece of meta-information, we collect multiple instructions for each type of task: at least 3 for *Multi-Choice QA*, 2 for *Yes/No QA*, 3 for *Caption Matching*, and 4 for *Caption Generation*. In this way, we collect a total of 7,540 instructions in our benchmark. In Appendix A.3, we show the detailed distribution of task instructions, video duration and answer distribution. In Appendix A.5, we present complete data examples including the video, meta-information, static content and instructions.

3.3 Quality Verification

After the data collection process described in Section 3.2, we randomly sample 200 task instructions to verify the data quality. These instructions and videos are presented to three human annotators to perform the task. Human annotators also have the option to label an instruction as "Cannot Answer", which indicates that the instruction is unreasonable. Among the 600 annotated results, only 5 are labeled as "Cannot Answer". Table 2 also show that the human annotators achieve near-perfect accuracy across most tasks and aspects, attesting to the high quality of the collected data. More details of quality verification can be found in Appendix A.4.

3.4 Automatic Evaluation

For *Multi-Choice QA*, *Yes/No QA* and *Caption Matching*, we adopt a hybrid approach that integrates rule-based methods and ChatGPT to automatically evaluate the responses generated by Video LLMs. To begin with, we check whether any candidate option (e.g., *A/B/C/D*, *Yes/No* or *Caption A/Caption B*) is explicitly mentioned in the response and compare it against the ground-truth answer. Hand-crafted matching rules are specifically designed for different types of tasks. Then, for responses that fail to match any candidate options, we resort to ChatGPT's language understanding ability to determine whether they are correct based on the

	Baseline		Image LLM			Video LLM								
	Human	Random	LLaVA-1.5 13B	SPHINX-v2 13B	Qwen-VL 7B	V-LLaVA 7B	LLaMA-VID 7B	mPLUG-Owl 7B	PandaGPT 13B	Valley 7B	VideoChat2 7B	V-ChatGPT 7B	V-LLaMA 13B	TimeChat 7B
Multi-Choice QA														
Action	100	28.9	71.3	**89.9**	85.8	70.4	58.6	66.6	35.5	47.0	<u>88.5</u>	47.0	54.1	82.5
Direction	96.7	27.8	31.6	**37.0**	<u>36.7</u>	32.2	29.9	29.3	27.8	29.3	36.4	31.6	24.5	36.4
Speed	90	32.1	36.0	**43.2**	42.3	38.2	29.3	32.2	29.3	32.5	42.0	28.4	28.1	**43.2**
Event Order	100	32.2	34.4	36.4	40.7	<u>41.4</u>	30.5	34.8	31.8	18.9	40.7	37.1	32.8	**42.1**
Attribute Change	100	28.5	38.9	45.1	44.8	39.9	26.0	35.4	30.9	29.9	<u>45.5</u>	30.9	28.5	**47.2**
Avg	97.3	29.9	42.8	<u>50.9</u>	50.6	44.7	35.3	40.0	31.1	31.8	**51.1**	35.2	33.9	50.7
Match Rate	-	-	84.2	99.6	46.8	37.9	62.9	3.1	6.4	3.5	**100.0**	1.3	0.6	<u>99.9</u>
Yes/No QA														
Action	96.7	50.0	74.7	<u>79.1</u>	**81.4**	74.3	63.0	64.4	53.0	58.1	72.8	52.5	68.1	66.2
Direction	83.3	50.0	48.8	51.2	51.6	51.8	48.8	50.6	49.6	<u>52.0</u>	**53.8**	50.0	46.0	51.6
Speed	96.7	50.0	49.0	<u>54.7</u>	**59.8**	50.3	49.2	51.2	50.8	52.5	53.8	49.5	48.8	47.7
Event Order	93.3	50.0	49.5	**54.5**	50.8	49.2	48.4	51.3	<u>53.7</u>	50.3	51.3	51.0	51.8	51.0
Attribute Change	100	50.0	**55.4**	50.4	49.1	51.1	52.7	52.0	52.2	52.9	53.8	50.0	50.9	<u>54.7</u>
Avg	94	50.0	56.4	<u>59.1</u>	**60.0**	56.4	53.0	54.4	51.8	53.5	58.0	50.7	53.7	54.7
Match Rate	-	-	**100.0**	**100.0**	<u>99.8</u>	**100.0**	99.1	95.6	**100.0**	98.7	18.8	**100.0**	95.1	**100.0**
Caption Matching														
Action	100	50.0	86.9	<u>89.2</u>	**90.2**	88.2	72.7	56.9	56.6	15.5	65.0	64.6	73.1	75.1
Direction	96.7	50.0	50.8	52.0	53.5	<u>53.8</u>	45.6	45.3	51.4	21.4	**53.8**	48.6	47.4	**55.7**
Speed	100	50.0	54.6	47.1	<u>55.0</u>	**61.9**	52.2	46.4	44.3	22.0	52.6	47.8	47.1	<u>55.0</u>
Event Order	100	50.0	55.0	53.0	**60.3**	57.0	49.0	49.3	55.0	28.3	53.0	49.3	52.0	<u>58.3</u>
Attribute Change	100	50.0	51.0	55.2	<u>56.9</u>	**58.3**	49.0	49.0	49.0	22.9	53.8	48.6	48.3	51.7
Avg	99.3	50.0	59.5	59.2	<u>63.1</u>	**63.7**	53.6	49.3	51.3	22.0	55.6	51.8	53.5	59.1
Match Rate	-	-	76.7	88.7	91.6	74.3	41.9	15.8	30.7	11.2	<u>95.3</u>	7.5	0.1	**100.0**
Caption Generation														
Action	100	28.8	<u>67.4</u>	**67.9**	62.6	50.8	53.0	46.5	23.7	24.7	54.0	40.9	54.3	56.1
Direction	86.7	28.4	<u>31.9</u>	19.0	27.8	28.7	28.0	28.2	25.7	20.4	31.0	28.4	21.3	**32.8**
Speed	100	32.4	24.7	20.4	29.6	23.2	21.9	<u>30.4</u>	26.0	21.9	**32.7**	24.5	13.9	29.4
Event Order	100	32.1	33.0	37.2	34.8	38.2	35.5	31.2	29.8	35.8	34.2	31.8	<u>38.5</u>	**38.8**
Attribute Change	100	28.6	35.4	31.0	32.3	33.6	35.9	<u>36.5</u>	32.6	29.4	**41.4**	33.9	33.9	37.8
Avg	97.3	30.0	38.4	34.9	37.3	34.8	34.8	34.4	27.5	26.3	<u>38.5</u>	31.8	32.2	**38.9**

Table 2: Accuracy of MLLMs on our TempCompass benchmark. “V-” in the model names stands for “Video-”. The best and second-best MLLM results are **bold** and <u>underlined</u>, respectively. "Match Rate" denotes the success rate of matching a predicted option from the MLLM’s response using hand-crafted rules. The complete results of all temporal aspects are reported in Appendix D.1.

task instruction and ground-truth answer. Details of the matching rules and the prompts for LLM-based evaluation are illustrated in AppendixB.

When it comes to the *Caption Generation* task, the rule-based evaluation method is ineffective because almost all Video LLM responses are free-form video captions. Therefore, we solely rely on ChatGPT for evaluation. Specifically, we prompt ChatGPT to answer the corresponding *Multi-Choice* question using the generated video caption as context. If the answer by ChatGPT is correct, then the generated caption is deemed as correct and vice versa. The motivation is that if the Video LLM selects an incorrect information to generate the caption, ChatGPT will consequently select an incorrect option. Considering the possibility that the generated caption may not involve any of the provided information, we include an extra option: “*None of the choices are correct*” in the *Multi-Choice* question. In case where ChatGPT selects this option, the generated caption is also deemed as incorrect.

4 Experiments

4.1 Evaluated Models

We conduct evaluation experiments on a total of 12 open-sourced state-of-the-art MLLMs, including TimeChat (Ren et al., 2023b), Video-LLaMA (Zhang et al., 2023), Video-ChatGPT (Maaz et al., 2023), Valley (Luo et al., 2023), VideoChat2 (Li et al., 2023d), mPLUG-Owl (Ye et al., 2023), PandaGPT (Su et al., 2023), Video-LLaVA (Lin et al., 2023a), LLaMA-VID (Li et al., 2023g), LLaVA-v1.5 (Liu et al., 2023a), SPHINX (Lin et al., 2023b; Gao et al., 2024) and Qwen-VL-Chat (Bai et al., 2023b). These models cover a wide range of Video LLMs and Image LLMs with different model architectures and training strategies. Inspired by (Li et al., 2023d), we append answer prompts to the task instructions to guide MLLMs generating responses in the desired formats (see Appendix C.2 for details). In addition to the MLLMs, we also incorporate random and human baselines. Details of the models and human baseline are described in Appendix C and Appendix A.4, respectively.

4.2 Main Results

Table 2 summarizes the results across the four tasks. We discuss the results from four perspectives:

Overall Performance. Existing MLLMs exhibit poor temporal perception ability. Five Video LLMs, i.e., LLaMA-VID, Panda-GPT, Valley, Video-ChatGPT, and Video-LLaMA, fail to convincingly surpass the random baseline across all tasks. Although TimeChat, Video-LLaVA and VideoChat2 exhibit improved performance, they still fall significantly short of the human. Notably, all Video LLMs struggle to consistently surpass SPHINX-v2 and Qwen-VL-Chat, two Image LLMs, highlighting a pervasive lack of temporal perception ability in current Video LLMs. This finding echoes with VITATECS (Li et al., 2023e), which reveals that current video-language models barely surpass random guesses in a task similar to our *Caption Matching*.

Performance Across Temporal Aspects. MLLMs demonstrate their highest proficiency in *Action* aspect, with the best model achieving near 90 accuracy on *Multi-Choice QA* and *Caption Matching*. The reason is that the type of action can largely be deduced from static visual cues alone. This observation indicates that existing MLLMs already demonstrate a strong understanding capability of static visual information, which is the foundation to develop temporal perception capabilities. In comparison, the performance are significantly worse on the remaining four aspects, as they are more dependent on the temporal information across frames. This finding implies that there is a pressing need for enhancing the current MLLMs' capabilities in perceiving *Speed*, *Direction*, *Event Order* and *Attribute Change*.

Performance Across Tasks. Comparing the results across all four tasks, we can see that there exists a significant variation in performance. This variation can be attributed to two factors. On the one hand, the inherent complexity of the tasks varies, as exemplified by the performance differences between *Multi-Choice QA* and *Caption Generation*. The latter generally yields worse results, because it necessitates not only selecting the correct information but also generating the caption accordingly. On the other hand, individual models have innate strengths and weaknesses in different tasks. For instance, Video-LLaVA takes a leading place in the *Speed* aspect on *Caption Matching*, while performing not better than random in the same temporal aspect on *Yes/No QA* and *Caption Generation*. These findings suggest that the temporal perception ability displayed by MLLMs is highly dependent on the form of evaluation tasks, which emphasizes the need to incorporate a diverse array of tasks in the assessment process.

Question: What is happening in the video?
A. A person drops down the pineapple
B. A person pushes forward the pineapple
C. A person rotates the pineapple
D. A person picks up the pineapple

Original Video

Ground-Truth: D. A person picks up the pineapple
SPHINX-v2: D (✓)
Video-LLaVA: D (✓)
Video-ChatGPT: The best option is to pick up the pineapple. (✓)

Reversed Video

Ground-Truth: A. A person drops down the pineapple
SPHINX-v2: D (✗)
Video-LLaVA: C (✗)
Video-ChatGPT: The best option would be to pick up the pineapple and place it on the table. (✗)

Table 3: Example of MLLM responses to a multi-choice question, given a pair of conflicting videos. ✓and ✗are assessed by our automatic evaluation method.

