FAVOR-Bench: A Comprehensive Benchmark for Fine-Grained Video Motion Understanding

Chongjun Tu[1†] **Lin Zhang**[1,3†] **Pengtao Chen**[1†] **Peng Ye**[2] **Xianfang Zeng**[3♣]
Wei Cheng[3] **Gang Yu**[3] **Tao Chen**[1,4*]
[1] College of Future Information Technology, Fudan University
[2] The Chinese University of Hong Kong [3] StepFun [4]Shanghai Innovation Institute

Abstract

Multimodal Large Language Models (MLLMs) have shown impressive video content understanding capabilities but struggle with fine-grained motion comprehension. To comprehensively assess the motion understanding ability of existing MLLMs, we introduce FAVOR-Bench, which comprises 1,776 videos from both ego-centric and third-person perspectives and enables assessment through both close-ended and open-ended tasks. For close-ended evaluation, we carefully design 8,184 multiple-choice question-answer pairs spanning six distinct sub-tasks. For open-ended evaluation, we employ the GPT-assisted evaluation and develop a novel cost-efficient LLM-free assessment method, where the latter can enhance benchmarking interpretability and accessibility. Comprehensive experiments with 21 state-of-the-art MLLMs reveal significant limitations in their ability to comprehend and describe detailed temporal dynamics in video motions. To alleviate this limitation, we further build FAVOR-Train, a dataset of 17,152 videos with fine-grained motion annotations. Finetuning Qwen2.5-VL on FAVOR-Train yields consistent improvements on motion-related tasks across TVBench, MotionBench and our FAVOR-Bench. Our assessment results demonstrate that the proposed FAVOR-Bench and FAVOR-Train provide valuable tools for the community to develop more powerful video understanding models.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated remarkable video understanding capabilities [6, 56, 52]. The emergence of high-quality video-text datasets [32, 4, 46] has further promoted their development and powered various downstream applications like motion recognition, caption generation and video generation [53, 18, 26, 61, 2, 41, 57]. To effectively evaluate the capabilities of these models, kinds of benchmarks with different focuses have been developed, such as comprehensive capabilities [11, 24], long-video understanding [60, 12], and video reasoning [47, 49].

Despite these advances in video understanding benchmarks, the evaluation of fine-grained video motion understanding remains under-explored, particularly across diverse viewing perspectives and evaluation tasks. This capability is critical for fields that require precise understanding and control (such as embodied imitation learning [42, 10] and text-image to video generation (TI2V) [15, 33]). As shown in Figure 1, while MLLMs can identify overall behaviors in videos, they struggle with problems related to fine-grained motions. For open-ended description tasks, even when explicitly

♣Xianfang Zeng is the project leader.
†Equal contribution. *Corresponding author.
Work was done when interned at StepFun.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

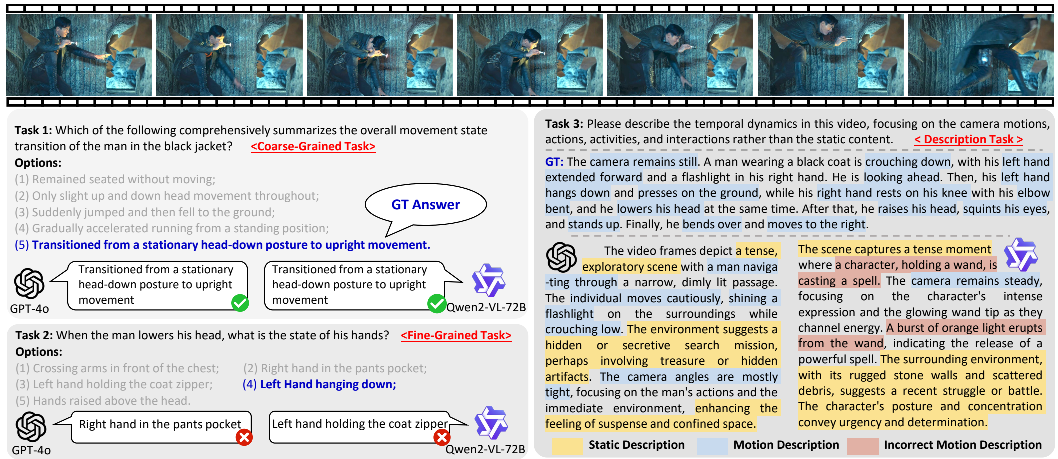

Figure 1: Illustration of motion understanding capabilities of proprietary and open-source MLLMs. Both models correctly answer the coarse-grained summarization question (Task 1), but fail to resolve the fine-grained action detail question (Task 2). For the open-ended description task (Task 3), despite being required to focus on temporal dynamics, the responses emphasize static content, and the motion descriptions are either coarse-grained or contain errors.

instructed to focus on temporal dynamics, models predominantly emphasize static content and often lack fine-grained analysis of the motions and activities. Traditional datasets like ActivityNet-QA [54] and motion-related subsets in recent benchmarks [24, 60] primarily focus on the event-level granularity. The concurrent MotionBench [14] evaluates motion-level perception through multiple-choice questions but is limited to third-person perspective videos and lacks open-ended assessment.

