TemporalBench: BENCHMARKING FINE-GRAINED TEMPORAL UNDERSTANDING FOR MULTIMODAL VIDEO MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding fine-grained temporal dynamics is crucial for video understanding. Yet, popular video benchmarks, such as MSRVTT and TGIF, often fail to effectively evaluate AI models' temporal reasoning abilities due to the lack of fine-grained temporal annotations. As a result, text-based models, leveraging strong language priors, often perform comparably to video models, and image-trained models have been reported to outperform their video-trained counterparts on MSRVTT and TGIF. This paper introduces a new *TemporalBench* benchmark for fine-grained temporal event understanding in videos. TemporalBench, sourced from a diverse video datasets, consists of $\sim$10K pairs of video description questions, derived from $\sim$2K high-quality human-annotated video captions. Uniquely, our benchmark provides fine-grained temporal annotations to evaluate models' temporal reasoning abilities. Our results show that state-of-the-art models like GPT-4o achieve only 38.0% multiple binary QA accuracy on TemporalBench, demonstrating a significant human-AI gap in temporal understanding. We hope that TemporalBench is instrumental to fostering research on improving models' temporal reasoning capabilities.

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 (OpenAI, 2024) and Gemini (Gemini Team, 2024) as well as open-sources ones (Liu et al., 2023a; Zhu et al., 2024b; Bai et al., 2023), that have demonstrated impressive results on existing video benchmarks (Xu et al., 2016; Chen & Dolan, 2011; Yu et al., 2019a; Mangalam et al., 2024). 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 (Antol et al., 2015). For example, prior works (Tan et al., 2024; Li et al., 2023a) show that language models such as Flan-T5 (Chung et al., 2024) and Llama-2/3 (Touvron et al., 2023) perform comparably to video models on EgoSchema (Mangalam et al., 2024) and Seed-Bench (Li et al., 2023a) 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 (Lei et al., 2023). 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) (Liu et al., 2024a;b) that are trained on image-level datasets, including FreeVA (Wu, 2024), IG-VLM (Kim et al., 2024) and M^3 (Cai et al., 2024b), often outperform their video counterparts on popular video question answering benchmarks such as MSRVTT (Xu et al., 2016), MSVD (Xu et al., 2017), and TGIF (Jang et al., 2017).

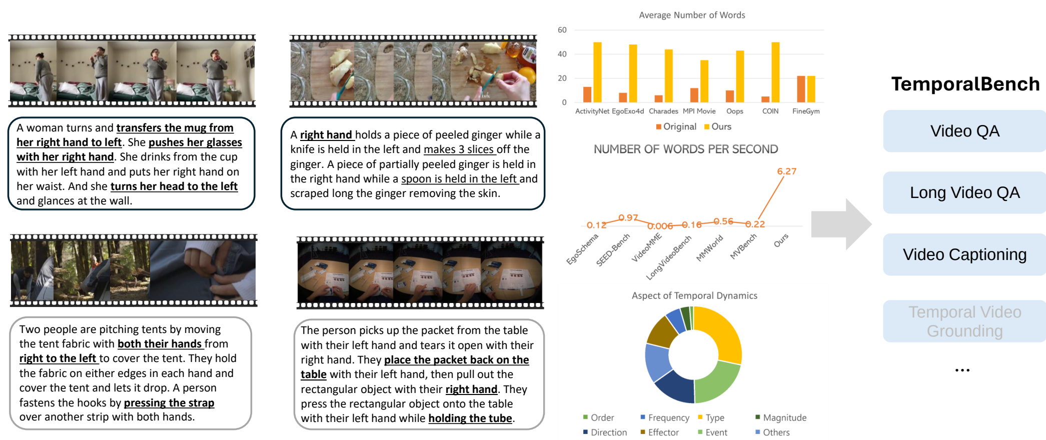


Figure 1: **The tasks of *TemporalBench***. *TemporalBench* starts from fine-grained video descriptions and supports diverse video understanding tasks including video QA, video captioning, long video understanding, *etc*. It differs from existing benchmarks by the average number of words per video (middle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

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 (Tang et al., 2019), human activities (Krishna et al., 2017; Gao et al., 2017) and ego-centric videos (Grauman et al., 2024). 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, 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 (>3 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

