MindZero: LEARNING ONLINE MENTAL REASONING WITH ZERO ANNOTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective real-world assistance requires AI agents with robust Theory of Mind (ToM): inferring human mental states from their behavior. Despite recent advances, several key challenges remain, including (1) online inference with robust uncertainty updates over multiple hypotheses; (2) efficient reasoning suitable for real-time assistance; and (3) the lack of ground-truth mental state annotations in real-world domains. We address these challenges by introducing *MindZero*, a self-supervised reinforcement learning framework that trains multimodal language models to perform efficient and robust online mental reasoning. During training, the model is rewarded for generating mental state hypotheses that maximize the likelihood of observed actions estimated by a planner, similar to model-based ToM reasoning. This method thus eliminates the need for explicit mental state annotations. After training, *MindZero* internalizes model-based reasoning, and performs mental inference in a single forward pass at test time. We evaluate *MindZero* across challenging mental reasoning and AI assistance tasks in gridworld and household domains. *MindZero* matches the robustness of explicit model-based methods while significantly accelerating inference, outperforming state-of-the-art methods by a large margin. These results demonstrate that mental reasoning can be learned as a self-supervised skill, bridging the gap between robustness and efficiency in ToM modeling.

1 INTRODUCTION

To proactively assist human users in the real world, AI agents must understand users' minds and anticipate their needs. This requires strong Theory of Mind (ToM), i.e., the ability to infer users' mental states (such as desires, beliefs, and goals) from their behavior. Recent advances in large language models (LLMs) have sparked growing interest in machine Theory of Mind (Wimmer & Perner, 1983; Ullman, 2023; Wilf et al., 2024; Sclar et al., 2023; Jin et al., 2024). However, much of the existing work focuses on question-answering-based ToM evaluation and development, which is insufficient for real-world assistance. In practice, an assistive agent must continuously update its inferences about a user's mental state and track uncertainty over multiple competing hypotheses. This form of online mental-state reasoning can guide agent planning, enabling proactive assistance, adaptation to changing contexts, and more effective collaboration with users.

For instance, in Figure 1, as the agent observes a human's actions in a household setting, it maintains and updates a probability distribution over multiple possible goal hypotheses in real time, and plans assistance under uncertainty (e.g., fetching tableware before the user asks) Puig et al. (2023).

However, training models for online mental reasoning remains challenging. Human mental states are latent and often ambiguous. They are also dynamically changing over time in sequential tasks. For many real-world applications, such as household or web assistance, it is extremely difficult and costly to collect large-scale training data with reliable annotations of ground-truth mental states. As a result, prior works on learning-based ToM methods have been limited to controlled settings (Rabinowitz et al., 2018; Rhinehart et al., 2019; Bortoletto et al., 2024a;b), lacking open-endedness and scalability.

To circumvent these data and annotation challenges, recent work has explored inference-time reasoning methods that leverage the generality and strong reasoning ability of LLMs for ToM, without requiring model training. In particular, when integrated with model-based ToM methods, such as

Bayesian inverse planning (BIP), inference-time scaling has demonstrated strong performance on challenging ToM reasoning tasks Jin et al. (2024); Shi et al. (2025); Zhang et al. (2025); Ying et al. (2023); Kim et al. (2025). These methods leverage LLMs to propose and evaluate mental state hypotheses, achieving robust and scalable mental reasoning. However, they are computationally prohibitive in online mental reasoning required for real-world assistance tasks. These challenges call for a new type of ToM approach that retains the deliberative structure of model-based reasoning while better leveraging the efficiency and learning capacity of LLMs.

To address these limitations, we introduce *MindZero*, a novel Theory of Mind reasoning framework that trains multimodal language models to perform robust and efficient online mental reasoning without requiring mental state annotations. During training, the model explicitly generates hypotheses about mental states (e.g., beliefs and goals) and is rewarded when these hypotheses assign high likelihood to the actions people actually take. We term this Self-Supervised Reinforcement Learning (SSRL). Unlike common RL-based language model training, the reward in our SSRL method is entirely calculated via self-supervised signals; it also encourages the model to produce explicit mental state hypotheses with robust uncertainty estimates. This method eliminates the need for ground-truth mental state labels, allowing the model to learn directly from behavior and internalize ToM reasoning patterns that explain actions in context. The trained *MindZero* model infers mental states in a single forward pass, while remaining grounded in a model-based objective that preserves robustness and interpretability.

