

000 001 *TemporalBench*: EVALUATING FINE-GRAINED TEMPO- 002 RAL DYNAMICS UNDERSTANDING FOR MULTIMODAL 003 MODELS 004 005

006 **Anonymous authors**
007 Paper under double-blind review
008
009
010

011 ABSTRACT 012

013 Understanding fine-grained temporal dynamics is crucial for multimodal video
014 comprehension and generation. Due to the lack of fine-grained temporal annotations,
015 existing video benchmarks mostly resemble static image benchmarks and
016 are insufficient at evaluating models for temporal understanding. In this paper,
017 we introduce *TemporalBench*, a benchmark dedicated to evaluating **fine-grained**
018 **temporal understanding** in videos. *TemporalBench* consists of $\sim 15K$ video
019 question-answer pairs, derived from $\sim 2K$ high-quality human annotations detailing
020 the temporal dynamics. As a result, our benchmark provides a unique testbed
021 for evaluating various temporal understanding and reasoning abilities such as *ac-*
022 *tion frequency, motion magnitude, event order, etc.* Moreover, it enables evalua-
023 tions on various tasks such as both short and long video understanding, as well
024 as different models including multimodal embedding models and text generation
025 models. Furthermore, we notice a critical pitfall for multi-choice QA where LLMs
026 can detect the subtle changes in negative captions and find a “centralized” descrip-
027 tion as a cue for its prediction, and we propose Multiple Binary Accuracy (MBA),
028 a new metric for dense temporal understanding, to correct such bias. Results show
029 that state-of-the-art models like Gemini-2.5-Pro achieve only 43.6% question an-
030 swering accuracy on *TemporalBench* short video QA, demonstrating a significant
031 gap ($\sim 24\%$) between humans and AI in temporal understanding. We hope that
032 *TemporalBench* can foster research on improving models’ temporal reasoning ca-
033 pabilities. Both dataset and code will be available.
034

035 1 INTRODUCTION 036

037 Videos capture the temporal evolution of visual scenes through sequences of frames. This temporal
038 dimension is fundamental to video understanding—distinguishing it from static image comprehen-
039 sion. Yet recent evaluations reveal a surprising trend: vision-language models trained solely on
040 images often match or exceed the performance of video models on popular benchmarks (Xu et al.,
041 2016; Jang et al., 2017). Even more striking, models processing just a single frame can achieve
042 competitive accuracy on tasks ostensibly requiring temporal reasoning (Wu, 2024; Kim et al., 2024).
043 This single-frame bias suggests that existing benchmarks fail to genuinely evaluate temporal under-
044 standing.

045 The root cause lies in the coarse-grained nature of current video annotations. Consider a typical
046 question-answer pair: “What is the person doing?” with answer “cooking.” Such descriptions cap-
047 ture only high-level semantics that are often visible in any individual frame. The temporal dynam-
048 ics—how the cooking unfolds, the sequence of actions, the precise manipulations—remain unanno-
049 tated. Without these fine-grained temporal details, benchmarks reduce to static scene recognition
050 tasks where temporal reasoning becomes optional rather than essential.

051 This limitation has profound implications. Models achieve high scores without learning to process
052 temporal information, creating an illusion of video understanding. The benchmarks meant to drive
053 progress in temporal reasoning instead reward static visual recognition. The gap between benchmark
performance and genuine temporal comprehension remains unmeasured and unaddressed.

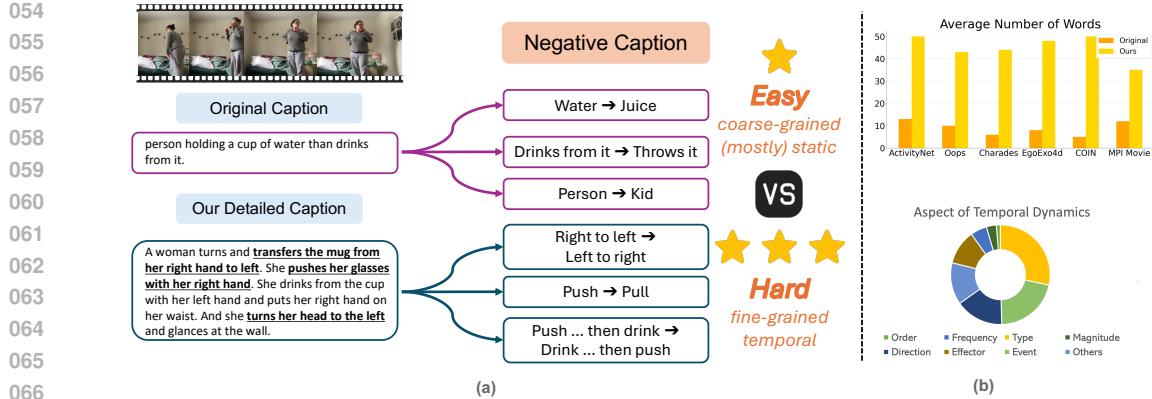


Figure 1: (a) **Comparison between negative captions generated from the original caption and our detailed caption in *TemporalBench*.** With fine-grained details, the negatives are more temporal centric and difficult. (b) *TemporalBench* differs from existing benchmarks by the average number of words per video (left), and the coverage of various temporal aspects (right).

To address this fundamental issue, we introduce *TemporalBench* (Figure 1), a benchmark specifically designed to evaluate fine-grained temporal understanding in videos. Our benchmark consists of $\sim 15K$ question-answer pairs (10K short, 5K long) derived from $\sim 2K$ meticulously annotated video captions that capture rich temporal dynamics. Unlike existing benchmarks, our annotations detail the precise sequence of actions, their duration, frequency, and temporal relationships—information that cannot be inferred from a single frame.

The key insight behind *TemporalBench* is simple: by annotating temporal dynamics with fine-grained detail, we create evaluation tasks that inherently require temporal reasoning. When a caption specifies “slices the ginger three times,” distinguishing it from “slices the ginger twice” demands processing multiple frames. When describing “transfers the mug from right hand to left, then pushes glasses up,” understanding the action sequence requires temporal comprehension. These detailed annotations transform video understanding from optional to essential.

As shown in Figure 2, we collect videos spanning diverse domains: procedural videos (Tang et al., 2019), human activities (Krishna et al., 2017; Gao et al., 2017), ego-centric videos (Grauman et al., 2024), movie clips (Rohrbach et al., 2015), professional gymnastics (Shao et al., 2020), and unexpected events (Epstein et al., 2020). Each video receives detailed temporal annotation through a rigorous two-stage process involving qualified annotators and expert refinement. We then generate challenging negative captions using large language models, creating minimal but semantically significant variations that test precise temporal understanding.

TemporalBench offers three defining characteristics:

- **Fine-grained temporal distinctions.** Our negative captions probe specific temporal aspects: action frequency (“twice” vs “three times”), ordering (“then” vs “before”), and dynamics (“pushes” vs “pulls”)—details invisible in single frames.
- **Short and long video evaluation.** We test both immediate temporal understanding (<20 seconds) and extended temporal reasoning (<20 minutes) by composing multiple clip descriptions.
- **Unified evaluation framework.** *TemporalBench* supports both discriminative models (XCLIP (Ni et al., 2022), ImageBind (Girdhar et al., 2023)) and generative models (GPT-4o, Gemini), enabling comprehensive assessment across architectures.

During evaluation, we identified a critical flaw in standard multi-choice question answering: when all incorrect options derive from small modifications to the correct answer, models can exploit the “centralized” pattern without understanding content. We address this with Multiple Binary Accuracy (MBA), decomposing each multi-choice question into independent binary decisions that eliminate such shortcuts.

Our results reveal the severity of the temporal understanding gap. State-of-the-art GPT-4o achieves only 38.5% MBA on short videos, compared to 67.9% human performance—a $\sim 30\%$ deficit. Performance degrades further on long videos, confirming that current models, despite high scores on

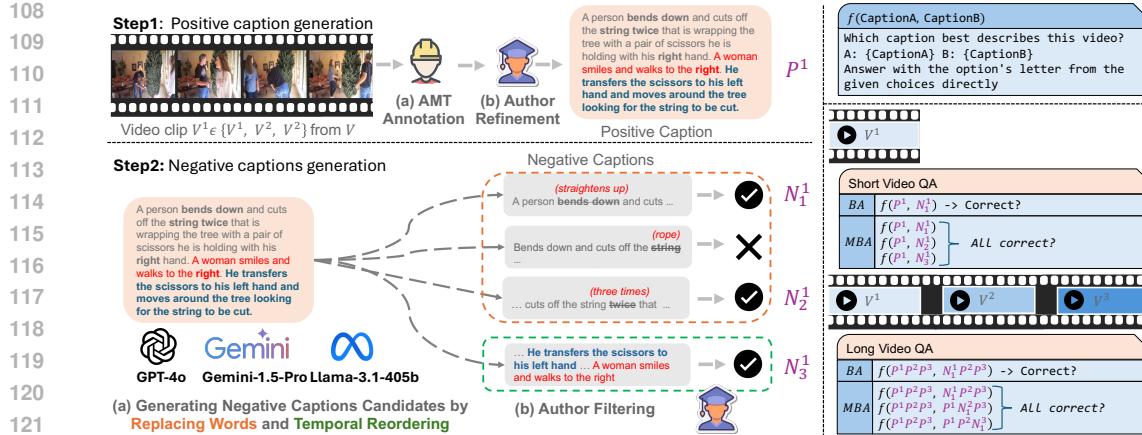


Figure 2: **Overview of the annotation and evaluation pipeline of *TemporalBench*.** **Left:** In step 1, we first collect high-quality captions for the videos using qualified AMT annotators followed by refining them. In step 2, we leverage existing LLMs to generate negative captions by replacing select words and reordering the sequence of actions before filtering them ourselves. **Right:** *TemporalBench* supports both short and long video QA, where the descriptions of the long video are concatenated from its video clips.

existing benchmarks, lack genuine temporal reasoning capabilities. *TemporalBench* thus provides both a diagnosis of current limitations and a path toward developing models with true temporal understanding.

