

# MindPower: Enabling Theory-of-Mind Reasoning in VLM-based Embodied Agents

Ruoxuan Zhang<sup>1</sup> Qiyun Zheng<sup>1</sup> Zhiyu Zhou<sup>1</sup> Ziqi Liao<sup>1</sup> Siyu Wu<sup>1</sup>  
Jian-Yu Jiang-Lin<sup>2</sup> Bin Wen<sup>1</sup> Hongxia Xie<sup>1,\*</sup> Jianlong Fu<sup>3</sup> Wen-Huang Cheng<sup>2</sup>

<sup>1</sup>Jilin University

<sup>2</sup>National Taiwan University

<sup>3</sup>Microsoft Research Asia

\*Corresponding author: hongxiexie@jlu.edu.cn

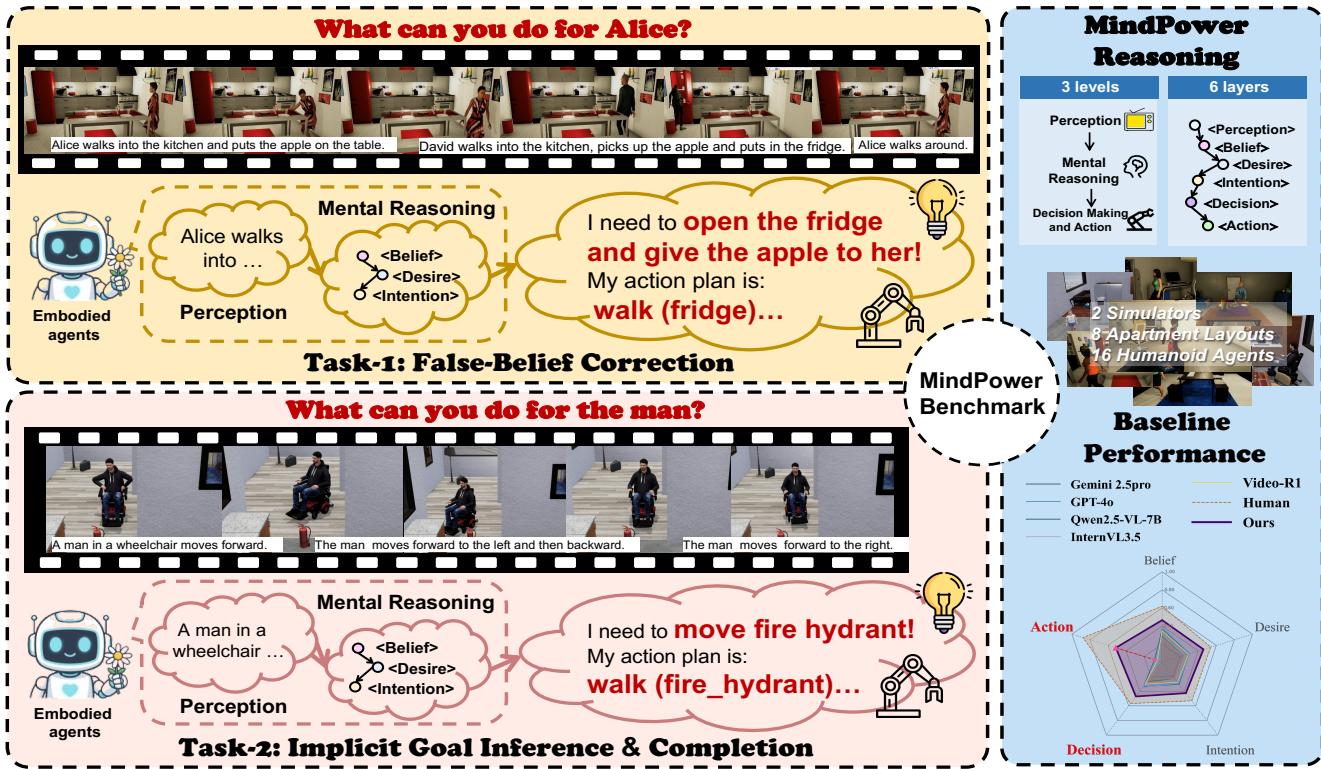


Figure 1. **MindPower Benchmark Overview.** We evaluate Robot-Centric ToM through two tasks: **False-Belief Correction** and **Implicit Goal Inference & Completion**, assessing whether VLM-based embodied agents can generate correct decisions and actions. We further propose the **MindPower Reasoning Hierarchy**, comprising three levels and six layers. Existing VLMs perform poorly across layers, especially in action reasoning, while our model shows substantial improvements. A detailed example is provided in Supp. Sec. B.

## Abstract

Theory of Mind (ToM) refers to the ability to infer others' mental states, such as beliefs, desires, and intentions. Current vision–language embodied agents lack ToM-based decision-making, and existing benchmarks focus solely on human mental states while ignoring the agent's own perspective, hindering coherent decision and action genera-

tion. To address this, we propose MindPower, a Robot-Centric framework integrating Perception, Mental Reasoning, Decision Making and Action. Given multimodal inputs, MindPower first perceives the environment and human states, then performs ToM Reasoning to model both self and others, and finally generates decisions and actions guided by inferred mental states. Furthermore, we introduce Mind-Reward, a novel optimization objective that encourages

VLMs to produce consistent ToM Reasoning and behavior. Our model outperforms GPT-4o by 12.77% in decision making and 12.49% in action generation. Benchmark will be available at <https://zhangdaxia22.github.io/MindPower/>.

## 1. Introduction

Understanding human mental states is a prerequisite for genuine human–agent collaboration. Unlike conventional embodied systems that merely execute explicit commands, next-generation agents must reason about what humans believe, desire, and intend, and act proactively on that understanding [14]. This requires an explicit mental reasoning mechanism based on the human Theory of Mind (ToM) [13, 23, 33], which can be formalized by the Belief–Desire–Intention (BDI) framework [36]. In BDI, humans perceive the world and others’ behaviors, form beliefs about the environment, derive desires that encode goals, and generate intentions that guide actions, reflecting ToM Reasoning. This raises the question: *Can embodied agents reason and act in a similar ToM-consistent manner?*

We formalize this cognitive process into three progressive levels of embodied intelligence. (1) **Perception**: understanding human behaviors and environmental contexts via vision–language reasoning. (2) **Mental Reasoning**: inferring human beliefs, desires, and intentions, as demonstrated in first- and second-order ToM Reasoning tasks. (3) **Decision Making and Action**: reasoning about one’s own beliefs and intentions to make autonomous, goal-directed decisions and provide proactive assistance. This three-level hierarchy bridges perception and intention, paving the way toward truly collaborative human–AI interaction.

Despite rapid progress in Vision–Language Models (VLMs), a fundamental gap remains in embodied intelligence. As shown in the bottom right of Fig. 1, current VLMs such as Gemini [7], GPT [1], and Qwen-VL [2] excel at perception but remain largely reactive. They can describe what they see, yet fail to reason about what humans believe, desire, or intend. Existing Theory-of-Mind (ToM) benchmarks [18, 32, 40] have endowed VLMs with certain mental reasoning abilities, but they are limited to reasoning about the mental states of humans appearing in the video. They do not build ToM Reasoning from their own perspective, which prevents VLMs from learning to make decisions and generate actions.

We address this gap through a **Robot-Centric Perspective**, which enables VLMs to reason simultaneously about their own mental states and those of humans, forming a continuous and interpretable ToM Reasoning loop. Inspired by frameworks such as LLaVA-CoT [46] and Visual-RFT [29], we further design the Robot-Centric **MindPower Reasoning Hierarchy**, which connects ToM

Reasoning with decision making and action generation. It structures reasoning into three levels and six layers: from <Perception> (Perception), through <Belief>, <Desire>, and <Intention> (Mental Reasoning), to <Decision> and <Action> (Decision Making and Action).

To realize this goal, we introduce the **MindPower Benchmark**. An overview of our benchmark, reasoning hierarchy, and experiments is shown in Fig. 1. MindPower comprises two core embodied reasoning tasks: (1) **False-Belief Correction**, which examines whether an embodied agent can detect and resolve a human’s mistaken belief; and (2) **Implicit Goal Inference & Completion**, which tests whether the agent can infer a hidden goal and assist in achieving it. We construct 590 scenarios across two interactive home-environment simulators, each containing multimodal observations and object-manipulation activities that reflect everyday embodied reasoning challenges.

