TemporalBench: Benchmarking Fine-grained Temporal Understanding for Multimodal Video Models

Mu Cai[1,*], Reuben Tan[2], Jianrui Zhang[1], Bocheng Zou[1], Kai Zhang[3], Feng Yao[4], Fangrui Zhu[5], Jing Gu[6], Yiwu Zhong[7], Yuzhang Shang[8], Yao Dou[9], Jaden Park[1], Jianfeng Gao[2,†], Yong Jae Lee[1,†], Jianwei Yang[2,†]

[1]University of Wisconsin-Madison [2]Microsoft Research, Redmond
[3] Ohio State University [4] University of California, San Diego [5] Northeastern University
[6] University of California, Santa Cruz [7] Chinese University of Hong Kong
[8] Illinois Institute of Technology [9] Georgia Institute of Technology

https://TemporalBench.github.io/

Abstract

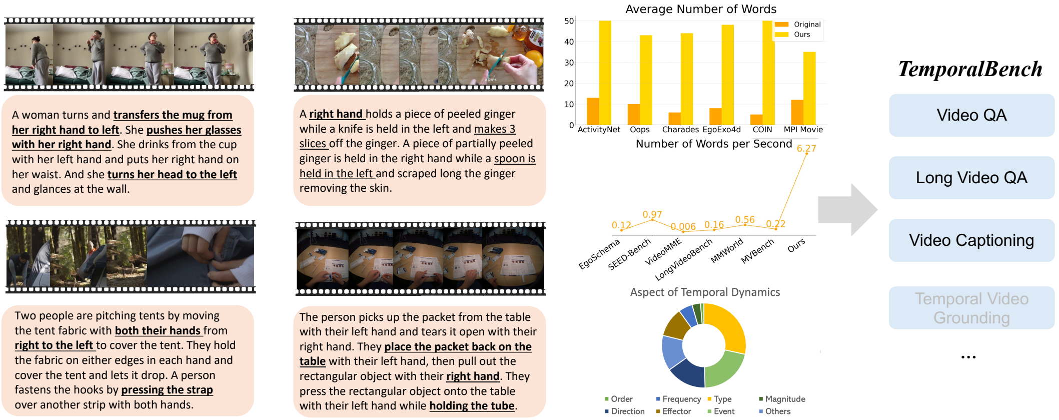
Understanding fine-grained temporal dynamics is crucial for multimodal video comprehension and generation. Due to the lack of fine-grained temporal annotations, existing video benchmarks mostly resemble static image benchmarks and are incompetent at evaluating models for temporal understanding. In this paper, we introduce *TemporalBench*, a new benchmark dedicated to evaluating **fine-grained temporal understanding** in videos. *TemporalBench* consists of $\sim$10K video question-answer pairs, derived from $\sim$2K high-quality human annotations detailing the temporal dynamics in video clips. As a result, our benchmark provides a unique testbed for evaluating various temporal understanding and reasoning abilities such as *action frequency, motion magnitude, event order, etc.* Moreover, it enables evaluations on various tasks like both video question answering and captioning, both short and long video understanding, as well as different models such as multimodal video embedding models and text generation models. Results show that state-of-the-art models like GPT-4o achieve only 38.5% question answering accuracy on *TemporalBench*, demonstrating a significant gap ($\sim$ 30%) between humans and AI in temporal understanding. Furthermore, we notice a critical pitfall for multi-choice QA where LLMs can detect the subtle changes in negative captions and find a "centralized" description as a cue for its prediction, where we propose Multiple Binary Accuracy (MBA) to correct such bias. We hope that *TemporalBench* can foster research on improving models' temporal reasoning capabilities. Both dataset and evaluation code will be made available.

1 Introduction

The ability to understand and reason about events in videos is a crucial aspect of artificial intelligence, with applications ranging from activity recognition and long-term action anticipation to perception for autonomous driving and robotics. Recently, there has been an emergence of highly capable multimodal generative models, including proprietary ones such as GPT-4o [51] and Gemini [17] as well as open-sources ones [37, 86, 4], that have demonstrated impressive results on existing video

[*]Work done during the internship at Microsoft Research, [†] Equal Advisory Contribution.

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).



A woman turns and **transfers the mug from her right hand to left**. She **pushes her glasses with her right hand**. She drinks from the cup with her left hand and puts her right hand on her waist. And she **turns her head to the left** and glances at the wall.

A **right hand** holds a piece of peeled ginger while a knife is held in the left and makes 3 slices off the ginger. A piece of partially peeled ginger is held in the right hand while a spoon is held in the left and scraped long the ginger removing the skin.

Two people are pitching tents by moving the tent fabric with **both their hands** from **right to the left** to cover the tent. They hold the fabric on either edges in each hand and cover the tent and lets it drop. A person fastens the hooks by **pressing the strap** over another strip with both hands.

The person picks up the packet from the table with their left hand and tears it open with their right hand. They **place the packet back on the table** with their left hand, then pull out the rectangular object with their **right hand**. They press the rectangular object onto the table with their left hand while **holding the tube**.

Figure 1: **The tasks of *TemporalBench***. *TemporalBench* starts from fine-grained video descriptions and supports diverse video understanding tasks including video QA, video captioning, long video understanding, *etc*. It differs from existing benchmarks by the average number of words per video (middle top), word density (center) and the coverage of various temporal aspects (middle bottom).

benchmarks [69, 7, 75, 43]. However, these benchmarks often do not truly evaluate the abilities of the aforementioned models to understand video content due to their generally *coarse-grained* annotations.

The lack of fine-grained temporal details in the annotations often leads to existing video understanding benchmarks suffering from a strong language prior bias. This is similar to observations in visual question answering with images [3]. For example, prior works [60, 31] show that language models such as Flan-T5 [11] and Llama-2/3 [62] perform comparably to video models on EgoSchema [43] and Seed-Bench [31] without using any information from videos. Furthermore, the lack of fine-grained temporal details often results in the single frame bias of current video understanding benchmarks [29]. These benchmarks are often biased toward spatial reasoning, where static information from a single frame suffices to achieve high performance. They often fail to test a model's ability to reason about temporal sequences, leading to inflated evaluations of AI models that are not genuinely capable of understanding temporal events. Specifically, vision-language models (VLMs) [38, 39] that are trained on image-level datasets, including FreeVA [66], IG-VLM [27] and M^3 [6], often outperform their video counterparts on popular video question answering benchmarks such as MSRVTT [69], MSVD [68], and TGIF [25].

To address this limitation, we propose *TemporalBench* (Figure 1), a new video understanding benchmark that evaluates multimodal video models on understanding fine-grained activities, and consists of ∼**10K** question and answer pairs curated from ∼**2K** high-quality human-annotated captions with rich activity details. Unlike static image-based tasks, video understanding requires models to reason effectively about both spatial and temporal information. The temporal dynamics inherent in videos introduce significant complexity, as actions and events often unfold over time and cannot be captured in a single frame.

With this in mind, we designed our benchmark to focus on areas where current models often struggle, emphasizing annotations related to long-range dependencies, fine-grained visual observations, and event progression.

As shown in Figure 2, we first collect video clips from existing video grounding benchmarks that span diverse domains, including procedural videos [61], human activities [28, 16], ego-centric videos [19], movie descriptions [56], professional gymnasium videos (FineGym from [57]), and unexpected humor videos [12]. The positive captions include *rich* and *fine-grained* details about actions and activities, which are annotated by highly qualified Amazon Mechanical Turk (AMT) workers and authors of this paper. Then, we generate the negative captions with respect to the actions using powerful Large Language Models (LLMs) and filter them according to our defined rules. Our resulting *TemporalBench* contains ∼10K video descriptions and matching questions of high quality. Furthermore, the rich temporal context of annotations in our diverse corpus creates a solid foundation for the development of additional benchmarks in related tasks such as spatio-temporal localization and causal inference. We hope that our benchmark can pave the road for further development of multimodal video models capable of fine-grained video understanding and reasoning.

In contrast to existing video benchmarks, *TemporalBench* has the following defining characteristics:

- **Emphasis on fine-grained action understanding**. Due to the highly descriptive video captions, our negative captions highlight fine-grained temporal differences shown in Figure 3, such as *"sliced the ginger three times"* versus *"sliced the ginger twice"*, and *"put on the eyeglasses"* versus *"push the eyeglasses"*.
- **Evaluations on both short (<20 seconds) and long (<20 minute) videos.** Since the videos clips are sampled from existing videos, our benchmark can also support evaluations on long video understanding by concatenating the descriptions of multiple and non-overlapping video clips from the same source video.
- **Extends to video captioning, video grounding, and video generation.** Besides the task of video question answering, the nature of the positive captions in our benchmark allows it to seamlessly extend to evaluation of other tasks such as video temporal grounding and dense captioning.
- **Evaluations of both video embedding and question-answering models.** Given the annotated positive and negative captions in *TemporalBench*, it also supports the evaluation of discriminative and contrastive learning-based models such as XCLIP [47], ImageBind [18] as well as multimodal generative models such as GPT-4o and Gemini.

Furthermore, we notice a critical pitfall for multi-choice QA. If every negative answer choice is generated by changing a small part of the correct answer, the LLM can detect those changes to find a "centralized" description and use that cue for its prediction. Therefore, we propose Multiple Binary Accuracy (MBA) to correct such bias.

Among other observations, our empirical evaluations show that state-of-the-art multimodal video models like GPT-4o only achieve an average accuracy of 38.5% on our benchmark (short videos) using our proposed multiple binary QA accuracy metric, compared to 67.9% obtained by humans. Models show even worse results on long videos. This result highlights that the aforementioned models are able to understand static visual concepts but are still limited in reasoning about the fine-grained temporal relationships of objects and events in videos. More significantly, we highlight a critical issue with using LLMs to answer multi-choice QA.

2 Related Work

Large Multimodal Models. Large Language Models (LLMs) like ChatGPT [49], GPT-4 [50], and Llama [62] have demonstrated impressive reasoning and generalization capabilities for text. The introduction of models that integrate visual data has brought about a significant shift in the landscape of LLMs, such as GPT-4V(ision)[48]. Building upon open-source LLMs [62, 10], a wide range of multimodal models has achieved remarkable progress, led by pioneering models such as LLaVA [37, 38] and MiniGPT-4 [86], which combine LLMs' capabilities with a CLIP [54] based image encoder. Recently, a growing number of LMMs have been developed to handle a wider range of tasks and modalities, such as region-level LMMs [5, 83, 8, 53, 81], 3D LMMs [23], and video LMMs [34, 80, 84].

Multimodal Understanding Benchmarks. The recent significant advancements have resulted in more versatile multimodal models, making it imperative to thoroughly and extensively evaluate their visual understanding and reasoning abilities. Conventional multimodal benchmarks like VQA [3], GQA [24] and VizWiz [20] have been revitalized and used for evaluating the general visual question answering performance for LMMs. Some other question answering benchmarks like TextVQA [58], DocVQA [44] and InfoVQA [45] have also been employed to validate the text-oriented understanding. Recent studies have introduced a variety of new benchmarks, such as SEED-Bench [31], MMBench [40] and MM-Vet [74] for evaluating the models' integrated problem-solving capabilities, and MMMU [77] and MathVista [42] for scientific and mathematical reasoning. In addition, the commonly known hallucination problem also appears in LMMs, and is also investigated in POPE [33], MMHal-Bench [59] and Object HalBench [73], *etc*.