2 RELATED WORK

Large Multimodal Models. The proliferation of Large Multimodal Models (LMMs), from proprietary systems like GPT-4V (OpenAI, 2023) and Gemini Gemini Team (2024), to open-source models like LLaVA Liu et al. (2023a) and Qwen-VL series (Bai et al., 2023), has created a need for more challenging evaluation benchmarks. While these models have demonstrated impressive capabilities in image understanding, their ability to reason about fine-grained temporal dynamics remains under-explored.

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), MMHall-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 Shangguan et al. (2024); Zhou et al. (2018); Kesen et al. (2024); Li et al. (2024d); Chen et al. (2024); Cores et al. (2024). 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; Li et al., 2024d; Shangguan et al., 2024) 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). VideoCon Bansal et al. (2024) curated

negative captions using LLM from the original caption, yet lacks fine-grained details shown in Figure 1. MV-Bench (Li et al., 2024c) 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. YouCook2 Zhou et al. (2018) emphasizes procedure videos but lacks fine-grained temporal understanding. VITATECS Li et al. (2024d), TempCompass Liu et al. (2024c), TOMATO Shangguan et al. (2024) and TVBench Cores et al. (2024) work towards better model temporal dynamics via the counterfactual manner, but still lack the dense captions for fine-grained details. 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.

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 existing video understanding benchmark is the question, “*What action is happening in the video?*” with the answer, “*Cooking*.” 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 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. For example, in Figure 1, models only need to choose between “*The man is cooking*” and “*The man is exercising*”, which can be answered without understanding the temporal aspects.

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 multimodal 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*”.
- **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 (Ni et al., 2022). Given a positive caption and several negative captions, we can evaluate whether CLIP (Radford et al., 2021) based

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

Dataset	# Samples	Org # words	Ours # words
ActivityNet	281	13.03	49.55
EgoExo4D	307	7.73	47.79
Charades	298	6.21	44.16
Movie Description	326	12.39	35.33
Oops	294	10.06	43.27
COIN	385	5.01	50.06
FineGym	288	21.92	21.92
<i>TemporalBench</i> (ours)	2179	10.91	42.02

216

217

218
219
220
Table 2: Comparsion between *TemporalBench* and other video understanding benchmarks. # Words/s
denote the average number of words per second, and A & M denotes Auto & Manual during the
annotation process.

221

222

223

224

225

226

227

228

Dataset	Num QA	# Video	Open-domain	Annotation	Avg Sec	# Words/s
MVBBench	4,000	3,641	✓	Auto	16	0.22
EgoSchema	5,000	5,063	✗	A & M	180	0.12
LongVideoBench	6,678	3,763	✓	Manual	473	0.16
Video-MME	2,700	900	✓	Manual	1018	0.006
SEED-Bench	3,757	2,320	✗	A & M	110	0.97
MMWord	6,627	1,910	✓	A & M	102	0.56
<i>TemporalBench</i> (short)	9,867	2,179	✓	A & M	8	6.27
<i>TemporalBench</i> (long)	5,485	1,574	✓	A & M	123	2.62

229

230

231

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).

232

233

234

235

Our details video captions support tasks like fine-grained video captioning shown in the Supp., and enable future tasks such as video grounding from detailed text descriptions and text-to-video generation with detailed prompts.

236

237

3.1 VIDEO COLLECTION

238

We collect video clips from a wide range of sources across *diverse* domains, where the majority comes from existing video grounding dataset test/val split. 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, comparsion with existing benchmarks are shown in Table 2 and also Table 6 in the appendix.

248

249

250

251

252

253

254

We intentionally filter out video clips that (1) are mostly static by leveraging optical flow (Farnebäck, 2003), (2) contain multiple scene transitions by detecting shot changes in videos 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 the Supp. 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.

255

256

3.2 VIDEO CAPTION ANNOTATION PROCESS

257

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, 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.

258

259

260

261

262

263

264

We first use 3 probing video captioning questions with 2 in-context examples as the onboarding task for AMT master workers, where prompts are shown in the Supp. 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.

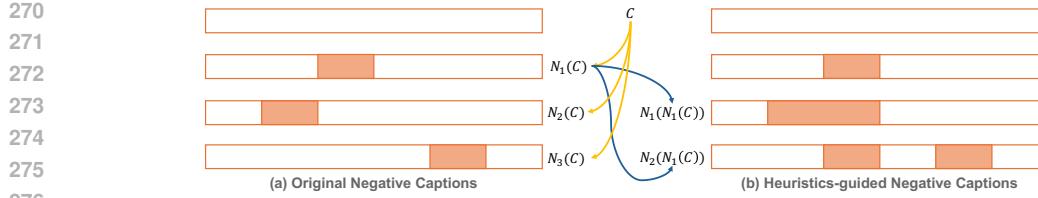


Figure 3: An illustration of multi-choice QA with (a) original and (b) heuristics-guided negative captions. Orange blocks indicate the altered contents from the positive option (green box).

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 (Dubey et al., 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. Prompts are shown in the Supp.

Afterwards, the authors of this work review those negative caption candidates in the format of multi-choice QA. Essentially, we build a reliable evaluation pipeline w.r.t. the detailed captioning, alleviating the limitation of previous This process results in our complete *TemporalBench* dataset with $\sim 2K$ high-quality human-annotated video captions and $\sim 10K$ (short) video question-answer pairs.

3.3 A PITFALL IN MULTI-CHOICE QUESTION ANSWERING

Multi-choice QA is standard for evaluating LMMs (MMMU (Yue et al., 2024a), MathVista (Lu et al., 2024), EgoSchema (Mangalam et al., 2024)). However, recent studies (Cai et al., 2024; Yue et al., 2024b) show that pure LLMs often match or exceed LMM performance without visual input. This is attributed to (1) poorly designed questions answerable without the video, or (2) data contamination.

We identify another critical pitfall. If all negatives are slight modifications of the correct answer, LLMs can detect a “centralized” description as a cue (Figure 3). To test this, given positive caption C and negative $N(C)$, we derive further negatives from $N(C)$ (e.g., $N_1(N(C))$, $N_2(N(C))$). With options $[C, N(C), N_1(N(C)), N_2(N(C))]$, $N(C)$ becomes “centralized”. Surprisingly, text-only GPT-4o selects $N(C)$ 66.4% of the time, and C only 6.4%. This aligns with psychology findings (Furman & Wang, 2008) where humans exploit similarity cues in multiple choice.

Motivated by this, we propose decomposing multi-choice QA into multiple binary QAs, eliminating the “centralized option.” We introduce Multiple Binary Accuracy (MBA), which reflects the compounded difficulty of answering *all* sub-questions: $\text{MBA} = \prod_{j=1}^M \text{acc}_j$, where acc_j is the correctness of sub-question j (e.g., $92\%^5 = 65.9\%$). Given M negatives, random chance drops from $\frac{1}{M+1}$ (multi-choice) to $(\frac{1}{2})^M$ (MBA). Since $(\frac{1}{2})^M < \frac{1}{M+1}$ for $M > 1$, MBA is a significantly more challenging metric.

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

We evaluate both (1) multimodal video text generation models, including GPT-4o, o1, o4-mini (OpenAI, 2024), Gemini-2.5-Pro (Gemini Team, 2024), Claude-3.5-Sonnet (Anthropic, 2024), Qwen2.5VL, Qwen2VL (Wang et al., 2024), LLaVA-OneVision (Li et al., 2024a), LLaVA-Next-Video (Zhang et al., 2024b), MA-LMM (He et al., 2024a), VideoLLaVA (Lin et al., 2023), InternLM-Xcomposer-2.5 (Zhang et al., 2024a), Matryoshka Multimodal Models (Cai et al., 2024), and (2) multimodal video embedding models, including XCLIP (Ni et al., 2022), ImageBind (Girdhar et al., 2023), and LanguageBind (Zhu et al., 2024). We exponentially increase the number of frames to study its effect on video understanding. *TemporalBench* takes comparable inference com-

324
 325
 326
 327
 328
 329
 330
 331
 332
 333
 334
 335
 336
 337
 338
 339
 340
 341
 342
 343
 344
 345
 346
 347
 348
 349
 350
 351
 352
 353
 354
 355
 356
 357
 358
 359
 360
 361
 362
 363
 364
 365
 366
 367
 368
 369
 370
 371
 372
 373
 374
 375
 376
 377
 Table 3: *TemporalBench* performance of various mmodels 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	-	11.1	13.8	6.2	12.1	5.6	11.2	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										
Gemini-2.5-Pro	1FPS	81.1	78.0	68.7	81.9	81.5	78.8	81.5	78.7	43.6
OpenAI o1	32	82.0	75.9	70.2	80.6	79.7	80.0	82.1	78.6	43.0
OpenAI o4-mini	32	78.7	74.5	71.0	78.7	79.6	78.4	81.7	77.5	40.7
Gemini-2.5-Flash	1FPS	78.4	71.8	64.0	78.4	79.2	78.0	78.6	75.5	36.7
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
Qwen2.5-VL-7B	1 FPS	71.3	65.7	56.0	71.0	69.1	68.0	70.0	67.2	26.1
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-NeXT-Video-34B	32	30.6	26.8	10.4	24.8	18.0	25.2	17.3	64.0	22.0
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

pute as existing video understanding benchmarks such as NextQA Xiao et al. (2021). Details are shown in the Supp.

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 FlanT5 (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.