Ability to Respond in Desired Format. Despite the use of answer prompts, some MLLM usually fail to respond in the desired format, as reflected by the low match rate in Table 2. This phenomenon demonstrates the limitation of rule-based matching in evaluating MLLM responses and underlines the necessity of LLM-based evaluation. We also observe that the design of answer prompt has a non-negligible impact on the match rate. Please refer to Appendix D.2 for the analytical study.

4.3 Qualitative Results

Table 3 illustrates the responses from three MLLMs, given a pair of conflicting videos of the *Direction* aspect. We can see that all three MLLMs

Multi-Choice	Yes/No	Caption Matching	Caption Generation
99.67	98.33	99.0	79.33

Table 4: Accuracy of the automatic evaluation results, benchmarked against human evaluation as ground-truth.

Model	Multi-Choice	Yes/No	Caption Matching	Caption Generation
LLaVA-1.5				
w/ Conflicting	35.1	50.6	52.8	31.3
w/o Conflicting	**41.1**	**52.8**	**57.3**	**33.6**
SPHINX-v2				
w/ Conflicting	40.3	52.7	51.8	26.7
w/o Conflicting	**52.8**	**58.3**	**62.8**	**31.5**
Qwen-VL-Chat				
w/ Conflicting	41.0	53.2	56.4	31.0
w/o Conflicting	**49.0**	**59.5**	**63.5**	**32.9**
Random	30.1	50.0	50.0	30.3

Table 5: The performance of Image LLMs w/ and w/o conflicting videos. The results are averaged over all temporal aspects except for the *Action* aspect, for which we do not construct conflicting videos.

accurately respond to the question when presented with the original video; however, they fail to deliver correct answers when confronted with the reversed version. This result indicates the inherent inability of the models to perceive and understand the direction of movement. The automatic evaluation results also showcase that our LLM-based evaluation method is able to deal with the free-form response from MLLMs. More qualitative results on other temporal aspects and task formats can be found in Appendix D.3.

4.4 Automatic Evaluation Accuracy

To validate the reliability of the proposed automatic evaluation method, we compare its results with human evaluation. The details of evaluation setups is described in Appendix B.3. Table 4 shows the percentage of automatic evaluation results that agree with human judgements, averaged over three human evaluators. We can see that our automatic evaluation method achieves very high consistency with humans in *Multi-Choice QA*, *Yes/No QA* and *Caption Matching*. In terms of *Caption Generation*, roughly 20% of the LLM-based evaluation are inconsistent with humans. This is because the MLLMs may hallucinate some contents irrelevant to the video, which is hard to detect for the pure-text GPT3.5-Turbo. In Appendix D.3, we present qualitative examples to better illustrate the pros and cons of our automatic evaluation method.

4.5 Effect of the Conflicting Videos

Table 5 compares the performance of Image LLMs on all videos and when excluding the constructed conflicting videos (i.e., on the raw videos). Evidently, the Image LLMs notably outperform the random baseline on raw video samples, especially in *Multi-Choice QA*, *Yes/No QA* and *Caption Matching*. This implies that, to a considerable degree, questions about raw videos can be answered by leveraging the single-frame bias and language priors. With the introduction of conflicting video, the performance of Image LLMs is clearly closer to random baseline, effectively alleviating the impact of biases. The effect of conflicting videos is also illustrated by the example cases in Table 3, 20, 21, 22, 23.

5 Conclusions

In this work, we propose the TempCompass benchmark to evaluate the temporal perception ability of Video LLMs. Our benchmark introduces ten temporal aspects and four distinct types of task formats, which offers a comprehensive view to investigate the temporal perception capability. Two innovative strategies are devised in the data collection process, including (1) the construction of conflicting videos to mitigate the influence of single-frame bias and language-priors and (2) the collaboration of human annotation and LLM generation to efficiently collect high-quality task instructions. We also propose an automatic evaluation method based on ChatGPT, which is able to accurately assess the free-form Video LLM responses. Based on TempCompass, we extensively evaluate 9 SOTA Video LLMs and 3 Image LLMs. Our evaluation results reveal the pressing need to enhance the temporal perception ability of Video LLMs.

6 Limitations

Despite the contributions made by TempCompass, this work is still limited in two perspective. First, despite our effort in constructing the conflicting videos, the influence of single-frame and languages-priors persist. This is evident from the fact that Image LLMs continue to perform clearly above random baselines in specific tasks and temporal aspects. Second, our automatic evaluation method encounters challenges in accurately assessing certain generated video captions, which, although consistent with the ground-truth candidate information, incorporate elements of hallucinated content.

Acknowledgements

We thank all the anonymous reviewers for their constructive comments. This work is supported in part by a Huawei Research Grant, National Natural Science Foundation of China (No. 62176002) and The Fundamental Research Funds for the Central Universities. Xu Sun is the corresponding author of this paper.

References

Piyush Bagad, Makarand Tapaswi, and Cees G. M. Snoek. 2023. Test of time: Instilling video-language models with a sense of time. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2503–2516.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *ArXiv*, abs/2309.16609.

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

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xing Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023c. Touchstone: Evaluating vision-language models by language models. *ArXiv*, abs/2308.16890.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1708–1718.

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 *NeurIPS*.

S. Buch, Cristobal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the “video” in video-language understanding. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2907–2917.

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. *ArXiv*, 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 Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Yuxin Fang, Wen Wang, Binhui Xie, Quan-Sen Sun, Ledell Yu Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2022. Eva: Exploring the limits of masked visual representation learning at scale. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, Wenqi Shao, Chao Xu, Conghui He, Junjun He, Hao Shao, Pan Lu, Hongsheng Li, and Yu Qiao. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *ArXiv*, abs/2402.05935.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190.

Rohit Girdhar and Deva Ramanan. 2019. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *ArXiv*, abs/1910.04744.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter N. 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. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2018. Localizing moments in video with temporal language. In *EMNLP*, pages 1380–1390. Association for Computational Linguistics.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

De-An Huang, Vignesh Ramanathan, Dhruv Kumar Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. 2018. What makes a video a video: Analyzing temporal information in video understanding models and datasets. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7366–7375.

Y. Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1359–1367.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2023. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *ArXiv*, abs/2311.08046.

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. 2024. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. In *International Conference on Learning Representations (ICLR)*.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2022. Revealing single frame bias for video-and-language learning. *ArXiv*, abs/2206.03428.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *ArXiv*, 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. 2023d. Mvbench: A comprehensive multi-modal video understanding benchmark. *ArXiv*, abs/2311.17005.

Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. 2023e. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. *ArXiv*, abs/2311.17404.

Shuailin Li, Yuang Zhang, Yucheng Zhao, Qiuyue Wang, Fan Jia, Yingfei Liu, and Tiancai Wang. 2023f. Vlm-eval: A general evaluation on video large language models. *ArXiv*, abs/2311.11865.

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

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

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Jiao Qiao. 2023b. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *ArXiv*, abs/2311.07575.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. 2022a. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3032–3041.

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

Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022b. A convnet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Ming-Hui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *ArXiv*, abs/2306.07207.

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

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. *ArXiv*, abs/2311.16103.

OpenAI. 2022. Introducing chatgpt. *CoRR*.

Maxime Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193.

Viorica Puatruaucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan S. Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yezhou Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexander Fréchette, Hanna Klimczak, R. Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. *ArXiv*, abs/2305.13786.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou. 2023a. Testa: Temporal-spatial token aggregation for long-form video-language understanding. *ArXiv*, abs/2310.19060.

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

Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. 2019. Only time can tell: Discovering temporal data for temporal modeling. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 535–544.

Yongliang Shen, Kaitao Song, Xu Tan, Dong Sheng Li, Weiming Lu, and Yue Ting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *ArXiv*, abs/2303.17580.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *ArXiv*, abs/2305.16355.

D'idac Sur'is, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11854–11864.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

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

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Chenfei Wu, Sheng-Kai Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *ArXiv*, abs/2303.04671.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9772–9781.

D. 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. *Proceedings of the 25th ACM international conference on Multimedia*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Jiao Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *ArXiv*, abs/2306.09265.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *ArXiv*, abs/2303.11381.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan

Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178.

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 *ICLR*. OpenReview.net.

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

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 *AAAI*, 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. *ArXiv*, abs/2306.02858.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Liejie Yuan. 2023a. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *ArXiv*, abs/2310.01852.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592.

A More Details of Data

A.1 Static Contents

Our benchmark covers nine categories of static contents, including *people, animals, plants, food, natural objects, vehicles, artifacts, buildings, abstract*. *Natural objects* denotes lifeless natural objects and scenery. *Artifacts* encompasses human-made objects, excluding large-size objects like vehicles and buildings. *Abstract* refers to abstract geometric shapes and symbols. For better understanding, please refer to the example videos with annotated categories in Table 16,17,18,19.

A.2 Instruction Collection

We collect the task instructions in four steps:

1. Generating *Multi-Choice QA* instructions based on meta-information, using ChatGPT.
2. Manually review and rectify [2] the generated *Multi-Choice QA* instructions.
3. Generating instructions for the other three tasks based on manually rectified *Multi-Choice QA* instructions, using ChatGPT.
4. Manually review and rectify the generated instructions.

The detailed collection process for each type of task is described as follows:

Multi-Choice QA The task instructions are directly generated from the annotated meta-information. We also design some in-context learning examples to help ChatGPT better understand the task to accomplish. The detailed prompt is shown in Table 12. For each piece of meta-information, we prompt ChatGPT to generate five *Multi-Choice QA* instructions. To prevent bias towards any specific option position, we randomly shuffle the order of the options. Following this step, the generated instructions undergo meticulous review and refinement by the authors, ensuring that a minimum of three high-quality instructions are retained within in the benchmark.

Yes/No QA. Based on the manually rectified *Multi-Choice QA* questions, we prompt ChatGPT to directly generate an equal number of Positive and Negative questions, as shown in Table 13.

Caption Matching. Based on the manually rectified *Multi-Choice QA* questions, we first prompt ChatGPT to generate a True caption and three False captions, which are subsequently integrated into several templates to construct the task instructions. To eliminate bias stemming from caption position, we randomize the sequence in which True and False captions are displayed for each instruction. The caption generation prompt and instruction templates are shown in Table 14.

Caption Generation. As shown in Table 15, the instructions for this task consists of a task description and several pieces of information similar to the meta-information. We first manually compose a task description and paraphrase it using ChatGPT. Then, an instruction "Ensure that the generated video caption is brief" is appended to the two task descriptions, resulting in four task descriptions in total. The candidate information are derived from the meta-information and the manually rectified *Multi-Choice* question.

[2]All human annotation and evaluation in this study were done by the authors.

