
EgoExoBench: A Benchmark for First- and Third-person View Video Understanding in MLLMs

Yuping He^{1,2*}, Yifei Huang^{2,3*}, Guo Chen¹, Baoqi Pei^{2,4}, Jilan Xu^{2,5}
Tong Lu^{1†} and Jiangmiao Pang^{2†}

¹Nanjing University, ²Shanghai AI Laboratory, ³The University of Tokyo

⁴Zhejiang University, ⁵Fudan University

<https://github.com/ayiyayi/EgoExoBench>

Abstract

Transferring and integrating knowledge across first-person (egocentric) and third-person (exocentric) viewpoints is intrinsic to human intelligence, enabling humans to learn from others and convey insights from their own experiences. Despite rapid progress in multimodal large language models (MLLMs), their ability to perform such cross-view reasoning remains unexplored. To address this, we introduce EgoExoBench, the first benchmark for egocentric-exocentric video understanding and reasoning. Built from publicly available datasets, EgoExoBench comprises over 7,300 question–answer pairs spanning eleven sub-tasks organized into three core challenges: semantic alignment, viewpoint association, and temporal reasoning. We evaluate 13 state-of-the-art MLLMs and find that while these models excel on single-view tasks, they struggle to align semantics across perspectives, accurately associate views, and infer temporal dynamics in the ego-exo context. We hope EgoExoBench can serve as a valuable resource for research on embodied agents and intelligent assistants seeking human-like cross-view intelligence.

1 Introduction

Understanding and transferring knowledge between first-person (egocentric) and third-person (exocentric) perspectives is a fundamental aspect of human intelligence [17, 27]. We routinely learn new skills by observing third-person demonstrations and mapping those actions onto our own egocentric experience [3, 20, 53, 24, 21]. Conversely, we can adopt an exocentric view to articulate insights and guide others [9, 5]. Beyond simple transfer, humans excel at cross-view joint reasoning: in procedural tasks, we align live egocentric observations with exocentric examples to anticipate next steps [27, 28]; in navigation, we fuse egocentric experience with a global map to plan efficient routes [6, 13, 33]. This dual-view reasoning enables robust perception and skill acquisition in dynamic environments.

Replicating this ego–exo reasoning in artificial systems offers a promising path toward more capable embodied agents. Robots that could align large-scale exocentric knowledge, such as instructional video, can execute tasks more effectively and better perform imitation learning [58, 1, 57]. Bridging first- and third-person perspectives is also essential for seamless human–robot collaboration [41, 66].

Recent benchmarks [45, 14, 84, 38, 8] have driven significant advances in single-view video understanding for multimodal large language models (MLLMs) [85, 62, 81, 35, 78]. For example, EgoSchema [45] evaluates temporal reasoning in egocentric streams, Video-MME [14] evaluates multi-modal QA on diverse third-person footage, and MLVU [84] tests long-form video comprehension. While each benchmark pushes the boundary of MLLMs in video understanding from different perspectives, they all operate on either egocentric or exocentric data in isolation. There is

*equal contribution. † corresponding authors

no systematic evaluation of a model’s ability to integrate information across ego–exo viewpoints, leaving the question of whether MLLMs can, like humans, perform cross-view reasoning.

To fill this gap, we introduce EgoExoBench, the first benchmark designed specifically for cross-view video understanding in MLLMs. EgoExoBench is built from publicly available paired egocentric–exocentric video sources and curated through a rigorous annotation protocol. It comprises over 7,300 question–answer pairs that target three key dimensions: Ego-Exo Matching, Ego-Exo View Transition, and Ego-Exo Temporal Reasoning. Ego-Exo Matching evaluates semantic alignment by testing whether models recognize the same entities or actions across first- and third-person views. Ego-Exo View Transition probes spatial correspondence, measuring a model’s ability to translate between egocentric and exocentric viewpoints. Ego-Exo Temporal Reasoning examines temporal sequence reasoning by asking models to align and predict event sequences across asynchronous or overlapping video streams. 11 subtasks span these three dimensions, forming a comprehensive evaluation suite for ego–exo view understanding and reasoning.

We conduct extensive experiments on 13 open- and closed-source MLLMs, including the GPT-4o [29], GPT-o4-mini [47], Claude 3.7 Sonnet [2], Qwen2.5-VL [62], and InternVL3 [85]. The results reveal a clear pattern: models that perform strongly on single-view tasks experience a significant drop when confronted with cross-view challenges. Further analysis indicates that even models equipped with explicit reasoning capabilities struggle to interleave text-based reasoning with information from dual-view videos. These results highlight a significant gap between current capabilities and the human-like cross-view understanding required for embodied agents and human–robot collaboration. We hope EgoExoBench will serve as a comprehensive evaluation suite for ego–exo view reasoning and inspire new architectures and training strategies to bridge this gap.

2 Related Work

Video benchmarks. With the rapid application of Multimodal Large Language Models (MLLMs) into video understanding [37, 81, 48], many efforts have been made to benchmark the video understanding capabilities of MLLMs. Benchmarks such as Next-QA [69], Next-GQA [70], MLVU [84], LV-Bench [64], Video-MME [14], and CG-Bench [8] evaluate multiple aspects from general QA to multimodal understanding. Egocentric datasets [10, 16, 42, 51, 4, 11, 31, 80, 60] spur EgoSchema [45] to probe egocentric long-form video understanding, while HourVideo [7] tests spatial intelligence. However, these efforts each focus on a single viewpoint, either egocentric or exocentric, and none assess a model’s ability to integrate both. EgoExoBench fills this gap by offering the first benchmark for cross-perspective video understanding, measuring semantic, spatial, and temporal reasoning between first- and third-person streams.

Ego-exo datasets. Prior work has produced a variety of paired egocentric–exocentric video collections. Charades-Ego [59] and Home Action Genome [52] capture synchronized first- and third-person video in home environments, while Assembly101 [56] and LEMMA [30] offer both synchronous recordings of multi-step tasks. Collections like Ego-Exo4D [17] and EgoExoLearn [27] offer extensive scale and rich annotations, but lack any downstream QA or reasoning benchmarks. Robotics-focused datasets like iGibson [36] and H2O [34] record fixed-camera arrays for pose estimation and navigation, yet remain narrowly scoped to those tasks. Built atop these datasets, EgoExoBench provides a unified testbed for evaluating cross-view video understanding and reasoning of MLLMs.

Ego-exo video understanding. Egocentric video understanding methods often leverage larger-scale exocentric data [46, 18, 61] to compensate for the limited size of first-person corpora. Prior work falls into three main categories: joint view-invariant learning [75, 63, 72, 71, 23, 25], domain adaptation [68, 79, 73, 26], and knowledge distillation [39, 50, 22]. While these approaches improve recognition performance, they mainly focus on recognition problems and do not evaluate a model’s ability to reason across perspectives, which is the aspect EgoExoBench is specifically designed to benchmark.

3 Benchmark

3.1 Task Suite

EgoExoBench is a large-scale benchmark for evaluating cross-view video understanding in multimodal large language models. It covers diverse environments and activities, probing three key

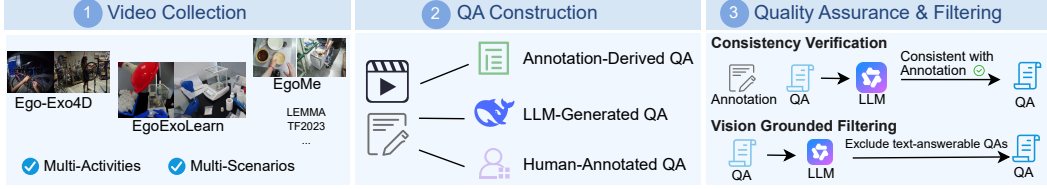


Figure 1: The Construction Pipeline of EgoExoBench.

dimensions of ego–exo reasoning: semantic alignment, spatial correspondence, and temporal reasoning. While open-ended question answering closely mirrors human dialogue, automating the evaluation of free-form responses can be difficult and error-prone. To enable reliable, scalable assessment of cross-view video understanding, EgoExoBench adopts a multiple-choice question (MCQ) format. In the following sections, we detail our task suite and question–answer generation pipeline, both crafted to produce diverse, high-quality multiple-choice questions that rigorously probe semantic alignment, spatial correspondence, and temporal reasoning.

Creating a benchmark for cross-view video understanding presents unique challenges: questions must not only span multiple temporal segments but also require synthesizing information between first- and third-person perspectives. EgoExoBench addresses this by first establishing three core dimensions that capture the essence of ego–exo reasoning: semantic alignment, spatial correspondence, and temporal relation. We then leverage rich, publicly available egocentric–exocentric datasets, mining their synchronized and asynchronous multi-view recordings, spatial annotations (e.g., poses, bounding boxes), and temporal action labels to construct high-quality MCQs.

Our task suite comprises 11 subtasks organized under 3 key dimensions. Ego–Exo Matching evaluates semantic alignment across perspectives; Ego–Exo View Transition measures spatial correspondence between egocentric and exocentric coordinate frames; and Ego–Exo Temporal Reasoning examines sequence integration by requiring models to align and predict event flows across paired video streams. Together, these subtasks form a comprehensive evaluation suite for ego–exo view video understanding and reasoning.

3.2 Data Construction Pipeline

As shown in Figure 1, our QA creation follows a three-stage pipeline tailored to the specific demands of cross-view evaluation. Below, we describe the general construction process. We provide task-specific details in subsequent sections.

Video Collection. Our benchmark aggregates videos from six ego-exo datasets: Ego-Exo4D [17], EgoExoLearn [27], LEMMA [30], EgoMe [49], TF2023, and CVMHAT [19]. These datasets span a wide range of environments (e.g., kitchen, laboratory, sports field) and activities (e.g., cooking, sports, repair). Ego-Exo4D [17], LEMMA [30], TF2023 [82], and CVMHAT [19] include synchronized multi-view videos, while EgoExoLearn [27] and EgoMe [49] feature asynchronous demonstration-follower recordings. This diverse setting allows for a comprehensive evaluation of models’ ability to relate and understand both synchronized and asynchronous cross-view scenarios.

Question-Answer Construction. To accommodate the diverse nature of tasks in our benchmark, we employ multiple QA construction strategies. (1) *Annotation-Derived QA*. For tasks with structured and deterministic annotations, we construct QA pairs directly from the annotations using predefined templates. (2) *LLM-Generated QA*. For cases that require open-ended reasoning in QA generation, we utilize LLMs [40, 76] to generate context-aware QA pairs. The LLMs are provided with task definitions, detailed annotations, exemplar QA pairs, and task-specific constraints. (3) *Human-Annotated QA*. For tasks that require fine-grained spatial understanding, current MLLMs fail to produce reliable results. Therefore, we employ human annotators to construct the corresponding QA pairs. All QA pairs are formatted as 4-way multiple-choice questions.