4.3 FINE-GRAINED VIDEO QUESTION ANSWERING ON SHORT VIDEOS

The results for multimodal generative models and embedding models are shown in Table 3. 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 the Supp. Several interesting findings arise:

The performance of any video model is far from human performance. As shown in Table 3, humans show an average performance of 67.9% on MBA, which is significantly higher than the best models, Gemini 2.5 Pro, by $\sim 24\%$. Therefore, there is a large gap between models’ 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 videos like FineGym.

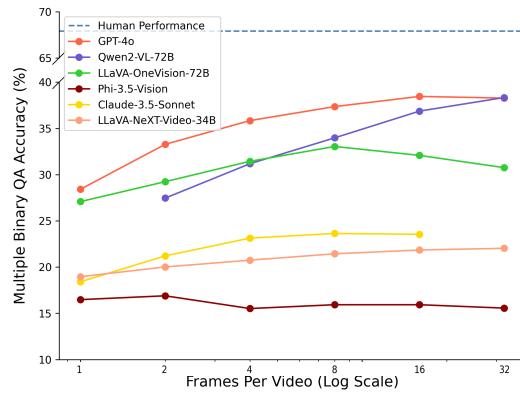


Figure 4: **Model performance on TemporalBench** with varying numbers of frames for short video understanding.

378 **Models show limited performance gains with more frames.** Shown in Figure 4, with more frames,
 379 multimodal video models usually show better performance. However, performance generally sat-
 380 urates around 8-16 frames, meaning that models struggle to improve fine-grained activity under-
 381 standing even with more frames. This is a clear contrast with human performance, showing that there is
 382 a large space for multimodal video models to improve.

383 **Multiple Binary QA is a more challenging metric.** Multiple Binary QA, as proposed in Sec-
 384 tion 3.3, prevents a model from exploiting cues in the answer choices, and evaluates whether a
 385 model truly understands the temporal dynamics in the video by splitting a single $M + 1$ -way multi-
 386 ple choice question into M binary choice questions. For example, Gemini-2.5-Pro receives 78.7%
 387 accuracy but only 43.6% on multiple binary accuracy, showing a huge gap. These results indicate
 388 that understanding the fine-grained temporal dynamics is still a challenging task for current propri-
 389 etary models and open-sourced models.

390 **Video embedding models show near chance performance.** All multimodal video embedding
 391 models, including XCLIP, LanguageBind, and ImageBind show near random chance performance.
 392 One reason could be that their small embedding size (typically a vector of size 768-2048) is insuffi-
 393 cient to capture fine-grained temporal details.

395 4.4 LONG VIDEO UNDERSTANDING

396 Since our benchmark is annotated at the video clip level, we can easily extend it to long video un-
 397 derstanding by concatenating the captions of different video clips within the same original video. In
 398 our study, we choose video datasets from AcitivityNet, Charades, EgoExo4D, COIN and FineGym.
 399 We randomly sample video clips within the same original video, and then crop a new video segment
 400 whose starting time corresponds to that of the earliest sampled video clip and whose ending time
 401 corresponds to that of the latest sampled video clip. We then concatenate all the sampled video cap-
 402 tions together to form a single long detailed description corresponding to the new video segment.
 403 Given this positive caption, we generate negative captions for it by replacing the positive caption of
 404 one of the sampled video clips with its negatives. The model is then tasked to choose the correct long
 405 caption out of multiple choices. We control the random chance multiple binary QA performance to
 406 be $\sim 9.5\%$, resulting in an apple-to-apple comparison with in Sec 4.3. In this way, we investigate
 407 whether multimodal video models can understand and distinguish fine details in a long video. Note
 408 that each caption for the short video is highly dense and informative, thus can be uniquely matched
 409 to its corresponding clip in the long video. Finally, we sampled 1,574 videos with durations ranging
 410 between $(0, 20]$ minutes, shown in the Supp.

411 We show in Figure 6 and table in the Supp., that all multimodal video models show a significant
 412 performance drop for this task compared to short video understanding. This is also reflected in all
 413 models performing better on relatively shorter videos (e.g., Charades) compared to longer videos
 414 (e.g., FineGym). These results indicate that finding the subtle temporal dynamic differences in
 415 a long video is indeed an extremely difficult task. It is similar in nature to the needle-in-the-sea
 416 task (Kamradt, 2023) in NLP except in the temporal domain. We hope that *TemporalBench* for
 417 long video understanding can serve as a very challenging task for future video understanding model
 418 development.

420 5 IN-DEPTH ANALYSIS

421 **Why multiple binary QA instead of multi-
 422 choice QA?** As discussed in Section 3.3, in
 423 the standard multi-choice QA setting, if neg-
 424 atives are all slightly variations of the pos-
 425 itive caption, we find that LLMs can determine
 426 the “centralized” caption, and take a shortcut
 427 to achieve better performance. To demon-
 428 strate this, based on one negative caption $N(C)$ in *TemporalBench*, we intentionally generate two negative
 429 captions derived from $N(C)$ (instead of C), resulting in $N_1(N(C))$ and $N_2(N(C))$. Given two set
 430 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 Fig-
 431 ure 3, text-only GPT-4o displays different behaviors. As shown in Table 4, under the intentionally

422 Table 4: Effect of the “Centralized” Caption on
 423 text-only GPT-4o.

Percentage of Predictions Aligned with \rightarrow	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

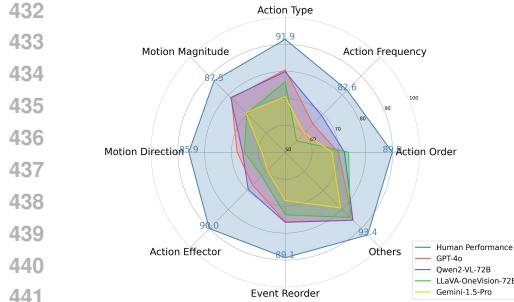


Figure 5: Per-category binary QA accuracy (BA) of various video models on short video QA.

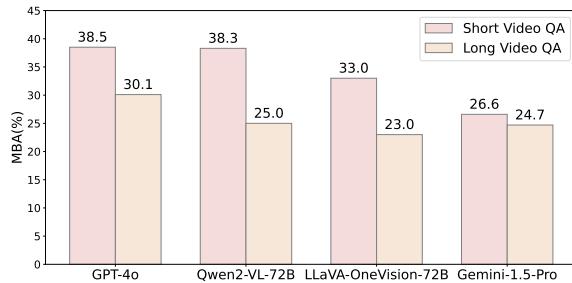


Figure 6: Models’ MBA performance on short and long video QA.

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.

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, where the statistics is shown in Table 5 in the appendix. Results in Fig. 5 indicate that multimodal video models show 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 shows that they do not memorize repeated occurrences well. Results indicate that emphasizing on fine-grained temporal aspects is critical for the development of multimodal video models, especially from the training data perspective. The visualizations of failure cases in GPT-4o are shown in Fig. 7.

6 CONCLUSION AND FUTURE WORK

We propose *TemporalBench*, a novel video understanding benchmark, to evaluate fine-grained temporal understanding abilities of multimodal video models. The video captions in our benchmark are significantly denser than those in 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 this 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.

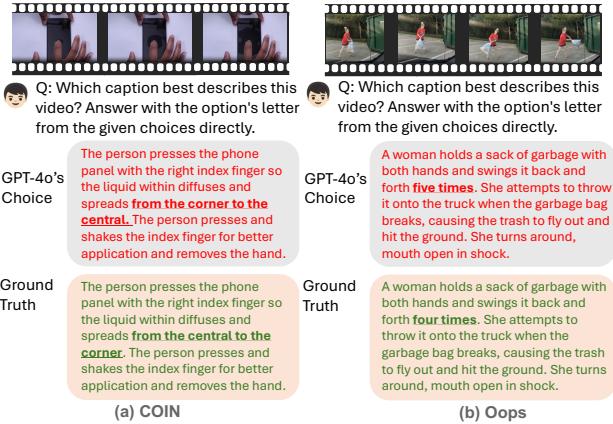


Figure 7: **GPT-4o failure cases in *TemporalBench***: not understanding fine-grained details like motion direction, action frequency, etc..

486 REPRODUCIBILITY STATEMENT
487488 We visualize part of the dataset in the submission’s supplementary materials. We will also publicly
489 release it along with the code used to evaluate the LMMs upon the paper’s acceptance.
490491 ETHICS STATEMENT
492493 This research primarily utilizes publicly available video datasets, which have been collected and
494 annotated by qualified annotators and authors, ensuring compliance with ethical standards. We have
495 made every effort to ensure that the data used respects privacy and contains no personally identifi-
496 able information. Furthermore, we acknowledge the potential implications of fine-grained video
497 understanding, especially in sensitive applications such as surveillance and autonomous systems. As
498 such, we advocate for responsible and ethical use of this research, urging caution in deploying these
499 models in real-world scenarios to avoid harmful or unintended consequences.
500501 REFERENCES
502503 Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany
504 Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical re-
505 port: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*,
506 2024.507 Anthropic. Claude-sonnet-3.5, 2024.
508509 Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zit-
510 nick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international*
511 *conference on computer vision*, pp. 2425–2433, 2015.512 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang
513 Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, local-
514 ization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.515 Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust
516 video-language alignment via contrast captions. In *Proceedings of the IEEE/CVF Conference on*
517 *Computer Vision and Pattern Recognition (CVPR)*, pp. 13927–13937, June 2024.519 Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *arXiv*
520 *preprint arXiv:2405.17430*, 2024.
521522 David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In
523 *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-*
524 *2011)*, Portland, OR, June 2011.525 Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong
526 Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and genera-
527 tion with better captions. *arXiv preprint arXiv:2406.04325*, 2024.528 Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng,
529 Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An
530 open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
531533 Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano.
534 Tvbench: Redesigning video-language evaluation, 2024.535 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
536 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
537 *arXiv preprint arXiv:2407.21783*, 2024.539 Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video.
CVPR, 2020.

