参赛学生姓名：莫霁然

中学：北京市第一0一中学

省份：北京市

国家/地区：中国

指导老师姓名：钟方威 周宇辰

指导老师单位：北京师范大学

北京市第一0一中学

论文题目：Structured Higher-Order Mental State Inference for Multi-Modal Machine Theory of Mind

Structured Higher-Order Mental State Inference for Multi-Modal Machine Theory of Mind

Student:
Jiran Mo
Beijing No.101 High School

Supervisors:
Fangwei Zhong
Beijing Normal University
Yuchen Zhou
Beijing No.101 High School

Abstract

Theory of Mind (ToM) enables reasoning about others' beliefs, intentions, and knowledge, especially with higher order forms being crucial for complex social interaction. Large language and vision language models have shown weak ToM capabilities in recursive inference, multi-modal grounding and contextual continuity. Current approaches such as AutoToM and MuMA-ToM, while being able to tackle lower-order reasoning decently, are unable or not enough competent to process higher-order tasks. To address these challenges, we propose an innovative approach for structured multi-modal higher-order mental state inference which consists of the following components: Sequential Monte Carlo for belief propagation designed to break down complex social interactions and reduce the negative effects of the interactions' extent on reasoning accuracy, semantic retrieval augmented reasoning designed to retrieve examples that closely resemble the one being processed and to thus improve reasoning accuracy, and reflective memory management which utilizes previous reasoning sessions to expand the ground truth database. To address the gaps in the field of multi-modal higher-order ToM dataset, we construct a novel photorealistic dataset of multi-agent scenarios with zeroth- to fourth-order reasoning tasks enabling diverse actions and communications with partial observability. Experimental results show that our approach can complete the complex higher-order tasks, especially third- and fourth-order ones, which are not supported by the state-of-the-art approaches. Ablation studies further demonstrate the effectiveness and unique value of the proposed technologies for tasks with different orders.

Keywords—Theory of Mind, higher-order reasoning, sequential Monte Carlo, semantic retrieval augmented reasoning, reflective memory management

Table of Contents

1 INTRODUCTION

Theory of Mind (ToM) refers to the ability to identify mental states, such as beliefs, goals, intentions, and emotions of oneself and others (Premack & Woodruff, 1978). ToM plays a crucial role in social interactions such as cooperation, deception, persuasion, behavior prediction, and conflict navigation. Mental states can be categorized into lower-order states, which are limited to direct perceptions and beliefs about the environment, and higher-order states, which are recursive embeddings of others' mental states, an "order" corresponding to a layer of embedding. For example, a first-order belief is a simple attribution such as "I think that she is going left," a second-order belief embeds another person's perspective as in "I think he thinks that I am going left," and a third-order belief adds yet another layer, e.g., "I think she thinks that I think that she is going left". A typical example is illustrated in Figure 1.

Higher-order ToM is especially significant because many real-world interactions depend on such recursive reasoning. In law and politics, reasoning about others' nested beliefs can determine judgments and negotiation strategies; in finance and commerce, higher-order ToM could be very useful for facilitating cooperation; and in literature, higher-order ToM enables the comprehension of irony, deception, and complex character motivations.

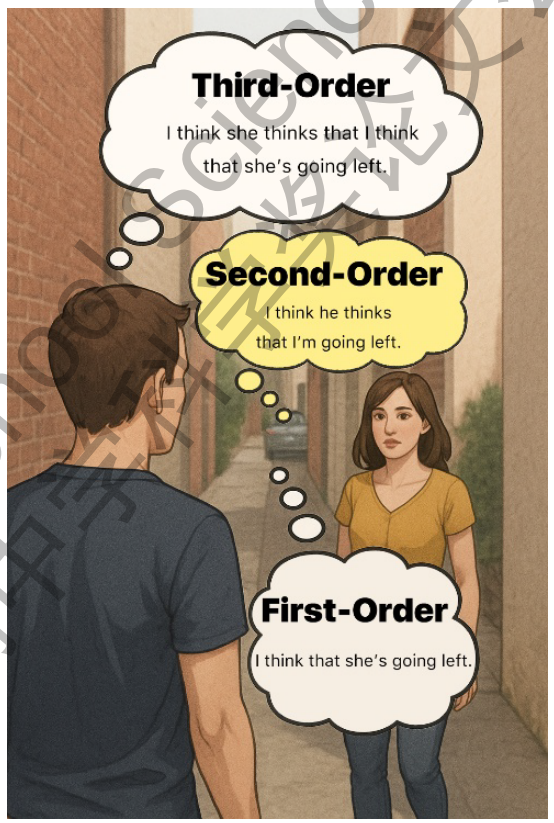

Figure 1. Illustration of higher-order ToM through recursive mental-state embeddings. This visualization highlights the progression from lower-order reasoning, which concerns direct perceptions and beliefs, to higher-order reasoning, which involves increasingly complex embeddings of others' mental states.

As AI becomes increasingly involved in our daily lives, equipping AI with ToM capabilities becomes critical for better human-AI interactions. Achieving practical machine ToM requires methods capable of processing multi-modal inputs, including text, video, and other sensory information, with continuity and depth. Many attempts have been made to achieve machine ToM, the earliest being symbolic reasoning (e.g., Bolander & Andersen, 2011; Stuhlmüller & Goodman, 2014). While such methods offer transparency and explicit logic, they suffer from labor-intensive, handcrafted modelling that impedes scalability and flexibility, particularly for recursive higher-order reasoning. Large Language Model (LLM) and Vision-Language Model (VLM) have experienced significant and rapid advances in recent years. While studies (Kosinski et al., 2023; Ullman et al., 2023; Zhang et al., 2024) show a positive correlation between the improvement in language abilities and the improvement in ToM abilities, state-of-the-art LLMs still exhibit significant limitations in ToM reasoning. Specifically, they often fail to distinguish between intended actions, sub-optimal decisions, and failed attempts, leading to misinterpretations; moreover, AI neglects temporal continuity and critical contextual information while processing videos.

Recently, several research studies aiming to improve LLMs' ToM abilities have introduced Bayesian Inverse Planning (BIP). This new approach addresses some scalability issues but introduces new challenges. Methods like MMToM-QA (Jin et al., 2024) and MuMA-ToM (Shi et al., 2024) often struggle to distinguish nuanced intentions and fail to handle the escalating complexity of higher-order logic. More advanced methods, such as AutoToM (Zhang et al., 2024), which incorporate structured recursive frameworks, remain vulnerable to hallucinations. Furthermore, most of these methods operate from an external, observer-centric perspective, limiting their ability to handle partial observability and information asymmetry. Additionally, existing methods that support multi-modal input often have to also depend on information of other modalities,