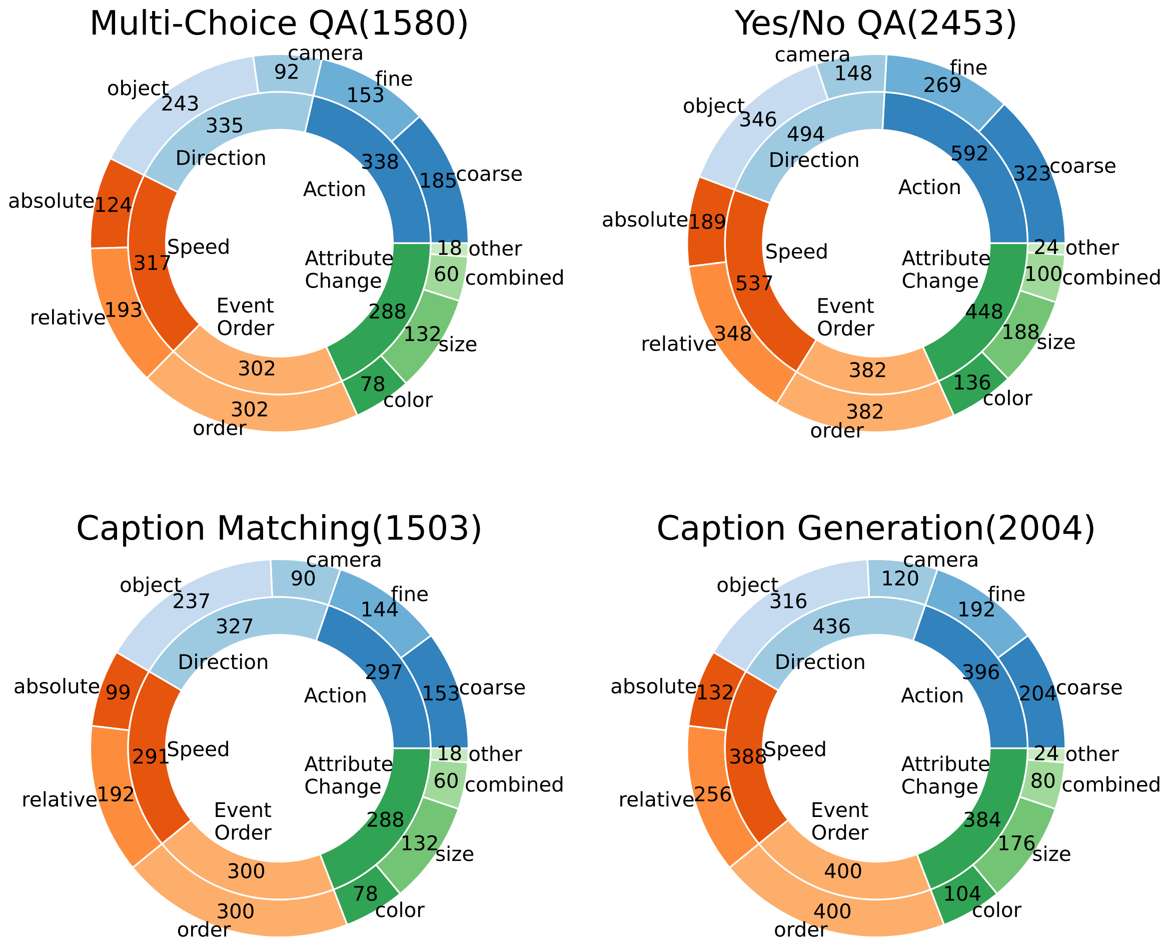


Figure 5: Distribution of task instructions over the temporal aspects.

A.3 Data Statistics

Task Instructions. Figure 5 presents the distribution of task instructions. We can see that each type of task involves at least 1,500 instructions and every basic temporal aspect has a balanced number of these instructions.

Answers. As can be seen in Figure 6, the distribution of ground-truth answers within our benchmark dataset is balanced across all options. An exception is the option "*D*" in *Multi-Choice QA*, which appears less frequently compared to the other three options. This is because not all *Multi-Choice* question includes four options. when we restrict our analysis to questions that offer exactly four options (a total of 675 instances), the frequency of "*D*" as the correct answer (occurring 157 times) aligns closely with the frequencies of the remaining three options.

Video Duration. Figure 7 shows the distribution of video duration. Our benchmark primarily focuses on short and medium-length videos within 30 seconds.

A.4 Quality Verification and Human Baseline

We randomly sample 200 task instructions, with a balanced distribution of 10 instructions for each temporal aspect across every task (i.e., 50 instructions per task). For *Multi-Choice QA*, *Yes/No QA* and *Caption Matching*, human annotators are directly asked to select an option, instead of generating a free-form answer as the MLLMs. The selected option is then compared with the ground-truth answer. For *Caption Generation*, human annotators follow the same instructions presented to MLLMs to generate video captions, which are then evaluated in the manner described in Appendix B. In addition to performing the task, human annotators have another option to label a task instruction

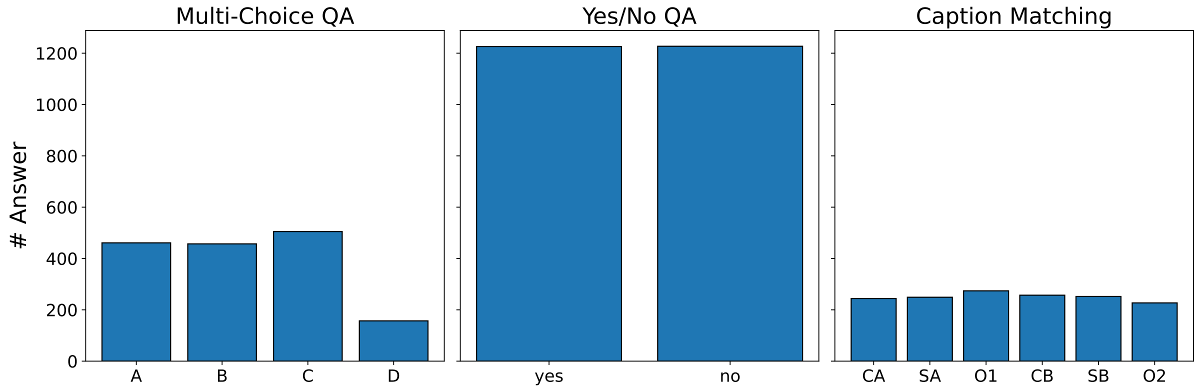


Figure 6: Distribution of answers. CA, CB, SA, SB, O1, O2 stand for Caption A, Caption B, Sentence A, Sentence B, Option 1, Option 2, respectively.

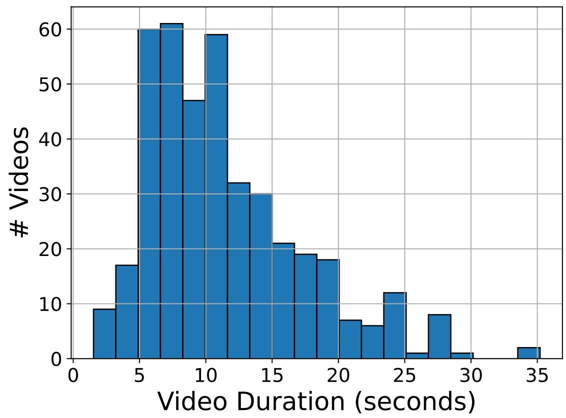


Figure 7: Distribution of video duration.

as "Cannot Answer". In this case, the answer is considered as incorrect when evaluating the human performance. Figure 8 shows the interface to collect human answers. The final results are obtained by averaging among three human annotators.

A.5 Data Examples

Table 16,17,18,19 illustrate complete data examples in our benchmark. Each example contains the video, meta-information, static content categories and task instructions.

A.6 Comparison With Related Benchmarks

Table 6 summarizes the specific temporal aspects and task formats involved in related benchmarks. We can see that the majority of existing benchmarks lack a comprehensive categorization of temporal aspects. By contrast, VITATECS, Perception Test, ViLMA and MVBench introduce a variety of temporal aspects, which are complementary to the ones presented in our TempCompass. Meanwhile, the variation in performance across different task formats cannot be reflected by most current benchmarks. While Perception Test considers both multiple task formats and temporal aspects, it is constrained to indoor videos that focusing on people and artifacts. In comparison, our proposed TempCompass uniquely stands out by emphasizing a rich variety of temporal dimensions, diverse task formats and open-domain videos. This design enables TempCompass to provide a more holistic assessment of Video LLM's temporal perception capabilities.

It is worth noting that the definition of temporal aspects and task formats vary among different studies. For the sake of clarity, we unify their naming in Table 6. Here we explain some definitions as follows:

Temporal Aspects. "Repetition" measures the ability to count the number of repeating activities. "Object Interaction" focuses on the relationship between different objects participating in the same event. "Temporal Localization" require the model to identify the temporal position of specific events in the video. "X Change" encompasses various changes over time, including attribute change, scene change, etc.

Task Formats. "Free-form QA" may involve different formats of task instructions but a proper categorization is not provided in the benchmark. In "Action Recognition", the model is required to classify videos into a predetermined set of actions. Notably, the original SSv2 dataset does not offer explicit task instructions for this classification process. "Grounded Video QA" demands that the model tracks the objects meeting specific condi-

Benchmark	Temporal Aspects	Task Formats	Open Domain
Conventional Video Understanding Benchmarks			
MSVD-QA (Xu et al., 2017)	-	Free-form QA	✓
MSRVTT-QA (Xu et al., 2017)	-	Free-form QA	✓
TGIF-QA (Jang et al., 2017)	Repetition,Event Order	Free-form QA	✓
SSv2 (Goyal et al., 2017)	-	Action Recognition	✗
SSv2-label (Lei et al., 2022)	-	Caption Matching	✗
CLEVRER (Yi et al., 2020)	-	MC QA,Free-form QA	✗
ActivityNet-QA (Yu et al., 2019)	Action,Event Order	Free-form QA	✗
NEXT-QA (Xiao et al., 2021)	Action,Event Order	MC QA,Free-form QA	✗
ViLMA (Kesen et al., 2024)	Action,Direction, X Change,Repetition	Caption Matching	✓
VITATECS (Li et al., 2023e)	Action,Event Order, Speed,Direction Object Interaction	Caption Matching	✓
Perception Test (Puatruaucean et al., 2023)	Event Order,Repetition, Direction,Action,X Change Temporal Localization	MC QA,Grounded Video QA Object Tracking,Point Tracking Action Localization	✗
Video LLM Benchmarks			
SEEDBench (Li et al., 2023a)	Action,Event Order	MC QA	✗
Video-Bench (Ning et al., 2023)	-	MC QA	✓
VLM-Eval (Li et al., 2023f)	-	Free-form QA,Retrieval, Caption Generation	✓
AutoEval-Video (Chen et al., 2023)	Event Order,Direction, Attribute Change	Free-form QA	✓
MVBench (Li et al., 2023d)	Action,Repetition,Direction, Temporal Localization, X Change,Event Order Object Interaction	MC QA	✓
TempCompass (Ours)	Action,Speed,Direction, Attribute Change,Event Order	MC QA,Y/N QA, Caption Matching, Caption Generation	✓

Table 6: Comparison with related benchmarks. The temporal aspects focus on basic temporal perception ability while excluding the aspects that require reasoning skills. "MC QA" and "Y/N QA" represent multi-choice QA and Yes/No QA, respectively. Video LLMs cannot be directly tested on the gray task formats because they lack textual task instructions. Detailed definition of some temporal aspects and task formats are explained in Appendix A.6.

tions by pinpointing them within bounding boxes throughout the video. "Object Tracking" and "Point Tracking" require tracking the bounding boxes and points, without providing a textual task instruction. "Retrieval" encompasses text-to-video (T2V) retrieval and video-to-text (V2T) retrieval. Taking V2T as example, the Video LLM first generates a description of the video, which is then used to retrieve the relevant texts.

B More Details of Evaluation Setups

B.1 Rule-based Evaluation

For *Multi-Choice QA*, we map the Video LLM response to an option if the response matches with the complete option (e.g., "A. melting") or the option indicator (e.g., "A"). For *Caption Matching*, we match the Video LLM response with the complete option (e.g., "Caption A: Ice cream is melting."), the option sentence (e.g., "Ice cream is melting.") or the option indicator (e.g., "Caption A"). In terms of *Yes/No QA*, we check if the Video LLM response starts with "yes" or "no". Once the Video LLM response has been mapped to a specific option, we proceed by comparing that chosen option with the ground-truth answer to assess the correctness of the response.

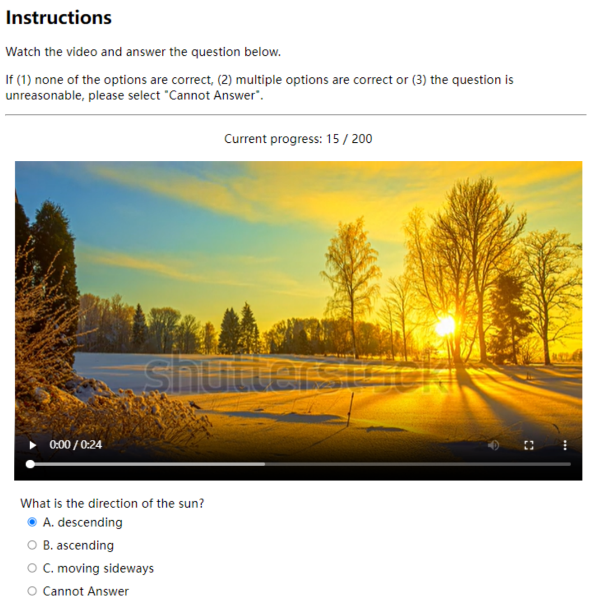


Figure 8: Screenshot of the interface to collect human answers.

