MA-EgoQA: Question Answering over Egocentric Videos from Multiple Embodied Agents

Kangsan Kim[1], Yanlai Yang[2], Suji Kim[1,3], Woongyeong Yeo[1], Youngwan Lee[1,4], Mengye Ren[2†], and Sung Ju Hwang[1,5†]

[1] KAIST [2] New York University
[3] Samsung Electronics [4] ETRI [5] DeepAuto.ai
kangsan.kim@kaist.ac.kr

Abstract. As embodied models become powerful, humans will collaborate with multiple embodied AI agents at their workplace or home in the future. To ensure better communication between human users and the multi-agent system, it is crucial to interpret incoming information from agents in parallel and refer to the appropriate context for each query. Existing challenges include effectively compressing and communicating high volumes of individual sensory inputs in the form of video and correctly aggregating multiple egocentric videos to construct system-level memory. In this work, we first formally define a novel problem of understanding multiple long-horizon egocentric videos simultaneously collected from embodied agents. To facilitate research in this direction, we introduce MultiAgent-EgoQA (MA-EgoQA), a benchmark designed to systemically evaluate existing models in our scenario. MA-EgoQA provides 1.7k questions unique to multiple egocentric streams, spanning five categories: social interaction, task coordination, theory-of-mind, temporal reasoning, and environmental interaction. We further propose a simple baseline model for MA-EgoQA named EgoMAS, which leverages shared memory across embodied agents and agent-wise dynamic retrieval. Through comprehensive evaluation across diverse baselines and EgoMAS on MA-EgoQA, we find that current approaches are unable to effectively handle multiple egocentric streams, highlighting the need for future advances in system-level understanding across the agents. The code and benchmark are available at https://ma-egoqa.github.io.

1 Introduction

Recent advances are increasingly transferring progress from text-based environments to physical embodied domains, such as autonomous driving [11, 17] and industrial robotics [21, 48]. Just as large language models (LLMs) like ChatGPT have reshaped how humans interact with and reason over knowledge, we can expect a future where intelligent agents increasingly assist and augment our physical activities. As these agents become common and multiple agents begin to operate within a shared environment, a growing line of research explores how

† Equal advising.

Table 1: Comparison between MA-EgoQA and existing benchmarks.

Benchmarks	#QA	Avg Len	Ego	Days-Long	Cross-Video	ToM
EgoMemoria [40]	7k	∼60m	✓	✗	✗	✗
MuMA-ToM [31]	0.9k	36s	✗	✗	✗	✓
EgoSchema [27]	5.1k	180s	✓	✗	✗	✗
EgoThink [8]	0.7k	-	✓	✗	✗	✓
EgoToM [26]	1k	300s	✓	✗	✗	✓
EgoLifeQA [39]	6k	44h	✓	✓	✗	✗
CVBench [49]	1k	106s	✗	✗	✓	✗
EgoExoLearn [18]	2.2k	13m	✓	✗	✓	✗
MA-EgoQA	1.7k	266h	✓	✓	✓	✓

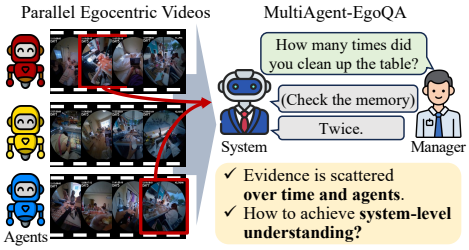


Fig. 1: Problem formulation of MultiAgent-EgoQA and the associated challenges.

they can interact and collaborate effectively as a system [10]. Multi-agent system (MAS) can achieve the goals efficiently by decomposing and executing tasks in parallel, and collective reasoning enables agents to leverage broader context and arrive at more optimal solutions [6, 16]. At the same time, multi-agent settings pose unique challenges, such as designing robust communication protocols and developing strategies for sub-task generation and allocation. Addressing these challenges is essential for developing embodied agent systems that can operate at scale and effectively handle the complexities of real-world environments.

However, contextual understanding and question answering (QA) remain largely underexplored in the study of multiple embodied agents. Prior works have predominantly focused on goal-directed algorithms, task allocation, and completion, particularly within the robotics domain [14, 44, 46]. While these directions are critical for system operation, effective actions ultimately depend on accurately interpreting event histories, environmental state, and inter-agent communication. This capability is especially critical in QA scenarios, such as when a human manager queries the system to monitor progress, as shown in Fig. 1. Importantly, the system must be able to integrate experiences across agents and retrieve relevant episodes to provide a correct answer. Such functionality has broad application in the real world, for example, querying anomalies across multiple bodycam feeds from police officers, or checking when household robots last cleaned the bathroom. As QA is fundamental for making multi-agent systems transparent, controllable, and manageable, it calls for systematic investigation and comprehensive evaluation to inform better system design.

A main challenge in video QA with multiple agents lies in managing extremely long egocentric video streams and aggregating the experience of each agent. Since embodied agents continuously capture their egocentric video streams during operation, models must be capable of reasoning over very long temporal horizons [28, 33, 41]. In realistic scenarios, embodied agents may operate for days, with each agent generating massive video histories, making it difficult to locate the query-related events and deliver accurate answers. In practice, current video LLMs are still limited to processing video [23] at the scale of only a few hours [34, 36], falling short of the requirements for multi-agent QA. Furthermore, the system should go beyond understanding the history of a single agent and integrate events across multiple agents to achieve system-level comprehension.

Therefore, we construct a new benchmark evaluating the models' understanding capability in multi-agent egocentric videos, named MultiAgent-EgoQA (MA-EgoQA). MA-EgoQA consists of 1.7k question-answer pairs based on the EgoLife egocentric video dataset [39]. Captured by 6 people living in a shared house for 7 days, the dataset provides five categories unique and fundamental in multi-agent scenarios: Social Interaction, Task Coordination, Theory of Mind, Temporal Reasoning, and Environmental Interaction. Questions and answers are generated using GPT-based pipelines, refined through LLM filtering, and ultimately validated by human annotators to ensure quality. We also propose a baseline model called EgoMAS, designed to illustrate the potential of multi-agent video QA. EgoMAS aggregates information from all agents into a shared memory, and upon receiving a query, it selectively retrieves the relevant information from the appropriate agents. Compared to the naive method concatenating all information from agents, selective retrieval based on the shared memory of EgoMAS is remarkably token-efficient and superior in QA performance.

We evaluate a range of existing LLMs and video LLMs on MA-EgoQA, including our EgoMAS model. Results reveal that current models struggle to handle multiple egocentric video streams effectively: even state-of-the-art LLMs and video LLMs fail to capture the complexities of multi-agent egocentric video understanding. In contrast, EgoMAS, despite its simplicity, outperforms the baseline by 4.48% within the same Gemini-2.5-Flash [9] backbone, and even EgoMAS with Qwen3VL-8B-Thinking [2] achieves a performance comparable to the Gemini baseline operating at a 1M-token context. These findings underscore both the difficulty of the task and the promise of specialized approaches like EgoMAS.

2 Related Work

2.1 Multiple Embodied Agents System

Existing works on multiple embodied agent systems have been developed, focusing on studying effective approaches to cooperate in diverse environments. CoELA [46] integrates perception, memory, and execution in a modular framework and coordinates plans across agents in a natural language using LLM, and Co-NavGPT [42] employs vision-language model (VLM) as a global planner to enable multiple robots to explore complex environments. Meanwhile, several papers evaluate models under limited communication resources and partial observability, which is closer to real-world settings [19, 43]. It is also shown that structured and leadership prompting enhances teamwork efficiency and reduces unnecessary communication [14], and PARTNR [5] built a large-scale human–robot collaboration benchmark, demonstrating that even state-of-the-art LLM-based systems still face limitations in planning and coordination. However, while these studies have primarily focused on optimizing action execution and cooperation strategies, the problem of integrating egocentric experiences collected over long periods by multiple agents to perform QA remains insufficiently addressed.

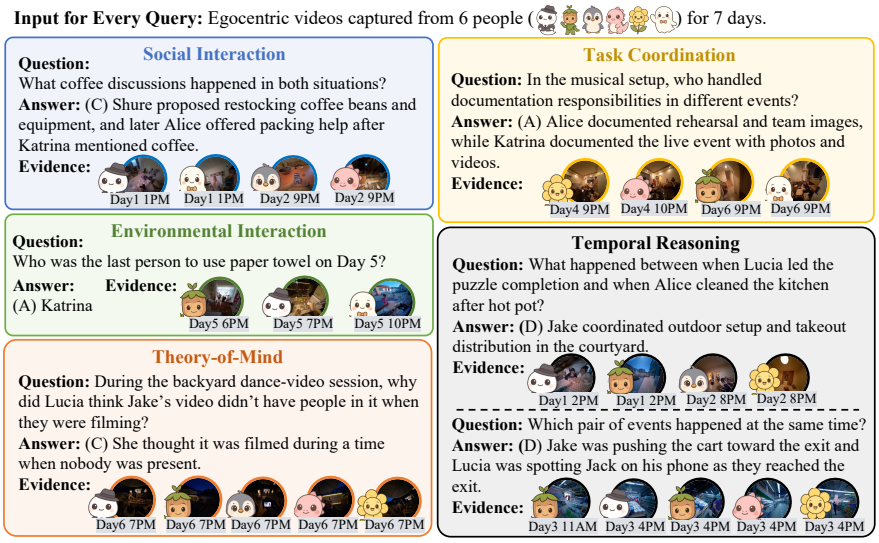


Fig. 2: Examples from MA-EgoQA across five categories. MA-EgoQA is the first multiple embodied agents egocentric video QA benchmark, requiring comprehension of six egocentric videos spanning seven days per query. False options are omitted.