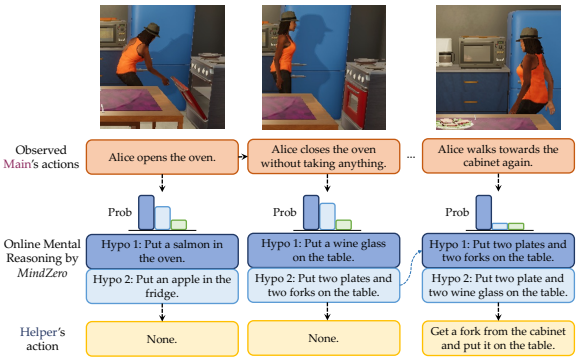


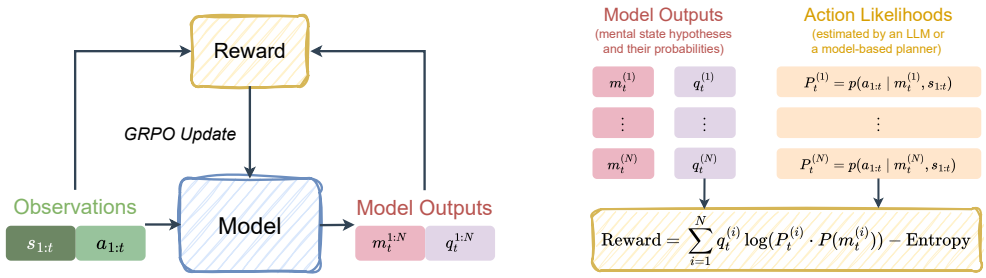
Figure 1: An example of online mental reasoning for proactive assistance. In our Household Proactive Assistance setting, as the helper observes Alice's actions over time, *MindZero* continuously updates a probability distribution over multiple goal hypotheses. Based on the multiple possible hypotheses maintained at each step, the helper decides whether to act and proactively assists by fetching relevant tableware and placing it on the table.

In our experiments, we compared *MindZero* against state-of-the-art ToM methods on question answering and proactive assistance tasks in both gridworld Jha et al. (2024) and household environments Puig et al. (2023). Small multimodal language models trained with our *MindZero* method significantly outperformed baselines in all tasks, matching the robustness of model-based methods while significantly reducing the computational cost. These results suggest that mental reasoning can be learned as a self-supervised skill, narrowing the gap between robust but slow model-based inference and fast but error-prone reasoning by a small multimodal language model.

In sum, our main contributions include: (1) a self-supervised RL method, *MindZero*, that trains multimodal language models to conduct robust and efficient online mental reasoning without mental state annotations; (2) systematic evaluation of *MindZero* and recent ToM methods in a suite of challenging online mental reasoning and proactive AI assistance benchmarks.

2 Related Work

Theory of Mind Methods. Existing methods for ToM reasoning fall into three main categories. (1) *Prompting-based* approaches (Wilf et al., 2024; Jung et al., 2024; Huang et al., 2024; Hou et al., 2024; Sclar et al., 2023) improve upon base LLMs but still exhibit systematic errors in long-context understanding, complex behaviors, and recursive reasoning. (2) *Model-based* approaches, especially Bayesian inverse planning (BIP) (Baker et al., 2009; Ullman et al., 2009), explicitly model agents' mental states and their influence on behavior. Recent work integrates BIP with LLMs (Jin et al., 2024; Shi et al., 2025; Zhang et al., 2025), combining structured reasoning with flexible language understanding. However, these methods are often computationally expensive, as they require searching large hypothesis spaces at test time. (3) *Learning-based* methods train neural networks for mental-state inference (Rabinowitz et al., 2018; Liang et al., 2024; Sclar et al., 2024; Lu et al., 2025), but they rely on costly and unreliable ground-truth annotations, limiting their scalability and



(a) Self-Supervised Reinforcement Learning. (b) Reward Computation.

Figure 2: (a) Overview of our Self-Supervised Reinforcement Learning (SSRL) framework. Given observations of states $s_{1:t}$ and actions $a_{1:t}$ up to timestep t, the model outputs a set of N mental state hypotheses $m_t^{1:N}$ along with their probabilities $q_t^{1:N}$. Unlike standard RL-based language model training, SSRL derives rewards entirely from self-supervised signals based on observations and model outputs, which are used to guide GRPO updates. (b) Reward computation in SSRL. Given the model outputs, an action likelihood evaluator (either an LLM or a model-based planner) estimates the likelihood of the observed action under each mental state hypothesis. The reward is computed as the probability-weighted log-likelihood of the observed action minus an entropy regularization term.

applicability. To address these limitations, *MindZero* learns mental reasoning directly from human behavior data. Our approach improves over prompting-based methods, avoids the computational overhead of model-based inference, and eliminates the need for explicit mental state annotations required by prior learning-based approaches.

ToM-Guided Assistance. Recent ToM methods and evaluations mainly focus on question-answering settings (Le et al., 2019; Gandhi et al., 2023; Kim et al., 2023; Wu et al., 2023; Xu et al., 2024; Jin et al., 2024; Shi et al., 2025; Fan et al., 2025), where the ToM model reads a story and answers questions about mental states. In contrast, ToM-guided assistance is more challenging: it requires agents to continuously infer and update mental states, often under uncertainty and across long-horizon interactions. For example, Online Watch-And-Help (O-WAH) (Puig et al., 2021; 2023) tasks a helper agent with observing another agent's actions, inferring its underlying goal, and assisting it to reach the goal more efficiently in realistic household environments. MindCraft (Bara et al., 2021) introduces a Minecraft-based collaborative task with asymmetric knowledge and skills, and periodically collects players' self-reported beliefs about the world and each other, enabling in-situ modeling of belief evolution and common ground in situated dialogue. ToM-SSI (Bortoletto et al., 2025) further expands situated evaluation to richer social dynamics by introducing a multi-modal benchmark with up to four interacting agents and tasks spanning cooperative, obstructive, and mixed interactions, requiring perspective-taking and joint reasoning over percepts, beliefs, and intentions. Taken together, these settings emphasize that assistance requires online mental-state tracking under uncertainty from multimodal interaction traces. ToM-SWE (Zhou et al., 2025) explores how ToM may help challenging open-ended coding tasks. While there has been prior work on online mental reasoning shown to be effective in ToM-guided assistance (e.g., Puig et al., 2023; Wang et al., 2021; Shvo et al., 2022; Zhi-Xuan et al., 2024; Ying et al., 2025; Zhang et al., 2025), they have strong assumptions about human behavior and/or require high computational costs for complex tasks. *MindZero* directly targets this gap by training a small multimodal language model to efficiently and robustly conduct online mental reasoning that can support downstream assistance tasks in a scalable way.