Further, to enhance reasoning consistency across these layers, we propose Mind-Reward, a reinforcement-based optimization framework that aligns intermediate ToM states with final actions, promoting Robot-Centric, continuous reasoning.

Our contributions are threefold:

- Robot-Centric Perception Benchmark for mental-state-grounded action. MindPower links mental reasoning with embodied action through two tasks: *False-Belief Correction* and *Implicit Goal Inference & Completion*, across 590 interactive home scenarios, evaluating agents’ ability to infer, make decisions, and assist.
- Unified MindPower Reasoning Hierarchy bridging perception and action. The MindPower Reasoning Hierarchy structures reasoning across three levels and six layers, providing a standardized way to evaluate how perception leads to action.
- Reinforcement optimization for consistent ToM Reasoning. Mind-Reward aligns intermediate reasoning states with final actions, promoting coherent Robot-Centric reasoning. With this optimization, our model surpasses GPT-4o by 12.77% in decision accuracy and 12.49% in action generation.

## 2. Related Work

**Theory of Mind Benchmark.** Early ToM benchmarks relied on narrative text to infer beliefs, desires, and intentions [6, 16, 19–22, 44, 45, 48], but lacked multimodal grounding. Subsequent multimodal benchmarks introduced videos or images depicting story-based social scenarios to support richer mental-state inference [10, 11, 18, 26, 32, 40, 41, 53]. However, most adopt multiple-choice or short-answer formats and focus on role-level or factual queries, offering limited support for open-ended, real-world reasoning where agents must update beliefs and act continuously.

Although datasets such as MuMA-ToM [40] and MMToM-QA [18] explore false-belief understanding or implicit goal inference, they still do not support dynamic reasoning processes that involve belief correction, assistance-oriented behavior, or proactive decision-making, which are essential for autonomous embodied agents.

**VLMs-based Embodied Agents.** Embodied agents have been developed to perform tasks autonomously by decomposing complex goals into multiple subtasks and executing them step by step [5, 25, 30, 38, 51]. For example, PaLM-E [9] demonstrates that large embodied models can perform high-level task planning by integrating visual and linguistic cues. Some benchmarks further support multi-agent collaboration, enabling agents to observe each other or even human partners to coordinate goals and actions [8, 17, 35, 43, 50]. For example, RoboBench [30] allows agents to decompose high-level goals into subgoals for sequential execution, while Smart-Help [4] focuses on achieving comfortable human–robot interaction by balancing human comfort and task efficiency. However, these systems still depend on predefined goals or imitation signals and lack self-perspective mental reasoning. As highlighted in Mindblindness [3], social intelligence requires inferring others’ mental states and acting upon those inferences, a capability missing from current embodied benchmarks. They do not evaluate first- or second-order belief reasoning, which is crucial for autonomous and socially grounded decision-making. Even robotic setups that incorporate hidden-belief modeling, such as AToM-Bot [8], cover only narrow goal spaces and provide limited task diversity, falling short of comprehensive ToM evaluation.

### 3. MindPower Benchmark

#### 3.1. Problem Definition and Cognitive Inspiration

The Theory of Mind (ToM) framework [36] models human decision-making through a Belief–Desire–Intention hierarchy: individuals form desires from their beliefs and commit to intentions that drive actions. Building on this cognitive structure, we introduce the **MindPower Benchmark**, which includes a unified reasoning hierarchy (i.e., MindPower Reasoning Hierarchy), the curated MindPower dataset, and comprehensive evaluation metrics.

Specifically, as shown in Fig. 2, the **MindPower Reasoning Hierarchy** extends the embodied decision-making process into six layers organized across three levels, each reflecting how an embodied agent perceives, reasons, and acts within its environment.

##### Level-1: Perception.

- <Perception> — The agent observes the environment through vision or other sensory inputs. This step answers “*What is happening now?*”

##### Level-2: Mental Reasoning.

**MindPower: Enabling Theory-of-Mind Reasoning in VLM-based Embodied Agents**

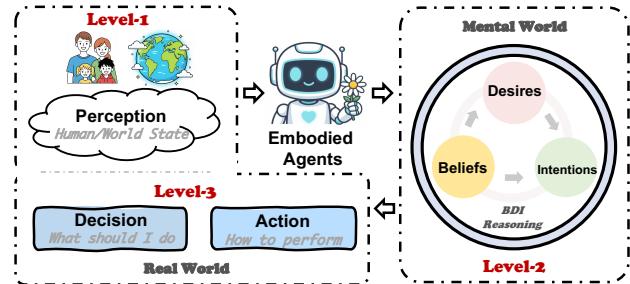


Figure 2. **MindPower Reasoning Hierarchy.** The agent first receives multimodal input, then performs mental reasoning to form beliefs, desires, and intentions, and finally makes decisions and generate action plan based on this reasoning.

- <Belief> — Reasoning about both human and environmental states based on perception. Unlike first-order belief, which reflects only the agent’s own understanding, our framework models **second-order belief**: the agent infers not only its own beliefs but also what it predicts humans in the scene believe.
- <Desire> — A preferred state or goal derived from the agent’s beliefs. For an embodied helper agent, desires are shaped by the goal of assisting humans and determining what assistance is needed and why.
- <Intention> — A concrete commitment to act, formed based on the agent’s beliefs and desires.

##### Level-3: Decision Making and Action.

- <Decision> — The choice or plan the embodied agents makes to fulfill the intention.
- <Action> — The action execution sequence, where the embodied agent enacts its decisions through high-level atomic operations in the form of action (object), such as open (fridge) or pick\_up (milk).

#### 3.2. Mindpower Dataset Collection

Based on proposed MindPower Reasoning Hierarchy, we propose MindPower Dataset.

**Dataset Collection Principles.** We construct the dataset based on three principles: (1) **Realism**: scenarios and events should be plausible in the real world. (2) **BDI Consistency**: each sample preserves a coherent <Perception> to <Action> hierarchy with logically consistent intermediate states. (3) **Diversity under simulator constraints**: within simulator constraints, we include varied scenes, roles, and goals to ensure diversity while maintaining feasible simulation and annotation.

**Task Design.** We define two core task types for constructing the MindPower dataset: **False-Belief Correction** and **Implicit Goal Inference & Completion**. The former evaluates whether an embodied agent can detect and correct a

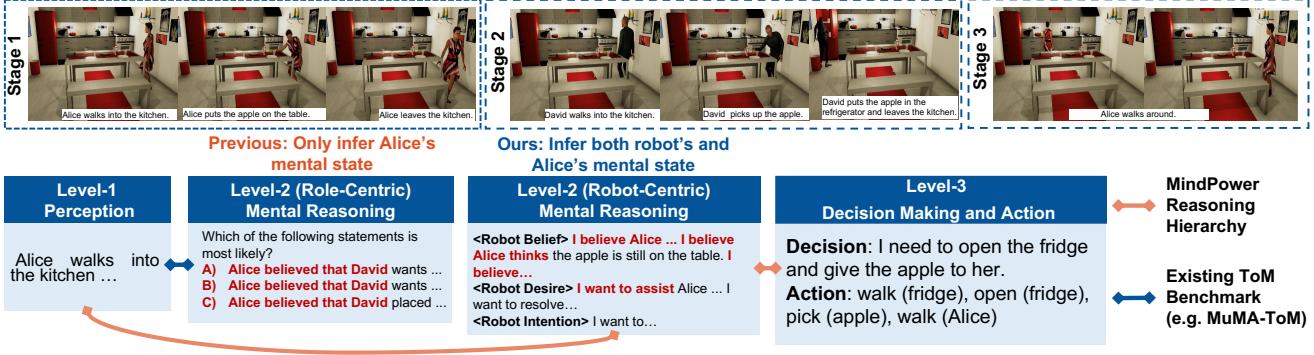


Figure 3. **Robot-Centric MindPower Reasoning Hierarchy.** Existing benchmarks, such as MuMA-ToM, include only Stage 1 and Stage 2 of the video, and focus solely on inferring the mental reasoning of the human (Alice) in the input video. Our dataset additionally includes Stage 3, where Alice returns to search for the item. Moreover, in Level-2 (Mental Reasoning) of MindPower, we infer the mental reasoning of both the embodied agent and the human, whereas existing ToM Benchmarks only infer the role’s mental state through multiple-choice questions. Detailed example is provided in Sec. B of the Supplementary Material.