2.2 Egocentric Video Understanding Benchmarks

Since egocentric video understanding is crucial for real-world applications, a number of benchmarks [18, 40] have been introduced to evaluate video QA in this setting. Ego4D [13] released 3,670 hours of egocentric video along with multiple tasks, including episodic memory and hand–object interaction, establishing a fundamental resource for egocentric video research. Building on this foundation, EgoSchema [27] evaluates minute-level video understanding with a focus on long-term context, while EgoThink [8] defines six core capabilities and twelve dimensions of egocentric reasoning to assess the interpretive and inferential capacity of VQA models. Extending beyond recognition, EgoPlan-Bench [7] introduced tasks for embodied planning, examining how models can connect visual observations with action planning in egocentric settings. However, in all of these benchmarks, the video duration per sample remains shorter than one hour, which is not reflective of embodied agents that operate continuously for days.

To address this limitation, EgoLife [39] constructed a super-long egocentric video dataset in which six individuals wore camera-equipped glasses to capture their daily experiences over seven consecutive days in a shared house. While this dataset breaks the length barrier of prior work, its QA benchmark, EgoLifeQA, is designed under a single-agent assumption, and questions can be answered by referencing only one individual's memory. Meanwhile, MA-EgoQA is the first to evaluate the QA task on multiple, super-long, and temporally aligned egocentric videos. Tab. 1 compares existing benchmarks with MA-EgoQA.

3 MA-EgoQA Benchmark

We now introduce MA-EgoQA, a new benchmark designed to evaluate the ability to comprehend and reason over multiple egocentric video streams from embodied agents. We first formally define the task in Sec. 3.1 and describe the categories in Sec. 3.2. Finally, we provide a detailed analysis of MA-EgoQA, presenting dataset statistics and qualitative examples in Sec. 3.3. We present QA examples for each category in Fig. 2, highlighting that the questions are unique to multi-agent settings and require references to multiple timestamps.

3.1 Task Definition

We formally define the **Multi-Agent Egocentric Video Question Answering** task as follows. Let there be N embodied agents, where each agent A_i continuously records an egocentric video stream V_i for T hours. The complete multi-agent video collection is denoted as $\mathcal{V} = \{V_1, V_2, \ldots, V_N\}$, resulting in a total of $N \times T$ hours of video. Given a user query, the system should generate a response based on information from $\mathcal{V}$. In this task, a query requires information from more than two agents, causing two main challenges. First, the system must build a coherent global understanding across agents. To achieve this, each agent's information should be stored in a way that is temporally and contextually aligned with others. Second, the model must effectively retrieve relevant events from the saved representations to answer each query. Since queries may require referencing different timestamps across multiple agents, retrieval becomes significantly more complex and challenging than in a single-agent setting.

3.2 Benchmark Categories

We design five categories in the **MA-EgoQA benchmark** prior to generating the question-answer pairs, ensuring that the data generation pipeline is optimized for each. In selecting these categories, our primary goal is to capture aspects that are unique to the multi-agent setting and essential for human-agent system interactions in the real world. The categories are described as follows.

Social Interaction (SI) This category evaluates the ability to accurately localize and ground casual conversations or affiliative behaviors across video streams. This category contains questions related to how people engage or respond to others, group behaviors without specific goal sequences involving multiple people interacting, and meaningful information during the conversation.

Task Coordination (TC) As highlighted in Sec. 2.1, research on MAS has largely focused on collaboration to achieve shared goals. To reflect this practical importance, we define this category. Questions in this category address how roles were assigned, responsibilities divided, and actions sequenced toward goal completion, and how decisions were made throughout the execution of tasks.

Theory of Mind (ToM) ToM refers to the cognitive ability to reason about the mental states of others, including their thoughts, beliefs, desires, and emotions. Effective use of diverse perspectives is crucial for capturing contextual cues

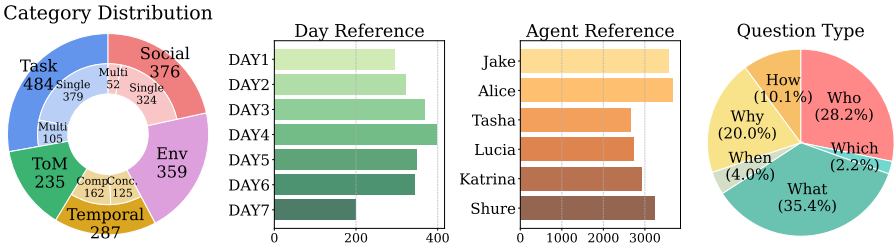


Fig. 3: Statistics for MA-EgoQA; *(Left)* number of samples per category; *(Center)* day/agent reference counts of all categories; *(Right)* question type distribution.

and improving performance on ToM questions. This category includes queries about what an agent believed or misunderstood, what information they could or could not perceive, and the intentions underlying their actions.

Temporal Reasoning (TR) Accurate understanding of multi-agent experiences requires aligning timelines from different egocentric video streams into a coherent global view. To evaluate this ability, the TR category is divided into two subcategories: *concurrency*, which focuses on what one agent was doing while others performed different activities at the same time, and *comparison*, which examines the relative temporal ordering of events across agents.

Environmental Interaction (EI) Agents also engage with their surrounding environment, such as operating a vacuum cleaner or turning on a faucet. Since these interactions are distributed across multiple agents, aggregating them is essential for tracking environmental states and planning appropriate actions to achieve shared goals. Questions in this category explore object usage, including frequency, first-time use, and which agent engaged with them most.

For SI and TC categories, we additionally design multi-span questions that require reasoning across multiple non-contiguous temporal windows, thereby extending MA-EgoQA into a genuinely long-horizon benchmark. Benchmark examples for each category are shown in Fig. 2 and Suppl. D.

3.3 Benchmark Statistics

MA-EgoQA contains 1,741 questions across five categories, and each question must be answered based on multiple egocentric videos captured by six agents over seven days. We analyze the distribution of questions in MA-EgoQA across question types, categories, days, and agents, and present the statistics in Fig. 3. The results show that the questions are well distributed across the five categories, the seven recorded days, and the six agents. This balance ensures that the evaluation is not dominated by a single skill or event but provides a diverse set of challenges and requires the model to track and reason about all agents effectively. The benchmark includes questions with diverse linguistic formulations. The majority of questions begin with `What` and `Who`, while questions beginning with `Why`, `How`, `When`, and `Which` broaden the range of question types. Overall, the broad coverage of questions in MA-EgoQA ensures a comprehensive evaluation of multimodal reasoning in egocentric multi-agent scenarios.

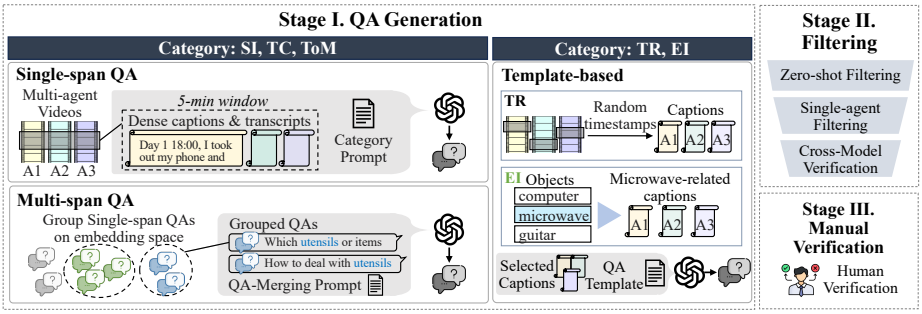


Fig. 4: Benchmark construction pipeline. QA pairs are generated for each category through its dedicated process, and refined through LLM filtering and manual check.

4 Benchmark Construction

4.1 Data Generation

We construct candidate question–answer pairs using GPT-based generation. Since the nature of questions differs across categories, we adopt two complementary strategies. For SI, TC, and ToM categories, where question formats are diverse and open-ended, we generate a large pool of samples and subsequently filter low-quality pairs using the LLM-based pipeline described in Sec. 4.2. In contrast, for TR and EI, which typically involve more structured queries, we design predefined templates and generate QA pairs by instantiating them with contextual information, followed by filtering and verification. For convenient evaluation, MA-EgoQA is designed to be multiple-choice questions with five options. We describe the candidate generation process in detail below. The prompts and templates used in each generation stage are shown in Suppl. E and F.

Single-span Question Generation Each long egocentric video from an agent is divided into fixed-length segments of 5 minutes. For each 5-minute window, we collect dense captions and transcripts from all available agents during that period and provide them to GPT-4o, together with a system prompt and a category-specific prompt for SI, TC, and ToM categories. The prompts instruct the model to understand the events within the 5-minute context and to generate a question, an answer, a rationale, the names and timestamps of the referenced agents, and four wrong options. Each generated question must be grounded by more than two agents and relevant to its assigned category. In total, we generate 33.4k, 31.6k, and 34.1k QA samples for the SI, TC, and ToM categories, respectively.

Multi-span Question Generation Multi-span questions are constructed by grouping semantically similar single-span questions, ensuring that each multi-span question addresses the same topic across different timestamps. This design ensures that the resulting questions are coherent and contextually plausible. To be specific, we concatenate each single-span question q and its paired answer a, and obtain a sentence embedding e using a text encoder $E(\cdot)$, which is represented as $e = E([q; a])$. We then define pairwise cosine similarity between two samples as $s_{ij} = \langle e_i, e_j \rangle / \|e_i\| \|e_j\|$. A similarity graph is constructed where each

vertex corresponds to a QA pair and there is an edge (i, j) if $s_{ij} \geq \delta$, with δ a predefined threshold. Connected components are extracted from this graph as groups, where each group contains at least two semantically related QA pairs. For each group, all single span questions and their correct answers in the group are given to GPT-5 with the system prompt and category-wise prompt, and the model synthesizes multi-span question and answer only, not false options at this stage. To ensure difficulty and quality, false options are created separately, rather than being inherited from the single-span samples. We generated 15.9k samples for SI multi-span questions and 16.3k samples for TC multi-span questions.

