EgoThinker: Unveiling Egocentric Reasoning with Spatio-Temporal CoT

Baoqi Pei[1,2], **Yifei Huang**[1,3]*, **Jilan Xu**[1,4], **Yuping He**[5], **Guo Chen**[5], **Fei Wu**[2], **Yu Qiao**[1], **Jiangmiao Pang**[1]
[1]Shanghai Artificial Intelligence Laboratory, [2]Zhejiang University,
[3]The University of Tokyo, [4]Fudan University, [5]Nanjing University
peibaoqi@gmail.com; hyf@iis.u-tokyo.ac.jp

Abstract

Egocentric video reasoning centers on an unobservable agent behind the camera who dynamically shapes the environment, requiring inference of hidden intentions and recognition of fine-grained interactions. This core challenge limits current multimodal large language models (MLLMs), which excel at visible event reasoning but lack embodied, first-person understanding. To bridge this gap, we introduce **EgoThinker**, a novel framework that endows MLLMs with robust egocentric reasoning capabilities through spatio-temporal chain-of-thought supervision and a two-stage learning curriculum. First, we introduce **EgoRe-5M**, a large-scale egocentric QA dataset constructed from 13M diverse egocentric video clips. This dataset features multi-minute segments annotated with detailed CoT rationales and dense hand–object grounding. Second, we employ SFT on EgoRe-5M to instill reasoning skills, followed by reinforcement fine-tuning (RFT) to further enhance spatio-temporal localization. Experimental results show that EgoThinker outperforms existing methods across multiple egocentric benchmarks, while achieving substantial improvements in fine-grained spatio-temporal localization tasks. Full code and data are released at https://github.com/InternRobotics/EgoThinker.

1 Introduction

Humans possess a remarkable ability to reason, plan, and execute complex, goal-oriented behaviors within dynamic real-world environments. Recent works in Multimodal Large Language Models (MLLMs) have advanced the field of visual understanding [55, 1, 77, 54, 12]. Techniques such as chain-of-thought prompting [75, 89] and reinforcement fine-tuning [39, 27] further underscore the potential of MLLMs in high-level reasoning. However, existing approaches mainly address visual reasoning from an observer-centric, third-person viewpoint. This perspective fails to capture the embodied cognitive processes central to human reasoning, which naturally occur from the egocentric perspective.

Egocentric reasoning differs fundamentally from conventional visual reasoning due to the presence of the observer as an active participant in the scene. Thus, models must infer not only visible events but also the internal cognitive states, intentions, and future behaviors of the individual behind the camera. This reasoning process poses several unique challenges: (1) **Reasoning for complex tasks:** Understanding the rationale behind an action and predicting what comes next demands explicit cause-effect chains rather than isolated event recognition. (2) **Human–object interaction recognition:** Successful reasoning hinges on accurately localizing hands and manipulated objects.

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

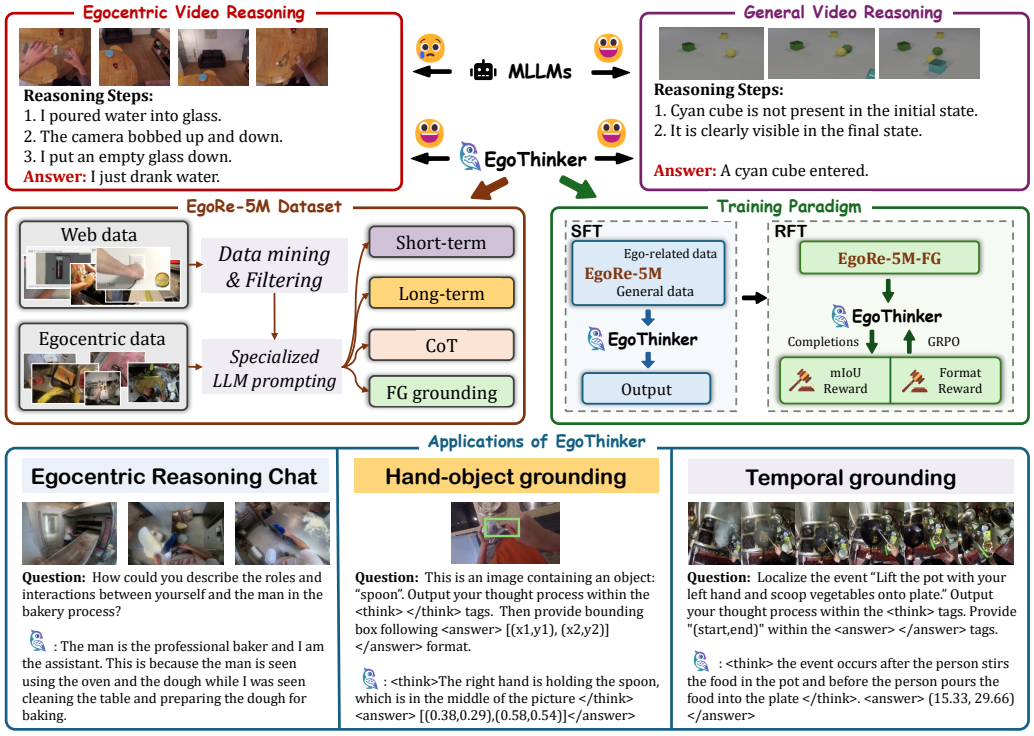

Figure 1: **Overview of our EgoThinker.** Unlike general video reasoning, egocentric video reasoning poses unique challenges because it must infer an unobservable camera wearer's interactions and intentions. EgoThinker addresses this by curating EgoRe-5M, a large-scale egocentric reasoning dataset, and applying a two-stage supervised and reinforcement fine-tuning paradigm. This design empowers robust egocentric reasoning chat, hand–object grounding, and temporal grounding, making EgoThinker a promising foundation for wearable assistants and embodied AI.

(3) **Multi-horizon temporal integration:** Egocentric video streams span minutes to hours, requiring models to track evolving context and retain fine-grained details across thousands of frames.

Existing egocentric datasets [97, 102] are mainly derived from Ego4D [25] and EgoExo4D [26]. They provide extensive collections of egocentric videos but lack explicit reasoning chains, temporally spanned annotations, and detailed fine-grained grounding data. Consequently, existing MLLMs [85, 101, 3], although successful in general visual understanding tasks, often struggle when reasoning about complex tasks in long-term egocentric videos.

In this work, we propose **EgoThinker**, a novel framework designed to enable robust egocentric reasoning in MLLMs. As illustrated in Figure 1, to overcome dataset limitations, we develop a pipeline to extract egocentric videos from large-scale web data to capture diverse real-world scenarios. Leveraging this, we construct EgoRe-5M, a large-scale egocentric QA dataset, featuring diverse questions spanning from seconds to several minutes. In the dataset, we incorporate detailed Chain-of-Thought (CoT) annotations to explicitly model the causal relationships underlying complex human activities, enabling models to emulate human-like causal inference and planning. Furthermore, recognizing the critical role of hand-object interaction in egocentric reasoning, we additionally introduce specialized data focusing on fine-grained spatio-temporal hand-object interactions.

Inspired by recent advances in reinforcement fine-tuning (RFT) [39, 27], EgoThinker employs a two-stage training strategy. Initially, the model undergoes supervised fine-tuning (SFT) [85, 101], utilizing EgoRe-5M to establish foundational understanding and reasoning. Subsequently, we employ RFT using spatio-temporal grounding data via GRPO [27] method. This paradigm significantly enhances the model's capability in fine-grained localization, temporal reasoning, and causal inference. Experiments demonstrate that EgoThinker outperforms existing state-of-the-art methods across multiple egocentric benchmarks, showcasing strong performance gains in egocentric QA [42, 21, 56], long-term video reasoning [15, 64], and spatio-temporal hand-object interaction localization [36, 19].

In summary, our contributions are: (1) EgoRe-5M, a large-scale egocentric QA dataset with chain-of-thought and hand-object annotations curated from diverse video sources. (2) A two-stage training regime combining supervised fine-tuning and reinforcement fine-tuning via GRPO to effectively couple high-level reasoning with low-level grounding. (3) EgoThinker, an MLLM setting new state-of-the-art on multiple egocentric video benchmarks, demonstrating coherent egocentric reasoning and precise spatial and temporal grounding.