3 Problem Formulation

We formalize the problem of online mental state inference (Section 3.1) and characterize how inferred mental states can be leveraged to enable proactive assistance (Section 3.2). Our formulation provides a unified probabilistic framework for reasoning about users' latent beliefs and goals from sequential observations, and for translating this uncertainty-aware reasoning into effective assistive decision making in dynamic environments.

3.1 Online Mental Reasoning

Given a sequence of observed user behavior up to time step t, including states $s_{1:t}$ and actions $a_{1:t}$, a ToM model infers the latest mental state of the user m_t, which could include different mental variables such as beliefs b_t and goals g_t. Inspired by Bayesian inverse planning (BIP) Baker et al. (2009; 2017); Zhi-Xuan et al. (2020), a model-based ToM inference method, we formalize online mental state inference as following Bayesian inference:

$$\underbrace{P(m_t \mid s_{1:t}, a_{1:t})}_{\text{posterior}} \propto \underbrace{P(a_{1:t} \mid m_t, s_{1:t})}_{\text{action likelihood}} \cdot \underbrace{P(m_t)}_{\text{prior}}, \tag{1}$$

Unlike prior work Zhi-Xuan et al. (2020), this formulation goes beyond the typical Markovian assumptions behind BIP, modeling all past behavior jointly. In real-world domains, this Bayesian inference can be computationally intractable due to an infinite hypothesis space and costly action likelihood estimation (which is achieved via forward planning conditioned on hypothetical mental states). Our *MindZero* method aims to overcome these computational bottlenecks by training a multimodal language model to directly output quality hypothesis samples and their posterior probabilities without explicit Bayesian inference.

3.2 Proactive Assistance Guided by Online Mental Reasoning

In online mental reasoning, the model must continuously update multiple mental state hypotheses $\{m_t\}$ at every step t and estimate their probabilities $\{q_t\}$ given a user's behavior history $(s_{1:t}, a_{1:t})$. Given the top hypotheses of a user's mental state, an assistive agent can then plan for the assistive actions to best help the user. Let a_t^A be the assistive action at time step t. We define the assistive agent's policy as

$$P(a_t^A \mid s_{1:t}, a_{1:t}) = \sum_{m_t} P(a_t^A \mid s_t, m_t) P(m_t | s_{1:t}, a_{1:t}). \tag{2}$$

Such assistive decision making must consider the uncertainty in the mental inference, which requires a robust estimate of the confidence of multiple hypotheses. It also needs to frequently update plans based on the most recent user behavior, and thus needs a fast inference to support real-time replanning. *MindZero* aims to achieve this via training a small multimodal language model with low computational cost and latency.

4 *MindZero*

We introduce *MindZero*, a self-supervised reinforcement learning framework that trains multimodal language models to perform efficient and robust online mental reasoning. *MindZero* learns directly from behavioral data using self-supervised signals, addressing the lack of ground-truth mental state labels in real-world domains (Section 4.1 and Figure 2a). The core of *MindZero* is its reward design: the model is rewarded for generating mental state hypotheses that maximize the likelihood of observed actions, as estimated by a planner, in a manner similar to model-based ToM reasoning (Section 4.2 and Figure 2b). Through this process, *MindZero* internalizes the Bayesian inverse planning procedure in Equation (1) and enables real-time planning for proactive assistance as formulated in Equation (2).

4.1 Self-Supervised RL for Mental Reasoning

Standard supervised approaches to mental reasoning rely on ground-truth mental state annotations, which are scarce and difficult to collect. Existing self-supervised methods for sequential modeling, such as next-token prediction (Bengio et al., 2003; Radford et al., 2018) and return-conditioned behavioral cloning (Chen et al., 2021), emphasize forward prediction and learn by mimicking future words or actions from past context. In contrast, mental reasoning requires inverse modeling: explicitly inferring the mental state that causes the observed behavior. This capability is not explicitly learned by existing self-supervised objectives, which are optimized for prediction rather than explanation.