Video Understanding Benchmarks. Recently, an increasing amount of research is transitioning its focus from the image to the video domain. Videos differ from images in that they possess more complex content with temporal dynamics. This unique aspect calls for a different set of metrics and

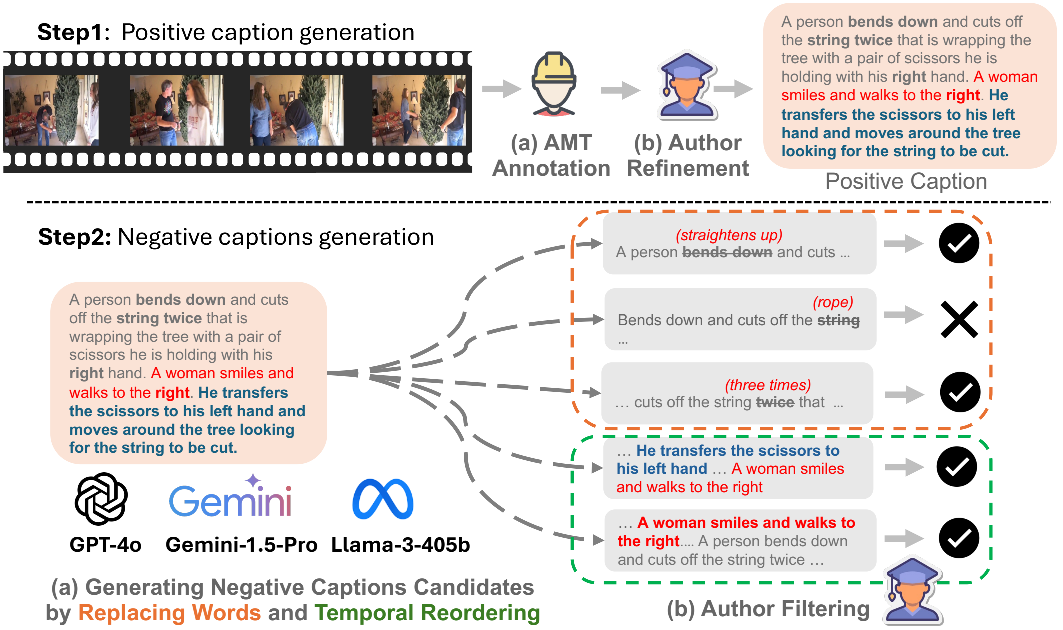

Step1: Positive caption generation
(a) AMT Annotation
(b) Author Refinement
A person bends down and cuts off the string twice that is wrapping the tree with a pair of scissors he is holding with his right hand. A woman smiles and walks to the right. He transfers the scissors to his left hand and moves around the tree looking for the string to be cut.
Positive Caption
Step2: Negative captions generation
A person bends down and cuts off the string twice that is wrapping the tree with a pair of scissors he is holding with his right hand. A woman smiles and walks to the right. He transfers the scissors to his left hand and moves around the tree looking for the string to be cut.
GPT-4o
Gemini-1.5-Pro
Llama-3-405b
(a) Generating Negative Captions Candidates by Replacing Words and Temporal Reordering
(straightens up)
A person ~~bends down~~ and cuts ...
✓
(rope)
Bends down and cuts off the ~~string~~ ...
✗
(three times)
... cuts off the string ~~twice~~ that ...
✓
... He transfers the scissors to his left hand ... A woman smiles and walks to the right
✓
... A woman smiles and walks to the right.... A person bends down and cuts off the string twice ...
✓
(b) Author Filtering

Figure 2: **Overview of the annotation pipeline for *TemporalBench*.** In step 1, we fist collect high-quality captions for the videos using qualified AMT annotators followed by refining them. In step 2, we leverage existing LLMs to generate negative captions by replacing select words and reordering the sequence of actions before filtering them ourselves.

benchmarks. Many efforts have leveraged existing video question answering benchmarks [68, 76, 67] built on top of video-text datasets [7, 69, 79]. More recently, several LMM-oriented benchmarks have been proposed for different aspects such as long-form egocentric understanding with EgoSchema [43], and temporal understanding and ordering like Tempcompass [41]. MV-Bench [32] compiles existing video annotations from different disciplines into a new benchmark, while Video-MME [14] and MMWorld [22] claim to support a comprehensive evaluation of video understanding and world modeling, respectively. Our *TemporalBench* serves the common goal of evaluating models for video understanding but differs in several aspects. On the one hand, we exhaustively curate videos from different domains and ask human annotators to annotate the visual contents with as much detail as possible. On the other hand, we particularly focus on temporal dynamics such as human actions and human-object interactions that exist exclusively in videos and which are crucial for video understanding, reasoning and forecasting. While the ShareGPT4Video dataset [9] also contains long captions, theirs differ from ours by being entirely generated by GPT-4o instead of annotated by humans.

3 *TemporalBench*

Compared to static images, videos inherently contain significantly more fine-grained temporal information, as they capture the unfolding of actions and events over time. Existing multimodal video understanding benchmarks [69] mostly evaluate models' coarse-level understanding of videos. An example from the recent Seed-Bench dataset is the question, *"What action is happening in the video?"* with the answer, *"moving something up."* However, such types of coarse-level video questions have been demonstrated to be easily solved with just a single frame [66] or even by a text-only LLM [60, 43].

Such phenomena arises due to a fundamental limitation in the text descriptions in those benchmarks. As a result of their coarseness, the positive and negative options for video question-answering can usually be distinguished without understanding the temporal dynamics, such as the models only needing to choose between *"The man is cooking"* and *"The man is exercising"*.

To address this limitation, we carefully design a human annotation pipeline to curate highly detailed descriptions about the activities in the videos. Given the detailed video clip descriptions, such as

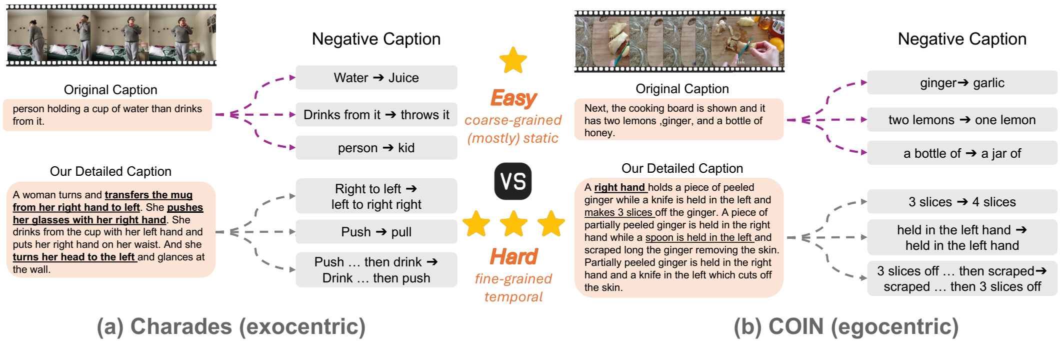

Negative Caption
Water ➔ Juice
Original Caption
person holding a cup of water than drinks from it.
Drinks from it ➔ throws it
person ➔ kid
Easy
coarse-grained
(mostly) static
Our Detailed Caption
A woman turns and transfers the mug from her right hand to left. She pushes her glasses with her right hand. She drinks from the cup with her left hand and puts her right hand on her waist. And she turns her head to the left and glances at the wall.
Right to left ➔
left to right right
Push ➔ pull
Push … then drink ➔
Drink … then push
VS
Hard
fine-grained
temporal
(a) Charades (exocentric)
Original Caption
Next, the cooking board is shown and it has two lemons ,ginger, and a bottle of honey.
Negative Caption
ginger➔ garlic
two lemons ➔ one lemon
a bottle of ➔ a jar of
Our Detailed Caption
A right hand holds a piece of peeled ginger while a knife is held in the left and makes 3 slices off the ginger. A piece of partially peeled ginger is held in the right hand while a spoon is held in the left and scraped long the ginger removing the skin. Partially peeled ginger is held in the right hand and a knife in the left which cuts off the skin.
3 slices ➔ 4 slices
held in the left hand ➔
held in the left hand
3 slices off … then scraped➔
scraped … then 3 slices off
(b) COIN (egocentric)

Figure 3: **Comparison of negative captions generated from the original captions and our detailed captions in *TemporalBench*.** With fine-grained details, the negatives are more difficult and temporal centric.

A right hand holds a piece of peeled ginger while a knife is held in the left and makes 3 slices off the ginger., the negative captions can be curated to truly reflect whether a model understands the temporal dynamics, such as changing *"three slices"* into *"two slices"*. In a nutshell, such highly detailed temporal annotations can be used to carefully examine whether a multimodel video model truly understands the temporal state transition in videos.

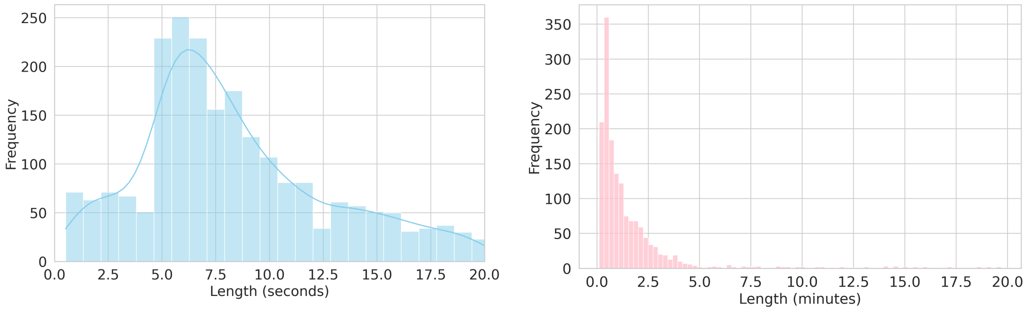
Our benchmark enriches several fundamental video understanding tasks due to its detailed captions:

- **Fine-grained video question answering.** Given a detailed positive caption, multimodal video models need to distinguish it from the associated negative where a slight modification is made to temporal descriptions, *e.g., "push the eyeglasses up"* versus *"pull the eyeglasses down"*, or *"cut 3 slices off"* versus *"cut 2 slices off"*.
- **Fine-grained video captioning.** Our detailed video captions can naturally enrich the video captioning task, different from current video captioning tasks such as MSRVTT [69] which focus on coarse-level descriptions.
- **Long video understanding with fine-grained activity inspection.** Since the video clips are extracted from a long source video, the respective video clip descriptions can be concatenated to form a longer video description which can be pivoted to the long video understanding task, where we find that all current multimodal video models suffer.
- **Dense video-text matching and retrieval.** Our detailed video captions can be naturally employed to evaluate video-language embedding models such as XCLIP [47]. Given a positive caption and several negative captions, we can evaluate whether CLIP [54] based video embedding models can distinguish the subtle differences in captions. In addition, given a set of positive video-text pairs, video retrieval performance can be evaluated, similar to image retrieval on COCO [36] and Flickr30K [72].
- **Video grounding from detailed text descriptions.** Since the video clips are cropped from the source video, with the documented starting and ending time, our benchmark can serve as a fine-grained moment localizing benchmark from text descriptions. This is different from existing video grounding datasets such as Charades-STA [16], COIN [61], Ego4D [19] where the text descriptions are usually very short, possibly resulting in low temporal localization performance due to the vague and coarse descriptions.
- **Text-to-Video (T2V) generation with detailed prompts.** Given our highly detailed description, a T2V generation model can be evaluated by verifying if the generated videos reflect the fine-grained action details.

Next, we detail the dataset curation and evaluation setup for *TemporalBench*.

3.1 Video Collection

We collect video clips from a wide range of sources across diverse domains, where the majority comes from existing video grounding benchmarks. Our dataset includes a wide spectrum of video



(a) Short videos **(b) Long videos**

Figure 4: Video length distribution of (a) short video clips and (b) long videos in *TemporalBench*.

types from seven sources, including (1) procedure videos *e.g.,* COIN [61], (2) human activities *e.g.,* ActivityNet-Captions [75] and Charades [28], (3) ego-centric videos *e.g.,* EgoExo4D [19], (4) movie descriptions [56], (5) professional gymnasium videos *e.g.,* FineGym [57], and (6) unexpected humor videos Oops [12]. We sample around 300 video clips from the validation and test sets of each video dataset, which results in 2K videos. The statistics of *TemporalBench* is shown in Table 1.

We intentionally filter out video clips that (1) are mostly static by leveraging optical flow [13], (2) contain multiple scene transitions by leveraging PySceneDetect [2] and (3) last longer than 20 seconds. We observe that the large amount of information in long videos make it difficult for annotators to provide detailed action descriptions. The distribution of video lengths is shown in Figure 4 (a). Additionally, we remove the audio from the videos during annotation to ensure that all informative signals come solely from the visual frames, preventing the answers from being influenced by the audio.