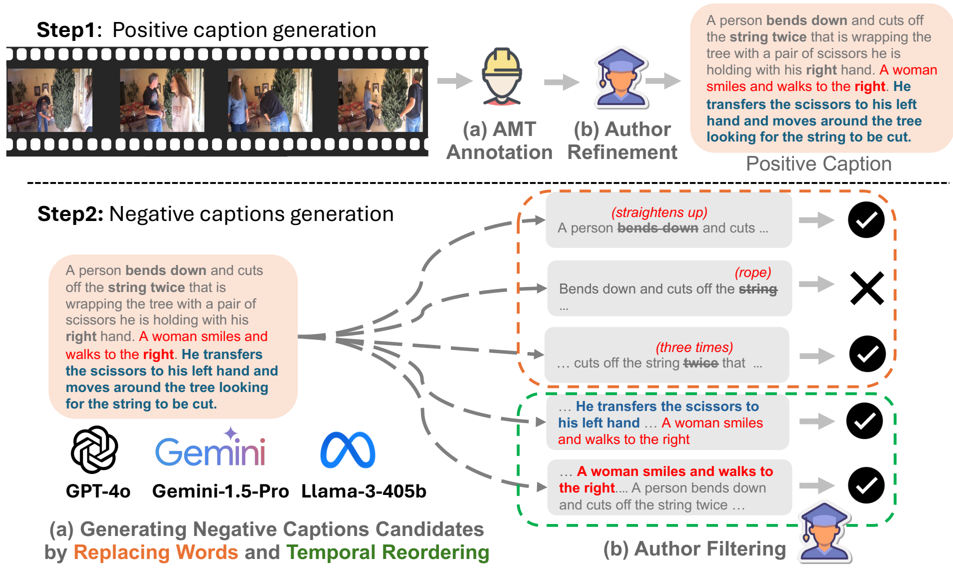


Figure 2: **Overview of the annotation pipeline for *TemporalBench*.** In step 1, we fist collect high-quality captions for the videos using qualified AMT annotators followed by refining them. In step 2, we leverage existing LLMs to generate negative captions by replacing select words and reordering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

and contrastive learning-based models such as XCLIP (Ni et al., 2022), ImageBind (Girdhar et al., 2023) as well as multimodal generative models such as GPT-4o and Gemini.

Among other observations, our empirical evaluations show that state-of-the-art multimodal video models like GPT-4o only achieve an average accuracy of 38.0% on our benchmark using our proposed multiple binary QA accuracy metric, compared to 67.9% obtained by humans. 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 (OpenAI, 2023b), GPT-4 (OpenAI, 2023c), and Llama (Touvron et al., 2023) 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)(OpenAI, 2023a). Building upon open-source LLMs (Touvron et al., 2023; Chiang et al., 2023), a wide range of multimodal models has achieved remarkable progress, led by pioneering models such as LLaVA (Liu et al., 2023a; 2024a) and MiniGPT-4 (Zhu et al., 2024b), which combine LLMs' capabilities with a CLIP (Radford et al., 2021) 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 (Cai et al., 2024a; Zhang et al., 2023c; Chen et al., 2023; Peng et al., 2023; Zhang et al., 2023b), 3D LMMs (Hong et al., 2023), and video LMMs (Lin et al., 2023; Zhang et al., 2023a; 2024b).

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 (Antol et al., 2015), GQA (Hudson & Manning, 2019) and VizWiz (Gurari et al., 2018) have been revitalized and used for evaluating the general visual question answering performance for LMMs. Some other question answering benchmarks like TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021) and InfoVQA (Mathew et al., 2022) have also been employed to validate the text-oriented understanding. Recent studies have introduced a variety of new benchmarks, such as SEED-Bench (Li et al., 2023a), MMBench (Liu et al., 2023b) and MM-Vet (Yu et al., 2024b) for evaluating the models' integrated problem-solving capabilities, and MMMU (Yue et al., 2024a) and MathVista (Lu et al.,

2024) for scientific and mathematical reasoning. In addition, the commonly known hallucination problem also appears in LMMs, and is also investigated in POPE (Li et al., 2023b), MMHal-Bench (Sun et al., 2023) and Object HalBench (Yu et al., 2024a), *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 benchmarks. Many efforts have leveraged existing video question answering benchmarks (Xu et al., 2017; Yu et al., 2019b; Xiao et al., 2021) built on top of video-text datasets (Chen & Dolan, 2011; Xu et al., 2016; Zhang et al., 2019). More recently, several LMM-oriented benchmarks have been proposed for different aspects such as long-form egocentric understanding with EgoSchema (Mangalam et al., 2024), and temporal understanding and ordering like Tempcompass (Liu et al., 2024c). MV-Bench (Li et al., 2024b) compiles existing video annotations from different disciplines into a new benchmark, while Video-MME (Fu et al., 2024) and MMWorld (He et al., 2024b) 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 (Chen et al., 2024) 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 (Xu et al., 2016) 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 (Wu, 2024) or even by a text-only LLM (Tan et al., 2024; Mangalam et al., 2024).