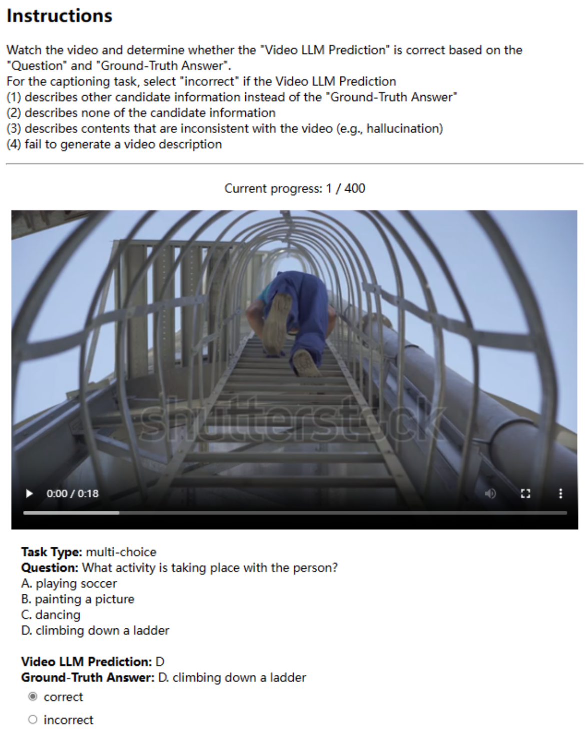


Figure 9: Screenshot of the human evaluation interface.

B.2 LLM-based Evaluation

If a Video LLM response fails to match any of the options, we resort to LLM-based evaluation. For *Multi-Choice QA*, *Yes/No QA* and *Caption Matching*, we present task instruction, Video LLM response and ground-truth answer to ChatGPT and prompt it to determine whether the response is correct. The detailed prompts are shown in Table 10.

Regarding the task of *Caption Generation*, we engage ChatGPT to tackle the corresponding *Multi-Choice QA* task, using the caption generated by Video LLM contextual reference, as described in Section 3.4. To enhance the accuracy of ChatGPT in answering the *Multi-Choice* questions, we present it with several in-context learning examples and prompt it to generate an extra reasoning step prior to obtaining the final answer. Table 11 presents a clear illustration of the prompt structure used in this process.

B.3 Human Evaluation

To conduct human evaluation, we randomly sample 400 responses from SPHINX-v2 and Video-LLaVA, ensuring that each of the four tasks contains an equal share of 100 samples. The video, MLLM response, task instruction and ground-truth answer are presented to three human annotators, who then assign binary labels indicating the correctness of the MLLM response. For the *Caption Generation* task, an MLLM response is deemed as incorrect if it (1) describes other candidate information instead of the "Ground-Truth Answer", (2) describes none of the candidate information, (3) describes contents that are inconsistent with the video (e.g., hallucination), or (4) fail to generate a video description. Figure 9 illustrates the interface used in our human evaluations.

C More Details of Evaluated Models

C.1 Model Architecture

We evaluate the performance of nine Video LLMs and three Image LLMs on TempCompass. All the evaluated models follow the prevalent MLLM paradigm and contain three primary components: a visual encoder, a vision-language connector, and an LLM. The details of these methods are as follows.

TimeChat (Ren et al., 2023b) introduces a time-sensitive frame encoder (ViT (Dosovitskiy et al., 2020) + Q-Former) to encode timestamp knowledge into visual tokens, then uses a sliding video Q-Former to aggregate the representations of individual frames. It takes LLaMA-2-7B as the LLM backbone.

Video-LLaMA (Zhang et al., 2023) employs the same visual encoder as used in BLIP-2 (Li et al., 2023b) (ViT (Dosovitskiy et al., 2020) + Q-Former) and introduces a trainable video Q-Former to aggregate the representations of individual frames. Both the vision encoder and the LLM are frozen during training. We choose "Video-LLaMA-2-13B" for evaluation which is based on LLaMA-2-13B (Touvron et al., 2023b).

Video-ChatGPT (Maaz et al., 2023) proposes to use spatial pooling and temporal pooling to aggregate frame features from a frozen image encoder (CLIP-ViT-L/14 (Radford et al., 2021)). A single linear layer is utilized to connect the pooled features to a frozen LLM (Vicuna-v1.1-7B (Chiang et al., 2023)). Unlike most MLLMs, Video-ChatGPT only performs single-stage instruction tuning on video-text data.

Valley (Luo et al., 2023) uses a similar pooling strategy as Video-ChatGPT and further incorporates a temporal modeling module into the vision encoder. In Valley, the LLM parameters are also fine-tuned during instruction tuning to achieve stronger performance. Our evaluation is carried out on "Valley2-7b" with LLaMA-2-7B as the base LLM.

VideoChat2 (Li et al., 2023d) adopts UMT-L (Liu et al., 2022a) as the vision encoder, Vicuna-v0-7B as the LLM, and utilizes a Q-Former to connect both modalities. It follows a progressive three-stage training strategy including vision-language alignment, vision-language connection, and instruction tuning.

mPLUG-Owl (Ye et al., 2023) proposes to use a visual abstractor similar to the Q-Former to connect the vision encoder and the LLM. It incorporates both language-only data and multimodal data into the instruction tuning procedure. Its video version, "mPLUG-Owl-video-7B", uses LLaMA-7B (Touvron et al., 2023a) as the LLM and introduces additional temporal query tokens into the visual abstractor for temporal modeling.

PandaGPT (Su et al., 2023) adopts ImageBind (Girdhar et al., 2023) as the visual encoder which is pre-trained for multi-modal alignment. Similar to LLaVA (Liu et al., 2023b), the vision-language connector consists only of a linear projection. Only the projection and additional LoRA (Hu et al., 2021) weights on LLM attention modules are updated during single-stage instruction tuning. We test "pandagpt-13b-max-len-400" on our dataset, which uses Vicuna-v0-13B as the LLM.

Video-LLaVA (Lin et al., 2023a) uses LanguageBind (Zhu et al., 2023a) to encode visual inputs and a linear layer to project visual features into the LLM space. LanguageBind and the LLM (Vicuna-v1.5-7B) are both frozen during the two-stage training.

LLaMA-VID (Li et al., 2023g) represents each frame with two tokens, a text-guided context token and a visual content token, which significantly reduces computational cost when increasing the number of sampled frames. We evaluate the performance of "llama-vid-7b-full-224-video-fps-1", which is based on EVA-ViT-G (Fang et al., 2022) and Vicuna-v1.5-7B.

LLaVA-1.5 (Liu et al., 2023a) is an Image LLM built upon the pioneering framework of LLaVA. It replaces the original linear connector with an MLP and includes additional training data to enhance its capabilities. The version we test is "LLaVA-1.5-13B", which adopts Vicuna-v1.5-13B as the LLM.

SPHINX (Lin et al., 2023b) achieves high performance on many Image LLM benchmarks by mixing visual embeddings from various vision backbones including ViT, ConvNeXt (Liu et al., 2022b), DINOv2 (Oquab et al., 2023), and Q-Former. We evaluate "SPHINX-v2" on our benchmark which is built upon LLaMA-2-13B.

Qwen-VL-Chat (Bai et al., 2023b) utilizes a single-layer cross-attention module with learnable query embeddings as the vision-language connector. It undergoes extensive vision-language pre-training before fine-tuned on multi-modal instruction data. The LLM of Qwen-VL-Chat is initialized with Qwen-7B (Bai et al., 2023a).

C.2 Inference Settings

Table 7 shows the detailed inference settings for the MLLMs. The frame sampling strategies of Video LLMs and the LLM decoding strategies of all the evaluated MLLMs are determined according to the recommended inference script in their corresponding codebases. For Image LLMs, we extract the middle frame of each video as the visual input to these models.

Inspired by (Li et al., 2023d), we append answer prompts to the task instructions to guide MLLMs generating responses in the desired formats. In terms of *Multi-Choice QA* and *Caption Matching*, we employ the prompt "*Best Option:*" for most

	Frame sampling		LLM decoding	
	strategy	# frame	strategy	parameter
Image LLM				
LLaVA-1.5 (13B)	Middle frame	1	Random	$T = 0.7$
SPHINX-v2 (13B)	Middle frame	1	Top-p	$T = 0.9, p = 0.8$
Qwen-VL-Chat (7B)	Middle frame	1	Top-p	$p = 0.3$
Video LLM				
TimeChat (7B)	Uniform	96	Top-p	$p = 0.8$
Video-LLaVA (7B)	Uniform	8	Random	$T = 0.1$
LLaMA-VID (7B)	1fps	variable	Random	$T = 1.0$
mPLUG-Owl (7B)	Uniform	8	Top-k	$k = 5$
PandaGPT (13B)	See Girdhar et al. (2023)	10	Top-p	$p = 0.8$
Valley (7B)	Uniform	8	Greedy	
VideoChat2 (7B)	Uniform	16	Greedy	
Video-ChatGPT (7B)	Uniform	100	Random	$T = 0.2$
Video-LLaMA (13B)	Uniform	8	Top-p	$p = 0.8$

Table 7: Inference settings for the evaluated MLLMs.

models except for VideoChat2 and TimeChat. Regarding VideoChat2, an additional left bracket is appended (i.e., "*Best Option: (*") following the original paper (Li et al., 2023d). For TimeChat, we utilize "Please directly give the best option:". In the case of *Yes/No QA*, we introduce the prompt of "*Please answer yes or no:* ". Lastly, for *Caption Generation*, we use "*Generated Caption:*". Unless otherwise specified, all the results in this paper are obtained using the above answer prompts.

D More Experimental Results

D.1 Results on Fine-Grained Temporal Aspects

Table 8 summarizes the evaluation results on all fine-grained temporal aspects.

D.2 Effect of Answer Prompt

Table 9 reports the results on *Multi-Choice QA* and *Caption Matching* when using two different answer prompts, i.e., "*Best Option:*" and "*Please directly give the best option:*". The following observations can be derived: (1) The selection of answer prompt has a non-negligible impact on the match rate. The latter answer prompt, which is more detailed, can substantially increase the match rate for most MLLMs that already achieve >30% match rate using the former answer prompt, on both two tasks. By contrast, the match rate of VideoChat2 significantly drops from near 100% to near 0%. This reveals that while VideoChat2 can respond in desired format (i.e., directly selecting an option) by identifying the left-bracket, it is not robust to the variation of answer prompts. (2) Compared with match rate, the accuracy is relatively insensitive to the chance of answer prompt. For instance, the change in accuracy of Video-LLaVA and Qwen-VL-Chat on *Multi-Choice QA* is less than 1%, despite a 50% 60% increase of match rate.

D.3 Qualitative Results

Table 22, 23, 24, 25, 26, 27, 28, 29 demonstrate examples of MLLM responses alongside our automated assessment results. We can find that: (1) It is evident that the models demonstrate a deficiency in genuine temporal perception skills in terms of *speed*, *direction*, *event order* and *attribute change*. While they manage to provide accurate answers for most questions in certain videos, their performance falters when confronted with corresponding conflicting videos. (2) The proposed automatic evaluation method is reliable. Despite the arbitrary form of MLLM responses, our method can offer accurate assessment in most cases. (3) Our LLM-based evaluation method mistakenly assesses a small portion of incorrect captions as correct (Table 26, 27, 28), which echoes with the results in Table 4. We find that such inaccurate evaluation is mostly caused by the failure to detect hallucinated contents in the captions. Table 30 presents two more detailed evaluation examples with intermediate reasoning steps by ChatGPT. As we can see, ChatGPT is able to select the correct option in *Multi-Choice QA*, despite the existence of hallucinated content in the generated captions, thereby leading to inaccurate assessment.