To mitigate these concerns, we introduce FAVOR-Bench, a comprehensive benchmark for **f**ine-gr**a**ined **v**ideo m**o**tion unde**r**standing. FAVOR-Bench spans both ego-centric and third-person perspectives with 1,776 videos from diverse fields. Furthermore, our evaluation framework encompasses close-ended and open-ended tasks to assess motion understanding capabilities thoroughly. For close-ended evaluation, we carefully curate 8,184 challenging multiple-choice QA pairs across six distinct tasks using a semi-automated pipeline. Specifically, we employ the powerful DeepSeek-R1 [13] to generate initial QA pairs based on structured motion annotation metadata, followed by blind filtering and single-frame filtering to remove questions that can be solved through common sense or isolated frames. Subsequently, all QA pairs undergo manual verification to ensure quality. For open-ended evaluation, we construct fine-grained motion-level captions for each video to assess models' generative capabilities. We follow existing benchmarks with open-ended evaluation [60, 39] to incorporate the widely-used GPT-assisted evaluation. While relatively reliable, it comes with substantial costs when evaluating numerous models or on large-scale datasets. Therefore, we propose a novel LLM-free framework as a complementary evaluation approach with better interpretability and reproducibility, making the assessment of open-ended motion understanding more accessible.

Comprehensive results on FAVOR-Bench reveal significant limitations in current video understanding models' fine-grained motion comprehension capabilities. In close-ended evaluation, Gemini-1.5-Pro achieves the highest overall accuracy (49.77%), while Qwen2.5-VL-72B leads among open-source models (47.96%). Most models perform better on ego-centric videos than third-person videos, suggesting that capturing complex camera motions and interactions among multiple subjects in third-person perspectives remains challenging. For open-ended evaluation, Tarsier2-Recap-7B demonstrates the strongest generative performance in GPT-assisted (4.60/4.38) and LLM-free (56.58%) evaluations. However, the performance is still below the practical deployment expectations for real-world applications. Notably, our proposed LLM-free framework shows a strong correlation with GPT-assisted evaluations, validating its effectiveness as a more accessible evaluation alternative. To promote the development of fine-grained motion comprehension, we further build FAVOR-Train, comprising 17,152 videos with fine-grained manual annotations spanning both third-person and ego-centric perspectives. By performing supervised fine-tuning (SFT) on Qwen2.5-VL [52] with FAVOR-Train, we achieve consistent improvements across all metrics: +1.07% in close-ended accuracy, +0.27/+0.12 points in GPT-assisted evaluation, and +7.87% in LLM-free evaluation. We further evaluate the fine-tuned model on TVBench [9] and MotionBench [14], also enhancing the performance on motion-related tasks. These evaluations demonstrate that FAVOR-Train can effectively strengthen fine-grained motion comprehension capabilities, showcasing its value for developing more powerful video understanding models.

Our contribution can be concluded from three aspects:

- We present FAVOR-Bench, the first fine-grained video motion understanding benchmark that spans both ego-centric and third-person perspectives with comprehensive evaluation including both close-ended QA tasks and open-ended descriptive tasks.
- We propose a novel LLM-free evaluation framework that complements GPT-assisted assessment, providing a more accessible, interpretable, and cost-effective approach for evaluating open-ended motion descriptions.
- We construct FAVOR-Train, a training dataset covering third-person and ego-centric videos with fine-grained motion annotations, which can improve MLLMs' motion understanding capabilities and promote the development of more powerful models.

2 Related Work

2.1 MLLMs for Video Understanding

Recent advancements in Multimodal Large Language Models (MLLMs) have enhanced video understanding capabilities through module designs and scalable training paradigms. Models include VideoLLaMA3 [8, 56] pioneer vision-centric designs with dynamic tokenization to capture fine-grained spatial details and temporal dynamics. The Tarsier series [43, 55] combines CLIP encoders with LLMs and temporal alignment techniques, enabling precise video description and causal reasoning. Qwen2-VL [45] and InternVL 2.5 [6] unify multimodal processing through dynamic resolution mechanisms and propose advanced positional embeddings, which support high-resolution inputs and better multimodal integration. Additionally, performance across various video tasks of MLLMs also benefits from larger-scale training data and parameters [52, 7]. These rapid advancements underscore the necessity for more challenging benchmarks to evaluate specific aspects of MLLMs' video understanding capabilities.

2.2 Video Understanding Benchmarks

With the development of video understanding models, various benchmarks have been constructed to evaluate their capabilities. Traditional benchmarks primarily evaluate basic video understanding capabilities [50, 3]. Subsequent works reveal the single frame bias [16, 22] and emphasize evaluating temporal dynamics from diverse aspects [35, 28]. Recently, benchmarks focusing on different aspects have been constructed, as depicted in Table 1. Comprehensive benchmarks like MVBench [24] and Video-MME [11] evaluate general video understanding capabilities. Besides, specialized benchmarks focused on specific challenging scenarios, such as long-form videos [60, 30], counter-intuitive reasoning [49], and ego-centric video understanding [30].

For evaluation methods, multiple-choice questions remain the most common approach [24, 9, 30, 60, 14], while open-ended evaluation is increasingly adopted. Early benchmarks employ similarity-based metrics for short generative tasks [48, 54] while descriptive tasks with long responses [29, 60, 5] commonly utilize GPT-assisted evaluation. However, recent studies raise concerns about the interpretability and evaluation costs on large-scale datasets [9, 44], highlighting the need for more accessible evaluation frameworks.

For motion understanding in videos, ActivityNet-QA [3] introduces VideoQA with manually annotated datasets. NExT-QA [48] further evaluates causal and temporal reasoning abilities. Several comprehensive benchmarks include action and motion understanding subsets [60, 24, 9]. However, these works primarily evaluate at the event level, lacking consideration for more fine-grained motions. The concurrent MotionBench [14] attempts to bridge this gap through multiple-choice evaluation of motion-level perception but is limited to third-person perspectives and close-ended evaluation. Compared to existing works, our FAVOR-Bench provides a more comprehensive evaluation for fine-grained motion understanding from both video perspectives and evaluation tasks. Furthermore, we propose an LLM-free framework complementary to GPT-assisted evaluation that benefits the accessibility of open-ended motion understanding evaluation.