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 *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 (Xu et al., 2016) which focus on coarse-level descriptions.
- **Long video understanding and 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 VideoCLIP (Xu et al., 2021). Given a positive caption and several negative captions, we can evaluate whether CLIP (Radford et al., 2021) 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 (Lin et al., 2014) and Flickr30K (Young et al., 2014).
- **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 (Gao et al., 2017), COIN (Tang et al., 2019), Ego4D (Grauman et al., 2024) 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 types from seven sources, including (1) procedure videos *e.g.,* COIN (Tang et al., 2019), (2) human activities *e.g.,* ActivityNet-Captions (Yu et al., 2019a) and Charades (Krishna et al., 2017), (3) ego-centric videos *e.g.,* EgoExo4D (Grauman et al., 2024), (4) movie descriptions (Rohrbach et al., 2015), (5) professional gymnasium videos *e.g.,* FineGym (Shao et al., 2020), and (6) unexpected humor videos Oops (Epstein et al., 2020). 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 (Farnebäck, 2003), (2) contain multiple scene transitions by leveraging PySceneDetect [1] 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 3. 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.

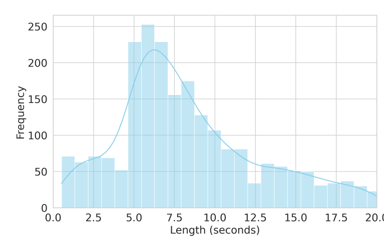


Figure 3: Video length distribution of *TemporalBench*.

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 (Shao et al., 2020) 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.

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

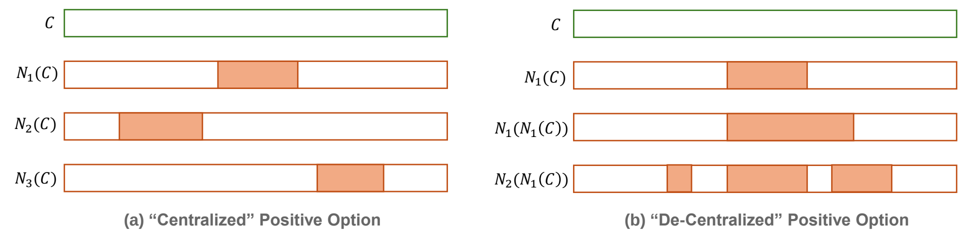


Figure 4: An illustration of multi-choice QA with (a) "centralized" and (b) "de-centralized" positive option. Orange blocks indicate the altered contents from the positive option (green box).

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 (OpenAI, 2024), Gemini-1.5-Pro (Gemini Team, 2024) and Llama-3.1-405b (Meta, 2024) 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 $\sim$2K high-quality human-annotated video captions and $\sim$10K video question-answer pairs.

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 (Yue et al., 2024a), MathVista (Lu et al., 2024), EgoSchema (Mangalam et al., 2024) etc. However, indicated by recent studies by (Cai et al., 2024b) and (Yue et al., 2024b), 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. 4). Surprisingly, we find that 62% of text-only GPT-4o's predictions correspond to $N(C)$, while only 18% of its predictions correspond to C. Our findings also align with human behavior analysis from psychology (Furman & Wang, 2008), 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 (OpenAI, 2024), Gemini-1.5-Pro (Gemini Team, 2024), Claude-3.5-Sonnet (Anthropic, 2024), Qwen2VL (Wang et al., 2024), LLaVA-OneVision (Li et al., 2024a), LLaVA-Next-Video (Zhang et al., 2024b), Phi-

3.5-Vision (Abdin et al., 2024), MiniCPM-2.6 (Yao et al., 2024), MA-LMM (He et al., 2024a), VideoLLaVA (Lin et al., 2023), InternLM-Xcomposer-2.5 (Zhang et al., 2024a), and (2) multimodal video embedding models, including XCLIP (Ni et al., 2022), ImageBind (Girdhar et al., 2023), and LanguageBind (Zhu et al., 2024a). We exponentially increase the number of frames to study its effect on video understanding. More details can be found in Appendix C.

To study the effect of single frame bias and text bias, we also evaluate models trained on single images, including LLaVA-1.5 (Liu et al., 2024a), LLaVA-NeXT (Liu et al., 2024b), and Phi-3V (Abdin et al., 2024). In the latter case, we evaluate the LLMs including GPT-4o (OpenAI, 2024), Gemini-1.5-Pro (Gemini Team, 2024), Yi-34B (Young et al., 2024), Vicuna (Chiang et al., 2023) and Flan-T5 (Wei et al., 2021) 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.

4.3 Fine-grained Video Question Answering

The results for multimodal generative models and embedding models are shown in Table 2. 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 5, 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.

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 76.0% accuracy but only 38.0% 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.