540 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.

541

542

543

544 Chaoyou Fu, Yuhua 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>.

545

546

547

548

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

550

551

552 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.

553

554

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

556

557 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.

558

559 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.

560

561

562

563

564 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.

565

566

567

568 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.

569

570

571

572 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>.

573

574

575

576

577 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.

578

579

580 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.

581

582

583

584 Gregory Kamradt. Needle in a haystack - pressure testing llms, 2023. Accessed: 2024-10-01.

585

586 Ilker Keser, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, et al. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. *ICLR*, 2024.

587

588

589 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.

590

591

592 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.

593

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

597

598 Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Bench-
 599 marking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*,
 600 2023a.

601 Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin
 602 Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model.
 603 *arXiv preprint arXiv:2410.05993*, 2024b.

604

605 Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen,
 606 Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video under-
 607 standing benchmark, 2024c. URL <https://arxiv.org/abs/2311.17005>.

608

609 Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou.
 610 Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In
 611 *European Conference on Computer Vision*, pp. 331–348. Springer, 2024d.

612

613 Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating ob-
 614 ject hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika
 615 Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language
 616 Processing*, pp. 292–305, Singapore, December 2023b. Association for Computational Linguis-
 617 tics. doi: 10.18653/v1/2023.emnlp-main.20. URL <https://aclanthology.org/2023.emnlp-main.20>.

618

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

621

622 Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization
 623 Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguis-
 624 tics. URL <https://aclanthology.org/W04-1013>.

625

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

630

631 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*,
 632 2023a.

633

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

636

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

639

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

643

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

646

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

648 Kartikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic bench-
 649 mark for very long-form video language understanding. In *Adv. Neural Inform. Process. Syst.*,
 650 2024.

651

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

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

658

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

662 OpenAI. Gpt-4v(ision) system card, 2023.

663

664 OpenAI. Gpt-4o, 2024.

665

666 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic
 667 evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.),
 668 *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.
 669 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguis-
 670 tics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>.

671 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
 672 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
 673 models from natural language supervision. In *International conference on machine learning*, pp.
 674 8748–8763. PMLR, 2021.

675

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

679

680 Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie de-
 681 scription. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*
 682 (*CVPR*), 2015.

683

684 Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Ar-
 685 man Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation
 686 models. *arXiv preprint arXiv:2410.23266*, 2024.

687

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

691

692 Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun
 693 Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo
 694 J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra.
 695 Longyu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv*
 696 *preprint arXiv:2410.17434*, 2024.

697

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

701

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 rlfh. *arXiv preprint arXiv:2309.14525*, 2023.

702 Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer,
 703 Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. In *Proceedings*
 704 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13581–13591,
 705 2024.

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

709

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

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

717

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

721 Wenhao Wu. Freeva: Offline mllm as training-free video assistant. *arXiv preprint arXiv:2405.07798*,
 722 2024.

723

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

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

730

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

734

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

740

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

744

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

749

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

753

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

757

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

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

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

764

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

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

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

778

779 Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video
 780 instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024c.

781 Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web
 782 instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32,
 783 2018.

784

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

789

790 A MORE VISUALIZATIONS OF OUR BENCHMARK

791

792 In this section, we present comprehensive visualizations of our fine-grained annotations with both
 793 positive and negative descriptions. For each benchmark mentioned in Table 1 in the main paper,
 794 we provide one video example with its positive annotation and one of the corresponding negative
 795 descriptions (there are more than one negative for a single video in our dataset) in Figures 8 & 9.
 796 The video examples (a - g) are displayed in the same order as their sources in Table 1 in the main
 797 paper (7 in total).

798 799 Table 5: *TemporalBench* statistics on negative caption types.

800

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

803

804 B MORE TASKS THAT CAN BE SUPPORTED BY *TemporalBench*

805

806 B.1 VIDEO CAPTIONING

807

808 Our detailed video captions also enables analyzing a model’s fine-grained video captioning capabili-
 809 ties. For this, we prompt multimodal video models to generate a caption for an input video, with

810

811

812

813

814

815

816

817

818

819

Positive (a)



Holding a hose in their left hand, a person is gently spraying 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 .

820

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.

821

822

823

824

825

826

827

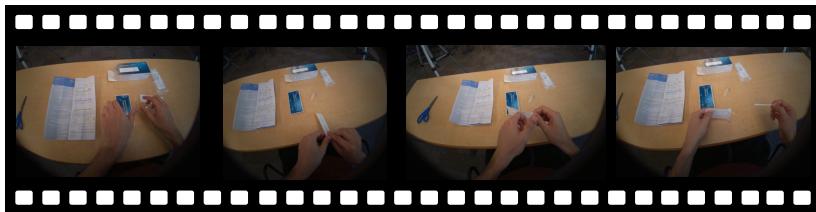
828

829

830

831

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.

832

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.

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

Figure 8: Visualizations (I) of our fine-grained annotations of the videos with both positive and negative descriptions.

852

853

854

855

856

Table 6: Comparison of video understanding benchmarks upon annotation type and density.

857

858

859

860

861

862

863

Benchmark	Annotation Type	Annotators	Fine-Grained	# Words/Sec*
TempCompass	Caption	LLM+Human	Some	2.12
YouCook2	Instructions	Human	Procedural	0.44
VITATECS	Caption	LLM+Human	Yes	2.48
TOMATO	QA Pairs	Human	Some	1.67
TVBench	QA Pairs	Parse+Human	Some	0.07
TemporalBench	Captions	Human	Yes	6.27

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917



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.



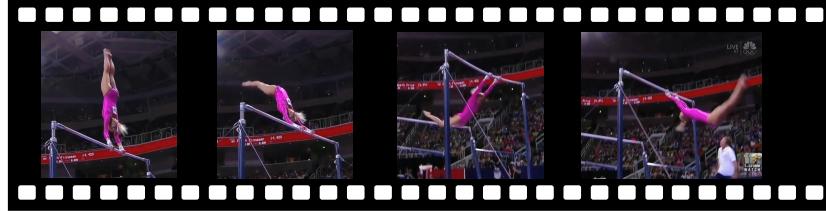
Positive Two deer come out of the trees and run along a road into the trees on the other side. A **third** deer trips as it approaches the road, then turns back around and **goes back to where it came from**.

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**.



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.



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

918
 919 Table 7: Comparison of models for video captioning using Caption Similarity, CIDEr, BLEU, and
 920 ROUGE metrics. Cosine similarity using sentence transformer reflects the captioning quality the
 921 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

941
 942 3 captioning examples in the prompt as guidance to mimic the style of our detailed video captions.
 943 Note that we remove the FineGym captions due to its different structure compared to other video
 944 captions, resulting in 1891 samples. We evaluate the resulting video captioning performance using
 945 classical image captioning metrics, CIDEr (Vedantam et al., 2015), BLEU (Papineni et al., 2002)
 946 at different n-gram levels, ROUGE (Lin, 2004), as well as the embedding similarity with sentence
 947 transformer (Reimers & Gurevych, 2019) between the ground truth caption and the generated cap-
 948 tion. Note that we for each model, we use the same number of frames as in Section 4.3 in the main
 949 paper.

950 Results in Table 7 show that GPT-4o achieves the best performance. Interestingly, the results indicate
 951 that the embedding similarity aligns most closely with the video QA task results from Sec 4.3 in the
 952 main paper. Other classical captioning metrics show inconsistent results. For example, GPT-4o
 953 obtains similar performance with one compared to 64 frames on both CIDEr and BLEU scores (e.g.,
 954 for BLEU_1 24.1 vs. 25.1). On the other hand, all models show similar ROUGE scores. Thus, for
 955 the zero-shot captioning task, our findings indicate that text embedding similarity may be the most
 956 reliable metric.

958 B.2 FUTURE TASKS

960 Due to the detailed annotation of *TemporalBench*, more tasks can be supported.

961

- 962 • **Video grounding from detailed text descriptions.** Since the video clips are cropped
 963 from the source video, with the documented starting and ending time, our benchmark can
 964 serve as a fine-grained moment localizing benchmark from text descriptions. This is dif-
 965 ferent from existing video grounding datasets such as Charades-STA (Gao et al., 2017),
 966 COIN (Tang et al., 2019), Ego4D (Grauman et al., 2024) where the text descriptions are
 967 usually very short, possibly resulting in low temporal localization performance due to the
 968 vague and coarse descriptions.
- 969 • **Text-to-Video (T2V) generation with detailed prompts.** Given our highly detailed de-
 970 scription, a T2V generation model can be evaluated by verifying if the generated videos
 971 reflect the fine-grained action details.

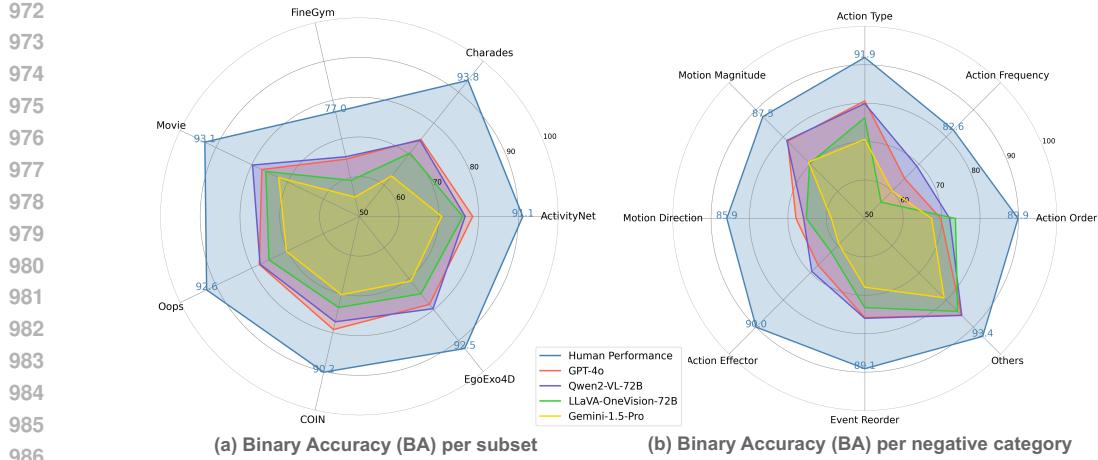


Figure 10: Visualization of binary accuracy for short video QA per (a) subset and (b) negative type. Human performance is much better than GPT-4o, Qwen2-VL-72B, LLaVA-OneVision-72B, and Gemini-1.5-Pro.