Quality Assurance and Filtering. To ensure the reliability of the constructed QA pairs, we implement a quality assurance and filtering process. (1) *Consistency Verification*. We perform an automated consistency check by prompting an LLM [76] to verify the logical correctness of each question-answer pair against the original video annotations. Questions are discarded if they are

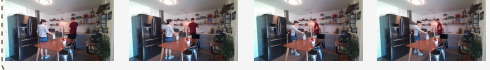
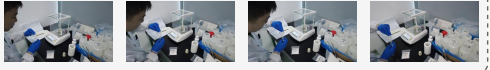
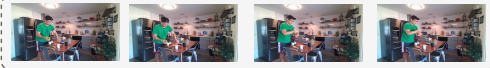
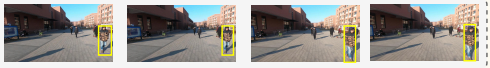
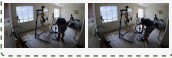
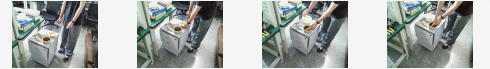
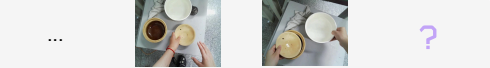
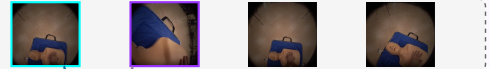
Ego-Exo Matching			
Task Matching		Action Matching	
Query Video:  Q: Which candidate video most accurately matches the activities in the query video? A.  B.  C.  D. 		Query Video:  Q: Which candidate video most accurately matches the actions in the query video? A.  B.  C.  D. 	
Object Matching		Person Matching	
Query Video:  Q: Which candidate video most accurately matches the interacted object in the query video? A.  B.  C.  D. 		Query Video:  Q: Which candidate video most accurately matches the person (within the bbox) in the query video? A.  B.  C.  D. 	
Ego-Exo View Transition			
Egocentric Wearer Identification	Body Part Action Understanding		Direction Prediction
Egocentric Video:  Q: Which boxed person in the third-person view is the camera wearer who captured the egocentric video? A.  B.  C.  D. 	Video1:  Video2:  Q: What is the difference in the actions involving the person's feet in the two videos? A. In Video 1, the person lifts left foot... B. In Video 1, the person steps on the solid base with left foot; in Video 2, the person does it with right foot. C. In Video 1, the person places right foot on the pedal... D. In Video 1, the person steps on the solid base ...		Egocentric Video:  Q: In which view does the person move to the foreground of the scene? A.  B.  C.  D. 
Ego-Exo Temporal Reasoning			
Action Prediction		Sequence Alignment	
Reference Video:  Egocentric Video:  Q: Based on the reference video, what is the next action in the egocentric video? A. The right hand picks up the pot and pours the water into the bowl. B. The left hand holds the lid and cleans it with water in the pot. C. The right hand pours some water from the bowl into the pot. D. The left hand holds the lid and places it on a yellow pot.		Video1:  Video2:  Q: What is the key difference in the action sequences between Video 1 and Video 2? A. Video 1 confirms patient consciousness, while Video 2 does not. B. Video 1 includes chest compressions, while Video 2 does not. C. Video 1 places the hand on top of the first before confirming consciousness, while Video 2 does it after. D. Video 2 includes a conversation before confirming patient...	
Action Ordering		Skill Evaluation	
Video1:  Video2:  Q: What is the correct temporal order between the actions in these two videos? A. Actions in Video 2 occur before actions in Video 1. B. Actions in Video 1 are part of the sequence of actions in Video 2. C. Actions in Video 2 are part of the sequence of actions in Video 1. D. Actions in Video 1 occur before actions in Video 2.		Reference Video:  Q: Which candidate video performs the task best according to the reference video? A.  B.  C.  D. 	

Figure 2: Example MCQs in EgoExoBench. The correct answers are bold-highlighted.

ambiguous, inconsistent with annotations, or admit multiple plausible answers. (2) *Vision-Grounded Filtering*. To guarantee that each question necessitates visual understanding, we present text-only questions to an LLM [76] and discard QA pairs that can be answered solely based on textual input.

3.3 Task-Specific Details

3.3.1 Ego-Exo Matching

The Ego–Exo Matching dimension evaluates a model’s ability to semantically align content across first- and third-person views, which is a critical foundation for any cross-view reasoning. We design four subtasks: task-, action-, person-, and object-level matching, aiming for a comprehensive assessment.

Task Matching (TM). We draw on synchronized and asynchronous demonstrations in Ego-Exo4D [17], EgoExoLearn [27], and LEMMA [30]. We pair egocentric clips of a high-level activity (*e.g.*, assembling a device in a kitchen) with exocentric videos labeled with the same task ID. To prevent trivial scene-matching, ground-truth pairs come from different environments within the same scenario, while negative candidates depict other tasks in visually similar contexts. QA pairs are generated via annotation-derived templates, converting task IDs into four-choice MCQs, and are filtered through automated consistency checks and vision-grounded prompts to ensure reliance on visual input.

Action Matching (AM). Given a video clip, this task aims to identify a corresponding clip from another viewpoint that captures the same action. It requires bridging perspective-induced variations, such as camera motion and occlusion, to establish fine-grained action correspondence. We construct QA pairs using videos from LEMMA [30] and EgoExoLearn [27]. For LEMMA [30], the ground-truth clip is temporally aligned with the query but recorded from a different viewpoint. For EgoExoLearn [27], the correct matching clip is drawn from a different video depicting the same action.

Object Matching (OM). Given a query video, this task aims to identify the candidate video from another viewpoint that involves interaction with the same object. We curate videos from LEMMA [30] and derive object interaction from the annotations. To increase difficulty, no textual cues about the object are given. The model must first infer the object interacted with in each video before matching across views. The correct video is temporally aligned with the query but from another viewpoint, while negative candidates depict interactions with different objects in the same environment.

Person Matching (PM). Given a video clip, this task aims to identify the same individual from a different viewpoint. We curate data from CVMHAT [19], which provides synchronized egocentric and top-view recordings in outdoor scenarios. In each query, the target individual is highlighted using a bounding box. The correct answer corresponds to the same person observed at the same timestamp but from another viewpoint. To prevent shortcuts based on scene or temporal cues, negative candidates are drawn from the same timestamp but depict different individuals.

3.3.2 Ego-Exo View Transition

The subtasks in Ego–Exo View Transition assess a model’s ability to translate spatial information between first- and third-person perspectives. We define three subtasks: egocentric wearer identification, direction prediction, and body part action understanding, each constructed with tailored data sources and QA strategies.

Egocentric Wearer Identification (EWI). Given an egocentric video, this task aims to identify the camera wearer in a third-person view. Compared to the Person Matching task, this setting is more challenging, as the egocentric wearer is typically absent from their own viewpoint. With limited appearance cues, the model must infer spatial relationships by analyzing the relative positions of surrounding people and objects, and then map these relations from the egocentric to the exocentric perspective. We build upon the TF2023 dataset [82], which provides synchronized egocentric–exocentric image pairs and annotated human bounding boxes. We retain only samples containing four or more individuals to increase spatial complexity. To construct MCQ candidates, we select a different person based on the same exocentric frame.

Direction Prediction (DP). Given an egocentric video depicting the motion of a person or an interacted object, the task is to identify, among four provided third-person views, which view shows

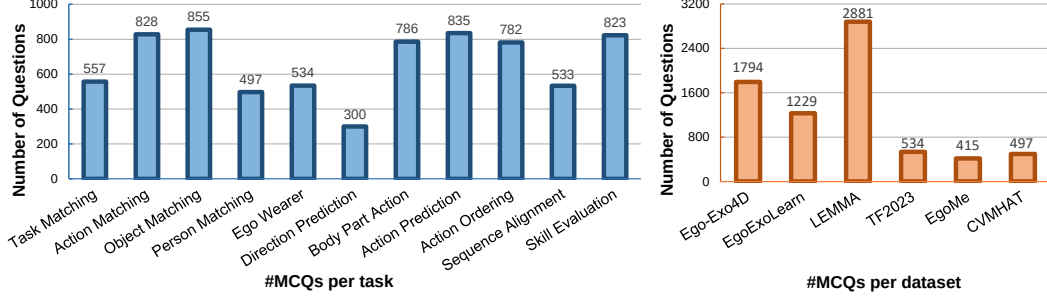


Figure 3: Distribution of MCQs in EgoExoBench. Left: number of questions per subtask. Right: number of questions per dataset.

the motion in a specified direction (e.g., left to right). The direction is specified in the textual question. Using Ego-Exo4D’s [17] synchronized multi-view recordings, we first identify segments with clear entity movements via Qwen2.5 32B-filtered narrations. Expert annotators then label the movement direction of the wearer or an object in the egocentric view. Each question offers multiple exocentric clips, only one of which matches the specified directional condition. Manual verification ensure that only unambiguous, visually grounded examples remain.

Body Part Action Understanding (BPA). This task probes fine-grained spatial mapping of limb movements across the two views. We extract segments from Ego-Exo4D [17] and EgoExoLearn [27] where narrations explicitly reference body parts such as the left or right hand. Qwen2.5-32B [76] identifies associated verbs and objects, and DeepSeek-V3 [40] generates distractors that share at least one action or object cue. We generate questions asking to pair egocentric and exocentric clips showing the same body-part activity. Finally, we apply a Qwen2.5-32B-based filter that discards any instances solvable via text alone.

3.3.3 Ego-Exo Temporal Reasoning

The Ego-Exo Temporal Reasoning evaluates a model’s ability to align and infer the flow of events across egocentric and exocentric video streams. We define four subtasks—next-action prediction, action ordering, sequence alignment, and skill evaluation.

Action Prediction (AP). We pair a partial egocentric clip with a longer exocentric demonstration and ask models to forecast the subsequent egocentric action. To construct data, we collect videos from LEMMA [30] and EgoMe [49]. For LEMMA [30], we utilize the action annotations to extract ground-truth next actions and negative candidates. Specifically, we select ten actions surrounding the current action in the egocentric sequence, excluding the true next action, and prompt Qwen2.5-32B [76] to identify the three most plausible distractors from this set. In asynchronous settings from EgoMe [49], we rely on annotated action descriptions from the egocentric view and use DeepSeek-V3 [40] to generate three distractors accordingly. Finally, we use Qwen2.5-32B [76] to filter out questions that can be answered based solely on textual input.

Action Ordering (AO). In this task, the model is given two short clips, one egocentric, one exocentric, and is required to judge their temporal relationship. To construct data, we begin with raw action annotations from LEMMA [30]. Each question instance is formed by selecting two temporally continuous action clips. To ensure the validity of temporal ordering, we prompt Qwen2.5-32B [76] with textual descriptions of the actions to filter out ambiguous or unordered action pairs.

Sequence Alignment (SA). Sequence alignment extends ordering to multi-step activities by asking whether an egocentric and an exocentric video share the same action ordering, are reversed, or differ at key steps. To construct the data, we first curate videos from Ego-Exo4D [17] that include keystep annotations. Video pairs are selected from the same activity category, with each pair containing at least two shared steps. For each pair, we organize the step descriptions and use DeepSeek-V3 [40] to generate questions that emphasize key differences in action sequences. To ensure question quality, we employ Qwen2.5-32B [76] to verify that the correct answer aligns with the annotated steps. Finally, we used Qwen2.5-32B [76] to filter out questions that could be answered with only textual input.

Skill Evaluation (SE). Given an expert demonstration video as reference, this task assesses the skill level of other videos from different viewpoints. It is formulated as a multiple-choice question where the model selects the best or worst performance among candidates. Compared to textual guidelines, expert demonstrations offer fine-grained behavioral cues that are often missed by rule-based methods. We construct data from EgoExoLearn [27] and Ego-Exo4D [17]. For EgoExoLearn [27], we use exocentric demonstration videos as references and build transitive ranking chains from pairwise skill annotations to establish a clear hierarchy among candidate videos. For Ego-Exo4D [17], we select "Late Expert" videos as references and curate candidates with varying proficiency scores from the same activity. Egocentric videos with poor visibility of actions are manually excluded.

3.4 Benchmark Statistics

EgoExoBench comprises a total of 7,330 multiple-choice questions (MCQs), each designed in a 4-way format. These questions are constructed from six ego-exo paired datasets, encompassing a wide range of scenarios and activities. As illustrated in Figure 2, the benchmark comprises three task categories encompassing a total of eleven subtasks. Figure 3 presents the distribution of MCQs across subtasks and datasets.

4 Experiment

4.1 Experiment Setup

We evaluate EgoExoBench in a zero-shot multiple-choice question answering (MCQ) format using a diverse set of MLLMs. Closed-source models include Claude 3.7 Sonnet [2], GPT-4o [29], and GPT-o4-mini [47]. For open-source models, we assess Qwen2.5-VL [62], InternVL3 [85], LLaVA-OV [35], LLaVA-Video [81], NVILA [43], and EgoGPT [78], covering a diverse set of architectures and parameter scales. Each model receives the same standardized prompt, which presents the question stem followed by labeled options and instructs the model to return only the letter of its chosen answer. We apply a rule-based approach to extract the predicted answer. All evaluations use accuracy as the primary metric, and no fine-tuning or gradient updates are performed to ensure fairness. For closed-source models, we conduct evaluations via official APIs. For open-source models, all experiments are performed using four A100 GPUs with 80GB memory each.

4.2 Main Result

Prior to the main experiments, we first confirmed that MLLMs reliably distinguish multiple video streams as separate inputs, validating our multi-video experimental setup (Details in supplementary).

Table 1 summarizes the performance of both open-source and closed-source models on EgoExoBench. Among open-source models, Qwen2.5-VL-72B [62] achieves the highest overall accuracy at 47.0%, GPT-o4-mini [47] leads the closed-source group with 48.0%. In the Ego-Exo Matching category, GPT-o4-mini [47] outperforms Qwen2.5-VL-72B [62] by 5% on average, indicating that open-source architectures still lag behind state-of-the-art closed-source systems in explicit semantic alignment across views. For Ego-Exo View Transition and Ego-Exo Temporal Reasoning, Qwen2.5-VL-72B [62] achieves average accuracies of 47.3% and 37.4%, respectively. Its relative strength in spatial and video grounding likely contributes to better performance on these spatial correspondence and sequence integration tasks. Notably, EgoGPT [78], despite being fine-tuned on a large egocentric video corpus (EgoIT-99K) [78], offers only marginal gains over similarly sized counterparts, suggesting that specialized pretraining alone is insufficient to master cross-perspective reasoning without task-specific objectives.