To bridge this gap, we formulate mental reasoning as a self-supervised reinforcement learning (SSRL) problem centered on explanatory consistency. Instead of treating actions as prediction

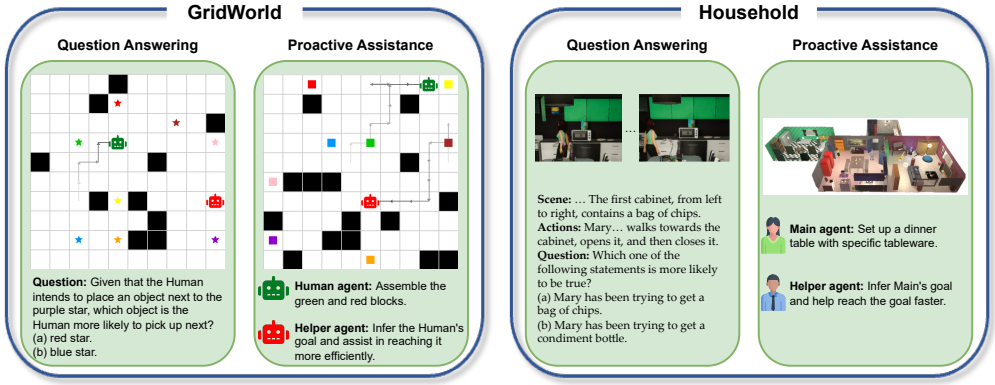


Figure 3: Our experimental settings for mental state reasoning and proactive assistance: (1) GridWorld Question Answering (Section 5.1); (2) GridWorld Proactive Assistance (Section 5.2); (3) Household Question Answering (Section 5.3); and (4) Household Proactive Assistance (Section 5.4).

targets, we view them as evidence. In *MindZero*, the model is rewarded not for predicting actions directly, but for generating mental state hypotheses that maximize the likelihood of user actions, thereby providing coherent explanations of agent behavior. As illustrated in Figure 2a, unlike common RL-based language model training, the reward in our SSRL method is entirely calculated via self-supervised signals from user behavior (without ground-truth mental state annotations) and model outputs. Based on this reward, we then use GRPO Shao et al. (2024); Guo et al. (2025) to train the model, closing the self-supervised learning loop.

4.2 Reward Design

Formally, given a sequence of user behavior $(s_{1:t}, a_{1:t})$, we optimize a multimodal language model Q_θ to approximate the posterior of mental states m_t via variational inference Bishop (2006). As traversing the full hypothesis space is intractable, we maximize the Evidence Lower Bound (ELBO) Kingma & Welling (2014). The optimization objective can be formalized as the following reward function:

$$\mathcal{J}(\theta) = \mathbb{E}_{Q_\theta}[\log(\underbrace{P(a_{1:t} \mid m_t, s_{1:t})}_{\text{action likelihood}} \cdot \underbrace{P(m_t)}_{\text{prior}})] + \underbrace{H(Q_\theta)}_{\text{entropy}} \tag{3}$$

where the P terms denote estimators of the **action likelihood** and **mental state prior** in Equation (1), which are described in detail below. The entropy term $H(Q_\theta)$ encourages exploration over mental state hypotheses and prevents premature collapse to a single mode, thereby promoting robust and diverse posterior approximations.

In practice, the model $Q_\theta(\cdot \mid s_{1:t}, a_{1:t})$ generates a finite set of N mental state hypotheses $\mathcal{M}_t = \{m_t^{(1)}, \ldots, m_t^{(N)}\}$, along with their normalized posterior probabilities $\mathcal{Q}_t = \{q_t^{(1)}, \ldots, q_t^{(N)}\}$ such that $\sum_{i=1}^{N} q^{(i)} = 1$. We treat these N candidates as the effective support of the variational posterior. Consequently, the likelihood, prior, and entropy terms in Equation (3) are computed as weighted sums:

$$R(\mathcal{M}_t, \mathcal{Q}_t) = \sum_{i=1}^{N} q_t^{(i)} [\log(\underbrace{P(a_{1:t} \mid m_t^{(i)}, s_{1:t})}_{\text{action likelihood}} \cdot \underbrace{P(m_t^{(i)})}_{\text{prior}})] \underbrace{- \sum_{i=1}^{N} q_t^{(i)} \log q_t^{(i)}}_{\text{entropy}} . \tag{4}$$

Action Likelihoods. Action likelihoods measure how probable the observed actions are under a given mental state hypothesis. Specifically, $P_t^{(i)} = P(a_{1:t} \mid m_t^{(i)}, s_{1:t})$ computes the likelihood of the action sequence up to time t , given the observed states $s_{1:t}$ and a proposed mental state hypothesis $m_t^{(i)}$. This likelihood can be estimated using either a model-based planner or an LLM.

Mental State Priors. Mental state priors $P(m_t)$ represent the prior probabilities assigned to different mental state hypotheses m_t. These priors can be either uniform or non-uniform to incorporate prior knowledge from symbolic rules or LLMs, helping constrain the hypothesis space. For example, in a household environment, goals such as placing food into a dishwasher or setting the table with vastly mismatched numbers of plates and cutlery would be assigned a low prior probability. This effectively prevents the model from generating hypotheses that violate common sense at the proposal stage.

In summary, to produce hypotheses with high action likelihoods, high mental state priors, and consequently, high rewards, the proposed mental states must be explicit and meaningful for both estimators for the action likelihood and the mental state prior. This then encourages the model to learn to propose explicit and meaningful mental states through RL training. In the meantime, with the entropy bonus objective, the hypothesis distribution would remain diverse and robust. As a result, the model can learn to conduct explicit online mental reasoning without the need for ground-truth mental state annotations.