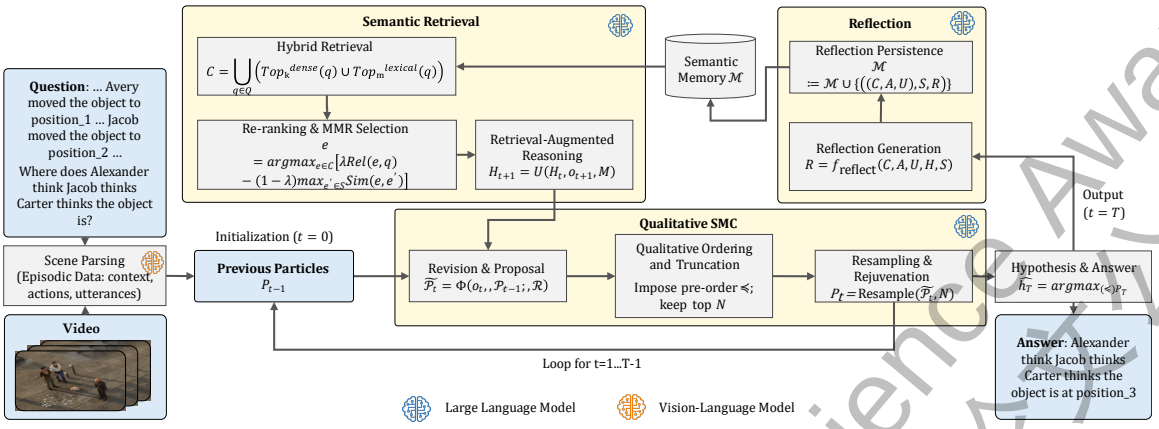

Figure 2. Overview of the proposed reasoning pipeline. Multi-modal inputs, including videos and questions, are first parsed by a pre-trained VLM into structured textual representations. The parsed story is then processed through retrieval, Sequential Monte Carlo (SMC), and reflective memory management, which iteratively retrieves relevant examples, generates or updates hypotheses, and performs reflection, by collaborating with LLM until the story is complete. Finally, the reasoning outputs are aggregated to answer higher-order ToM questions.

sacrificing their applicability in real-world scenarios. In summary, current machine ToM methods exhibit critical shortcomings, including (1) extremely limited higher-order reasoning capability, (2) inability to reliably identify failed attempts and sub-optimal behaviors, (3) low generalizability, and (4) loss of temporal and contextual continuity in multi-modal inputs.

To address the above persistent challenges, we propose a structured higher-order mental state inference approach for multi-modal ToM reasoning in multi-agent scenarios. Our method can handle partial observability and information asymmetry inherent in multi-agent interactions very well. Our method integrates a Sequential Monte Carlo (SMC) module, which dynamically updates hypotheses about agents' intentions and beliefs. Semantic retrieval augmented reasoning is designed to retrieve examples that closely resemble the one being processed and to thus improve reasoning accuracy. A reflective memory management component is designed to utilize previous reasoning sessions to expand the ground truth database. Furthermore, to address the gaps in this field and support the development and evaluation of our framework, we also create a new photorealistic dataset featuring multi-agent social scenarios requiring higher-order mental state reasoning using UnrealZoo (Zhong et al., 2024). The pipeline of the proposed method is illustrated in Figure 2, with details described in Section 4. Experimental results show that our approach can complete the complex higher-order tasks, especially third- and fourth-order ones, which are not supported by the state-of-the-art approaches. It also significantly outperforms them on second-order tasks. Ablation studies further demonstrate the effectiveness and unique value of the proposed technologies for tasks with different orders.

Our main contributions can be summarized as follows:

- We introduce an automatic data generation pipeline built on photo-realistic virtual worlds and construct a Multi-modal dataset for higher-order multi-agent ToM reasoning.
- We propose a structured higher-order mental state inference approach for multi-modal ToM reasoning in multi-agent scenarios, including Sequential Monte Carlo for belief propagation, semantic retrieval-augmented reasoning, and reflective memory management.
- We conduct experiments for evaluating the higher-order multi-modal ToM reasoning capability of different methods and the value of the proposed technologies for different tasks with specific orders, and demonstrating the usability of our dataset and the significant improvements of our approach.

2 RELATED WORK

Numerous approaches from different perspectives have been developed to enhance machine ToM reasoning, including psychological principles, reasoning approaches with or without LLM or VLM, datasets and tools.

Psychological Principles Psychological and cognitive science literature has long been a theoretical backbone for ToM modelling. Perner and Wimmer (1985) conducted a seminal study demonstrating that children begin to understand second-order beliefs, such as "Mary thinks that John thinks…", around seven. Moore et al. (1990) extended this line of research by exploring children's grasp of speaker-listener dynamics, showing that advanced ToM plays a critical role in pragmatic language understanding. Kinderman et al. (1998) examined how higher-order ToM influences adult causal attributions, especially in clinical populations. Miller (2009) reviewed various developmental findings and emphasized the roles of language, executive function, and social experience in fostering higher-order ToM. That same year, Apperly and Butterfill (2009) proposed a dual-systems theory, suggesting that humans rely on a fast, implicit system for basic mental state tracking and a slower, cognitively demanding system for recursive, higher-order reasoning. Most recently, Osterhaus and Koerber (2021) used structural modelling to reveal that middle childhood ToM reasoning involves multiple interrelated components, highlighting its cognitive complexity and layered nature.

Approaches Early approaches such as symbolic logic-based methods (Bolander & Andersen, 2011; Stuhlmüller & Goodman, 2014), BIP without large models (Baker et al., 2009; 2011), I-POMDP framework (Gmytrasiewicz & Doshi, 2005), and script-based reasoning (Schank & Abelson, 1997) laid the foundation for machine ToM. However, they struggle with flexibility and generalization to real-world complexity. With the rise of LLMs and VLMs, more recent efforts can be broadly divided into two categories: prompt-based (Kosinski, 2023) and structurally augmented approaches (Jin et al., 2024; Shi et al., 2024; Zhang et al., 2024). These newer methods demonstrate improved performance in simulating mental state reasoning. Still, significant challenges remain. Prompt-based methods often show limited gains beyond basic prompting, while structurally augmented methods encounter difficulty when handling nuanced higher-order beliefs, particularly in dynamic multi-agent environments. Further innovation is needed to close the gap between human and machine ToM. In addition, recent studies have explored alternative paradigms to enhance LLMs' reasoning. Reflexion introduces verbal reinforcement learning to let language agents iteratively improve via reflective feedback (Shinn et al., 2023). Retrieval-Augmented Generation (RAG) integrates external knowledge retrieval into the generation process to reduce hallucinations and enhance reasoning, especially in knowledge-intensive tasks (Gao et al., 2023).

Datasets To support training and evaluation, many ToM related datasets have been created (Rabinowitz et al., 2018; Jain et al., 2019; Sap et al., 2019; Kosinski, 2023; Zhang et al., 2023; Zhang et al., 2024). However, these datasets present several key limitations. Many are either unimodal or single agent, for example, Social IQa (Sap et al., 2019) and ToM Tasks (Kosinski, 2023) are purely textual, lacking perceptual grounding, while SCoNe (Jain et al., 2019) remains focused on written narratives. Datasets like CLEVR-Mental-State (Zhang et al., 2023) attempt to introduce visual inputs but rely on synthetic, static scenes with limited ecological validity, and higher-order ToM is also rarely addressed. Most datasets target only first-order inferences, and even those that support higher-order reasoning, like Hi-ToM (He et al., 2023), suffer from low-fidelity visuals and rigid character actions. Additionally, agent embodiment and real-time interactivity are often absent, as in Rabinowitz et al. (2018), where agents operate in abstract environments without grounded social exchange. These factors limit how much models trained on such datasets can generalize to complex, real-world mental state reasoning.