Human Performance. To contextualize model performance, we randomly sample 30 questions from each subtask, resulting in 330 questions, and measure human accuracy on the same MCQs. Human evaluators are instructed to derive their answers within 1 minute (quick responses) and 3 minutes (deliberate responses), respectively. As shown in Table 1, humans performing under the deliberate condition (3 minutes per question) achieve an average accuracy of 90.1%, outperforming the best model by 42%. This gap indicates that MLLMs still fall significantly short of human-level in cross-view video understanding. The largest gap appears in the Egocentric Wearer Identification (EWI) subtask, where all human evaluators can successfully infer the identity of the egocentric camera wearer in third-person views based on spatial relationships between people and objects. In contrast,

Table 1: Performance of various open-source and closed-source MLLMs on EgoExoBench. The best result among all models is highlighted in bold, and the second-best is underlined.

	Avg.	Matching				View Transition			Temporal Reasoning			
		TM	AM	OM	PM	EWI	DP	BPA	AP	AO	SA	SE
Human Performance (330 Questions)												
Human (1min)	64.6	72.2	58.9	62.2	70.0	91.1	78.9	53.3	68.9	61.1	60.0	34.4
Human (3min)	90.1	95.5	90.0	92.2	94.4	100.0	96.7	85.5	91.1	87.8	88.9	68.9
Qwen2.5-VL-72B	48.5	56.7	46.7	66.7	56.7	56.7	43.3	46.7	46.7	43.3	43.3	26.7
Claude-3.7-Sonnet	32.8	47.1	30.0	36.7	46.7	33.3	30.0	26.7	20.0	30.0	30.0	30.0
GPT-4o	38.5	53.8	30.0	60.0	56.7	33.3	26.7	30.0	26.7	33.3	43.3	30.0
Open-source MLLMs												
Qwen2.5-VL-7B	32.8	40.2	34.4	45.5	36.0	26.4	30.7	34.5	19.2	31.0	37.1	26.1
Qwen2.5-VL-32B	39.7	43.3	40.1	50.4	42.3	41.7	34.0	42.4	38.6	31.1	45.8	27.5
Qwen2.5-VL-72B	44.7	51.0	43.5	56.6	49.7	56.7	37.0	48.1	39.9	33.6	46.1	29.9
InternVL3-8B	31.3	36.8	30.9	37.0	27.2	16.5	33.7	37.2	27.4	29.8	47.2	20.9
InternVL3-14B	35.1	38.8	31.9	43.5	31.8	30.5	29.0	41.1	33.9	32.9	48.1	24.2
InternVL3-78B	41.4	50.6	37.3	48.9	39.0	46.9	32.3	38.0	51.0	31.5	50.6	29.5
LLaVA-OV-7B	29.5	30.7	28.7	34.0	28.0	22.7	27.0	31.4	21.7	29.3	44.1	26.8
LLaVA-Video-7B	31.2	33.9	29.1	35.4	27.0	29.0	27.0	36.4	23.5	28.9	43.8	28.9
NVILA-8B	29.6	30.0	25.0	30.8	31.6	23.6	27.0	37.7	24.4	26.2	44.5	24.8
EgoGPT-7B	29.6	29.8	29.5	35.6	30.4	22.3	26.3	32.1	22.3	29.4	40.1	27.6
Closed-source MLLMs												
Claude-3.7-Sonnet	31.3	33.4	33.0	35.1	30.6	37.8	30.0	33.3	26.0	28.5	30.2	26.4
GPT-4o	38.5	52.9	44.6	57.4	48.0	41.6	24.1	37.1	27.7	26.1	37.4	26.3
GPT-o4-mini	48.0	65.3	45.2	56.8	53.7	73.8	30.0	40.0	46.4	35.9	51.5	30.1

current MLLMs struggle to reason about such spatial configurations across perspectives. The smallest discrepancy arises in Skill Evaluation, where even humans find domain-specific assessments (*e.g.*, judging basketball proficiency) challenging. This suggests that effective cross-view skill assessment requires MLLMs to integrate both domain knowledge and multi-view reasoning.

4.3 Can Reasoning Improve MLLM’s Performance?

Prompting techniques [32, 67] have been shown to enhance the performance of MLLMs on various reasoning tasks [67, 83, 15, 55]. To examine whether similar reasoning prompts can improve the cross-perspective understanding, we sample 100 questions from each subtask and compare baseline zero-shot MCQ performance against CoT-augmented prompts. Following [32, 65], we append “Let’s think step by step” to each prompt and evaluate four models representing diverse architectures and scales: LLaVA-OV-7B [35], Qwen2.5-VL-32B [76], Qwen2.5-VL-72B [62], and GPT-4o [29].

Figure 4 shows the accuracy difference ($Acc_{CoT} - Acc_{baseline}$) across all subtasks. CoT prompting degrades performance on most tasks. In particular, for the Person Matching and Action Matching subtasks, Qwen2.5-VL-72B [62], which already performs well without CoT, suffers a substantial drop in accuracy (20% and 19%). One reason for the performance drop is that the tasks in EgoExoBench demand that a model alternate between interpreting one video stream, translating that understanding to language, and then applying it to a second stream. Standard CoT cues, which focus purely on textual decomposition, disrupt this interleaved visual–linguistic workflow, leading to degraded performance.

Figure 5 provides a concrete example from Egocentric Wearer Identification with GPT-4o [29] under CoT prompting. The model correctly deduces from the first-person clip that the wearer stands next to someone in a blue jacket (highlighted in green). However, during the subsequent reasoning step, it misidentifies the person in the blue jacket as the camera wearer, rather than someone nearby (highlighted in red). This failure illustrates how purely text-oriented CoT reasoning can break the continuity of cross-view inference, underscoring the need for new prompting or architectural techniques that integrate visual and linguistic reasoning in tandem.

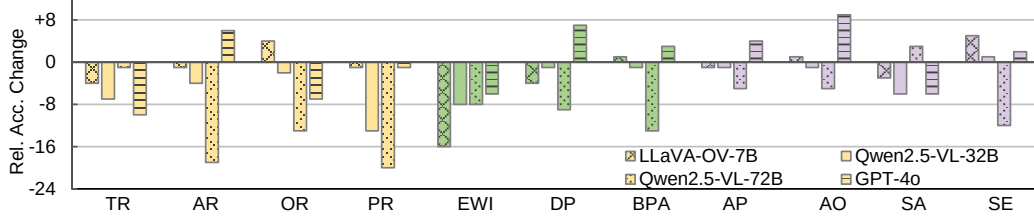


Figure 4: Relative performance changes after applying CoT. On average, CoT leads to a performance drop across tasks. This suggests that EgoExoBench cannot be effectively addressed by linguistic reasoning alone.

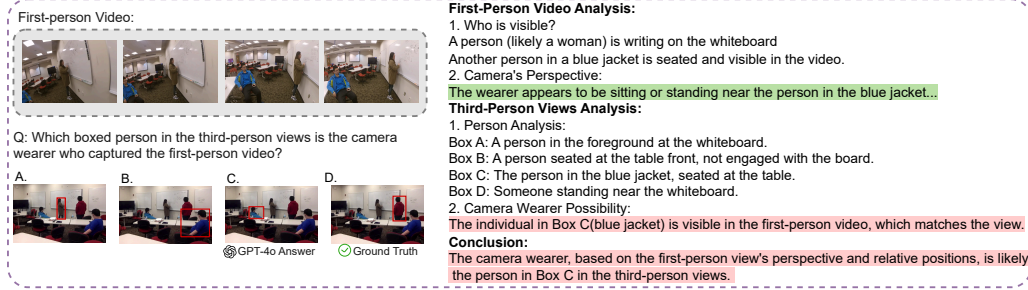


Figure 5: A failure case of GPT-4o with CoT prompting. The model fails to maintain consistency across cross-perspective visual-linguistic reasoning.

4.4 Can MLLMs Leverage Cross-Perspective Guidance?

To determine whether MLLMs can leverage additional cross-view context, we augment the Action Prediction and Skill Evaluation tasks with an additional video from a different perspective as a reference. For each question, the model receives both the primary clip (*e.g.*, egocentric for Action Prediction) and the reference clip (*e.g.*, exocentric), then selects the answer. Meanwhile, we perform an ablation by removing the reference clip and measuring the change in accuracy. The results are presented in Table 2. In the Action Prediction task, providing the reference video consistently boosts performance. For instance, Qwen2.5-VL [62] achieves a 8.2% gain. This improvement confirms the practical value of cross-perspective cues for better understanding actions. In contrast, results on the Skill Evaluation task are less consistent. While GPT-4o [29] and InternVL3 [85] show small performance differences, Qwen2.5-VL [62] experiences a 1.5% drop when the reference video is included. We hypothesize that this task’s reliance on domain-specific expertise and subtle quality judgments, such as assessing proficiency in sports, outweighs the benefits of extra visual context. Under current model capabilities, cross-view input alone may be insufficient to enhance fine-grained skill assessment without deeper incorporation of domain knowledge.

Table 2: Ablation study on the use of reference videos.

	Action Prediction		Skill Evaluation	
	w.o./ref	w./ref	w.o./ref	w./ref
Qwen2.5-VL-72B	31.7	39.9 (+8.2)	31.4	29.9 (-1.5)
InternVL3-78B	41.9	51.0 (+9.1)	30.4	29.5 (-0.9)
GPT-4o	26.1	27.7 (+1.1)	26.5	26.3 (-0.2)

4.5 What Makes the Egocentric Viewpoint Special?

To examine the uniqueness of egocentric viewpoints, we conduct exo–exo ablation studies across multiple tasks. In each task, the egocentric video in a QA pair is replaced with its synchronized third-person counterpart, enabling direct comparison between egocentric and exocentric settings. We use videos from LEMMA and EgoExo4D, which provide temporally aligned ego–exo recordings.

Table 3: Ego–Exo ablation study on ego–exo matching, with results grouped by scenario type.

Model	View	Cooking	Household	Entertainment	Sport	Assembly
Qwen2.5-VL-7B	Ego–Exo	47	41	46	27	25
	Exo–Exo	49 (+2)	33 (-8)	46 (0)	27 (0)	40 (+15)
Qwen2.5-VL-72B	Ego–Exo	61	68	54	30	45
	Exo–Exo	58 (-3)	56 (-12)	46 (-8)	36 (+6)	60 (+15)
InternVL3-14B	Ego–Exo	47	44	42	26	35
	Exo–Exo	44 (-3)	41 (-3)	50 (+8)	34 (+8)	43 (+8)
InternVL3-78B	Ego–Exo	53	65	59	22	36
	Exo–Exo	52 (-1)	51 (-14)	48 (-11)	30 (+8)	48 (+12)
LLaVA-Video-7B	Ego–Exo	37	48	30	22	20
	Exo–Exo	36 (-1)	33 (-15)	26 (-4)	28 (+6)	30 (+10)
LLaVA-OV-7B	Ego–Exo	36	31	35	26	23
	Exo–Exo	35 (-1)	27 (-4)	26 (-9)	29 (+3)	33 (+10)

Table 4: Ego–Exo ablation study on BPA, AO, SA, and SE tasks.

Model	BPA		AO		SA		SE	
	Ego–Exo	Exo–Exo	Ego–Exo	Exo–Exo	Ego–Exo	Exo–Exo	Ego–Exo	Exo–Exo
Qwen2.5-VL-7B	29	37 (+8)	26	28 (+2)	37	39 (+2)	27	32 (+5)
InternVL3-14B	47	48 (+1)	32	52 (+20)	47	48 (+1)	29	37 (+8)
LLaVA-Video-7B	33	35 (+2)	28	37 (+9)	44	45 (+1)	29	30 (+1)
Qwen2.5-VL-72B	48	58 (+10)	34	48 (+14)	46	48 (+2)	40	39 (-1)
InternVL3-78B	38	51 (+13)	32	46 (+14)	52	54 (+2)	30	45 (+15)
LLaVA-OV-7B	25	25 (0)	29	32 (+3)	45	46 (+1)	27	27 (0)

Table 3 reports the results for the Ego–Exo Matching category. We group the results by scene type. Egocentric views exhibit advantages in Cooking, Household, and Entertainment scenarios, where fine-grained hand–object interactions and manipulation details are best observed from the actor’s perspective. In contrast, exocentric views outperform in Sports and Assembly scenarios, which benefit from stable, wide-angle coverage of full-body motion and the overall workspace. We further extend this analysis to other tasks, as summarized in Table 4. Egocentric views are less effective for tasks that require global motion cues or temporal consistency(e.g., AO, SA, SE) due to issues like motion blur and limited field of view. These gaps underscore the need for improved egocentric representation learning and temporal modeling to enhance cross-view understanding in MLLMs.