2 Related work

Egocentric Video Understanding. Egocentric video understanding has recently garnered increasing research attention, with the introduction of large-scale egocentric datasets such as Ego4D. Previous works mainly focus on action understanding [72, 33, 24, 32, 34, 69, 29, 10], vision-language pre-training [56, 35, 108, 70] and hand-object interaction understanding [100, 9, 22, 93]. Recently, some methods [5, 97, 102, 30, 37] have attempted to use MLLMs for egocentric video understanding, and some targeting long-form egocentric videos [94, 87]. Different from these works, EgoThinker is the first model that enables reasoning and precise hand–object understanding in first-person videos.

Multimodal Large Language Models. Recent advancements in MLLMs [55, 1, 77, 31, 54, 12, 6] have shown robust comprehension and perception abilities. Video-LLaVA [55] and Videochat2 [51] enable MLLMs with general video understanding and temporal localization capability. Recent works extend to diverse domains including long-form video understanding [104, 11, 83, 44], robot learning [90, 45], and spatio-temporal perception [67, 62]. Chain-of-Thought (CoT) prompting [89] has proven effective in eliciting multi-step reasoning in LLMs. Recently, some works [13, 75, 14] use CoT techniques to enhance MLLM's visual reasoning capabilities. In our work, we construct large-scale QA samples with causal CoT captions to equip MLLMs with egocentric reasoning ability.

Reinforcement Learning for MLLMs. Recently, OpenAI-o1 [39] and DeepSeek-R1 [27] have shown that reward-driven fine-tuning can substantially enhance LLM reasoning. For MLLMs, many works [110, 59, 99, 20, 71, 60, 95, 103, 61] focused on leveraging reinforcement learning (RL) techniques with verifiable rewards to enhance visual reasoning capabilities with only a small amount of data. Videochat-R1 [52] enhances MLLM's temporal perception ability via RFT for general video understanding. Building on these advances, we construct fine-grained spatio-temporal grounding datasets and apply GRPO approach to endow MLLMs with precise hand–object localization and long-horizon causal inference in egocentric videos.

3 EgoThinker: Enhancing MLLMs Egocentric Reasoning Ability

In this section, we introduce our EgoThinker framework, designed to equip MLLMs with egocentric reasoning capabilities. We begin with the curation of EgoRe-5M, a large-scale egocentric instruction tuning dataset. We then describe how EgoRe-5M fuels our two-stage learning curriculum: initial supervised fine-tuning to establish foundational reasoning skills, followed by a reinforcement fine-tuning paradigm to sharpen hand-object grounding and temporal reasoning ability.

3.1 EgoRe-5M

Recent works [97, 102, 112] have constructed egocentric QA datasets but lack data for long-term causal reasoning and fine-grained spatio-temporal localization. This deficiency hinders the development of models capable of egocentric reasoning. To address this limitation, we propose EgoRe-5M, a large-scale egocentric QA dataset designed with rich causal reasoning and grounding data.

3.1.1 Egocentric Video Collection

Accurate egocentric reasoning demands a vast and diverse collection of egocentric data. While existing datasets like Ego4D [25] and EgoExo4D [26] are valuable, their scale is notably smaller compared to web-sourced data. To overcome this scale bottleneck, we develop a multi-stage filtering pipeline to mine high-quality egocentric clips from web-sourced videos.

(1) Web-scale Mining. We choose the large instructional video dataset HowTo100M [65] as the data source in which numerous instructional video are recorded by a head-mounted or handheld

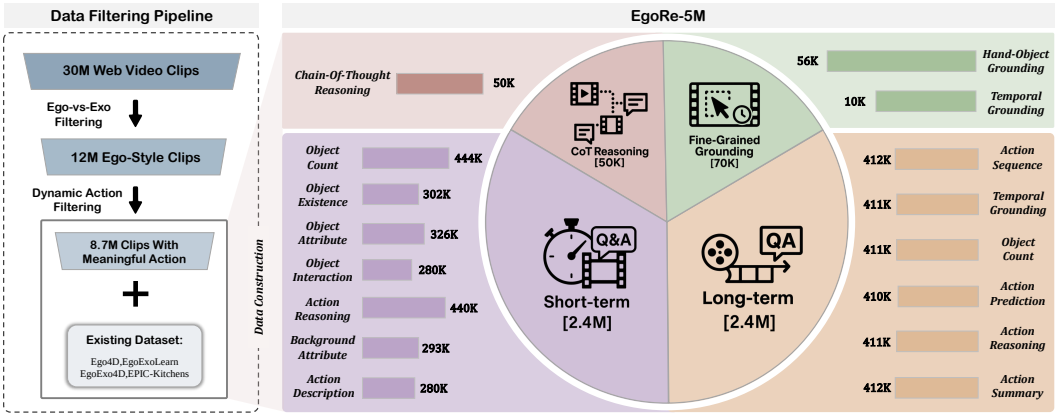

Figure 2: **Data Filtering Pipeline and EgoRe-5M Statistics.** With our multi-stage filtering pipeline, we construct EgoRe-5M, a large-scale QA dataset to facilitate egocentric reasoning in MLLMs.

camera featuring fine-grained hand-object interactions. Moreover, the task diversity and naturalistic narration within HowTo100M make it especially rich in egocentric perspectives. In particular, we select HTM-AA [28], a dataset containing a large number of temporally aligned video-narration pairs and Howto-Interlink7M [81], which additionally provides high-quality video caption annotations as our primary data sources, resulting in a total of 30M initial video clips. The duration of the video clips spans from a few seconds to several minutes.

(2) Ego-vs-Exo Filtering. To distinguish true egocentric footage from exocentric content, we train a classification model on balanced sets of manually labeled ego- and exo-centric clips. The classification model utilizes InternVideo backbone [85] followed by a two layer MLP. The model achieves 92% accuracy and 89% AUC on held-out validation data. Applying this model reduces our pool to 12M clips that exhibit clear first-person camera motion and range from seconds to minutes.

(3) Dynamic Interaction Filtering. Even after egocentric filtering, many clips remain static or depict group activities, offering limited value for reasoning about egocentric activities. Since most egocentric activities are centered on hand-object interactions, we run a pre-trained hand-object detector [74] to identify frames containing both a visible hand and an active object. We design several rules to filter the clips that exhibit hand-object interaction and dynamic changes, with each clip having a minimum duration of 2 seconds. This refinement yields 8.7M high-quality egocentric clips, each containing rich, dynamic interactions suitable for downstream QA annotation.

As a result, we combine our filtered 8.7M web data with existing egocentric datasets (Ego4D [25], EPIC-Kitchens [18], EgoExoLearn [36] and EgoExo4D [26]) to form a collection of 13M egocentric video clips in total. For details regarding the pipeline, please refer to the Supplementary A.

3.1.2 Egocentric Reasoning Data Construction

Prior egocentric QA datasets focus primarily on short video clips and raw narrations, falling short on long-term causal chains and spatial-temporal grounding, which are key components of egocentric reasoning. To address these gaps, we build EgoRe-5M, an automatically generated QA corpus containing four complementary task dimensions: short-term perception, long-term causal reasoning, chain-of-thought rationales, and fine-grained grounding. Figure 2 shows an overview of the data source of our EgoRe-5M. Since some video clips have no corresponding text annotations, or the text annotations come from low-quality automatic speech recognition, we employ Videochat2-HD [46], an efficient and robust video caption model, to annotate these videos with a sampling rate of 1 fps. Questions are formulated by specialized LLMs per split, detailed as follows.

Short-term Data. To instill foundational egocentric perception skills, we generate a large-scale short-term QA split covering clips of 1–10 seconds. We design seven perceptual question types to capture immediate scene understanding: object existence, object attribute, object count, object interaction, action description, action reasoning and background attribute. For each clip, we combine the original text annotation with VideoChat2-HD captions and apply DeepSeek-V3 to instantiate and

answer these question templates. This process yields 2.4M QA pairs, ensuring broad coverage of objects, interactions, and immediate causal cues essential for downstream model pretraining.