3.2 Video Caption Annotation Process

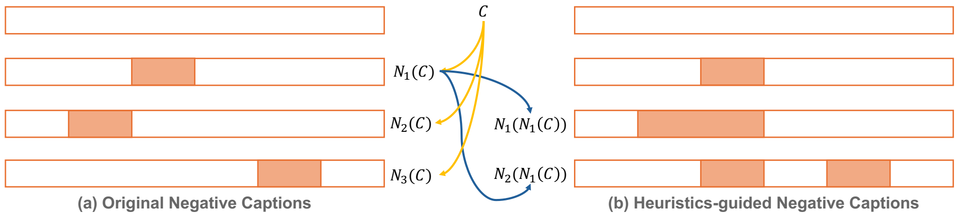
Positive Captions Annotation. We employ a two-stage human labeling process for curating video captions with fine-grained activity descriptions, where the qualified Amazon Mechanical Turk (AMT) workers are first instructed to give a detailed video caption. Then, the authors of this work refine the caption by correcting the mistakes and adding missing details *w.r.t.* the actions. The overall pipeline is shown in Figure 2. All video clips are annotated following the same pipeline except for Finegym [57] as it has already provided accurate and detailed action descriptions for professional gymnasium videos. Consequently, we reuse its annotations.

We first use 3 probing video captioning questions with 2 in-context examples as the onboarding task for AMT master workers. We manually inspect the soundness and amount of temporal details of the AMT worker captions to select high quality AMT video captioning workers. During the annotation process by AMT workers, we also continue to remove the unqualified workers based on the ratio of the captions that authors in this paper refined. In this way, we ensure that the AMT provides a high quality initial point for positive captions.

Negative Caption Annotation. Our negative captions are aimed at confusing multimodal video models with respect to fine-grained activity details, such as changing *"cut a ginger twice using a knife"* to *"cut a ginger three times using a knife"*. We construct negatives upon two granularities: word level and event level. Specifically, word level negatives denote the case where a certain word or phrase is replaced while event level negatives denote the case where the order of two events are reversed. Empirically, we find that LLMs can produce more creative and diverse negatives compared to AMT workers and authors. Therefore, we leverage three leading LLMs, GPT-4o [51], Gemini-1.5-Pro [17] and Llama-3.1-405b [46] to curate a diverse set of negative caption candidates instructed by 3 in-context examples, with up to 9 negatives at word level and 6 negatives at event level.

Afterwards, the authors of this work review those negative caption candidates in the format of multi-choice QA, which results in our complete *TemporalBench* dataset with ∼2K high-quality human-annotated video captions and ∼10K video question-answer pairs.

[2] https://www.scenedetect.com/



(a) Original Negative Captions

(b) Heuristics-guided Negative Captions

Figure 5: An illustration of multi-choice QA with (a) original and (b) heuristics-guided negative captions. Orange blocks indicate the altered contents from the positive option (green box).

3.3 A Pitfall in Multi-choice Question Answering

A conventional approach to evaluate large multimodal models is using the multi-choice question-answering format, which is adopted by the majority of current benchmarks including MMMU [77], MathVista [42], EgoSchema [43] etc. However, indicated by recent studies by [6] and [78], a pure LLM can achieve comparable or even stronger performance on those benchmarks without looking at the visual content at all. Recent studies argue that (1) some questions are not designed well so that the question can be answered without looking at the visual content, or (2) the model memorizes the QA pairs, *i.e.,* data contamination occurs.

While developing our benchmark, we notice another previously ignored but critical pitfall for multi-choice QA. Specifically, if every negative answer choice is generated by changing a small part of the correct answer, the LLM can detect those changes to find a "centralized" description and use that cue for its prediction. To study this, given a positive caption C and its associated negative caption $N(C)$, we intentionally derive a few negatives from $N_1(C)$ (instead of for C), resulting in $N_1(N_1(C))$ and $N_2(N_1(C))$, resulting in $[C, N_1(C), N_1(N_1(C)), N_2(N_1(C))]$ as options, so that $N_1(C)$ becomes the "centralized" description (see Fig. 5). Surprisingly, we find that 66.4% of text-only GPT-4o's predictions correspond to $N(C)$, while only 6.4% of its predictions correspond to C. Our findings also align with human behavior analysis from psychology [15], where humans can achieve better than random chance performance on multi-choice QAs using similar cues.

Motivated by this findings, we propose to decompose a single multi-choice QA into multiple binary QAs. In this case, we eliminate the "centralized option" due to the fact that there are only two options to choose from. As a result, given M negatives, the multiple binary QAs will query a model M times, where the random chance performance changes from $\frac{1}{M+1}$ to $(\frac{1}{2})^M$. Given that $(\frac{1}{2})^M > \frac{1}{M+1}$ for every $M > 2$, multiple binary QA is a more difficult task than multi-choice QA.

4 Experiments

4.1 Experiment Setup

We evaluate both (1) multimodal video text generation models, including GPT-4o [51], Gemini-1.5-Pro [17], Claude-3.5-Sonnet [2], Qwen2VL [64], LLaVA-OneVision [30], LLaVA-Next-Video [84], Phi-3.5-Vision [1], MiniCPM-2.6 [70], MA-LMM [21], VideoLLaVA [34], InternLM-Xcomposer-2.5 [82], Matryoshka Multimodal Models (M^3) [6], and (2) multimodal video embedding models, including XCLIP [47], ImageBind [18], and LanguageBind [85]. We exponentially increase the number of frames to study its effect on video understanding. More details can be found in Appendix F.

To study the effect of single frame bias and text bias, we also evaluate models trained on single images, including LLaVA-1.5 [38], LLaVA-NeXT [39], and Phi-3V [1]. In the latter case, we evaluate the LLMs including GPT-4o [51], Gemini-1.5-Pro [17], Yi-34B [71], Vicuna [10] and Flan-T5 [65] without using videos at all.

4.2 Human Performance

We use Amazon Mechanical Turk to evaluate human performance. Note that we exclude the positive caption annotators to ensure that there is no data contamination. Again, we use an onboarding test using a held out binary video QA evaluation set which has clear answers. Next, we show the performance on each task.

Table 1: Dataset characteristics including number of samples, average number of words in original captions and our fine-grained captions.

Dataset	Number of Samples	Org. Avg. # words	Ours Avg. # words
ActivityNet [28]	281	13.03	49.55
EgoExo4D [19]	307	7.73	47.79
Charades [16]	298	6.21	44.16
MPI Movie Description [56]	326	12.39	35.33
Oops [12]	294	10.06	43.27
COIN [61]	385	5.01	50.06
FineGym [57]	288	21.92	21.92
TemporalBench (ours)	2179	10.91	41.72

4.3 Fine-grained Video Question Answering on Short Videos

The results for multimodal generative models and embedding models are shown in Table 2 and Figure 7 (a). Note that we show the result with the best average multiple binary QA (MBA) performance for each model with respect to the number of frames. Results under different frames can be found in Appendix F. Several interesting findings arise:

The performance of any video model is far from human performance. As shown in the table, humans show an average performance of 67.9%, which is significantly higher than the best models, GPT-4o and Qwen2VL-72B, by ∼30%. Therefore, there is a large gap between model's performance and human performance. Note that we are employing standard AMT workers instead of domain experts, meaning that the expert-level accuracy can be even higher, especially for professional video understanding like FineGym.

Models show limited performance gains with more frames. As shown in Figure 6, with more frames, multimodal video models usually show better performance. However, performance generally saturates around 8-16 frames, meaning that models struggle to improve fine-grained activity understanding even with more frames. This is a clear contrast with human performance, showing that there is still a large space for multimodal video models to improve.

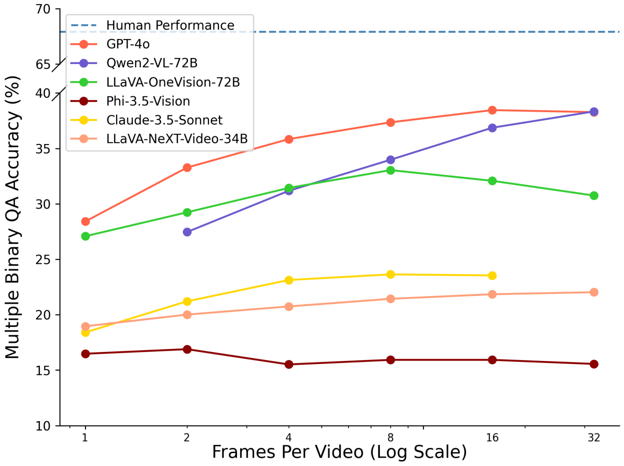


Figure 6: **Model performance on *TemporalBench*** with varying frames.

Multiple Binary QA is a more challenging metric. Multiple Binary QA, as proposed in Section 3.3, prevents a model from exploiting cues in the answer choices, and evaluates whether a model truly understands the temporal dynamics in the video by splitting a single $M+1$-way multiple choice question into M binary choice questions. For example, GPT-4o receives 75.7% accuracy but only 38.5% on multiple binary accuracy, showing a huge gap. These results indicate that understanding the fine-grained temporal dynamics is still a challenging task for current proprietary models and open-sourced models.

Video Embedding models show near chance performance. All multimodal video embedding models, including XCLIP, LanguageBind, and ImageBind show near random chance performance. One reason could be that their small embedding size (typically a vector with size around 768-2048) is insufficient to capture fine-grained temporal details.

Low single-frame bias and language bias. As shown in Figure 6 and Table 10, the performance of models like GPT-4o gradually increases with more frames. Excluding GPT-4o, all remaining VLMs trained with single images *e.g.,* LLaVA-1.5, Phi-3V, and text-only LLMs such as Yi-34B and Vicuna-7B show poor performance.

Table 2: *TemporalBench* performance of various multimodal generative models and embedding models under the binary QA accuracy (BA) and multiple binary QA settings (MBA) for short videos. The prefix "T-" indicates MBA performance for the annotated subset in our *TemporalBench*. We show the result with the best average MBA performance for each model with respect to the number of frames, denoted as # Frames.