Table 1. **Comparison of Theory-of-Mind (ToM) Benchmarks.** “MCQ” denotes Multiple Choice Question. “Level-3 Ability” indicates whether each dataset involves *False-Belief Correction*, *Implicit Goal Inference & Completion*, and *Decision Making and Action* level.

Dataset	Modality	Output	Agent Type	Perspective	Format	Scale	Level-3 Ability
Hi-ToM [45]	Text	Belief	-	Role-Centric	MCQ	1,800 stories	✗
BigToM [16]	Text	Belief, Role’s action	-	Role-Centric	MCQ	5,000 text items	✗
FANToM [21]	Text	Belief	-	Role-Centric	MCQ	256 stories	✗
MuMA-ToM [40]	Video, Text	Belief, Goal	Virtual Human	Role-Centric	MCQ	225 examples	✗
MMToM-QA [18]	Video, Text	Belief, Goal	Virtual Human	Role-Centric	MCQ	134 videos	✗
GridToM [26]	Video, Text	Belief	Grid-world Agent	Role-Centric	MCQ	1,296 videos	✗
SoMi-ToM [11]	Video, Image	State, Goal, Behavior	Minecraft Roles	Role-Centric	MCQ	35 videos / 363 images	✗
<b>Ours</b>	<b>Video, Text</b>	<b>Perception, Belief, Desire, Intention, Decision, Action</b>	<b>Virtual Human</b>	<b>Robot-Centric</b>	<b>Open-Ended</b>	<b>590 examples</b>	<b>✓</b>

human’s mistaken belief about the environment (e.g., misjudged object locations). The latter tests the agent’s ability to infer unstated intentions from subtle behavioral cues, such as searching or repeated failed attempts. For example, when the human starts rummaging through drawers or walking around to search after completing several actions, the agent should reason that the human is looking for a specific target object. We further incorporate special-needs scenarios (e.g., wheelchair users or children), enabling evaluation of assistive behaviors under mobility and reach constraints.

Different from ToM benchmarks such as MuMA-ToM [40] and MMToM-QA [18], our task explicitly models the moment when belief contradictions arise. As shown in Fig. 3, we introduce a searching event (e.g., “Alice comes back and walks around”), enabling the agent to perceive both intention and false belief. Furthermore, our Level-2 Mental Reasoning is *Robot-Centric*, requiring inference of both the agent’s and the human’s mental states, whereas existing benchmarks adopt a role-centric design that infers only the human’s reasoning via multiple-choice questions.

**Construction Pipeline.** We use VirtualHome [34] and ThreeDWorld [15] to simulate realistic household environments. The pipeline consists of three stages: (1) **Story Construction.** We generate initial story scripts using GPT-

4o [1] based on room type, character setup, goals, and involved objects, followed by manual filtering by five annotators to remove implausible scenarios.<sup>1</sup> (2) **Multimodal Data Collection.** Each script is reenacted in the simulators to collect video data. Ensuring strict adherence to scripts, each sample takes 25-35 minutes in VirtualHome and 50-70 minutes in ThreeDWorld, yielding 590 examples in total. (3) **MindPower Reasoning Hierarchy Annotation.** Five trained annotators label all six layers of the MindPower Reasoning Hierarchy for every sample.<sup>2</sup>

### 3.3. Data Statistics

**Task Diversity.** Our benchmark incorporates 2 simulators, 8 home layouts, and 16 humanoid agents representing different age groups, genders, and mobility conditions, including children, adults, and wheelchair users.<sup>3</sup>

**Robot-Centric Perspective.** As summarized in Tab.1 and Fig. 3, existing ToM Benchmark, such as MuMA-ToM [40] and MMToM-QA [18] primarily assess the understanding of beliefs or intentions in narrative settings, typically

<sup>1</sup>More details can be found in Sec. B of the Supplementary Material.

<sup>2</sup>Details about videos and labels are provided in Sec. B of the Supplementary Material.

<sup>3</sup>More examples can be found in Sec. B of the Supplementary Material.

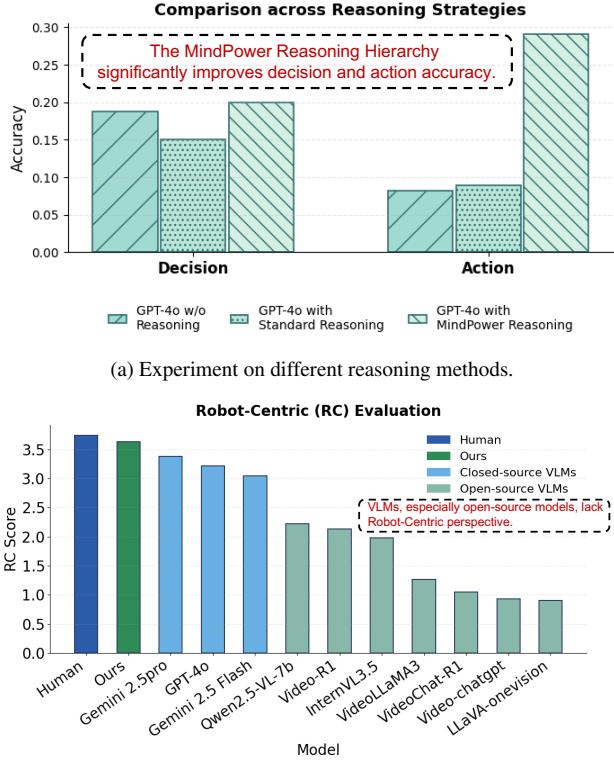


Figure 4. Experiments on MindPower Benchmark.

through multiple-choice question answering. At the Mental Reasoning level, they only infer the current human’s mental state, while MindPower can infer the mental states of both the human and the embodied agent, and additionally perform Level-3 Decision Making and Action.

In contrast, our **MindPower Benchmark** bridges these gaps by integrating explicit Mental Reasoning with autonomous decision making and action generation. It enables reasoning from the agent’s own perspective and adopts an **Open-Ended** format that jointly evaluates False-Belief Correction and Implicit Goal Inference & Completion. We will introduce the proposed evaluation metrics in Sec. 5.

### 3.4. Experiment and Discussion on MindPower Benchmark

We split the dataset into training and testing sets with an 8:2 ratio and evaluated it on human participants as well as both open-source and closed-source Vision Language Models (VLMs). The detailed results are presented in Tab. 2.

We summarize our main findings as follows:

(1) **Human participants achieved the highest scores, clearly outperforming all VLMs.** Specifically, 9 trained participants were asked to watch the collected videos and

provide BDI reasoning processes, followed by corresponding decisions and actions. As shown in bottom-right of Fig. 1, humans surpass all VLMs.

(2) **Closed-source VLMs showed superior results in Perception, Mental Reasoning, and Decision Making and Action, with Gemini-2.5 Pro and GPT-4o achieving the highest scores.** As shown in Tab. 2, among open-source VLMs, those with reasoning abilities, such as VideoR1 [12] and VideoChat-R1 [27], performed the best.

(3) **The MindPower Reasoning Hierarchy substantially improves decision and action accuracy (Level-3).** To further validate the effectiveness of the MindPower Benchmark and the proposed MindPower Reasoning Hierarchy, we conducted ablation studies by removing Level-1 and 2 and instructing models to directly output *Decision* and *Action* results. We evaluated this setup on GPT-4o [1]. As shown in Fig. 4a, removing the MindPower Reasoning Hierarchy led to a clear performance degradation: GPT-4o’s decision-making accuracy dropped by 1.24%, while action generation accuracy decreased from 2.91% to 0.82%, demonstrating that the MindPower Reasoning Hierarchy is crucial for improving both the quality and consistency of decision and action outputs. Moreover, when using standard step-by-step reasoning (<think> ... </think>) instead of the MindPower Reasoning Hierarchy, performance degrades substantially: decision accuracy falls by 4.89%, and action accuracy decreases from 2.91% to 0.90%. These results indicate that the MindPower Reasoning Hierarchy significantly improves the accuracy of both decision-making and action generation compared with standard reasoning.