Long-term Data. Human activities often progress through multiple steps, which makes long-form understanding essential for egocentric reasoning. To capture such extended causal chains, we aggregate consecutive clips into segments of 15–120 seconds and integrate their narrations into a single, coherent caption. We then design six question types to assess temporal and causal understanding within each segment: action sequence, temporal grounding, object count, action prediction, action summary and action reasoning. Utilizing DeepSeek-V3 on these concatenated captions, we automatically generate 2.5 million QA pairs. This aims to enhance models' abilities to connect events over long durations and deduce causal relationships.

Chain-of-thought Data. Recent advances in chain-of-thought (CoT) prompting [50, 107] have shown that explicitly conditioning LLMs on intermediate reasoning steps can significantly boost performance on complex inference tasks such as mathematics and coding. Visual CoT [75] extends this idea by supplying models with region-level hints to guide step-by-step reasoning. Thanks to the success of the open-source DeepSeek R1 [27] model, it is possible to utilize LLMs to generate reasoning processes. Similar to long-term data, we select video clips with dense captions and concatenate them to form new captions. Each description is then fed to DeepSeek R1, producing a question and an accompanying step-by-step rationale. We design a prompt to allow the model to decide whether a given segment warrants a CoT question, ensuring that only clips amenable to multi-step inference are annotated. The resulting split comprises 50K high-fidelity CoT QA pairs, each pairing a complex egocentric scenario with an explicit chain-of-thought process.

Fine-Grained Grounding Data. Hand–object interactions lie at the heart of egocentric reasoning. Existing specialized methods [100] perform well on these tasks, but existing MLLMs exhibit poor performance. To remedy this, we construct QA data for two complementary grounding tasks: *Hand-object Grounding* and *Temporal Grounding*.

First, leveraging EK-Visor's pixel-level masks for hands and active objects [19], we generate questions asking about the spatial positions of hands/objects. The input can take the form of an image. The model must first articulate its intermediate reasoning and then output a normalized bounding box. This split trains models to map hand and object visual cues to precise coordinates.

For temporal grounding, using EgoExoLearn's fine-grained temporal annotations [36], we pose questions that require selecting the exact time interval in a clip that contains the evidence needed to answer. Models need to provide step-by-step reasoning and output start–end times in seconds. This formulation emphasizes the ability to pinpoint moments of interest for downstream inference.

3.1.3 Data Statistics

As shown in Figure 2, our EgoRe-5M dataset comprises 5M question-answer annotations across four complementary splits. To verify the annotation quality, we randomly sample 500 QA pairs and check the correctness of answers and logical coherence of the intermediate rationales. Over 95% of reviewed samples meet our standards for accuracy and annotation quality. We provide qualitative examples in the supplementary to demonstrate the dataset's quality and range in object categories, action types, causal chains, and precise grounding scenarios.

3.2 Training EgoThinker

We employ a two-stage curriculum to imbue MLLMs with robust egocentric reasoning capabilities. First, we perform supervised fine-tuning on a carefully balanced mixture of general visual, egocentric, and QA datasets, including EgoRe-5M's short, long, and CoT splits, to establish core capabilities in object perception, causal inference, and multi-step planning (see details in the supplementary). Second, we refine spatio-temporal grounding via Reinforcement Fine-Tuning (RFT) using GRPO approach on EgoRe-5M's fine-grained grounding data. Together, these stages transform an off-the-shelf MLLM into our EgoThinker capable of egocentric reasoning.

Table 1: Overview of fine-tuning datasets used during training.

Stage	Domain	Data Type	Amount	Training Dataset
SFT	General	Caption	100K	*Llavar[105], Sharegpt4o[17], Sharegpt4video[13], YC2[111], Webvid[84], Videochatgpt[63]*
		VQA	70K	*Next-QA [91], TVQA [48], Clevrer-QA [98], TGIF-QA [40], Star-QA [80]*
	Ego	Ego-Related	390K	*SSV2 [109], Ego-QA [8], EgotimeQA [21]*
		EgoRe-5M	860K	*EgoRe-5M-Short, EgoRe-5M-Long, EgoRe-5M-CoT*
RFT	Ego	Grounding	70K	*EgoRe-5M-FG*

3.2.1 Supervised Fine-tuning

Supervised fine-tuning establishes the foundational reasoning, perception, and language skills of EgoThinker. To also preserve general visual understanding and conversational fluency, we assemble a 1.5M sample fine-tuning corpus drawn from a diverse mix of datasets of 4 categories as Table 1.

The details of the SFT training data are as follows: (1) **High quality caption datasets.** To maintain the general visual understanding and conversation ability, we select five high-quality datasets (llavar-gpt4-20k [105], sharegpt4o [17], sharegpt4video [13], webvid-2M [84], YouCook2 (YC2) [111] and videochatgpt [63]) containing image/video captions and conversational data. (2) **VQA datasets:** We use five commonly used QA datasets (Next-QA [91], TVQA [48], Clevrer-QA [98], TGIF-QA [40], and STAR-QA [80]) to sharpen QA skills. (3) **Ego-related datasets.** To emphasize the egocentric perspective understanding, we include existing ego-related datasets. Something-Something V2 (SSV2) [109] is relevant to action recognition and contain numerous ego-style clips. EgoTimeQA [21] and Ego-QA [51] are constructed from egocentric datasets, covering action recognition, and long-term video comprehension tasks. (4) **EgoRe-5M:** We sample from our EgoRe-5M short-term, long-term, and CoT splits to instill egocentric reasoning patterns. For details about SFT training, please refer to Supplementary B.

3.2.2 Reinforcement Fine-Tuning via GRPO

To enhance EgoThinker's spatio-temporal reasoning ability, we employ reinforcement fine-tuning (RFT) on the EgoRe-5M-FG split utilizing Group Relative Policy Optimization (GRPO) approach. Unlike traditional reward-model approaches [66], we adopt the rule-based scoring functions outlined in [47, 39, 76] to score the outputs of MLLMs directly. GRPO [27] then optimizes the policy by comparing groups of candidate responses without a separate critic network in.

Preliminary. Reinforcement Learning with Verifiable Reward (RLVR) replaces learned reward models with rule-based scorers that classify each output as correct or incorrect. DeepSeek R1-Zero's GRPO algorithm builds on RLVR by generating N distinct candidate responses $o = \{o_1, \ldots, o_N\}$ for an input question q through policy sampling. Each candidate is evaluated by a reward function to produce corresponding scores $\{r_1, \ldots, r_N\}$. To compare answers fairly, GRPO computes a normalized advantage for each response:

$$A_i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^N)}{\text{std}(\{r_i\}_{i=1}^N)}, \tag{1}$$

where A_i represents the relative quality of the i-th answer. The policy update maximizes expected advantage-weighted likelihood, with a KL-divergence penalty to the reference model π_{ref}:

$$\max_{\pi_\theta} \mathbb{E}_{o\sim\pi_{\theta_{\text{old}}}(p)} \Big[\sum_{i=1}^{N} \frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}}(o_i)} \cdot A_i - \beta \, \mathrm{D}_{\text{KL}}\Big(\pi_\theta \,\|\, \pi_{\text{ref}}\Big) \Big], \tag{2}$$

where π_θ is the policy model, π_{ref} is the reference model before optimization and β is a regularization coefficient to control the KL-divergence. This objective encourages the model to allocate higher probability to top-ranked candidates while maintaining stability through KL regularization.

Reward function design. The design of the reward model is a critical step. We design two complementary rule-based rewards tailored to assist MLLM in egocentric reasoning.

Format reward. Following Deepseek-R1, we employ a format reward function to enforce the model to output the prediction value and thinking process in the specified format. Specifically, we expect the model to output its thinking process within `<think>...</think>` and the answer with

Table 2: **Results on egocentric video benchmarks.** For EgoTaskQA, we convert the dataset into a multiple-choice question format following [102].