Model	# Frames	T-ActivityNet	T-Charades	T-FineGym	T-Movie	T-Oops	T-COIN	T-EgoExo4D	BA	MBA
Human Performance	-	**68.7**	**82.2**	**36.1**	**74.2**	**69.7**	**70.6**	**71.0**	**89.7**	**67.9**
Random Chance	-	11.0	13.7	6.1	12.0	5.6	11.1	5.6	50.0	9.5
Video Embedding Models: Text + Multiple Frames as Input										
XCLIP	8	14.2	16.1	7.3	19.9	8.8	15.6	6.8	51.6	12.9
ImageBind	2	17.4	16.8	7.3	19.0	11.2	16.1	9.1	53.0	14.0
LanguageBind	8	22.4	15.1	6.6	19.3	10.9	15.6	11.1	52.8	14.5
Video Multimodal Generative Models : Text + Multiple Frames as Input										
GPT-4o	16	**48.8**	**42.6**	**18.8**	41.7	31.6	**46.5**	36.5	75.7	**38.5**
Gemini-1.5-Pro	1FPS	34.9	24.5	8.3	35.6	22.8	34.3	21.8	67.5	26.6
Claude-3.5-Sonnet	8	29.9	27.5	11.1	28.2	16.3	29.6	20.5	65.5	23.6
Qwen2-VL-72B	32	43.8	42.6	16.7	**45.1**	**36.7**	43.6	**37.1**	**75.8**	38.3
Qwen2-VL-7B	32	32.4	32.2	4.9	35.9	18.4	25.5	21.8	64.4	24.7
LLaVA-OneVision-72B	8	45.2	36.2	11.8	41.1	31.0	34.5	30.3	72.1	33.0
LLaVA-OneVision-7B	32	30.2	23.2	5.9	27.3	18.0	25.5	16.3	61.9	21.2
LLaVA-NeXT-Video-34B	32	30.6	26.8	10.4	24.8	18.0	25.2	17.3	64.0	22.0
LLaVA-NeXT-Video-7B	8	33.5	32.6	10.8	28.2	17.3	22.9	19.9	65.1	23.6
InternLM-XC2.5	1FPS	25.3	21.5	8.7	24.8	11.9	18.4	14.0	58.8	17.9
VideoLLaVA	8	35.2	29.2	13.5	25.5	20.7	32.5	20.2	67.1	25.5
MiniCPM-V2.6	1FPS	33.1	25.8	8.0	29.1	13.6	23.4	16.0	62.3	21.4
Phi-3.5-Vision	2	25.3	20.1	5.2	22.7	12.2	18.2	13.7	58.0	16.9
MA-LMM	4	12.5	16.4	3.5	11.0	5.1	11.4	4.9	48.0	9.4
M^3	6	21.0	20.1	6.6	19.6	10.2	15.1	10.4	56.4	14.8
Large Multimodal Models (LMMs): Text + 1 Frame as Input										
GPT-4o	1	32.0	30.2	15.3	31.3	26.5	33.8	27.7	70.0	28.4
LLaVA-1.5-13B	1	16.0	17.1	9.4	16.6	6.1	16.4	9.1	55.7	13.1
LLaVA-1.5-7B	1	25.3	25.8	8.7	19.3	9.2	21.8	16.6	60.5	18.3
LLaVA-NeXT-34B	1	20.6	22.5	9.4	21.5	15.3	21.6	13.7	60.5	18.0
Phi-3-Vision	1	23.1	19.8	4.5	17.8	8.5	17.7	13.7	54.4	15.1
Large Language Models (LLMs): Text as Input										
GPT-4o	0	30.2	31.9	16.7	27.9	22.8	27.5	28.0	67.7	26.5
Gemini-1.5-Pro	0	22.4	20.5	4.5	19.9	10.2	16.9	17.9	58.1	16.1
Yi-34B	0	17.4	27.5	10.4	21.8	11.2	23.4	16.9	59.9	18.7
Vicuna7b-1-5	0	11.4	17.4	6.6	11.3	5.1	12.2	7.8	50.5	10.4
Flan-T5-XL	0	24.9	23.5	5.6	19.9	11.9	23.4	14.0	57.9	17.9
Flan-T5-XXL	0	19.2	16.8	8.3	18.1	7.8	19.7	14.0	55.1	15.1

4.4 Video Captioning

Our detailed video captions also enables analyzing a model's fine-grained video captioning capabilities. For this, we prompt multimodal video models to generate a caption for an input video, with 3 captioning examples in the prompt as guidance to mimic the style of our detailed video captions. Note that we remove the FineGym captions due to its different structure compared to other video captions, resulting in 1891 samples. We evaluate the resulting video captioning performance using classical image captioning metrics, CIDEr [63], BLEU [52] at different n-gram levels, ROUGE [35], as well as the embedding similarity with sentence transformer [55] between the ground truth caption and the generated caption. Note that we for each model, we use the same number of frames as in Section 4.3.

Results in Table 4 show that GPT-4o achieves the best performance. Interestingly, the results indicate that the embedding similarity aligns most closely with the video QA task results from Sec 4.3. Other classical captioning metrics show inconsistent results. For example, GPT-4o obtains similar performance with one compared to 64 frames on both CIDEr and BLEU scores (e.g., for BLEU_1 24.1 vs. 25.1). On the other hand, all models show similar ROUGE scores. Thus, for the zero-shot captioning task, our findings indicate that text embedding similarity may be the most reliable metric.

5 Conclusion and Future Work

We propose TemporalBench, a novel video understanding benchmark, to evaluate the fine-grained temporal understanding abilities of multimodal video models. The video captions in our benchmark are significantly denser than existing datasets such as MSRVTT and TGIF, offering detailed temporal annotations. TemporalBench also provides a more challenging set of tasks that push current multimodal models beyond coarse-level understanding. The empirical results reveal a substantial gap between human performance and current state-of-the-art models. We also found a critical pitfall for multi- choice QA, where we devise multiple binary accuracy (MBA) to address thi issue. We hope that this benchmark fosters further research in developing models with enhanced temporal reasoning capabilities. Our benchmark could also be easily utilized for other fundamental video tasks such as spatio-temporal localization and text-to-video generation with fine-grained prompts.

Acknowledgement

This work was supported in part by NSF IIS2404180, and Institute of Information & communications Technology Planning & Evaluation(IITP) grants funded by the Korea government(MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration) and (No. RS2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training), and Microsoft Accelerate Foundation Models Research Program.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] Anthropic. Claude-sonnet-3.5. https://www.anthropic.com/news/claude-3-5-sonnet, 2024.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

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

[5] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[6] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*, 2024.

[7] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR, June 2011.

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

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

[10] 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. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[12] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. *CVPR*, 2020.

[13] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pp. 363–370. Springer, 2003.

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

[15] Moran Furman and Xiao-Jing Wang. Similarity effect and optimal control of multiple-choice decision making. *Neuron*, 60(6):1153–1168, 2008.

[16] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275, 2017.

[17] Gemini Team. Gemini: A family of highly capable multimodal models, 2024.

[18] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

[19] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

[20] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

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

[22] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Kevin Lin, William Yang Wang, Lijuan Wang, and Xin Eric Wang. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos, 2024. URL https://arxiv.org/abs/2406.08407.

[23] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *NeurIPS*, 2023.

[24] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

[25] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2758–2766, 2017.

[26] Gregory Kamradt. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023. Accessed: 2024-10-01.

[27] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *arXiv preprint arXiv:2403.18406*, 2024.

[28] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

[29] Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 487–507, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.29. URL https://aclanthology.org/2023.acl-long.29.

[30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[31] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[32] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024. URL https://arxiv.org/abs/2311.17005.

[33] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20.

[34] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[35] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024.

[39] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[40] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[41] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos?, 2024. URL https://arxiv.org/abs/2403.00476.

[42] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL https://arxiv.org/abs/2310.02255.

[43] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Adv. Neural Inform. Process. Syst.*, 2024.

[44] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

[45] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

[46] Meta. Llama-3. https://ai.meta.com/blog/meta-llama-3/, 2024.

[47] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision (ECCV)*, 2022.

[48] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023.

[49] OpenAI. Chatgpt. https://openai.com/blog/chatgpt/, 2023.

[50] OpenAI. Gpt-4 technical report. 2023.

[51] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

[52] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

[53] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[55] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

[56] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[57] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[58] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

[59] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[60] Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13581–13591, 2024.

[61] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1207–1216, 2019.

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[63] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

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

[65] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.

[66] Wenhao Wu. Freeva: Offline mllm as training-free video assistant. *arXiv preprint arXiv:2405.07798*, 2024.

[67] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

[68] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

[69] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

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

[71] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai. *arXiv*, 2024.

[72] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[73] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

[74] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning*, 2024.

[75] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, volume 33, pp. 9127–9134, 2019.

[76] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering, 2019. URL https://arxiv.org/abs/1906.02467.

[77] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

[78] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

[79] Da Zhang, Xiyang Dai, and Yuan-Fang Wang. Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pp. 712–728. Springer, 2019.

[80] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[81] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, and Jianwei Yang. Llava-grounding: Grounded visual chat with large multimodal models, 2023.

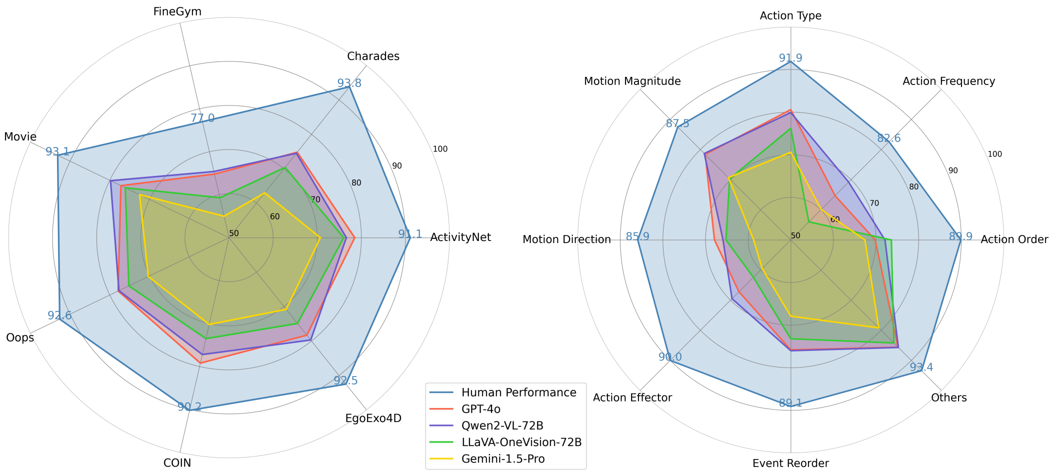
[82] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.

[83] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

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

[85] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024.

[86] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ICLR*, 2024.



(a) Binary Accuracy (BA) per subset

(b) Binary Accuracy (BA) per negative category

Figure 7: Visualization of binary accuracy for short video QA per (a) subset and (b) negative type. Human performance is much better than GPT-4o, Qwen2-VL-72B, LLaVA-OneVision-72B, and Gemini-1.5-Pro.

Reproducibility Statement

We attach part of the dataset in the submission's supplementary materials. We will also publicly release it along with the code used to evaluate the LMMs upon the paper's acceptance.

Limitations. One cannot fully analyze the behavior of proprietary models included in this paper due to the lack of access to these models, which are GPT-4o, Gemini-1.5-Pro and Claude 3.5 Sonnet.

A Long Video Understanding

Since our benchmark is annotated at the video clip level, we can easily extend it to long video understanding by concatenating the captions of different video clips within the same original video. In our study, we choose video datasets from AcitivityNet, Charades, EgoExo4D, COIN and FineGym. We randomly sample video clips within the same original video, and then crop a new video segment whose starting time corresponds to that of the earliest sampled video clip and whose ending time corresponds to that of the latest sampled video clip. We then concatenate all the sampled video captions together to form a single long detailed description corresponding to the new video segment. Given this positive caption, we generate negative captions for it by replacing the positive caption of one of the sampled video clips with its negatives. The model is then tasked to choose the correct long caption out of multiple choices. We control the random chance multiple binary QA performance to be $\sim$9.5%, resulting in an apple-to-apple comparsion with in Sec 4.3. In this way, we investigate whether multimodal video models can understand and distinguish fine details in a long video. Finally, we sampled 1,574 videos with durations ranging between $[0, 20]$ minutes, as shown in Figure 4.

We show in Table 5, that all multimodal video models show a significant performance drop for this task compared to short video understanding. This is also reflected in all models performing better on relatively shorter videos (*e.g.,* Charades) compared to longer videos (*e.g.,* FineGym). These results indicate that finding the subtle temporal dynamic differences in a long video is indeed an extremely difficult task. It is similar in nature to the needle-in-the-sea task [26] in NLP except in the temporal domain. We hope that *TemporalBench* for long video understanding can serve as a very challenging task for future video understanding model development.

Table 3: Effect of the "Centralized" Caption on text-only GPT-4o.

Percentage of Predictions Aligned with ⟶	C	$N_1(C)$
$[C, N_1(C), N_2(C)), N_3(C))]$	**83.3**	6.4
$[C, N_1(C), N_1(N_1(C)), N_2(N_1(C))]$	17.7	**66.4**

B In-Depth Analysis

B.1 Why multiple Binary QA instead of multi-choice QA?

As discussed in Section 3.3, in the standard multi-choice QA setting, if negatives are all slightly variations of the positive caption, we find that LLMs can determine the "centralized" caption, and take a shortcut to achieve better performance. To demonstrate this, based on one negative caption $N(C)$ in *TemporalBench*, we intentionally generate two negative captions derived from $N(C)$ (instead of C), resulting in $N_1(N(C))$ and $N_2(N(C))$. Given two set of options $[C, N_1(C), N_2(C)), N_3(C))]$ and $[C, N_1(C), N_1(N_1(C)), N_2(N_1(C))]$ shown in Figure 5, text-only GPT-4o displays different behaviors. As shown in Table 3, under the intentionally designed negative options, GPT-4o will choose $N_1(C)$ under 66.4% cases. This again demonstrates the necessity and advantage of our multiple binary QA accuracy (MBA) metric design over the standard multi-choice QA setting.

