

---

# EgoThinker: Unveiling Egocentric Reasoning with Spatio-Temporal CoT Supplementary Material

---

Anonymous Author(s)

Affiliation

Address

email

## 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 [3], EPIC-Kitchens [2], EgoExoLearn [5], and EgoExo4D [4] to form a comprehensive dataset totaling 13 million clips. Specifically, we utilize 4M clips from [9] 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:

$$\text{Filter Clip} \iff N_{hands} > 2 \quad (1)$$

*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  $\alpha \times N$ :

$$\text{Filter Clip} \iff \sum_{i=1}^N N_{objects}^{(i)} < \alpha \times N \quad (2)$$

*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 ( $\min(H, W)$ ):

$$\text{Filter Clip} \iff \sum_{t1=1, t2=1}^{H, W} \sqrt{(x_{t1} - x_{t2})^2 + (y_{t1} - y_{t2})^2} > 0.1 \times \min(H, W) \quad (3)$$

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 1 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

Table 1: **Statistics of the proposed EgoRe-5M.** The table shows 4 data splits and 16 question types, along with the corresponding example question-answer pairs.

Data Split	Question Type	Number	Example
Short-term (0-10s)	Object Existence	302K	<i>What object is the person interacting with in the video?</i> The person is interacting with a grass-trimming tool.
	Object Attribute	326K	<i>What is the state of the garlic during the slicing process?</i> The garlic is being sliced and held in place by the left hand.
	Object Count	444K	<i>How many people are in the video?</i> There are three people in the video.
	Object Interaction	280K	<i>What is the chef doing with the white soup bowl?</i> The chef is rinsing the white soup bowl under running water to ensure it is thoroughly clean.
	Action Description	280K	<i>What actions are being performed by the hands in the video?</i> The left hand moves downward to grasp an object, while the right hand pulls a handle to the side.
	Action Reasoning	440K	<i>Why does the left hand move to the center of the frame?</i> The left hand moves to the center of the frame to pick up a book.
	Background Attribute	293K	<i>What is the setting or background of the video?</i> The background appears to be an outdoor area with grass, likely a lawn or garden.
Long-term (15-120s)	Action Sequence	412K	<i>What is the sequence of actions performed by the left and right hands when cutting and placing mango pieces?</i> 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	<i>When does person z demonstrate a clear preference for using their left hand?</i> 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	<i>How many objects are being interacted with across the video clips?</i> 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	<i>After person c places a piece of cloth into the bag, what is the likely next action?</i> The likely next action is person c sorting the clothes in the bag.
	Action Summary	411K	<i>What are the key actions involving the left hand across the video clips?</i> 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.
CoT (15-200s)	Action Reasoning	412K	<i>Why does the person use both hands to apply plaster to the ceiling?</i> 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.
	Chain-of-thought Reasoning	50K	<i>How did the person adjust their hand positioning while operating the lawnmower?</i> 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	<i>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 &lt;think&gt;&lt;/think&gt; tags. Then, provide the start and end times (in seconds, precise to two decimal places) in the format "(start,end)" within the &lt;answer&gt;&lt;/answer&gt; tags.</i>
	Hand-Object Grounding	56K	<i>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 &lt;think&gt;&lt;/think&gt; tags. Then provide your bounding box within the &lt;answer&gt;&lt;/answer&gt; tags, following &lt;answer&gt;(x_min,y_min),(x_max,y_max) &lt;/answer&gt; format. The bounding box coordinates are normalized to the range [0, 1], relative to the width and height of the image.</i>

25 detailed and meaningful reasoning processes to the question. For the Fine-Grained split, we introduce  
 26 two special tokens, `<think>` and `<answer>`, to structurally format the reasoning process and final  
 27 answers, which facilitates the execution of our reward function during model training.

## 28 B Training

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

35 **Training Details.** We use QwenVL2-7B as our baseline for training. For SFT, we adpot  
 36  $max\_pixels = 200704$ ,  $min\_pixels = 3136$ ,  $lr = 1e - 6$ ,  $epoch = 1$  for training. We utilize

32 A100 GPUs and train for 30 hours. For RFT, we adopt  $lr = 1e - 5$ ,  $epoch = 1$  for training.  
 38 We utilize 8 A100 GPUs and train for 12 hours. Notably, during RFT training phase, we first train  
 39 hand-object grounding task and then train temporal grounding task.