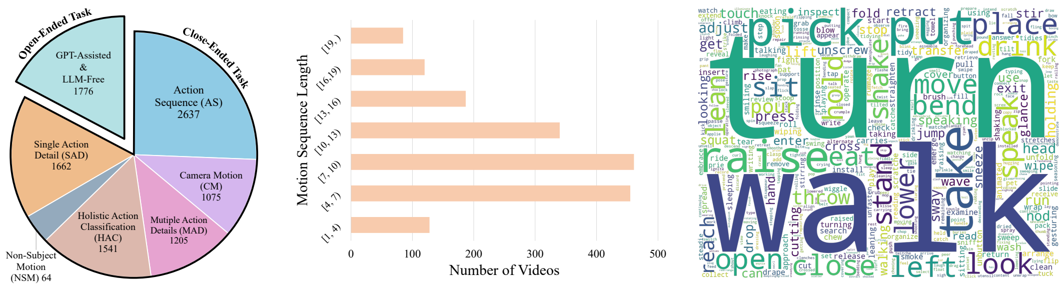

Figure 2: Data statistics of FAVOR-Bench. **Left:** Task type distribution across close-ended and open-ended evaluation in FAVOR-Bench. **Middle:** Distribution of annotated motion sequence length per video. **Right:** The word cloud statistics of motion vocabularies in FAVOR-Bench.

Table 1: Comparison of FAVOR-Bench with existing video understanding benchmarks. **#Videos** and **#Close-Ended QA** refer to the number of videos and close-ended question-answer pairs respectively. FAVOR-Bench covers wide video types (Third-Person, Ego-Centric, Simulation) while focusing on fine-grained motion understanding. Moreover, FAVOR-Bench provides comprehensive evaluation, including close-ended QA and open-ended tasks (both GPT-assisted and a novel LLM-Free evaluation).

Benchmarks	#Videos	Video Type			Fine-Grained Motion	#Close-Ended QA	Open-Ended Evaluation	
		Third-Person	Ego-Centric	Simulation			GPT-Assisted	LLM-Free
MVBench [24]	4,000	✓	✗	✓	✗	4,000	✗	✗
TVBench [9]	2,525	✓	✗	✓	✗	2,525	✗	✗
AutoEval-Video [5]	327	✓	✓	✓	✗	–	✓	✗
EgoSchema [30]	5,031	✗	✓	✗	✗	5,031	✗	✗
EgoTaskQA [20]	2,315	✗	✓	✗	✓	40,322	✗	✗
MLVU [60]	1,730	✓	✗	✓	✗	3,102	✓	✗
MovieChat-1K [38]	130	✓	✗	✗	✗	1,950	✗	✗
MotionBench [14]	5,385	✓	✗	✓	✓	8,052	✗	✗
FAVOR-Bench	1,776	✓	✓	✓	✓	8,184	✓	✓

3 FAVOR-Bench: Fine-Grained Video Motion Understanding Benchmark

This section presents FAVOR-Bench, a comprehensive benchmark for fine-grained video motion understanding. We start with a brief overview of FAVOR-Bench. Then, we provide detailed descriptions of the dataset curation and evaluation tasks.

3.1 Overview

FAVOR-Bench consists of 1,776 carefully curated videos spanning diverse domains and perspectives. The video durations add up to 10.2 hours, with an average of 20.6 seconds. Using a semi-automatic pipeline, we construct 8,184 challenging QA pairs based on fine-grained structured annotations, challenging models through six motion-centric tasks. In addition, FAVOR-Bench includes open-ended evaluation, comprising the widely adopted GPT evaluation and our novel LLM-free evaluation framework. Through these tasks, we comprehensively assess models' fine-grained motion understanding and description capabilities. Figure 2 shows data statistics of FAVOR-Bench, including task distribution, annotated motion sequence length, and motion vocabulary distribution.

3.2 Dataset Curation

This section elaborates on FAVOR-Bench's dataset curation process, including data collection, filtering, and manual annotation.

Data Collection and Filtering. The raw videos we collected consist of four types: daily-life records, TV series, animations and ego-centric videos. To ensure fine-grained motion understanding and annotation quality, we choose videos with rich motions and manageable durations. For daily-life record, 868 videos are sampled from Charades [37] with the highest quality scores (provided by Charades). For TV series and animations, video clips with high motion quality are acquired with a comprehensive pipeline including scene-aware cropping, optical flow-based filtering, and manual curation. 574 clips from TV-series and 138 clips from animations are selected. For Egocentric videos, we select EgoTaskQA [20] as the data source and randomly sample 196 videos. More details of video curation and filtering are provided in Section C of the Supplementary Materials.