C EXTENDED COMPARISON OF NEGATIVE CAPTIONS

Here we show the extended comparison for the negative captions generated from the original captions and our detailed captions in *TemporalBench* in Figure 11. With fine-grained details, the negatives are more difficult and temporal centric.

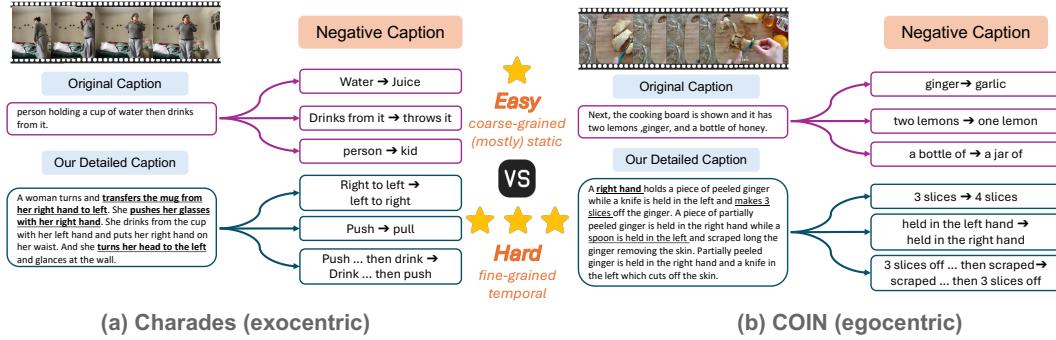


Figure 11: Comparison of negative captions generated from the original captions and our detailed captions in *TemporalBench*. With fine-grained details, the negatives are more difficult and temporal centric.

D *TemporalBench* PERFORMANCE UNDER EACH CATEGORY

Multimodal video models' performance under each category is shown in Table 8 and Figure 10 (b). Results indicate that multimodal video models show 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 shows that they do not memorize repeated occurrences well. Results indicate that emphasizing on fine-grained temporal aspects is critical for the development of multimodal video models, especially from the training data perspective. The visualizations of failure cases in GPT-4o are shown in Figure 12.

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042



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 into 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 into 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 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**.

(c) FineGym

(d) EgoExo4D

Figure 12: **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.

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

Table 8: *TemporalBench* short QA 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 Language 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

E PER SUBSET RESULTS FOR SHORT AND LONG VIDEO QA

The per subset results (denoted as “T-”) for short and long video QA under Binary Accuracy (BA) are shown in Table 11, and Table 13, respectively. We also provide a visualization of per subset short video QA performance in Figure 10 (a). The per subset results (denoted as “T-”) for long video QA under Multiple Binary Accuracy (MBA) is shown in Table 12. Still, human achieves much better performance than all multimodal videos. Interestingly, both human and AI models show worse performance on Finegym, the professional video understanding dataset.

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 frames that leads to the best performance. Here we extend the results to show the results of more frames in Table 9 and Table 10. Note that we include results for models including LLaVA-Video Zhang et al. (2024c), Aria Li et al. (2024b) and LongVU Shen et al. (2024). Generally, with more frames, performance 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 a large space for multimodal video models to improve.

G DATA ANNOTATION PIPELINE AND PLATFORM

Our **six-stage** annotation process includes: (1) AMT workers filtering through an exam, (2) Initial positive captions by filtered AMT workers, (3) author (experts) refined positive captions, (4) LLM-generated negatives with targeted temporal changes, (5) meticulous author filtering, (6) AMT

1134 Table 9: *TemporalBench* performance of various models under binary QA accuracy (BA) and multi-
 1135 ple binary QA accuracy (MBA) setting for short and long question answering with different number
 1136 of frames. “Overall” denotes the average performance of short and long video QA performance.
 1137 (Part 1)

Model	# Frames	Overall MBA	Overall BA	Short MBA	Short BA	Long MBA	Long BA	Captioning
Human Performance	-	-	-	67.9	89.7	-	-	-
Random Chance	-	9.5	50.0	9.5	50.0	9.5	50.0	-
XCLIP	8	12.0	51.7	12.9	51.6	11.1	51.7	-
ImageBind	2	12.4	52.0	14.0	53.0	10.7	51.0	-
LanguageBind	8	13.3	52.2	14.5	52.8	12.0	51.6	-
GPT-4o	64	35.4	73.3	38.0	76.0	32.7	70.5	63.5
	32	32.9	71.5	38.3	75.9	27.4	67.0	63.2
	16	34.3	72.8	38.5	75.7	30.1	69.8	61.3
	8	32.9	72.0	37.4	75.1	28.3	68.8	60.3
	4	31.9	71.2	35.8	74.4	28.0	68.0	58.8
	2	30.3	70.3	33.3	72.7	27.3	67.8	55.3
	1	26.5	67.4	28.4	70.0	24.5	64.7	52.3
	0	27.4	67.7	26.5	67.7	28.2	67.6	-
Gemini-1.5-Pro	1FPS	25.7	66.4	26.6	67.5	24.7	65.2	56.5
	0	18.7	60.2	16.1	58.1	21.2	62.2	-
Claude-3.5-Sonnet	16	23.2	64.2	23.5	65.9	22.9	62.4	54.1
	8	24.1	65.1	23.6	65.5	24.5	64.6	53.1
	4	23.2	64.2	23.1	64.8	23.3	63.6	51.9
	2	21.1	62.2	21.2	61.9	20.9	62.4	48.2
	1	18.7	58.9	18.4	58.5	18.9	59.3	41.0
Qwen2-VL-72B	32	31.7	70.2	38.3	75.8	25.0	64.5	56.1
	16	31.5	70.1	36.9	74.6	26.1	65.5	54.1
	8	30.1	68.9	34.0	73.1	26.2	64.7	51.4
	4	28.6	68.4	31.2	71.5	26.0	65.3	48.3
	2	27.3	67.6	27.5	69.2	27.1	66.0	43.9
Qwen2-VL-7B	32	21.8	62.1	24.7	64.4	18.8	59.7	51.9
	16	21.2	61.5	23.6	63.3	18.7	59.7	50.3
	8	19.2	59.4	21.1	61.1	17.2	57.7	48.4
	4	17.4	58.5	19.3	59.5	15.4	57.5	46.1
	2	16.4	56.9	17.7	57.8	15.0	56.0	42.0
LLaVA-OneVision-72B	32	26.6	66.6	30.7	70.5	22.4	62.7	53.9
	16	27.2	67.3	32.1	71.2	22.3	63.4	54.2
	8	28.1	67.9	33.0	72.1	23.1	63.6	55.0
	4	27.6	67.3	31.4	71.2	23.8	63.4	54.2
	2	25.7	66.3	29.2	69.6	22.1	63.0	51.1
	1	23.3	64.1	27.1	67.9	19.5	60.2	48.6
LLaVA-OneVision-7B	32	18.7	59.4	21.2	61.9	16.2	56.9	50.1
	16	17.9	58.8	20.1	60.9	15.6	56.6	50.4
	8	17.3	57.8	19.5	59.9	15.0	55.7	50.2
	4	16.4	56.3	18.9	58.9	13.9	53.7	49.7
	2	14.8	54.4	16.8	56.1	12.7	52.7	47.1
	1	12.0	51.4	13.4	53.3	10.6	49.5	44.1
LLaVA-NeXT-Video-34B	32	19.9	61.1	22.0	64.0	17.7	58.2	53.1
	16	20.3	61.1	21.8	63.7	18.7	58.4	53.3
	8	20.5	61.9	21.4	63.3	19.5	60.4	53.4
	4	20.4	61.7	20.7	63.0	20.0	60.3	52.5
	2	19.9	61.2	20.0	61.8	19.7	60.6	48.9
	1	19.0	59.8	19.0	60.5	18.9	59.1	46.2

1173
 1174 worker/expert performance as validation. This ensures high-quality, challenging negatives. Specifi-
 1175 cally,

1180 **Positive Captions** We use Amazon Mechanical Turk (AMT) with in-context examples for positive
 1181 caption annotation, shown in Figure 13. Then we use Label Studio to let authors refine the caption.
 1182 As shown in Figure 14, authors can edit the caption from AMT workers. Also, we provide the
 1183 original short video captions to let people better understand our task.

1184
 1185 **Negative Captions Filtering Platform** We first prompt LLMs (GPT-4o, Gemini, and Llama-3.1-
 1186 405b) to get initial negative captions, and then ask authors to choose the negatives that can reflect
 1187 the temporal dynamic. The visualization of the multi-choice platform is shown in Figure 15.