E Licencing and Intended Use

Our TempCompass benchmark is under CC-BY 4.0 license. The videos and textual annotation in this work should only be used for research purposes.

	Action		Direction		Speed		Event Order	Attribute Change				Avg
	fine	coarse	object	camera	absolute	relative	order	color	size	combined	other	
Multi-Choice QA												
Baseline												
Human	100		96.7		90		100	100				97.3
Random	28.4	29.3	28.3	26.3	30.6	33.0	32.2	30.1	28.9	26.4	25.9	29.9
Image LLM												
LLaVA-1.5 (13B)	56.2	83.8	32.5	29.3	44.4	30.6	34.4	42.3	35.6	38.3	50.0	42.8
SPHINX-v2 (13B)	**85.0**	94.1	36.2	**39.1**	48.4	39.9	36.4	51.3	40.2	46.7	50.0	50.9
Qwen-VL-Chat (7B)	82.4	88.6	37.4	34.8	46.0	39.9	40.7	**52.6**	40.9	43.3	44.4	50.6
Video LLM												
TimeChat (7B)	73.9	89.7	37.0	34.8	44.4	**42.5**	**42.1**	46.2	**47.7**	**48.3**	44.4	50.7
Video-LLaVA (7B)	54.9	83.2	31.7	33.7	46.0	33.2	41.4	39.7	40.2	35.0	**55.6**	44.7
LLaMA-VID (7B)	34.6	78.4	30.0	29.3	30.6	28.5	30.5	23.1	25.0	28.3	38.9	35.3
mPLUG-Owl (7B)	49.7	80.5	28.8	30.4	36.3	29.5	34.8	30.8	37.1	35.0	44.4	40.0
PandaGPT (13B)	40.5	31.4	29.6	22.8	20.2	35.2	31.8	30.8	33.3	25.0	33.3	31.1
Valley (7B)	33.3	58.4	34.2	16.3	31.5	33.2	18.9	39.7	26.5	26.7	22.2	31.8
VideoChat2 (7B)	80.4	**95.1**	**39.1**	29.3	**54.0**	34.2	40.7	**52.6**	43.9	43.3	33.3	**51.1**
Video-ChatGPT (7B)	28.8	62.2	33.7	26.1	28.2	28.5	37.1	26.9	31.1	35.0	33.3	35.2
Video-LLaMA (13B)	40.5	65.4	26.7	18.5	28.2	28.0	32.8	26.9	25.0	33.3	44.4	33.9
Yes/No QA												
Baseline												
Human	96.7		83.3		96.7		93.3	100				94
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Image LLM												
LLaVA-1.5 (13B)	65.4	82.4	48.3	50.0	48.1	49.4	49.5	55.1	52.7	**62.0**	50.0	56.4
SPHINX-v2 (13B)	72.1	84.8	50.6	52.7	58.7	52.6	**54.5**	44.1	51.6	55.0	58.3	59.1
Qwen-VL-Chat (7B)	**74.0**	**87.6**	51.4	52.0	**61.9**	**58.6**	50.8	45.6	51.6	52.0	37.5	**60.0**
Video LLM												
TimeChat (7B)	55.8	74.9	51.4	52.0	47.1	48.0	51.0	**58.1**	53.2	53.0	54.2	54.7
Video-LLaVA (7B)	58.4	**87.6**	51.4	52.7	50.8	50.0	49.2	52.2	50.0	53.0	45.8	56.4
LLaMA-VID (7B)	53.9	70.6	48.3	50.0	53.4	46.8	48.4	54.4	50.0	55.0	54.2	53.0
mPLUG-Owl (7B)	54.6	72.4	50.9	50.0	52.4	50.6	51.3	57.4	52.7	49.0	29.2	54.4
PandaGPT (13B)	53.9	52.3	50.3	48.0	47.6	52.6	53.7	53.7	52.7	47.0	**62.5**	51.8
Valley (7B)	50.2	64.7	**52.3**	51.4	57.7	49.7	50.3	57.4	51.1	52.0	45.8	53.5
VideoChat2 (7B)	62.5	81.4	**52.3**	**57.4**	59.3	50.9	51.3	50.7	54.3	58.0	50.0	58.0
Video-ChatGPT (7B)	50.2	54.5	50.0	50.0	49.7	49.4	51.0	50.0	50.0	50.0	50.0	50.7
Video-LLaMA (13B)	58.7	75.9	45.1	48.0	53.4	46.3	51.8	43.4	**55.3**	54.0	45.8	53.7
Caption Matching												
Baseline												
Human	100		96.7		100		100	100				99.3
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Image LLM												
LLaVA-1.5 (13B)	82.6	90.8	48.9	55.6	61.6	51.0	55.0	39.7	50.0	**66.7**	55.6	59.5
SPHINX-v2 (13B)	82.6	95.4	54.0	46.7	54.5	43.2	53.0	56.4	50.8	63.3	55.6	59.2
Qwen-VL-Chat (7B)	**86.1**	94.1	56.5	45.6	55.6	54.7	**60.3**	**64.1**	55.3	51.7	55.6	63.1
Video LLM												
TimeChat (7B)	64.6	85.0	53.6	**61.1**	60.6	52.1	58.3	41.0	54.5	58.3	55.6	59.1
Video-LLaVA (7B)	79.2	**96.7**	53.2	55.6	**69.7**	**57.8**	57.0	60.3	**56.1**	60.0	**61.1**	**63.7**
LLaMA-VID (7B)	61.8	83.0	47.7	40.0	56.6	50.0	49.0	51.3	47.7	46.7	55.6	53.6
mPLUG-Owl (7B)	54.9	58.8	45.6	44.4	48.5	45.3	49.3	42.3	50.8	51.7	55.6	49.3
PandaGPT (13B)	54.2	58.8	50.6	53.3	45.5	43.8	55.0	55.1	47.0	46.7	44.4	51.3
Valley (7B)	16.7	14.4	23.2	16.7	27.3	19.3	28.3	21.8	22.0	26.7	22.2	22.0
VideoChat2 (7B)	56.9	72.5	**57.0**	45.6	56.6	50.5	53.0	57.7	54.5	46.7	55.6	55.6
Video-ChatGPT (7B)	61.1	68.0	48.1	50.0	45.5	49.0	49.3	47.4	46.2	56.7	44.4	51.8
Video-LLaMA (13B)	65.3	80.4	48.1	45.6	52.5	44.3	52.0	50.0	43.2	55.0	55.6	53.5
Caption Generation												
Baseline												
Human	100		86.7		100		100	100				97.3
Random	28.3	29.2	28.8	27.2	30.8	33.2	32.1	29.5	28.8	26.7	29.2	30.0
Image LLM												
LLaVA-1.5 (13B)	**56.2**	77.9	**36.1**	20.8	25.0	24.6	33.0	41.3	34.1	35.0	20.8	38.4
SPHINX-v2 (13B)	54.2	**80.9**	23.7	6.7	14.4	23.4	37.2	32.7	29.5	32.5	29.2	34.9
Qwen-VL-Chat (7B)	47.4	77.0	29.4	23.3	25.0	**32.0**	34.8	28.8	34.7	30.0	37.5	37.3
Video LLM												
TimeChat (7B)	43.2	68.1	**36.1**	24.2	25.0	31.6	**38.8**	32.7	**41.5**	36.2	37.5	**38.9**
Video-LLaVA (7B)	33.3	67.2	29.7	25.8	18.2	25.8	38.2	31.7	36.9	26.2	41.7	34.8
LLaMA-VID (7B)	38.5	66.7	28.8	25.8	18.9	23.4	35.5	36.5	35.2	37.5	33.3	34.8
mPLUG-Owl (7B)	38.0	54.4	28.8	**26.7**	35.6	27.7	31.2	33.7	38.6	35.0	37.5	34.4
PandaGPT (13B)	26.0	21.6	28.2	19.2	21.2	28.5	29.8	30.8	36.4	25.0	37.5	27.5
Valley (7B)	25.0	24.5	23.7	11.7	19.7	23.0	35.8	31.7	29.0	25.0	37.5	26.3
VideoChat2 (7B)	45.8	61.8	32.9	25.8	**40.2**	28.9	34.2	**43.3**	38.1	**47.5**	37.5	38.5
Video-ChatGPT (7B)	26.0	54.9	30.4	23.3	20.5	26.6	31.8	36.5	34.1	30.0	33.3	31.8
Video-LLaMA (13B)	41.7	66.2	23.1	16.7	15.2	13.3	38.5	28.8	34.7	33.8	**50.0**	32.2

Table 8: Results of the evaluation experiments. The best and second-best MLLM results are **bold** and underlined, respectively.

	Image LLM			Video LLM							
	LLaVA-1.5 13B	SPHINX-v2 13B	Qwen-VL-Chat 7B	V-LLaVA 7B	LLaMA-VID 7B	mPLUG-Owl 7B	PandaGPT 13B	Valley 7B	VideoChat2 7B	V-ChatGPT 7B	V-LLaMA 13B
Multi-Choice QA											
Prompt 1											
Avg Acc	42.8	50.9	50.6	44.7	35.3	40.0	31.1	31.8	**51.1**	35.2	33.9
Match Rate	84.2	99.6	46.8	37.9	62.9	3.1	6.4	3.5	**100.0**	1.3	0.6
Prompt 2											
Avg Acc	47.4	50.6	**51.1**	45.6	38.0	36.4	34.4	29.6	42.9	37.7	31.3
Match Rate	99.9	**100.0**	98.5	**100.0**	97.0	13.7	3.9	0.4	0.0	0.2	3.3
Caption Matching											
Prompt 1											
Avg Acc	59.5	59.2	63.1	**63.7**	53.6	49.3	51.3	22.0	55.6	51.8	53.5
Match Rate	91.2	89.3	91.6	76.6	44.5	15.8	30.7	11.2	**95.3**	7.5	0.1
Prompt 2											
Avg Acc	**64.3**	**64.3**	64.1	63.3	56.0	48.5	51.6	34.6	53.7	53.7	54.2
Match Rate	98.2	**99.9**	96.0	99.5	68.3	63.3	22.5	3.7	1.5	16.5	0.5

Table 9: Accuracy and match rate when using different answer prompts. Prompt 1 is "*Best Option: (*" for VideoChat2 and "*Best Option:*" for the remaining MLLMs. Prompt 2 is "*Please directly give the best option:*".

You will receive a [X] question, the ground-truth answer and the prediction from a question answering (QA) model. Your task is to determine whether QA model prediction is correct, based on the question and ground-truth answer. If the prediction is correct, respond "Correct". If the prediction is incorrect, respond "Incorrect". [X] Question: [question] Ground-Truth Answer: [ground_truth_answer] Model Prediction: [video_llm_prediction]

Table 10: The prompt used to evaluate *Multi-Choice QA*, *Yes/No QA* and *Caption Matching*, where $[X] \in$ {Multi-Choice, Yes/No, Caption Matching}.

You will receive a video description and a multi-choice question. Your task is to choose the correct answer and briefly explain the reason why you choose the answer. If none of the choice candidates are correct or the video description lacks enough information to answer the question, just answer "None of the choices are correct". Please organize your response in this format:
“‘
Reasoning: [Your reason to obtain the answer]
Answer: [Your answer]
“‘

Here are some examples of video description, multi-choice question and the expected answer:
“‘
Video Description: A person is palying football.
Multi-Choice Question:
What is the person doing in the video?
A. cooking B. palying football C. playing basketball D. reading book
Reasoning: The video description mentions that the person is playing football.
Answer: B. palying football

Video Description: A bird is flying clockwise.
Multi-Choice Question:
In which direction is the bird flying? A. backwark B. counter-clockwise C. clockwise D. downward
Reasoning: The video description mentions that the bird is flying clockwise
Answer: C. clockwise

Video Description: An air balloon is inflating.
Multi-Choice Question:
What is happening to the air balloon? A. exploding B. getting smaller C. flying
Reasoning: The video description mentions that the air balloon is inflating, while none of the choices can be explained as inflating.
Answer: None of the choices are correct
“‘

Video Description:
[video_llm_prediction]
Multi-Choice Question:
[multi_choice_question]
Answer:

Table 11: The prompt used to answer the [multi_choice_question] using the generated video caption as context. The answer from ChatGPT is compared with the ground-truth to assess the correctness of generated caption.