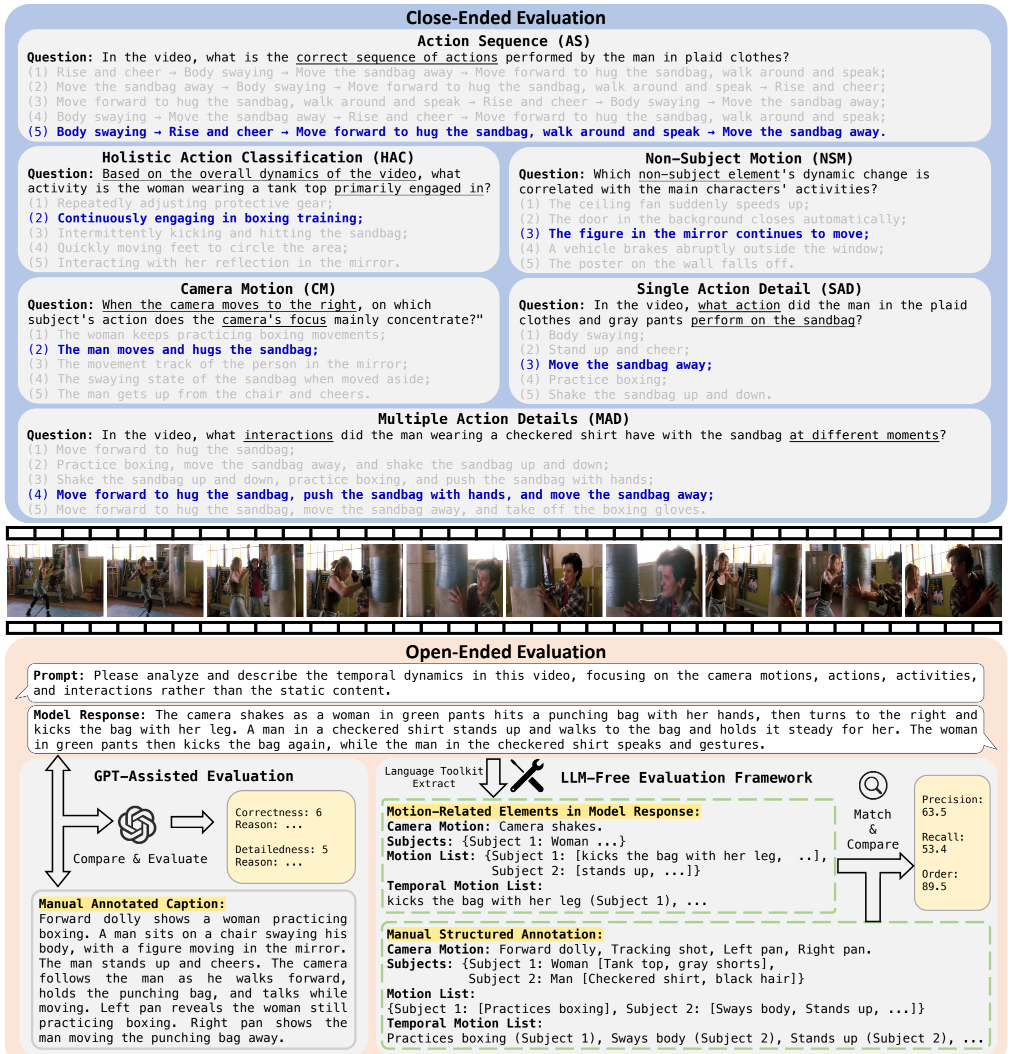

Figure 3: Overview of evaluation tasks. FAVOR-Bench comprises close-ended and open-ended evaluations. The close-ended evaluation includes six tasks focusing on different aspects. The open-ended evaluation comprises a GPT-assisted evaluation and a novel LLM-free framework. The former directly compares model responses with manual captions, while the latter parses structured motion elements from responses and compares them with the structured annotations.

Manual Annotation. We hired eight highly educated personnel for a two-week full-time labeling process with a manual inspection-revision mechanism. For each video, structured annotations include: 1) Subjects (such as man, dog, first-person subject, etc.) involved in motion or action, with up to three attributes (like wearing, color, etc.) 2) Per-subject motion lists with timestamps specified in seconds. 3) Camera motions with timestamps specified in seconds. 4) Comprehensive video captions, which includes all the annotated subjects, motions, and camera motions mentioned above.

3.3 Close-Ended Evaluation

3.3.1 Task Definition

Close-ended tasks are widely adopted as a quantitative evaluation of specific capabilities. FAVOR-Bench examines six critical dimensions of fine-grained motion understanding through carefully designed tasks, formatted as multiple-choice questions (illustrated in the upper part of Figure 3). The detailed explanations of each task are as follows:

Action Sequence (AS). The action sequence task focuses on understanding the temporal dynamics in the video. In this task, one or more subjects in the video may perform a series of complex actions, and the models are required to compare which action occurs first or answer the complete action sequence of a specific subject in the video.

Holistic Action Classification (HAC). The holistic action classification task requires models to answer the core action of the subjects in the video. This task is similar to traditional video action classification and recognition tasks, focusing on the ability to summarize global action.

Single Action Detail (SAD). This task examines the moment-specific detail recognition ability. The model will be asked about the subjects' states at a specific moment and the interaction between the subject and an object.

Multiple Action Details (MAD). This task focuses on evaluating the ability to compare and analyze details across multiple moments. The model will be required to answer the changes in the actions and states of the subject over time, or the interactions between the subject and multiple objects.

Camera Motion (CM). The camera motion task examines the understanding of viewpoint dynamics, focus shifts, and their coordination with subject actions in the video. This capability is equally essential for fine-grained action understanding, as camera motion may affect the visibility of the subjects or cause a subject switch.

Non-Subject Motion (NSM). This task focuses on evaluating the environmental context awareness of models, such as the movements and behaviors of non-subject elements (including background objects and passersby) in the video. Non-subject motions can serve as peripheral cues to refine the understanding of primary motion, especially in real-world scenarios.

3.3.2 QA Generation

Based on the curated video dataset and structured motion annotations, we adopt a three-stage semi-automatic pipeline to construct QA pairs: 1) Automatic Question-Answer Generation, 2) Blind Filtering and Single-Frame Filtering, and 3) Manual Verification. After the above steps, FAVOR-Bench constructs an average of 4.6 multiple-choice questions for each video.

Automatic Question-Answer Generation. For each of the six tasks, we design distinct prompt templates to generate QA pairs from the annotation metadata using DeepSeek-R1 [13]. Our guidelines for automatic QA generation include two critical requirements: 1) maximizing question diversity to cover more video content, and 2) crafting challenging distractors without compromising the uniqueness of the correct answers. The specific prompt templates can be found in Section C of the Supplementary Materials. Through this process, we obtain 20,402 multiple-choice QA pairs in total.