Template-based Question Generation To improve the efficiency of QA generation, we employ predefined template-based methods for the TR and EI categories, as questions in these categories share common objectives within each category, such as ordering event timestamps or counting object usage. To generate TR questions, we first create captions for every agent's videos using temporal windows of 30 seconds, 10 minutes, and 1 hour to encourage coverage across multiple temporal scales. We then define two subcategories within TR: the *comparison* type, which focuses on ordering events across multiple agents, and the *concurrency* type, which requires temporal alignment among agents. After constructing all captions, we generate each QA instance by selecting several captions from multiple agents at the same temporal scale and providing them to GPT-5 along with a prompt. For the *comparison* type, we randomly sample different timestamps for each agent. For the *concurrency* type, we randomly select a single timestamp and gather captions from all agents at that time. GPT-5 then generates a question with five answer options, including one correct answer.

For the EI category, we first manually list the objects with which agents interact in the shared environment. For each object and for each six hour or one day interval, we collect all dense captions from all agents that include the object. The collected captions are then paired with predefined QA templates and provided to GPT-5, which analyzes interactions between the agents and the object within the interval and generates a question with five answer options.

4.2 LLM Filtering and Human Verification

We employ an automatic LLM-based filtering prior to the human verification stage. Each generated question–answer pair is verified by prompting an LLM with a restricted input context and evaluating whether the model can correctly infer the correct answer. If the model predicts the answer under these constraints, the sample is deemed insufficiently challenging and excluded from the benchmark. All filtering steps are conducted using GPT-5 in the following order, which serves as the backbone for quality control in our data generation pipeline.

Zero-shot Filtering In this stage, each question is evaluated without providing any context, thereby testing whether the correct answer can be trivially inferred. If the model consistently predicts the correct option under such a zero-context setting, the sample is considered unchallenging and excluded. Concretely, we

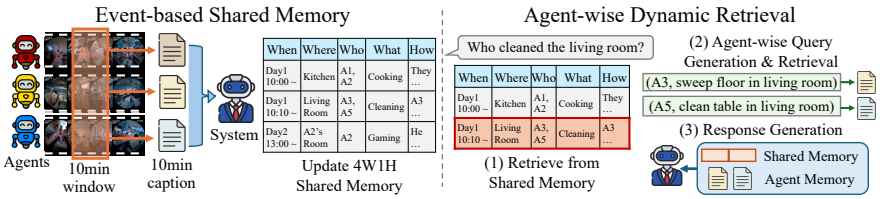


Fig. 5: Overview of EgoMAS. EgoMAS builds an event-based shared memory from multi-agent videos and dynamically retrieve relevant information from it.

query the model three times per question, and if it produces the correct answer in more than two trials, the corresponding sample is discarded from the dataset.

Single Agent Filtering Since MA-EgoQA is designed to evaluate reasoning in a multi-agent context, we remove any sample that can be answered using the memory of a single individual. To enforce this criterion, we extract all human names appearing in the question and the correct answer, and perform inference using each corresponding individual's memory. If no human name is present, we randomly select one from the available agents and conduct inference with their memory. If the model produces the correct answer under any single-agent memory condition, the sample is eliminated from the benchmark.

Cross-model Validation All preceding filtering stages rely on GPT-5, which may introduce model-specific biases. To address this, we include an additional verification using two external models, Gemini-2.5-Flash and Claude-Sonnet-4. Each sample is re-examined under the same contextual information available at generation time. This step not only checks the correctness of the answer, but also validates the question and the false options, ensuring that the query is non-trivial. If either external model flags a sample as invalid, it is removed.

Human Verification Following all automatic filtering steps, a final human verification stage is conducted to ensure the quality and validity of the remaining candidates. Human verifiers are granted access to the complete set of information, including captions, transcripts, and videos from all agents. Each verifier evaluates the samples according to the same guidelines used during data generation, ensuring consistency with the original design criteria. In total, four human verifiers reviewed 3,436 candidates, and 1,741 high-quality samples were selected.

5 Egocentric Video Reasoning in Multi-Agent System

Beyond introducing the new benchmark, we also propose a simple yet effective training-free baseline, named EgoMAS (Egocentric Multi-Agent System), to stimulate future research on MA-EgoQA. EgoMAS is a centralized MAS that is specifically designed to tackle the unique challenges of multi-agent egocentric reasoning. First, it leverages an event-based shared memory that enables a system-level global understanding by integrating fragmented events across

agents. Second, EgoMAS dynamically selects which agent's memory to reference and adapts the search query for each agent, allowing fine-grained reasoning across multiple perspectives. Prompts for each step are provided in Suppl. G.

5.1 Event-based Shared Memory

At every 10-minute interval, each embodied agent provides a caption summarizing its own observations during the preceding time span. A centralized manager then integrates these individual captions into a system-level summary. Rather than producing a flat textual condensation, the manager first identifies key events across agents and, for each event, explicitly records the corresponding 4W1H fields: `When`, `What`, `Where`, `Who`, and `How`, producing a coherent global memory that aligns agent perspectives while preserving critical details for reasoning.

5.2 Agent-wise Dynamic Retrieval

Given a query q, EgoMAS first retrieves top-n system-level memories from the shared memory $\mathcal{M}_{\text{shared}}$ using BM25 ranking

$$\mathcal{R}_{\text{sys}}(q) = \text{Top-}n\ \{(m, s(m,q)) \mid m \in \mathcal{M}_{\text{shared}}\},$$

where $s(m,q)$ denotes the BM25 score between memory m and query q.
From the retrieved system-level context, EgoMAS generates a set of agent-specific retrieval requests $\mathcal{Q}_{\text{agent}} = \{(a_j, q_j)\}_{j=1}^{J}$, where each request consists of an agent identifier a_j and a sub-query q_j. For each (a_j, q_j), EgoMAS performs agent-level retrieval from the agent's memory $\mathcal{M}_{a_j}$:

$$\mathcal{R}_{a_j}(q_j) = \text{Top-}k\ \{(m, s(m,q_j)) \mid m \in \mathcal{M}_{a_j}\}.$$

To ensure relevance, we filter out memories with scores below a threshold τ:

$$\widetilde{\mathcal{R}}_{a_j}(q_j) = \{(m, s(m,q_j)) \in \mathcal{R}_{a_j}(q_j) \mid s(m,q_j) \geq \tau\}.$$

Finally, the system generates the final response by conditioning on both the retrieved system-level context $\mathcal{R}_{\text{sys}}(q)$ and the aggregated agent-level results $\widetilde{\mathcal{R}} = \bigcup_{j=1}^{J} \widetilde{\mathcal{R}}_{a_j}(q_j)$:

$$\hat{y} = F\big(q, \mathcal{R}_{\text{sys}}(q), \widetilde{\mathcal{R}}\big),$$

where $\hat{y}$ and F denote the response and response generation function.

6 Experimental Setup and Results

6.1 Experimental Setup

We evaluate a range of competitive open-source and proprietary LLMs and video LLMs on MA-EgoQA. Proprietary LLMs include Gemini-2.5-Flash [9] and GPT-5 [32]. Open-source LLMs include Llama-3.1 [12], Qwen2.5-1M [38], Qwen3 [37],

Table 2: Evaluation results of 16 baselines and EgoMAS on MA-EgoQA.

Models	Input context	SI	TC	ToM	TR	EI	Avg.
Random	-	20.00	20.00	20.00	20.00	20.00	20.00
All Caption Concat Baselines							
Gemini-2.5-Flash [9]	1M	41.22	36.36	24.26	<u>46.59</u>	33.98	36.93
Llama-3.1-Nemotron-8B [4]	1M	20.48	22.93	20.43	22.65	21.17	21.65
Qwen2.5-7B-Instruct-1M [38]	500k	27.13	28.93	20.85	24.04	26.18	26.08
GPT-5 [32]	272k	36.17	33.88	22.55	39.71	38.72	34.81
Qwen3-30b-a3b-instruct [12]	240k	29.52	27.69	16.60	23.00	26.46	25.56
gpt-oss-120b [1]	128k	28.72	27.48	22.55	28.22	25.63	26.82
gpt-oss-20b [1]	128k	19.15	28.10	23.83	26.83	25.35	24.81
All Frame Concat Baselines							
VideoChat-Flash [25]	32k	23.67	25.00	20.00	24.39	20.73	23.06
VideoXL-2 [29]	32k	20.21	19.83	19.15	20.91	21.73	20.39
Qwen2.5-VL-7B [3]	128k	26.33	26.45	21.28	22.22	25.07	25.22
RAG Baselines							
VideoRAG [20]	8.0k	27.13	26.03	20.85	25.09	25.63	25.33
EgoRAG [39]	-	22.34	21.07	18.30	15.33	18.87	19.57
Ego-R1-Agent-3B [33]	-	22.34	25.62	18.30	26.83	28.41	24.70
WorldMM-8B [41]	4.1k	29.26	34.50	21.70	25.09	22.56	27.63
BM25 [30]	8.1k	**44.68**	37.60	<u>30.21</u>	33.45	30.64	36.01
DPR [22]	7.8k	31.12	28.31	18.72	26.48	22.84	26.19
Ours							
EgoMAS (Gemini-2.5-Flash)	4.6k	41.49	**41.32**	**33.62**	39.37	**48.19**	**41.41**
EgoMAS (Qwen3VL-8B-Thinking)	5.4k	38.03	<u>39.88</u>	28.94	**47.39**	<u>44.85</u>	<u>40.26</u>
EgoMAS (Qwen3VL-8B-Instruct)	7.4k	<u>43.09</u>	39.05	24.26	35.54	40.67	37.68
EgoMAS (Qwen2.5VL-7B-Instruct)	5.4k	37.77	38.64	25.96	34.15	36.49	35.55
Oracle (Gemini-2.5-Flash)	-	86.44	97.60	70.21	88.15	81.34	83.80
Oracle (Qwen3VL-8B-Instruct)	-	76.86	80.99	57.02	77.35	69.92	73.98

and gpt-oss [1]. Video LLMs include VideoChat-Flash [25], Video-XL-2 [29], and Qwen2.5-VL-7B [3]. For text-only LLMs, we concatenate captions from all agents in chronological order and directly pass them to the LLM along with the question. For video LLMs, we concatenate all video frames from all agents in chronological order and uniformly sample the number of frames that each model can receive without exceeding the model's input limits.