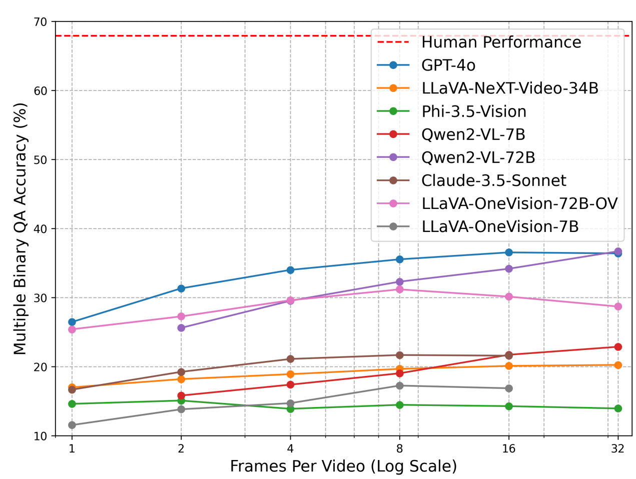


Figure 5: **Model performance on *TemporalBench*** with varying frames.

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

Table 1: Dataset characteristics including number of samples, average length, single image bias, and language bias.

Dataset	Number of Samples	Org. Avg. # words	Ours Avg. # words
ActivityNet (Krishna et al., 2017)	281	13.03	49.55
EgoExo4d (Grauman et al., 2024)	307	7.73	47.79
Charades (Gao et al., 2017)	298	6.21	44.16
MPI Movie Description (Rohrbach et al., 2015)	326	12.39	35.33
Oops (Epstein et al., 2020)	294	10.06	43.27
COIN (Tang et al., 2019)	385	5.01	50.06
FineGym (Shao et al., 2020)	288	21.92	21.92
TemporalBench (ours)	2179	10.9	41.72

Table 2: *TemporalBench* performance of various multimodal generative models and embedding models under the binary QA accuracy (BA) and multiple binary QA settings (MBA). The prefix "T-" indicates the annotated subset in our *TemporalBench*.

Model	T-ActivityNet	T-Charades	T-FineGym	T-Movie	T-Oops	T-COIN	T-EgoExo	BA	MBA
Human Performance	**69.4**	**81.9**	**35.8**	**74.5**	**69.7**	**70.6**	**70.7**	**89.7**	**67.9**
Random Chance	11.0	13.7	6.1	12.0	5.6	11.1	5.6	50.0	9.4
Video Embedding Models: Text + Multiple Frames as Input									
XCLIP	14.2	16.1	7.3	19.9	8.8	15.2	6.8	51.6	12.8
ImageBind	17.4	16.8	7.3	19.0	11.2	16.1	9.1	53.0	14.0
LanguageBind	22.4	15.1	6.3	19.3	10.9	15.6	11.1	52.8	14.5
Video Multimodal Generative Models : Text + Multiple Frames as Input									
GPT-4o	48.8	42.6	16.7	43.9	34.4	42.9	34.5	76.0	38.0
Gemini-1.5-Pro	34.9	24.5	8.0	35.6	22.8	34.0	21.8	67.4	26.4
Claude-3.5-Sonnet	29.5	27.5	13.2	29.1	14.6	27.8	21.2	65.9	23.5
InternLM-XC2.5	25.3	34.9	19.4	38.7	25.9	18.2	16.6	58.7	25.2
VideoLLaVA	34.9	29.2	13.5	25.5	20.7	32.5	20.2	67.2	25.5
MA-LMM	12.1	16.8	3.1	11.7	4.8	11.9	4.9	48.0	9.3
Phi-3.5-Vision	24.9	20.1	5.2	22.7	12.2	18.2	13.7	58.0	16.8
MiniCPM-V2.6	33.1	25.8	7.6	29.1	13.6	22.9	16.0	62.2	21.3
LLaVA-NeXT-Video-7B	33.5	32.6	10.8	28.2	17.3	22.6	19.9	65.1	23.5
LLaVA-NeXT-Video-34B	30.6	26.8	10.4	24.8	18.0	24.9	17.3	64.0	22.0
LLaVA-OneVision-7B	30.2	27.5	7.6	25.8	16.0	22.1	14.3	60.0	19.7
LLaVA-OneVision-72B	43.8	34.2	11.5	35.3	27.9	33.0	28.3	70.5	30.7
Qwen2-VL-7B-Instruct	32.4	31.9	4.5	35.9	18.4	25.2	21.8	64.6	24.9
Qwen2-VL-72B-Instruct	43.4	42.6	16.7	45.1	36.4	43.4	37.1	75.8	38.2
Large Multimodal Models (LMMs): Text + 1 Frame as Input									
GPT-4o	32.0	30.2	15.3	31.0	26.5	33.8	27.7	70.0	28.4
LLaVA-1.5-13B	16.0	17.1	9.4	16.6	6.1	16.1	9.1	55.6	13.1
LLaVA-1.5-7B	25.3	25.8	8.7	19.3	9.2	22.1	16.6	60.5	18.3
Phi-3-Vision-128k-Instruct	22.8	19.8	4.5	17.8	8.5	17.7	14.7	54.4	15.3
Large Language Models (LLMs): Text as Input									
GPT-4o	30.2	31.9	16.7	27.9	22.8	27.5	28.0	67.7	26.5
Gemini-1.5-Pro	22.4	20.5	4.5	19.9	10.2	16.6	17.9	58.0	16.0
Yi-34B	20.6	27.5	10.4	21.8	11.2	23.4	16.9	59.9	18.3
Vicuna7b-1-5	19.2	17.4	6.6	11.0	5.1	12.5	7.8	50.4	9.8
Flan-T5-XL	24.6	23.5	5.6	19.9	11.9	23.1	14.0	57.8	17.8