1188 Table 10: *TemporalBench* performance of various models under binary QA accuracy (BA) and multi-
 1189 ple binary QA accuracy (MBA) setting for short and long question answering with different number
 1190 of frames. “Overall” denotes the average performance of short and long video QA performance.
 1191 (Part 2)

1192	Model	# Frames	Overall MBA	Overall BA	Short MBA	Short BA	Long MBA	Long BA	Captioning
1193	LLaVA-NeXT-Video-7B	32	15.9	57.1	17.3	59.5	14.5	54.7	51.6
		16	19.3	59.9	22.4	64.0	16.1	55.7	49.9
		8	20.5	61.2	23.6	65.1	17.3	57.2	50.1
		4	20.0	60.7	23.0	64.2	17.0	57.2	49.2
		2	19.2	60.3	21.5	63.1	16.8	57.4	46.8
		1	17.6	59.1	19.1	62.0	16.1	56.1	44.0
1194	LLaVA-Video-72b	128	33.3	72.1	36.3	74.9	30.2	69.4	53.7
		64	33.8	72.1	37.1	75.3	30.5	68.9	54.2
		32	33.7	72.4	37.7	75.9	29.6	68.8	54.8
		16	33.1	72.0	37.4	75.8	28.8	68.3	54.9
		8	32.7	71.9	36.3	75.6	29.0	68.3	54.7
		4	32.5	71.4	35.9	74.9	29.2	68.0	53.1
1195	1200	2	31.3	70.4	33.0	73.4	29.7	67.5	50.2
		1	29.5	69.5	30.9	71.6	28.1	67.4	47.0
		128	21.9	62.1	20.2	61.2	23.5	63.0	51.8
		64	22.8	63.0	22.0	62.7	23.6	63.3	52.5
		32	22.9	63.6	22.9	63.3	22.9	63.9	52.1
		16	21.6	63.0	21.8	62.9	21.5	63.1	51.9
1196	1201	8	20.9	62.0	21.5	61.8	20.3	62.2	50.9
		4	18.8	60.5	18.4	59.8	19.1	61.2	49.6
		2	18.2	59.6	18.9	58.6	17.5	60.7	46.6
		1	16.0	57.0	16.2	56.7	15.9	57.4	43.1
		128	25.1	66.0	27.2	68.7	23.1	63.4	50.6
		64	24.2	65.9	26.3	68.4	22.0	63.4	51.1
1197	1202	32	25.0	65.9	26.6	68.4	23.4	63.5	51.5
		16	24.4	65.6	26.2	68.3	22.6	62.9	51.4
		8	24.4	65.4	26.5	68.5	22.3	62.3	51.7
		4	23.5	64.8	26.0	67.8	21.0	61.7	50.8
		2	22.3	62.2	25.1	66.5	19.6	57.8	48.7
		1	17.9	54.8	22.2	64.5	13.6	45.1	45.7
1198	LongVU	1FPS	18.9	58.5	20.9	61.7	16.9	55.3	40.5
1199	InternLM-XC2.5	1FPS	16.8	57.3	17.9	58.8	15.6	55.8	52.4
1200	VideoLLaVA	8	20.3	61.6	25.5	67.1	15.1	56.0	46.0
1201	MiniCPM-V2.6	1FPS	20.4	61.3	21.4	62.3	19.3	60.3	47.2
1202	Phi-3.5-Vision	32	14.1	54.3	15.6	56.8	12.6	51.7	48.4
		16	14.7	55.1	15.9	57.2	13.5	53.0	48.9
		8	14.9	55.6	15.9	57.4	13.8	53.7	48.3
		4	15.0	56.0	15.5	57.5	14.5	54.5	44.0
		2	15.5	56.2	16.9	58.0	14.1	54.4	42.9
		1	14.5	55.9	16.5	57.7	12.5	54.0	42.1
1203	MA-LMM	4	9.1	47.4	9.2	48.0	9.0	46.9	38.7
1204	M3	6	13.3	54.7	14.8	56.4	11.8	53.1	47.8
1205	LLaVA-1.5-13B	1	13.7	55.1	13.1	55.7	14.2	54.5	47.9
1206	LLaVA-1.5-7B	1	15.3	56.8	18.3	60.5	12.3	53.2	45.7
1207	LLaVA-NeXT-34B	1	19.0	60.5	18.0	60.5	19.9	60.5	49.1
1208	Phi-3-Vision	1	15.4	55.2	15.1	54.4	15.6	56.0	42.0
1209	Gemini-1.5-Pro	0	18.6	60.1	16.1	58.1	21.2	62.2	-
1210	Yi-34B	0	18.5	59.7	18.7	59.9	18.4	59.5	-
1211	Vicuna7b-1-5	0	10.1	50.8	10.4	50.5	9.9	51.1	-
1212	Flan-T5-XL	0	18.6	59.0	17.9	57.9	19.4	60.1	-
1213	Flan-T5-XXL	0	15.9	56.0	15.1	55.1	16.7	56.9	-

H VIDEO STATISTICS

1234 Video length distribution of (a) short video clips and (b) long videos in *TemporalBench* is shown in
 1235 Figure 16. For benchmarks without caption, we computed the number of words per second using
 1236 the length of question and answers.
 1237

1238 For ~10K short video QA, Qwen2VL-7B (16 frames) requires 34 minutes for MBA evaluation on 4
 1239 A6000 GPUs, comparable to existing benchmarks (e.g., 32 minutes for NextQA Xiao et al. (2021)
 1240 which owns multichoice QA format).

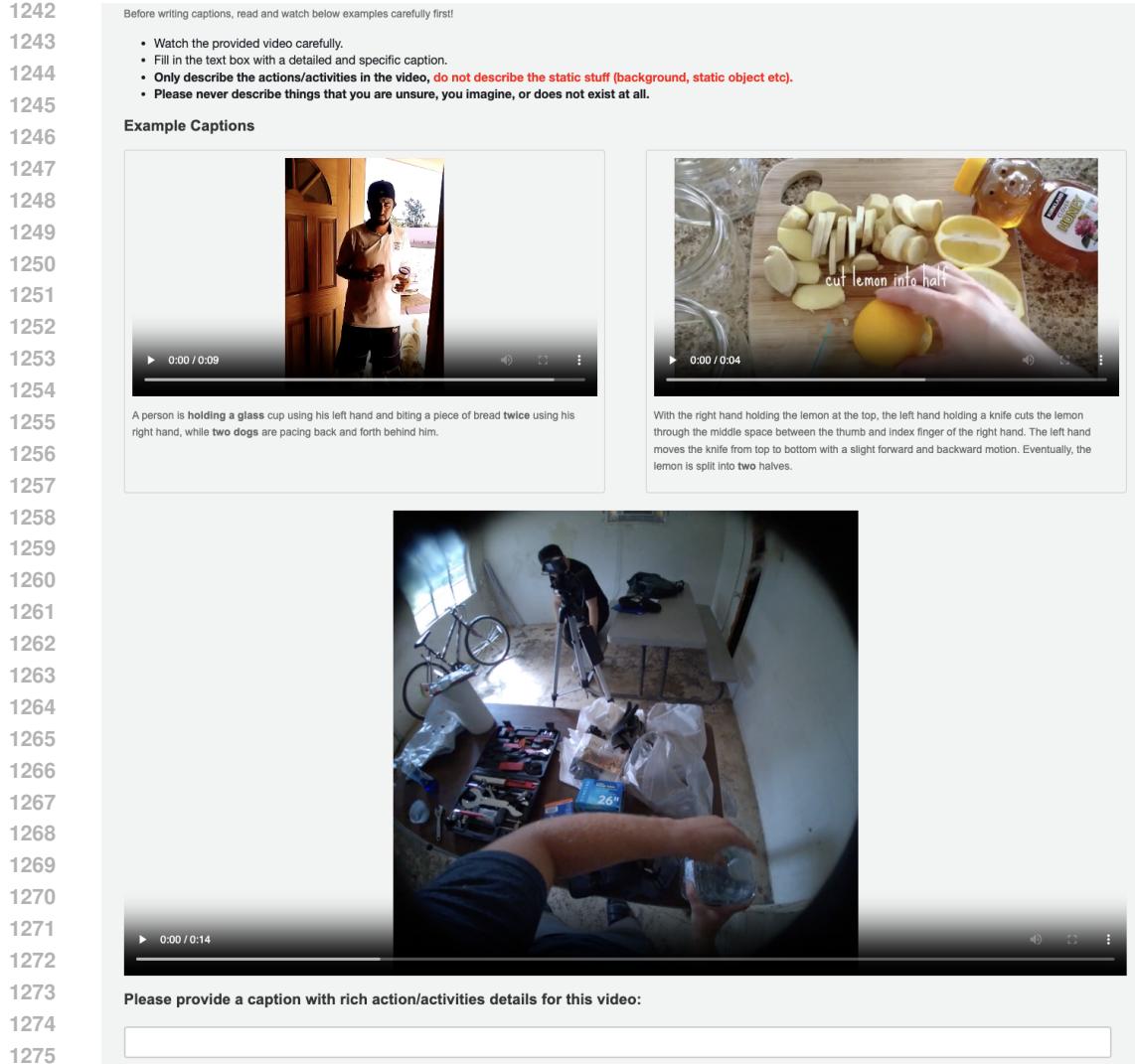


Figure 13: Positive caption annotation platform for AMT workers.

In *TemporalBench*, each sub-question is answered with high human accuracy (e.g., $\sim 92\%$ BA on most categories), but the *product* of probabilities will lead to an overall lower success rate ($92\%^5 = 65.9\%$). SOTA models (GPT-4o) still lag significantly (38.5% on short video MBA), highlighting *TemporalBench*'s difficulty.