B.2 Performance on categories

Broadly, *TemporalBench* evaluates word level replacement and event level re-ordering. Here we further breakdown the word level replacement into following categories: (1). Action order (change the order); (2). Action frequency (1 times *v.s.* two times); (3). Action type (put *v.s.* pull); (4). Motion magnitude (slightly *v.s.* intensively); (5). Motion Direction/Orientation (forward *v.s.* backward, circular *v.s.* back-and-forth). (6). Action effector (cutting with left hand *v.s.* cutting with right hand) (7). Others. We prompt GPT-4o to perform 7-way classification and show the per-category performance in Table 7 and Figure 7 (b). Results indicate that multimodal video models shows better performance on "others" category rather than the other categories related to actions. Among the seven categories, models struggle most on action frequency (counting), which show that they do not memorize repeated occurrences well. The visualizations of failture cases in GPT-4o is shown in Figure 8.

Ethics Statement

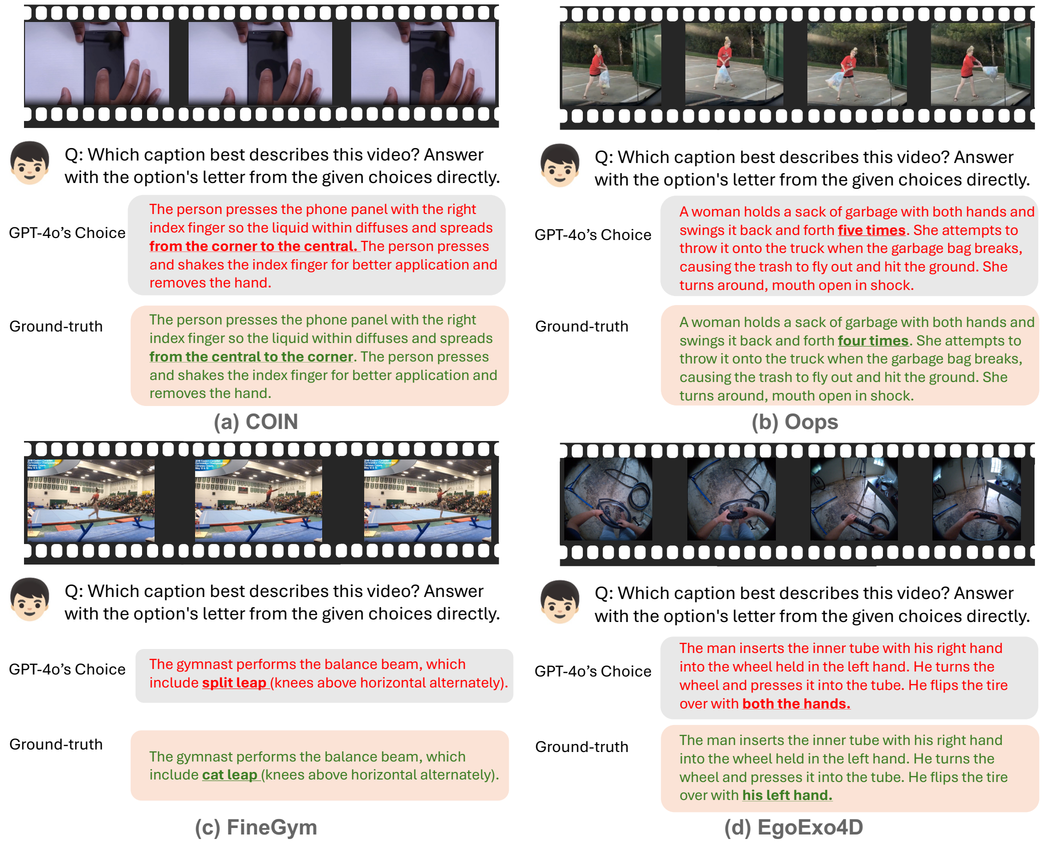
This research primarily utilizes publicly available video datasets, which have been collected and annotated by qualified annotators and authors, ensuring compliance with ethical standards. We have made every effort to ensure that the data used respects privacy and contains no personally identifiable information. Furthermore, we acknowledge the potential implications of fine-grained video understanding, especially in sensitive applications such as surveillance and autonomous systems. As such, we advocate for responsible and ethical use of this research, urging caution in deploying these models in real-world scenarios to avoid harmful or unintended consequences.

C Broader Impact

TemporalBench, a comprehensive benchmark for video understanding, has the potential to significantly advance research in this field by offering improved metrics for model evaluation. Our work aims to enhance the temporal reasoning capabilities of future video understanding models. However, the broader impact of more advanced video understanding technologies raises important societal concerns, including the risk of mass surveillance, privacy violations, and the development of harmful applications like autonomous weapons. Therefore, we strongly encourage thoughtful consideration when deploying these models in real-world scenarios to mitigate negative or unintended consequences.

D More Visualizations of Our Benchmark

In this section, we present comprehensive visualizations of our fine-grained annotations with both positive and negative descriptions. For each benchmark mentioned in Table 1, we provide one video



Q: Which caption best describes this video? Answer with the option's letter from the given choices directly.

GPT-4o's Choice: The person presses the phone panel with the right index finger so the liquid within diffuses and spreads **from the corner to the central.** The person presses and shakes the index finger for better application and removes the hand.

Ground-truth: The person presses the phone panel with the right index finger so the liquid within diffuses and spreads **from the central to the corner**. The person presses and shakes the index finger for better application and removes the hand.

(a) COIN

Q: Which caption best describes this video? Answer with the option's letter from the given choices directly.

GPT-4o's Choice: A woman holds a sack of garbage with both hands and swings it back and forth **five times**. She attempts to throw it onto the truck when the garbage bag breaks, causing the trash to fly out and hit the ground. She turns around, mouth open in shock.

Ground-truth: A woman holds a sack of garbage with both hands and swings it back and forth **four times**. She attempts to throw it onto the truck when the garbage bag breaks, causing the trash to fly out and hit the ground. She turns around, mouth open in shock.

(b) Oops

Q: Which caption best describes this video? Answer with the option's letter from the given choices directly.

GPT-4o's Choice: The gymnast performs the balance beam, which include **split leap** (knees above horizontal alternately).

Ground-truth: The gymnast performs the balance beam, which include **cat leap** (knees above horizontal alternately).

(c) FineGym

Q: Which caption best describes this video? Answer with the option's letter from the given choices directly.

GPT-4o's Choice: The man inserts the inner tube with his right hand into the wheel held in the left hand. He turns the wheel and presses it into the tube. He flips the tire over with **both the hands.**

Ground-truth: The man inserts the inner tube with his right hand into the wheel held in the left hand. He turns the wheel and presses it into the tube. He flips the tire over with **his left hand.**

(d) EgoExo4D

Figure 8: **The failure cases of GPT-4o in *TemporalBench*.** GPT-4o does not understand the fine-grained details well, including motion direction, action frequency, action type, and motion direction.

example with its positive annotation and one of the corresponding negative descriptions (there are more than one negative for a single video in our dataset) in Figures 9 & 10. The video examples (*a* - *f*) are displayed in the same order as their sources in Table 1 (7 in total).

E Per subset Results for Short and Long Video QA under Binary Accuracy (BA)

The per subset results (denoted as "T-") for short and long video QA under Binary Accuracy (BA) are shown in Table 8, and Table 9, respectively. Still, human achieve much better performance than all multimodal videos. Interestingly, both human and Finegym, the professional subset,

F More Results with Extended Frames

In the main paper, we only report the performance of each multimodal video models with the the number of frams that leads to the best performance. Here we extend the results to show the results of more frames in Table 10.

G Data Annotation Platform

Positive Captions We use Amazon Mechanical Turk (AMT) [3] for positive caption annotation, and then use Label Studio [4] to let authors refine the caption. As shown in Figure 11, authors can edit the

[3] https://www.mturk.com/

[4] https://labelstud.io/

Table 4: Comparison of models for video captioning using Caption Similarity, CIDEr, BLEU, and ROUGE metrics. Cosine similarity using sentence transformer reflects the captioning quality the best.

Model	Similarity	CIDEr	ROUGE	BLEU_1	BLEU_2	BLEU_3	BLEU_4
Video Multimodal Generative Models : Text + Multiple Frames as Input							
GPT-4o	**61.3**	7.3	**19.6**	24.1	**11.8**	**5.8**	**3.0**
Gemini-1.5-Pro	56.5	**10.9**	19.1	19.0	9.2	4.5	2.4
Claude-3.5-Sonnet	54.1	8.6	17.1	24.4	10.3	4.4	2.1
Qwen2-VL-72B	56.1	9.3	19.1	15.7	8.0	4.1	2.2
Qwen2-VL-7B	51.9	6.9	18.0	12.5	6.1	3.0	1.6
LLaVA-OneVision-72B	55.0	9.7	18.7	23.7	11.3	5.6	2.9
LLaVA-OneVision-7B	50.1	0.3	14.5	11.1	5.1	2.2	1.1
LLaVA-NeXT-Video-34B	53.1	5.3	15.9	21.4	9.2	4.0	1.8
LLaVA-NeXT-Video-7B	50.1	2.3	15.8	18.1	7.0	2.6	1.1
InternLM-XC2.5	52.4	2.3	15.9	17.8	7.1	2.8	1.2
VideoLLaVA	46.0	4.5	16.9	12.6	5.4	2.3	1.0
MiniCPM-V2.6	47.2	1.5	14.2	15.5	5.4	1.9	0.8
Phi-3.5-Vision	42.9	3.7	16.5	20.4	8.4	3.4	1.6
MA-LMM	38.7	3.1	15.0	10.1	4.8	2.2	1.1
M^3	47.8	3.0	16.4	16.7	6.9	2.8	1.2
Large Multimodal Models (LMMs): Text + 1 Frame as Input							
GPT-4o	52.3	7.3	17.1	**25.1**	11.1	5.0	2.4
LLaVA-1.5-13B	47.9	4.9	18.0	22.6	9.8	4.2	2.0
LLaVA-1.5-7B	45.7	6.9	17.8	22.0	9.5	4.2	2.0
LLaVA-NeXT-34B	49.1	6.2	16.7	24.2	10.4	4.6	2.2
Phi-3-Vision	42.0	4.0	16.1	19.9	8.3	3.4	1.6

caption from AMT workers. Also, we provide the original short video captions to let people better understand our task.

Negative Captions We first prompt LLMs (GPT-4o, Gemini, and Llama-3.1-405b) to get initial negative captions, and then ask authors to choose the negatives that can reflect the temporal dynamic. The visualization of the multi-choice platform in shown in Figure 12.

Table 5: *TemporalBench* performance of various multimodal generative models and embedding models under **long video** understanding with binary QA accuracy (BA) and multiple binary QA accuracy (MBA). The MBA performance under each dataset is also included. We show the result with the best average MBA performance for each model with respect to the number of frames, denoted as # Frames.