5 Conclusion

EgoExoBench is the first large-scale benchmark explicitly designed to evaluate cross-view video understanding in multimodal LLMs. By aggregating paired egocentric–exocentric recordings and crafting over 7,300 high-quality multiple-choice questions across eleven subtasks, EgoExoBench probes three fundamental dimensions of ego–exo reasoning: semantic alignment, spatial correspondence, and sequence integration. Our extensive evaluation of both open- and closed-source models reveals that, despite strong single-view performance, current MLLMs struggle to bridge perspectives. Chain-of-thought prompting and additional cross-perspective guidance offer limited improvements, underscoring the need for novel architectures and training paradigms that can interleave visual and linguistic inference across multiple viewpoints. While the benchmark spans diverse tasks, it may not fully reflect the breadth of real-world ego–exo scenarios, which we leave for future work. We hope EgoExoBench will serve as a valuable resource to spur research on embodied agents and collaborative systems that require human-like cross-view intelligence.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (Grant No.62372223 and U24A20330), and in part by Nanjing University-China Mobile Communications Group Co., Ltd. Joint Institute under Grant NJ20250037, the National Key R&D Program of China (2022ZD0160201), JSPS KAKENHI JP25K24384, and Shanghai Artificial Intelligence Laboratory.

References

- [1] Cihan Acar, Kuluhan Binici, Alp Tekirdağ, and Yan Wu. Visual-policy learning through multi-camera view to single-camera view knowledge distillation for robot manipulation tasks. *IEEE Robotics and Automation Letters*, 9(1):691–698, 2023.
- [2] Anthropic. Claude 3.7 sonnet and claude code, 2025. URL <https://www.anthropic.com/news/claude-3-7-sonnet>.
- [3] Albert Bandura. Observational learning. *The international encyclopedia of communication*, 2008.
- [4] Siddhant Bansal, Chetan Arora, and C.V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In *Eur. Conf. Comput. Vis.*, 2022.
- [5] Cristina Becchio, Marco Del Giudice, Olga Dal Monte, Luca Latini-Corazzini, and Lorenzo Pia. In your place: neuropsychological evidence for altercentric remapping in embodied perspective taking. *Social cognitive and affective neuroscience*, 8(2):165–170, 2013.
- [6] Neil Burgess. Spatial memory: how egocentric and allocentric combine. *Trends in cognitive sciences*, 10(12):551–557, 2006.
- [7] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.
- [8] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024.
- [9] Massimiliano Conson, Elisabetta Mazzarella, Dalila Esposito, Dario Grossi, Nicoletta Marino, Angelo Massagli, and Alessandro Froli. “put myself into your place”: Embodied simulation and perspective taking in autism spectrum disorders. *Autism Research*, 8(4):454–466, 2015.
- [10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.
- [11] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022.
- [12] DeepMind. Gemini 2.5 pro, 2025. URL <https://deepmind.google/models/gemini/pro/>.
- [13] Arne D Ekstrom, Derek J Huffman, and Michael Starrett. Interacting networks of brain regions underlie human spatial navigation: a review and novel synthesis of the literature. *Journal of neurophysiology*, 118(6):3328–3344, 2017.
- [14] Chaoyou Fu, Yuhan Dai, Yondong 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.
- [15] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022.
- [16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.
- [17] 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.
- [18] Chunhui Gu et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6047–6056, 2018. doi: 10.1109/CVPR.2018.00633.

- [19] Ruize Han, Wei Feng, Yujun Zhang, Jiewen Zhao, and Song Wang. Multiple human association and tracking from egocentric and complementary top views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5225–5242, 2022. doi: 10.1109/TPAMI.2021.3070562.
- [20] Nicola J Hodges, A Mark Williams, Spencer J Hayes, and Gavin Breslin. What is modelled during observational learning? *Journal of sports sciences*, 25(5):531–545, 2007.
- [21] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *European Conference on Computer Vision*, 2018.
- [22] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.
- [23] Yifei Huang, Minjie Cai, and Yoichi Sato. An ego-vision system for discovering human joint attention. *IEEE Transactions on Human-Machine Systems*, 50(4):306–316, 2020.
- [24] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14024–14034, 2020.
- [25] Yifei Huang, Lijin Yang, and Yoichi Sato. Compound prototype matching for few-shot action recognition. In *European conference on computer vision*, pages 351–368. Springer, 2022.
- [26] Yifei Huang, Lijin Yang, and Yoichi Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18908–18918, 2023.
- [27] 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.
- [28] Yifei Huang, Jilan Xu, Baoqi Pei, Lijin Yang, Mingfang Zhang, Yuping He, Guo Chen, Xinyuan Chen, Yaohui Wang, Zheng Nie, et al. Vinci: A real-time smart assistant based on egocentric vision-language model for portable devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–33, 2025.
- [29] 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.
- [30] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *European Conference on Computer Vision*, pages 767–786. Springer, 2020.
- [31] 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.
- [32] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [33] Lukas Kunz, Armin Brandt, Peter C Reinacher, Bernhard P Staresina, Eric T Reifensstein, Christoph T Weidemann, Nora A Herweg, Ansh Patel, Melina Tsitsiklis, Richard Kempter, et al. A neural code for egocentric spatial maps in the human medial temporal lobe. *Neuron*, 109(17):2781–2796, 2021.
- [34] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Int. Conf. Comput. Vis.*, 2021.
- [35] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [36] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021.
- [37] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

- [38] 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.
- [39] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021.
- [40] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [41] Yueyue Liu, Zhijun Li, Huaping Liu, and Zhen Kan. Skill transfer learning for autonomous robots and human-robot cooperation: A survey. *Robotics and Autonomous Systems*, 128:103515, 2020.
- [42] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022.
- [43] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024.
- [44] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025.
- [45] 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.
- [46] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.
- [47] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 07 2024. URL <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- [48] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023.
- [49] Heqian Qiu, Zhaofeng Shi, Lanxiao Wang, Huiyu Xiong, Xiang Li, and Hongliang Li. Ego-me: Follow me via egocentric view in real world. *arXiv preprint arXiv:2501.19061*, 2025.
- [50] Camillo Quattrocchi, Antonino Furnari, Daniele Di Mauro, Mario Valerio Giuffrida, and Giovanni Maria Farinella. Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs, 2023. *arXiv:2312.02638*.
- [51] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021.
- [52] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Nibbles. Home action genome: Cooperative compositional action understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.
- [53] Richard Ramsey, David M Kaplan, and Emily S Cross. Watch and learn: the cognitive neuroscience of learning from others’ actions. *Trends in neurosciences*, 44(6):478–491, 2021.
- [54] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <https://arxiv.org/abs/1908.10084>.
- [55] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

- [56] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.
- [57] Jinghuan Shang and Michael S Ryoo. Self-supervised disentangled representation learning for third-person imitation learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 214–221. IEEE, 2021.
- [58] Pratyusha Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. *Advances in Neural Information Processing Systems*, 32, 2019.
- [59] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018.
- [60] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Adv. Neural Inform. Process. Syst.*, 2024.
- [61] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. *arXiv:1212.0402*.
- [62] Qwen Team. Qwen2.5-vl, January 2025. URL <https://qwenlm.github.io/blog/qwen2.5-vl/>.
- [63] Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Int. Conf. Comput. Vis.*, 2023.
- [64] Weihang Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.
- [65] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- [66] Yeping Wang, Gopika Ajaykumar, and Chien-Ming Huang. See what i see: Enabling user-centric robotic assistance using first-person demonstrations. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 639–648, 2020.
- [67] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [68] Haifeng Xia, Pu Wang, and Zhengming Ding. Incomplete multi-view domain adaptation via channel enhancement and knowledge transfer. In *Eur. Conf. Comput. Vis.*, 2022.
- [69] 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.
- [70] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024.
- [71] Boshen Xu, Sipeng Zheng, and Qin Jin. Pov: Prompt-oriented view-agnostic learning for egocentric hand-object interaction in the multi-view world. In *ACM Int. Conf. Multimedia*, pages 2807–2816, 2023.
- [72] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024.
- [73] Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. *arXiv preprint arXiv:2504.11732*, 2025.
- [74] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

- [75] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022.
- [76] 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.
- [77] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024.
- [78] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*, 2025.
- [79] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.
- [80] Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutsu, and Yoichi Sato. Fine-grained affordance annotation for egocentric hand-object interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.
- [81] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.
- [82] Ziwei Zhao, Yuchen Wang, and Chuhua Wang. Fusing personal and environmental cues for identification and segmentation of first-person camera wearers in third-person views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16477–16487, 2024.
- [83] Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*, 2023.
- [84] 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.
- [85] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

A Additional details and benchmark statistics

We first present additional statistics of EgoExoBench. Next, we detail the QA construction process for each task. Finally, we provide detailed information on the human evaluation performance.

A.1 Benchmark Statistics

EgoExoBench is constructed from six publicly available ego–exo datasets and encompasses three task categories, comprising a total of 11 subtasks. Figure 6 summarizes the number of QA pairs contributed by each dataset for every subtask.

Ego-Exo Matching			
Task Matching(557)	Action Matching(828)	Object Matching(855)	Person Matching(497)
LEMMA: 212 EgoExoLearn: 108 Ego-Exo4D: 237	LEMMA: 612 EgoExoLearn: 216	LEMMA: 855	CVMHAT: 497

Ego-Exo View Transition		
Direction Prediction(300)	Egocentric Weaer Identification (534)	Body Part Action Understanding (786)
Ego-Exo4D: 497	TF2023: 534	Ego-Exo4D: 325 EgoExoLearn: 461

Ego-Exo Temporal Reasoning			
Action Prediction(835)	Action Order(782)	Sequence Alignment (533)	Skill Evaluation(823)
LEMMA: 420 EgoMe: 415	LEMMA: 782	Ego-Exo4D: 533	EgoExoLearn: 444 Ego-Exo4D: 379

Figure 6: Overview of QA distribution in EgoExoBench.

A.2 Ego-Exo Matching

Ego–Exo Matching evaluates a model’s ability to associate semantically similar visual content across first- and third-person perspectives. Below, we detail the QA construction process for the task-, action-, person-, and object-level matching subtasks. Examples of QA instances for each subtask are illustrated in Figures 7 and 8.

Task Matching. We curate videos from the LEMMA [30], EgoExoLearn [27], and EgoExo4D [17] datasets. Since LEMMA [30] contains multi-agent scenarios involving multiple concurrent tasks, we include only videos that depict a single task to avoid ambiguity in task identification. For each QA pair, the ground-truth candidate video illustrates the same task as the query video but from a different viewpoint (e.g., egocentric vs. exocentric). To construct negative candidates, we avoid trivial distinctions based solely on environmental differences. In particular, distractor videos are selected from scenes of the same general type as the query video (e.g., all from kitchen environments). This design ensures that the model must rely on task-relevant visual cues rather than background differences. In total, we construct 212, 108, and 237 QA pairs from LEMMA [30], EgoExoLearn [27], and Ego-Exo4D [17], respectively.

Action Matching. We construct QA pairs using videos from the LEMMA [30] and EgoExoLearn [27] datasets. For LEMMA [30], we utilize the provided action annotations. The ground-truth (GT) video is drawn from the same source video as the query, capturing the same temporal segment from a different viewpoint. Negative candidates are selected from different temporal segments of the same video, corresponding to different actions. For EgoExoLearn [27], we use QA pairs from the

association benchmark. The GT video is directly provided. To construct negative candidates, we begin with the 20 candidate videos included in the benchmark. We first filter for candidates that share at least one verb or noun with the query. Then, we use SentenceTransformer [54] to compute the similarity between the query narration and each candidate’s narration. The three most similar candidates (excluding the GT) are selected as negative options. This ensures distractors are relevant, maintaining task difficulty. In total, we construct 612 and 216 QA pairs from LEMMA [30] and EgoExoLearn [27], respectively.

Object Matching. We curate videos from the LEMMA [30] dataset. The interacted object associated with the person is extracted based on action annotations. Similar to the Action Matching subtask, for each query video (e.g., from the egocentric view), the ground-truth candidate is a temporally aligned clip from the corresponding third-person view. Negative candidates are sampled from the same third-person video but correspond to different time segments where the person interacts with different objects.