Method	EgoTaskQA	QAEgo4D	ERQA	EgoPlan-Val	EgoSchema	EgoMCQ		VLN-QA	RES
	Acc.	Acc.	Acc.	Acc.	Acc.	Inter-Acc.	Intra-Acc.	Acc.	Acc.
LLaVA-Video-7B [106]	55.0	**68.6**	-	39.8	57.3	86.7	37.3	32.0	31.1
LLaVA-OneVision-7B [49]	55.8	65.7	-	37.3	60.1	84.5	36.0	34.0	25.3
VideoLLaMA3 [7]	56.6	62.4	-	36.4	61.1	77.3	29.8	34.5	21.3
VideoChat2-HD [51]	45.5	52.0	-	35.7	55.8	87.1	36.4	43.0	24.5
InternVL2-8B [16]	61.0	66.4	-	34.2	64.2	79.0	31.0	46.0	30.0
Exo2Ego-7B [102]	48.1	62.1	-	42.7	61.3	88.4	41.2	44.5	-
Qwen2-VL-7B [82]	57.9	60.3	37.0	38.3	63.3	86.4	34.1	42.0	26.3
EgoThinker	**64.4**	66.2	41.8	**47.1**	**67.6**	**89.3**	**41.4**	**54.0**	**39.5**

<answer>...</answer>, and we design a format reward R_{format}. We utilize regular expression matching to determine whether the output of the MLLM matches our specified format. If matches, we assign $R_{\text{format}} = 1$; otherwise, we assign $R_{\text{format}} = 0$.

IoU Reward in Spatio-Temporal Grounding. The mIoU (mean intersection over union) metric offers clear guidance for the grounding task, making it an appropriate choice for the reward function. Therefore, using the fine-grained annotations from EgoRe-5M-FG, we calculate the spatial/temporal IoU between the predicted boxes/intervals and the ground-truth boxes/intervals as a reward function $R_{\text{o}_{\text{iou}}}/R_{\text{t}_{\text{iou}}}$. Thus, for hand-object grounding task, we get $R_{\text{og}} = R_{\text{format}} + R_{\text{o}_{\text{iou}}}$, and for temporal grounding task, we get $R_{\text{tg}} = R_{\text{format}} + R_{\text{t}_{\text{iou}}}$.

During RFT, we use the combined rewards to strengthen EgoThinker's ability to generate well-structured reasoning traces and accurate spatial and temporal groundings.

4 Experiments

The main experiments are all conducted based on Qwen2-VL-7B [82]. To train EgoThinker, we first perform SFT on our curated datasets, then apply RFT via GRPO.

Benchmarks. To thoroughly assess performance, we organize our evaluation benchmarks as follows: *(1) Egocentric Benchmarks*. We evaluate core first-person reasoning capabilities on six established datasets: EgotaskQA [42], QAEgo4D [21], EgoPlan [15], EgoSchema [64], EgoMCQ [56], ERQA [78] and VLN-QA [51]. *(2) Cross-View Skill Transfer (RES)*. To gauge the model's ability to generalize learned skills across perspectives, we introduce the Referenced Egocentric Skill (RES) benchmark of 4-way MCQs using paired clips from EgoExoLearn [36] and EgoExo4D [26]. *(3) Fine-grained Spatio-Temporal Grounding*. We assess spatial and temporal grounding on two newly constructed specialized HOI benchmarks derived from EK-Visor [19] and EgoExoLearn [36]. *(4) General Video Understanding*. Finally, to confirm that EgoThinker maintains broad applicability, we report results on three general video benchmarks including MVBench [51], Perception Test [2], and VideoMME [23]. Please refer to the Supplementary for details about these benchmarks.

4.1 Quantitative Evaluation of Egothinker

Egocentric Benchmarks. Table 2 shows a comprehensive comparison of EgoThinker against our baseline Qwen2-VL-7B and other leading MLLMs on the egocentric benchmarks. EgoThinker establishes new state-of-the-art performance across all tasks, achieving a 4.4% absolute gain on EgoPlan, 3.4% on EgoSchema, and 8.0% on VLN-QA. This clearly demonstrates EgoThinker's strong capacity for long-horizon planning, semantic inference, and goal-oriented question answering in egocentric video. In contrast, baseline models exhibit inconsistent strengths: LLaVA-Video leads on QAEgo4D while InternVL2 excels on EgoSchema and EgotaskQA, revealing an apparent lack of generalization in existing MLLMs to egocentric perception and reasoning.

On the Referenced Egocentric Skill (RES) cross-view benchmark, where prior MLLMs perform at near-random levels, EgoThinker outperforms the second best model by 8.4%, underscoring its unique ability to transfer learned skills between perspectives. These results confirm that our EgoThinker effectively unleashes the egocentric reasoning capabilities of MLLMs.

Table 3: EgoThinker results on hand-object and temporal grounding tasks.

Method	EK-Visor		EgoExoLearn	
	mIoU	Loc-Acc.	mIoU	R1@0.05
LLaVA-Video [106]	46.7	67.7	1.30	7.8
Qwen2VL-7B [82]	56.7	64.5	1.53	5.4
Qwen2.5VL-72B [79]	64.1	71.7	21.1	49.9
EgoThinker	**53.7**	**80.3**	**25.2**	**63.9**

Table 4: EgoThinker results on general video benchmarks.

Method	MVBench	Perception Test	VideoMME
MiniCPM-V2.6 [96]	67.1	58.1	60.9
InternVL2 [16]	66.4	60.1	54.0
LLaVA-Video [106]	58.6	67.9	63.3
InternVideo2.5 [86]	**74.0**	**76.2**	**65.1**
Qwen2VL [82]	68.2	70.3	62.9
EgoThinker	70.0 (+1.8)	72.7 (+2.4)	62.9 (+0.0)

Spatio-Temporal Grounding Benchmarks. For spatial grounding of hand–object, we measure mean Intersection over Union (mIoU) and Localization Accuracy (Loc-Acc). For Loc-Acc, we determine the correctness by checking whether the predicted box's center is within the ground truth box. As shown in Table 3, off-the-shelf 7B-parameter MLLMs achieve only modest mIoU and Loc-Acc on EK-Visor. Scaling up to Qwen2.5-VL-72B improves these scores; however, EgoThinker, after our two-stage training paradigm, surpasses this baseline by a wide margin.

For temporal grounding, we employ the temporal window's mIoU and R1@0.05 as metrics. In Table 3, results in EgoExoLearn show that 7B-parameter MLLMs perform at near-zero levels, indicating almost no temporal localization capability. Qwen2.5-VL-72B improves the performance, but still underperforms our EgoThinker. These improvements confirm that our training paradigm effectively equips MLLMs to reason about "when" and "where" in dynamic egocentric video.

General Benchmarks. We further evaluated EgoThinker on three standard video understanding benchmarks. As shown in Table 4, EgoThinker exhibits no degradation across any of these tasks, matching or exceeding the Qwen2-VL-7B baseline. Notably, it achieves clear gains on MVBench and Perception Test, underscoring that our two-stage paradigm not only unlocks robust egocentric reasoning but also enhances broad video understanding capabilities.

4.2 Ablation Studies and Discussions

Table 5: **Ablations on EgoRe-5M.** We use different splits of the data for training to validate the effectiveness of our dataset.

Method	Training Data				EgoTaskQA	QAEgo4D	Egoschema(val)	EK-Visor	
	Short	Long	CoT	FG	Acc.	Acc.	Acc.	mIoU	Loc-Acc.
Baseline	-	-	-	-	57.7	60.3	68.2	28.6	64.5
+SFT	✓	-	-	-	61.6	63.1	69.1	29.1	64.9
	✓	✓	-	-	64.2	63.7	71.1	28.9	64.5
	✓	✓	✓	-	64.3	**67.2**	**71.9**	28.5	64.4
+RFT	✓	✓	✓	✓	**64.4**	66.1	71.8	**53.7**	**80.3**

Training data. To systematically evaluate the contribution of each component in EgoRe-5M dataset, we perform ablations by applying SFT on the short-term, long-term, and CoT splits, while employing RFT on the fine-grained split. As shown in Table 5, the short-term split boosts performance on three QA benchmarks especially EgotaskQA, highlighting the value of dense, scenario-diverse clips. The long-term split improves EgoSchema as expected, while incorporating the CoT split leads to marked enhancements in QAEgo4D, a dataset focusing on episodic-memory-based question answering. This shows that explicit causal reasoning traces can benefit memory-driven tasks. Finally, the FG split significantly enhances spatial and temporal grounding. Overall, these results verify that each split in EgoRe-5M provides essential, complementary information for different aspects of egocentric reasoning, and that jointly leveraging them enables more holistic understanding of first-person activities.