Utility Tools for creating multi-modal datasets have evolved alongside advancements in machine ToM. VirtualHome (Puig et al., 2018) enables the simulation of household activities through scripted sequences of symbolic actions, supporting studies on task planning and high-level execution. However, its abstraction of physical interactions and low-fidelity video output limit its utility for tasks requiring motor precision or rich sensory input. iGibson 2.0 (Li et al., 2021) improves on this by offering a physically grounded, object-centric environment with interactive manipulation and realistic visuals. It supports diverse sensory and control modalities, making it better suited for embodied learning. Still, its computational demands can hinder scalability and integration with higher-level reasoning. UnrealZoo, built on Unreal Engine, represents the most advanced platform. It combines photo-realism, dynamic interactions, and diverse scene layouts, overcoming key shortcomings of its predecessors. Unlike VirtualHome's symbolic abstractions, UnrealZoo

(a) Before Step 5

(b) Step 5

(c) Step 14 & 15

A third-order episode with 5 agents and 15 steps

Step 1: Carter, Avery, Jacob, Jackson and Alexander were on the rooftop.
Step 2: The turkey was on the trash bin.
Step 3: Carter made no movements and stayed on the rooftop for 1 minute.
Step 4: Carter went downstairs.
Step 5: Avery moved the turkey near the vent.
Step 6: Avery went downstairs.
Step 7: Jacob moved the turkey under the patio umbrella.
Step 8: Jacob went downstairs.
Step 9: Jackson made no movements and stayed on the rooftop for 1 minute.
Step 10: Jackson went downstairs.
Step 11: Alexander made no movements and stayed on the rooftop for 1 minute.
Step 12: Alexander went downstairs.
Step 13: Carter, Avery, Jacob, Jackson and Alexander entered the living room.
Step 14: Jacob publicly claimed that the turkey was on the trash bin.
Step 15: Alexander privately told Jacob that the potato was in the kitchen drawer.
Question: *Where does Alexander think Jacob thinks Carter thinks the potato is?*
Answer: *on the trash bin*

Figure 3. A typical example interaction from our dataset that shows the complexity of higher-order ToM reasoning. For one to reach the correct answer, they must comprehend the notion of partial observability and keep track of the mental states of multiple agents.

enables both visual and physical realism, and it offers broader, more flexible environments than iGibson 2.0, making it especially well-suited for next-generation multi-modal ToM research.

3 HIGHER-ORDER MULTI-MODAL DATASET GENERATION

We construct a multi-agent simulation dataset using UnrealZoo that covers zeroth- to fourth-order problems. Our dataset contains high resolution videos at 60fps that show social interactions happened in photorealistic 3D environments with multiple rooms and interactive objects. The stories in our dataset are based on the Hi-ToM (He et al., 2023) dataset, each containing approximately 15 actions or utterances. An example is shown in Figure 3.

Within this simulation, agents engage in rich multi-modal interactions. They can manipulate objects, e.g., pick up, move, or drop items in containers and communicate with one another either publicly or privately. A public speech is audible to all agents in the vicinity. In contrast, a private utterance is directed at a specific agent. Meanwhile, all agents continuously observe each other's actions when they are within line-of-sight or the same room, which means an event, such as an object being moved or a statement being made, is only witnessed by those present. This controlled communication and observability structure leads to knowledge asymmetry: some agents gain information that others lack, setting the stage for higher-order ToM reasoning. The process of data generation is shown in Algorithm 1.

We generate the dataset using the aforementioned environment and interactions. Each scenario is presented as a short video with synchronized dialogue transcript and visual events, accompanied by a ToM reasoning task. The question with multiple choices focuses on an agent's nested beliefs about others. For example, a

Algorithm 1 General procedure for UnrealZoo/UnrealCV dataset recording
(Initialize environment, control agents and objects, record annotated frames)

Connect to UnrealCV API via `gym.make(env_id)`; reset environment
Initialize capture settings (resolution, viewmode, flags, FOV); set camera pose
Spawn agent Blueprints (`BP_Character_C_*`) in circle layout; assign `app_id`
for each frame block **do**
  Set or keep camera; advance env with NOOP; sleep to meet target FPS
  if navigation is required **then**
    Issue `nav_to_goal/obj` to move BP to target (x, y, z)
    Loop `tick()` until within tolerance or timeout; **break** on stall
  end if
  if pickable object interaction is required **then**
    Attempt `pick_up`: repeat {approach → `set_pickup` → check `is_picked`} up to N times
    To drop: navigate to drop (x, y, z); toggle `set_pickup`; verify not picked
  end if
  Optionally print public/private utterances; short hold for annotation
end for
Close environment and disconnect

Algorithm 1. Algorithmic workflow for episode generation. Starting from a Hi-ToM–derived scenario seed, the procedure (1) instantiates an UnrealZoo scene with agents and interactable objects, (2) schedules a stepwise script of actions and utterances, and (3) renders the video and aligns a time stamped dialogue transcript.

question may ask what an agent thinks of the thought about another agent of yet another agent. This question-answer format enables quantitative evaluation of a method or human participant's understanding of the scenario's mental state dynamics.

Each video scenario consists of approximately 15 discrete steps, depicting a sequence of agent actions and communications that unfold the story. The scenarios are explicitly designed to require higher-order ToM inferences: each one targets a zeroth-, first-, second-, third-, or fourth-order belief reasoning challenge. In all cases, the correct answer to the scenario's question hinges on understanding these nested beliefs rather than just simple beliefs or facts.

Throughout these events, the agents' beliefs are formed and updated step by step, resulting in a complex tapestry of who knows what – and who knows that others do not know. Crucially, by the end of this scenario, the participants have misaligned mental states that require higher-order reasoning to untangle. This design of the dataset ensures that evaluating on our 300 episodes rigorously tests a method's higher-order ToM reasoning in multi-modal, dynamic social environments.

4 METHOD

Higher-order ToM reasoning in a multi-agent environment is challenging. In order to track and infer the evolving mental states of multiple agents, we design a reasoning mechanism that maintains a hypothesis space of possible beliefs for each agent and updates these hypotheses as new observations of actions or utterances arrive. It consists of three components: (1) Sequential Monte Carlo (SMC) mechanism to initialize and rejuvenate a set of belief hypotheses over time, (2) a Semantic Retrieval Augmented Reasoning (SRAR) component that leverages stored examples of ground truth and successful past experiences to guide future hypothesis updates, and (3) a Reflective Memory Management (RMM) component that generates feedback after each inference episode and stores it as metadata for future retrieval. This section details the algorithms and notations of our method, describing how it propagates beliefs and learns from each scenario.

4.1 Belief Modeling for Multi-Agent Higher-Order ToM

At the core, our method infers higher-order beliefs, enabling reasoning about what agents believe about others' beliefs. In this section, we define symbols and notations for belief modeling that would be later used to explain our algorithm and provide an overview of how we model nested beliefs. A first-order belief $B_i(p)$ denotes

(a) Episodes and Scenes

(b) Pickable Objects

(c) Human Characters

Figure 4. Illustration of the diversity of our dataset: agents operate in photorealistic, multi-location environments with over 20 types of interactive objects. (a) Various episodes and scenes. (b) Pickable Objects: A diverse set of interactable objects used in the scenarios, enabling rich manipulation actions. (c) Human Characters: The pool of animated agents used to enact multi-agent social scenarios with varied appearances and roles.

agent i's belief in proposition p, while a second-order belief $B_i\left(B_j(p)\right)$ denotes i's belief about j's belief in p.

In realistic social settings, no agent possesses complete information about the environment. Our method therefore maintains a separate knowledge state for each agent, capturing only the events they directly observed or were informed of. Each event or utterance is tagged with the set of perceiving agents, ensuring belief updates occur only from appropriate perspectives. This design encodes information asymmetry, allowing the model to represent divergent beliefs and track how false beliefs persist when an agent misses critical observations; that is, our method takes agents' ignorance into consideration, such as deducing that an agent A does not know p if they missed the relevant event. This design helps to ensure coherence: when an agent

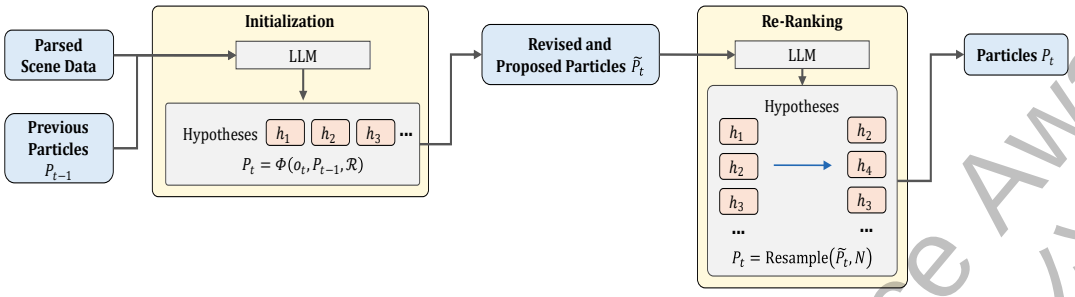

Figure 5. Sequential Monte Carlo for Belief Propagation

learns new information, only the beliefs of agents who observed this learning are updated, while others retain their prior false or outdated beliefs.