## 40 C Benchmark Details

41 **EgoTaskQA.** EgoTaskQA [7] is a large-scale egocentric video question-answering dataset designed  
 42 to evaluate models’ understanding of goal-oriented human tasks. It is derived from LEMMA  
 43 dataset [6], focusing on aspects such as action effects, intent, multi-agent collaboration, and object  
 44 interactions. The dataset emphasizes reasoning types including spatio-temporal understanding, causal  
 45 dependencies, and task planning, supported by 30K annotated state transitions. It includes a variety  
 46 of question formats, such as binary and open-ended queries, to ensure a balanced and unbiased  
 47 evaluation. To evaluate this dataset, we reformulate the original open-ended QA samples into a  
 48 multiple-choice question through a systematic conversion process. Specifically, we first aggregate  
 49 all potential answers into a list. For each question, BERT [1] is used to compute semantic similarity  
 50 scores between the ground-truth answer and all candidate answers in the pool. The four most  
 51 semantically similar answers were then selected to construct the new multiple-choice question, with  
 52 the ground-truth answer serving as the correct option.

53 **QAEgo4D.** QAEgo4D represents a specialized benchmark for assessing episodic memory through  
 54 video-based question answering. This dataset, derived from the Ego4D, measures the ability of  
 55 vision-language models to comprehend and reason about dynamic visual sequences. Each entry  
 56 consists of four key components: (1) an egocentric video clip, (2) a manually crafted question, (3) its  
 57 corresponding answer, and (4) precise temporal localization of the relevant visual evidence. To ensure  
 58 annotation quality, the dataset employs redundant textual descriptions that undergo cross-verification.  
 59 QAEgo4D provides researchers with a robust framework for investigating memory-related video  
 60 understanding tasks. To evaluate the dataset, we select the closed-set QA split parsed by [11].

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

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

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

82 **VLN-QA.** VLN-QA represents a specialized evaluation benchmark for assessing multimodal  
 83 navigation understanding in indoor environments through question-answering tasks. Derived from the  
 84 VLN-CE framework, this dataset comprises thousands of carefully annotated multiple-choice items  
 85 paired with egocentric video sequences that replicate authentic navigation scenarios. The benchmark  
 86 specifically examines a system’s ability to interpret visual-spatial information and correlate it with

87 textual queries. For our implementation, we utilize the preprocessed dataset version established in  
88 VideoChat2’s experimental setup.

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

97 **Grounding Benchmark.** For the grounding benchmark construction, we select existing annotations  
98 to derive our evaluation dataset. Specifically, for the hand-object grounding benchmark, we curated  
99 our dataset from EK-Visor, which provides bounding box annotations. Our methodology involved  
100 extracting bounding boxes from segmentation masks in the validation set, serving as ground-truth  
101 references. This process yielded a comprehensive collection of 13,000 object queries for evaluation  
102 purposes. For the temporal grounding task, we strategically selected the EgoExoLearn dataset due to  
103 its unique dual-level annotation structure, which makes it suitable for temporal localization tasks. We  
104 select L1-level (coarse-grained) video clips as our primary video sources and L2-level (fine-grained)  
105 temporal windows as precise ground truth annotations. To this end, we curate an evaluation set of  
106 3,000 test instances.

## 107 D Additional Experiments

108 **Effects Of Extra Video Sources.** Table 2 provides a comparative analysis of model performance  
109 with and without the inclusion of QA samples sourced from the HowTo100M dataset. The results  
110 indicate consistent performance improvements across all evaluated benchmarks when leveraging the  
111 HowTo100M-derived data, with particularly notable gains on long-term understanding tasks such as  
112 QAEgo4D and EgoPlan. We attribute these improvements to the long-term split in EgoRe-5M, which  
113 is primarily derived from HowTo100M, significantly enhancing the model’s capacity for extended  
114 temporal reasoning. These findings underscore the effectiveness of our data curation strategy in  
115 enabling robust egocentric reasoning ability.

Table 2: Ablations on our EgoRe-5M. We evaluate the impact of incorporating data filtered from the HowTo100M dataset on performance.

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

116 **Results On General Grounding Task.** To further validate EgoThink’s grounding capabilities,  
117 we conduct additional experiments on the COCO dataset [8]. As evidenced by Table 3, our model  
118 demonstrates significant performance improvements despite never being trained on COCO data,  
119 which substantiates its strong generalization ability for object grounding tasks.

Table 3: 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)

120 **Additional Visualization Results.** Figure 1 compares temporal grounding outputs for the baseline,  
121 Gemini 2.5-Pro [10], and EgoThinker on a sample query. We observe that Qwen2-VL fails to

122 identify the query and lacks a reasoning process in both examples. Gemini2.5-Pro provides dense  
 123 video captions and can accurately align the captions with the corresponding events. However, the  
 124 experimental results reveal that the model exhibits inferior localization performance when processing  
 125 scenes with hand-object interactions. EgoThinker combines an explicit reasoning trace with precise  
 126 temporal window prediction, accurately identifying the start and end times of the queried event.  
 127 These visualizations underscore EgoThinker’s unique ability to deliver both interpretable reasoning  
 128 and robust spatio-temporal grounding in first-person video.