You will receive a piece of meta-information in the form of JSON dictionary. The meta-information consists of a "subject" and a temporal dimension (related to "action", "speed", "direction", "order" or "attribute change"). Your task is to generate 5 multi-choice questions and a correct answer based on the meta-information. Ensure that the 5 questions are diverse in language, diverse in format and diverse in the set of choices.
Ensure that the question can be answered from the given meta-information.

Here are some examples of meta-information and generated questions:
"'
[in_context_examples]
"'

Meta-information: [meta_information]
Generate 5 multi-choice questions and correct answers related to "[temporal_aspect]". Generate the correct answer after every generated question. Separate the questions with the string "[SEP]" and don't list the number of questions.

[in_context_examples]:
Meta-information: {"subject": "boy", "action": "playing basketball"}
Multi-Choice Question:
What is the boy doing in the video?
A. cooking B. singing C. playing basketball
Correct Answer: C. playing basketball

Meta-information: {"subject": "entire video", "speed": "normal speed"}
Multi-Choice Question:
What is the speed of the video? A. normal speed B. time-lapse C. slow motion
Correct Answer: A. normal speed

Meta-information: {"subject": "car", "direction": "turning left"}
Multi-Choice Question:
In which direction is the car driving? A. straightforward B. leftwards C. rightwards
Correct Answer: B. leftwards

Meta-information: {"subject": "girl", "event1": "dressing up", "event2": "leaving the room"}
Multi-Choice Question:
What is the girl doing? A. dressing up and then leaving the room B. entering the room and dressing up C. turning off clothes and then leaving the room D. entering the room and then turning off clothes
Correct Answer: A. dressing up and then leaving the room

Meta-information: {"subject": "balloon", "attribute_change": "exploding"}
Multi-choice Question:
What is happening to the balloon? A. shrinking B. stay in the same shape C. exploding
Correct Answer: C. exploding

Table 12: The prompt used to generate *Multi-Choice QA* instructions. [meta_information] and [temporal_aspect] are dependent on the given meta-information. [in_context_examples] are fixed for all *Multi-Choice QA* instructions.

You will receive information about several multi-choice questions in the form of JSON dictionary. The dictionaries consist of a "question" that describes the question and choices and an "answer" that describes the correct answer. Your task is to generate a positive question and a negative question for each multi-choice question. The positive questions, which are related to the correct "answer" of multi-choice question, should be answered with "yes". The negative questions, which are related to other choices except for the correct "answer", should be answered with "no". Ensure that the generated questions are diverse in language and do NOT fabricate information that does not exist in the given multi-choice question.

Here is an example of multi-choice questions and generated positive and negative questions:
"'
Multi-Choice Questions:
{"question": "What is the person doing? A. singing B. cooking C. sleeping", "answer": "B. cooking"}
{"question": "What is the primary action of the person? A. playing football B. cooking C. sleeping", "answer": "B. cooking"}
{"question": "Which of the following actions best describes the person? A. singing B. cooking C. drinking tea", "answer": "B. cooking"}
Positive Questions:
{"question": "Is the person cooking?", "answer": "yes"}
{"question": "Is the primary action of the person about cooking?", "answer": "yes"}
{"question": "Is cooking best describes the person's action?", "answer": "yes"}
Negative Questions:
{"question": "Is the person sleeping?", "answer": "no"}
{"question": "Is the primary action of the person about playing football?", "answer": "no"}
{"question": "Is drinking tea best describes the person's action?", "answer": "no"}
"'

Multi-Choice Questions:
[multi_choice_questions]
Generate the positive and negative questions in JSON format as shown in the above example:

Table 13: The prompt used to generate *Yes/No QA* instructions. [multi_choice_questions] are generated by ChatGPT and rectified by humans.

Prompt to Generate True/False Captions

You will receive information about two multi-choice questions in the form of JSON dictionary. The dictionary consists of a "question" that describes the question and choices and an "answer" that describes the correct answer. Your task is to generate 1 true caption and 3 false captions. The true caption describes the correct "answer" of multi-choice question. The false captions describe other choices except for the correct "answer".

Ensure that the generated captions are diverse in language and do NOT fabricate information that does not exist in the given multi-choice questions.

Here is an example of multi-choice questions and generated true and false captions:

‘‘‘

Multi-Choice Questions: {"question": "What is the person doing? A. singing B. cooking C. sleeping", "answer": "B. cooking"}

{"question": "What is the action shown in the video? A. drawing B. cooking C. reading", "answer": "B. cooking"}

True Caption:

A person is cooking.

False Captions:

A person is sleeping.

A video showing a person singing.

The person is reading.

‘‘‘

Multi-Choice Questions:

[multi_choice_questions]

Generate the true and false captions, do NOT show the multi-choice question in your response:

Instruction Templates for Caption Matching

- Which caption matches the video better? Caption A: [caption_a] Caption B: [caption_b]
- Which description is a more suitable match for the video? Option 1: [caption_a] Option 2: [caption_b]
- Which sentence better captures the essence of the video? Sentence A: [caption_a] Sentence B: [caption_b]

Table 14: The prompt used to generate Ture/False captionis (upper) and the instruction templates for *Caption Matching* (lower). True and False captions are randomly inserted into [caption_a] or [caption_b]. [multi_choice_questions] are generated by ChatGPT and rectified by humans.

Instruction Template for Caption Generation:
[task_description]
Information A: {"subject": [subject], "[temporal_aspect]": [option_a]}
Information B: {"subject": [subject], "[temporal_aspect]": [option_b]}
Information C: {"subject": [subject], "[temporal_aspect]": [option_c]}
Information D: {"subject": [subject], "[temporal_aspect]": [option_d]}

Templates for [task_description]:
- You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly.
- A video and multiple pieces of information will be provided to you. One of these pieces of information matches the content of the video, while the remaining ones do not. Your objective is to pinpoint the information that is in harmony with the video and craft a suitable video caption.
- You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly. Ensure that the generated video caption is brief.
- A video and multiple pieces of information will be provided to you. One of these pieces of information matches the content of the video, while the remaining ones do not. Your objective is to pinpoint the information that is in harmony with the video and craft a suitable video caption. Ensure that the generated video caption is brief.

Table 15: *Caption Generation* instruction templates. [subject] and [temporal_aspect] are obtained from the meta-information. The [options] are derived from the *Multi-Choice QA* instructions. Every [task_description] template will be combined with the candidate information to construct different task instructions.

Meta-Information:
{"subject":"ice cream", "attribute change":"melting", "fg-aspect":"size & shape change"}

Static Contents: food

Multi-Choice QA:
- What is happening to the ice cream? A. melting B. freezing C. evaporating D. solidifying
- How is the ice cream changing? A. changing color B. being eaten out C. turning into liquid D. solidifying into a ball
- Which term best describes the state change of the ice cream? A. evaporation B. solidifying C. melting

Yes/No QA:
- Is the ice cream melting?
- Is the ice cream turning into liquid?
- Is the ice cream freezing?
- Is the ice cream evaporating?

Caption Matching:
- Which caption matches the video better? Caption A: The ice cream is melting. Caption B: The ice cream is freezing.
- Which description is a more suitable match for the video? Option 1: The ice cream is evaporating. Option 2: The ice cream is melting.
- Which sentence better captures the essence of the video? Sentence A: The ice cream is melting. Sentence B: The ice cream is solidifying.

Caption Generation:
- You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly.

  Information A: {'subject':'ice cream', 'attribute_change':'melting'}
  Information B: {'subject':'ice cream', 'attribute_change':'freezing'}
  Information C: {'subject':'ice cream', 'attribute_change':'evaporating'}
  Information D: {'subject':'ice cream', 'attribute_change':'solidifying'}

Table 16: One data example in TempCompass. For each task type, we collect multiple instructions. Due to space limitation, only one instruction is shown for the caption generation task.

Meta-Information:
{"subject":"clock hands", "direction":"moving counter-clockwise", "fg-aspect":"object direction"}

Static Contents: artifacts

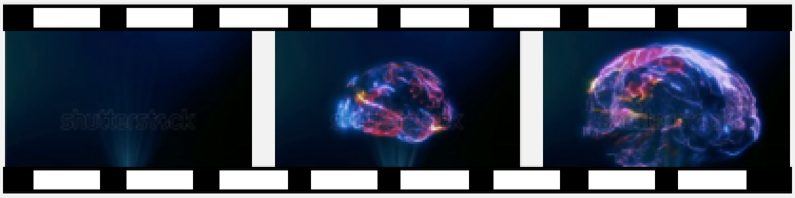
Multi-Choice QA:
- What is the direction in which the clock hands are moving? A. counterclockwise B. clockwise C. stationary
- How are the clock hands changing their orientation? A. standing still B. moving clockwise C. moving anti-clockwise
- In what direction are the clock hands moving? A. clockwise B. alternating between clockwise and counterclockwise C. counterclockwise

Yes/No QA:
- Are the clock hands moving counterclockwise?
- Are the clock hands moving clockwise?

Caption Matching:
- Which description is a more suitable match for the video? Option 1: The clock hands are moving clockwise. Option 2: The clock hands are moving counterclockwise.
- Which sentence better captures the essence of the video? Sentence A: The clock hands are moving counterclockwise. Sentence B: The clock hands are moving clockwise.
- Which caption matches the video better? Caption A: The clock hands are rotating counterclockwise. Caption B: The clock hands are rotating clockwise.

Caption Generation:
- You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly.
  Information A: {'subject':'clock hands', 'direction':'counterclockwise'}
  Information B: {'subject':'clock hands', 'direction':'clockwise'}
  Information C: {'subject':'clock hands', 'direction':'stationary'}

Table 17: One data example in TempCompass. For each task type, we collect multiple instructions. Due to space limitation, only one instruction is shown for the caption generation task.



Meta-Information: {"subject":"entire video", "direction":"zoom into a 3d digital brain", "fg-aspect":"camera direction"}

Static Contents: abstract

Multi-Choice QA:
- What is happening in the video? A. Zoom out from a 3D digital brain B. Standing still before a 3D digital brain C. Zoom into a 3D digital brain
- In which direction is the camera moving in the video? A. zooming in B. downwards C. upwards D. zooming out
- How would you describe the trajectory of the video's direction? A. panning right B. panning left C. zooming out D. zooming in

Yes/No QA:
- Is the camera zooming into a 3D digital brain?
- Is the camera moving in the video by zooming in?
- Is the camera standing still before a 3D digital brain?
- Is the camera zooming out in the video?

Caption Matching:
- Which caption matches the video better? Caption A: The camera is standing still before a 3D digital brain. Caption B: The camera is zooming into a 3D digital brain.
- Which description is a more suitable match for the video? Option 1: The camera is zooming out from a 3D digital brain. Option 2: The camera is zooming into a 3D digital brain.
- Which sentence better captures the essence of the video? Sentence A: The camera is zooming into a 3D digital brain. Sentence B: The camera is moving downwards.

Caption Generation:
- You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly.
  Information A: {'subject':'entire video', 'direction':'Zoom out from a 3D digital brain'}
  Information B: {'subject':'entire video', 'direction':'Standing still before a 3D digital brain'}
  Information C: {'subject':'entire video', 'direction':'Zoom into a 3D digital brain'}

Table 18: One data example in TempCompass. For each task type, we collect multiple instructions. Due to space limitation, only one instruction is shown for the caption generation task.