Person Matching. We construct data from the CVMHAT [19] dataset, which provides synchronized egocentric and top-view recordings across five outdoor scenes. To indicate the target person in each video, we overlay bounding boxes derived from the dataset’s annotations. The query video is an egocentric clip captured by a specific individual, while all candidate videos are sourced from the corresponding top-view frame at the same timestamp. The difference among candidates lies in the identity of the person highlighted by the bounding box.

A.3 Ego–Exo View Transition

The subtasks in Ego–Exo View Transition evaluate a model’s ability to translate spatial information between first- and third-person perspectives. Figure 9 illustrates QA examples from each subtask. Below, we provide more details on the construction of QA pairs for each subtask.

Egocentric Wearer Identification. We adopt an annotation-derived strategy to construct QA pairs. We curate data from TF2023 [82] dataset, which provides synchronized egocentric–exocentric image pairs. Given an egocentric video, the goal is to identify the corresponding wearer in the third-person view. Candidates are constructed following the same method as in the Person Matching subtask. For each option, the target person is marked with a bounding box. To increase spatial complexity, we select only samples that contain four or more individuals in the scene.

Direction Prediction. We adopt a human-annotated strategy to construct the data. The process consists of five main steps: (1) *Video Selection*. We select synchronized multi-view videos from the Ego-Exo4D [17]. (2) *Action Segment Filtering*. We use the atomic descriptions annotations provided in Ego-Exo4D and retain only segments where the action is marked as visible in the egocentric view. We then apply Qwen2.5-32B [76] to determine whether the action description contains directional information (e.g., moving forward). Only segments with explicit directional content are kept. (3) *QA Annotation*. We hire three student annotators to create QA pairs based on the filtered segments. Each annotator accesses synchronized egocentric and third-person videos and is instructed to focus on the movement direction of the actor or the object being interacted with. If the action is unclear in the egocentric view or the movement direction cannot be judged from any third-person view, the sample is discarded. To construct question, the annotator selects a movement direction observed in one third-person view and writes a question referring to that direction. Movement descriptions are written in free-form natural language. Each question is designed to have exactly one correct answer among the candidates. (4) *Question Polishing*. We use Qwen2.5-32B [76] to refine the questions for clarity and consistency. (5) *Quality Assurance*. We hire an additional student, independent from the annotation process, to review each QA pair. The reviewer ensures that each question has only one correct answer, and the described direction accurately corresponds to the correct video.

Body Part Action Understanding. We adopt an LLM-based approach to construct QA pairs. Each question is derived from a pair of videos that share at least one verb or object. We input the textual descriptions of both videos into DeepSeek-V3 [40] and prompt it to generate a question that compares the actions, with an explicit focus on body parts involved. The prompting strategy is shown in Figure 10. To ensure the questions require visual information, we further use Qwen2.5-32B [76] to filter out those that can be answered without visual input. The filtering prompt is shown in Figure 11.

Ego-Exo Matching

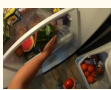
Task Matching

Query Video:



Question:
Which candidate video most accurately reflects the activities observed in the query video?

Candidate Videos:

Video1: 

Video2: 

Video3: 

Video4: 

Query Video:



Question:
Which candidate video most accurately reflects the activities observed in the query video?

Candidate Videos:

Video1: 

Video2: 

Video3: 

Video4: 

Query Video:



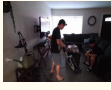
Question:
Which candidate video most accurately reflects the activities observed in the query video?

Candidate Videos:

Video1: 

Video2: 

Video3: 

Video4: 

Action Matching

Query Video:



Question:
Which candidate video most accurately reflects the actions observed in the query video?

Candidate Videos:

Video1: 

Video2: 

Video3: 

Video4: 

Query Video:



Question:
Which candidate video most accurately reflects the actions observed in the query video?

Candidate Videos:

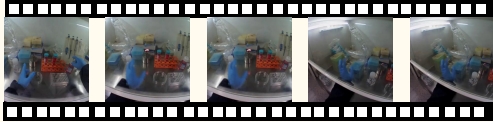
Video1: 

Video2: 

Video3: 

Video4: 

Query Video:



Question:
Which candidate video most accurately reflects the actions observed in the query video?

Candidate Videos:

Video1: 

Video2: 

Video3: 

Video4: 

Figure 7: Examples on Ego-Exo Matching. Correct answers are highlighted in green.

Ego-Exo Matching

Object Matching

Query Video:



Question:
Which candidate video most accurately matches the interacted object seen in the query video?

Candidate Videos:

Video1: 

Video2: 

Video3: 

Video4: 

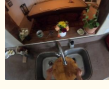
Query Video:



Question:
Which candidate video most accurately matches the interacted object seen in the query video?

Candidate Videos:

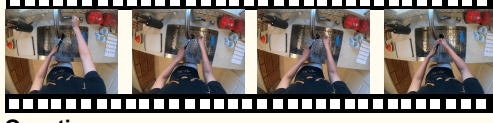
Video1: 

Video2: 

Video3: 

Video4: 

Query Video:

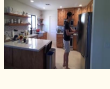


Question:
Which candidate video most accurately matches the interacted object seen in the query video?

Candidate Videos:

Video1: 

Video2: 

Video3: 

Video4: 

Person Matching

Query Video:

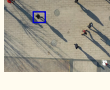


Question:
Which candidate video most accurately matches the person (within the bbox) in the query video?

Candidate Videos:

Video1: 

Video2: 

Video3: 

Video4: 

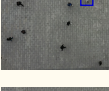
Query Video:



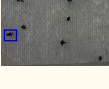
Question:
Which candidate video most accurately matches the person (within the bbox) in the query video?

Candidate Videos:

Video1: 

Video2: 

Video3: 

Video4: 

Query Video:

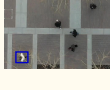


Question:
Which candidate video most accurately matches the person (within the bbox) in the query video?

Candidate Videos:

Video1: 

Video2: 

Video3: 

Video4: 

Figure 8: Examples on Ego-Exo Matching. Correct answers are highlighted in green.

Ego-Exo View Transition
Egocentric Wearer Identification

First-person Video:



Question:
Which boxed person in the third-person views is the camera wearer who captured the first-person video?

Third-person views:

A.



B.



C.



D.



First-person Video:



Question:
Which boxed person in the third-person views is the camera wearer who captured the first-person video?

Third-person views:

A.



B.



C.

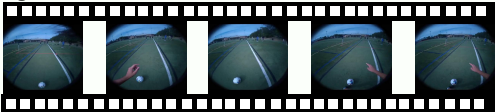


D.



Direction Prediction

EgoVideo:



Question:
In which view does the man run diagonally forward towards the left side of the scene?

Options:

A.



B.



C.



D.



EgoVideo:



Question:
In which view does the person approach the foreground of the scene?

Options:

A.



B.



C.



D.



Body Part Action Understanding

Video1:



Video2:



Options:

A. In Video 1, the right hand transfers an object, while in Video 2, it interacts with the object.

B. In both videos, the right hand remains passive and does not interact with the object.

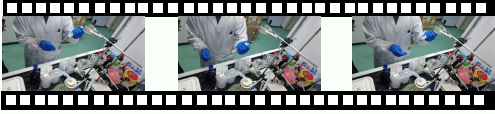
C. In both videos, the right hand is responsible for cutting an object.

D. In Video 1, the right hand stabilizes the object, while in Video 2, it manipulates the object.

Video1:



Video2:



Options:

A. In Video 1, the left hand picks up items and the right hand holds them, while in Video 2, both hands perform actions simultaneously.

B. In Video 1, both hands are used for picking up items, while in Video 2, the right hand holds an item.

C. In Video 1, both hands perform actions simultaneously, while in Video 2, the left hand holds an item and the right hand picks it up.

D. In Video 1, the left hand holds items and the right hand picks them up, while in Video 2, the left hand holds an item and the right hand performs an action.

Figure 9: Examples on Ego–Exo View Transition. Correct answers are highlighted in green.

You will be provided with two text descriptions of different video scenes. Each description involves a person performing actions that engage specific body parts. Based on these descriptions, you are required to generate a set of question-and-answer (QA) pairs.

Requirements:

1. Each QA pair should include one question, four answer options, and one correct answer.
2. The question must clearly address details from both descriptions, particularly focusing on the roles of the body parts, without directly referencing specific actions or objects to prevent giving away hints.
3. Ensure there is one correct answer, and the other three options should be plausible but incorrect to increase the difficulty.
4. Use 'Video 1' to refer Video 1, use 'Video 2' to refer Video 2.
5. The output must be in JSON format as follows:

```
```json
{
 "Question": "The question text goes here",
 "Options": {
 "A": "Option A text",
 "B": "Option B text",
 "C": "Option C text",
 "D": "Option D text"
 },
 "Answer": "Correct option letter (e.g., 'A')"
}
...
```
```

Example Input:

Video 1: I hold the knife with my left hand and press the garlic with my right hand to slice it.

Video 2: Hold the chili pepper with your left hand and cut it into pieces with a knife in your right hand.

Output Example Based on the Above Input:

```
```json
{
 "Question": "How do the tasks performed by the left and right hands differ in the two videos?",
 "Options": {
 "A": "In Video 1, the left hand holds the knife and the right hand slices; in Video 2, the left hand holds the chili and the right hand cuts.",
 "B": "In Video 1, the left hand presses the garlic and the right hand holds the knife; in Video 2, the left hand cuts the chili and the right hand holds the knife.",
 "C": "In Video 1, the left hand holds the knife and the right hand presses the garlic; in Video 2, the left hand holds the knife and the right hand cuts the chili.",
 "D": "In Video 1, the left hand cuts the garlic and the right hand holds the knife; in Video 2, the left hand holds the chili and the right hand cuts."
 },
 "Answer": "A"
}
...
```
```

Video1: [description]

Video2: [description]

Figure 10: Prompt for QA generation in the Body Part Action Understanding subtask.

Carefully read the question and all the options, and based on the information provided, select and provide the most likely correct answer.

[Question]

[Options]

The answer is:

Figure 11: Prompt used to filter out QA pairs that can be correctly answered using text-only input.

A.4 Ego-Exo Temporal Reasoning

Ego-Exo Temporal Reasoning evaluates a model’s ability to align and infer event sequences across egocentric and exocentric video streams. Figures 16 and 17 present example QA pairs for each subtask. Below, we provide further details on the QA construction process for each subtask.

Action Prediction. We adopt an annotation-derived strategy to construct QA data from the LEMMA [30] and EgoMe [49] datasets. For LEMMA [30], we utilize the provided action annotations. To generate candidate answers, ten actions surrounding the current one in the egocentric sequence are selected, excluding the true next action. These are then input into Qwen2.5-32B [76], which selects the three most plausible distractors. The prompting strategy is illustrated in Figure 12. For EgoMe [49], we use fine-grained step annotations. The previous, current, and next actions in the egocentric sequence are given to DeepSeek-V3 [40] to generate three distractors. The corresponding prompt is shown in Figure 13. To ensure data quality, we retain only samples where the current action appears uniquely in the sequence, eliminating ambiguity in next-step prediction. Additionally, Qwen2.5-32B [76] is used to confirm that the distractors do not include the correct answer. Following the approach in the Body Part Action Understanding task, we also apply Qwen2.5-32B [76] to remove any QA pairs that can be answered correctly using text alone.

For the current action, which of the following actions is most likely to be the next action?
Current action: [action name]
Actions: [action names], [action names], ...
The answer is:

Figure 12: Prompt for generating negative options for Action Prediction questions constructed from the LEMMA dataset.

Action Order. We adopt an annotation-derived strategy to construct QA data based on the LEMMA [30] dataset. Each question instance consists of a pair of temporally adjacent action clips: one from the exocentric view and the other from the corresponding egocentric view. To ensure valid temporal relationships, we prompt Qwen2.5-32B [76] with the descriptions of the two actions to filter out pairs that do not exhibit a plausible temporal order. The prompting strategy is shown in Figure 14.

Sequence Alignment. We adopt an LLM-based strategy to construct QA data based on the Ego-Exo4D [17] dataset. For each video pair, we prompt DeepSeek-V3 [40] with the corresponding keystep annotations to generate QAs that highlight key differences in the action sequences. The prompting strategy is illustrated in Figure 15. For quality control, we use Qwen2.5-32B [76] to verify each option against the keystep annotations and discard QAs with multiple valid answers. Finally, as in the Action Prediction subtask, we filter out QAs that can be correctly answered using text input alone.