5 EXPERIMENTS

As shown in Figure 3, we systematically evaluate *MindZero* and baseline methods across four experimental settings: (1) GridWorld Question Answering (Section 5.1), (2) GridWorld Proactive Assistance (Section 5.2), (3) Household Question Answering (Section 5.3), and (4) Household Proactive Assistance (Section 5.4). The question answering settings focus on directly answering ToM-related questions about humans' mental states, whereas the assistance settings require fast, online mental reasoning about human behavior to provide proactive and accurate support.

Table 1: Results of *MindZero* and baselines in GridWorld Question Answering. *MindZero* consistently outperforms all baselines in both accuracy and efficiency.

Method	Accuracy ↑	TFLOPs ↓
Qwen3-VL-4B	37.7	3.6
Qwen3-VL-8B	43.3	7.2
Qwen3-VL-235B-A22B	39.3	21.9
GPT-5.2	50.7	N/A
GPT-5.2-Think	50.7	N/A
ThoughtTracing Kim et al. (2025)		
+ Qwen3-VL-4B	50.3	31.0
+ Qwen3-VL-8B	56.7	54.3
+ Qwen3-VL-235B-A22B	53.0	169.8
+ GPT-5.2	57.3	N/A
+ Gemini-3-Flash	64.0	N/A
AutoToM Zhang et al. (2025)		
+ Qwen3-VL-4B	49.3	344.4
+ Qwen3-VL-8B	52.3	741.2
+ Qwen3-VL-235B-A22B	44.7	1089.7
+ GPT-5.2	57.3	N/A
+ Gemini-3-Flash	47.0	N/A
MindZero (Ours)		
+ Qwen3-VL-4B	63.7	**3.6**
+ Qwen3-VL-8B	**65.0**	7.2

5.1 GRIDWORLD QUESTION ANSWERING

Setting. We adapt the *Construction* environment (Jha et al., 2024), a fully observable 2D grid world where agents navigate around obstacles (e.g., walls) and can carry colored objects to various locations. In this setting, a human agent aims to assemble two blocks of specific colors by picking one up and moving it toward the other. The challenge for a ToM model in this environment is to infer the human's intended goal – specifically, which two colored blocks the human intends to assemble – given a partial trajectory. In addition to mental reasoning, the task requires **visual grounding**, which makes the problem particularly challenging even for state-of-the-art pretrained multimodal language models: the model must ground the question and trajectory to the correct block colors within the scene. This goes beyond prior ToM QA benchmarks, which are predominantly story-based and do not require vision-language grounding. To evaluate the model's performance in such scenarios, we first designed a question answering dataset that describes diverse human action patterns.

Models and Baselines. We evaluate *MindZero* using two base models: Qwen3-VL-4B and Qwen3-VL-8B (Yang et al., 2025). We compare against several baselines: (1) **Pretrained Models**, including Qwen3-VL-4B, Qwen3-VL-8B, Qwen3-VL-235B-A22B, and GPT-5.2; (2) **ThoughtTracing** (Kim et al., 2025), a test-time reasoning algorithm that traces mental states by generating hypotheses and weighting them based on observations, using four different pretrained models as backends; and (3) **AutoToM** (Zhang et al., 2025), a model-based mental inference method based on automated agent modeling. We run *ThoughtTracing* and *AutoToM* with four different pretrained models as backends.

Since they do not support visual contexts, we provide the textual transcript of the GridWorld as input.

For the GridWorld domain, we assume a uniform prior for the reward defined in Equation (4) when training *MindZero*. We use a model-based planner to estimate action likelihoods.

Results. Table 1 summarizes the performance of *MindZero* and baseline methods on GridWorld question answering. Our approach consistently outperforms all pretrained and test-time scaling baselines, including GPT-5.2 and large Qwen3-VL variants, despite using much smaller models. In particular, *MindZero* achieves the highest accuracy with Qwen3-VL-4B and Qwen3-VL-8B while matching the inference cost of standard pretrained models and substantially reducing TFLOPs relative to test-time reasoning methods. These results demonstrate that *MindZero* delivers a superior accuracy–efficiency trade-off without relying on expensive inference-time computation.

5.2 GRIDWORLD PROACTIVE ASSISTANCE

Setting. Using the same *Construction* environment as in Section 5.1, we define a proactive assistance task in which a human agent aims to assemble two blocks of specific colors, while a helper agent must continuously observe the human's actions, infer the intended goal, and assist in completing it more efficiently. A key challenge is *online goal inference under ambiguity*: the assistant must identify the goal early enough to provide timely help, yet not so early that it acts on an incorrect hypothesis. Delayed inference limits the opportunity for effective assistance, while premature but incorrect inference leads to *large penalties* when the assistant helps toward the wrong goal and later revises its belief. Moreover, the *large goal space* further increases the difficulty of accurately inferring the user's intent in a timely manner.

Table 2: Results of *MindZero* and baselines in GridWorld Proactive Assistance. *MindZero* achieves a substantial improvement in task completion speed, while all baselines yield little to no speedup.