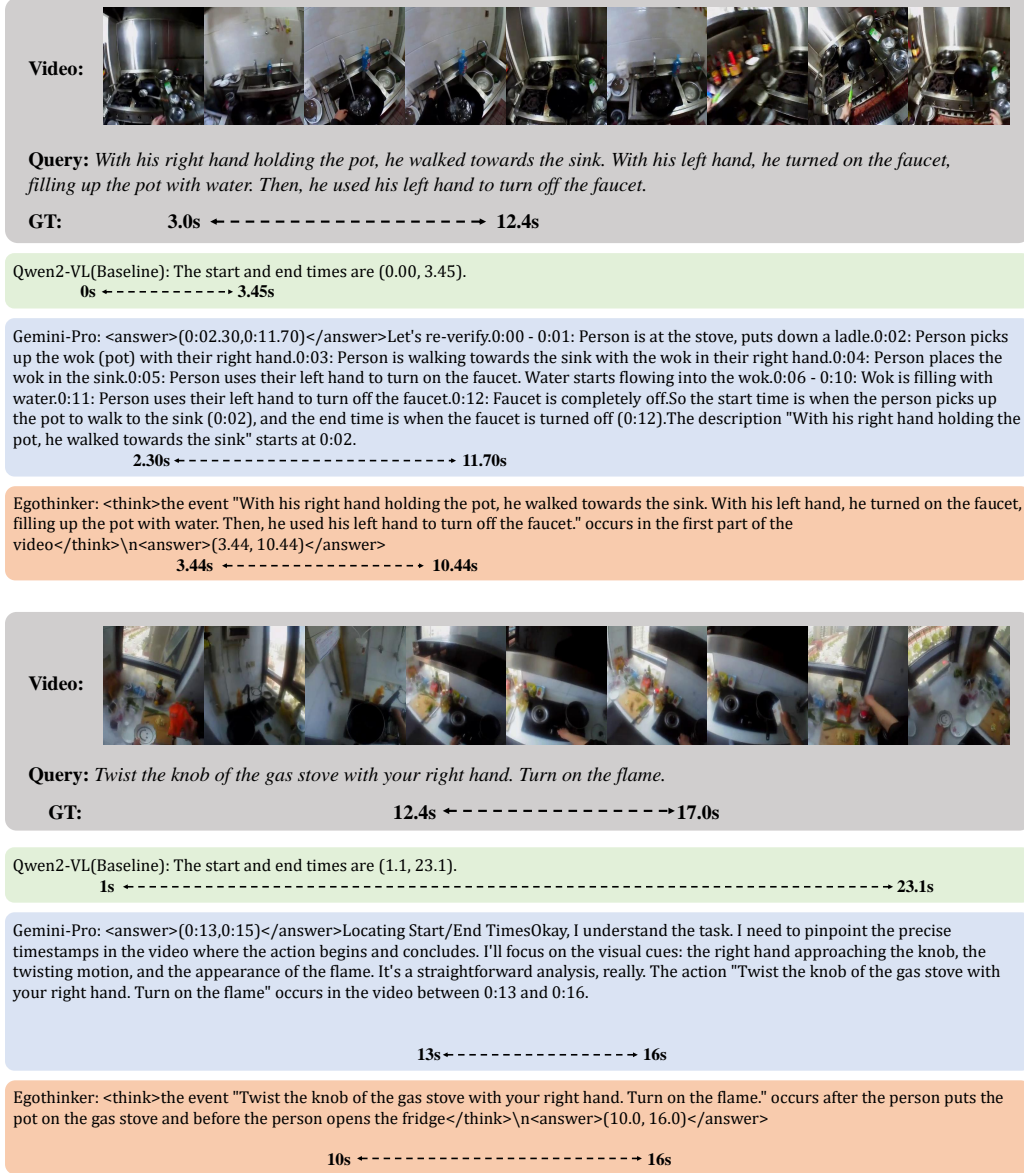


Figure 1: Temporal grounding visualization on the EgoExoLearn dataset. We compare our method to baseline Qwen2-VL and one of the strongest MLLM Gemini2.5-Pro.

## References

- [1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoxolearn: 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.
- [6] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *European Conference on Computer Vision*, pages 767–786. Springer, 2020.
- [7] 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.
- [8] 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.
- [9] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Ego4d-v2: 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.
- [10] 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.
- [11] 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.