Comparison of SFT and RFT. Table 6 presents a direct comparison of SFT and RFT on the EgoRe-5M-FG split. We can observe that compared to SFT, utilizing RFT significantly enhances the model's

performance across all tasks. Additionally, utilizing SFT affects performance on Egoschema and QAEgo4D, whereas RFT maintains robust performance across these benchmarks. This demonstrates that our method not only enhances spatio-temporal perception but also preserves the model's ability to generalize across diverse egocentric tasks.

Table 6: Ablations on different training paradigms.

Stage	Method	Data	EK-Visor		Egoexolearn		EgoSchema(val)	QAEgo4D
			mIoU	Loc-Acc.	mIoU	R1@0.05	Acc.	Acc.
Stage1	Baseline	-	28.6	64.5	1.53	5.4	68.2	60.3
	+ SFT	SFT Data	28.5	64.4	1.81	7.0	**71.9**	**67.2**
Stage2	Stage1 + SFT	EgoRe-5M-FG	38.9	74.1	9.84	24.9	71.4	62.1
	Stage1 + RFT	EgoRe-5M-FG	**53.7**	**80.3**	**25.2**	**63.9**	71.8	66.1

Input Frame Number. We investigate the impact of input sequence length by adjusting the number of sampled frames from 1 to 64 during inference, with the results presented in Figure 3. Across all benchmarks, it is evident that as the number of frames increases, the performance improves progressively. Notably, a marked decline in performance is observed when only a single frame is used, especially for EgoSchema. This indicates that long-term reasoning is an intrinsic need in egocentric reasoning.

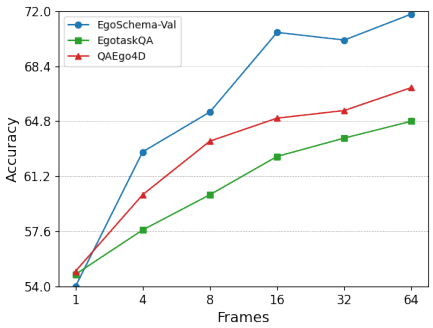

Figure 3: Ablations on number of frames.

We investigate the impact of input sequence length by adjusting the number of sampled frames from 1 to 64 during inference, with the results presented in Figure 3. Across all benchmarks, it is evident that as the number of frames increases, the performance improves progressively. Notably, a marked decline in performance is observed when only a single frame is used, especially for EgoSchema. This indicates that long-term reasoning is an intrinsic need in egocentric reasoning.

Ablations on hallucination detection. To validate whether our model exhibits hallucinations, we conduct evaluations on two widely-adopted benchmarks: (1) VideoHallucer [88] for assessing MLLMs, which contains object-relation, temporal, semantic detail and extrinsic hallucination detection. (2) POPE (Polling-based Object Probing) [53] for object hallucination detection.

The experimental results in Table 7 demonstrate a marginal performance decrease (0.6%) on VideoHallucer, while achieving a significant improvement of 3.2% on the POPE benchmark. We attribute this differential performance to our model's enhanced hand-object grounding capability, which particularly enhances its robustness against object hallucination.

Table 7: Results on hallucination detection benchmarks.

Method	VideoHallucer	POPE
Qwen2VL [82]	47.6	83.6
EgoThinker	47.0	86.8
GPT4o	53.3	-

Grounding Data and Egocentric Video Understanding. To find the relationship between the spatio-temporal grounding data and egocentric video understanding, we conduct further analysis on these benchmarks. We found that benchmarks such as EgoTaskQA, QAEgo4D, EgoSchema, and EgoPlan focus on high-level understanding, planning, and reasoning. They only require the identification of simple objects and do not involve fine-grained interaction details. Our baseline model, Qwen2-VL, already possesses a certain level of capability, which explains why our grounding data did not yield significant improvements.

Previous works [68, 43] demonstrated that using hand and object as supervisory signals enhances egocentric video understanding abilities. Therefore, we select three more fine-grained benchmarks: ERQA, EgoMCQ and EGTEA for evaluation.

As shown in Table 8, the inclusion of grounding data significantly improved model performance across all benchmarks. These datasets involve tasks directly related to our hand-object grounding data, such as hand/robotic arm motion, object identification, and object classification.

Table 8: Ablations on the grounding data.

Method	ERQA	EgoMCQ	EGTEA
Qwen2-VL	37.0	86.4/34.1	32.4
EgoThinker(without grounding data)	40.1	87.6/38.3	33.1
EgoThinker	41.8	89.3/41.4	35.4

4.3 Qualitative Results

In Figure 4, we show the hand–object grounding and reasoning traces for 4 methods. Our baseline Qwen2-VL performs poorly on both hand and object grounding. GPT-4o [38] generates a coherent chain-of-thought but misidentifies the target object (mislabeling the knife) and inaccurately localizes the hand. Grounding-DINO [58] is an expert model specialized in object grounding, however, it cannot distinguish left from right hand. In contrast, EgoThinker first approximates the object location and then outputs an accurate bounding box after thinking, demonstrating its superior fine-grained spatial reasoning in egocentric contexts.

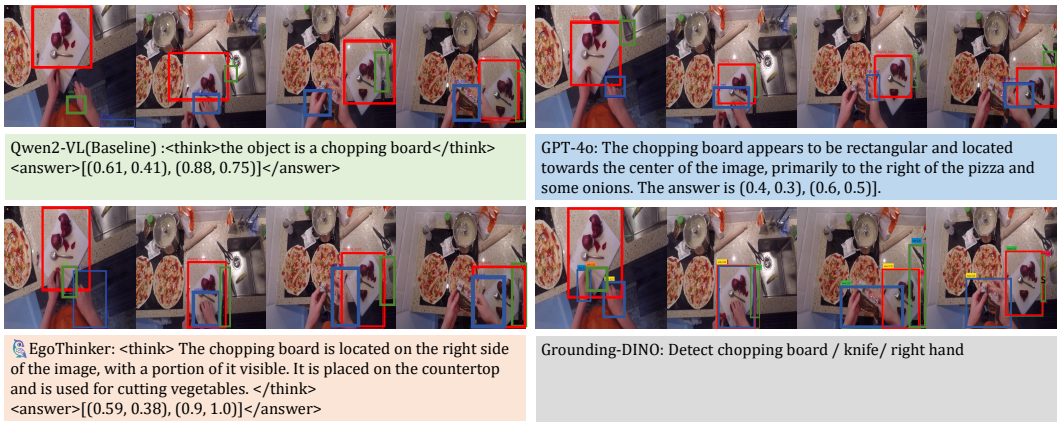

Figure 4: Hand-object grounding visualization on EK-Visor dataset. We compare our method to baseline Qwen2-VL, GPT-4o and expert model Grounding-DINO. We utilize different prompts tailored to each model and for each image, we use ”chopping board”, ”knife”, ”right hand” as query for grounding.

5 Conclusion

We introduce EgoThinker, a framework toenhance egocentric reasoning in MLLMs. By constructing EgoRe-5M, a large-scale egocentric instruction-tuning dataset, we provide the rich, structured data required for human-like spatio-temporal understanding and reasoning. To efficiently leverage our data, we employ the SFT-RFT combined paradigm to further equip EgoThinker with robust causal planning, long-horizon context integration, and precise localization capabilities. Experimental results demonstrate that EgoThinker achieves state-of-the-art performance across multiple egocentric benchmarks and significantly improves performance on spatio-temporal perception tasks, while maintaining general video understanding abilities. Despite these advantages, EgoThinker possesses certain limitations, such as reliance on extensive annotations and offline fine-tuning, and cannot perform real-time inference in resource-constrained environments. Future work will focus on real-time adaptation, richer multimodal integration, and self-supervised learning to further enhance its robustness and efficiency and we hope to extend EgoThinker into the field of Embodied AI, enabling more interactive agent behaviors in real-world environments.