Method	Speedup ↑	TFLOPs ↓
Random Goal	0.0	0.0
Qwen3-VL-4B	3.9	**201.9**
Qwen3-VL-235B-A22B	0.0	1024.6
MindZero (Ours)		
+ Qwen3-VL-4B	**41.3**	212.0

Models and Baselines. We evaluate *MindZero* on the base model Qwen3-VL-4B (Yang et al., 2025). We compare against several baselines: (1) **Random Goal**, which randomly samples a goal; and (2) **Pretrained Models**, including Qwen3-VL-4B, Qwen3-VL-235B-A22B, and GPT-5.2.

Results. Table 2 reports performance on GridWorld proactive assistance in terms of speedup and computational cost. *MindZero* achieves a substantial improvement in task completion speed, delivering over 40% speedup with Qwen3-VL-4B, while all pretrained baselines, including large open-weight and proprietary models, yield little to no speedup. Despite its strong performance, *MindZero* maintains comparable TFLOPs to standard pretrained models and remains significantly more efficient than large-scale models. These results demonstrate that *MindZero* enables effective and timely proactive assistance without relying on expensive inference-time computation.

5.3 HOUSEHOLD QUESTION ANSWERING

Setting. We evaluate household question answering using MMToM-QA (Jin et al., 2024), a multimodal benchmark that includes questions covering the beliefs and goals of a person searching for an object (e.g., a remote controller) in a household environment. The task is challenging because it requires joint inference of both beliefs and goals with both visual and textual inputs.

Models and Baselines. We evaluate *MindZero* on three base models: Llama-3.1-8B, Llama-3.2-3B (Dubey et al., 2024), and Qwen3-4B (Yang et al., 2025). We compare against several baselines: (1) **Pretrained Models**, including open-weight models Llama-3.1-8B, Llama-3.2-3B, Qwen3-4B, Qwen3-4B-Think, Qwen3-235B-A22B, and Qwen3-235B-A22B-Think, as well as proprietary models GPT-5.2, GPT-5.2-Think, Gemini-3-Flash, and Gemini-3-Pro; (2) **ThoughtTracing** (Kim et al., 2025); and (3) **AutoToM** (Zhang et al., 2025). We run *ThoughtTracing* and *AutoToM* with five different pretrained models as backends.

For the household domain, we use the same pretrained LLM to estimate both the prior and action likelihood terms for the reward defined in Equation (4) when training *MindZero*. In particular, for the prior term, the LLM produces the log prior probabilities directly by judging whether a goal is

plausible in the context of a household task. This essentially incorporates commonsense knowledge from a pretrained LLM as a prior into the model training, helping the model limit the goal space.

Training Data. We use the MMToM-QA training set to construct training data for *MindZero*. Since the test questions use binary choices, valid hypotheses may often lie outside the provided candidate set. To better match this format, we apply hypothesis filtering to construct binary options instead of sampling from the full hypothesis space. For goal-related questions, we form choices by pairing a randomly sampled observed object with an unobserved one. For belief-related questions, we sample an unobserved object–container pair to create a binary verification task. Applying this filtering strategy to the 953 training episodes yields a final dataset of 4,866 examples.

Table 3: Results of *MindZero* and baselines in Household Question Answering. *MindZero* consistently outperforms all baselines in both accuracy and efficiency.

Method	Accuracy ↑	TFLOPs ↓
Llama-3.1-8B	41.3	12.9
Llama-3.2-3B	34.8	4.0
Qwen3-4B	42.8	10.9
Qwen3-235B-A22B	54.5	80.4
GPT-5.2	65.0	N/A
GPT-5.2-Think	73.5	N/A
Gemini-3-Flash	67.2	N/A
Gemini-3-Pro	60.8	N/A
ThoughtTracing Kim et al. (2025)		
+ Llama-3.1-8B	44.3	571.7
+ Llama-3.2-3B	43.5	232.9
+ Qwen3-4B	54.5	291.2
+ Qwen3-235B-A22B	59.8	2097.9
+ GPT-5.2	68.0	N/A
AutoToM Zhang et al. (2025)		
+ Llama-3.1-8B	22.5	136.3
+ Llama-3.2-3B	8.5	23.4
+ Qwen3-4B	54.7	177.5
+ Qwen3-235B-A22B	67.5	389.9
+ GPT-5.2	76.5	N/A
+ Gemini-3-Flash	80.2	N/A
MindZero (Ours)		
+ Llama-3.1-8B	73.7	12.9
+ Llama-3.2-3B	**76.0**	**4.4**
+ Qwen3-4B	73.8	13.1

Results. Table 3 summarizes performance on Household Question Answering. Note that we follow Kaplan et al. (2020) to calculate FLOPs for each model to estimate the computational costs. *MindZero* consistently achieves strong accuracy across all base models, matching or surpassing large proprietary and test-time scaling methods while using orders of magnitude fewer TFLOPs. In particular, *MindZero* with Llama-3.2-3B attains the highest accuracy among open-weight and test-time scaling methods, and is competitive with the best proprietary models, despite operating at minimal inference cost. Compared to *ThoughtTracing* and *AutoToM*, which require substantial test-time computation, *MindZero* delivers superior accuracy–efficiency trade-offs, demonstrating its effectiveness for scalable and practical household reasoning. Notably, *ThoughtTracing* and *AutoToM*, when equipped with much larger LLM backend models, still perform worse than *MindZero* driven by a small model.