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 5 and Table 6, the performance of models like GPT-4o gradually increases with more frames. Excluding GPT-4o, all remaining VLMs are trained with single images *e.g.,* LLaVA-1.5, Phi-3V, and text-only LLMs such as Yi-34B and Vicuna-7B.

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. We evaluate the resulting video captioning performance using classical image captioning metrics,

Table 3: 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	**63.47**	6.59	**19.99**	23.70	**11.74**	**5.90**	**3.09**
Gemini-1.5-Pro	56.54	**10.98**	19.11	18.96	9.19	4.53	2.36
Claude-3.5-Sonnet	54.13	8.64	17.14	24.35	10.32	4.43	2.05
VideoLLaVA	45.97	4.49	16.95	12.59	5.44	2.29	1.03
MA-LMM	38.72	3.07	14.99	10.09	4.81	2.24	1.06
Phi-3.5-Vision	42.93	3.67	16.54	20.36	8.38	3.40	1.58
MiniCPM	47.24	1.50	14.18	15.53	5.45	1.92	0.79
LLaVA-NeXT-Video-7B	50.09	2.31	15.84	18.07	6.98	2.60	1.05
LLaVA-NeXT-Video-34B	53.13	5.33	15.92	21.43	9.17	4.02	1.83
LLaVA-OneVision-7B	50.33	1.43	16.08	16.17	6.99	2.92	1.33
LLaVA-OneVision-72B	53.90	8.00	18.23	22.08	10.63	5.31	2.78
Qwen2-VL-7B-Instruct	51.93	6.87	18.03	12.45	6.07	3.00	1.56
Qwen2-VL-72B-Instruct	56.13	9.31	19.11	15.71	8.03	4.14	2.24
Large Multimodal Models (LMMs): Text + 1 Frame as Input							
GPT-4o	52.32	7.29	17.10	**25.07**	11.09	5.04	2.41
LLaVA-1.5-13B-HF	47.92	4.90	18.04	22.62	9.78	4.23	2.03
LLaVA-1.5-7B-HF	45.68	6.87	17.82	21.95	9.53	4.17	1.98
Phi-3-Vision-128k-Instruct	41.96	4.00	16.10	19.86	8.29	3.42	1.59

CIDEr (Vedantam et al., 2015), BLEU (Papineni et al., 2002) at different n-gram levels, ROUGE (Lin, 2004), as well as the embedding similarity with sentence transformer (Reimers & Gurevych, 2019) between the ground truth caption and the generated caption.

Results in Table 3 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 better performance with one compared to 64 frames on both CIDEr and BLEU scores (e.g., for CIDEr 7.29 vs. 6.59). 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.

4.5 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 whose original length is both short (AcitivityNet and Charades, average length < 3 minutes) and long (COIN and FineGym, > 20 minutes). 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 set the number of negative options to be $\sim$4, resulting in a similar random chance performance as in Sec 4.3. In this way, we investigate whether multimodal video models can understand and distinguish fine details in a long video.

We show in Table 7 (supplemental), 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 (ActivityNet and Charades) compared to longer videos (COIN and 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 (Kamradt, 2023) 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 5: *TemporalBench* statistics on each category on binary QA accuracy.

Action Order	Action Frequency	Action Type	Motion Magnitude	Motion Direction	Action Effector	Event Reorder	Others	Overall
130	531	2812	321	1554	1118	2105	1347	9918

5 In-Depth Analysis

5.1 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 4, text-only GPT-4o displays different behaviors. As shown in Table 4, under the intentionally designed negative options, GPT-4o will choose $N_1(C)$ with 66.4% probability. This again demonstrates the necessity and advantage of our multiple binary QA accuracy (MBA) metric design over the standard multi-choice QA setting.

5.2 Performance on categories

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

Percentage of Predictions Aligned with ->	C	$N_1(C)$
"Centralized" Negative	83.3	6.4
"De-Centralized" Negatives	17.7	66.4

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 vs pull); 4. Motion magnitude (slightly vs intensively); 5, Motion Direction/Orientation (forward vs backward, circular vs back-and-forth). 6. Action effector (cutting with left hand vs cutting with right hand) 7. Others. We prompt GPT-4o to perform 7-way classification and show the per-category performance in Table 8 (supplemental). 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.