Model	# Frames	T-ActivityNet	T-Charades	T-FineGym	T-COIN	T-EgoExo4D	BA	MBA
Random Performance	-	9.3	9.8	10.1	11.4	9.3	50.0	9.5
Video Embedding Models: Text + Multi-Frames as Input								
XCLIP	8	11.1	12.4	6.5	10.8	11.8	51.7	11.1
ImageBind	2	10.2	8.1	9.3	10.8	12.4	51.0	10.7
LanguageBind	8	11.7	10.8	10.3	11.0	14.1	51.6	12.0
Video Multimodal Generative Models : Text + Multi-Frames as Input								
GPT-4o	64	**40.0**	**37.8**	16.8	**32.7**	29.3	**70.5**	**32.7**
Gemini-1.5-Pro	1FPS	32.1	18.4	18.7	24.8	23.8	65.2	24.7
Claude-3.5-Sonnet	8	28.9	22.2	16.8	22.2	26.7	64.6	24.5
Qwen2-VL-72B	8	32.4	20.5	21.5	18.9	33.1	64.7	26.2
Qwen2-VL-7B	32	22.2	20.0	9.3	18.3	18.7	59.7	18.8
LLaVA-OneVision-72B	4	28.6	19.5	18.7	16.5	30.9	63.4	23.8
LLaVA-OneVision-7B	32	21.3	13.0	13.1	11.4	19.8	56.9	16.2
LLaVA-NeXT-Video-34B	4	23.5	22.2	19.6	17.9	19.2	60.3	20.0
LLaVA-NeXT-Video-7B	8	18.1	21.6	10.3	18.5	15.6	57.2	17.3
InternLM-XC2.5	1FPS	21.0	18.4	20.6	14.0	11.4	55.8	15.6
VideoLLaVA	8	20.0	16.8	15.9	9.8	16.6	56.0	15.1
MiniCPM-V2.6	1FPS	14.3	16.8	6.5	17.1	14.1	60.3	19.3
Phi-3.5-Vision	4	23.2	11.9	19.6	10.2	13.3	54.5	14.5
MA-LMM	4	10.2	9.2	2.8	11.4	11.6	47.1	9.2
M^3	6	10.8	8.6	12.1	13.0	12.4	53.1	11.8
Large Multimodal Models (LMMs): Text + 1 frame as Input								
GPT-4o	1	27.9	23.2	19.6	25.2	22.9	64.7	24.5
LLaVA-1.5-13B	1	14.3	11.9	10.3	15.4	14.7	54.8	14.2
LLaVA-1.5-7B	1	9.2	11.9	10.3	12.8	14.5	53.2	12.3
LLaVA-NeXT-34B	1	21.6	20.5	19.6	18.9	19.8	60.5	19.9
Phi-3-Vision	1	18.1	12.4	15.0	15.4	15.6	56.0	15.6
Large Language Models (LLMs): Text as Input								
GPT-4o	0	27.6	32.4	17.8	24.2	**33.5**	67.6	28.2
Gemini-1.5-Pro	0	22.9	19.5	17.8	19.3	23.4	62.2	21.2
Yi-34B	0	19.7	19.5	14.0	15.9	20.6	59.5	18.4
Vicuna7b-1-5	0	6.3	9.2	9.3	10.6	12.0	51.1	9.9
Flan-T5-XL	0	21.6	15.7	**23.4**	18.1	19.8	60.1	19.4
Flan-T5-XXL	0	20.0	11.9	18.7	15.7	17.1	56.9	16.7

Table 6: *TemporalBench* statistics on negative caption types.

Action	Action	Action	Motion	Motion	Action	Event		Overall
Order	Frequency	Type	Magnitude	Direction	Effector	Reorder	Others	
129	530	2,802	320	1,536	1,109	2,099	1,342	9,867

Table 7: *TemporalBench* performance under each category under BA. Multimodal videos models struggle on certain tasks such as action frequency. We show the result with the best average MBA performance for each model with respect to the number of frames.

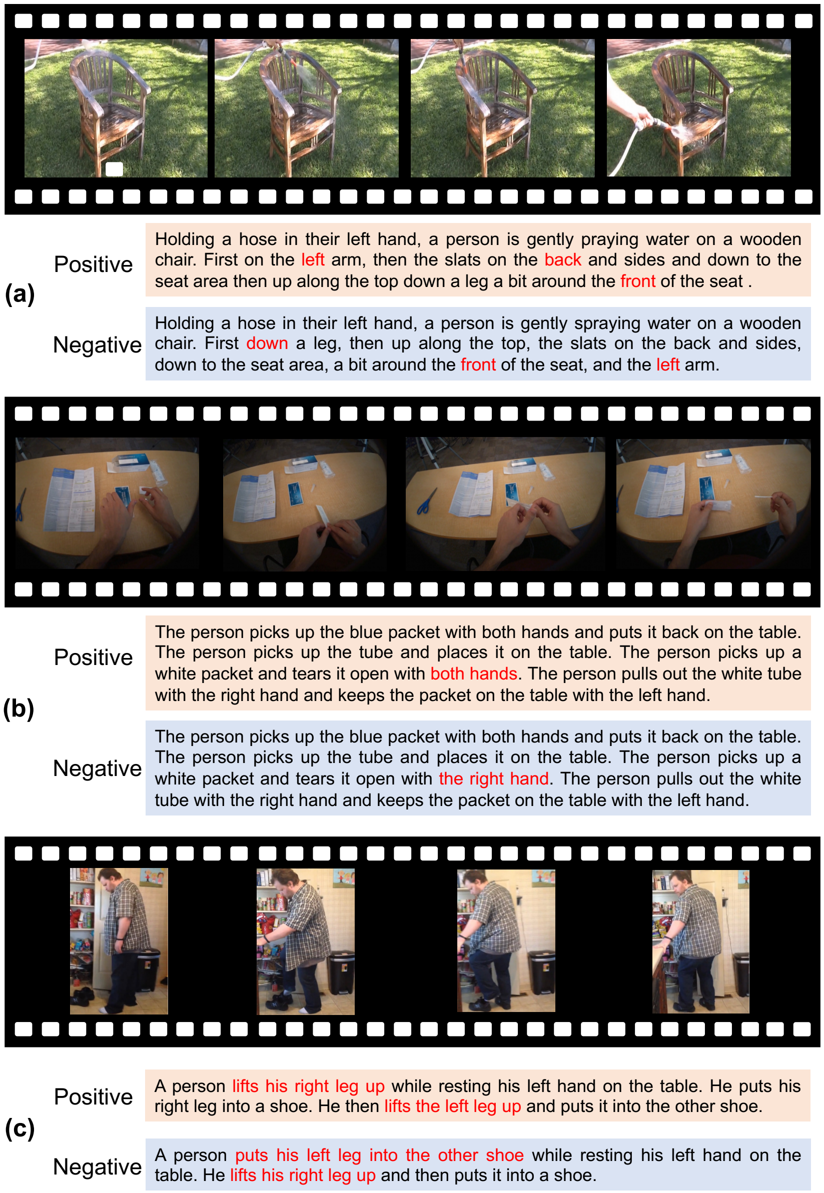
Model	The Number of Frames	Action Order	Action frequency	Action Type	Motion Magnitude	Motion Direction	Action Effector	Event Reorder	Others	Average
Human Performance	-	**89.9**	**82.6**	**91.9**	**87.5**	**85.9**	**90.0**	**89.1**	**93.4**	**89.7**
Random Chance	-	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Video Embedding Models: Text + Multi-Frames as Input										
XCLIP	8	46.5	50.8	50.9	56.9	51.2	51.7	50.1	55.6	51.6
ImageBind	2	44.2	44.7	55.4	50.9	52.5	50.5	48.6	61.8	53.0
LanguageBind	8	43.4	41.5	53.4	55.0	51.4	46.6	51.0	65.9	52.8
Video Multimodal Generative Models : Text + Multi-Frames as Input										
GPT-4o	16	69.8	64.7	**80.6**	78.4	**67.9**	67.2	75.8	85.6	75.7
Gemini-1.5-Pro	1FPS	67.4	60.1	70.6	70.7	58.7	59.5	67.9	79.2	67.5
Claude-3.5-Sonnet	8	62.0	57.4	70.7	70.3	60.0	57.8	61.3	76.2	65.5
Qwen2-VL-72B	32	72.1	69.2	79.9	**78.7**	65.9	69.5	**76.0**	**85.7**	**75.8**
Qwen2-VL-7B	32	65.9	45.8	67.3	66.1	54.6	54.7	69.7	75.7	64.4
LLaVA-OneVision-72B	8	**73.6**	56.0	76.2	70.3	65.2	62.4	73.2	84.2	72.1
LLaVA-OneVision-7B	32	63.6	45.5	62.9	56.9	52.8	54.0	66.5	77.1	61.9
LLaVA-NeXT-Video-34B	32	61.2	56.0	66.4	61.6	58.5	59.3	63.4	74.1	64.0
LLaVA-NeXT-Video-7B	8	69.0	65.7	68.2	62.2	66.5	68.6	52.2	74.3	65.1
InternLM-XC2.5	1FPS	55.8	42.5	62.7	62.5	52.6	51.1	58.3	70.7	58.8
VideoLLaVA	8	69.8	**70.2**	71.4	70.0	70.6	**70.2**	50.5	75.5	67.1
MiniCPM-V2.6	1FPS	59.4	52.3	65.5	62.5	54.1	53.3	63.5	74.7	62.3
Phi-3.5-Vision	2	53.5	55.3	60.1	55.9	54.0	52.2	55.3	69.4	58.0
MA-LMM	4	54.3	43.0	48.0	47.8	46.3	48.8	48.6	49.6	48.0
M^3	6	51.9	53.6	58.9	56.3	52.2	53.7	50.8	68.6	56.4
Large Multimodal Models (LMMs): Text + 1 frame as Input										
GPT-4o	1	67.4	65.1	74.1	70.3	64.2	62.6	68.7	78.4	70.0
LLaVA-1.5-13B	1	57.4	51.9	57.6	53.8	50.4	53.9	54.2	63.1	55.7
LLaVA-1.5-7B	1	62.0	61.5	62.2	54.1	61.4	64.9	51.0	67.9	60.5
LLaVA-NeXT-34B	1	51.2	55.7	61.2	60.0	54.8	53.0	65.0	67.5	60.5
Phi-3-Vision	1	46.5	45.5	56.0	55.6	48.8	49.2	56.9	62.1	54.4
Large Larguage Models (LLMs): Text as Input										
GPT-4o	0	65.1	59.8	73.7	70.0	61.5	60.1	69.3	68.6	67.7
Gemini-1.5-Pro	0	54.3	42.5	60.4	62.2	53.6	53.3	64.8	57.4	58.1
Yi-34B	0	51.9	62.3	60.1	60.3	57.1	55.1	65.4	58.0	59.9
Vicuna7b-1-5	0	55.8	47.2	51.7	48.4	50.1	49.4	49.9	51.4	50.5
Flan-T5-XL	0	53.5	57.7	60.2	59.7	56.1	56.9	54.9	60.7	57.9
Flan-T5-XXL	0	55.8	62.5	59.0	58.4	54.2	48.2	49.3	58.9	55.1

Table 8: *TemporalBench* performance of various multimodal generative models and embedding models under the binary QA accuracy (BA) and multiple binary QA settings (MBA) for short videos. The prefix "T-" indicates **BA** performance for the annotated subset in our *TemporalBench*. We show the result with the best average MBA performance for each model with respect to the number of frames, denoted as # Frames.