We also evaluate several RAG-based baselines, including simple text-based retrieval methods BM25 [30] and DPR [22] with Qwen3-VL-8B-Instruct [2] and VideoRAG [20] with InternVideo2 [35] encoder. There are three baselines that work on the EgoLife dataset, but in a single-agent setting: EgoRAG [39], Ego-R1-Agent-3B [33], and WorldMM-8B [41]. Since they do not support multi-agent scenarios, we only use videos from a single person in EgoRAG, and use an LLM router to select which individual's memory to refer to in other models. RAG-based baselines leverage both frames and captions as memories.

For EgoMAS, we evaluate the system with four different backbone models: Gemini-2.5-Flash [9], Qwen3VL-8B-Thinking [2], Qwen3VL-8B-Instruct [2], and Qwen2.5VL-7B-Instruct [3]. The hyperparameters are set to $n = 20$, $k = 5$, $\tau = 10$. Lastly, we also evaluate an Oracle setting, which generates answers using the same context used to generate the questions. More details including frame sampling and memory retrieval in baselines are described in Suppl. B.

Table 3: Performance comparison on sub-categories in SI, TC, and TR.

Models	SI		TC		TR	
	Single	Multi	Single	Multi	Concur.	Compare
Qwen2.5-7B-Instruct-1M	27.78	23.08	28.23	31.43	28.00	20.99
GPT-5	37.65	26.92	35.62	27.62	42.40	37.65
gpt-oss-120b	30.86	15.38	29.55	20.00	25.60	30.25
VideoChat-Flash	24.07	21.15	26.65	19.05	27.20	22.22
VideoRAG	29.94	9.62	27.97	19.05	28.00	22.84
BM25	41.05	23.08	39.84	18.10	38.40	22.84
EgoMAS (Qwen3-VL-8B-Inst)	46.30	23.08	43.27	23.81	36.00	35.19

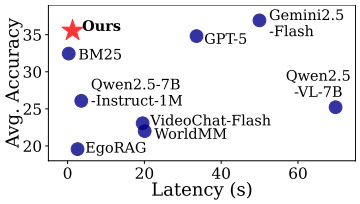


Fig. 6: Comparison of inference latency and accuracy.

6.2 Evaluation Results

We present the MA-EgoQA evaluation results in Tab. 2. Overall, the results show that MA-EgoQA is a highly challenging benchmark for current models. Even the strongest model, Gemini-2.5-Flash, achieves an average accuracy of only 36.93%, and many baselines, including gpt-oss models and VideoXL-2, perform only marginally above random chance, highlighting the difficulty of the task. Specifically, in `All Caption Concat Baselines`, only proprietary models achieve accuracies above 30%, while smaller open-source models show below 27% despite handling extremely large input context (128k to 1M tokens). These results indicate that providing all captions without retrieval tends to distract the model with irrelevant information while requiring substantial computation costs. Meanwhile, `All Frame Concat Baselines` achieve the lowest average performance among all baseline categories, likely due to the absence of transcript information and the presence of many irrelevant frames.

In contrast, `RAG Baselines`, including BM25 and WorldMM-8B, outperform other non-retrieval open-source models while using significantly shorter input contexts. This demonstrates that retrieval is crucial for efficiently understanding multiple long egocentric videos. Finally, EgoMAS significantly outperforms all baselines. EgoMAS (Gemini-2.5-Flash) achieves 41.41% accuracy, which is 4.48% higher than the Gemini-2.5-Flash baseline. Notably, even Qwen3VL-8B-based EgoMAS models surpass Gemini-2.5-Flash and GPT-5, highlighting the effectiveness of the EgoMAS framework. We also observe that EgoMAS performance scales with the capability of the backbone model, improving from Qwen2.5VL-7B-Instruct to Gemini-2.5-Flash backbone, suggesting that stronger reasoning ability enables more effective retrieval strategies and better integration of information retrieved from multiple agents. However, all models exhibit a substantial gap compared to the oracle setting, indicating significant room for improvement.

7 Analysis

7.1 Analysis on the MA-EgoQA Benchmark

Does MA-EgoQA require multi-agent memory? To verify that MA-EgoQA requires access to multiple agents' memories, we evaluate EgoMAS using

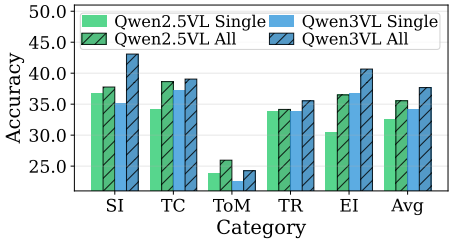


Fig. 7: Evaluation with single-agent memory and all agents memory.

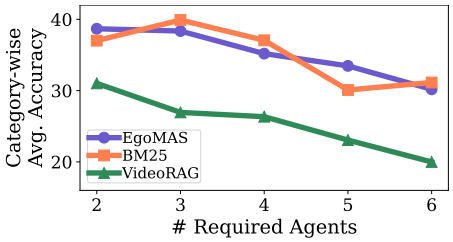


Fig. 8: MA-EgoQA performance with varying numbers of required agents.

only a single agent's memory and compare the results with the original performance. As shown in Fig. 7, restricting EgoMAS to a single agent leads to a substantial performance drop, demonstrating that models should utilize memories from multiple agents to correctly answer MA-EgoQA queries.

Multi-span and multi-agent reasoning as key challenges We analyze which aspects of MA-EgoQA make the task challenging from two perspectives: long-horizon reasoning and multi-agent knowledge fusion. First, we compare model performance on queries that require context from a single timestamp versus multiple timestamps across SI, TC, and TR categories. As shown in Tab. 3, models achieve substantially lower accuracy on queries grounded in multiple spans (Multi-span in SI and TC, and Comparison in TR). This suggests that identifying and relating multiple events across seven days of video requires stronger retrieval capabilities. Second, we analyze performance with respect to the number of agents required to answer each query in Fig. 8. Performance consistently decreases as the number of required agents increases, indicating that current approaches to multi-agent knowledge fusion remain limited. This highlights the need for future research on constructing a coherent global understanding beyond simple memory sharing among agents. As a result, both aspects significantly contribute to increasing the task difficulty.

ToM is the hardest category in MA-EgoQA As shown in Tab. 2, ToM consistently exhibits the lowest accuracies across most models. Although EgoMAS substantially improves ToM performance, it still lags behind other categories. This is because ToM requires inferring latent mental states, such as goals and beliefs, rather than extracting explicitly observable visual or textual cues. This observation corroborates prior work identifying ToM as exceptionally challenging due to its inherent ambiguity, the need for recursive mental-state reasoning [31], and complex goal–belief–action inference [26].

7.2 Analysis on EgoMAS

Efficiency of EgoMAS Fig. 6 shows the average latency measured over 100 randomly selected MA-EgoQA samples and overall accuracy of baseline models and EgoMAS (Qwen2.5VL-7B-Instruct). Non-retrieval-based models show substantially higher computational overhead, whereas retrieval-based models, including EgoMAS, exhibit notably lower latency. In particular, EgoMAS achieves

Table 4: Ablation study on EgoMAS structure.

Shared Memory	Dynamic Retrieval	Acc.
✗	✗	27.80
✗	✓	28.20
✓	✗	30.04
✓	✓	**35.55**

Table 5: Ablation studies on sub-modules of EgoMAS.

(a) Shared Memory Structure

Shared Memory Structure	Acc.
Summary	30.67
Triplet	30.44
Chunk	25.96
Graph	31.99
Ours (4W1H)	**35.55**

(b) Memory Retriever

Memory Retriever	Acc.
DPR	28.67
Qwen3-Embed-0.6B	33.03
NV-Embed-v2 7B	**37.91**
Ours (BM25)	35.55

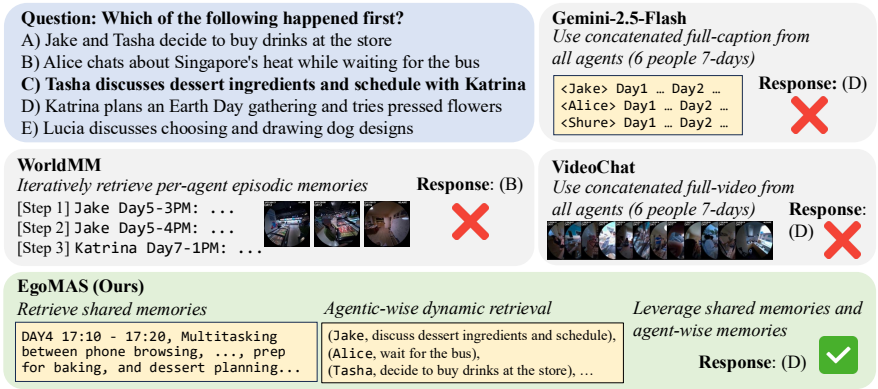


Fig. 9: Case Study: Comparative analysis across models on MA-EgoQA

the highest accuracy among retrieval-based models while requiring only 1.3 seconds per query. These results highlight that EgoMAS is a practical approach for building a multi-agent egocentric video QA system in real-world settings.

Efficacy of Two Methods in EgoMAS To validate the efficacy of each method in EgoMAS, we perform an ablation study using EgoMAS (Qwen2.5VL-7B-Instruct), as shown in Tab. 4. For the model without shared memory, we use 10-minute captions for all agents as a retrieval source. Meanwhile, the models without agent-wise dynamic retrieval leverage generate the response only with shared memories. Using both methods together achieves the highest accuracy, highlighting that aggregating memories from all agents into shared memory is effective, and a dynamic retrieval method helps the model to find the most informative and related knowledge from the corpus. Ablation studies on the hyperparameter for each module is presented in Suppl. A.1.