6 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 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.

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.

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.

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.

References

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.

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

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.

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.

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*, 2024a.

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

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.

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.

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.

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/.

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.

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

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.

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.

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

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.

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

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.

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.

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.

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)*, 2024a.

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, 2024b. URL https://arxiv.org/abs/2406.08407.

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.

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.

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.

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

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.

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.

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.

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*, 2024a.

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*, 2023a.

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, 2024b. URL https://arxiv.org/abs/2311.17005.

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 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20.

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.

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.

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.

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

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

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 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

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*, 2023b.

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?, 2024c. URL https://arxiv.org/abs/2403.00476.

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.

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.

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.

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.

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

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.

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

OpenAI. Chatgpt. https://openai.com/blog/chatgpt/, 2023b.

OpenAI. Gpt-4 technical report. 2023c.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021.

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.

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.

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.

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.

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, 2024a.

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*, 2024b.

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, 2019a.

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, 2019b. URL https://arxiv.org/abs/1906.02467.

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*, 2024a.

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*, 2024b.

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.

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

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, 2023b.

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*, 2024a.

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*, 2023c.

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 2024b. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

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*, 2024a.

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

A 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.

B 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 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 6 & 7. The video examples (*a* - *f*) are displayed in the same order as their sources in Table 1 (7 in total).

C More results with extended frames

In the main paper, we only report the performance of each multimodal video models with the the number of frames that leads to the best performance. Here we extend the results to show the results of more frames in Table 6.

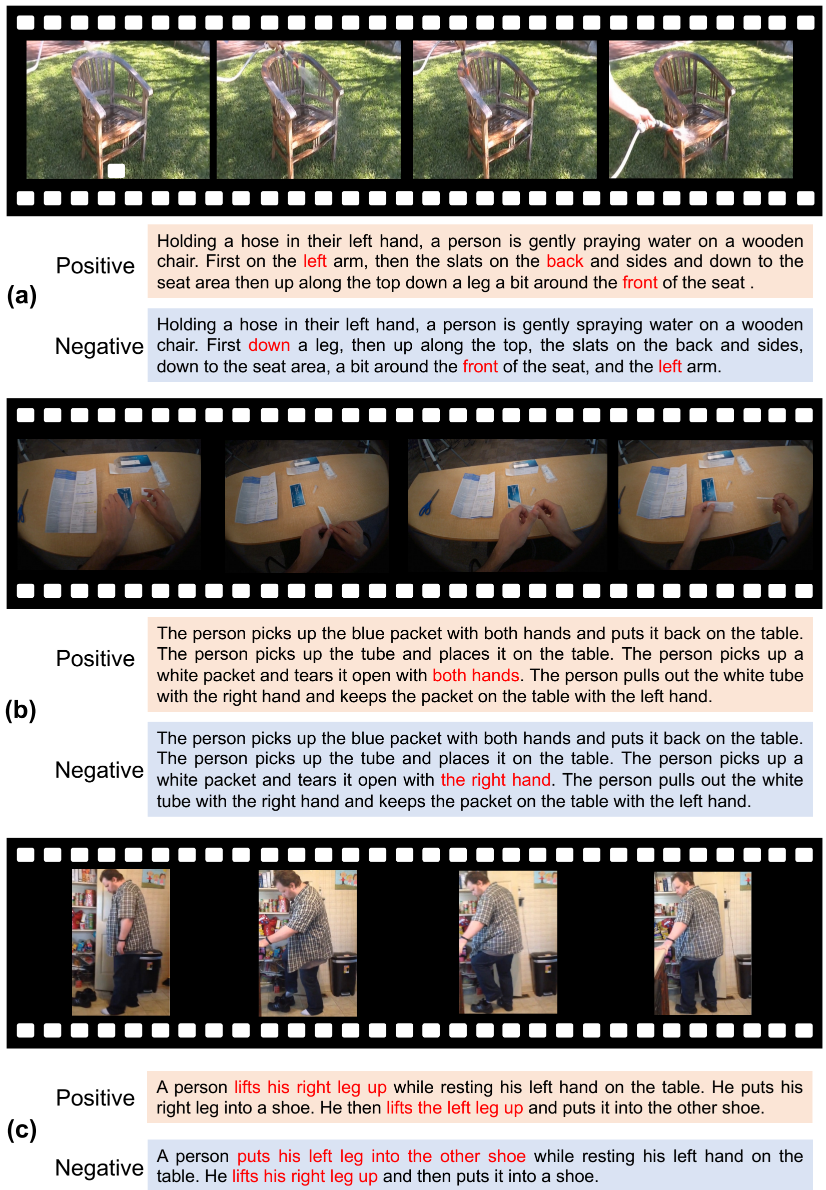
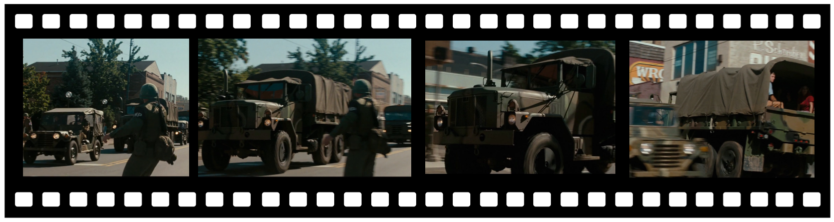