4.2 Sequential Monte Carlo for Belief Propagation

Taking into consideration the importance of continuity, our study leverages the principles of SMC to structure the belief update process as illustrated in Figure 5. Specifically, we use an LLM to qualitatively generate, assess, and rejuvenate particles rather than doing so quantitatively as in the classical approach.

The LLM assumes the core functional roles traditionally handled by mathematical operations in SMC. Instead of proposing new particles from a probabilistic motion model, the LLM generates new belief hypotheses based on its understanding of the evolving narrative. Rather than assigning and updating numerical weights based on observation likelihoods, the LLM qualitatively assesses the plausibility of each hypothesis in light of new events, effectively ranking them. Finally, the resampling and rejuvenation steps are replaced by a logical process where less plausible hypotheses are discarded and replaced with new, more coherent alternatives generated by the LLM, ensuring the diversity and relevance of the hypothesis pool.

In our framework, each "particle" is not a state vector but rather a rich, text-based hypothesis representing a candidate mental state for an agent. For example, a particle might encapsulate a complex belief such as, "Agent A incorrectly believes that Agent B thinks the object is still in the box, because A did not witness B seeing the object being moved". By manipulating a set of these descriptive hypotheses over time, our method can track the intricate, nested, and often non-obvious mental states of multiple agents in a dynamic environment. The following subsections detail the specific algorithms for this qualitative belief propagation process.

4.2.1 Initialization

We employ SMC algorithm to perform belief propagation for each agent across a sequence of observations. In this context, each “particle” represents a candidate hypothesis about an agent’s mental state, for example, what that agent believes about a particular fact or situation. As agents interact in the scenario, their beliefs may involve nested and higher-order reasoning about each other. We denote an n-th-order belief concerning a proposition p using a nested notation. A first-order belief is written as $B_i(p)$, a second-order belief as $B_i\left(B_j(p)\right)$, and a general n-th-order belief is defined as:

$$B_{(i_1)}\left(B_{(i_2)}\left(\dots\left(B_{(i_n)}(p)\right)\dots\right)\right) \tag{1}$$

where, for example, $B^2_{a,b}(p)$ means “agent a believes that agent b believes p”, and so on. The hypothesis space maintained by our SMC module includes both lower-order beliefs as direct beliefs about the environment and higher-order beliefs as beliefs about others’ beliefs.

At the start of an episode (time $t = 0$), the model draws an initial set of hypotheses for each agent’s mental state based on the context. Formally, let $\mathcal{H}_{i,0} = \{H^{(p)}_{i,0}\}_{p=1}^{N}$ be the set of N hypothesis particles for agent i’s

state, e.g. what agent i believes about key facts at the beginning of the scenario. These initial particles $\{H_{i,0}^{(p)}\}$ are generated by a function implemented via an LLM prompt that infers plausible mental states consistent with the scenario. This serves as an approximate prior distribution over agent i's beliefs before any dynamic observations. We rank each hypothesis $H_{i,0}^{(p)}$ by plausibility as determined by the language model.

4.2.2 Belief Update with New Observations

Maintaining consistency of beliefs over time is critical. In dynamic multi-agent environments, new information arrives sequentially, and agents' beliefs must be updated accordingly. A major challenge is ensuring that these updates do not introduce contradictions either within an agent's own belief set or across the nested beliefs of different agents.

An example of maintaining consistency is handling a change-of-state with limited observability. Suppose initially agent A and agent B both believe a box is in Room1. Then the box is moved to Room2 while B watches but A does not. After this event, our model updates B_B^1("Box in Room2") to true, and correspondingly B's belief that the box is in Room1 to false. For A, however, $\mathrm{B}_{\mathrm{A}}^1$("Box in Room2") remains false (since A didn't see the move, A continues to believe the box is in Room1). Now consistency requires that B's second-order belief about A reflects A's ignorance: specifically, $\mathrm{B}_{\mathrm{B,A}}^2$("Box in Room2") should be false, meaning B believes that "A does not know the box's new location". Our model ensures this alignment in its representation of B's beliefs about A. Later, if A is informed or observes that the box is in Room2, we update A's beliefs accordingly. We also update $\mathrm{B}_{\mathrm{B,A}}^2$("Box in Room2") once B becomes aware that A has learned the new location, for instance, if B saw A open the box in Room2 or heard someone tell A. Through these carefully coordinated updates, the model maintains a globally coherent picture of all agents' minds over time.

Let $\mathcal{H}_{i,t-1} = \left\{h_{i,t-1}^{(1)}, \cdots, h_{i,t-1}^{(K)}\right\}$ be the agent-centric hypothesis set at time $t-1$, where each h is a complete, perspective-aware assignment over first and higher order beliefs, e.g., whether $B_a^1(p)$ or $B_{b,a}^2(p)$ holds, consistent with all prior events. When a new observation O_t as an action or utterance arrives, the model does not compute numeric likelihoods or weights. Instead, it prompts a language model to qualitatively examine each hypothesis for coherence with O_t, revise those that can be reconciled by incorporating the newly revealed information and visibility constraints, and propose alternatives where reconciliation is untenable. We summarize this step by an abstract update operator:

$$\mathcal{H}_{i,t} = \mathrm{Update}\big(\mathcal{H}_{i,t-1}, O_t, E_t\big), \tag{2}$$

where E_t represents a set of relevant examples retrieved from the memory store, which are incorporated into the language model's prompt to guide more accurate qualitative reasoning, as detailed in Section 4.3. Operationally, the update proceeds in the following two phases:

- **Revision and Proposal**: applies a perspective-aware transformation to each $h \in \mathcal{H}_{i,t-1}$ and injects new candidates only when needed;
- **Qualitative Ordering and Truncation**: imposes an LLM-internal pre-order $\preceq_t$ (a "most-to-least plausible" ranking with no numeric scores) and keeps the top K.

We define:

$$\widetilde{\mathcal{H}_{i,t}} = \left\{\mathrm{Reflect}_{E_t}(h, O_t) \;\middle|\; h \in \mathcal{H}_{i,t-1}\right\} \cup \mathrm{Propose}_{E_t}\big(O_t, \mathcal{H}_{i,t-1}\big), \mathcal{H}_{i,t} = \mathrm{Select}_K!\left(\mathrm{Rank}_{o_t}!\left(\widetilde{\mathcal{H}_{i,t}}\right)\right). \tag{3}$$

here, $\mathrm{Reflect}_{E_t}(h, o_t)$ edits only those particles that the new event licenses from the correct perspective, e.g., a private perception changes $B_v^1(\cdot)$ for the viewer v but leaves non-viewers' beliefs intact; a public assertion updates $B_{r,j}^2(\cdot)$ for each listener r about speaker j's belief. If a prior h cannot be reconciled without violating perspective safety, $\mathrm{Propose}_{E_t}$ introduces a fresh explanation that honors who saw or heard what. The LLM then produces $\preceq_t$ implicitly—articulating which explanations are more compelling after O_t—and Select_K clips the list to capacity, ensuring $\left|\mathcal{H}_{i,t}\right| = K$. In effect, the update functions as a particle filter's correction