Analysis on Sub-modules of EgoMAS We validate the contribution of each sub-module with ablation studies in Tab. 5. In Tab. 5a, we evaluate multiple memory construction strategies introduced in prior works [15,45] and 4W1H memory. The event-based 4W1H memory structure clearly outperforms alternative methods, demonstrating its strong ability to abstract and fuse the events across agents. Tab. 5b presents a comparison between our choice of memory retriever and other dense retrievers [24,47]. Although NV-Embed-v2 achieves the highest accuracy, it contains 7B parameters and consequently incurs substantial computational overhead. In contrast, BM25 employs lightweight keyword-based retrieval and delivers competitive performance, highlighting its practicality.

Case Study We further analyze the baselines and EgoMAS through the qualitative example in Fig. 9. As discussed earlier, Gemini-2.5-Flash and VideoChat-Flash receive excessive information, which prevents them from focusing on the relevant events and leads to incorrect predictions. WorldMM attempts to retrieve memories, but since memories from agents are not aggregated, it retrieves each agent's memory iteratively and ultimately fails to answer the query. In contrast, EgoMAS successfully locates target events using shared memory, retrieves additional details through agent-wise dynamic retrieval, and responses correctly.

8 Conclusion

In this work, we present MA-EgoQA, a novel benchmark for question answering over multiple long-horizon egocentric video streams from embodied agents, spanning five categories that capture the core challenges of multi-agent reasoning. Alongside, we propose EgoMAS, a simple training-free baseline that combines shared memory with agent-wise dynamic retrieval, achieving superior performance to recent frontier models with significantly long context despite using a much smaller model. Our evaluations show that current LLMs and video LLMs struggle with a multi-agent egocentric setting, particularly in the theory of mind and multi-agent reasoning, underscoring the complexity of the scenario and highlighting MA-EgoQA as a promising future direction for multiple agents.

References

1. Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R.K., Bai, Y., Baker, B., Bao, H., et al.: gpt-oss-120b & gpt-oss-20b model card. arXiv preprint arXiv:2508.10925 (2025) 11
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 3, 11
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 11
4. Bercovich, A., Levy, I., Golan, I., Dabbah, M., El-Yaniv, R., Puny, O., Galil, I., Moshe, Z., Ronen, T., Nabwani, N., et al.: Llama-nemotron: Efficient reasoning models. arXiv preprint arXiv:2505.00949 (2025) 11
5. Chang, M., Chhablani, G., Clegg, A., Cote, M.D., Desai, R., Hlavac, M., Karashchuk, V., Krantz, J., Mottaghi, R., Parashar, P., et al.: Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. arXiv preprint arXiv:2411.00081 (2024) 3
6. Chen, J., Yu, C., Zhou, X., Xu, T., Mu, Y., Hu, M., Shao, W., Wang, Y., Li, G., Shao, L.: Emos: Embodiment-aware heterogeneous multi-robot operating system with llm agents. arXiv preprint arXiv:2410.22662 (2024) 2
7. Chen, Y., Ge, Y., Ge, Y., Ding, M., Li, B., Wang, R., Xu, R., Shan, Y., Liu, X.: Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. CoRR (2023) 4

8. Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H., Liu, Y.: Egothink: Evaluating first-person perspective thinking capability of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14291–14302 (2024) 2, 4
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) 3, 10, 11
10. Feng, Z., Xue, R., Yuan, L., Yu, Y., Ding, N., Liu, M., Gao, B., Sun, J., Zheng, X., Wang, G.: Multi-agent embodied ai: Advances and future directions. arXiv preprint arXiv:2505.05108 (2025) 2
11. Fu, D., Li, X., Wen, L., Dou, M., Cai, P., Shi, B., Qiao, Y.: Drive like a human: Rethinking autonomous driving with large language models. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). pp. 910–919. IEEE (2024) 1
12. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 10, 11
13. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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. pp. 18995–19012 (2022) 4
14. Guo, X., Huang, K., Liu, J., Fan, W., Vélez, N., Wu, Q., Wang, H., Griffiths, T.L., Wang, M.: Embodied llm agents learn to cooperate in organized teams. arXiv preprint arXiv:2403.12482 (2024) 2, 3
15. Gutiérrez, B.J., Shu, Y., Qi, W., Zhou, S., Su, Y.: From RAG to memory: Non-parametric continual learning for large language models. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=LWH8yn4HS2 14
16. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems **35**, 4874–4886 (2022) 2
17. Huang, Y., Sansom, J., Ma, Z., Gervits, F., Chai, J.: Drivlme: Enhancing llm-based autonomous driving agents with embodied and social experiences. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3153–3160. IEEE (2024) 1
18. Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., Zhang, H., Dong, L., Wang, Y., Wang, L., 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. pp. 22072–22086 (2024) 2, 4
19. Jain, U., Weihs, L., Kolve, E., Farhadi, A., Lazebnik, S., Kembhavi, A., Schwing, A.: A cordial sync: Going beyond marginal policies for multi-agent embodied tasks. In: European Conference on Computer Vision. pp. 471–490. Springer (2020) 3
20. Jeong, S., Kim, K., Baek, J., Hwang, S.J.: Videorag: Retrieval-augmented generation over video corpus. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 21278–21298 (2025) 11
21. Kannan, S.S., Venkatesh, V.L., Min, B.C.: Smart-llm: Smart multi-agent robot task planning using large language models. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 12140–12147. IEEE (2024) 1

22. Karpukhin, V., Oguz, B., Min, S., Lewis, P.S., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: EMNLP (1). pp. 6769–6781 (2020) 11
23. Kim, K., Park, G., Lee, Y., Yeo, W., Hwang, S.J.: Videoicl: Confidence-based iterative in-context learning for out-of-distribution video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3295–3305 (2025) 2
24. Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., Ping, W.: Nv-embed: Improved techniques for training llms as generalist embedding models. arXiv preprint arXiv:2405.17428 (2024) 14
25. Li, X., Wang, Y., Yu, J., Zeng, X., Zhu, Y., Huang, H., Gao, J., Li, K., He, Y., Wang, C., et al.: Videochat-flash: Hierarchical compression for long-context video modeling. arXiv preprint arXiv:2501.00574 (2024) 11
26. Li, Y., Veerabadran, V., Iuzzolino, M.L., Roads, B.D., Celikyilmaz, A., Ridgeway, K.: Egotom: Benchmarking theory of mind reasoning from egocentric videos. arXiv preprint arXiv:2503.22152 (2025) 2, 13
27. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023) 2, 4
28. Pei, B., Chen, G., Xu, J., He, Y., Liu, Y., Pan, K., Huang, Y., Wang, Y., Lu, T., Wang, L., et al.: Egovideo: Exploring egocentric foundation model and downstream adaptation. arXiv preprint arXiv:2406.18070 (2024) 2
29. Qin, M., Liu, X., Liang, Z., Shu, Y., Yuan, H., Zhou, J., Xiao, S., Zhao, B., Liu, Z.: Video-xl-2: Towards very long-video understanding through task-aware kv sparsification. arXiv preprint arXiv:2506.19225 (2025) 11
30. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2009) 11
31. Shi, H., Ye, S., Fang, X., Jin, C., Isik, L., Kuo, Y.L., Shu, T.: Muma-tom: Multi-modal multi-agent theory of mind. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1510–1519 (2025) 2, 13
32. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025) 10, 11
33. Tian, S., Wang, R., Guo, H., Wu, P., Dong, Y., Wang, X., Yang, J., Zhang, H., Zhu, H., Liu, Z.: Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. arXiv preprint arXiv:2506.13654 (2025) 2, 11
34. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y., et al.: Lvbench: An extreme long video understanding benchmark. arXiv preprint arXiv:2406.08035 (2024) 2
35. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: European Conference on Computer Vision. pp. 396–416. Springer (2024) 11
36. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024) 2
37. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 10

38. Yang, A., Yu, B., Li, C., Liu, D., Huang, F., Huang, H., Jiang, J., Tu, J., Zhang, J., Zhou, J., et al.: Qwen2. 5-1m technical report. arXiv preprint arXiv:2501.15383 (2025) 10, 11
39. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al.: Egolife: Towards egocentric life assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28885–28900 (2025) 2, 3, 4, 11
40. Ye, H., Zhang, H., Daxberger, E., Chen, L., Lin, Z., Li, Y., Zhang, B., You, H., Xu, D., Gan, Z., et al.: Mm-ego: Towards building egocentric multimodal llms for video qa. arXiv preprint arXiv:2410.07177 (2024) 2, 4
41. Yeo, W., Kim, K., Yoon, J., Hwang, S.J.: Worldmm: Dynamic multimodal memory agent for long video reasoning. arXiv preprint arXiv:2512.02425 (2025) 2, 11
42. Yu, B., Kasaei, H., Cao, M.: Co-navgpt: Multi-robot cooperative visual semantic navigation using large language models. arXiv preprint arXiv:2310.07937 (2023) 3
43. Yu, C., Yang, X., Gao, J., Chen, J., Li, Y., Liu, J., Xiang, Y., Huang, R., Yang, H., Wu, Y., et al.: Asynchronous multi-agent reinforcement learning for efficient real-time multi-robot cooperative exploration. arXiv preprint arXiv:2301.03398 (2023) 3
44. Yuan, P., Ma, A., Yao, Y., Yao, H., Tomizuka, M., Ding, M.: Remac: Self-reflective and self-evolving multi-agent collaboration for long-horizon robot manipulation. arXiv preprint arXiv:2503.22122 (2025) 2
45. Zeng, R., Fang, J., Liu, S., Meng, Z.: On the structural memory of llm agents. arXiv preprint arXiv:2412.15266 (2024) 14
46. Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J.B., Shu, T., Gan, C.: Building cooperative embodied agents modularly with large language models. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=EnXJfQqy0K 2, 3
47. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025) 14
48. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1702–1713 (2025) 1
49. Zhu, N., Dong, Y., Wang, T., Li, X., Deng, S., Wang, Y., Hong, Z., Geng, T., Niu, G., Huang, H., et al.: Cvbench: Evaluating cross-video synergies for complex multimodal understanding and reasoning. arXiv preprint arXiv:2508.19542 (2025) 2

Supplementary Materials

This supplementary information includes additional analyses (Section A), more details on the experimental setups (Section B), and a discussion of limitations and future work (Section C). We further provide benchmark details including category examples (Section D), prompts used for benchmark generation (Section E), and question templates for the TR and EI categories (Section F). Lastly, we present additional analysis on the performance with different input context modalities on MA-EgoQA (Section H).