Model	# Frames	T-ActivityNet	T-Charades	T-FineGym	T-Movie	T-Oops	T-COIN	T-EgoExo4D	BA	MBA
Human Performance	-	**91.1**	**93.8**	**77.0**	**93.1**	**92.6**	**90.2**	**92.5**	**89.7**	**67.9**
Random Chance	-	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Video Embedding Models: Text + Multiple Frames as Input										
XCLIP	8	52.7	52.8	49.0	53.9	53.5	52.3	48.1	51.6	12.9
ImageBind	2	52.9	52.6	47.5	55.4	56.8	52.4	53.4	53.0	14.0
LanguageBind	8	56.5	50.1	48.2	55.8	55.1	51.1	52.8	52.8	14.5
Video Multimodal Generative Models : Text + Multiple Frames as Input										
GPT-4o	16	78.5	74.8	64.8	77.2	77.9	79.2	78.3	75.7	38.5
Gemini-1.5-Pro	1FPS	70.7	63.0	55.0	72.5	70.3	70.2	70.8	67.5	26.6
Claude-3.5-Sonnet	8	68.5	62.4	62.7	68.2	64.2	65.4	66.8	65.5	23.6
Qwen2-VL-72B	32	76.6	74.5	65.4	79.8	77.7	77.2	79.7	75.8	38.3
Qwen2-VL-7B	32	67.0	65.2	49.9	70.5	66.5	66.5	66.6	64.4	24.7
LLaVA-OneVision-72B	8	76.0	70.4	59.3	76.1	75.2	73.5	74.9	72.1	33.0
LLaVA-OneVision-7B	32	66.5	60.0	49.4	68.0	61.6	64.6	64.4	61.9	21.2
LLaVA-NeXT-Video-34B	32	67.5	62.9	56.3	68.0	66.1	63.4	64.5	64.0	22.0
LLaVA-NeXT-Video-7B	8	68.0	66.5	56.7	69.9	66.1	65.2	65.0	65.1	23.6
InternLM-XC2.5	1FPS	61.0	57.9	50.6	63.5	60.3	59.2	59.7	58.8	17.9
VideoLLaVA	8	71.8	63.4	61.6	68.2	68.5	68.9	67.3	67.1	25.5
MiniCPM-V2.6	1FPS	66.1	59.6	54.1	68.0	63.1	62.7	62.7	62.3	21.4
Phi-3.5-Vision	2	62.0	55.8	50.0	64.1	58.2	57.7	58.9	58.0	16.9
MA-LMM	4	49.8	48.8	42.3	48.0	49.9	49.0	48.8	48.0	9.4
M^3	6	59.5	54.9	51.1	60.9	58.9	54.9	55.2	56.4	14.8
Large Multimodal Models (LMMs): Text + 1 Frame as Input										
GPT-4o	1	69.1	67.1	64.8	71.0	71.9	71.0	74.0	70.0	28.4
LLaVA-1.5-13B	1	57.6	54.3	51.9	56.8	53.2	58.1	57.8	55.7	13.1
LLaVA-1.5-7B	1	64.2	58.6	55.7	61.0	57.5	62.7	63.9	60.5	18.3
LLaVA-NeXT-34B	1	59.7	60.3	55.0	61.8	62.0	61.0	63.7	60.5	18.0
Phi-3-Vision	1	57.4	54.5	45.2	57.5	52.8	55.8	58.9	54.4	15.1
Large Language Models (LLMs): Text as Input										
GPT-4o	0	66.2	67.4	65.6	65.6	68.9	67.8	71.7	67.7	26.5
Gemini-1.5-Pro	0	58.5	57.6	50.6	59.8	57.6	58.6	64.3	58.1	16.1
Yi-34B	0	59.1	62.3	54.9	59.7	57.7	63.1	63.6	59.9	18.7
Vicuna7b-1-5	0	49.7	49.5	50.2	50.7	50.5	50.0	52.1	50.5	10.4
Flan-T5-XL	0	60.5	59.2	50.5	60.7	56.8	58.7	60.3	57.9	17.9
Flan-T5-XXL	0	56.7	49.3	52.0	59.0	54.6	56.1	56.2	55.1	15.1

Table 9: *TemporalBench* performance of various multimodal generative models and embedding models under **long video** understanding with binary QA accuracy (BA) and multiple binary QA accuracy (MBA). The **BA** performance under each dataset is also included. We show the result with the best average MBA performance for each model with respect to the number of frames, denoted as # Frames.

Model	# Frames	T-ActivityNet	T-Charades	T-FineGym	T-COIN	T-EgoExo4D	BA	MBA
Random Performance	-	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Video Embedding Models: Text + Multi-Frames as Input								
XCLIP	8	51.9	48.7	47.9	52.6	52.8	51.7	11.1
ImageBind	2	50.3	52.6	47.9	51.3	51.3	51.0	10.7
LanguageBind	8	51.9	46.4	48.2	52.0	53.7	51.6	12.0
Video Multimodal Generative Models : Text + Multi-Frames as Input								
GPT-4o	64	74.8	73.8	61.2	70.1	68.7	70.5	32.7
Gemini-1.5-Pro	1FPS	67.0	61.6	60.6	65.9	65.9	65.2	24.7
Claude-3.5-Sonnet	8	66.8	63.7	56.7	63.1	66.6	64.6	24.5
Qwen2-VL-72B	8	68.5	59.6	62.5	59.6	70.0	64.7	26.2
Qwen2-VL-7B	32	60.7	58.0	49.9	59.8	61.9	59.7	18.8
LLaVA-OneVision-72B	4	67.0	63.5	61.2	55.8	69.3	63.4	23.8
LLaVA-OneVision-7B	32	60.0	53.6	57.6	53.2	59.8	56.9	16.2
LLaVA-NeXT-Video-34B	4	59.4	63.0	57.6	59.5	61.4	60.3	20.0
LLaVA-NeXT-Video-7B	8	60.9	58.6	51.5	56.7	56.1	57.2	17.3
InternLM-XC2.5	1FPS	59.6	58.9	57.0	54.9	52.8	55.8	15.6
VideoLLaVA	8	61.2	57.0	59.5	50.1	57.3	56.0	15.1
MiniCPM-V2.6	1FPS	53.7	58.6	41.3	54.8	53.9	60.3	19.3
Phi-3.5-Vision	4	60.3	52.3	58.1	50.3	55.1	54.5	14.5
MA-LMM	4	47.4	51.7	36.4	50.1	51.2	47.1	9.2
M^3	6	52.5	52.9	51.0	53.4	53.6	53.1	11.8
Large Multimodal Models (LMMs): Text + 1 frame as Input								
GPT-4o	1	67.6	64.3	62.8	65.9	62.0	64.7	24.5
LLaVA-1.5-13B	1	55.1	52.3	52.9	55.0	54.8	54.5	14.2
LLaVA-1.5-7B	1	51.2	53.4	51.5	51.8	56.2	53.2	12.3
LLaVA-NeXT-34B	1	60.6	60.8	57.0	59.8	61.8	60.5	19.9
Phi-3-Vision	1	56.9	53.9	52.1	55.6	57.6	56.0	15.6
Large Larguage Models (LLMs): Text as Input								
GPT-4o	0	67.1	68.1	63.6	65.1	71.3	67.6	28.2
Gemini-1.5-Pro	0	62.8	59.4	55.6	60.7	65.7	62.2	21.2
Yi-34B	0	59.0	60.2	56.5	59.5	60.4	59.5	18.4
Vicuna7b-1-5	0	49.0	52.4	49.3	51.2	52.2	51.1	9.9
Flan-T5-XL	0	61.3	57.7	59.8	58.8	61.7	60.1	19.4
Flan-T5-XXL	0	59.4	53.6	59.5	56.3	56.5	56.9	16.7



(a)

Positive

Holding a hose in their left hand, a person is gently praying water on a wooden chair. First on the left arm, then the slats on the back and sides and down to the seat area then up along the top down a leg a bit around the front of the seat .

Negative

Holding a hose in their left hand, a person is gently spraying water on a wooden chair. First down a leg, then up along the top, the slats on the back and sides, down to the seat area, a bit around the front of the seat, and the left arm.

(b)

Positive

The person picks up the blue packet with both hands and puts it back on the table. The person picks up the tube and places it on the table. The person picks up a white packet and tears it open with both hands. The person pulls out the white tube with the right hand and keeps the packet on the table with the left hand.

Negative

The person picks up the blue packet with both hands and puts it back on the table. The person picks up the tube and places it on the table. The person picks up a white packet and tears it open with the right hand. The person pulls out the white tube with the right hand and keeps the packet on the table with the left hand.

(c)

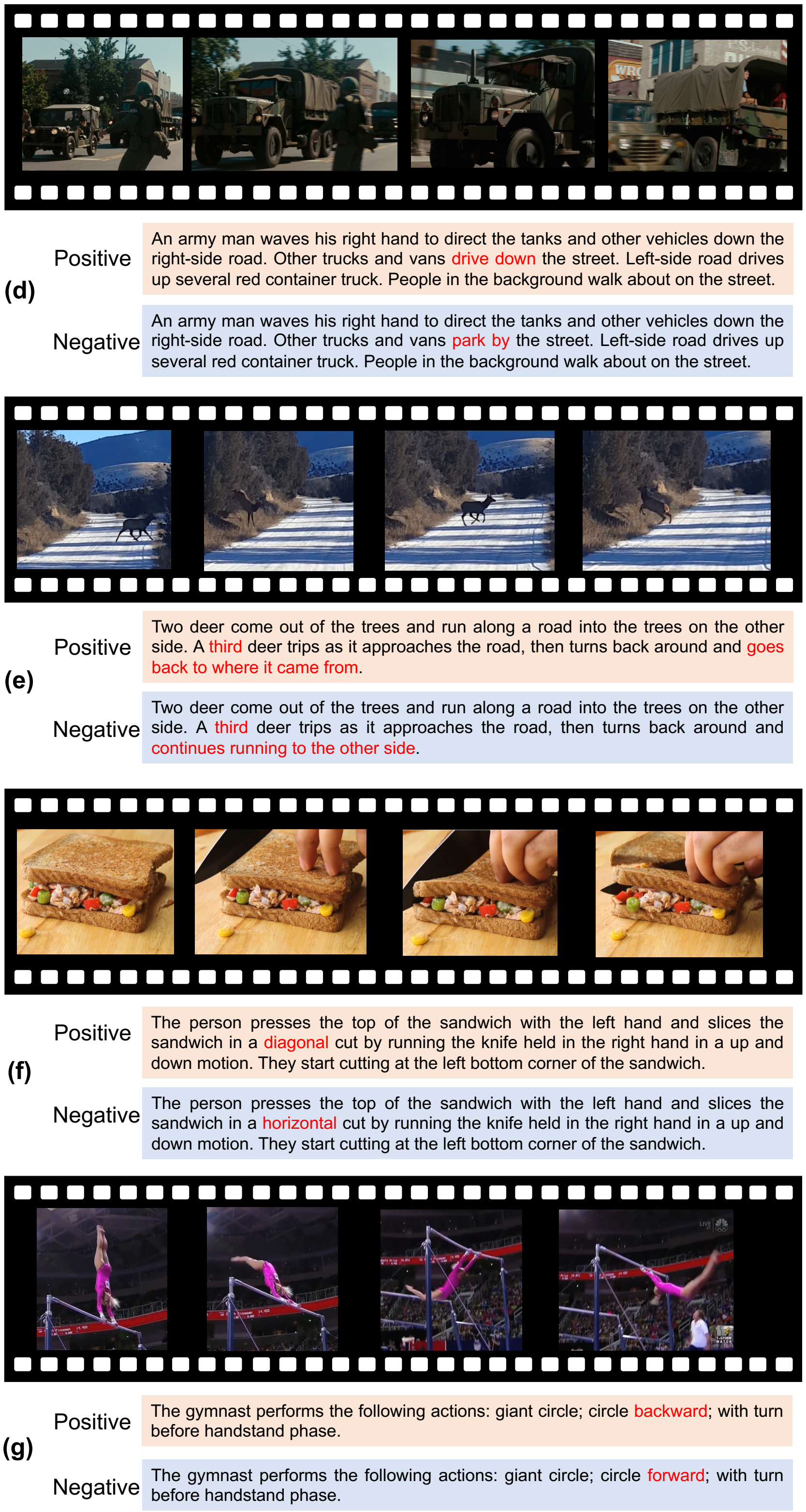
Positive

A person lifts his right leg up while resting his left hand on the table. He puts his right leg into a shoe. He then lifts the left leg up and puts it into the other shoe.

Negative

A person puts his left leg into the other shoe while resting his left hand on the table. He lifts his right leg up and then puts it into a shoe.

Figure 9: Visualizations (I) of our fine-grained annotations of the videos with both positive and negative descriptions.



(d)

Positive: An army man waves his right hand to direct the tanks and other vehicles down the right-side road. Other trucks and vans drive down the street. Left-side road drives up several red container truck. People in the background walk about on the street.

Negative: An army man waves his right hand to direct the tanks and other vehicles down the right-side road. Other trucks and vans park by the street. Left-side road drives up several red container truck. People in the background walk about on the street.

(e)

Positive: Two deer come out of the trees and run along a road into the trees on the other side. A third deer trips as it approaches the road, then turns back around and goes back to where it came from.