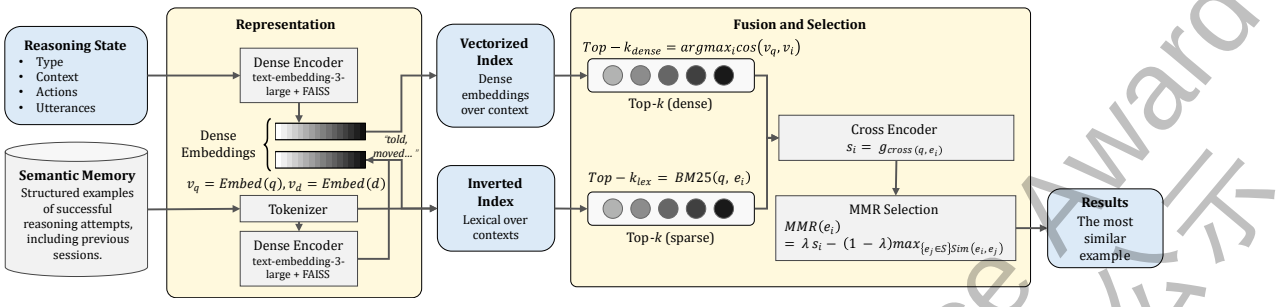

Figure 6. Semantic Retrieval Augmented Reasoning

step without numeric scoring: the model keeps, edits, or replaces hypotheses purely by qualitative reasoning over content and visibility, optionally nudged by retrieved analogies E_t.

4.2.3 Resampling and Rejuvenation

Following revision, the model executes a qualitative analog of resampling. It retains the revised hypotheses that remain coherent after O_t and rejuvenates the rest by replacing them with newly proposed alternatives, again with no numeric weights involved. Conceptually, the procedure partitions the previous set into those that survive reflection and those that are discarded, then fills any vacancies with new candidates tailored to O_t optionally conditioned on E_t to keep diversity at capacity K. We summarize the pool-construction step by:

$$\widetilde{\mathcal{H}_{i,t}} = \underbrace{\left\{ \text{Reflect}_{E_t}(h, O_t) \;\middle|\; h \in \mathcal{H}_{i,t-1},\ \text{coherent under } O_t \right\}}_{\text{retained \& revised}} \cup \underbrace{\text{Propose}_{E_t}!\left(O_t,\ \mathcal{H}_{i,t-1}\right)}_{\text{rejuvenated replacements}}. \tag{4}$$

The final $\mathcal{H}_{i,t}$ is obtained by qualitatively ordering $\widetilde{\mathcal{H}_{i,t}}$ (the LLM's implicit $\preceq_t$) and truncating to the top K, as already expressed above. In practice, this resampling-with-rejuvenation functions as a "keep the good, replace the rest" cycle that continually refreshes the hypothesis set. Retained items provide temporal stability and carry forward long-range constraints, e.g., persistent ignorance or entrenched false beliefs, while rejuvenated items explore new explanatory avenues unlocked by O_t, e.g., alternative attributions of intent or knowledge transfer paths. Because proposals are context-conditioned and optionally memory-augmented via E_t, the injected diversity is targeted: it expands exactly where the prior set failed to reconcile the latest event. Over time, repeating this cycle yields a top K pool that remains both focused on the most defensible interpretations and adaptable to unexpected turns in the narrative, all without invoking numeric likelihoods, scores, or probabilities.

4.3 Semantic Retrieval Augmented Reasoning

While the SMC mechanism maintains temporal coherence of beliefs, it may still suffer from limited generalization and vulnerability to complex multi-agent interactions. To address these shortcomings, we introduce a semantic retrieval module that supplements SMC with relevant prior experiences, enabling the model to ground its updates in both current observations and analogous past scenarios, as shown in Figure 6.

By utilizing semantic vector store, we maintain a long-term memory $\mathcal{M}$ which is a set of stored episodes, each indexed by a dense semantic embedding, as well as associated metadata, such as a brief context summary, the final correct mental state, and an abstract reflection. When a new reasoning task defined by the current scenario context, ongoing actions and utterances and the subject of inference is given, we construct one or more textual queries capturing the essence of the task. For example, if the subject of inference is "each character's belief about the location of the object" and the current context involves certain actions, the model might form a query combining keywords from the context including agents' names, object names, key events

along with phrases indicating "belief" or "thought about [object]". We denote the set of query variants as $Q = \{q_1, q_2, \dots, q_m\}$, designed to cover different lexical formulations of the information need.

4.3.1 Hybrid Retrieval with Dense and Lexical Approaches

For each query $q \in Q$, we perform two kinds of search over $\mathcal{M}$:

- **Dense Vector Similarity Search**: the query q is embedded into the same vector space as the memory entries, using the embedding model of the vector store, and we find the entries with highest cosine similarity to q.
- **Lexical Search**: we use a BM25-based search over the textual metadata of entries to find those with common keywords. Let $\mathrm{TopK}_{\mathrm{dense}}(q)$ be the top K retrieved items by semantic similarity, and $\mathrm{TopK}_{\mathrm{lex}}(q)$ the top K items by lexical matching.

We take the union of results from all query variants and both methods, then eliminate duplicates, yielding a candidate set D_{cand} of potentially relevant memory entries:

$$D_{\mathrm{cand}} = \bigcup_{i=1}^{m} \left(\mathrm{TopK}_{\mathrm{dense}}(q_i) \cup \mathrm{TopK}_{\mathrm{lex}}(q_i)\right). \tag{5}$$

Each candidate entry $d \in D_{\mathrm{cand}}$ comes with a stored content, e.g. a short description of a scenario or a distilled outcome, and metadata. At this stage, $\mathrm{D}_{\mathrm{cand}}$ might still be large and contain some less relevant items due to noise from broad semantic matches or keyword overlaps. We therefore apply a reranking and filtering process to select the most relevant and diverse examples to actually use in reasoning.

4.3.2 Relevance Re-Ranking and Maximal Marginal Relevance Selection

For relevance re-ranking, we first score each candidate $\mathrm{d} \in \mathrm{D}_{\mathrm{cand}}$ for relevance to the current query using a cross-encoder model F_{CE}. This model takes the pair $(q, \mathrm{content}_d)$ and produces a relevance score $R(q, d)$. Compared to the initial vector similarity, $R(q, d)$ is a more precise semantic relevance measure (e.g., capturing whether the memory entry truly addresses a similar question or situation as our current task). We compute $R(q, d)$ for all candidates, then sort $\mathrm{D}_{\mathrm{cand}}$ by this score in descending order.

We also implement Maximal Marginal Relevance (MMR) selection. Rather than simply picking the top K highest $\mathrm{R}(\mathrm{q}, \mathrm{d})$ entries, which could be redundant, e.g. multiple entries that are near duplicates of each other, we use a Maximal Marginal Relevance strategy to ensure diversity among the retrieved examples. We will select a final subset $\mathrm{E} = \{\mathrm{d}_1, \mathrm{d}_2, \dots, \mathrm{d}_{\mathrm{K}}\}$ of K examples iteratively. Initialize an empty set $S = \emptyset$. At each selection step, choose the candidate $\mathrm{d} \in \mathrm{D}_{\mathrm{cand}} \setminus \mathrm{S}$ that maximizes a trade-off between relevance and dissimilarity to already-selected items. The MMR objective for selecting the next example d^* can be written as:

$$d^* = \arg\max_{d \in D_{\mathrm{cand}} \setminus S} \left(\lambda R(q, d) - (1 - \lambda) \max_{d' \in S} \mathrm{sim}\,(d, d')\right). \tag{6}$$

Here $\mathrm{sim}(\mathrm{d}, \mathrm{d}')$ is the cosine similarity between the embedding vectors of two candidate entries, and $0 \leq \lambda \leq 1$ is a parameter, e.g. $\lambda = 0.7$, controlling the relevance-vs-diversity balance. Intuitively, this criterion prefers entries that have high relevance $R(q, d)$ to the query while penalizing those that are overly similar to any already chosen entry to promote informational diversity. We add d^* to S and repeat until K entries are selected (or until $\mathrm{D}_{\mathrm{cand}}$ is exhausted). The outcome is a set $\mathrm{E} = \mathrm{d}_1, \dots, \mathrm{d}_{\mathrm{K}}$ of the most pertinent and diverse examples from memory for the task at hand.