Meta-Information: {"subject":"entire video", "event1":"a person is kneading dough", "event2":"a girl is jumping into water"}

Static Contents: food, people, natural objects

Multi-Choice QA:
- What is the sequence of events in the video? A. a person is kneading dough followed by a girl jumping into water B. a girl jumping into water followed by a person kneading dough C. a girl jumping into water while a person kneading dough
- What is happening first in the video? A. Both events occur at the same time B. a person is kneading dough C. a girl is jumping into water
- What event occurs second in the video? A. a person is kneading dough B. a girl is jumping into water C. Both events occur at the same time

Yes/No QA:
- Is the sequence of events in the video a person kneading dough followed by a girl jumping into water?
- Is the sequence of events in the video a girl jumping into water followed by a person kneading dough?

Caption Matching:
- Which description is a more suitable match for the video? Option 1: In the video, a person is kneading dough followed by a girl jumping into water. Option 2: A girl jumping into water while a person kneading dough.
- Which sentence better captures the essence of the video? Sentence A: A girl is jumping into water, followed by a person kneading dough. Sentence B: In the video, a person is kneading dough followed by a girl jumping into water.
- Which caption matches the video better? Caption A: In the video, a person is kneading dough followed by a girl jumping into water. Caption B: A person is kneading dough after a girl jumping into water.

Caption Generation:
- You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly.

Information A: {'subject':'entire video', 'order':'a person is kneading dough followed by a girl jumping into water'}

Information B: {'subject':'entire video', 'order':'a girl jumping into water followed by a person kneading dough'}

Information C: {'subject':'entire video', 'order':'a girl jumping into water while a person kneading dough'}

Table 19: One data example in TempCompass. For each task type, we collect multiple instructions. Due to space limitation, only one instruction is shown for the caption generation task.

Multi-Choice QA:
Q: What is happening in the video? A. A person drops down the pineapple B. A person pushes forward the pineapple C. A person rotates the pineapple D. A person picks up the pineapple
Ground-Truth: D. A person picks up the pineapple
SPHINX-v2: D (✓)
Video-LLaVA: D (✓)
Video-ChatGPT: The best option is to pick up the pineapple. (✓)

Yes/No QA:
Q: Is a person picking up the pineapple?
Ground-Truth: yes
SPHINX-v2: Yes (✓)
Video-LLaVA: Yes (✓)
Video-ChatGPT: Yes, a person is picking up the pineapple. (✓)

Caption Matching:
Q: Which sentence better captures the essence of the video? Sentence A: A person is dropping down the pineapple. Sentence B: A person is picking up the pineapple.
Ground-Truth: Sentence B: A person is picking up the pineapple.
SPHINX-v2: Sentence B (✓)
Video-LLaVA: Sentence B: A person is picking up the pineapple. (✓)
Video-ChatGPT: Sentence B captures the essence of the video better as it shows the person picking up the pineapple, which is the main focus of the video. (✓)

Caption Generation:
Ground-Truth: Information D: {'subject': 'person', 'direction': 'A person picks up the pineapple'}
SPHINX-v2: A person is holding a pineapple on a pink table. (✓)
Video-LLaVA: A person picks up the pineapple and drops it down. (✓)
Video-ChatGPT: In the video, a person is seen dropping a pineapple, pushing forward a pineapple, rotating a pineapple, and picking up a pineapple. The person is seen using their hands to interact with the pineapple. The video shows the person's actions and the pineapple's movement, but the exact reason for the actions is not clear from the video. (✗)

Table 20: An example of MLLM responses and evaluation results of the *Direction* aspect. The ✓and ✗in the parentheses are assessed by our automatic evaluation method. The *Caption Generation* task instruction is discarded for simplicity.

Multi-Choice QA:
Q: What is happening in the video? A. A person drops down the pineapple B. A person pushes forward the pineapple C. A person rotates the pineapple D. A person picks up the pineapple
Ground-Truth: A. A person drops down the pineapple
SPHINX-v2: D (✗)
Video-LLaVA: c (✗)
Video-ChatGPT: The best option would be to pick up the pineapple and place it on the table. (✗)

Yes/No QA:
Q: Is a person picking up the pineapple?
Ground-Truth: no
SPHINX-v2: Yes (✗)
Video-LLaVA: Yes (✗)
Video-ChatGPT: Yes, a person is picking up the pineapple. (✗)

Caption Matching:
Q: Which sentence better captures the essence of the video? Sentence A: A person is dropping down the pineapple. Sentence B: A person is picking up the pineapple.
Ground-Truth: Sentence A: A person is dropping down the pineapple.
SPHINX-v2: Sentence B (✗)
Video-LLaVA: Sentence B: A person is picking up the pineapple. (✗)
Video-ChatGPT: Sentence B captures the essence of the video better as it shows the person picking up the pineapple, which is the main action in the video. (✗)

Caption Generation:
Ground-Truth: Information A: {'subject': 'person', 'direction': 'A person drops down the pineapple'}
SPHINX-v2: A person is holding a pineapple on a pink surface. (✗)
Video-LLaVA: A person drops down the pineapple, and then a person pushes forward the pineapple. After that, a person rotates the pineapple, and finally, a person picks up the pineapple. (✗)
Video-ChatGPT: In the video, a person is seen dropping a pineapple, pushing forward a pineapple, rotating a pineapple, and picking up a pineapple. The person is seen using their hands to interact with the pineapple. The video shows the person's actions and the pineapple's movement, but the exact reason for the actions is not clear from the video. (✗)

Table 21: An example of MLLM responses and evaluation results of the *Direction* aspect. The ✓and ✗in the parentheses are assessed by our automatic evaluation method. The *Caption Generation* task instruction is discarded for simplicity.

Multi-Choice QA:
Q: What is happening to the ice cream? A. melting B. freezing C. evaporating D. solidifying
Ground-Truth: A. melting
SPHINX-v2: A (✓)
Video-LLaVA: A. melting (✓)
Video-ChatGPT: The best option is A. melting, as the ice cream is melting in the image. (✓)

Yes/No QA:
Q: Is the ice cream melting?
Ground-Truth: yes
SPHINX-v2: Yes (✓)
Video-LLaVA: Yes. (✓)
Video-ChatGPT: Yes, the ice cream is melting. (✓)

Caption Matching:
Q: Which sentence better captures the essence of the video? Sentence A: The ice cream is melting. Sentence B: The ice cream is solidifying.
Ground-Truth: Sentence A: The ice cream is melting.
SPHINX-v2: The ice cream is melting. (✓)
Video-LLaVA: Sentence B: The ice cream is solidifying. (✗)
Video-ChatGPT: Sentence A, T̈he ice cream is melting,c̈aptures the essence of the video more accurately as it shows the ice cream melting and dripping off the table. (✓)

Caption Generation:
Ground-Truth: Information A: {'subject': 'ice cream', 'attribute_change': 'melting'}
SPHINX-v2: A scoop of ice cream melting on a pink table. (✓)
Video-LLaVA: A scoop of ice cream melts on a pink background. (✓)
Video-ChatGPT: The video shows a close-up of a melting ice cream cone, which is a consistent piece of information with the video. (✓)

Table 22: An example of MLLM responses and evaluation results of the *Attribute Change* aspect. The ✓and ✗in the parentheses are assessed by our automatic evaluation method. The *Caption Generation* task instruction is discarded for simplicity.

Multi-Choice QA:
Q: What is happening to the ice cream? A. melting B. freezing C. evaporating D. solidifying
Ground-Truth: D. solidifying
SPHINX-v2: A (✗)
Video-LLaVA: A. melting (✗)
Video-ChatGPT: The best option is A. melting, as the ice cream is shown melting in the video.(✗)

Yes/No QA:
Q: Is the ice cream melting?
Ground-Truth: no
SPHINX-v2: Yes (✗)
Video-LLaVA: Yes. (✗)
Video-ChatGPT: Yes, the ice cream is melting. (✗)

Caption Matching:
Q: Which sentence better captures the essence of the video? Sentence A: The ice cream is melting. Sentence B: The ice cream is solidifying.
Ground-Truth: Sentence B: The ice cream is solidifying.
SPHINX-v2: The ice cream is melting. (✗)
Video-LLaVA: Sentence A: The ice cream is melting. (✗)
Video-ChatGPT: Sentence A, T̈he ice cream is melting,c̈aptures the essence of the video better as it shows the ice cream melting and dripping off the table. (✗)

Caption Generation:
Ground-Truth: Information D: {'subject': 'ice cream', 'attribute_change': 'solidifying'}
SPHINX-v2: A scoop of ice cream melting on a pink table. (✗)
Video-LLaVA: The video shows a close-up of ice cream melting on a pink surface. The ice cream is shown melting in a circular motion, with the liquid dripping off the surface. The video emphasizes the process of melting and evaporating, with the ice cream gradually solidifying. (✗)
Video-ChatGPT: The video shows a close-up of an ice cream cone melting. This information is consistent with the video and is reflected in the generated caption. (✗)

Table 23: An example of MLLM responses and evaluation result of the *Attribute Change* aspect. The ✓and ✗in the parentheses are assessed by our automatic evaluation method. The *Caption Generation* task instruction is discarded for simplicity.

Multi-Choice QA:
Q: In relation to the speedometer on the left, how does the speedometer in the center appear? A. stationary B. moving slower C. moving faster
Ground-Truth: C. moving faster
SPHINX-v2: C (✓)
Video-LLaVA: The speedometer in the center appears to be moving faster than the one on the left. (✓)
Video-ChatGPT: The speedometer in the center appears to be moving slower than the one on the left. (✗)

Yes/No QA:
Q: Is the speedometer in the center moving faster than the one on the left?
Ground-Truth: yes
SPHINX-v2: yes (✓)
Video-LLaVA: Yes. (✓)
Video-ChatGPT: Yes, the speedometer in the center is moving faster than the one on the left. (✓)

Caption Matching:
Q: Which description is a more suitable match for the video? Option 1: The speedometer in the center is moving faster than the one on the left. Option 2: The speedometer in the center is moving slower than the one on the left.
Ground-Truth: Option 1: The speedometer in the center is moving faster than the one on the left.
SPHINX-v2: The speedometer in the center is moving faster than the one on the right. (✗)
Video-LLaVA: Option 1: The speedometer in the center is moving faster than the one on the left. (✓)
Video-ChatGPT: Option 1 is the most suitable description for the video, as the speedometer in the center is moving faster than the one on the left. (✓)

Caption Generation:
Ground-Truth: Information C: {'subject': 'the speedometer in the center', 'speed': 'moving faster'}
SPHINX-v2: A speedometer with the needle pointing to the number 0. (✗)
Video-LLaVA: The video shows a speedometer in the center of the screen, which is stationary. The speedometer is then shown moving slower, and finally moving faster. The video does not provide any additional information about the speedometer. (✗)
Video-ChatGPT: The video shows a speedometer in the center, with two other speedometers on either side. The speedometer in the center is stationary, while the speedometers on either side are moving slower and faster, respectively. (✗)

Table 24: An example of MLLM responses and evaluation results of the *Speed* aspect. The ✓and ✗in the parentheses are assessed by our automatic evaluation method. The *Caption Generation* task instruction is discarded for simplicity.