Figure 6: Visualizations (I) of our fine-grained annotations of the videos with both positive and negative descriptions.



(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)

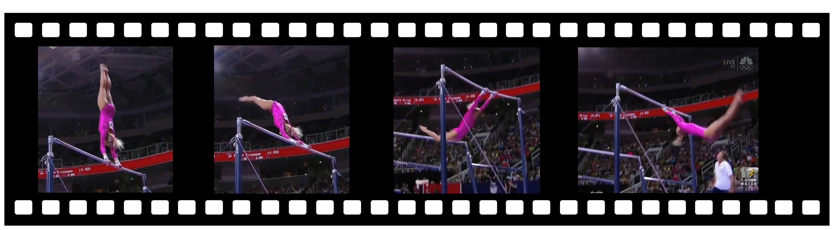
Positive: Two deer come out of the trees and run along a road into the trees on the other side. A third deer trips as it approaches the road, then turns back around and goes back to where it came from.

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

Table 6: *TemporalBench* performance of various models under binary QA and multiple binary QA setting.

Model	Frames Per Video	Multiple Binary Accuracy (%)	Binary QA Accuracy (%)
Human	-	**67.9**	**89.7**
Random Chance	-	9.4	50.0
XCLIP	8	12.8	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	38.4	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	26.4	67.4
	0	16.0	58.0
Claude-3.5-Sonnet	16	23.5	65.9
	8	23.6	65.4
	4	23.0	64.7
	2	21.2	61.8
	1	18.4	58.4
InternLM-XC25	1fps	25.2	58.7
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	61.9
	1	18.8	60.5
LLaVA-NeXT-Video-7B-DPO	32	17.2	59.6
	16	22.3	64.0
	8	23.5	65.1
	4	22.9	64.2
	2	21.4	63.1
	1	19.0	62.0
VideoLLaVA	8	25.5	67.2
Phi-3.5-Vision-Instruct	15.5	56.7	
	16	15.9	57.2
	8	15.9	57.4
	4	15.5	57.5
	2	16.8	58.0
	1	16.4	57.8
Qwen2-VL-7B-Instruct	32	24.9	64.6
	16	23.5	63.2
	8	20.9	60.9
	4	19.2	59.5
	2	17.6	57.8
Qwen2-VL-72B-Instruct	38.2	75.8	
	35.5	74.4	
	33.8	73.0	
	31.0	71.4	
	27.3	69.1	
MiniCPM-V-2.6	64	21.3	62.2
LLaVA-1.5-13B-HF	1	13.1	55.6
LLaVA-1.5-7B-HF	1	18.3	60.5
Phi-3-Vision-128k-Instruct	1	15.3	54.4
Vicuna7B-1.5	0	9.8	50.4
Yi34BNousYi	0	18.3	59.9
FastChat-FlanT5	0	11.9	52.2
Flan-T5-XL	0	17.8	57.8

Table 7: *TemporalBench* performance of various multimodal generative models and embedding models under long video understanding with multiple binary QA accuracy (MBA).