Acknowledgement This work is funded in part by the National Key R&D Program of China (2022ZD0160201), JSPS KAKENHI JP25K24384, and Shanghai Artificial Intelligence Laboratory.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Susan P Ashdown and Marilyn DeLong. Perception testing of apparel ease variation. *Applied Ergonomics*, 26(1):47–54, 1995.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[5] Jing Bi, Yunlong Tang, Luchuan Song, Ali Vosoughi, Nguyen Nguyen, and Chenliang Xu. Eagle: Egocentric aggregated language-video engine. *arXiv preprint arXiv:2409.17523*, 2024.

[6] Aaron Blakeman, Aarti Basant, Abhinav Khattar, Adithya Renduchintala, Akhiad Bercovich, Aleksander Ficek, Alexis Bjorlin, Ali Taghibakhshi, Amala Sanjay Deshmukh, Ameya Sunil Mahabaleshwarkar, et al. Nemotron-h: A family of accurate and efficient hybrid mamba-transformer models. *arXiv preprint arXiv:2504.03624*, 2025.

[7] Zesen Cheng Zhiqiang Hu Yuqian Yuan Guanzheng Chen Sicong Leng Yuming Jiang Hang Zhang Xin Li Peng Jin Wenqi Zhang Fan Wang Lidong Bing Deli Zhao Boqiang Zhang, Kehan Li. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. URL https://arxiv.org/abs/2501.13106.

[8] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[9] Dibyadip Chatterjee, Fadime Sener, Shugao Ma, and Angela Yao. Opening the vocabulary of egocentric actions. *Advances in Neural Information Processing Systems*, 36:33174–33187, 2023.

[10] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *arXiv preprint arXiv:2211.09529*, 2022.

[11] 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.

[12] Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, et al. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. *arXiv preprint arXiv:2504.15271*, 2025.

[13] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

[14] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*, 2024.

[15] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. *CoRR*, 2023.

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

[17] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

[19] 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.

[20] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.

[21] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12943, 2024.

[22] Victor Escorcia, Ricardo Guerrero, Xiatian Zhu, and Brais Martinez. Sos! self-supervised learning over sets of handled objects in egocentric action recognition. In *European Conference on Computer Vision*, pages 604–620. Springer, 2022.

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

[24] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021.

[25] 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.

[26] 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.

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

[28] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment network for long-term video. In *CVPR*, 2022.

[29] Yuping He, Yifei Huang, Guo Chen, Lidong Lu, Baoqi Pei, Jilan Xu, Tong Lu, and Yoichi Sato. Bridging perspectives: A survey on cross-view collaborative intelligence with egocentric-exocentric vision. *arXiv preprint arXiv:2506.06253*, 2025.

[30] Yuping He, Yifei Huang, Guo Chen, Baoqi Pei, Jilan Xu, Tong Lu, and Jiangmiao Pang. Egoexobench: A benchmark for first-and third-person view video understanding in mllms. *arXiv preprint arXiv:2507.18342*, 2025.

[31] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

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

[41] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In *European Conference on Computer Vision*, pages 767–786. Springer, 2020.

[42] 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.

[43] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11107–11116, 2021.

[44] Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, et al. Token-efficient long video understanding for multimodal llms. *arXiv preprint arXiv:2503.04130*, 2025.

[45] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[46] Yi Wang Yizhuo Li Wenhai Wang Ping Luo Yali Wang Limin Wang KunChang Li, Yinan He and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[47] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

[48] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

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

[50] Jia Li, Ge Li, Yongmin Li, and Zhi Jin. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–23, 2025.

[51] 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.

[52] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

[53] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[54] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025.

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

[56] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[58] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[59] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

[60] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[61] Lidong Lu, Guo Chen, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms. *arXiv preprint arXiv:2506.05328*, 2025.

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

[63] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

[64] 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.

[65] 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.

[66] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[67] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761, 2023.

[68] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, et al. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024.

[69] Baoqi Pei, Yifei Huang, Guo Chen, Jilan Xu, Yali Wang, Limin Wang, Tong Lu, Yu Qiao, and Fei Wu. Guiding audio-visual question answering with collective question reasoning. *International Journal of Computer Vision*, pages 1–18, 2025.

[70] Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, et al. Modeling fine-grained hand-object dynamics for egocentric video representation learning. *arXiv preprint arXiv:2503.00986*, 2025.

[71] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[72] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19935–19947, 2022.

[73] 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.

[74] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[75] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv e-prints*, pages arXiv–2403, 2024.

[76] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

[78] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[79] Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

[80] Daniel Theron. Sentence transformer for audit retrieval question-answering (star-qa), Feb 2024. URL `https://huggingface.co/dptrsa/STAR-QA`.

[81] Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. Cosmo: Contrastive streamlined multimodal model with interleaved pre-training. *arXiv preprint arXiv:2401.00849*, 2024.

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

[83] Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Li, Jose M Alvarez, Lei Zhang, and Zhiding Yu. Videoitg: Multimodal video understanding with instructed temporal grounding. *arXiv preprint arXiv:2507.13353*, 2025.

[84] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[85] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024.

[86] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

[87] Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*, 2023.

[88] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024.

[89] 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.

[90] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

[91] 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.

[92] 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.

[93] 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.

[94] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. Egolife: Towards egocentric life assistant, 2025. URL `https://arxiv.org/abs/2503.03803`.

[95] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

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

[97] Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. Mm-ego: Towards building egocentric multimodal llms. *arXiv preprint arXiv:2410.07177*, 2024.

[98] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[99] Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*, 2025.

[100] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Helping hands: An object-aware ego-centric video recognition model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13901–13912, 2023.

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

[102] Haoyu Zhang, Qiaohui Chu, Meng Liu, Yunxiao Wang, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, Yaowei Wang, and Liqiang Nie. Exo2ego: Exocentric knowledge guided mllm for egocentric video understanding. *arXiv preprint arXiv:2503.09143*, 2025.

[103] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

[104] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[105] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding, 2023.

[106] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. URL https://arxiv.org/abs/2410.02713.

[107] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

[108] Yue Zhao and Philipp Krähenbühl. Training a large video model on a single machine in a day. *arXiv preprint arXiv:2309.16669*, 2023.

[109] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018.

[110] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.

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

[112] Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. *arXiv preprint arXiv:2502.07411*, 2025.

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: We have written contributions and scope in the abstact.

   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: We discuss the limitations and our future works in the conclusion.

   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: We do 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: We provide the statistic of the data and hyperparameters in training.

   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: [No]

Justification: We provide the statistic of the data and training setting in the supplementary, but we will not provide data and code in the submission.

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: We provide these information in the supplementary.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: Error bars are not reported because it would be computationally expensive.

   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: We provide the training cost and time for SFT and RFT training in the experiment section and supplementary.

   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: We follow 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: [No]

    Justification: This work is a foundational research, and will not have societal impacts.

    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: Our work has no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We have cited the original paper that produced the code package or dataset.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: We do not release new assets.

    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: [NA]

    Justification: We do not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: This paper does not involve LLMs as any important, original, or non-standard components.

Guidelines:

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

A Details About EgoRe-5M

Video Source. After obtaining our filtered 8.7M video clips from web data, we combine them with existing egocentric datasets including Ego4D [25], EPIC-Kitchens [18], EgoExoLearn [36], and EgoExo4D [26] to form a comprehensive dataset totaling 13 million clips. Specifically, we utilize 4M clips from [73] within Ego4D. For EPIC-Kitchens, we select 56,000 frames from the EK-Visor dataset. From EgoExoLearn, we carefully curated 10,000 L1-level video clips for temporal grounding training and 400 samples for the RES benchmark. Regarding EgoExo4D, we incorporated 500 samples for the RES benchmark evaluation.