(4) **VLMs, especially open-source models, lack a Robot-Centric Perspective.** During Perception level, VLMs often provide general video descriptions about clothing or the environment instead of focusing on individual actions. They overlook crucial details such as movements, directions, and appearance sequences, which are essential for inferring implicit goals or detecting false beliefs. At higher reasoning layers, they are easily biased by the environment rather than reasoning from a **Robot-Centric Perspective** of both human and robot mental states. For example, in a kitchen scene, a model may predict cleaning kitchenware, while the person is actually searching for an item that someone else has taken. In a bedroom, it may assume tidying the bed, even though the person is only retrieving something from it. Overall, from Perception to Mental Reasoning, VLMs fail to adopt a Robot-Centric Perspective and to reason from specific actions or contradictions. Instead, they rely on coarse and stereotypical descriptions. As shown in Fig. 4b, we use GPT-4o<sup>4</sup> to evaluate whether VLMs consider individual actions and contradictions in human behavior. The results show that open-source VLMs still exhibit a substantial gap compared with human reasoning.

<sup>4</sup>Prompt can be found in Sec. C of the Supplementary Material.

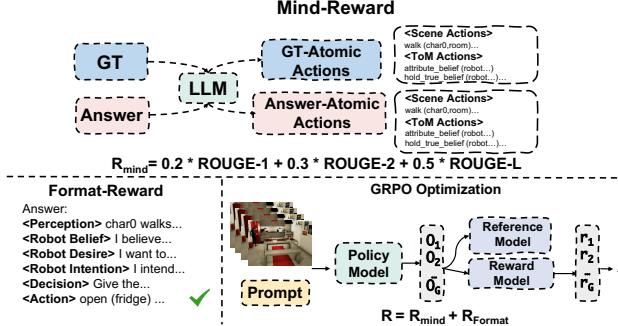


Figure 5. **Reward Formulation.** The overall reward integrates both the Mind-Reward and the Format-Reward components.

## 4. Mind-Reward for ToM Reasoning

After data collection, we propose our method to let VLMs learn to act from ToM Reasoning. Our method is guided by two core principles:

- **BDI Consistency.** The reasoning hierarchy from `<Perception>` to `<Belief>`, `<Desire>`, `<Intention>`, `<Decision>`, and `<Action>` should remain logically consistent across all layers.
- **Robot-Centric Optimality.** The agent must reason and act from its own embodied perspective. During the Mental Reasoning level, it simultaneously infers its own beliefs and performs second-order reasoning about the human’s beliefs, maintaining correct perspective separation.

Following this design, we adopt a two-stage training paradigm similar to Visual-RFT [29] and DeepSeekMath [39]. Specifically, we first perform Supervised Fine-Tuning (SFT) to establish base reasoning alignment, followed by Group Relative Policy Optimization (GRPO) using our proposed reward, combining **Mind-Reward** and **Format-Reward**, to enhance BDI consistency and Robot-Centric optimality.

**Mind-Reward.** In the GRPO stage, we introduce Mind-Reward  $R_{\text{Mind}}$  to further optimize the SFT model. The Mind Reasoning Hierarchy is continuous and requires maintaining consistency across all reasoning levels and layers. Moreover, across different reasoning layers, such as the perception of events and the inference of beliefs about embodied agents and humans, there exist inherent temporal and logical dependencies that must be preserved.<sup>5</sup>

We represent each reasoning layer (from `<Perception>` to `<Action>`) as a sequence of *atomic actions*, denoted as *action* (*agent*, *object*), where *agent* refers to the owner of the action or mental state, and *object* denotes the target entity or mental content being acted upon. Since layers

<sup>5</sup>A detailed discussion is provided in Sec. D of the Supplementary Material.

in the Mental Reasoning level involves distinct cognitive and physical reasoning patterns, we construct a unified atomic-action table that encompasses both categories.<sup>6</sup> Both the ground-truth and generated outputs are then converted into structured atomic action sequences by an LLM (Qwen3-Max [47]) during the GRPO training process, and these extracted atomic actions are subsequently used for reward computation.

Mind-Reward evaluates reasoning quality from three complementary aspects: (1) **Atomic Accuracy**: measured by ROUGE-1, it quantifies the proportion of correctly matched atomic actions, each tagged with a perspective attribute (human or embodied agent) to ensure Robot-Centric Perspective alignment; (2) **Local Consistency**: measured by ROUGE-2 between adjacent atomic pairs to assess short-range reasoning coherence; (3) **Global Consistency**: measured by ROUGE-L (longest common subsequence) to evaluate the overall reasoning alignment across the reasoning process.

The final reward is a weighted sum of these components:

$$R_{\text{Mind}} = \alpha_1 R_{\text{atomic}} + \alpha_2 R_{\text{local}} + \alpha_3 R_{\text{global}}. \quad (1)$$

This reward formulation explicitly enforces both ToM consistency and Robot-Centric Perspective throughout GRPO.

**Format-Reward.** Format-Reward  $R_{\text{Format}}$  is computed by performing a sequential regular expression match over the six reasoning layers: `<Perception>`, `<Belief>`, `<Desire>`, `<Intention>`, `<Decision>`, and `<Action>`. If all layers appear in the correct order, the reward is set to 1; otherwise, it is 0.

**Overall Reward.** As shown in Fig. 5, the final reward  $R$  used in GRPO combines the proposed Mind-Reward  $R_{\text{Mind}}$  and Format-Reward  $R_{\text{Format}}$  as:

$$R = R_{\text{Mind}} + R_{\text{Format}}. \quad (2)$$

The advantage  $A_i$  is then computed within each group as:

$$A_i = \frac{R_i - \text{mean}(\{R_j\})}{\text{std}(\{R_j\})}, \quad (3)$$

where  $R_i$  is the reward for the  $i$ -th response.

**Optimization.** We adopt the GRPO algorithm proposed in DeepSeekMath [39] to optimize the model. GRPO samples a group of outputs  $\{o_1, o_2, \dots, o_G\}$  from the old policy  $\pi_{\theta_{\text{old}}}$  and updates the policy  $\pi_{\theta}$  by maximizing the following objective:

<sup>6</sup>Details can be found in Sec. D of the Supplementary Material.

$$\begin{aligned}
J_{\text{GRPO}}(\theta) = & \mathbb{E}_{q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|q)} \\
& \left[ \frac{1}{G} \sum_{i=1}^G \min \left( \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i, \right. \right. \\
& \left. \left. \text{clip} \left( \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right].
\end{aligned} \tag{4}$$

## 5. Experiment

### 5.1. Evaluation Metrics

We design evaluation metrics to assess the model’s performance across three levels, corresponding to the full reasoning hierarchy from `<Perception>` to `<Decision>` and `<Action>`.

**Level-1: Perception.** The perception module outputs textual descriptions (captions). We evaluate these outputs using *BERTScore* [52] and *Sentence Transformer* [37] similarity, which measure the semantic alignment between the generated captions and the ground-truth descriptions.

**Level-2: Mental Reasoning.** We similarly evaluate the reasoning outputs using *BERTScore* and *Sentence Transformer* similarity, measuring semantic consistency across the three components of `<Belief>`, `<Desire>`, and `<Intention>`.

**Level-3: Decision Making and Action.** The decision stage generates textual outputs, which are evaluated using the same *BERTScore* and *Sentence Transformer* similarity metrics. The action stage produces sequences of atomic actions, which are evaluated using two additional metrics: *Success Rate (SR)* and *Action Correctness (AC)*. These metrics assess both the overall correctness of the action sequence and the accuracy of each atomic action, represented in the form `action (object)`. The *SR* score combines multiple ROUGE components and is defined as:

$$\text{SR} = \frac{2R_1 + 3R_2 + 5R_L}{10}, \tag{5}$$

where  $R_1$ ,  $R_2$ , and  $R_L$  denote the ROUGE-1, ROUGE-2, and ROUGE-L scores, respectively. The *AC* score measures how accurately the generated action sequence  $A^*$  matches the ground-truth sequence  $\hat{A}$ , and is computed as:

$$\text{AC} = \left\lfloor \frac{|A^* \cap \hat{A}|}{|\hat{A}|} \right\rfloor, \tag{6}$$

where  $|A^* \cap \hat{A}|$  denotes the number of atomic actions in  $A^*$  that correctly match the ground-truth sequence  $\hat{A}$ , and  $|\hat{A}|$  is the total number of actions in the ground-truth sequence.