4.3.3 Augmenting Reasoning with Retrieved Examples

The selected examples E are then incorporated into the LLM's prompt to guide hypothesis generation and rejuvenation. Each example $\mathrm{d} \in \mathrm{E}$ is represented in a concise textual form, typically including a brief description of the prior scenario's context and the relevant outcome, for example, what the true state was,

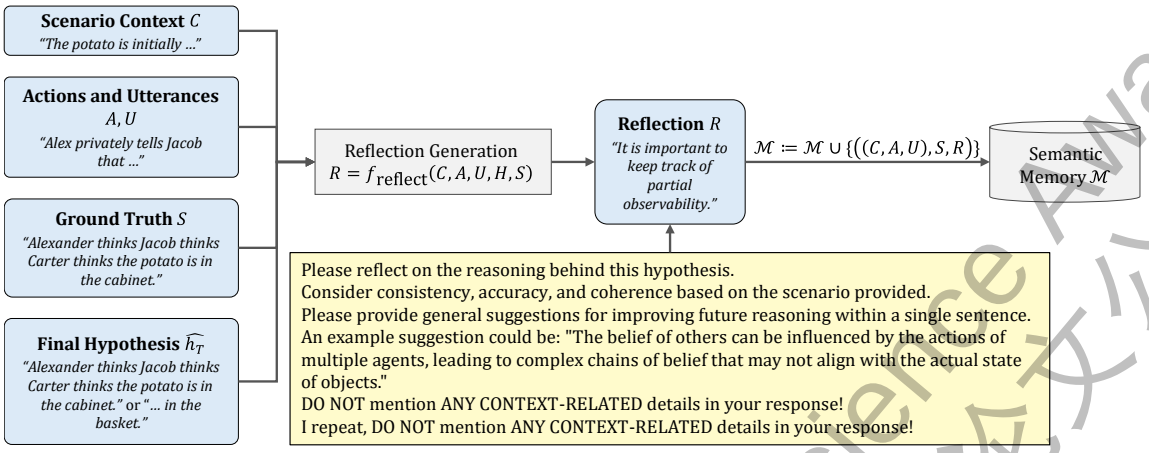

Figure 7. Reflective Memory Management

along with any stored reflection or lesson from that scenario as will be detailed in the next subsection. An example might be presented as: "Example 1: In the context of '<scenario summary>', <outcome>. Here is a piece of advice: <reflection>." Listing a few such examples before the current scenario prompt provides the model with analogical cases. This retrieval-augmented prompt acts as a set of implicit constraints or hints, helping the model to avoid pitfalls and leverage patterns learned from prior similar situations.

For instance, if the memory contains a scenario where an agent was misled by another agent's action to a false belief, and our current task has a structurally similar deception, a retrieved example might remind the model that "the belief of others can be influenced by the actions of multiple agents, leading to complex chains of belief that may not align with reality." By seeing this hint, the model is more likely to correctly hypothesize the presence of a false belief in the current scenario.

Formally, the hypothesis update function from the SMC step is now augmented with memory E: instead of relying solely on O_t, the update uses the combined information (O_t, E) when revising and re-ranking the particles. This can be seen as conditioning the proposal distribution for new hypotheses on both the latest observation and the retrieved knowledge. The result is a semantics-augmented SMC update that is better informed and less prone to prior mistakes, since the model can "remember" how similar belief reasoning problems were resolved in the past.

4.4 Reflective Memory Management

A reflection mechanism is designed to allow the model to learn from each completed inference and accumulate abstract knowledge for future use, as illustrated in Figure 7. The SMC updates and SRAR have led to a final set of hypotheses, and typically a final answer to the query, such as the most likely mental state for each agent. After that, the model performs a post-hoc reflection on its reasoning process. This component generates an abstract feedback statement for each hypothesis–answer pair and stores this information into the semantic vector store as metadata.

4.4.1 Generating Abstract Feedback

The reflection process takes as input the scenario context C, the sequence of actions A and utterances U that occurred, the final hypothesis H produced by the model for the query, e.g. the model's inferred answer about an agent's belief, and the correct answer S, the ground-truth mental state or outcome, if known from the simulation or annotation. Using these, we prompt the LLM to reflect on the reasoning behind the hypothesis H, focusing on aspects like consistency, accuracy, and coherence with the scenario.

Crucially, the prompt explicitly instructs the model not to mention any scenario-specific details in its reflection. Instead, it must distill a general insight or suggestion for improving future reasoning in similar tasks, phrased in a single sentence. For example, if the model's hypothesis was incorrect because it failed to

consider that an agent could be misinformed by another, the reflection might be a general statement like: "An agent's belief can be incorrect if it is based on another agent's misleading action, so future reasoning should consider the possibility of deception." This sentence does not reference any character or object by name; it is an abstract lesson extracted from the particular scenario.

Formally, we can denote the reflection generation as a function f_{reflect} applied to the scenario and outcome:

$$R = f_{\text{reflect}}(C, A, U, H, S), \tag{8}$$

where R is the resulting reflection sentence. The model f_{reflect} is the same LLM but used in a special "reflection mode" with a crafted prompt as described. Because H and S are given as part of the input, the model can analyze the relationship between its hypothesis and the true state, effectively performing an error analysis or self-evaluation. The output R is a high-level critique or advice that could help avoid the mistake or reinforce the correct reasoning in the future. Notably, we design this reflection to be agnostic to the concrete story details so that it may apply broadly: it functions as a piece of transferable knowledge.

4.4.2 Persisting Reflection as Metadata

Once the reflection R is generated, the model stores a new entry into the semantic memory $\mathcal{M}$ (the vector store). This entry includes the key aspects of the just-completed scenario and the reflection. We represent an entry as a tuple $\big((C, A, U),\ S,\ R\big)$, consisting of a representation of the context and events (C, A, U), the final correct state S, and the reflection sentence R. In practice, the context (C, A, U) may be summarized or embedded rather than stored verbatim if it's long, but the entry retains enough information to be searchable via semantic vectors and keywords later. The vector store's metadata fields for this entry record the type of inference, e.g. whether S was a "belief" state, an "intention," etc., corresponding to the subject of inference, as well as the text of R. We denote the storage operation as:

$$\mathcal{M} \coloneqq \mathcal{M} \cup \{\big((C, A, U), S, R\big)\}. \tag{9}$$

By adding this entry, the model learns from its experience. In subsequent reasoning episodes, the retrieval module can search this memory and potentially retrieve the newly stored example if the future query is semantically similar. The reflection R then serves as a piece of meta-knowledge or advice, attached to a scenario reminiscent of the new one. Over time, as more scenarios are processed and more reflections stored, the model's semantic memory grows into a repository of distilled Theory-of-Mind reasoning insights. This contributes to improved performance: the model increasingly benefits from past lessons, thus gradually mitigating repeated errors.

The Reflection mechanism essentially implements a form of iterative self-improvement: each run of the ToM reasoning loop not only produces an answer but also a training signal, the reflection, for the next loops. Because the reflections are abstract and generalized, the model avoids overfitting to specific past contexts and instead accumulates broadly applicable wisdom, e.g. recognizing common pitfalls like ignoring an agent's false belief or misinterpreting an ambiguous action.