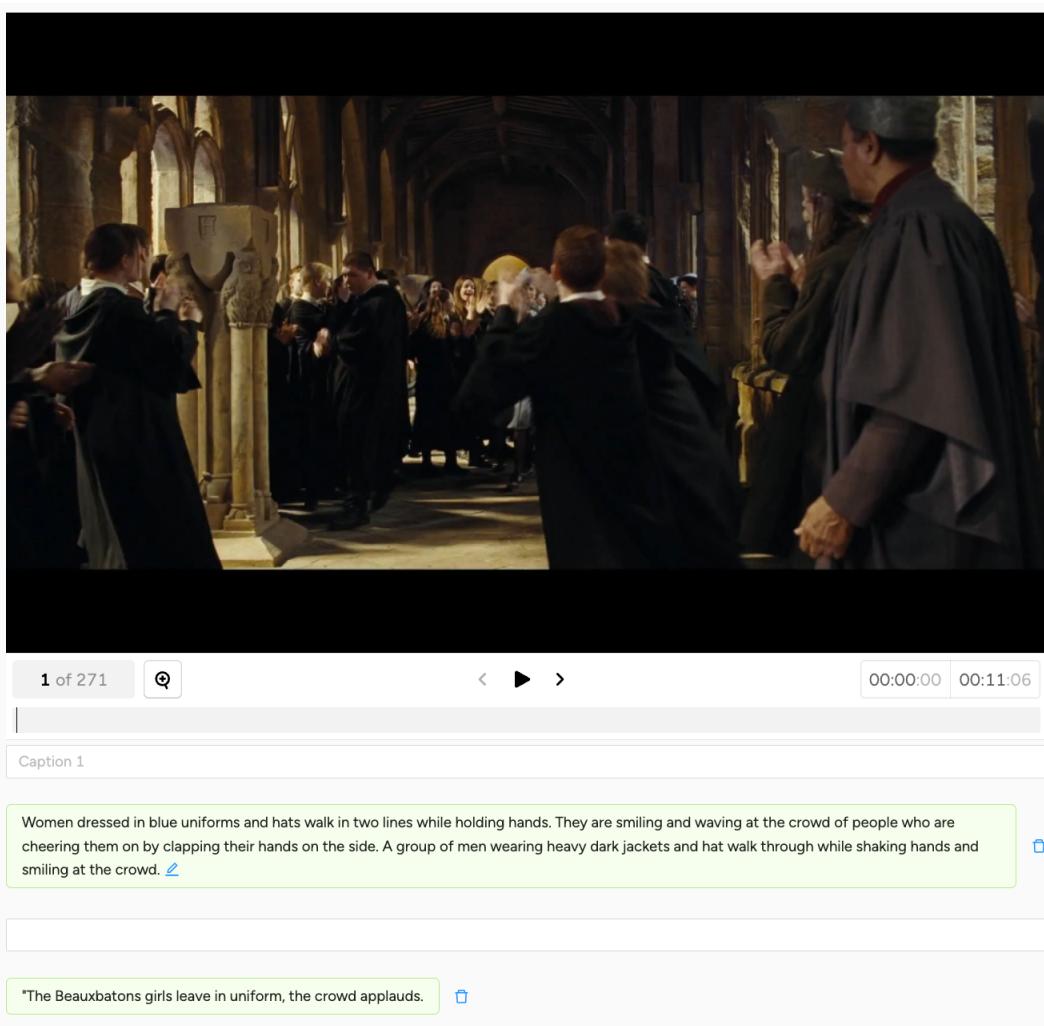
I MORE RELATED WORKS

Existing works explored video understanding along different directions. YouCook2 Zhou et al. (2018) emphasizes procedure videos but lacks fine-grained temporal understanding. VITATECS Li et al. (2024d), TempCompass Liu et al. (2024c), TOMATO Shangguan et al. (2024) and TVBench Cores et al. (2024) work towards better model temporal dynamics via the counterfactual manner, but still lacks the dense captions for fine-grained details.

J PROMPTS FOR NEGATIVE CAPTION CURATION

We provide the prompts for the negative caption curation in Table 14 and Table 15 for the word level replacement and event level re-ordering for the negative caption curation process, respectively.

1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316



1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331

Figure 14: Positive caption refinement platform.

1332 We leverage such prompts to gather responses from LLMs including including GPT-4o, Llama-3.1-
1333 405B, and Gemini-1.5-Pro, followed by human filtering.

1334

K LIMITATIONS

1335

1336 One cannot fully analyze the behavior of proprietary models included in this paper due to the lack
1337 of access to these models, which are GPT-4o, Gemini-1.5-Pro and Claude 3.5 Sonnet.

1338

L BROADER IMPACT

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

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.

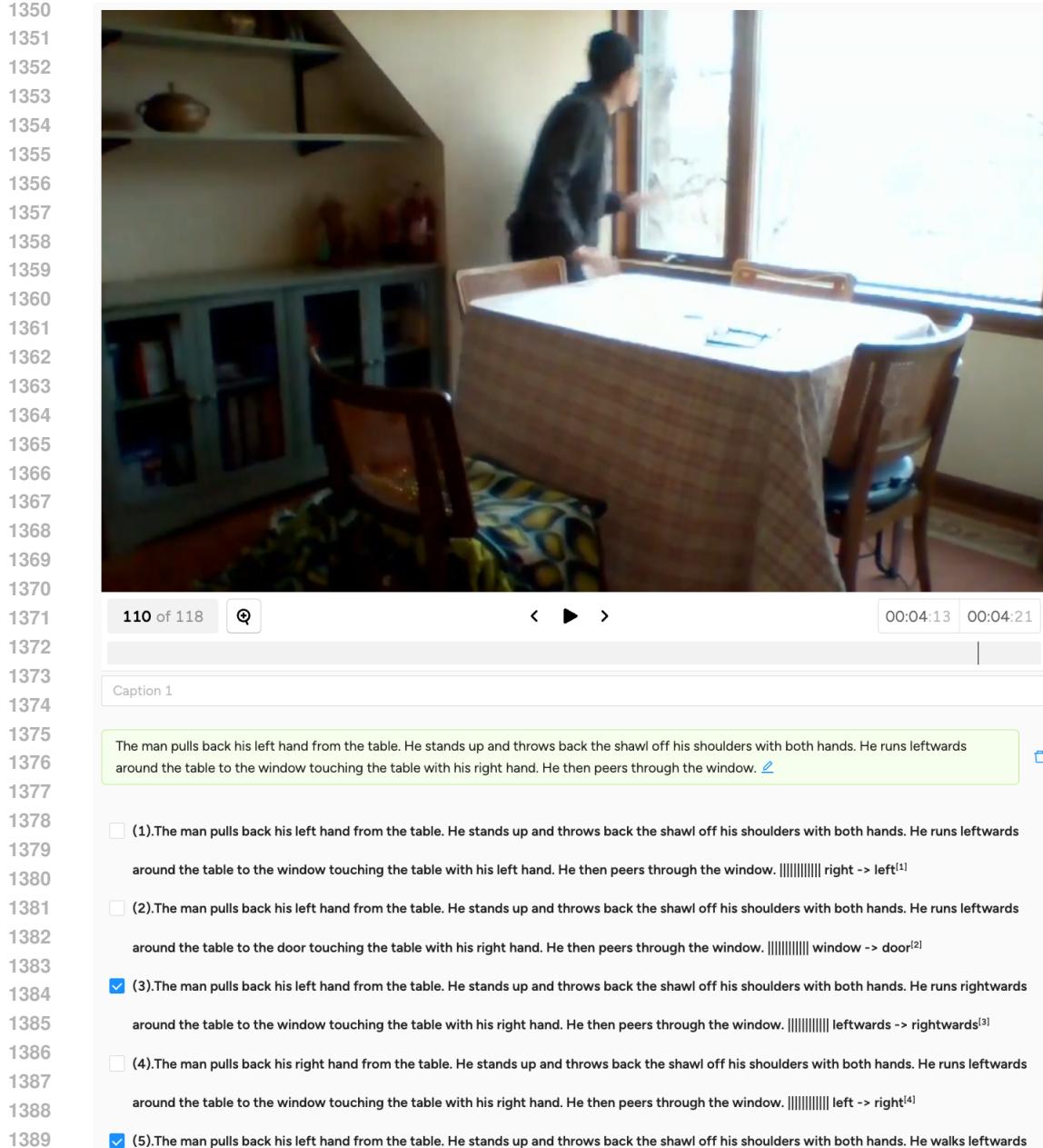


Figure 15: Negative caption annotation platform.

M THE USE OF LLMs IN THIS RESEARCH

LLMs including Gemini and Claude are to help polish this paper, LLMs are also used during the negative caption generation process.

N TOWARDS IMPROVING TEMPORAL FINE-GRAINED CAPABILITY

Here we explore enhancing video LLM's temporal fine-grained capability, where we focus on curating high-quality temporal negatives. By leveraging LLaVA-Video's dense captions and Gemini 2.5 Pro, we curate 5k high quality captioning binary QA problems. Finetuning upon Qwen-2.5-VL-

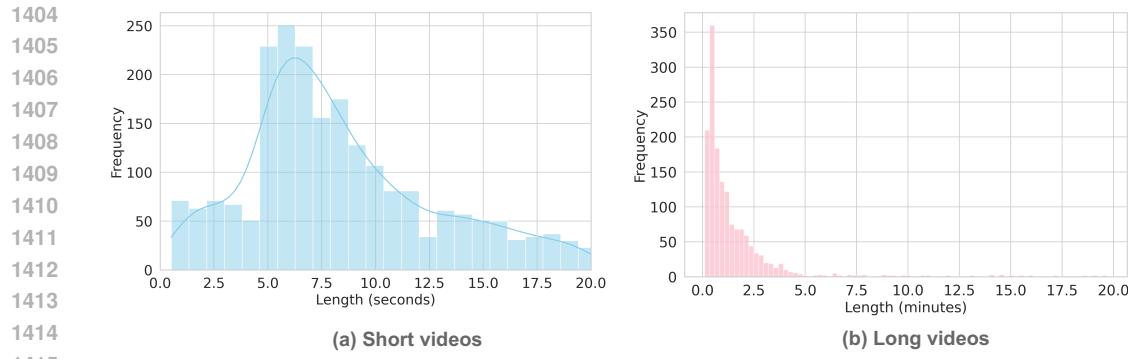


Figure 16: Video length distribution of (a) short video clips and (b) long videos in *TemporalBench*.