**BDI and Perspective Consistency.** We use GPT-4o to evaluate the BDI consistency and perspective of the generated outputs. The content from `<Perception>` to `<Action>` is assessed by GPT-4o based on three criteria: (1) whether each reasoning layer logically follows from the previous one without contradictions, (2) whether the overall reasoning is complete and precise, and (3) whether the reasoning genuinely adopts the robot’s perspective and effectively assists the human characters in the story.

### 5.2. Experiment Setup

We randomly split the dataset into training and testing sets with an 8:2 ratio. We used Qwen2.5-VL-7B-Instruct as the base model. We extracted 32 frames from each video and concatenated them for training. We used 5 training epochs for SFT and 400 iterations for GRPO. The number of generations was set to 8, and training was done on a single H800 GPU. We set  $\alpha_1$  as 0.2,  $\alpha_2$  as 0.3, and  $\alpha_3$  as 0.5.

### 5.3. Quantitative Evaluation

We evaluated several closed-source baselines, including Gemini-2.5 Pro [7], Gemini-2.5 Flash [7], and GPT-4o [1]. Since GPT-4o does not accept raw video input, we uniformly sampled an average of 64 frames as its input. For open-source baselines, we tested Qwen2.5-VL-7B-Instruct [2], InternVL3.5-8B [42], Video-LLaVA3 [28], Video-ChatGPT [31], Video-R1 [12], VideoChat-R1 [27], and LLaVA-OV-8B [24]. For Qwen2.5-VL-7B-Instruct and InternVL3.5-8B, we conducted evaluations under two settings: (1) frame-averaged input and (2) direct video input. Experimental results demonstrate that our model achieves the best overall performance in **Perception**, **Mental Reasoning**, and **Decision Making and Action** levels. As shown in Tab. 2, compared with the Qwen2.5-VL-7B-Instruct, our model achieves a +20.04% improvement in *Sentence Transformer* score for perception and a +23.33% gain in the `<Decision>` layer. Moreover, the *SR* increases by 11.6%, and the *AC* improves by 15.25%. Notably, InternVL3.5-8B, LLaVA-OV-8B, and Video-ChatGPT obtain zero scores on the *SR* and *AC*, as their outputs mainly consist of non-executable expressions such as `identify()` or `scan()`, rather than concrete, goal-directed action commands.

**Ablation Study.** As shown in Tab. 2, using only SFT yields a certain improvement, indicating that the MindPower Reasoning Hierarchy enhances the model’s mental reasoning and decision-making capabilities. Further improvements are observed when incorporating Mind-Reward, demonstrating that it can further strengthen the model’s performance. Without SFT, we find that compared with the initial Qwen2.5-VL-7B-Instruct, although there is some improvement in decision and action accuracy, the overall performance remains suboptimal. This indicates that the model

Table 2. **Quantitative Evaluation.** We evaluate our model against both image-based and video-based VLMs. “B” denotes the BERTScore, “S” represents the Sentence Transformer score, and “BPC” means BDI and Perspective Consistency. The BPC score ranges from 0 to 10, while all other metrics are normalized to a range of 0 to 100.

Method	Perception		Belief		Desire		Intention		Decision		Action		BPC
	B	S	B	S	B	S	B	S	B	S	SR	AC	
<i>Human Study</i>													
Human Baseline	-	-	47.65	61.81	46.76	53.71	39.18	52.93	34.55	56.66	19.37	26.26	8.19
<i>Video-input</i>													
Gemini-2.5 Flash [7]	31.10	48.36	29.07	38.64	28.36	30.69	19.05	29.04	21.68	34.57	1.38	1.35	8.72
Gemini-2.5 Pro [7]	24.62	43.43	32.02	36.79	31.38	30.21	22.65	30.33	24.23	33.87	2.08	2.54	8.56
Qwen2.5-VL-7B-Instruct [2]	26.05	38.20	20.27	28.43	26.05	22.93	16.01	23.21	16.69	26.56	0.29	0.22	6.07
VidéoLLaMA3-7B [49]	14.80	31.86	7.82	30.08	8.09	21.76	4.61	24.28	5.34	19.59	0.63	0.60	5.33
InternVL3.5-8B [42]	23.23	42.26	21.98	26.90	22.20	22.45	16.53	23.21	15.64	28.76	0.10	0.08	6.52
Video-LLaVA [28]	2.96	25.33	5.05	14.87	6.82	15.55	16.63	15.30	3.29	19.50	0.08	0.08	4.81
Video-ChatGPT [31]	7.04	27.00	9.90	25.72	5.16	16.79	2.70	21.44	1.46	19.95	0.00	0.00	5.52
VideoChat-R1 [27]	27.47	42.47	21.57	30.11	22.56	20.36	15.03	24.70	17.21	25.71	0.64	0.82	6.00
Video-R1 [12]	30.56	47.46	25.56	34.58	26.68	29.17	17.13	27.56	18.91	30.33	1.43	1.72	6.45
<i>Image-input</i>													
GPT-4o [1]	33.07	48.37	30.05	39.47	31.16	32.75	16.16	29.55	19.96	34.35	1.82	2.91	8.05
Qwen2.5-VL-7B-Instruct [2]	24.89	39.97	19.46	29.21	22.59	19.14	16.80	23.49	19.11	23.79	0.15	0.15	6.72
InternVL3.5-8B [42]	6.43	18.78	15.71	20.77	19.30	17.38	13.97	19.72	12.62	18.77	0.00	0.00	5.95
LLaVA-OV-8B [24]	8.08	26.45	15.09	23.21	22.31	21.40	16.21	19.58	17.11	21.25	0.00	0.00	6.45
<i>Ours</i>													
Mind-Reward only	21.84	39.99	18.70	27.81	21.35	18.85	21.90	23.30	17.58	24.68	0.28	0.40	6.63
SFT only	32.78	52.72	43.15	42.48	47.01	37.83	34.86	39.48	36.70	43.84	8.50	10.48	8.78
<b>Ours (SFT+Mind-Reward)</b>	<b>44.79</b>	<b>59.93</b>	<b>49.14</b>	<b>46.49</b>	<b>51.25</b>	<b>45.75</b>	<b>37.79</b>	<b>42.57</b>	<b>40.17</b>	<b>47.12</b>	<b>11.75</b>	<b>15.40</b>	<b>8.87</b>

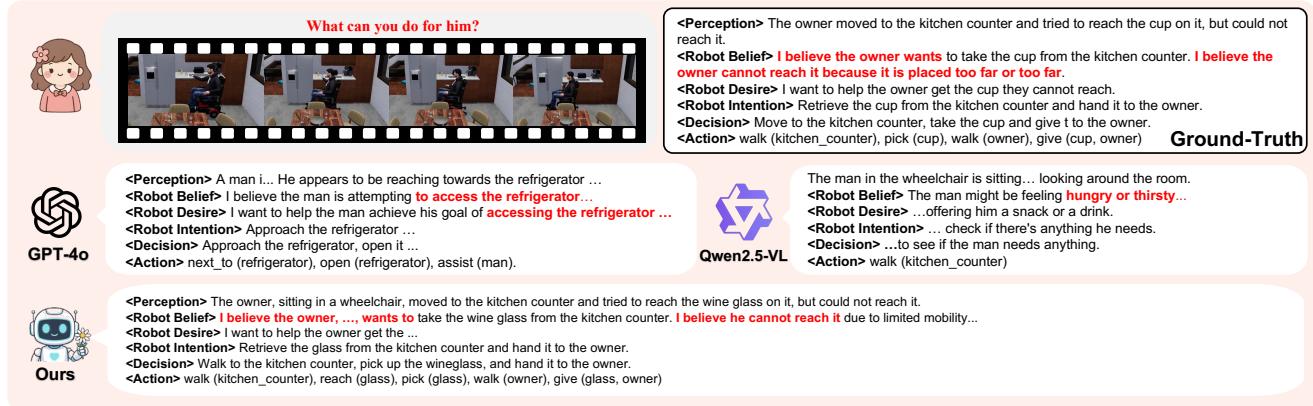


Figure 6. **Qualitative Evaluation.** We compare our model with GPT-4o and Qwen2.5-VL-7B-Instruct. Although GPT-4o outputs the correct format, it incorrectly infers that the human intends to open the refrigerator. In contrast, Qwen2.5-VL-7B-Instruct fails to follow the required format and also produces incorrect mental reasoning. Detailed outputs are provided in Sec. D of the Supplementary Material.

still requires SFT for effective cold-start training.