5 EXPERIMENTS

5.1 Experiment Settings

We evaluated the proposed method on the generated multi-modal dataset with totally 300 question–answering reasoning asks spanning zeroth-, first-, second-, third-, and fourth-order, with 60 tasks per order. Each task features a short photorealistic video depicting multi-agent interactions with partial observability, accompanied by a multiple-choice question requiring nested belief inference. The overall experimental goal is to assess both the scalability and the structural robustness of our model across increasing orders of recursive reasoning.

We employed Gemini 2.5 Pro as the vision–language model for extracting structured textual representations from videos, and GPT-5 and Qwen-3 Max as the core reasoning backend for all baselines and our proposed method. For the comparison experiments, we included GPT-5, GPT-5 Thinking, Qwen-3 Max, and the multi-modal benchmark, MuMA-ToM and AutoToM, tested in both GPT-5 and Qwen-3 Max configurations. Among

them, Qwen-3 Max and GPT-5 Thinking represent current frontier models for integrated reasoning, while MuMA-ToM and AutoToM serves as the SOTA methods of publicly available structured multi-modal baseline.

The ablation study is also performed with the components of SRAR, RMM and SMC removed incrementally. This setup isolates the contribution of each structural component to overall reasoning depth and generalization and evaluate the effectiveness of each one.

All methods received the same scene inputs, dialogue transcripts, and question formats, ensuring comparability. Each model was allowed to output only a single answer per question without iterative self-correction. Accuracy was computed as the percentage of correctly answered questions out of 60 for each order. We report both quantitative comparisons and ablation results to reveal how sequential belief propagation, retrieval augmentation, and reflective memory jointly enhance performance across reasoning orders.

5.2 Experiment Results

5.2.1 Quantitative Results

The quantitative comparison across reasoning orders is summarized in Table 1. The benchmark contains 60 tasks per order, and the results demonstrate clear and consistent improvements of our structured model over all baselines. Specifically, our full model with both GPT-5 and Qwen-3 Max achieves the highest accuracies on zeroth-, first-, second-, third-, and fourth-order reasoning respectively, representing a substantial advantage as the reasoning order increases. Especially, the results for third- and fourth-order reasoning tasks explicitly proved the overwhelming advantage against other methods.

For the simplest tasks with only zeroth- and first order, there is no significant different among all the tested methods, though our method achieves the highest accuracy. Even GPT-5 Thinking could achieve the similar accuracy with our method, marginally outperforms the others. From second-order tasks, the difference in results between our method and other methods become significant. For third-order task, MuMA-ToM could not support perform reasoning, and AutoToM achieves 0% without any right answer. Our method achieves the highest score, 51.67%, significantly outperforming the second one, GPT-5 Thinking with only 18.3%. For the most complex tasks with fourth-order, our model could achieve the accuracy of 21.67%, while the other methods could rarely generate the right answer, with only Qwen 3-Max and GPT-5 Thinking occasionally correctly answer 1-2 out of 60 questions.

Overall, our approach surpasses the best baseline by a wide margin—especially at third- and fourth-order levels, where all other models struggle to produce any correct answers. This evidences the necessity of explicit structured reasoning for multi-agent, multi-modal Theory-of-Mind tasks. The advantage of our method grows

Table 1 Results of Comparison Experiments.

Base Model	Method	Zeroth-Order	First-Order	Second-Order	Third-Order	Fourth-Order
Qwen-3 Max	Qwen-3 Max	98.33	90.00	75.00	11.67	1.67
	MuMA-ToM	96.67	91.67	80.00	Not supported	Not supported
	AutoToM	96.67	91.67	80.00	0.00	0.00
	Ours	**100.00**	**98.33**	**83.33**	**51.67**	**21.67**
GPT-5	GPT-5	**100.00**	95.00	78.33	18.33	0.00
	GPT-5 Thinking	**100.00**	**98.33**	78.33	18.33	3.33
	MuMA-ToM	96.67	93.33	83.33	Not supported	Not supported
	AutoToM	96.67	93.33	83.33	0.00	0.00
	Ours	**100.00**	**98.33**	**91.67**	**46.67**	**21.67**

rapidly with increasing order, validating its capability to manage deeply nested belief structures. Particularly at third- and fourth-order reasoning, the performance gap reaches over 33.4% comparing to GPT-5 Thinking, emphasizing the contribution of explicit belief propagation and retrieval-based augmentation.

To further validate the effects of each component, we conducted an ablation study disabled SRAR, RMM and SMC, with the results presented in Table 2. Disabling RMM has impact to the results of third-order and fourth-order tasks with 1.67%-3.34% right answers reduced. Disabling both SRAR and RMM has impact to the results of all the higher-order tasks with 8.34%-11.67% right answers reduced. Our method with only SMC component caused difference of all tasks with different orders. This demonstrates that all the 3 components play essential roles for high order reasoning. SRAR provides structured generalization through example grounding, and RMM consolidates the reasoning experience into reusable abstract insights. Their combination enables consistent reasoning depth and adaptability to unseen, asymmetric belief hierarchies.

Table 2 Results of Ablation Study.

Method	Zeroth-Order	First-Order	Second-Order	Third-Order	Fourth-Order
Ours	100.00	98.33	91.67	46.67	21.67
Ours w/o RMM	100.00	98.33	93.33	45.00	18.33
Ours w/o RMM and SRAR	100.00	98.33	81.67	35.00	13.33
GPT-5	100.00	95.00	78.33	18.33	0.00

5.2.2 Subjective Evaluation

Qualitative assessment of the model's responses further supports these quantitative findings. Compared with GPT-5 and GPT-5 Thinking, our model consistently provides shorter and more coherent justifications aligned with agents' perspectives. It avoids over-explanation and maintains internal consistency across nested beliefs. In high-order reasoning cases involving deception or missed observations, the structured belief propagation prevents logical leakage and ensures temporally consistent reasoning. Overall, the updated results affirm that explicitly structured belief modeling, retrieval-based grounding, and reflective meta-learning jointly yield stable and scalable higher-order ToM reasoning that baseline LLMs cannot achieve.

6 CONCLUSION

This work presents a structured higher-order mental state inference method that integrates a qualitative Sequential Monte Carlo mechanism, semantic retrieval augmented reasoning, and a reflective memory management component. Together with a photorealistic multi-agent dataset designed to induce knowledge asymmetries through partial observability and public versus private communication, the approach delivers a coherent pipeline for multi-modal ToM reasoning.

Empirically, the method substantially outperforms recent baselines on tasks that require recursive belief attribution. Ablation studies show that retrieval and reflection contribute materially to deeper recursion. Qualitative inspection further indicates that perspective-safe updates prevent omniscience leaks and help maintain temporal and cross-level consistency, enabling concise, defensible answers rather than brittle chains of reasoning.

These results support two conclusions. First, explicit structure is crucial: representing nested beliefs within agent-indexed state and updating them under visibility constraints is more reliable than treating ToM as unstructured text generation. Second, learned memory is beneficial: example-guided retrieval and distilled reflections reduce recurrent failure modes in multi-agent settings, especially when deception, missed observations, or asynchronous disclosures create belief divergence.

In summary, this study advances machine ToM by coupling principled structure with experience-driven augmentation. By making higher-order belief tracking more accurate, consistent, and sample-efficient, it

moves AI closer to robust social understanding in complex, real-time environments, with implications for collaborative robotics, education, healthcare, and assistive technologies.

REFERENCES

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970. https://doi.org/10.1037/a0016923

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief–desire attribution. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2469–2474).