A Additional Analysis on EgoMAS

A.1 Ablation on Hyperparameters

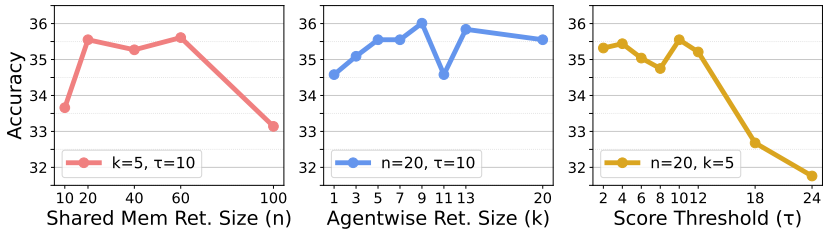


Fig. S1: Ablation studies on shared memory retrieval size (n), agent-wise dynamic retrieval size (k), and score threshold (τ).

There are three hyperparameters in EgoMAS: shared memory retrieval size (n), agent-wise dynamic retrieval size (k), and score threshold (τ). In Fig. S1, we report the performance of EgoMAS with different values of the hyperparameter. Shared memory retrieval size is optimal around 20 to 60, since if it is too small or big, then models suffer from finding meaningful information from the given context. However, the model shows robust performance over k, while high threshold τ decreases the accuracy by filtering most of the retrieved data.

A.2 Sensitivity Analysis to the Number of Available Agents

We investigate the sensitivity of MA-EgoQA to the number of agents by limiting the available agent views based on EgoMAS performance. As shown in Fig. S2, the results show a clear trend that using more agents leads to better accuracy from 31.99% to 35.55%. These results highlight that MA-EgoQA indeed requires models to incorporate information from multiple agents.

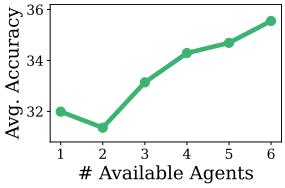


Fig. S2: Accuracy of EgoMAS with limited agents.

B Additional Details on Experimental Setups

For `All Caption Concat Baselines` in Tab. 2, we concatenate all captions from all agents and use them as context during response generation. For Gemini-2.5-Flash, Llama-3.1-Nemotron-8B, and Qwen2.5-7B-Instruct-1M, we concatenate captions from a 10-minute window for each agent, while for the other models we concatenate captions from a 1-hour window, according to the maximum context length of each model. If the resulting context exceeds the maximum context length, we truncate each caption until the length falls within the limit.

For `All Frame Concat Baselines`, we concatenate all frames captured by the agents over 7 days and uniformly sample frames until each model can process them without errors. As a result, 10k frames are sampled for VideoChat-Flash, 4k frames for VideoXL-2, and 1.9k frames for Qwen2.5VL-7B. For the VideoRAG model, we employ the InternVideo2 multimodal encoder to extract embeddings for each 30-second clip and its corresponding caption. We then retrieve five clips based on the cosine similarity between the query and each clip. From each retrieved clip, we uniformly sample 4 frames and provide them along with the caption to the response generation model. For BM25 and DPR, we select five 30-second clips using each retriever, and four frames uniformly sampled from each clip, together with their captions, are provided to the responder. For all `RAG Baselines` except Ego-R1, we use Qwen3-VL-8B-Inst as the response generator.

C Limitation and Future Work

This work has limitations as follows. First, while EgoLife provides 266 hours of egocentric videos from six people that cover a wide range of daily activities and environments, MA-EgoQA is based only on EgoLife and does not include other scenarios in different environments. This is because EgoLife is currently the only publicly available video dataset providing long-term, egocentric videos captured simultaneously from multiple agents. We hope more video datasets proposed in this direction and evaluate the model in various types of videos to ensure the generalization. Regarding EgoMAS, its performance still falls significantly behind the oracle performance (with a gap of around 42.4%), indicating that further research is required to effectively address the multi-agent egocentric video QA task. One promising direction is to employ more effective retrieval methods. As shown in Tab. 5b, stronger retrievers lead to better performance. Future work may explore more advanced retrieval strategies, for example hybrid retrieval methods that combine text embeddings with lexical matching.

D Benchmark QA Samples

We present example questions and options for each category following tables with the correct answers shown in bold.

Table S1: Social Interaction (SI) - Single Span category samples

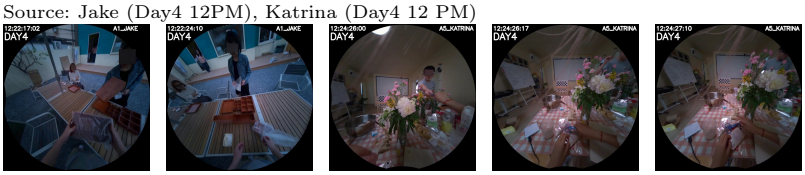
Category	**Social Interaction, Single Span**
Question	**Who helped each other find scissors, and what were the scissors used for?** (A) Jake lent scissors to Shure, who used them for prepping coffee beans. **(B) Jake handed scissors to Katrina for her package and flower crafts.** (C) Tasha and Lucia found scissors to cut plastic wrappings for food. (D) Alice shared scissors with Katrina for cutting papers. (E) Lucia requested scissors from Tasha for assembling wooden shelves Source: Jake (Day4 12PM), Katrina (Day4 12 PM)
Evidence	
Question	**Who asked for a power bank and how did Jake respond to their request?** (A) Lucia borrowed a power bank, but Jake didn't have any left. (B) Nicous asked for a power bank, but Jake refused to let him unplug anything. (C) Shure asked for a power bank, and Jake offered heavy assistance. **(D) Tasha borrowed a power bank, and Jake offered to find another one.** (E) Alice needed a power bank, and Jake handed his own immediately. Source: Jake (Day6 12PM)
Evidence	

Table S2: Social Interaction (SI) - Multi Span category samples

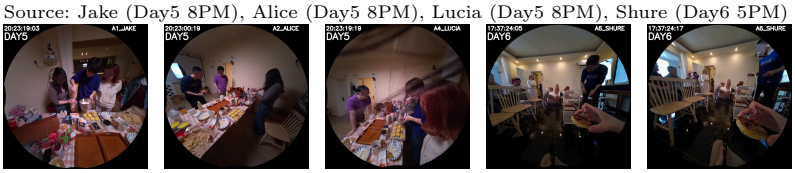
Category	**Social Interaction, Multi Span**
Question	**How did egg tart impressions shift between the sight and bite?** (A) Alice mistook the egg tart for a muffin, then Shure found the egg tart tasty despite looks. **(B) Jake mistook the egg tart for a pancake, then Shure found the egg tart tasty despite looks.** (C) Alice mistook the egg tart for a muffin, then Shure found the egg tart tasty despite looks. (D) Jake mistook the egg tart for a pancake, then Tasha found the egg tart bland despite looks. (E) Alice mistook the egg tart for a muffin, then Tasha found the egg tart bland despite looks. Source: Jake (Day5 8PM), Alice (Day5 8PM), Lucia (Day5 8PM), Shure (Day6 5PM)
Evidence	
Question	**What coffee discussions happened in both situations?** (A) Jake proposed restocking coffee beans and equipment and later Lucia offered packing help after Lucia mentioned coffee. **(B) Shure proposed restocking coffee beans and equipment, and later Alice offered packing help after Katrina mentioned coffee.** (C) Jake proposed restocking coffee beans and equipment, and later Lucia offered packing help after Lucia mentioned coffee. (D) Jake proposed restocking coffee beans and equipment, and later Alice offered packing help after Katrina mentioned coffee. (E) Shure proposed restocking coffee beans and equipment, and later Lucia offered packing help after Lucia mentioned coffee. Source: Jake (Day1 1PM), Shure (Day 1PM), Alice (Day7 6PM), Katrina (Day7 6PM)
Evidence	

Table S3: Task Coordination (TC) - Single span category samples

Category	**Task Coordination, Single Span**
Question	**How did they resolve concerns about the safety of the power strip?** (A) Lucia conducted a test and decided it was fine for usage. (B) Shure insisted safety wasn't a big issue and moved forward without discussion. (C) Shure reassured everyone it was brand-new and safe to use. (D) Katrina brought over a different power strip for better safety. **(E) Jake and Alice discussed using a reputable brand and reassured others.** Source: Jake (Day2 6PM), Alice (Day 6PM)
Evidence	
Question	**When they were trying to light the charcoal, how did the group coordinate and divide tasks to make it work?** **(A) Jake shielded the fire with a cardboard while Shure carefully transported it outside.** (B) Alice ignited the fire using dry leaves while Katrina monitored the process. (C) Lucia set up a fan to make the fire stronger while Tasha prepared water for safety. (D) Jake and Lucia used chemicals to ignite the charcoal while Alice held a bucket nearby. (E) Tasha managed the cooking station while Shure brought wood for the fire. Source: Jake (Day3 8PM)
Evidence	