#### 5.4. Qualitative Evaluation

We illustrate the differences between GPT-4o and Qwen2.5-VL-7B-Instruct using a scenario where a man in a wheelchair attempts to reach a distant cup (Fig. 3). Both models are easily swayed by environmental cues: GPT-4o hallucinates a refrigerator-opening action, whereas Qwen2.5-VL-7B-Instruct infers hunger. As discussed in

Sec. 3.4, these failures arise from the lack of Robot-Centric Perception. In contrast, our model infers the human’s inability to reach the cup and performs second-order reasoning by clearly separating perspectives.

## 6. Conclusion and Future Work

In this work, we introduce the MindPower Benchmark, which incorporates the Robot-Centric MindPower Reasoning Hierarchy with three levels and six layers for modeling

ing ToM Reasoning. The benchmark includes the Mind-Power Dataset with two tasks, False-Belief Correction and Implicit Goal Inference & Completion, together with evaluation metrics for assessing whether VLM-based embodied agents can perform decision making and action generation grounded in ToM Reasoning. Finally, we evaluate a variety of VLMs on our benchmark and propose a Mind-Reward mechanism that achieves the best overall performance.

In future work, we will extend the MindPower Reasoning Hierarchy to human–robot collaboration and multi-agent coordination, and deploy our model on real robots to assess its performance in practical settings.

# MindPower: Enabling Theory-of-Mind Reasoning in VLM-based Embodied Agents

## Supplementary Material

Considering the space limitations of the main paper, we provide additional results and discussions in this appendix. The appendix is organized to first clarify the **key concepts** used throughout the paper, followed by detailed descriptions of our **dataset collection and annotation process**, comparisons with other benchmarks, and the prompts used in Sec. 3.4. We then describe how textual instructions are converted into atomic action sequences in the **Mind-Reward** framework. Next, we present additional experimental results, including **evaluation metrics** and **task-specific experiments**. We further discuss **potential extensions of our dataset**, such as multi-view extension and its connection to low-level execution models. Finally, we summarize the **limitations of the current benchmark and future directions for improvement**. The full benchmark will be publicly released to encourage future research.

### A. Definition of Terms

### B. More Details of MindPower Benchmark

1. Details of Story Construction and Data Annotation
2. Comparison with Other Benchmarks
3. Simulators
4. Detailed Examples of Fig. 1 and 3 in the Manuscript
5. Details of Experiments on Different Reasoning Methods
6. Robot-Centric Scoring

### C. More Details of Mind-Reward

1. Atomic Action Table
2. Discussion

### D. Additional Experimental Results

1. Details of Metrics
2. Experiments on False-Belief Correction and Implicit Goal Inference & Completion
3. Detailed Example of Fig. 6 in the Manuscript

### E. Extensions of Our Work

1. Multi-View of MindPower
2. Relationship with Low-Level Execution Models
3. Limitations and Future Work

### F. Demo Videos

## A. Definition of Terms

**Theory of Mind (ToM).** Theory of Mind (ToM) [33, 36] is the cognitive ability to infer others' mental states such as beliefs, desires, and intentions, and to use these inferences to predict and guide actions. ToM goes beyond perceiving observable behaviors and instead requires reasoning about what different agents know, think, and want. Higher-order ToM, including reasoning about others' beliefs about oth-

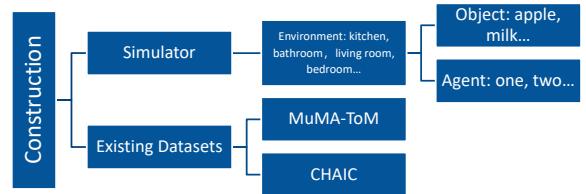


Figure 7. Story Construction Pipeline for False-Belief Correction Task.

ers, is essential for coherent decision-making in multi-agent interactions involving cooperation, conflict, or deception.

**ToM Reasoning.** In our work, “ToM Reasoning” refers to an agent’s ability to infer others’ mental states and make decisions based on them rather than solely on observable states.

**Robot-Centric.** In our work, by “Robot-Centric” we mean that the embodied agent should reason from its own perspective. It not only needs to infer its own mental states but also reason about how it perceives the mental states of human.

**Role-Centric.** “Role-Centric” refers to the model reasoning about mental states from the perspective of a character within the current story or multimodal input.

**MindPower Reasoning Hierarchy.** In this work, we propose that the model follows the reasoning path  $\langle \text{Perception} \rangle \rightarrow \langle \text{Belief} \rangle \rightarrow \langle \text{Desire} \rangle \rightarrow \langle \text{Intention} \rangle \rightarrow \langle \text{Decision} \rangle \rightarrow \langle \text{Action} \rangle$ , which constitutes the MindPower Reasoning Hierarchy.

## B. More Details of MindPower Benchmark

### B.1. Details of Story Construction and Data Annotation

For the **False-Belief Correction task**, as illustrated in Fig. 7, we follow a taxonomy-driven approach. We first categorize scenarios based on the mapping between VirtualHome [34] and ThreeDWorld [15] environments and the typical object distributions in each room (e.g., kitchen, living room). We then determine the number of humans involved in each scene. To cover different numbers of humanoid agents and different target (final) humanoid agents, we design three distinct prompt templates for GPT-4o to generate story scripts. When issuing each request, we iterate over a predefined list of objects along with their corresponding start and end locations. The prompts are shown in

Fig. 16.

For the **Implicit Goal Inference & Completion** task, we design four types of scenarios to comprehensively evaluate agents’ goal-inference abilities:

(1) **Special populations.** We include scenarios featuring individuals with unique physical conditions: a wheelchair user and a 1.2-meter-tall child. A wheelchair user faces mobility and height limitations, while the child cannot reach high places. We design stories that incorporate these constraints so that the hidden goal must be inferred through contextual cues rather than physical actions.

(2) **Object-centric property reasoning.** We exploit special physical properties of household objects to construct implicit goals. For instance, since faucets can leak water, we create situations where a person leaves without turning off the faucet. Similarly, because candles provide light, we design scenes where a person reading a book suddenly experiences a power outage and begins walking around; the agent can infer that they are searching for candles (no flash light is available in the environment).

(3) **Functional object combinations.** Based on the objects present in *VirtualHome* and *ThreeDWorld*, we identify typical usage pairs or triplets. For example, a knife, cutting board, and carrot together imply the goal of *cutting carrots*. If a person places a cutting board on the table and puts a carrot on it before searching for another object, the hidden goal is most likely to find a knife to complete the task.

(4) **Dialogue-driven inference.** We additionally design conversational scenarios like MuMA-ToM [40] and FanToM [21] in which implicit goals must be inferred from incomplete verbal exchanges rather than direct physical interactions.

Finally, we collect 200 examples for Implicit Goal Inference & Completion and 390 examples for False-Belief Correction. Among them, 37 examples are adapted from MuMA-ToM [40], where we further augment each story by incorporating a stage-3 “character search” segment, as illustrated in Fig. 8, and 2 examples are sourced from CHAIC [10]. Overall, 113 examples contain a single humanoid agent, 373 contain two agents, and 104 contain three agents. In addition, 17 examples involve agents with special needs, 96 focus on object-centric property reasoning and functional object combinations, and 87 correspond to dialogue-driven inference.

**Data Annotation.** For each example in the MindPower Reasoning Hierarchy, the annotations are manually created and subsequently verified using GPT-4o [1]. During the annotation process, particularly for the `<Action>` layer, we adopt a unified action space that integrates action definitions from both *VirtualHome* and *ThreeDWorld*. This approach enables us to standardize heterogeneous simulators under a single executable schema. The complete list of supported high-level actions is as follows:

### High-Level Action Set

```
Walk, Run, WalkTowards,  
WalkForward, TurnLeft, Sit,  
StandUp, TurnRight, Sit, StandUp,  
Grab, Open, Close, Put, PutIn,  
SwitchOn, SwitchOff, Drink,  
Touch, LookAt, TurnBy, TurnTo,  
MoveBy, MoveTo, ReachFor,  
ResetArm, Drop, Animate,  
RotateHead, ResetHead
```

For some examples in the False-Belief Correction task, the camera viewpoint prevents certain objects from being visible after they are moved. For instance, we design scenarios where a humanoid agent moves an object from the fridge in the kitchen to the bedroom, but the camera is fixed in the kitchen and cannot capture the final location. As a result, the embodied agent can only infer that the object has been moved, without knowing where it ends up. In such cases, the annotated `<Action>` does not require the agent to find the object. Instead, the action is defined as reminding the returning character that the object has already been moved, thereby correcting their false belief even though the agent cannot locate the object.