Instruct-7B leads to 3.5% improvements on the MBA score, which shows the positive signal towards our goal.

1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

1458
 1459
 1460
 1461
 1462
 1463
 1464
 1465
 1466
 1467
 1468
 1469

1470
 1471 Table 11: *TemporalBench* performance of various multimodal generative models and embedding
 1472 models under the binary QA accuracy (BA) and multiple binary QA settings (MBA) for **short**
 1473 **videos**. The prefix “T-” indicates BA performance for the annotated subset in our *TemporalBench*.
 1474 We show the result with the best average MBA performance for each model with respect to the
 1475 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

1500
 1501
 1502
 1503
 1504
 1505
 1506
 1507
 1508
 1509
 1510
 1511

1512
 1513
 1514
 1515
 1516
 1517
 1518
 1519
 1520
 1521
 1522
 1523
 1524
 1525
 1526
 1527
 1528
 1529
 1530
 1531
 1532
 1533
 1534
 1535
 1536
 1537
 1538
 1539
 1540
 1541
 1542
 1543
 1544
 1545
 1546
 1547
 1548
 1549
 1550
 1551
 1552
 1553
 1554
 1555
 1556
 1557
 1558
 1559
 1560
 1561
 1562
 1563
 1564
 1565

Table 12: *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 Languange 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

1566
 1567
 1568
 1569
 1570
 1571
 1572
 1573
 1574
 1575
 1576
 1577
 1578
 1579
 1580
 1581

Table 13: *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 Languange 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

1610
 1611
 1612
 1613
 1614
 1615
 1616
 1617
 1618
 1619

1620
 1621
 1622
 1623
 1624
 1625
 1626
 1627
 1628
 1629
 1630
 1631
 1632

Table 14: For each query, we illustrate the prompt construction process of the **word level replacement** for LLMs powered negative caption curation including GPT-4o, Llama-3.1-405B, and Gemini-1.5-Pro, to collect `query['response']` from `query['context']`, using few-shot in-context-learning, where examples are from `fewshot_samples`, each example including input `sample['context']` and output `sample['response']`. Note that `messages` is the final prompt. In this example, we provide the prompt used to generate the negative caption response.

1633
 1634
 1635
 1636
 1637
 1638
 1639
 1640
 1641
 1642
 1643
 1644
 1645
 1646
 1647
 1648
 1649
 1650
 1651
 1652
 1653
 1654
 1655
 1656
 1657
 1658
 1659
 1660
 1661
 1662
 1663
 1664
 1665
 1666
 1667
 1668
 1669
 1670
 1671
 1672
 1673

```

messages = [ {"role": "system", "content": f"""You are a helpful assistant in generating negative captions for
videos designed to output JSON.
You are given a video caption. You are tasked to: Generate negative captions that changes the action descriptions being discussed
in exactly one of the entities by only changing or swapping one word or phrase.
For example, you can change "open the drawer" to "close the drawer", or change from "pull" to "push", or "quickly" to
"slowly". You should generate negatives that can not be solved by feeding one random frame into a multimodal model. For
example, changing "wood chair" into "plastic chair" is a bad practice. But changing "man playing with one cat" to "man playing
with two cats" is good. Swapping "left leg" to "right leg" is also good if this caption updates completely changes the motion in
the video.
All in all, the new description must meet all of these requirements:
1. The change of action descriptions must be sufficiently different to make the new description inaccurate, but it should also be
somewhat related to be challenging to an AI model.
2. Compared to the original description, the new description must differ in only one aspect. All other details must be kept the
same.
3. The new description must mimic the sentence structure of the original description.
4. The new description must be fluent, logical, and grammatically correct.
5. Pose challenging(difficult enough) negative captions so that a large multimodal text generation model should struggle to
distinguish the original caption v.s. negative captions.
Here are some examples whose output format you can have a reference:"""}]
fewshot_samples = [
  { "context": """ Original Caption: "A person is holding a glass cup using his left hand and biting a piece of bread twice
using his right hand, while two dogs are pacing back and forth behind him."
Your answer: """,
    "response": """[ "A person is holding a glass cup using his left hand and biting a piece of bread three times using his right
hand, while two dogs are pacing back and forth behind him.",
    "A person is drinking water from a glass cup using his left hand and biting a piece of bread twice using his right hand, while
two dogs are pacing back and forth behind him.",
    "A person is holding a glass cup using his left hand and biting a piece of bread twice using his right hand, while one dog is
pacing back and forth behind him.", ]"""
  },
  { "context": """ Original Caption: "With the right hand holding the lemon at the top, the left hand holding a knife cuts the
lemon through the middle space between the thumb and index finger of the right hand. The left hand moves the knife from top to
bottom with a slight forward and backward motion. Eventually, the lemon is split into two halves."
Your answer: """,
    "response": """[ "With the right hand holding the lemon at the top, the left hand holding a knife cuts the lemon through the
middle space between the thumb and index finger of the right hand. The left hand moves the knife from top to bottom with a slight
forward and backward motion. Eventually, the lemon is split into four pieces.",
    "With the right hand holding the lemon at the top, the left hand holding a knife cuts the ginger through the middle space
between the thumb and index finger of the right hand. The left hand moves the knife from top to bottom with a slight forward
and backward motion. Eventually, the ginger becomes lots of pieces.",
    "With the left hand holding the lemon at the top, the right hand holding a knife cuts the lemon through the middle space
between the thumb and index finger of the right hand. The right hand moves the knife from top to bottom with a slight forward
and backward motion. Eventually, the lemon is split into two halves.", ]"""
  }
]
for sample in fewshot_samples:
  messages.append({"role": "user", "content": sample["context"]})
  messages.append({"role": "assistant", "content": sample["response"]})
messages.append({"role": "user", "content": '\n'.join(query)})

```

1674
1675
1676
1677
1678
1679

1680 Table 15: For each query, we illustrate the prompt construction process of the **event level re-**
1681 **ordering** for LLMs powered negative caption curation including GPT-4o, Llama-3.1-405B, and
1682 Gemini-1.5-Pro, to collect `query['response']` from `query['context']`, using few-shot
1683 in-context-learning, where examples are from `fewshot_samples`, each example including input
1684 `sample['context']` and output `sample['response']`. Note that `messages` is the final
1685 prompt. In this example, we provide the prompt used to generate the negative caption response.
1686

1687
1688
1689
1690
1691
1692
1693
1694
1695
1696

```

messages = [ {"role": "system", "content": f"""
You are a helpful assistant tasked with generating negative captions for videos. Your goal is to output results in JSON format. You are provided with a video caption, and your job is to:
- Generate a "negative" version of the caption by rearranging the sequence of actions in the original caption. This rearrangement should still make logical sense in real life.

Guidelines:
1. The changed order must still be reasonable. If reversing or rearranging the order of events is not logically possible, the negative caption is not valid. For instance, if the original caption is "AAAA then BBBB" (e.g., a fire starting before people use water to extinguish it), switching to "BBBB then AAAA" would be invalid.
2. The negative caption should differ only in the order of actions or events, while keeping the content of the events intact.
3. If there is only one event or all events happen at the same time (concurrently), output an empty array `[]`.
4. The negative caption must remain fluent, logical, and grammatically correct.
5. The negative captions should be challenging enough that a large multimodal model would find it difficult to distinguish between the original and negative captions.

Here are some examples whose output format you can have a reference:"""
}]

fewshot_samples = [
  { "context": """
Original Caption: "A person is spraying cleaning liquid on the wooden chair in the external environment, first from bottom to top on the right handrail, then top to bottom on one of the supporting pillars."
Your answer: """,
    "response": """
[ "A person is spraying cleaning liquid on the wooden chair in the external environment, first from top to bottom on one of the supporting pillars, then bottom to top on the right handrail", ]"""
  },
  { "context": """
Original Caption: "A person leans back pulling the rope as the boat sails forward. In the background more boats sail on the sea. The waves in the sea ebbs and flows. There is also a text "Accelerate" appearing at the bottom of the scene."
Your answer: """,
    "response": """
[ ]"""
  },
  { "context": """
Original Caption: "A man uses his left hand to put the unscrewed cap onto the table, then uses the same hand to move a red cup closer to his right hand. He is then seen pouring something from the right hand into the left cup. He then moves the cup back onto the table and picks the cap back up to screw it back on."
Your answer: """,
    "response": """
[ "A man uses his left hand to move a red cup closer to his right hand, then uses the same hand to put the unscrewed cap onto the table. He is then seen pouring something from the right hand into the left cup. He then moves the cup back onto the table and picks the cap back up to screw it back on.", "A man is seen pouring something from the right hand into the left cup. Then he uses his left hand to put the unscrewed cap onto the table, then uses the same hand to move a red cup closer to his right hand. He then moves the cup back onto the table and picks the cap back up to screw it back on.", "A man uses his left hand to put the unscrewed cap onto the table, then seen pouring something from the right hand into the left cup. He then uses the left hand to move a red cup closer to his right hand. He then moves the cup back onto the table and picks the cap back up to screw it back on.", "A man is seen pouring something from the right hand into the left cup. Then he uses the left hand to move a red cup closer to his right hand, then uses his left hand to put the unscrewed cap onto the table. He then moves the cup back onto the table and picks the cap back up to screw it back on." ]"""
  }
]

for sample in fewshot_samples:
  messages.append({ "role": "user", "content": sample["context"]})
  messages.append({ "role": "assistant", "content": sample['response'] })
messages.append({ "role": "user", "content": '\n'.join(query) })

```

1722
1723
1724
1725
1726
1727