Table S4: Task Coordination (TC) - Multi span category samples

Category	**Task Coordination, Multi Span**
Question	**In the flower tasks, what roles did participants take in different events?** (A) Jake observed and asked clarifying questions, later arranged the bouquets, while Katrina handled hands-on trimming and shaping. (B) Alice observed and asked clarifying questions, later asked about flower types, while Katrina handled hands-on trimming and shaping. **(C) Jake observed and asked clarifying questions, later asked about flower types, while Shure handled hands-on trimming and shaping.** (D) Alice observed and asked clarifying questions, later asked about flower types, while Shure handled hands-on trimming and shaping. (E) Jake observed and asked clarifying questions, later asked about flower types, while Katrina handled hands-on trimming and shaping. Source: Jake (Day1 11AM), JAKE (Day2 4PM), Shure (Day4 4PM)
Evidence	
Question	**In the delivery task, who coordinated unpacking and organizing, and who later managed packing in a different event?** (A) Alice coordinated unpacking and organizing, and later Lucia managed packing the delivery. **(B) Lucia and Katrina coordinated unpacking and organizing, and later on Jake managed packing the delivery.** (C) Alice coordinated unpacking and organizing, and later Jake managed packing the delivery. (D) Lucia and Katrina coordinated unpacking and organizing, and later Jake managed packing the delivery. (E) Lucia and Katrina coordinated unpacking and organizing, and later Lucia managed packing the delivery." Source: Lucia (Day4 4PM), Katrina (Day4 4PM)
Evidence	

Table S5: Theory of Mind (ToM) category samples

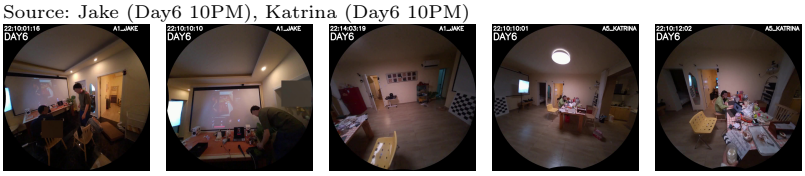
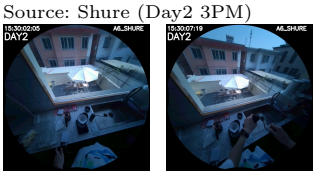
Category	**Theory of Mind**
Question	**What did Katrina wrongly assume while looking at Nicous and Violet filming?** **(A) That they were wrapping up their session immediately.** (B) That Jake wasn't involved in their filming plans. (C) That they were photographing desserts. (D) That filming had been done in the second-floor living room. (E) That Violet and Nicous would soon ask for her help. Source: Jake (Day6 10PM), Katrina (Day6 10PM)
Evidence	
Question	**Why did Shure think Choiszt might have misinterpreted the coffee instructions?** (A) Choiszt thought they needed a latte machine instead of an espresso machine. (B) Choiszt added too much water and ignored exact measurements. **(C) Choiszt confused the steps and thought milk was required.** (D) Choiszt started brewing with a wrong type of coffee grinder. (E) Choiszt assumed fruity coffee beans required special brewing techniques. Source: Shure (Day2 3PM)
Evidence	
Question	**Why didn't Katrina know where to put the big kitchen tools?** (A) Tasha was using the drawer for other purposes. (B) Katrina wasn't involved in cleaning the kitchen before. **(C) Shure had placed things randomly, making it hard to locate spaces.** (D) The storage arrangement in the kitchen had recently changed. (E) Alice moved things while reorganizing the kitchen. Source: Katrina (Day4 11AM)
Evidence	

Table S6: Temporal Reasoning (TR) - Comparison category samples

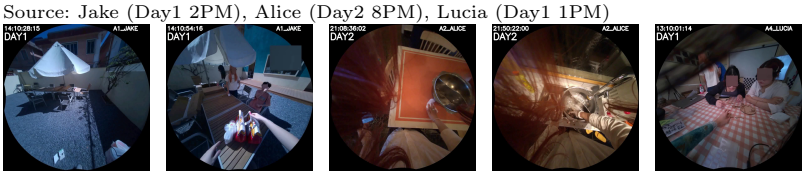
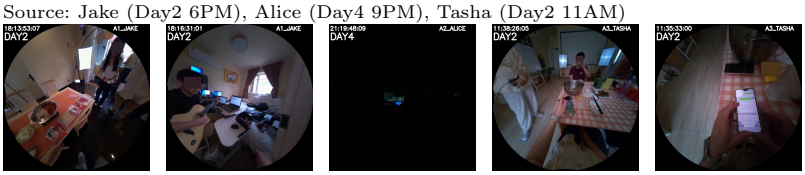
Category	**Temporal Reasoning, Comparison**
Question	**What happened between when Lucia led the puzzle completion and supply planning and when Alice deep-cleaned the kitchen after hot pot?** (A) Shure repaired a drone at dawn before anyone arrived. (B) Tasha oversaw final errands and device data consolidation discussions. (C) Lucia ran a midnight karaoke session after the meeting. **(D) Jake coordinated outdoor setup and takeout distribution in the courtyard.** (E) Katrina led themed room planning and decor decisions. Source: Jake (Day1 2PM), Alice (Day2 8PM), Lucia (Day1 1PM)
Evidence	
Question	**Which is the correct sequence of events?** A) Alice retrieved a power bank, doubled back for a forgotten item, and then took a ride-hailing car. B) Tasha wrapped up checking tickets and led a group discussion on May Day travel, even demonstrating a bus ticket booking. C) Jake set up lighting and equipment for meal prep and then worked on computer file transfers. (A) A-C-B (B) B-C-A (C) A-B-C **(D) C-A-B** (E) B-A-C Source: Jake (Day2 6PM), Alice (Day4 9PM), Tasha (Day2 11AM)
Evidence	

Table S7: Temporal Reasoning (TR) - Concurrency category samples

Category	**Temporal Reasoning, Concurrency**
Question	**Which pair of events happened at around the same time?** (A) Alice was checking out at the cashier and Shure was paying for bottled water at the self-checkout. (B) Lucia was loading boxes into a car and Jake was waiting outside the store. (C) Tasha was standing in line at the bakery counter and Katrina was asking a clerk about discounts. **(D) Jake was pushing the cart toward the exit and Lucia was spotting Jack on his phone as they reached the exit.** (E) Shure was sitting on a bench making a phone call and Alice was driving to the store. Source: Jake (Day5 4PM), Tasha (Day5 4PM)
Evidence	
Question	**When Katrina was fertilizing the flowers, and Alice was checking the dilution and water amounts, what was Jake doing?** (A) Pouring the nutrient solution into the vases himself (B) Going upstairs to do his nails and charge his phone. (C) Assembling a display rack by the whiteboard and taking inventory. **(D) Wiping down plates and the table with new wet wipes while tidying.** (E) Stepping outside to take all the trash out to the bin. Source: Jake (Day1 7PM), Alice (Day1 7PM), Katrina (Day1 7PM)
Evidence	

Table S8: Environmental Interaction (EI) category samples

Category	**Environmental Interaction**
Question	**Who used the microwave the most on DAY3?** (A) Katrina **(B) Tasha** (C) Lucia (D) Jake (E) Shure
Question	**When was the first time guitar was used on DAY2?** (A) 6 PM (B) 5 PM (C) 4 PM (D) 9 PM **(E) 3 PM**
Question	**How many people used oven on DAY1?** **(A) 4** (B) 5 (C) 1 (D) 2 (E) 3

E Benchmark Generation Prompts

You are an assistant tasked with generating meaningful and contextually grounded multiple-choice questions (MCQs) based on egocentric video data. The input consists of time-stamped captions and subtitles from multiple individuals who are recording their perspectives during the same 5-minute interval. They could be in different locations, so consider who is currently with whom based on the given context. Your job is to:

1. Read captions and subtitles of multiple people.

2. Generate **10 multiple-choice questions (MCQs)** that:

* Can **only** be answered by referencing the memories (captions + subtitles) of **multiple people** (randomly selected from two to five people).
* Generate questions **as diverse as possible**.
* Contain **5 answer options**.
* Have a **single correct answer**.
* Provide a **rationale** that explains **why the correct answer is correct**, citing **specific timestamps** from the relevant people.
* Include **a list of timestamp-name pairs** that indicate which memories should be reviewed to answer the question.

3. Return the final output as a **JSON dictionary** formatted as:

```json
{
    "qa1": {
        "question": "<Your question text here>",
        "options": ["<Option1>", "<Option2>", "<Option3>", "<Option4>", "<Option5>"],
        "answer": "<The correct option index starting from 0, e.g., 2>",
        "rationale": "<Detailed explanation with timestamp references>",
        "timestamps": [["<Name1>", "<Timestamp1>"], ["<Name2>", "<Timestamp2>"], ...]
    },
    "qa2": {
    ...
    }
}
```

Guidelines:

* Generate **a wide variety of questions** based on different events and locations. The types of questions should be as diverse as possible.
* Questions should be **specific and not abstract**. Do not put timestamps in the question and options.
* Theses are the names that you can use in the false options: "Jake", "Alice", "Tasha", "Lucia", "Katrina", "Shure". Try to use these names in the false options.
* Options should be in multiple words, not a single word or human name. False options must be longer than the correct answer.
* Options for each question should be **consistent in length** to avoid bias.
* Rephrase the questions **unformal and natural** as like a conversation looking back at the memories.

Fig. S3: System prompt for SI, TC, and ToM category QA generation.

You need to generate the questions in the "Social Interaction" category.

Ask questions about how people relate to each other in casual ways. Focus on informal moments like chatting, joking, walking together, teasing, or group fun—not tasks or goals.

Your questions should explore:

* How people engage or respond to others
* Group behaviors without a specific goal
* Who noticed, reacted to, or related to whom
* Sequences involving multiple people interacting
* Topics or meaningful information during the conversation

Avoid:

* Task- or plan-based questions
* Factual or time-based questions (e.g., “who did what first?”, "who was doing what while X was doing Y?")
* Belief or thought-based questions (e.g., “what did they think?”)