Dynamic Interaction Filtering. Since many video clips remain static or depict group activities, we need to filter out irrelevant clips for egocentric reasoning. After obtaining object bounding boxes for each frame using a hand-object detector, we design the following filtering rules to select video clips with dynamic interaction:

Step 1: For all video clips, we first examine the hand bounding boxes. If the number of hands detected exceeds two ($N_{hands} > 2$) indicating multi-person activities, we discard such clips:

$$\text{Filter Clip} \Longleftrightarrow N_{hands} > 2 \tag{3}$$

Step 2: For the remaining clips, given a clip C with N frames, we set a threshold $\alpha = 0.7$. The clip is discarded if the total number of object bounding boxes is less than $\alpha \times N$:

$$\text{Filter Clip} \Longleftrightarrow \sum_{i=1}^{N} N_{objects}^{(i)} < \alpha \times N \tag{4}$$

Step 3: After the first two steps, we obtain clips containing single-person hand-object interactions. To further select dynamic clips, we calculate the maximum inter-frame hand displacement. For clip C, we compute the hand center (x_t, y_t) in each frame t. The clip is kept if the maximum center displacement exceeds 10% of the image size ($min(H, W)$):

$$\text{Filter Clip} \Longleftrightarrow \sum_{t1=1,t2=1}^{H,W} \sqrt{(x_{t_1} - x_{t_2})^2 + (y_{t_1} - y_{t_2})^2} > 0.1 \times min(H, W) \tag{5}$$

Through these three steps, we filter out high-quality dynamic video clips, which can be used to train our egocentric reasoning MLLM.

Data Statistics. Table 9 presents the details of EgoRe-5M, including 16 question types and corresponding QA examples for each type. Notably, in the Chain-of-Thought (CoT) split, we provide detailed and meaningful reasoning processes to the question. For the Fine-Grained split, we introduce two special tokens, `<think>` and `<answer>`, to structurally format the reasoning process and final answers, which facilitates the execution of our reward function during model training.

B Training

SFT Data. To balance training efficiency and model performance, we carefully curated our training dataset as shown in Table 1. While using the complete dataset would lead to prohibitive computational costs and performance degradation due to data imbalance, We filter each dataset : for video caption dataset, we select 170K samples on total; for ego-related dataset, we select 390k QA samples in total; for our EgoRe-5M, we select 810K samples, including 410K from short-term splits, 400K from long-term split and 50K from CoT split.

Training Details. We use QwenVL2-7B as our baseline for training. For SFT, we adpot $max\_pixels = 200704$, $min\_pixels = 3136$, $lr = 1e-6$, $epoch = 1$ for training. We utilize 32 A100 GPUs and train for 30 hours. For RFT, we adpot $lr = 1e-5$, $epoch = 1$ for training. We utilize 8 A100 GPUs and train for 12 hours. Notably, during RFT training phase, we first train hand-object grounding task and then train temporal grounding task.

Data Split	Question Type	Number	Example
Short-term (0-10s)	Object Existence	302K	*What object is the person interacting with in the video?* The person is interacting with a grass-trimming tool.
	Object Attribute	326K	*What is the state of the garlic during the slicing process?* The garlic is being sliced and held in place by the left hand.
	Object Count	444K	*How many people are in the video?* There are three people in the video.
	Object Interaction	280K	*What is the chef doing with the white soup bowl?* The chef is rinsing the white soup bowl under running water to ensure it is thoroughly clean.
	Action Description	280K	*What actions are being performed by the hands in the video?* The left hand moves downward to grasp an object, while the right hand pulls a handle to the side.
	Action Reasoning	440K	*Why does the left hand move to the center of the frame?* The left hand moves to the center of the frame to pick up a book.
	Background Attribute	293K	*What is the setting or background of the video?* The background appears to be an outdoor area with grass, likely a lawn or garden.
Long-term (15-120s)	Action Sequence	412K	*What is the sequence of actions performed by the left and right hands when cutting and placing mango pieces?* The left hand moves downward to place the mango on the fruit cutter while the right hand holds the cutter steady. Then, the left hand stabilizes the mango while the right hand cuts it. after cutting, the left hand moves downward to place the mango piece while the right hand remains still.
	Temporal Grounding	411K	*When does person z demonstrate a clear preference for using their left hand?* Person z demonstrates a clear preference for using their left hand when they remove the cap and place it on the table, which occurs between 7.33s to 10.25s.
	Object Count	411K	*How many objects are being interacted with across the video clips?* The objects being interacted with include an art brush, paint, an art board, a piece of paper, tissue, and a paint palette, totaling six objects.
	Action Prediction	410K	*After person c places a piece of cloth into the bag, what is the likely next action?* The likely next action is person c sorting the clothes in the bag.
	Action Summary	411K	*What are the key actions involving the left hand across the video clips?* The key actions involving the left hand include person y picking up the phone, person z removing and placing the cap on the table, person z picking up a camera from person k, and the left hand moving in various directions while interacting with objects.
	Action Reasoning	412K	*Why does the person use both hands to apply plaster to the ceiling?* The person uses both hands to ensure precision and control while applying plaster. One hand scoops and guides the plaster, while the other maneuvers the scraper, allowing for an even and smooth application on the ceiling.
CoT (15-200s)	Chain-of-thought Reasoning	50K	*How did the person adjust their hand positioning while operating the lawnmower?* As described in 'Person uses both hands to lift a shovel'), then dug into the soil with it ('Person digs into the soil with a shovel, using both hands'). After digging, he spread the soil using his left hand while steadying a container with his right ('Man A spreads soil with his left hand, while his right hand steadies the container'). Finally, after completing the task, he carefully placed the shovel on the grass by guiding it down with his left hand and steadying the handle with his right ('Man A carefully places the shovel on the grass, his left hand guiding it down while his right hand steadies the handle'). This sequence shows a logical workflow: lifting the tool, executing the primary action (digging) distributing the material, and safely storing the tool afterward.
Fine-grained Grounding	Temporal Grounding	10K	*To accurately pinpoint the event "[QUESTION]" in the video, you need to identify a time interval from which the answer to the question can be deduced. Output your thought process within the <think></think>tags. Then, provide the start and end times (in seconds, precise to two decimal places) in the format "(start,end)" within the <answer></answer>tags.*
	Hand-Object Grounding	56K	*This is an image containing an object: "[OBJECT]" ,and output the bounding box of this object in the image. Output your thought process within the <think></think>tags. Then provide your bounding box within the <answer></answer>tags,following <answer>(x_min,y_min),(x_max,y_max) </answer>format. The bounding box coordinates are normalized to the range [0, 1], relative to the width and height of the image.*

Table 9: **Statistics of the proposed EgoRe-5M.** The table shows 4 data splits and 16 question types, along with the corresponding example question-answer pairs. More details can be found in Section 3.1.2.

C Benchmark Details

EgoTaskQA. EgoTaskQA [42] is a large-scale egocentric video questionanswering dataset designed to evaluate models' understanding of goal-oriented human tasks. It is derived from LEMMA dataset [41], focusing on aspects such as action effects, intent, multi-agent collaboration, and object interactions. The dataset emphasizes reasoning types including spatio-temporal understanding, causal dependencies, and task planning, supported by 30K annotated state transitions. It includes a variety of question formats, such as binary and open-ended queries, to ensure a balanced and unbiased evaluation. To evaluate this dataset, we reformulate the original open-ended QA samples into a multiple-choice question through a systematic conversion process. Specifically, we first aggregate all potential answers into a list. For each question, BERT [4] is used to compute semantic similarity scores between the ground-truth answer and all candidate answers in the pool. The four most semantically similar answers were then selected to construct the new multiple-choice question, with the ground-truth answer serving as the correct option.

QAEgo4D. QAEgo4D represents a specialized benchmark for assessing episodic memory through video-based question answering. This dataset, derived from the Ego4D, measures the ability of vision-language models to comprehend and reason about dynamic visual sequences. Each entry consists of four key components: (1) an egocentric video clip, (2) a manually crafted question, (3) its corresponding answer, and (4) precise temporal localization of the relevant visual evidence. To ensure annotation quality, the dataset employs redundant textual descriptions that undergo cross-verification. QAEgo4D provides researchers with a robust framework for investigating memory-related video understanding tasks. To evaluate the dataset, we select the closed-set QA split parsed by [92].