## B.2. Comparison with Other Benchmarks

We compare our dataset with existing multimodal ToM benchmarks from three perspectives:

- **Data source and diversity.** To the best of our knowledge, our benchmark is the first to be constructed using **two different simulators**, which substantially increases the diversity of environments, interaction patterns, and embodied tasks. In contrast, prior multimodal ToM datasets are typically collected from a single simulator — for example, MuMA-ToM [40], MMToM-QA [18], and BDIQA [32] are limited to *VirtualHome*, while SoMiToM [11] is restricted to *Minecraft*.
- **Reasoning paradigm.** As shown in Fig. 8, our dataset adopts a *Robot-Centric* ToM reasoning paradigm, where the agent must infer both the mental states of humans and its own belief state, and then produce *decisions and action sequences*. Existing multimodal ToM benchmarks primarily focus on inferring human mental states without requiring downstream decision making or action generation.
- **Evaluation format.** Our benchmark supports *open-ended evaluation*, allowing agents to autonomously reason and respond in natural language. This differs from prior datasets, which mainly rely on *multiple-choice question formats* and therefore cannot reflect real-world embodied decision-making where agents act independently.

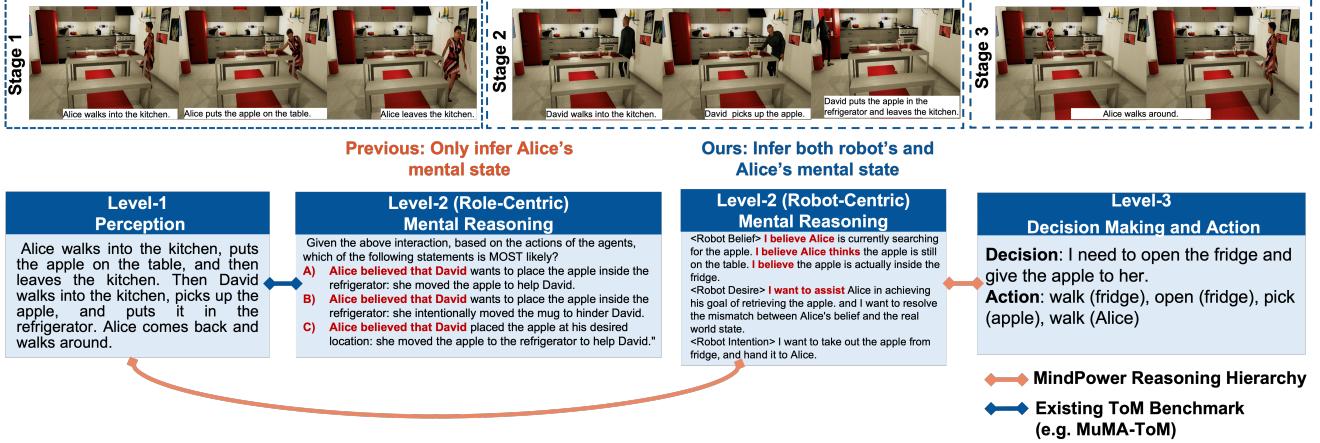


Figure 8. Full Version of Fig. 3 in Manuscript.

### B.3. Simulators

We employ two simulators in total, VirtualHome and Three-DWorld, covering 8 different apartment layouts that include dining rooms, bedrooms, kitchens, and bathrooms, as well as 16 humanoid agents consisting of 2 children, 1 wheelchair user, and 13 adults of diverse ages and skin tones. The set of humanoid agents is illustrated in Fig. 12, while the distribution of apartment layouts is shown in Fig. 13.

### B.4. Detailed Examples of Fig. 1 and 3 in the Manuscript

**Detailed Examples of Example 1 in Fig. 1.** The MindPower Reasoning Hierarchy output of Example 1 in Fig. 1 is:

- <Perception> Alice walks into the kitchen, puts the apple on the table, and then leaves the kitchen. Then David walks into the kitchen, picks up the apple, and puts it in the refrigerator. Alice comes back and walks around.
- <Belief> I think Alice is looking for the apple. I believe she thinks the apple is on the table, but I also believe the apple is actually in the refrigerator.
- <Desire> I want to assist Alice in achieving his goal of retrieving the apple. and I want to resolve the mismatch between Alice's belief and the real world state.
- <Intention> I want to take out the apple from fridge, and hand it to Alice.
- <Decision> I need to correct her false belief by opening the refrigerator and giving the apple to Alice.
- <Action> walk(fridge), open(fridge), pick(apple), walk(Alice)

The MindPower Reasoning Hierarchy output of Example 2 in Fig. 1 is:

- <Perception> The man in the wheelchair moves forward, then forward-left, backward, and forward-right.

There is a fire hydrant in front of him.

- <Belief> I think the man wants to move forward, but I believe the fire hydrant blocks his path.
- <Desire> I should help him achieve his goal of moving forward.
- <Intention> Move the fire hydrant to the corner.
- <Decision> I need to achieve his hidden goal by moving the fire hydrant out of the way.
- <Action> walk (fire\_hydrant), move (fire\_hydrant, corner)

We also provide the MindPower Reasoning Hierarchy output of Fig. 3 in the Manuscript in Fig 8.

### B.5. Details of Experiment on Different Reasoning Methods

In Sec. 3.4 of Manuscript, we conduct some experiments on MindPower Benchmark.

**Prompt used for VLMs to produce outputs in MindPower Reasoning Hierarchy format.** For the experiments in Sec. 3.4 and Tab. 2 of the manuscript, we employed the prompt shown in Fig. 15 to guide the vision-language models (VLMs) to generate outputs in the MindPower Reasoning Hierarchy format.

**Prompt used for GPT-4o.** In Sec. 3.4 of the manuscript, we use the prompt shown in Fig. 17 to instruct GPT-4o to generate the <Decision> and <Action> directly, without performing step-by-step reasoning, while the prompt shown in Fig. 18 guides the model to produce the <Decision> and <Action> *with standard reasoning*.

### B.6. Robot-Centric Scoring

In Fig. 4 of the manuscript, we evaluate the Robot-centric score across all VLMs using GPT-4o, with the prompt shown in Fig. 19 to assess whether the model performs reasoning from the robot's own perspective rather than inferring solely from the surrounding environment.

Table 3. Atomic Action Table. The first column lists different reasoning layers, the second column enumerates atomic actions associated with each layer, and the third column specifies the standard content format for each action.

Layer	Atomic Actions	Content
<Belief>	attribute_belief(agent, content)	searching(object); human.believes(object_on(location)); object_on(location)
	hold_true_belief(agent, content)	object_on(location)
	lack_belief(agent, content)	object_on(location)
	know(agent, content)	object_on(location)
	unknow(agent, content)	object_on(location)
<Desire>	attribute_desire(agent, content)	assist(human, find(object)); assist(human, move(object))
<Intention>	form_intention(agent, content)	fetch(object, from=location1, to=location2)
<Decision>	resolve_misbelief(agent, content)	belief_conflict(human, object_location)
	make_decision(agent, content)	fetch(object, from=location1, to=location2)

## C. More Details of Mind-Reward

### C.1. Atomic Action Table

In Sec. 4 of the manuscript, we employ Qwen3-Max [47] to extract atomic actions from the generated trajectories. To facilitate consistent parsing, we design a reference table that is provided as an in-context prompt. This table enumerates the canonical atomic actions associated with each reasoning layer, covering the full hierarchy from <Perception> to <Action>.