Do **not** use words like *social*, *interaction*, or *causal* in the question.
Questions should sound natural and always involve **at least three people**. Avoid excessive "who" questions; make a balance with "what," "where," "why," and "how" questions.

Fig. S4: Category prompt for SI.

You need to generate the questions in the "Task Coordination" category.

Ask questions about **goal-driven collaboration** between people working toward a **shared, meaningful objective**. Focus on how individuals **organized, contributed to, or discussed structured tasks**.

Your questions should explore:

* What roles people played in a shared task
* How responsibilities were divided or coordinated
* What steps were taken toward completing the goal
* How decisions were made during task execution

Avoid questions about:

* Casual or non-goal-directed actions (e.g. chatting, joking)
* Isolated facts not tied to collaboration
* Thoughts, beliefs, or mental states
* Timing or order of events ("who did X first?")

Questions should highlight **task structure, planning, and joint execution**, not social bonding or entertainment.

Fig. S5: Category prompt for TC.

You need to generate the questions in the "Theory of Mind" category.

Ask questions that require understanding **people's mental states** — what someone **knows, believes, sees, or intends**, especially when these differ from reality or from others' knowledge.

Focus your questions on:

* What someone **believed, assumed, or misunderstood**
* Whether someone was **aware of** something or not
* What someone **could/couldn't see or notice**
* **Differences in knowledge or beliefs** between people
* **Intentions** behind someone's actions or words

Avoid:

* Factual questions about events or actions
* Casual or social interaction questions
* Task planning or coordination questions
* Temporal sequencing questions (e.g., "what happened after X did Y?")

Prioritize reasoning about **internal, unseen mental states** — not just what happened, but what people **thought was happening**.

Fig. S6: Category prompt for ToM.

You are given multiple perspectives from different people, all describing events happening at the same time. Your task is to create a *time alignment multiple-choice question* in the following style:

Question format:
{question_format}

* Use exactly three people from the given perspectives.
* Each blank must be filled with a person's name from the provided perspectives and the action they were doing at that time.
* Ensure that all three parts come from the given perspectives so that the question tests whether the reader can align events happening simultaneously.
* Since the given captions are on the same timestamp, you can assume that the events happened at the same time if they are mentioned in the captions.

Answer format:

* Provide the correct answer as one of five options.
* Include four distractor options that are plausible but incorrect.
* Lengths of distractor options should be similar to the correct answer.
* The answer index should be the index of the correct option.
* Distractor options **should be generated newly (factually incorrect), not from the given perspectives.**
* When describing the actions, **focus on broader events across the person's perspective** rather than fine-grained actions from captions.

Output format (JSON dictionary):
{format_example}

Fig. S7: Category prompt for TR.

You are given:
1. A target **object** or **action** that the question should be about.
2. A **question template**.
3. A **context** that describes a situation, actions, or events related to the target object or action.

Your task is to:

* Generate a **single question** that fits the template and is grounded in the given context.
* Create **five multiple-choice options (MCQs)**:
  * **1 correct option**
  * **4 distractor options** (plausible but incorrect)
* Provide how are you confident about the correct answer in the question from 1 to 5.
* Ensure that the correct answer is based on the events when an **object that was directly used** (direct interaction, e.g., "He wrote with a pen").
* Indirect interaction, such as walking towards to the object, does **not** count as usage. Except for the object for watching or looking, physical interaction is required.

* Provide the final output in the following JSON dictionary format:

```json
{
        "question": "GENERATED_QUESTION",
        "options": ["OPTION0", "OPTION1", "OPTION2", "OPTION3", "OPTION4"],
        "answer_idx": CORRECT_OPTION_INDEX (starting from 0),
        "confidence": "CONFIDENCE_LEVEL" (from 1 to 5)
}
```

Constraints:
* Only one option can be correct.
* The `answer_idx` should match the position of the correct answer in the `options` list.

Fig. S8: Category prompt for EI.

F Question Templates

Table S9: Question templates for the TR category.

- When [Person A] was [doing X], and [Person B] was [doing Y], what was [Person C] doing?
- Which of the following happened while [Person A] was doing [Action A]?
- Which pair of events happened at around the same time?
- Which is the correct sequence of events?
A) Person A did X, B) Person B did Y, C) Person C did Z
- Which of the following happened first?
A) Person A did X, B) Person B did Y, C) Person C did Z, D) Person D did W, E) Person E did V
- Which of the following happened last?
A) Person A did X, B) Person B did Y, C) Person C did Z, D) Person D did W, E) Person E did V
- What happened between when [Person A] was [doing X] and when [Person B] was [doing Y]?

Table S10: Question templates for the EI category.

- When was the first time ([object] was used/[action] done) on DAY[day]?
- Who (used [object]/did [action]) the most on DAY[day]?
- How many people (used [object]/did [action]) on DAY[day]?
- Who was the last person to (use [object]/do [action]) on DAY[day]?

G Prompts for EgoMAS

You are helpful planner choosing which person's perspective should the response model refer to during answering the question.

Context: {context}

Question: {query}
Do not answer the question now. Follow the instructions below.

Based on the context, select one person (should be one of the list: Jake, Alice, Katrina, Lucia, Tasha, Shure) and generate query which memory should be referred to from that person. The question could be a sub-question for the original question.

For example, the answer should be like this: [{'name' : 'Jake', 'query' : 'order McDonalds for lunch'}, {'name' : 'Katrina', 'query : 'made pancakes'}, {'name' : 'Lucia', 'query : 'clean the kitchen'}]

Try to generate multiple elements in the list if necessary.

Fig. S9: Prompt for agentic-wise dynamic retrieval in EgoMAS.

These are captions of 10 minutes video captured from multiple people. Write a structured summary of the captions.
The output should consist of multiple items in the following format:

```
{
   "name": {
      "type": List[str],
      "description": "All names that involved in the event.",
   },
   "action": {
      "type": str,
      "description": "Action that happened in the event.",
   },
   "location": {
      "type": str,
      "description": "Location of the action.",
   },
   "detail": {
      "type": str,
      "description": "Important details of the action.",
   },
}
```

For example, if there are 3 key events in the context, the output should be like this:

```
[
   {
      "name": ["Jake", "Katrina"],
      "action": "Prepare for the lunch. Make pancakes and omelets.",
      "location": "Kitchen",
      "detail": "Jake first made batter for the pancakes using powder and water, saying 'I need to make batter for the pancakes.'. And then he turned on the stove ...",
   },
   {
      "name": ["Lucia", "Jake", "Katrina", "Tasha"],
      "action": "Have a lunch together. Discuss about the plan for the party.",
      "location": "Living room",
      "detail": "Lucia brought the plates and placed them on the table, saying 'Let's get started with lunch.'. Jake served the pancakes and omelets to everyone. Katrina poured some juice into glasses. Tasha sat down and thanked Jake for cooking. As they ate, Lucia asked, 'So, what's the plan for the party?'. ..."
   },
   {
      "name": ["Shure"],
      "action": "Play the piano.",
      "location": "Living room",
      "detail": "Shure sat down at the piano and started playing. He played a few notes and then started playing a song. He played the song for a while and then ..."
   }
]
```

Try to contain as many key events as possible. Do not miss any important events from the context. Also, put all the details of the action in the detail field.

Fig. S10: Prompt for shared memory construction in EgoMAS.

H Analysis on Input Context Modality

To investigate whether MA-EgoQA can benefit from information across different modalities, we compare the performance of the WorldMM-8B and EgoMAS models when using only text captions versus using both text captions and video frames as context during the response generation stage, as shown in Tab. S11. For WorldMM-8B, the performance consistently improves when video frames are provided together with text captions, indicating that the MA-EgoQA benchmark requires visual knowledge contained in the frames. This observation is further supported by the QA examples in Tab. S12. In these examples, actions such as placing items randomly, grabbing tissues, and switching to a thinner brush, as well as states such as being in another room upstairs, can be effectively identified from the video rather than from the transcripts. These results demonstrate that visual information is essential for answering certain questions in MA-EgoQA.

On the other hand, for the EgoMAS models, providing video frames improves performance in the SI, TR, and EI categories, but leads to performance drops in the TC and ToM categories. These results suggest that including video frames is not always beneficial and can even distract the model when unnecessary frames are provided, as also observed in the `All Frame Concat Baselines` in Tab. 2. Therefore, future research should focus on adaptively leveraging the necessary modalities and developing effective methods for frame selection.

Table S11: Analysis of input context modality on the MA-EgoQA benchmark.

Models	Text	Video	SI	TC	ToM	TR	EI
WorldMM-8B	✓	✗	27.13	30.37	17.87	24.04	19.78
	✓	✓	29.26	34.50	21.70	25.09	22.56
Δ			+2.13	+4.13	+3.83	+1.05	+2.78
EgoMAS (Qwen2.5-VL-7B-Inst)	✓	✗	37.77	38.64	25.96	34.15	36.49
	✓	✓	39.10	38.02	24.68	35.19	37.88
Δ			+1.33	-0.62	-1.28	+1.04	+1.39
EgoMAS (Gemini-3-preview)	✓	✗	50.53	47.52	41.70	51.92	44.29
	✓	✓	50.80	46.90	40.85	54.70	47.08
Δ			+0.27	-0.62	-0.85	+2.78	+2.79

Table S12: QA samples of MA-EgoQA that require visual information.

Q. Why didn't Katrina know where to put the big kitchen tools? A. Shure had placed things randomly, making it hard to locate spaces.
Q. Which pair of events happened at around the same time? A. Tasha was grabbing some tissues from the table and Alice was switching to a thinner brush to start her eye makeup.
Q. Why didn't Katrina notice the glasses charging issue earlier? A. She was upstairs in another room at the time.