EgoPlan. EgoPlan serves as a multimodal benchmark for evaluating human-like planning abilities in AI systems through egocentric video understanding. Derived from large-scale egocentric datasets including Epic-Kitchens and Ego4D, the benchmark comprises 4,939 rigorously validated multiple-choice questions, spanning 3,296 distinct task objectives and 3,185 executable action sequences across 419 diverse real-world environments. By simulating real-world decision-making scenarios, the benchmark facilitates progress in multimodal reasoning for practical planning applications. In our experiments, we adopt the dataset's predefined validation split for evaluating planning performance, as ground-truth annotations for the test set remain undisclosed.

EgoSchema. EgoSchema represents a novel benchmark framework for assessing long-form video comprehension in multimodal AI systems. Derived from the Ego4D video corpus, this evaluation suite comprises 5,000+ meticulously annotated multiple-choice question-answer pairs, sourced from 250+ hours of unscripted daily human activities captured in real-world settings. The benchmark presents a unique challenge where AI models must analyze three-minute video clips and select the most accurate response from five plausible alternatives, testing their capacity for sustained visual-temporal reasoning and contextual understanding.

EgoMCQ. EgoMCQ is a multiple-choice question-answering dataset designed to assess video-text alignment in egocentric vision systems. Derived from Ego4D, it includes 39,000 questions based on 468 hours of egocentric video covering a wide range of human activities. Each question involves selecting the correct video clip from five options based on a narration, with two settings: "inter-video", for distinguishing between different videos, and "intra-video", for fine-grained context within the same video.

VLN-QA. VLN-QA represents a specialized evaluation benchmark for assessing multimodal navigation understanding in indoor environments through question-answering tasks. Derived from the VLN-CE framework, this dataset comprises thousands of carefully annotated multiple-choice items paired with egocentric video sequences that replicate authentic navigation scenarios. The benchmark specifically examines a system's ability to interpret visual-spatial information and correlate it with textual queries. For our implementation, we utilize the preprocessed dataset version established in VideoChat2's experimental setup.

RES. To validate our model's cross-view reasoning capability, we developed a Cross-View Skill Transfer Benchmark named RES (Referenced Egocentric Skill). RES leverages paired exocentric–egocentric clips from the EgoExoLearn and EgoExo4D datasets. Each example presents one exocentric video as a reference and four candidate egocentric clips, and the model must identify which egocentric view corresponds to the reference. This multi-choice protocol rigorously tests the ability to transfer observed skills across perspectives. The final benchmark comprises 936 curated samples. Although RES is crafted to validate EgoThinker's cross-view reasoning, we anticipate it will become a valuable resource for the broader embodied AI community.

Grounding Benchmark. For the grounding benchmark construction, we select existing annotations to derive our evaluation dataset. Specifically, for the hand-object grounding benchmark, we curated our dataset from EK-Visor, which provides bounding box annotations. Our methodology involved extracting bounding boxes from segmentation masks in the validation set, serving as ground-truth references. This process yielded a comprehensive collection of 13,000 object queries for evaluation purposes. For the temporal grounding task, we strategically selected the EgoExoLearn dataset due to its unique dual-level annotation structure, which makes it suitable for temporal localization tasks. We select L1-level (coarse-grained) video clips as our primary video sources and L2-level (fine-grained)

temporal windows as precise ground truth annotations. To this end, we curate an evaluation set of 3,000 test instances.

D Additional Experiments

Effects Of Extra Video Sources. Table 10 provides a comparative analysis of model performance with and without the inclusion of QA samples sourced from the HowTo100M dataset. The results indicate consistent performance improvements across all evaluated benchmarks when leveraging the HowTo100M-derived data, with particularly notable gains on long-term understanding tasks such as QAEgo4D and EgoPlan. We attribute these improvements to the long-term split in EgoRe-5M, which is primarily derived from HowTo100M, significantly enhancing the model's capacity for extended temporal reasoning. These findings underscore the effectiveness of our data curation strategy in enabling robust egocentric reasoning ability.

Table 10: Ablations on our EgoRe-5M. We evaluate the impact of incorporating data filtered from the HowTo100M dataset on performance.

Data	**EgoTaskQA** Acc.	**QAEgo4D** Acc.	**EgoPlan-Val** Acc.	**VLN-QA** Acc.
Baseline	57.9	60.3	38.3	42.0
w/o Howto100M	62.2	61.6	41.3	50.0
w Howto100M	**64.4**	**66.2**	**47.1**	**54.0**

Results On General Grounding Task. To further validate EgoThink's grounding capabilities, we conduct additional experiments on the COCO dataset [57]. As evidenced by Table 11, our model demonstrates significant performance improvements despite never being trained on COCO data, which substantiates its strong generalization ability for object grounding tasks.

Table 11: Results on COCO detection dataset.

Method	**testA** mIoU	**testB** mIoU
Qwen2VL-7B	34.1	33.6
EgoThinker	55.2(+21.1)	57.8(+24.2)

Additional Visualization Results. Figure 5 compares temporal grounding outputs for the baseline, Gemini 2.5-Pro [77], and EgoThinker on a sample query. We observe that Qwen2-VL fails to identify the query and lacks a reasoning process in both examples. Gemini2.5-Pro provides dense video captions and can accurately align the captions with the corresponding events. However, the experimental results reveal that the model exhibits inferior localization performance when processing scenes with hand-object interactions. EgoThinker combines an explicit reasoning trace with precise temporal window prediction, accurately identifying the start and end times of the queried event. These visualizations underscore EgoThinker's unique ability to deliver both interpretable reasoning and robust spatio-temporal grounding in first-person video.

Video: 

Query: *With his right hand holding the pot, he walked towards the sink. With his left hand, he turned on the faucet, filling up the pot with water. Then, he used his left hand to turn off the faucet.*

GT: **3.0s** ←- - - - - - - - - - - - - - - - - - - → **12.4s**

Qwen2-VL(Baseline): The start and end times are (0.00, 3.45).
0s ←- - - - - - - - - - - → **3.45s**

Gemini-Pro: <answer>(0:02.30,0:11.70)</answer>Let's re-verify.0:00 - 0:01: Person is at the stove, puts down a ladle.0:02: Person picks up the wok (pot) with their right hand.0:03: Person is walking towards the sink with the wok in their right hand.0:04: Person places the wok in the sink.0:05: Person uses their left hand to turn on the faucet. Water starts flowing into the wok.0:06 - 0:10: Wok is filling with water.0:11: Person uses their left hand to turn off the faucet.0:12: Faucet is completely off.So the start time is when the person picks up the pot to walk to the sink (0:02), and the end time is when the faucet is turned off (0:12).The description "With his right hand holding the pot, he walked towards the sink" starts at 0:02.
2.30s ←- → **11.70s**

Egothinker: <think>the event "With his right hand holding the pot, he walked towards the sink. With his left hand, he turned on the faucet, filling up the pot with water. Then, he used his left hand to turn off the faucet." occurs in the first part of the video</think>\n<answer>(3.44, 10.44)</answer>
3.44s ←- - - - - - - - - - - - - - - - → **10.44s**

Video: 

Query: *Twist the knob of the gas stove with your right hand. Turn on the flame.*

GT: **12.4s** ←- - - - - - - - - - - - - - - → **17.0s**

Qwen2-VL(Baseline): The start and end times are (1.1, 23.1).
1s ←- → **23.1s**

Gemini-Pro: <answer>(0:13,0:15)</answer>Locating Start/End TimesOkay, I understand the task. I need to pinpoint the precise timestamps in the video where the action begins and concludes. I'll focus on the visual cues: the right hand approaching the knob, the twisting motion, and the appearance of the flame. It's a straightforward analysis, really. The action "Twist the knob of the gas stove with your right hand. Turn on the flame" occurs in the video between 0:13 and 0:16.

13s ←- - - - - - - - - - - - - - - - → **16s**

Egothinker: <think>the event "Twist the knob of the gas stove with your right hand. Turn on the flame." occurs after the person puts the pot on the gas stove and before the person opens the fridge</think>\n<answer>(10.0, 16.0)</answer>

10s ←- → **16s**

Figure 5: Temporal grounding visualization on the EgoExoLearn dataset. We compare our method to baseline Qwen2-VL and one of the strongest MLLM Gemini2.5-Pro.