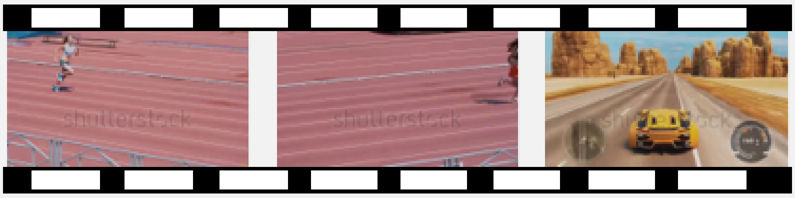
Multi-Choice QA:
Q: What is the speed of the video? A. time-lapse B. slow motion C. normal speed
Ground-Truth: B. slow motion
SPHINX-v2: C (✗)
Video-LLaVA: A. time-lapse (✗)
Video-ChatGPT: The correct answer is A. time-lapse. (✗)

Yes/No QA:
Q: Is the video playing in slow motion?
Ground-Truth: yes
SPHINX-v2: No (✗)
Video-LLaVA: Yes. (✓)
Video-ChatGPT: Yes, the video is playing in slow motion. (✓)

Caption Matching:
Q: Which description is a more suitable match for the video? Option 1: The video is playing in fast motion. Option 2: The video is playing in slow motion.
Ground-Truth: Option 2: The video is playing in slow motion.
SPHINX-v2: The video is playing in slow motion. (✓)
Video-LLaVA: Option 2: The video is playing in slow motion. (✓)
Video-ChatGPT: The video is playing in fast motion. (✗)

Caption Generation:
Ground-Truth: Information B: {'subject': 'entire video', 'speed': 'slow motion'}
SPHINX-v2: A man is charging his car with an electric charger. (✗)
Video-LLaVA: A man is seen charging his electric car in a parking lot. He uses a portable charger to charge his car. The video shows the process of charging the car and the man's actions. (✗)
Video-ChatGPT: Based on the provided information, the video is a time-lapse of a man pumping gas into a car. (✗)

Table 25: An example of MLLM responses and evaluation results of the *Speed* aspect. The ✓and ✗in the parentheses are assessed by our automatic evaluation method. The *Caption Generation* task instruction is discarded for simplicity.



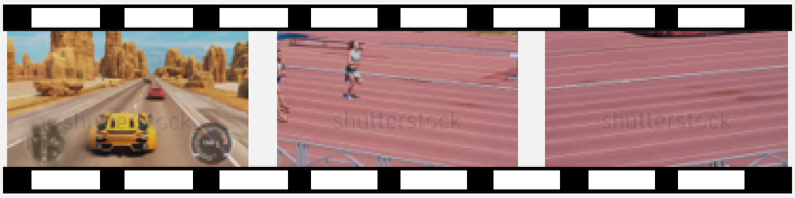
Multi-Choice QA:
Q: Which event happens first in the video? A. They happen simultaneously B. Cars are racing C. Women are running
Ground-Truth: C. Women are running
SPHINX-v2: C (✓)
Video-LLaVA: Cars are racing first in the video. (✗)
Video-ChatGPT: The events in the video occur in the following order: cars racing, then women running. (✗)

Yes/No QA:
Q: Do the events occur in the order of women running, then cars racing?
Ground-Truth: yes
SPHINX-v2: no (✗)
Video-LLaVA: Yes. (✓)
Video-ChatGPT: Yes, the events occur in the order of women running, then cars racing. (✓)

Caption Matching:
Q: Which description is a more suitable match for the video? Option 1: The video shows women running first, followed by cars racing. Option 2: Woman running and car racing happen simultaneously.
Ground-Truth: Option 1: The video shows women running first, followed by cars racing.
SPHINX-v2: Option 2 (✗)
Video-LLaVA: Option 2: The video shows a woman running and cars racing simultaneously. (✗)
Video-ChatGPT: Option 2 is the most suitable match for the video, as it shows both a woman running and cars racing simultaneously. (✗)

Caption Generation:
Ground-Truth: Information B: {'subject': 'entire video', 'order': 'Women running, then cars racing'}
SPHINX-v2: A woman running on a track next to a car race. (✗)
Video-LLaVA: The video starts with a woman running on a track, followed by several other women running. The scene then shifts to cars racing on a highway. The video then shows the women running again, followed by the cars racing. The video ends with the women running again. The video showcases the beauty of nature and the excitement of racing. The video is a perfect blend of nature and technology, making it a must-watch for all car enthusiasts. (✓)
Video-ChatGPT: The video shows a group of women running on a track while cars are racing nearby. The cars are racing in the background while the women are running in the foreground. This suggests that the video captures a dynamic scene where both the cars and the women are engaged in their respective activities. (✗)

Table 26: An example of MLLM responses and evaluation results of the *Event Order* aspect. The ✓and ✗in the parentheses are assessed by our automatic evaluation method. The *Caption Generation* task instruction is discarded for simplicity.



Multi-Choice QA:
Q: Which event happens first in the video? A. They happen simultaneously B. Cars are racing C. Women are running
Ground-Truth: B. Cars are racing
SPHINX-v2: C (✗)
Video-LLaVA: The video shows that the women are running first, followed by the cars racing.(✗)
Video-ChatGPT: The events in the video happen simultaneously. The women are running on the road, and at the same time, the cars are racing on the road. (✗)

Yes/No QA:
Q: Do the events occur in the order of cars racing, then women running?
Ground-Truth: yes
SPHINX-v2: no (✗)
Video-LLaVA: Yes (✓)
Video-ChatGPT: Yes. (✓)

Caption Matching:
Q: Which caption matches the video better? Caption A: The cars are racing first, followed by women running. Caption B: Woman running and car racing happen simultaneously.
Ground-Truth: Caption A: The cars are racing first, followed by women running.
SPHINX-v2: Caption A (✓)
Video-LLaVA: Caption B matches the video better as it shows the women running and the cars racing simultaneously. (✗)
Video-ChatGPT: Caption A is the best option as it accurately describes the events in the video. The cars are racing first, followed by women running. (✓)

Caption Generation:
Ground-Truth: Information C: {'subject': 'entire video', 'order': 'Cars racing, then women running'}
SPHINX-v2: The women are running and cars are racing. (✗)
Video-LLaVA: The video starts with a yellow sports car driving down a highway, followed by a group of women running on a track. The women are seen running in slow motion, and the video captures their movements in detail. The scene then shifts to a group of cars racing down the same highway, with the women running alongside them. The video captures the excitement and energy of both the car race and the women's run, showcasing the beauty of nature and the thrill of competition. (✓)
Video-ChatGPT: The information that is consistent with the video is that there are women running and cars racing in the video. The video shows a group of women running on a road, and at the same time, cars are racing around them. This suggests that the video captures a dynamic and exciting moment where both the women and the cars are engaged in their respective activities. (✗)

Table 27: An example of MLLM responses and evaluation results of the *Event Order* aspect. The ✓and ✗in the parentheses are assessed by our automatic evaluation method. The *Caption Generation* task instruction is discarded for simplicity.

Multi-Choice QA:
Q: Which direction is the camera moving before the stone building? A. rightwards B. upwards C. leftwards D. downwards
Ground-Truth: B. upwards
SPHINX-v2: B (✓)
Video-LLaVA: The camera is moving upwards before the stone building. (✓)
Video-ChatGPT: The camera is moving rightwards before the stone building. (✗)

Yes/No QA:
Q: Is the camera moving upwards before the stone building?
Ground-Truth: yes
SPHINX-v2: Yes (✓)
Video-LLaVA: Yes. (✓)
Video-ChatGPT: Yes, the camera is moving upwards towards the stone building. (✓)

Caption Matching:
Q: Which caption matches the video better? Caption A: The camera is moving downwards before the stone building. Caption B: The movement in the video is upwards.
Ground-Truth: Caption B: The movement in the video is upwards.
SPHINX-v2: Caption A (✗)
Video-LLaVA: Caption B: The movement in the video is upwards. (✓)
Video-ChatGPT: Caption A, which states that the camera is moving downwards before the stone building, is the best match for the video. (✗)

Caption Generation:
Ground-Truth: Information B: {'subject': 'entire video', 'direction': 'upwards'}
SPHINX-v2: The video is a great resource for understanding the subject matter. (✗)
Video-LLaVA: The video showcases a beautiful and historic location with a clock tower and a church. The camera pans around the area, capturing the stunning architecture and scenery. The video also includes a shot of a mountain in the background, adding to the overall beauty of the location. (✗)
Video-ChatGPT: The video features a large building with a clock tower, which is located in a mountainous area. The building is surrounded by a forest, and the clock tower is situated on top of a hill. The video also shows a person walking towards the building, and a car driving by. The building appears to be old and has a rustic appearance. The clock tower is visible in the background, and the person is walking towards it. The video is shot from a distance, capturing the overall landscape and architecture of the area. (✓)

Table 28: An example of MLLM responses and evaluation results of the *Direction* aspect. The ✓and ✗in the parentheses are assessed by our automatic evaluation method. The *Caption Generation* task instruction is discarded for simplicity.

Multi-Choice QA:
Q: What is the primary movement of the sun? A. descending B. standing still C. rising
Ground-Truth: C. rising
SPHINX-v2: C (✓)
Video-LLaVA: The sun is seen rising in the sky, which is the primary movement of the sun in the video. (✓)
Video-ChatGPT: The primary movement of the sun in the video is descending, as it is seen setting over the ocean. (✗)

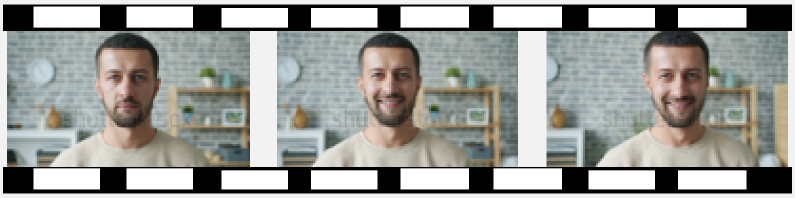
Yes/No QA:
Q: Is rising the primary movement of the sun?
Ground-Truth: yes
SPHINX-v2: No (✗)
Video-LLaVA: Yes. (✓)
Video-ChatGPT: Yes, the sun is rising in the sky. (✓)

Caption Matching:
Q: Which description is a more suitable match for the video? Option 1: The sun is moving downwards. Option 2: The sun is moving upwards.
Ground-Truth: Option 2: The sun is moving upwards.
SPHINX-v2: The sun is setting. (✓)
Video-LLaVA: Option 1: The sun is moving downwards. (✗)
Video-ChatGPT: The sun is moving downwards in the sky. (✗)

Caption Generation:
Ground-Truth: Information B: {'subject': 'sun', 'direction': 'rising'}
SPHINX-v2: The sun is setting in the sky. (✗)
Video-LLaVA: The sun is seen setting in the sky, with clouds moving around it. (✗)
Video-ChatGPT: The consistent information with the video is that the sun is setting, as it is described as a "sunset" in the video. The other two pieces of information, "sun rising" and "sun staying in one place," do not align with the visual content of the video, which shows the sun setting. Therefore, the caption should be "Sun setting in the sky." (✗)

Table 29: An example of MLLM responses and evaluation results of the *Direction* aspect. The ✓and ✗in the parentheses are assessed by our automatic evaluation method. The *Caption Generation* task instruction is discarded for simplicity.

Generated Caption (Video-LLaVA):
A group of people are dancing in a pool while a man is cooking in the background.
Multi-Choice Question:
What are the people doing? A. cooking B. sleeping C. dancing
ChatGPT Reasoning:
The video description mentions that a group of people are dancing in a pool, while cooking is happening in the background.
ChatGPT Answer: C. dancing
ChatGPT Rating: ✓
Human Rating: ✗



Generated Caption (Video-LLaVA):
The man in the video is seen smiling and then showing a serious expression. He then smiles again.
Multi-Choice Question:
In what order does the man's expression change? A. The man doesn't change his expression B. Smiling, then showing a serious expression C. Showing a serious expression, then smiling
ChatGPT Reasoning:
The video description states that the man is seen smiling and then showing a serious expression, followed by smiling again. This implies that the man's expression changes from smiling to showing a serious expression, and then back to smiling.
ChatGPT Answer: C. Showing a serious expression, then smiling
ChatGPT Rating: ✓
Human Rating: ✗

Table 30: Examples showing our LLM-evaluation fail to detect the unsatisfactory caption in terms of hallucination. The hallucinated content in MLLM response is highlighted in red. The correct answer to the multi-choice question is highlighted in green.