Bolander, T., & Andersen, M. B. (2011). Epistemic planning for single- and multi-agent systems. *Journal of Applied Non-Classical Logics, 21*(1), 9–34. https://doi.org/10.3166/jancl.21.9-34

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv.* https://arxiv.org/abs/2312.10997

Gmytrasiewicz, P. J., & Doshi, P. (2005). A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research, 24*, 49–79. https://doi.org/10.1613/jair.1600

He, Y., Wu, Y., Jia, Y., Mihalcea, R., Chen, Y., & Deng, N. (2023). HI-TOM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv.* https://arxiv.org/abs/2310.16755

Jain, U., Zhang, Z., Schwing, A. G., & Forsyth, D. (2019). Reasoning about object affordances in visual scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12354–12362).

Jin, C., Wu, Y., Cao, J., Xiang, J., Kuo, Y.-L., Hu, Z., Ullman, T., Torralba, A., Tenenbaum, J. B., & Shu, T. (2024). MMTOM-QA: Multi-modal theory of mind question answering. *arXiv.* https://arxiv.org/abs/2401.08743

Kinderman, P., Dunbar, R. I. M., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology, 89*(2), 191–204. https://doi.org/10.1111/j.2044-8295.1998.tb02677.x

Kosinski, M. (2023). Theory of mind emerged in large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2218523120. https://doi.org/10.1073/pnas.2218523120

Li, C., Xia, F., Martín-Martín, R., Lingelbach, M., Srivastava, S., Shen, B., Vainio, K., Gokmen, C., Dharan, G., Jain, T., Kurenkov, A., Liu, C. K., Gweon, H., Wu, J., Li, F.-F., & Fei-Fei, L. (2021). iGibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *Proceedings of the 5th Conference on Robot Learning* (pp. 1302–1319). https://proceedings.mlr.press/v164/li22a.html

Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychological Bulletin, 135*(5), 749–773. https://doi.org/10.1037/a0016854

Moore, C., Pure, K., & Furrow, D. (1990). Children's understanding of the speaker–listener relationship. *Child Development, 61*(3), 722–730. https://doi.org/10.2307/1130950

Osterhaus, C., & Koerber, S. (2021). A longitudinal analysis of the theory of mind development in middle childhood. *Child Development, 92*(1), 165–179. https://doi.org/10.1111/cdev.13469

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that…" Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology, 39*(3), 437–471. https://doi.org/10.1016/0022-0965(85)90051-7

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., & Torralba, A. (2018). VirtualHome: Simulating household activities via programs. *arXiv.* https://arxiv.org/abs/1806.07011

Rabinowitz, N. C., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 4218–4227). PMLR. https://proceedings.mlr.press/v80/rabinowitz18a.html

Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., & Choi, Y. (2019). Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 4463–4473). https://doi.org/10.18653/v1/D19-1454

Schank, R. C., & Abelson, R. P. (1997). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Shi, H., Ye, S., Fang, X., Jin, C., Isik, L., Kuo, Y.-L., & Shu, T. (2025). MuMA-ToM: Multi-modal multi-agent theory of mind. *arXiv.* https://arxiv.org/abs/2408.12574

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *arXiv.* https://arxiv.org/abs/2303.11366

Stuhlmüller, A., & Goodman, N. D. (2014). Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research, 28*, 80–99. https://doi.org/10.1016/j.cogsys.2013.05.001

Ullman, T. D., Spelke, E., Tenenbaum, J. B., & Gerstenberg, T. (2023). The emergence of intuitive psychology in large language models. *arXiv.* https://arxiv.org/abs/2305.16474

Zhang, H., Zhang, X., & Zhu, Y. (2023). CLEVR-Mental-State: Diagnosing theory of mind capabilities in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, Z., Jin, C., Jia, M., & Shu, T. (2024). AutoToM: Automated Bayesian inverse planning and model discovery for open-ended theory of mind. *arXiv.* https://arxiv.org/abs/2403.05530

Zhong, F., Wu, K., Wang, C., Chen, H., Ci, H., Li, Z., & Wang, Y. (2024). UnrealZoo: Enriching photorealistic virtual worlds for embodied AI. *arXiv.* https://arxiv.org/abs/2404.10660

ACKNOWLEGEMENT

The project is a culmination of the collaborative efforts between the student and the supervisors, and I would like to express my deepest gratitude and thankfulness for the invaluable guidance and support I received throughout the process. First and foremost, I would like to deliver my sincere thanks to the supervisors, Dr. Fangwei Zhong and Dr. Yuchen Zhou. Their expertise and encouragement played a pivotal role in helping us navigate the complexities of this research.

Source of the selected topic and research background

This topic of this project stems from our combined interests, a current hotspot of artificial intelligence research, and my continuous exploration in the field of machine Theory of Mind and cognitive psychology. It is further development and extension of our previous attempts in multi-agent learning and visual reasoning. We sought to contribute to this field by delving deeper into structured approaches for higher-order mental state inference.

The relationship between the supervisors and the student, the role supervisors played in the process of writing thesis, and whether the tutoring is paid

Dr. Fangwei Zhong is the off-campus supervisor of the Branch AI Laboratory of Center on Frontiers of Computing Studies Peking University, Talent Institute, Beijing No. 101 High School. Dr. Yuchen Zhou is the supervisor of the Branch AI Laboratory of Center on Frontiers of Computing Studies Peking University, Talent Institute, Beijing No. 101 High School. I'm a project team member of the laboratory. Both Dr. Zhong and Dr. Zhou provided their guidance voluntarily without compensation, driven solely by their passion for research and mentorship.

Research completed with the assistance of others

The project was completed independently under the supervision of Dr. Zhong and Dr. Zhou, with no external assistance apart from their mentorship. We are grateful to both for their patience and dedication.

Personal Profile

Jiran Mo, International Accelerated Class, Grade 11, Beijing No. 101 High School.

- **Math**
  - Earned AMC Distinction
  - Scored 5 on AP Calculus AB and AP Calculus BC
- **English**
  - Scored 114 and 117 on the TOEFL in Grades 9 and 10, respectively
- **Psychology**
  - Read major works such as *Thinking, Fast and Slow* and *The Anxious Generation*
  - Passionate about social psychology
- **AI**:
  - Familiar with PyTorch, CNN, RNN, ResNet and GNN, etc.
  - Attended the 2024 Large Language Model Elite Training Camp jointly organized by Peking University and Tencent, gained a systematic understanding of LLMs and hands-on experience in LoRA fine-tuning
  - Developed an ultrasonic algae removal robot equipped with a GNN-based autonomous path planning system
  - Created a UE5-Based 3D Shooting Game
- **Awards**
  - AMC Distinction
  - Conrad Challenge China 2025 Best Pitch Award

Supervisor Profile

Fangwei Zhong: Fangwei Zhong: Ph.D., an Associate Professor at the School of Artificial Intelligence, Beijing Normal University. Before that, he was a Boya Postdoctoral Researcher at Peking University and received Ph.D in Computer Science from EECS, Peking University, supervised by Prof. Yizhou Wang. His

current research interests are autonomous learning, multi-agent learning, and computer vision, particularly in building embodied agents with physical and social common sense.

Yuchen Zhou: Ph.D., Senior Researcher, supervisor of the Branch AI Laboratory of Center on Frontiers of Computing Studies Peking University, Talent Institute, Beijing No. 101 High School. He is a senior member of the ACM and IEEE, and former member of Technical Committee, Embedded System Society, China Computer Federation. With 20 years of technical innovation experience in IBM, he served as senior research manager of AI Perception in IBM Research - China, a member of IBM Academy of Science and Technology, IBM Master Inventor, the chair of technical committee and patent review committee of the center, etc. He won 3 outstanding technical achievement awards, published 1 book, contributed to 2 international standards, obtained around 50 international patents and published more than 30 papers.