Negative: Two deer come out of the trees and run along a road into the trees on the other side. A third deer trips as it approaches the road, then turns back around and continues running to the other side.

(f)

Positive: The person presses the top of the sandwich with the left hand and slices the sandwich in a diagonal cut by running the knife held in the right hand in a up and down motion. They start cutting at the left bottom corner of the sandwich.

Negative: The person presses the top of the sandwich with the left hand and slices the sandwich in a horizontal cut by running the knife held in the right hand in a up and down motion. They start cutting at the left bottom corner of the sandwich.

(g)

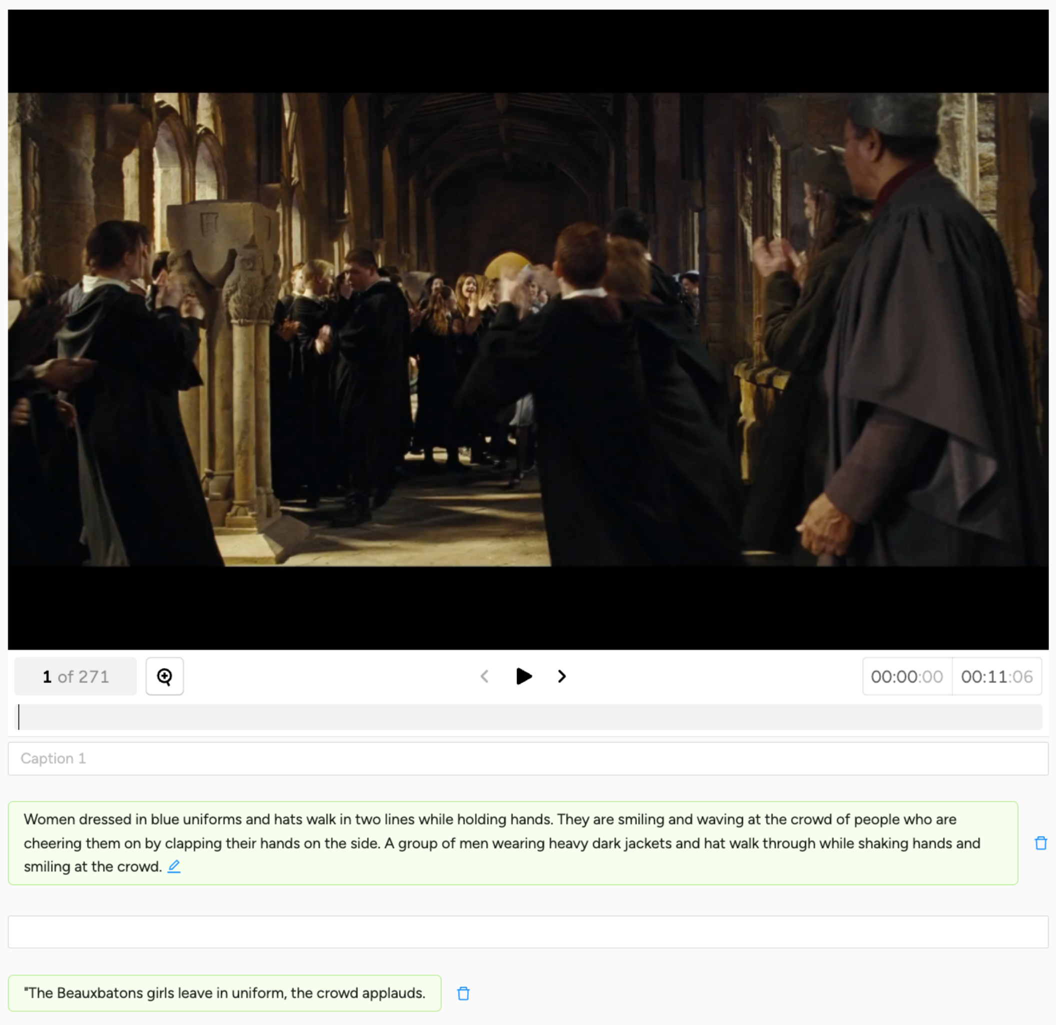
Positive: The gymnast performs the following actions: giant circle; circle backward; with turn before handstand phase.

Negative: The gymnast performs the following actions: giant circle; circle forward; with turn before handstand phase.

Figure 10: Visualizations (II) of our fine-grained annotations of the videos with both positive and negative descriptions.

Table 10: *TemporalBench* performance of various models under binary QA and multiple binary QA setting for short question answering with different number of frames.

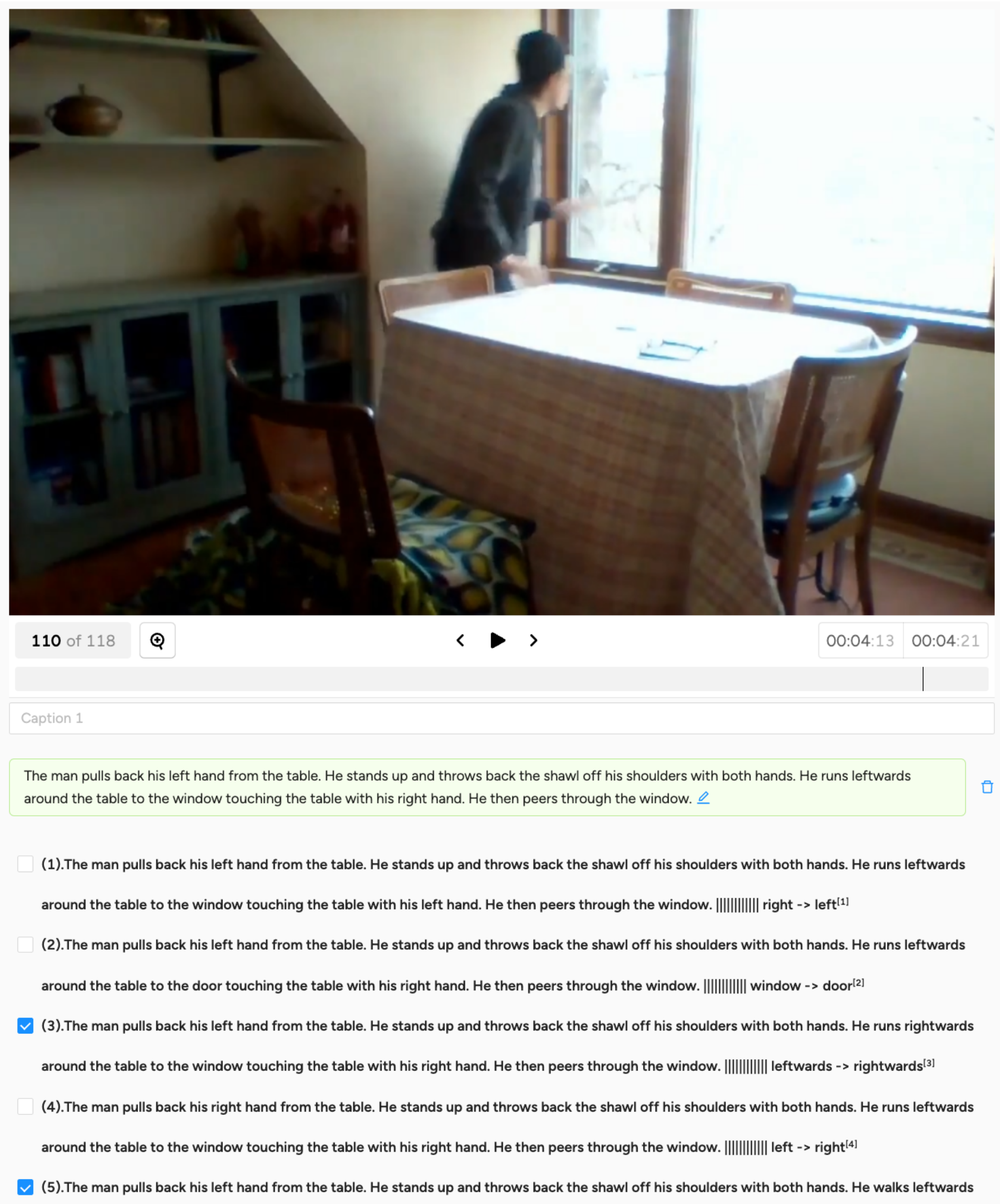
Model	Frames Per Video	Multiple Binary Accuracy (%)	Binary QA Accuracy (%)
Human	-	**67.9**	**89.7**
Random Chance	-	9.5	50.0
XCLIP	8	12.9	51.6
ImageBind	2	14.0	53.0
LanguageBind	8	14.5	52.8
GPT-4o	64	38.0	76.0
	32	38.2	75.9
	16	38.5	75.7
	8	37.3	75.1
	4	35.8	74.4
	2	33.2	72.7
	1	28.4	70.0
	0	26.5	67.7
Gemini-1.5-Pro	1FPS	26.6	67.5
	0	16.1	58.1
Claude-3.5-Sonnet	16	23.5	65.4
	8	23.6	65.5
	4	23.1	64.8
	2	21.2	61.8
	1	18.4	58.4
InternLM-XC25	1FPS	17.9	58.8
LLaVA-NeXT-Video-34B-DPO	32	22.0	64.0
	16	21.8	63.7
	8	21.4	63.3
	4	20.7	63.0
	2	19.9	61.8
	1	18.8	60.5
LLaVA-NeXT-Video-7B-DPO	32	17.2	59.5
	16	22.3	64.0
	8	23.6	65.1
	4	22.9	64.2
	2	21.4	63.0
	1	19.0	62.0
VideoLLaVA	8	25.5	67.1
Phi-3.5-Vision-Instruct	32	15.5	56.7
	16	15.9	57.2
	8	15.9	57.4
	4	15.5	57.5
	2	16.9	58.0
	1	16.4	57.7
Qwen2-VL-7B	32	24.7	64.4
	16	23.6	63.2
	8	21.0	60.9
	4	19.2	59.5
	2	17.6	57.8
Qwen2-VL-72B	32	38.3	75.8
	16	36.8	74.6
	8	33.8	73.0
	4	31.0	71.4
	2	27.3	69.1
MiniCPM-V-2.6	1FPS	21.4	62.3
LLaVA-1.5-13B	1	13.1	55.7
LLaVA-1.5-7B	1	18.3	60.5
Phi-3-Vision	1	15.1	54.4
Yi34B	0	18.7	59.9
Vicuna7B-1.5	0	10.4	50.55
Flan-T5-XL	0	17.9	57.9
Flan-T5-XXL	0	15.1	55.1



Women dressed in blue uniforms and hats walk in two lines while holding hands. They are smiling and waving at the crowd of people who are cheering them on by clapping their hands on the side. A group of men wearing heavy dark jackets and hat walk through while shaking hands and smiling at the crowd.

"The Beauxbatons girls leave in uniform, the crowd applauds.

Figure 11: Positive caption refinement platform.



The man pulls back his left hand from the table. He stands up and throws back the shawl off his shoulders with both hands. He runs leftwards around the table to the window touching the table with his right hand. He then peers through the window.

☐ (1).The man pulls back his left hand from the table. He stands up and throws back the shawl off his shoulders with both hands. He runs leftwards around the table to the window touching the table with his left hand. He then peers through the window. |||||||||||| right -> left[1]

☐ (2).The man pulls back his left hand from the table. He stands up and throws back the shawl off his shoulders with both hands. He runs leftwards around the table to the door touching the table with his right hand. He then peers through the window. |||||||||||| window -> door[2]

☑ (3).The man pulls back his left hand from the table. He stands up and throws back the shawl off his shoulders with both hands. He runs rightwards around the table to the window touching the table with his right hand. He then peers through the window. |||||||||||| leftwards -> rightwards[3]

☐ (4).The man pulls back his right hand from the table. He stands up and throws back the shawl off his shoulders with both hands. He runs leftwards around the table to the window touching the table with his right hand. He then peers through the window. |||||||||||| left -> right[4]

☑ (5).The man pulls back his left hand from the table. He stands up and throws back the shawl off his shoulders with both hands. He walks leftwards

Figure 12: Negative caption annotation platform.