Blind Filtering and Single-Frame Filtering. To ensure the benchmark's quality and challenge, we design a two-stage filtering framework comprising blind and single-frame filtering to help remove low-quality QA pairs. Specifically, our analysis reveals that a subset of generated questions can be resolved solely through common sense and language priors, without any visual information. Furthermore, the single-frame bias [16, 22], which refers to scenarios where a single frame suffices to answer the video-understanding question, can also impact evaluation. For blind filtering, we choose Qwen2-72B [51] as the representative LLM to answer the questions without visual inputs and remove correctly addressed questions. Subsequently, for single-frame filtering, the remaining QA pairs are fed into GPT-4o [19] together with five frames uniformly sampled from the corresponding video. The correctly answered questions are further filtered out. While this filtering framework might inadvertently exclude instances where the answers are correct by random chance, it effectively enhances the quality of the constructed close-ended tasks. After this process, 12,096 QA pairs remain.

Manual Verification. In this stage, annotators verify all QA pairs from three perspectives: question-video relevance, answer correctness, and answer uniqueness. Ambiguous or incorrect QA pairs are discarded. We further calibrate option distributions to mitigate position and frequency biases. FAVOR-Bench ultimately comprises 8,184 multiple-choice questions.

3.4 Open-Ended Evaluation

Beyond close-ended tasks that provide constrained options, FAVOR-Bench further challenges the models' comprehensive fine-grained motion understanding and description capabilities through generative tasks. Our open-ended evaluation includes two types of metrics: GPT-assisted evaluation and a novel LLM-free evaluation (illustrated in the lower part of Figure 3).

3.4.1 GPT-Assisted Evaluation

Following existing benchmarks with generative tasks [39, 60, 5], we employ GPT-4o [19] to assist in the open-ended evaluation. Specifically, we prompt the models being evaluated to describe the temporal dynamics in the video, including subject actions, camera motions, etc. Then, we input the generated responses and the manually crafted descriptions into GPT-4o for comparison and scoring from both correctness and detailedness perspectives. To reduce randomness and enhance the evaluation's robustness, we set the scoring range (from 1 to 10) in the prompt template and define the specific criteria for each score level. The prompt template is provided in Section C of the Supplementary Materials.

3.4.2 LLM-Free Evaluation

While using the powerful GPT models for evaluation has become a common practice, their high cost and limited interpretability pose challenges for reproducible benchmarking. These limitations have motivated our development of an LLM-free evaluation framework for open-ended, fine-grained video motion understanding.

We first develop a structured information extraction tool based on the NLTK library to obtain motion-related elements from the model's response, including camera motion, subject list, individual subject motion lists, and a comprehensive temporal motion list (including actions of all subjects). Specifically, our tool implements a hierarchical parsing approach with predefined lexicons and pattern-matching rules. It first tokenizes responses into sentences, then extracts camera motion using compiled patterns, before identifying subjects and their associated actions through part-of-speech tagging and context-aware heuristics. This structured representation bridges model responses with quantitative evaluation metrics while maintaining human interpretability.

With the extracted elements, we calculate the score of the model's response through hierarchical sequence comparison. Specifically, we adopt pre-trained models such as Sentence-BERT [34] to calculate the semantic similarity of the extracted and manually annotated subject attributes and complete action sequences, and combine these two types of similarities for subject matching. Next, we conduct sequence comparisons in three aspects to obtain the comprehensive score: the camera motion sequence, each subject's action sequence, and the comprehensive temporal action sequence. Taking the action sequence of one subject as an example, the predicted subject S is matched to the manually annotated subject G. Their action sequences are represented as $[a_1^s, ..., a_n^s]$ and $[a_1^g, ..., a_m^g]$ respectively. We construct an action similarity matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$ where each element M_{ij} represents the semantic similarity between predicted action a_i^s and ground truth action a_j^g:

$$M_{ij} = \text{sim}(a_i^s, a_j^g) \tag{1}$$

Based on this matrix and optimal matching, we calculate the similarity-weighted precision and recall:

$$P = \frac{|\text{matched predicted actions}|}{|\text{predicted actions}|} \cdot \overline{\text{sim}} \cdot L_f, \quad R = \frac{|\text{matched annotated actions}|}{|\text{annotated actions}|} \cdot \overline{\text{sim}} \cdot L_f, \tag{2}$$

where $\overline{\text{sim}}$ is the average similarity score of matched pairs. L_f is a length factor used to penalize the unfair comparisons that may occur due to a large discrepancy in sequence lengths (such as numerous repeated descriptions). For each pair of subjects, we further evaluate the order correctness using Kendall's Tau coefficient τ to measure the rank correlation between matched action indices. The calculated score for each subject pair is a weighted combination of multiple dimensions: $Score = w_pP + w_rR + w_oO$, where w_s, w_r and w_o are the weights of different indicators. The camera motion and the comprehensive temporal action sequence can be scored similarly. The final score of each model is obtained by averaging over all samples.

4 Experiments

4.1 Experimental Settings

We perform a comprehensive evaluation of 21 MLLMs through our FAVOR-Bench, which includes open-source and proprietary models. For models that are part of a series, we evaluate their most

Table 2: Comprehensive results of 21 MLLMs on FAVOR-Bench, including performance on overall benchmark, third-person videos, and ego-centric videos. Random selecting and human performance are also compared. For each category, we report closed-ended multiple choice (MCQA) and open-ended evaluation, including GPT-assisted scores (GPT-C / D) and LLM-free scores. GPT-C / D mean correctness and detailedness scores generated by GPT-4o. The highest and suboptimal results are **bolded** and underlined, respectively. Due to API response limitations, the video input of proprietary MLLMs is restricted to 16 frames if the video is longer than 16 seconds (demoted as "1 fps*"). Tarsier2-Recap-7B is a model specially designed for captioning and its close-ended performance is not compared.