Model	ActivityNet	Charades	FineGym	COIN	MBA
Video Embedding Models: Text + Multi Frame as Input					
XCLIP	2.99	5.34	1.87	2.92	3.27
ImageBind	2.69	3.40	4.27	3.50	3.40
LanguageBind	4.78	6.31	3.79	2.72	4.00
Video Multimodal Generative Models : Text + Multi Frame as Input					
GPT-4o	20.30	21.36	9.81	17.12	**17.12**
Gemini-1.5-Pro	15.52	9.71	8.66	16.54	14.58
Claude-3.5-Sonnet	19.10	10.68	4.78	5.64	9.70
VideoLLaVA	8.96	6.80	5.07	2.14	5.29
MA-LMM	7.76	6.80	3.28	8.37	5.60
Phi-3.5-Vision	8.06	2.43	6.57	3.50	5.23
MiniCPM	8.36	6.80	3.88	9.53	9.97
LLaVA-NeXT-Video-7B	10.45	8.74	2.69	7.39	8.00
LLaVA-NeXT-Video-34B	10.75	10.68	4.78	3.11	6.90
LLaVA-OneVision-7B	8.66	7.77	5.07	8.56	8.52
LLaVA-OneVision-72B	14.93	10.19	4.18	5.25	8.65
Qwen2-VL-72B-Instruct	14.33	10.68	11.04	14.40	14.56
Large Multimodal Models (LMMs): Text + 1 frame as Input					
GPT-4o	10.45	12.62	8.33	11.67	10.80
LLaVA-1.5-13B-HF	6.57	5.34	3.88	3.89	4.84
LLaVA-1.5-7B-HF	4.78	5.34	2.69	3.89	4.01
Phi-3-Vision-128k-Instruct	8.36	4.85	3.58	4.67	5.50
Large Language Models (LLMs): Text as Input					
GPT-4o	11.64	16.99	7.16	10.70	11.01
Gemini-1.5-Pro	11.64	8.74	2.99	7.98	7.77
Yi-34B	7.16	7.28	5.37	6.61	6.55
Vicuna7b-1-5	1.19	4.85	1.49	3.70	2.73
Flan-T5-XL	12.24	7.28	7.46	7.39	8.56

Table 8: *TemporalBench* performance under each category.

Model	Action Order	Action frequency	Action Type	Motion Magnitude	Motion Direction	Action Effector	Event Reorder	Others	Average
Video Embedding Models: Text + Multi Frame as Input									
XCLIP	46.2	50.8	50.9	56.9	51.2	51.6	50.2	55.5	51.6
ImageBind	43.8	44.8	55.4	51.1	52.5	50.4	48.6	62.0	53.0
LanguageBind	43.8	41.6	53.3	54.8	51.5	46.4	51.1	66.0	52.8
Video Multimodal Generative Models : Text + Multi Frame as Input									
GPT-4o	70.0	65.2	**80.8**	**78.8**	68.9	67.0	75.1	**87.3**	**76.0**
Gemini-1.5-Pro	66.9	60.1	70.8	70.7	58.6	59.5	67.7	79.0	67.4
Claude-3.5-Sonnet	63.8	58.0	71.1	68.2	60.0	57.4	62.5	76.7	65.9
InternLM-XC2.5	53.8	42.4	61.2	61.4	52.4	52.4	59.3	68.3	58.7
VideoLLaVA	70.0	**70.2**	71.4	70.1	**70.7**	**70.3**	50.5	75.6	67.2
MA-LMM	54.6	42.7	48.7	48.9	46.2	49.4	49.1	50.8	48.0
Phi-3.5-Vision	53.8	55.4	60.0	56.1	53.9	52.2	55.3	69.4	58.0
MiniCPM	58.5	52.4	65.6	62.3	54.1	53.2	63.3	74.7	62.2
LLaVA-NeXT-Video-7B	68.5	65.5	68.1	62.0	66.6	68.7	52.3	74.2	65.1
LLaVA-NeXT-Video-34B	60.8	56.1	66.4	61.7	58.4	59.5	63.3	74.3	64.0
LLaVA-OneVision-7B	60.8	44.6	61.4	53.0	50.1	48.2	66.0	74.9	59.8
LLaVA-OneVision-72B	68.5	53.7	74.6	67.9	63.7	62.0	71.2	83.0	70.5
Qwen2-VL-7B-Instruct	65.4	46.1	67.3	66.0	54.5	54.9	69.3	75.3	64.6
Qwen2-VL-72B-Instruct	**72.3**	69.3	80.0	**78.8**	65.9	69.4	**75.9**	85.5	75.8
Large Multimodal Models (LMMs): Text + 1 frame as Input									
GPT-4o	67.7	65.2	74.0	70.4	64.3	62.7	68.6	78.5	70.0
LLaVA-1.5-13B-HF	56.9	52.0	57.6	53.6	50.3	53.9	54.2	63.2	55.6
LLaVA-1.5-7B-HF	61.5	61.4	62.1	54.2	61.6	65.0	51.1	67.9	60.5
Phi-3-Vision-128k-Instruct	46.2	46.3	56.2	55.8	48.8	49.6	56.9	62.3	54.4
Large Language Models (LLMs): Text as Input									
GPT-4o	64.6	59.9	73.7	70.1	61.5	60.2	69.3	68.7	67.7
Gemini-1.5-Pro	53.8	42.4	60.3	62.3	53.5	53.2	64.8	57.4	58.0
Yi-34B	53.1	63.1	59.9	60.4	56.7	54.8	65.2	59.3	59.9
Vicuna7b-1-5	56.2	47.3	52.9	50.5	50.3	48.6	49.9	53.5	50.4
Flan-T5-XL	53.1	57.8	60.1	59.8	56.0	56.7	54.9	60.5	57.8