Skill Evaluation. We adopt an annotation-derived strategy to construct QA data using the EgoExoLearn [27] and Ego-Exo4D [17] datasets. For EgoExoLearn [27], we use exocentric demonstration videos as references. Then, we construct transitive ranking chains from pairwise skill annotations to derive candidate videos. For Ego-Exo4D [17], participant proficiency is categorized into Novice, Early Expert, Intermediate Expert, and Late Expert. We select Late Expert videos as references and sample candidates with varying proficiency levels from the same activity. In total, we curate 444 and 379 QA pairs from EgoExoLearn [27] and Ego-Exo4D [17], respectively.

A.5 Mitigation of Background Bias.

We implement strategies to minimize background interference in our benchmark. All questions and candidate videos are drawn from the same scene types, preventing models from relying on superficial background cues. For distinguishing tasks, we introduce stricter, task-specific controls, summarized in Table 5. Overall, these measures ensure that each task evaluates the intended skill rather than extraneous visual features.

Task Description:

You are tasked with generating multiple-choice distractor options (negative options) for a next-action prediction task. The goal is to create plausible and reasonable options that could be mistaken for the correct next action. Each distractor must adhere to the following criteria:

Logical Consistency: All options must align with the sequence of actions described in the "previous" and "current" actions. They should follow naturally from the context provided.

Reasonableness: Each option must represent a potential next action that could reasonably occur in the given scenario. Avoid introducing actions that are unrelated, overly complex, or implausible.

Non-redundancy: Distractors must not duplicate the correct answer or each other. Each option should introduce a distinct possibility for the next action.

No Obvious Errors: Ensure that none of the distractors contain logical flaws, contradictions, or actions that clearly violate the context.

Input Format:

You will receive the following inputs:

Previous Actions: A description of the actions that occurred before the current moment.

Current Action: A description of the action currently being performed.

Correct Next Action: The correct next action in the sequence (this will not be included in your output).

Output Format:

Generate 3 plausible distractor options that meet the criteria above. Each option should be concise and clearly describe a potential next action.

Example Input:

Previous Actions: "On the white table in front, there are two silver mice and a screwdriver. The left hand picks up the mouse on the left, and the right hand picks up the screwdriver on the right."

Current Action: "The right hand turns the screwdriver to loosen the screws on the mouse."

Correct Next Action: "The right hand rotates the screwdriver to tighten the screws on the mouse."

Example Output:

"The left hand steadies the mouse while the right hand inspects the loosened screws."

"The person pauses to examine the screwdriver for any signs of damage before continuing."

"The right hand uses the screwdriver to start adjusting another screw on the mouse."

Previous Actions: [action name], [action name], ...

Current Action: [action name]

Correct Next Action: [action name]

Figure 13: Prompt for generating negative options for Action Prediction questions constructed from the EgoMe dataset.

You are an assistant that understands the logical sequence of everyday actions. Please determine whether there is a clear cause-effect or sequential dependency between the following two actions:

Action1: [action name]

Action2: [action name]

Rules:

Respond with "True" if one action must precede the other.

Respond with "False" if the actions can be performed in any order or independently.

The answer is:

Figure 14: Prompt for filtering action pairs in the Action Order subtask to ensure plausible temporal relationships.

You are a professional test designer tasked with creating a multiple-choice question based on the following two video action sequences.

Design Requirements:

Question: Clearly focus on the differences in the order of actions or the presence/absence of specific actions between Video 1 and Video 2.

Options: Provide four logically clear options (A, B, C, D) directly sourced from the given action sequences. Each option must closely address the action sequences or specific actions within the two videos.

Correct Option: This option should precisely identify one specific difference between the two videos, whether it's about the sequence of actions or the presence/absence of an action.

Incorrect Options: These options must contradict the video content by:

1. Providing an incorrect sequence of actions;
2. Making false statements about the presence or absence of specific actions;

Accuracy: Avoid vague or overly subjective questions. Do not introduce actions or details that are not present in the provided sequences.

Naming Convention: Refer to the first video as "Video 1" and the second video as "Video 2".