5.4 Household Proactive Assistance

Setting. We evaluate household assistance using the embodied benchmark Online Watch-And-Help (O-WAH) (Puig et al., 2023), where a helper agent observes a human's actions, infers the intended goal, and assists in completing it more efficiently in realistic household environments. In this task, the helper agent must update its goal inference based on the latest observations in an online manner. At each step, we use the uncertainty-aware helping planner proposed in Puig et al. (2023) to generate assistance actions based on the inferred goals. To evaluate generalization, we use different apartments for training and testing. This setting introduces two key challenges beyond story-based evaluation: (1) reasoning must occur at *every timestep* rather than at a single queried moment, and (2) the model must generate *diverse yet plausible hypotheses from scratch* to capture uncertainty, rather than selecting from provided choices.

Table 4: Results of *MindZero* and baselines in Household Proactive Assistance. *MindZero* achieves a substantial improvement in task completion speed and the highest efficiency, while all baselines yield little to no speedup.

Method	Speedup ↑	TFLOPs ↓
Random Goal	-2.2	0.0
Llama-3.1-8B	2.3	656.1
Llama-3.2-3B	1.7	244.3
Qwen3-4B	7.6	213.1
MindZero (Ours)		
+ Llama-3.1-8B	4.3	667.7
+ Llama-3.2-3B	5.0	235.1
+ Qwen3-4B	**26.3**	**201.2**

Models and Baselines. We evaluate *MindZero* on three base models: Llama-3.1-8B, Llama-3.2-3B (Dubey et al., 2024), and Qwen3-4B (Yang et al., 2025). We compare against several baselines: (1) **Random Goal**, which assists based on a randomly sam-

pled goal from the goal space; and (2) **Pretrained Models**, including Llama-3.1-8B, Llama-3.2-3B, Qwen3-4B, and GPT-5.2. All models are paired with the same MCTS planner from Puig et al. (2023) to generate the helping actions conditioned on the corresponding online goal inference.

Results. Table 4 reports performance on Household Proactive Assistance in terms of speedup and computational cost. *MindZero* consistently outperforms the MCTS planner and pretrained baselines across all base models, achieving substantially higher speedup with comparable or lower TFLOPs. In particular, *MindZero* with Qwen3-4B attains a 26.3% speedup—over three times higher than the strongest baseline—while maintaining similar inference cost. In contrast, MCTS planning and standard pretrained models yield limited or even negative improvements. These results demonstrate that *MindZero* enables efficient and effective proactive assistance in complex household environments without relying on expensive planning or inference-time search.

5.5 Ablation Study

To understand the key components driving *MindZero*'s performance, we conduct comprehensive ablation studies on Qwen3-4B. We examine three critical design choices: prior modeling, multiple hypotheses, and entropy bonus. All experiments use the same training configuration as our main experiments.

Explicit Prior Modeling. In the household environment, humans are assumed to pursue a set of predefined goal types, such as setting up the dinner table or putting dishes in the dishwasher. We explicitly require an LLM to check whether each goal hypothesis is reasonable. For example, *putting an apple into the dishwasher* will be assigned a very low score. This constraint is key to generating plausible hypotheses and prevents reward hacking of action likelihood, e.g., including every possible item in the goal yields a high action-likelihood score but a low prior score. Compared to the full model (Row I), the speedup drops by 8.7% without explicit prior modeling (Row II).

Multiple Hypotheses. Maintaining a set of mental state hypotheses is important for capturing the uncertainty of understanding human behavior. For example, in the early stage of an episode, the assistant can only observe a limited human behavior, thus each hypothesis remains ambiguous and carries low confidence. Relying on a single estimation would lead to premature commitment to a potentially incorrect goal. By tracking a beam of hypotheses, the system can defer the decision until sufficient evidence is accumulated. Compared to the full model (Row I), the speedup drops for 16.0% comparing to generating a single most possible mental state (Row III). Accordingly, the token usage is the least.

Table 5: Ablation on Household Proactive Assistance using Qwen3-4B.

#	Method	Speedup ↑	TFLOPs ↓
I	*MindZero*	**26.3**	201.2
II	w/o prior modeling	17.0	200.5
III	w/o multiple hypotheses	10.3	**132.6**
IV	w/o entropy bonus	5.2	245.1

Entropy Bonus. Hypothesis distribution often suffers from mode collapse, where the model becomes overconfident in a single prediction too early. To mitigate this, the entropy regularization term in Equation (3) encourages the diversity of the hypothesis space. This bonus penalizes overly peaked distributions and ensures the model retains alternative possibilities during reasoning. Compared to the full model (Row I), the speedup drops by 21.1% without the entropy bonus (Row IV).