Methods	Date	Input	Overall			Third-Person			Ego-Centric		
			MCQA	GPT-C / D	LLM-Free	MCQA	GPT-C / D	LLM-Free	MCQA	GPT-C / D	LLM-Free
Full mark	–	–	100	10 / 10	100	100	10 / 10	100	100	10 / 10	100
Random	–	–	20	–	–	20	–	–	20	–	–
Human	–	–	90.92	7.29 / 6.24	74.25	90.91	7.33 / 6.21	74.12	91.28	7.00 / 6.45	75.26
Proprietary MLLMs											
Gemini-1.5-Pro [40]	2024-04	1 fps*	**49.77**	4.52 / **4.68**	52.91	**48.93**	4.56 / 4.71	53.62	56.58	**4.18** / **4.44**	45.37
GPT-4o [19]	2024-08	1 fps*	41.18	4.33 / 4.01	49.50	40.23	4.35 / 4.06	50.07	48.88	4.17 / 3.60	44.86
Claude-3.7-Sonnet [1]	2025-02	1 fps*	43.35	4.32 / 4.63	43.03	42.93	4.44 / **4.77**	43.92	46.76	3.34 / 3.48	35.75
Open-source MLLMs											
Video-LLaVA-7B [27]	2023-11	8 frms	23.45	2.18 / 2.31	41.36	23.27	2.22 / 2.37	41.69	24.89	1.84 / 1.86	38.71
LLaVA-NeXT-Video-7B [58]	2024-05	8 frms	22.45	2.57 / 2.02	29.48	22.17	2.64 / 2.08	29.57	24.67	1.99 / 1.57	28.84
LLaVA-NeXT-Video-34B [58]	2024-05	8 frms	29.51	2.83 / 2.67	39.41	29.57	2.85 / 2.70	39.39	29.02	2.64 / 2.42	39.58
Tarsier-7B [43]	2024-07	8 frms	14.04	3.47 / 2.80	46.25	13.41	3.53 / 2.87	46.87	19.20	2.99 / 2.31	41.22
Tarsier-34B [43]	2024-07	8 frms	26.94	3.79 / 2.97	47.13	26.37	3.81 / 3.01	47.17	31.58	3.56 / 2.63	46.76
Aria [23]	2024-10	8 frms	28.60	2.85 / 2.61	42.78	28.13	2.88 / 2.61	43.33	32.48	2.59 / 2.56	38.26
InternVL2.5-2B [6]	2024-12	8 frms	22.72	2.80 / 2.99	43.23	22.72	2.88 / 3.15	43.33	22.66	2.14 / 1.71	41.42
InternVL2.5-8B [6]	2024-12	8 frms	34.41	3.11 / 3.38	44.18	33.93	3.08 / 3.40	44.60	38.28	3.36 / 3.26	40.70
InternVL2.5-78B [6]	2024-12	8 frms	38.42	2.98 / 3.41	44.01	37.49	3.05 / 3.52	44.36	45.98	2.40 / 2.47	40.68
Tarsier2-Recap-7B [55]	2024-12	16 frms	–	**4.60** / 4.38	**56.58**	–	**4.66** / 4.48	**56.88**	–	4.10 / 3.62	**54.19**
LLaVA-Video-7B-Qwen2 [59]	2024-10	64 frms	38.39	3.57 / 3.40	45.41	37.53	3.58 / 3.42	45.96	45.42	3.43 / 3.23	40.95
LLaVA-Video-72B-Qwen2 [59]	2024-10	64 frms	45.81	3.42 / 3.42	46.06	44.73	3.47 / 3.46	46.89	54.58	3.01 / 3.05	39.53
VideoChat-Flash-Qwen2-7B [25]	2025-01	1 fps	43.52	3.25 / 2.55	40.82	43.19	3.38 / 2.68	41.42	46.21	2.21 / 1.54	35.98
VideoLLaMA3-2B [56]	2025-01	1 fps	32.76	3.14 / 2.98	39.29	32.26	3.17 / 3.03	39.86	36.83	2.90 / 2.62	34.86
VideoLLaMA3-7B [56]	2025-01	1 fps	41.23	3.64 / 3.24	48.63	40.60	3.68 / 3.28	49.32	46.32	3.31 / 2.87	43.08
Qwen2.5-VL-3B [52]	2025-01	1 fps	36.83	2.77 / 2.91	47.32	35.94	2.82 / 2.98	47.56	44.08	2.37 / 2.34	45.30
Qwen2.5-VL-7B [52]	2025-01	1 fps	40.49	3.28 / 3.41	48.46	39.81	3.30 / 3.44	48.64	46.09	3.15 / 3.18	46.98
Qwen2.5-VL-72B [52]	2025-01	1 fps	47.96	3.37 / 3.44	49.72	46.67	3.35 / 3.45	50.05	**58.48**	3.51 / 3.29	47.06
Qwen2.5-VL-7B+FAVOR-Train	–	1 fps	41.56	3.55 / 3.53	56.33	40.74	3.54 / 3.47	55.70	48.21	3.62 / 3.98	61.38

recently released versions like VideoLLaMA3 [56], InternVL2.5 [6] and Qwen2.5-VL [52]. All models are evaluated using either their official implementations or accessible APIs, and all assessments are conducted in a zero-shot manner. We employ either a uniform sampling strategy or a frame rate sampling strategy to form the vision input following the official examples of each model. For close-ended evaluation, we prompt models to choose from the provided options rather than merely output the chosen indices. For open-ended evaluation, models are prompted to focus more on temporal dynamics rather than static content. Experiments are conducted with 8×A800 GPUs.