The output must be in JSON format as follows:

```
```json
{
 "Question": "The question text goes here",
 "Options": {
 "A": "Option A text",
 "B": "Option B text",
 "C": "Option C text",
 "D": "Option D text"
 },
 "Answer": "Correct option letter (e.g., 'A')"
}
```

Video 1 Action Sequence: []

Video 2 Action Sequence: []

Figure 15: Prompt for QA generation in the Sequence Alignment subtask.

## A.6 Human Performance Evaluation.

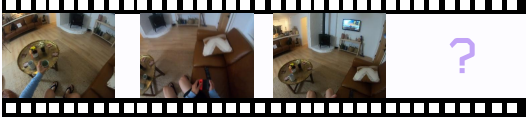
To estimate human-level performance on EgoExoBench, we randomly sample 30 QAs from each subtask, resulting in a total of 330 questions. Two graduate students are invited to complete this subset. To avoid annotation bias, the evaluators do not participate in the data construction process. Each evaluator receives the same input as MLLMs, including the question, options, and corresponding videos. Evaluators are instructed to answer all questions to the best of their ability. They are allowed to pause, replay, and watch the videos multiple times without time constraints. We report the average accuracy of the evaluators as the human performance baseline for this evaluation subset.

**Ego-Exo Temporal Reasoning**  
**Action Prediction**

**Exocentric Video:**



**Egocentric Video:**



**Question:**

Which option most accurately matches the next action expected to occur in the egocentric video?

**Options:**

- A. get controller from sofa with hand
- B. sit on sofa
- C. put controller to sofa with hand
- D. play game-console with controller

**Exocentric Video:**



**Egocentric Video:**



**Question:**

Which option most accurately matches the next action expected to occur in the egocentric video?

**Options:**

- A. Lift the top cover of the bag with the left hand
- B. Put the book into the open bag with the right hand
- C. Pick up a book from the shelf with both hands
- D. Use the left hand to shut the upper cover of the bag

**Skill Evaluation**

**Reference Video:**



**Question:**

Which candidate video performs the task best according to the reference video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

**Reference Video:**



**Question:**

Which candidate video performs the task worst according to the reference video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

**Reference Video:**



**Question:**

Which candidate video performs the task best according to the reference video?

**Candidate Videos:**

Video1: 

Video2: 

Video3: 

Video4: 

Figure 16: Examples on Ego–Exo Temporal Reasoning. Correct answers are highlighted in green.

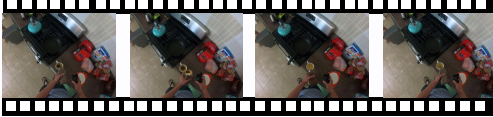
**Ego-Exo Temporal Reasoning**

**Action Order**

**Video1:**



**Video2:**



**Question:**  
What is the correct temporal relationship between the actions in these two videos?

- A. Actions in Video 1 occur before actions in Video 2
- B. Actions in Video 2 are part of the sequence of actions in Video 1
- C. Actions in Video 2 occur before actions in Video 1
- D. Actions in Video 1 are part of the sequence of actions in Video 2

---

**Video1:**



**Video2:**



**Question:**  
What is the correct temporal relationship between the actions in these two videos?

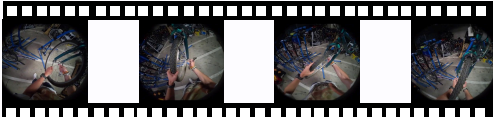
- A. Actions in Video 2 occur before actions in Video 1
- B. Actions in Video 1 are part of the sequence of actions in Video 2
- C. Actions in Video 2 are part of the sequence of actions in Video 1
- D. Actions in Video 1 occur before actions in Video 2

**Sequence Alignment**

**Video1:**



**Video2:**



**Question:**  
Which of the following statements accurately describes the difference in action sequences between Video 1 and Video 2?

- A. Both videos show the same sequence of actions without any differences
- B. Video 1 starts with rolling the wheel, while Video 2 starts with tightening both axle nuts
- C. Video 1 includes tightening the brake cable to the rear axle, while Video 2 does not
- D. Video 2 includes pushing the level inward and turning the axle clockwise, while Video 1 does not

---

**Video1:**



**Video2:**



**Question:**  
Which of the following statements accurately describes the difference in action sequences between Video 1 and Video 2?

- A. Video 1 includes reading the instructions, while Video 2 does not
- B. Video 2 includes arranging test material, while Video 1 does not
- C. Video 1 includes rotating and swirling the swab, while Video 2 does not
- D. Video 2 includes carefully opening the test tube seal paper, while Video 1 does not

Figure 17: Examples on Ego-Exo Temporal Reasoning. Correct answers are highlighted in green.

Table 5: Task-specific strategies to mitigate background interference. *Scene type* indicates high-level semantic categories (e.g., “kitchen”, “lab”), *Environment* denotes specific locations (e.g., kitchen 1 vs. kitchen 2), *Same environment* refers to videos from the same physical site but different viewpoints, and *Same video* means options from the same original file.

SubTask Type	Source	Scene Consistency	Remarks
OR, AR(LEMMA)	Same video	No background variation	All candidates from same video; GT-candidate from temporally-aligned clip; negatives differ temporally
PR, EWI	Same video	No background variation	All candidates from same video and clip; differ only by bounding boxes
AR (EgoExoLearn)	Same scene type, diff. env.	Candidates from same scene type (e.g., kitchens, labs) but diff. locations	Tests generalization across environments
DP, AO	Same env., diff. viewpoints	No background variation	All videos from same physical environment; vary only by viewpoint
TR	Same scene type, diff. env.	All videos from same scene type	Models rely on high-level task semantics, not background
SA, AP, SE	Same scene type	All videos from same scene type	Evaluates ability to align or predict actions/skills across similar environments

## B Experiments

We begin by designing an experiment to evaluate whether current MLLMs can accurately identify the boundaries between multiple concatenated video inputs. We then present additional results on the EgoExoBench benchmark and compare model performance on EgoExoBench with that on other video understanding benchmarks. Finally, we provide experimental details and qualitative examples related to the Chain-of-Thought (CoT) prompting strategies and the ablation study on the impact of using reference videos.

### B.1 Preliminary Study: Can MLLMs Distinguish Multiple Video Inputs?

Before conducting the main experiments, we verify whether MLLMs can differentiate multiple video inputs as distinct streams. This capability is critical, as EgoExoBench tasks require to compare or relate information across several videos simultaneously. To this end, we design an identical video pair identification task. Each question presents five video clips, two of which are exact duplicates. The model’s objective is to identify the matching pair. An example of this task is illustrated in Figure 18.

We curate video clips from Ego-Exo4D [17], including both egocentric and exocentric views. We construct a total of 100 question instances for evaluation. For each video, we uniformly sample 8 frames as input. Accuracy is used as the evaluation metric, where a response is considered correct only if the model precisely identifies the matching video pair. The expected accuracy under random guessing is 10%.

Table 6: Model performance on the preliminary experiment.

Model	Acc.
Qwen2.5-VL-7B	97
InternVL3-8B	85
LLaVA-Video-7B	77
NVILA-8B	78
EgoGPT	79
Claude-3.7-Sonnet	89

As shown in Table 6, all evaluated models achieve an accuracy above 77%, substantially exceeding random guess performance. This result confirms that current MLLMs are generally capable of distinguishing between multiple video inputs. This finding supports the feasibility of the results in our main experiments.

### B.2 Main Results

We additionally evaluate Gemini 2.5 Pro [12], Qwen2.5-Omni [74], Ola [44], on EgoExoBench. As shown in Table 7, Gemini 2.5 Pro [12] achieves the highest overall performance, with an average accuracy of 51.7%, outperforming the second-best model, GPT-4o-mini [47], by 3.7%. We also report the performance of MLLMs on several widely used video understanding benchmarks. As shown in Table 8, open-source models such as Qwen2.5VL-72B [62] and InternVL3-78B [85] outperform the closed-source GPT-4o [29] on certain benchmarks. Building on these overall results, we further highlight the importance of high-resolution visual encoding. Preserving fine-grained visual details is critical for tasks requiring precise spatial understanding, such as Ego-Exo View Transition and parts of Ego-Exo Matching. Models like InternVL3 and Qwen2.5-VL, which leverage dynamic resolution strategies to handle high-resolution inputs, consistently outperform those relying on fixed low-resolution inputs. This capability is especially beneficial for tasks like Egocentric Wearer Identification (EWI) and Body Part Action Understanding (BPA). Temporal modeling also plays a key role. Performance in the Ego-Exo Temporal Reasoning category correlates with the sophistication of a model’s temporal encoding. Qwen2.5-VL’s architecture introduces an absolute temporal encoding strategy. This provides the model with a physical sense of temporal flow and duration, a distinct advantage over simpler sequential encoding that merely counts frames. This architectural choice likely explains Qwen2.5-VL’s superior performance on Action Ordering (AO).

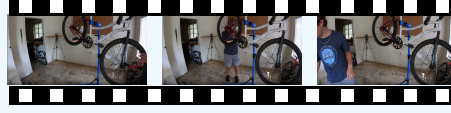
**[Task Instruction]**

You will be provided with 5 videos. Please carefully analyze the content of each video and identify the two videos that are most likely to be identical.

**Video1:**



**Video2:**



**Video3:**



**Video4:**



**Video5:**



**[Response Format]**

Output your answer in the format VideoX-VideoY, indicating the two videos you believe are most similar. The answer is:

**GT: Video1-Video3**

Figure 18: Example of the identical video pair identification task proposed in the preliminary experiment.

Table 7: Performance of various MLLMs on EgoExoBench. The best result among all models is highlighted in bold, and the second-best is underlined.

	Avg.	Matching				View Transition			Temporal Reasoning			
		TR	AR	OR	PR	EWI	DP	BPA	AP	AO	SA	SE
Qwen2.5-VL-72B	44.7	51.0	43.5	56.6	49.7	56.7	<u>37.0</u>	<b>48.1</b>	39.9	33.6	46.1	29.9
InternVL3-78B	41.4	50.6	37.3	48.9	39.0	46.9	32.3	38.0	<b>51.0</b>	31.5	<u>50.6</u>	29.5
Claude-3.7-Sonnet	31.3	33.4	33.0	35.1	30.6	37.8	30.0	33.3	26.0	28.5	30.2	26.4
GPT-4o	38.5	<u>52.9</u>	44.6	57.4	48.0	41.6	24.1	37.1	27.7	26.1	37.4	26.3
GPT-o4-mini	48.0	<b>65.3</b>	<u>45.2</u>	<u>56.8</u>	<u>53.7</u>	<u>73.8</u>	30.0	40.0	46.4	<b>35.9</b>	<b>51.5</b>	<u>30.1</u>
Gemini 2.5 Pro	<b>51.7</b>	63.1	<b>52.1</b>	<b>67.3</b>	<b>56.8</b>	<b>76.0</b>	<b>38.0</b>	<u>47.7</u>	<u>50.8</u>	<u>33.9</u>	49.3	<b>33.4</b>

Table 8: Comparison of MLLMs performance across EgoExoBench and existing video understanding benchmarks.

	Video-MME	MLVU	LongVideoBench	CG-Bench	Egoschema	EgoExoBench
Qwen2.5-Omni-7B	64.3/72.4	-	-	-	68.6	30.0
Ola	68.4/-	-	61.4	-	-	28.0
Qwen2.5-VL-7B	65.1/71.6	70.2	56.0	-	65.0	32.8
LLaVA-OV-7B	58.2/-	64.7	56.4	31.1/43.2	60.1	29.5
LLaVA-Video-7B	46.5/-	-	43.5	-	57.3	31.2
InternVL3-8B	66.3/68.9	71.4	58.8	38.6/55.2	-	31.3
InternVL3-14B	70.4/73.0	73.3	63.9	44.1/60.6	-	35.1
Qwen2.5-VL-72B	<b>73.3/79.1</b>	74.6	60.7	-	<b>76.2</b>	<b>44.7</b>
InternVL3-78B	72.7/75.7	<b>79.5</b>	65.7	<b>48.4/65.3</b>	-	41.4
GPT-4o	71.9/77.2	64.6	<b>66.7</b>	41.8/58.3	72.2	38.5



Given a prediction for a multiple-choice question, directly extract the selected answer while skipping the reasoning process. If the prediction explicitly chooses option (e.g., A, B, C, D), return the corresponding letter. If the prediction does not specify a choice or indicates that none of the options are correct, return None.

[Prediction]

Figure 19: Prompt for extracting the final answer from CoT prediction.

### B.3 Experimental Details of CoT Prompting

As described in the main paper, we evaluate the effectiveness of Chain-of-Thought (CoT) prompting on our benchmark. Following [77], we append the phrase "Let's think step by step." to each question. The decoding parameters are configured with temperature set to 0, top-p to 1, and top-k to 1. To extract the final answer from the model's output, we employ Qwen2.5-32B [76] as a parser to explicitly extract the answer. The prompt design is illustrated in Figure 19.