6 Conclusion

We introduced *MindZero*, a self-supervised reinforcement learning framework for training multimodal language models to perform robust and efficient online Theory of Mind reasoning without relying on mental state annotations. By rewarding hypotheses that best explain observed behavior, *MindZero* enables models to internalize the deliberative structure of model-based ToM while retaining the speed of single-pass inference. Extensive evaluations across question answering and proactive assistance tasks demonstrate that *MindZero* achieves strong robustness and uncertainty tracking comparable to explicit model-based methods, while substantially reducing computational cost. These results show that mental reasoning can be learned as a self-supervised skill grounded in behavioral evidence, bridging the long-standing gap between interpretability, robustness, and efficiency in ToM modeling. We believe *MindZero* provides a promising foundation for scalable, real-world assistive agents that can continuously reason about human intentions and adapt to dynamic environments.

REFERENCES

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4): 0064, 2017.

Cristian-Paul Bara, CH-Wang Sky, and Joyce Chai. Mindcraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1112–1125, 2021.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, 3(Feb):1137–1155, 2003.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. Explicit modelling of theory of mind for belief prediction in nonverbal social interactions. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pp. 866–873, 2024a.

Matteo Bortoletto, Lei Shi, and Andreas Bulling. Neural reasoning about agents' goals, preferences, and actions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 456–464, 2024b.

Matteo Bortoletto, Constantin Ruhdorfer, and Andreas Bulling. Tom-ssi: Evaluating theory of mind in situated social interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 32252–32277, 2025.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 15084–15097, 2021.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. Somitom: Evaluating multi-perspective theory of mind in embodied social interactions. In *Advances in Neural Information Processing Systems Datasets and Benchmarks (NeurIPS D&B)*, 2025.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 13518–13529, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In *Findings of the Association for Computational Linguistics: ACL*, pp. 11532–11547, 2024.

X Angelo Huang, Emanuele La Malfa, Samuele Marro, Andrea Asperti, Anthony G Cohn, and Michael J Wooldridge. A notion of complexity for theory of mind via discrete world models. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 2964–2983, 2024.

Kunal Jha, Tuan Anh Le, Chuanyang Jin, Yen-Ling Kuo, Joshua B Tenenbaum, and Tianmin Shu. Neural amortized inference for nested multi-agent reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 530–537, 2024.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 16077–16102, 2024.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 19794–19809, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 14397–14413, 2023.

Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B. Tenenbaum, and Yejin Choi. Hypothesis-driven theory-of-mind reasoning for large language models. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5872–5877, 2019.

Yancheng Liang, Daphne Chen, Abhishek Gupta, Simon S Du, and Natasha Jaques. Learning to cooperate with humans using generative agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 60061–60087, 2024.

Yi-Long Lu, Chunhui Zhang, Jiajun Song, Lifeng Fan, and Wei Wang. Do theory of mind benchmarks need explicit human-like reasoning in language models?, 2025. URL https://arxiv.org/abs/2504.01698.

Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Xavier Puig, Tianmin Shu, Joshua B Tenenbaum, and Antonio Torralba. Nopa: Neurally-guided online probabilistic assistance for building socially intelligent home assistants. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7628–7634. IEEE, 2023.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 4218–4227. PMLR, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Nicholas Rhinehart, Rowan McAllister, Kris Kitani, and Sergey Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2821–2830, 2019.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13960–13980, 2023.

Melanie Sclar, Jane Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. Explore theory of mind: Program-guided adversarial data generation for theory of mind reasoning. *arXiv preprint arXiv:2412.12175*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 1510–1519, 2025.

Maayan Shvo, Ruthrash Hari, Ziggy O'Reilly, Sophia Abolore, Sze-Yuh Nina Wang, and Sheila A McIlraith. Proactive robotic assistance via theory of mind. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9148–9155. IEEE, 2022.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023. URL https://arxiv.org/abs/2302.08399.

Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua Tenenbaum. Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 22, 2009.

Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8292–8308, 2024.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10691–10706, 2023.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8593–8623, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Lance Ying, Katherine M. Collins, Megan Wei, Cedegao E. Zhang, Tan Zhi-Xuan, Adrian Weller, Joshua B. Tenenbaum, and Lionel Wong. The neuro-symbolic inverse planning engine (NIPE): Modeling probabilistic social inferences from linguistic inputs. In *First Workshop on Theory of Mind in Communicating Agents*, 2023.

Lance Ying, Xinyi Li, Shivam Aarya, Yizirui Fang, Yifan Yin, Jason Xinyu Liu, Stefanie Tellex, Joshua B. Tenenbaum, and Tianmin Shu. Pragmatic embodied spoken instruction following in human-robot collaboration with theory of mind, 2025. URL https://arxiv.org/abs/2409.10849.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, Shunchi Zhang, and Tianmin Shu. Autotom: Scaling model-based mental inference via automated agent modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, and Vikash Mansinghka. Online bayesian goal inference for boundedly rational planning agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 19238–19250, 2020.

Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B Tenenbaum. Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. *arXiv preprint arXiv:2402.17930*, 2024.

Xuhui Zhou, Valerie Chen, Zora Zhiruo Wang, Graham Neubig, Maarten Sap, and Xingyao Wang. Tom-swe: User mental modeling for software engineering agents, 2025. URL `https://arxiv.org/abs/2510.21903`.