4.2 Results Analysis on FAVOR-Bench

We report results of 21 MLLMs on FAVOR-Bench in Table 2, evaluating their performance on both third-person and ego-centric videos through close-ended multiple-choice questions (MCQA) and open-ended tasks. The results reveal critical limitations in fine-grained video motion understanding. Task-specific results and analysis are provided in Section A of the Supplementary Materials.

Close-Ended Performance Analysis. As can be concluded from Table 2, proprietary MLLMs generally outperform open-source alternatives, though advanced open-source models are narrowing this gap. Gemini-1.5-Pro achieves the highest MCQA score (49.77%), yet remains substantially below practical deployment expectations for real-world applications. For open-source models, we observe evident scaling effects. For example, Qwen2.5-VL-72B scores 47.96%, exceeding its smaller variants by 7-11%. At the widely adopted 7B scale, VideoChat-Flash-Qwen2-7B [25] (43.52%) demonstrates strong capabilities . Most models perform better on ego-centric videos than third-person videos, suggesting that the complex camera motions and interactions among multiple subjects in third-person videos are more challenging for MLLMs.

Open-Ended Performance Analysis. The open-ended evaluation results in Table 2 show that describing fine-grained motions in videos poses challenges for existing MLLMs. Tarsier2-Recap-7B

Table 3: Comparison on TVBench and MotionBench with our proposed FAVOR-Train. AVG means the average score of all the 10 tasks of TVBench. ALL denotes the accuracy on all 4,018 questions of MotionBench-Dev. Qwen2.5-VL gains considerable performance improvement from fine-tuning with FAVOR-Train.

Methods	TVBench											MotionBench-Dev						
	AVG	AC	OC	AS	OS	ST	AL	AA	UA	ES	MD	ALL	MR	LM	CM	MO	AO	RC
Random	33.3	25.0	25.0	50.0	33.3	50.0	25.0	50.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Qwen2.5-VL-7B [52]	45.2	36.8	36.5	64.8	37.3	62.7	**38.8**	74.1	**41.5**	26.0	33.6	46.2	45.2	**45.4**	42.3	**66.7**	36.0	**33.0**
+ FAVOR-Train	**46.1**	**43.1**	**37.2**	**68.0**	**37.8**	**62.7**	34.4	**77.5**	37.8	**29.0**	**34.1**	**47.9**	**49.2**	**45.4**	**43.1**	66.2	**37.8**	32.3

achieves the highest performance in both GPT-assisted (4.60/4.38) and LLM-free (56.58) evaluations. Among general-purpose models, Gemini-1.5-Pro leads with 4.52/4.68 GPT scores and 52.91 LLM-free score. In comparison to MCQA, most models perform better on third-person videos in open-ended evaluations, suggesting that while MLLMs can understand ego-centric motions, they are not adept at describing them.

Reliability of LLM-Free Framework. Our proposed LLM-free metric exhibits strong correlations with GPT-assisted evaluations, achieving a Pearson correlation of 0.86 ($p < 0.001$) and a Spearman correlation of 0.77 ($p < 0.001$) with GPT-C/D scores. These correlations validate the LLM-free framework as a reliable and more accessible alternative to costly GPT-based evaluation for open-ended fine-grained motion description. More discussions about the LLM-free framework are provided in Section B of the Supplementary Materials.

Human Performance Baseline. We hired ten highly educated personnel to evaluate FAVOR-Bench, covering both MCQA and open-ended generative tasks. The human performance results are also presented in Table 2. Even the best-performing Gemini-1.5-Pro shows a substantial performance gap from human performance. This significant disparity quantifies the substantial room for improvement in current MLLMs' fine-grained video motion understanding abilities.

4.3 FAVOR-Train Set

To facilitate better video motion understanding and description, we further propose a training set, FAVOR-Train, including 17,152 videos covering third-person and ego-centric perspectives and the corresponding manual captions. All the 14,038 third-person videos are sourced from the Koala36M [46] dataset, which provides a rich variety of motions and interactions in vast scenarios. For the ego-centric portion with 3,114 videos, we curated data from four distinct datasets: EgoTaskQA [20], Charades-Ego [36], EgoExo4D [12], and EgoExoLearn [17]. To ensure a diverse range of activities and contexts within our dataset, specially designed sampling strategies are employed. There is no intersection between the videos in FAVOR-Train and FAVOR-Bench. Details of the sampling strategies can be found in Section D of the Supplementary Materials.

To validate the effectiveness of FAVOR-Train, we fine-tune the Qwen2.5-VL model using FAVOR-Train data and evaluate its performance on both FAVOR-Bench and existing benchmarks with motion-related tasks. As shown in Table 2, FAVOR-Train brings consistent improvements across all evaluation metrics. For the closed-ended MCQA tasks, the model achieves a 1.07% accuracy gain overall. The enhancement in open-ended evaluation is more significant, particularly for the LLM-Free score, which increases by 7.87% (from 48.46% to 56.33%). The GPT-assisted evaluation scores also improve across all categories, with gains of 0.27 and 0.12 points in correctness and detailedness, surpassing Qwen2.5-VL-72B. Table 3 demonstrates that FAVOR-Train also enhances performance on existing benchmarks. On TVBench [9], the average accuracy improves from 45.2% to 46.1%, with notable gains in action-related tasks like Action Count (AC: +6.3%) and Action Sequence (AS: +3.2%). On MotionBench-Dev [14], the overall accuracy increases by 1.7%, with the largest improvement in the Motion Recognition (MR) task (+4.0%).