**Case Analysis.** We present Chain-of-Thought (CoT) reasoning examples to illustrate how the model interprets videos and infers answer. Figure 20 shows a successful case of GPT-4o [29] on the Task Matching subtask. The model reasons by integrating environmental context and key actions of individuals. It identifies critical behaviors across videos, infers the underlying tasks, and compares candidate videos with the query video to select answer. Figure 21 illustrates a successful example on the Sequence Alignment subtask. The model first provides a step-by-step description of the action sequence for each video. Notably, it attends to the order in which participants interact with key objects (e.g., test tubes, testing plates). Furthermore, it distinguishes between fine-grained actions such as covering and unwrapping. This detailed understanding of action semantics and temporal ordering enables the model to effectively compare sequences across videos and infer the correct answer. Figure 22 presents a failure case of GPT-4o on the Body Part Action Understanding subtask. While GPT-4o successfully identifies the actions performed in both videos, it incorrectly attributes the action to the right hand in the egocentric video. In contrast, it correctly identifies the corresponding body part in the exocentric video. This suggests that GPT-4o still struggles to accurately localize left-right body part actions across different viewpoints. Figure 23 presents a failure case of GPT-4o on the Person Matching subtask. The model demonstrates inconsistency in the reasoning process. During the analysis of the query video, it infers that the target person is positioned at a distance from others. However, it subsequently selects an individual who is positioned closer to other people, contradicting its earlier observation. This inconsistency suggests that the model struggles to maintain reasoning continuity across cross-view inputs.

### B.4 Experimental Details on Reference Video Usage.

In the main paper, we evaluate whether models can effectively leverage cross-view information in the Action Prediction and Skill Evaluation subtasks. To assess the contribution of the reference video, we conduct an ablation study by removing it while retaining the original input video. To ensure fair comparison, we keep the input prompts as consistent as possible, modifying only the necessary components to reflect the absence of the reference video. All other experimental settings remain unchanged. Figure 24 and Figure 25 show the task prompts for the Action Prediction and Skill Evaluation subtasks, respectively. Figure 26 presents examples from the Action Prediction subtask, comparing GPT-4o's predictions with and without the reference video. Figure 27 shows similar comparisons for the Skill Evaluation subtask.

**Task Matching**

**Query Video:**



**Question:**  
Which candidate video most accurately reflects the activities observed in the query video? Let's think step by step.

---

 **GPT-4o Prediction:**

To determine which video closely matches the query video, let's analyze the query video step by step:

**Query Video Analysis:**  
 Setting: The video takes place in a kitchen.  
 Actions:  
 One person initially walks across the kitchen.  
 The second person is seen standing, then goes to the refrigerator.  
 The second person opens the refrigerator to take something out.  
 This person then proceeds to the counter and engages in food preparation near or at the stove.

**Comparing Candidate Videos:**  
 Let's examine each candidate video:

**Video 1:**  
 Setting: Different kitchen setup with a dark theme.  
 Actions: Involves actions like taking items from the refrigerator and using the microwave, which is not aligned with the query.

**Video 2:**  
 Setting: Similar kitchen setup to the query.  
 Actions:  
 A person goes to the refrigerator.  
 The person then uses the stove with utensils similar to those in the query, aligned with the sequence of opening the fridge, taking items, and preparing something on the stove.

**Video 3:**  
 Setting: Kitchen setup similar to the query.  
 Actions: Focuses on cutting a watermelon, which is not present in the query video.

**Video 4:**  
 Setting: Wooden kitchen cabinets, different style from query.  
 Actions: Involves using the refrigerator and pouring a drink. Does not involve using stove or food preparation activities.

**Conclusion:**  
 Video 2 fits the actions and setting closest to the query video, with the refrigerator and stove being used in a sequence consistent with the query. Thus, Video 2 most accurately reflects the activities observed in the query video. ✓

**Candidate Videos:**

Video1:



Video2:



Video3:



Video4:



Figure 20: A successful case of GPT-4o's CoT reasoning on the Task Matching subtask. The model compares key actions across videos to identify the matching candidate. Key actions are highlighted in green.

## Sequence Alignment

### Video1:



### Video2:



### Question:

What is the key difference in the action sequences between Video 1 and Video 2? Let's think step by step.

### Options:

- A. In Video 1, the test tube is covered after unwrapping the testing plate, whereas in Video 2, the test tube is covered before unwrapping the test tube cap.
- B. In Video 1, the test tube is covered before unwrapping the testing plate, whereas in Video 2, the test tube is covered after unwrapping the test tube cap.
- C. In Video 1, the swab is swirled after covering the test tube, whereas in Video 2, the swab is swirled before removing it from the testing tube.
- D. In Video 1, the testing plate is unwrapped before swirling the swab, whereas in Video 2, the testing plate is unwrapped after covering the test tube.

### GPT-4o Prediction:

To determine the key difference in the action sequences, let's analyze each video step-by-step based on the provided videos:

#### Video 1:

1. The person appears to be preparing materials on the table.
2. They seem to be handling a test tube, possibly **doing a covering or sealing action**.
3. There's a focus on opening or handling some instruction or paper.
4. **The testing plate is being unwrapped** and possibly manipulated.

#### Video 2:

1. The person is seen handling materials on a table.
2. They **unwrap the testing plate** and handle the test tube.
3. After unwrapping, they perform actions with the test tube, such as **covering it**.

#### Conclusion:

After reviewing the sequences, Option B correctly describes the actions. ✓

Figure 21: A successful case of GPT-4o's CoT reasoning on the Sequence Alignment subtask. The model captures the order of interactions with key objects and differentiates fine-grained actions, enabling accurate cross-video sequence comparison. Key reasoning steps are highlighted in green.

### Body Part Action Understanding

#### Video1:



#### Video2:



#### Question:

How do the roles of the left and right hands differ between Video 1 and Video 2? Let's think step by step.

#### Options:

- A. In Video 1, the left hand handles seasoning; in Video 2, the left hand supports the pot and the right hand handles scooping.
- B. In Video 1, the right hand handles seasoning; in Video 2, the right hand handles scooping and the left hand supports the pot.
- C. In Video 1, the right hand handles seasoning and the right hand assists; in Video 2, the right hand supports the pot and the left hand handles scooping,
- D. In Video 1, the left hand handles seasoning; in Video 2, the left hand handles scooping and the right hand supports the pot.



#### GPT-4o Prediction:

To determine the roles of the hands in each video, we analyze their actions separately.

#### Video 1:

The images show someone cooking over a stove. **The right hand is adding seasoning from a packet to the pan.** There is no clear indication that the left hand is supporting the pot, but it appears to assist during the process.

#### Video 2:

The cook is preparing a dish and serving it onto a plate. **The left hand is holding and supporting the pan.** **The right hand is using a utensil** to scoop food onto the plate.

#### Comparison of hand roles:

In Video 1, the right hand handles seasoning, and the left hand assists.

In Video 2, the right hand handles scooping, and the left hand supports the pot.

Based on this analysis, the correct option is: B. **X**

Figure 22: A failure case of GPT-4o's CoT reasoning on the Body Part Action Understanding subtask. The model misidentifies the active hand in the egocentric view but correctly recognizes it in the exocentric view, indicating challenges in maintaining consistent left-right localization across perspectives. Correct predictions are highlighted in green; incorrect ones in red.

**Person Matching**

**Query Video:**



**Question:**  
Which candidate video most accurately matches the person (within the bbox) in the query video? Let's think step by step.

**Candidate Videos:**

Video1:



Video2:



Video3:



Video4:



---

 **GPT-4o Prediction:**

To determine which candidate video most accurately matches the person in the query video, let's follow these steps:

- Query Video Analysis:**  
The query video features a person annotated with a bounding box who is dressed in blue and is standing in a more distant position relative to others.
- Candidate Videos Analysis:**

Video 1:  
The person within the bounding box is located in the middle area, positioned closer to others, and is facing towards the lower part of the image.

Video 2: The person within the bounding box is positioned at the bottom of the scene, facing upwards, and is in a different area relative to others.

Video 3: The person within the bounding box is located towards the top left of the image, facing right.

Video 4: The person within the bounding box is near the bottom, facing upwards.
- Matching Characteristics:**  
Considering the horizontal perspective of the query video, the person stands out due to their attire (blue) and position. Video 1's individual matches in positioning, general stance, and isolation relative to other figures more closely than the others. Therefore, Video 1 most accurately matches the person in the query video. ✗

Figure 23: A failure case of GPT-4o on the Person Matching subtask. The model infers the target person is distant from others in the query video but selects a candidate positioned closer to others, revealing inconsistency in cross-view reasoning. Correct reasoning steps are highlighted in green. Inconsistent reasoning steps are highlighted in red.

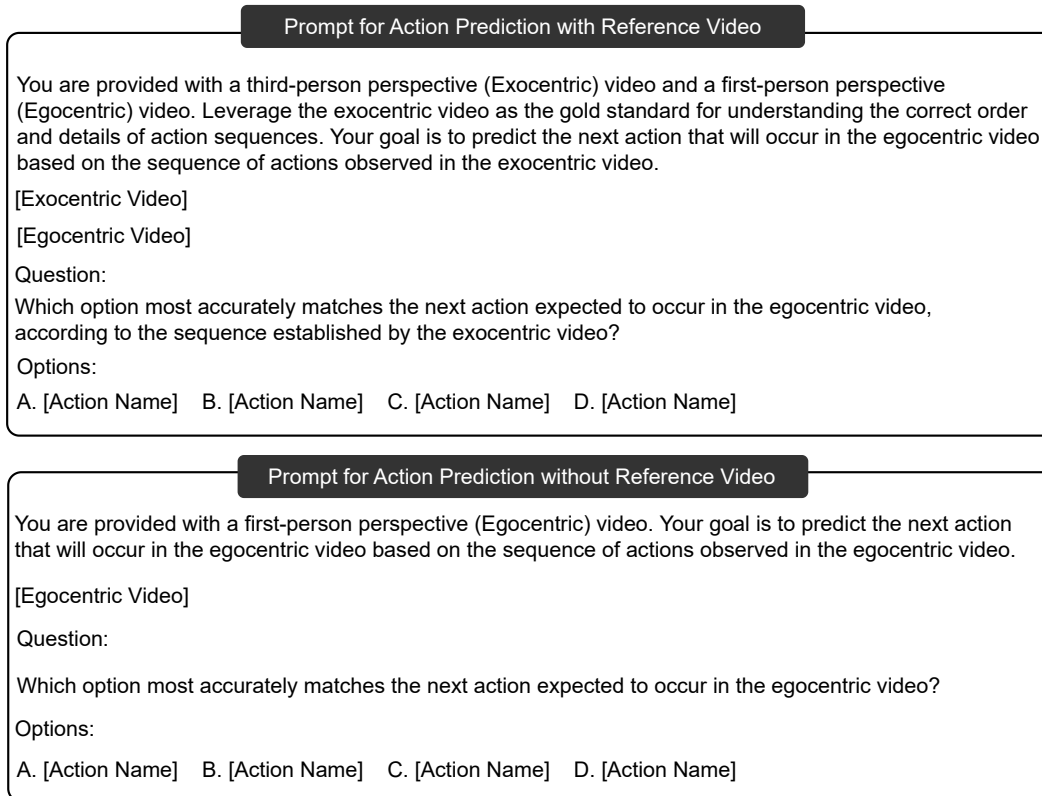


Figure 24: Prompts for the Action Prediction subtask, with and without the reference video.

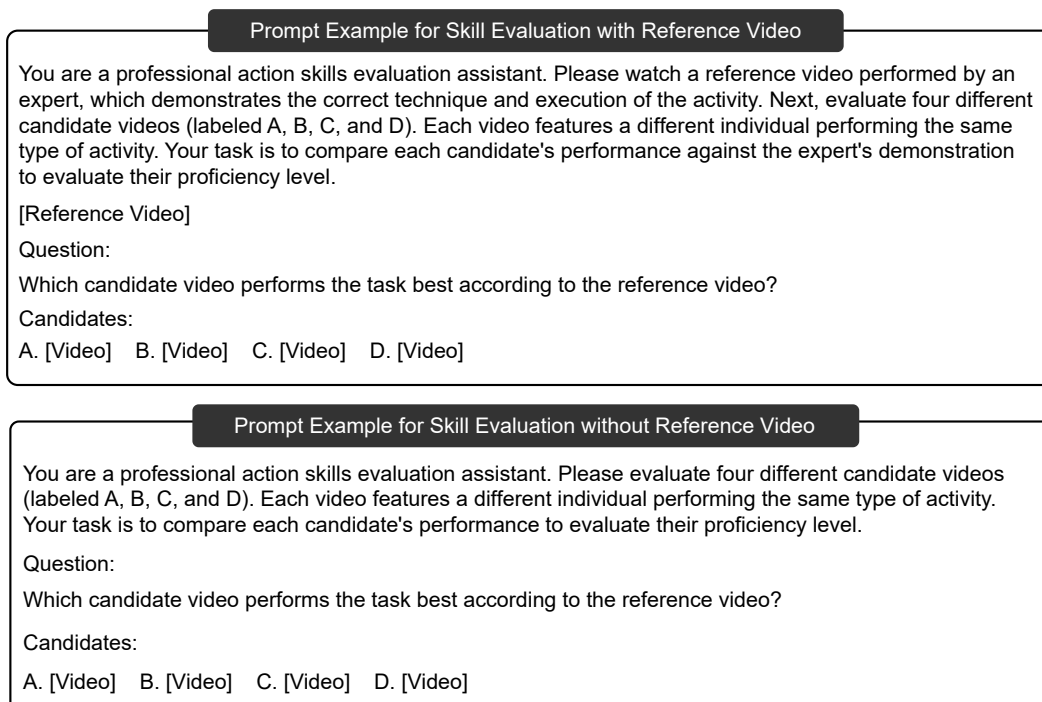


Figure 25: Prompts for the Skill Evaluation subtask, with and without the reference video.



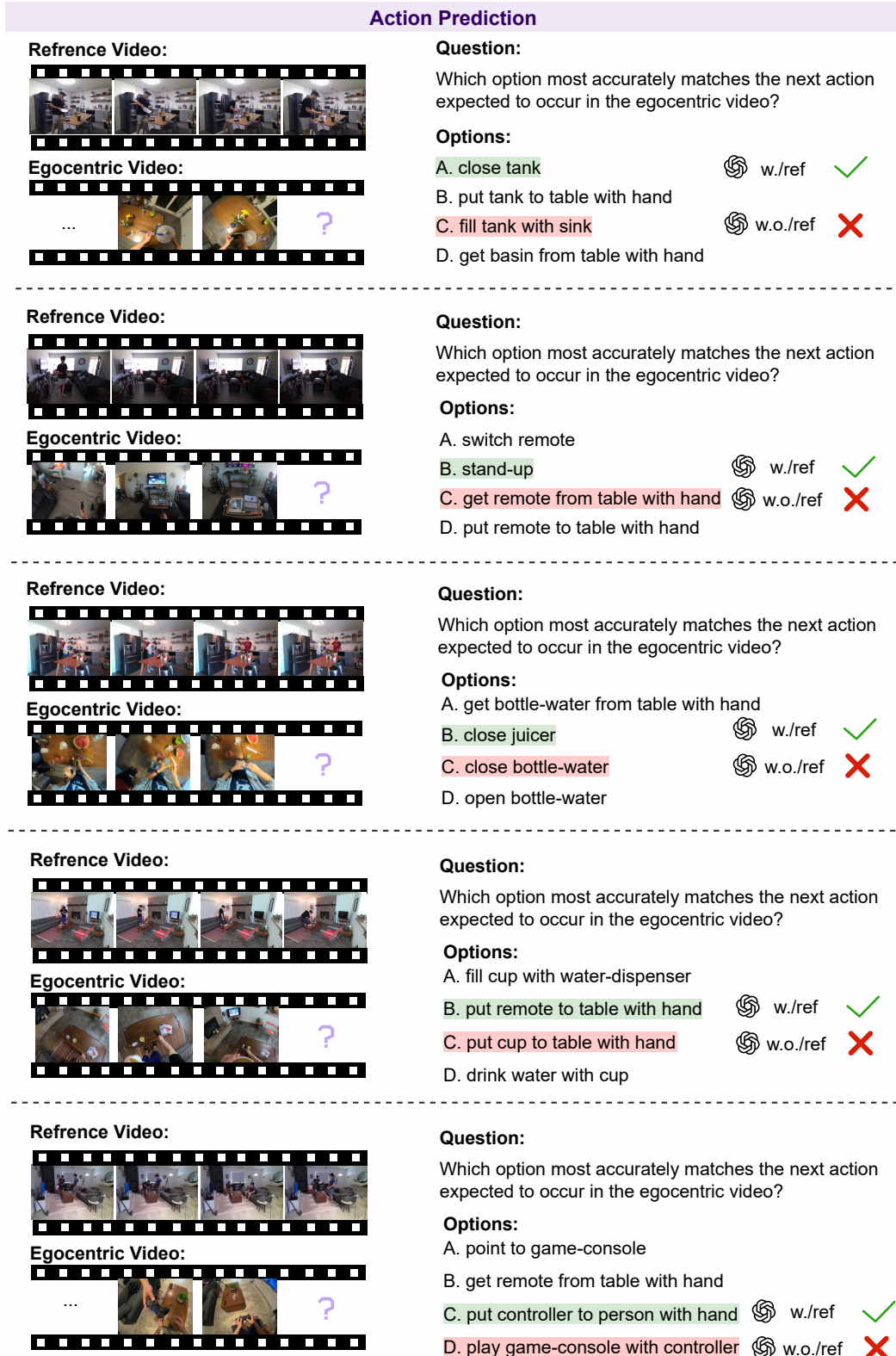


Figure 26: GPT-4o predictions on the Action Prediction subtask with and without the reference video. Correct predictions are highlighted in green; incorrect ones in red.

## Skill Evaluation

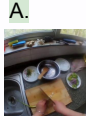
**Reference Video:**



**Question:**

which candidate video performs the task best?

**Candidates:**



🌀 w.o./ref ✗ 🌀 w./ref ✗

**Reference Video:**



**Question:**

which candidate video performs the task best?

**Candidates:**



🌀 w.o./ref ✗ 🌀 w./ref ✗

**Reference Video:**



**Question:**

which candidate video performs the task best?

**Candidates:**



🌀 w./ref ✗ 🌀 w.o./ref ✗

**Reference Video:**



**Question:**

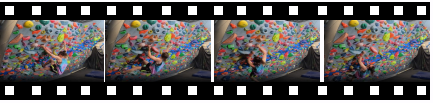
which candidate video performs the task worst?

**Candidates:**



🌀 w./ref ✓ 🌀 w.o./ref ✗

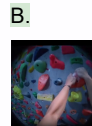
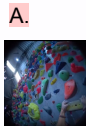
**Reference Video:**



**Question:**

which candidate video performs the task worst?

**Candidates:**



🌀 w./ref ✗ 🌀 w.o./ref ✓

Figure 27: GPT-4o predictions on the Skill Evaluation subtask with and without the reference video. Correct predictions are highlighted in green; incorrect ones in red.

## 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 abstract and introduction clearly state the paper's main contributions, which are consistently elaborated upon in the Benchmark and Experiment sections.

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: The paper discusses limitations.

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: The paper provides detailed descriptions of the experimental settings.

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: The paper provides both the code and dataset under open access, along with clear instructions and scripts in the supplemental material to facilitate faithful reproduction of the main experimental results.

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: Evaluation details are thoroughly documented in the experiment section, and the open-sourced code ensures reproducibility of the reported results.

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: [NA]

Justification: As a benchmark study, the paper avoids any fine-tuning or gradient updates to ensure fair comparison across models.

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: The experimental section includes details on computational resources, providing sufficient information to support reproducibility.

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: The research complies with 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: The paper discusses both the positive societal impacts and potential negative consequences.

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: [NA]

Justification: The data used in this work are sourced from publicly available open datasets that have established release protocols, and no additional safeguards are required or described in this paper.

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: All external datasets, and models used in this work are properly credited with citations and acknowledgments. The paper respects the licenses and terms of use as specified by the original creators.

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](https://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: The paper provides comprehensive documentation for all newly introduced assets, including detailed descriptions, usage instructions, and data format specifications, which are made available alongside the released code and benchmark.

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 paper provides full instructions given to annotators, including detailed guidelines and example screenshots.

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: [\[No\]](#)

Justification: The paper involves human annotations for benchmark construction. The annotation process posed minimal risk to participants and did not involve any collection of personal or sensitive information.

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: [Yes]

Justification: LLMs are used during the benchmark construction phase, which is described in the paper.

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.