5 Ethical Considerations and Societal Impact

Copyrights. FAVOR-Bench is provided as a research resource for non-commercial applications only. For open-sourced video datasets [37, 20, 36, 17, 12, 46], we carefully comply with their respective licenses. For self-collected video clips, users must acknowledge a non-commercial research agreement that prohibits the redistribution of any videos.

Broader Societal Impact. While FAVOR-Bench advances fine-grained video motion understanding with positive applications in accessibility technologies and human-computer interaction, we acknowledge potential limitations and risks. Our dataset may have cultural and regional under-representation and not fully represent global diversity in human actions. Besides, enhanced motion understanding could be misused for unauthorized behavior analysis. Therefore, we emphasize the research-only nature of our benchmark and encourage the community to prioritize beneficial applications that respect human dignity and privacy when building upon our work.

6 Conclusion

We present FAVOR-Bench, the first comprehensive benchmark for evaluating fine-grained video motion understanding in MLLMs across both ego-centric and third-person perspectives and both close-ended and open-ended tasks. For close-ended assessment, 8,184 multiple-choice QA pairs across six distinct tasks are carefully curated. The open-ended evaluation comprises GPT-assisted and our novel LLM-free evaluation. The comprehensive results of 21 state-of-the-art MLLMs reveal significant limitations in understanding detailed temporal dynamics. To narrow this gap, we construct FAVOR-Train with 17,152 videos spanning both video perspectives with fine-grained annotations, which effectively improves the motion understanding capabilities both on our proposed FAVOR-Bench and motion-related tasks of existing benchmarks. Through FAVOR-Bench and FAVOR-Train, we provide valuable tools for developing more powerful video understanding models.

Acknowledgments

This work is supported by National Key Research and Development Program of China (No. 2022ZD0160101), Shanghai Natural Science Foundation (No. 23ZR1402900), Shanghai Science and Technology Commission Explorer Program Project (24TS1401300), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103). The computations in this research were performed using the CFFF platform of Fudan University.

References

[1] Anthropic. Claude 3.7. https://www.anthropic.com/claude/sonnet, February 2025. 8, 3, 4

[2] Ali Athar, Xueqing Deng, and Liang-Chieh Chen. Vicas: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. *arXiv preprint arXiv:2412.09754*, 2024. 1

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 3, 5

[4] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 1

[5] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pages 179–195. Springer, 2024. 3, 4, 7

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 3, 8, 4

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 3

[8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 3

[9] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024. 2, 3, 4, 9

[10] Zichen Jeff Cui, Yibin Wang, Nur Muhammad Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncurated robot data. In *11th International Conference on Learning Representations, ICLR 2023*, 2023. 1

[11] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 1, 3

[12] 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*, pages 19383–19400, 2024. 1, 9

[13] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 6, 4

[14] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. *arXiv preprint arXiv:2501.02955*, 2025. 2, 3, 4, 9

[15] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18219–18228, 2022. 1

[16] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7366–7375, 2018. 3, 6

[17] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. 9

[18] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 1

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 7, 8, 3, 4

[20] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, 2022. 4, 9, 3

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[22] Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–507, 2023. 3, 6

[23] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 8, 3, 4

[24] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1, 2, 3, 4

[25] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 8, 3, 4

[26] Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*, 2024. 1

[27] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024. 8, 3, 4

[28] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 3

[29] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3

[30] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 3, 4

[31] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 5

[32] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 1

[33] Haomiao Ni, Bernhard Egger, Suhas Lohit, Anoop Cherian, Ye Wang, Toshiaki Koike-Akino, Sharon X Huang, and Tim K Marks. Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9015–9025, 2024. 1

[34] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. 7

[35] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 535–544, 2021. 3

[36] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 9

[37] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 4, 9, 2

[38] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 4

[39] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang Yan Gui, Yu Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 13088–13110. Association for Computational Linguistics (ACL), 2024. 2, 7

[40] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 8, 3, 4

[41] Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *arXiv preprint arXiv:2503.08342*, 2025. 1

[42] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. In *Conference on Robot Learning*, pages 201–221. PMLR, 2023. 1

[43] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 3, 8, 4

[44] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, 2023. 3

[45] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[46] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. *arXiv preprint arXiv:2410.08260*, 2024. 1, 9, 5

[47] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 1

[48] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 3

[49] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pages 39–57. Springer, 2024. 1, 3

[50] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 3

[51] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. 6

[52] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1, 2, 3, 8, 9, 4

[53] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685, 2023. 1

[54] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 2, 3

[55] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025. 3, 8, 4

[56] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 1, 3, 8, 4

[57] Lin Zhang, Xianfang Zeng, Kangcong Li, Gang Yu, and Tao Chen. Sc-captioner: Improving image captioning with self-correction by reinforcement learning. *arXiv preprint arXiv:2508.06125*, 2025. 1

[58] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. 8, 3, 4

[59] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 8, 3, 4

[60] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 1, 2, 3, 4, 7

[61] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252, 2024. 1

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section C the Supplementary Materials.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In the experiment section, we give detailed information about the experimental setup, evaluated models and evaluation metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide access to the data and evaluation code for reproduction.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental settings are indicated in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments compute resources are indicated in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are discussed in the abstract and introduction. We aim to provide valuable tools for the community to develop more powerful video understanding models.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: For the videos we have collected ourselves, we have carried out manual verification during the data filtering process to minimize the release of unsafe content.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited all papers. The license and copyright information related to data from existing datasets and benchmarks are discussed in Section C of the Supplementary Materials.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Details of the dataset and code, including the license and limitations, are discussed in the Supplementary Materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The labeling process is discussed in Section 3. The annotation guidelines are provided in Section C of the Supplementary Materials.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper only involves video annotation, and there are no potential risks.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.