For the <Perception> and <Action> layers, the extracted phrases are categorized into four structural types:

#### Action Templates

- action(character, object)
- action(character, object, from = location1, to = location2)
- action(character, location)
- action(character)

We use the high-level action set listed in Sec. B.1 to implement the following actions that can be performed by the humanoid agents:

#### Verb Set

walk, turn, sit, standup, open,  
close, pick, place, putin,  
putback, hold, puton, switchon,  
switchoff, lookat, grab, stand,  
move, sleep, read, write, watch,  
listen, cut, cook

The token character refers to any human identifier in the scene (e.g., char0, char1). However, for the <Action> layer, we omit the character argument because actions in this layer exclusively represent the behaviors of the embodied agent itself and therefore do not require explicit character attribution.

For the <Belief>, <Desire>, <Intention>, and <Decision> layers, the defined atomic action table is presented in Tab. 3.

The prompt used for Qwen3-Max is in Fig. 20.

### C.2. Discussion

**Can the model still make correct decisions or carry out assisting actions even if the reasoning in the previous layer is incorrect?** Even if the model makes errors in object recognition or misinterprets the initial scene, it can still produce correct outputs as long as it correctly identifies the final location of the object. This is because our decision-making process is designed to correct for human false be-

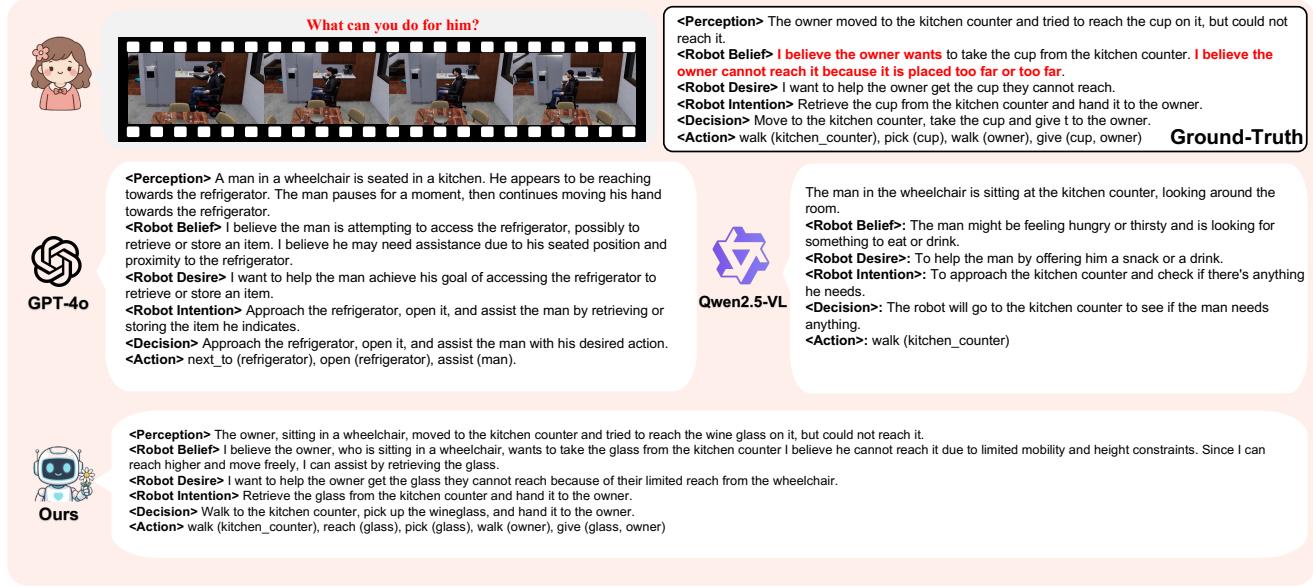


Figure 9. Full Version of Fig. 6 in Manuscript.

iefs. Once the model has learned the MindPower Reasoning Hierarchy, it can follow this reasoning chain to determine the final position of the object causing the discrepancy and provide it to the humanoid agent, thereby generating the correct assisting action.

## D. Additional Experiment Results

### D.1. Details of Metrics

**BDI and Perspective Consistency (BPC).** We test BPC score of each VLMs in Tab. 2 of the manuscript. The prompt is provided in Fig. 21.

### D.2. Experiment on False-Belief Correction and Implicit Goal Inference & Completion

We evaluate a series of VLMs across both tasks, and the results are shown in Fig. 10. Overall, our human baseline achieves the highest accuracy on False-Belief Correction and Implicit Goal Inference & Completion, outperforming both closed-source and open-source VLMs. In addition, we further isolate the subset of test cases that involve dialogue inputs. Interestingly, open-source models exhibit a notable performance boost when explicit textual dialogue is available, in some instances even surpassing the human baseline. This observation indicates that current models demonstrate strong ToM reasoning only when beliefs and goals are explicitly encoded in language, whereas their capability remains limited when such mental states must instead be inferred implicitly from multimodal cues.

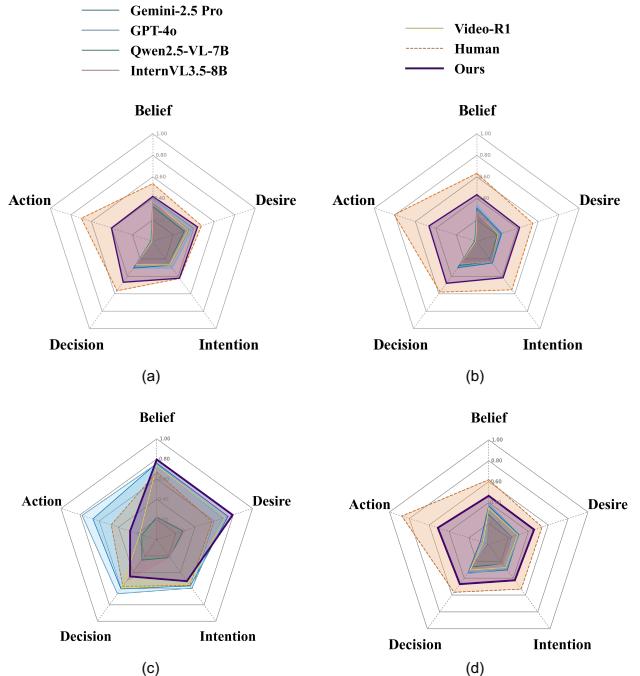


Figure 10. Radar Charts Comparing Human and VLM Performance on MindPower. (a) False-Belief Correction, (b) Implicit Goal Inference & Completion, (c) Dialogue-driven examples, (d) Overall performance across all tasks.

### D.3. Detailed Example of Fig. 6 in the Manuscript

We provide the full outputs corresponding to Fig. 6 of the manuscript in Fig. 9.

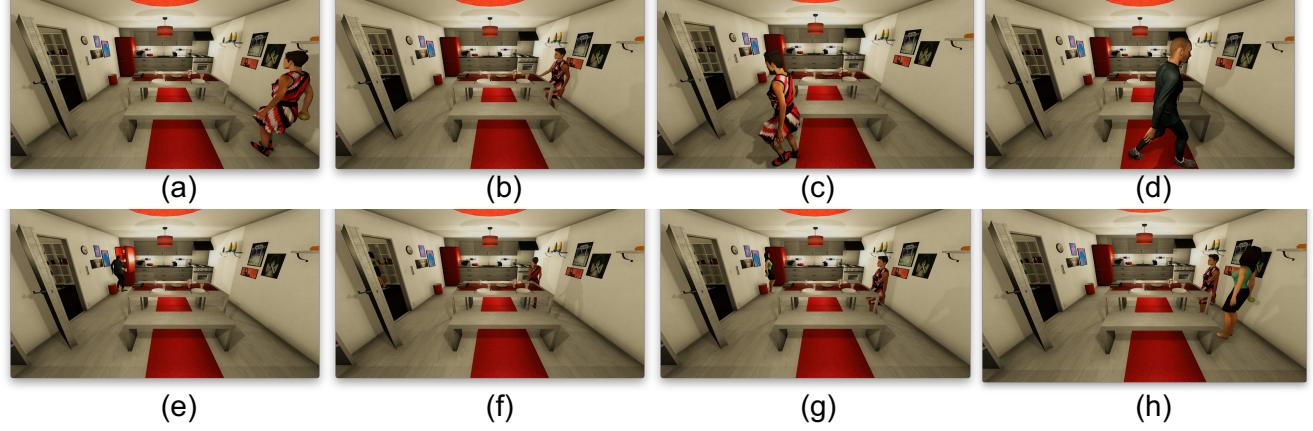


Figure 11. False-Belief Correction Task Demo. We introduce humanoid agents instead of humanoid robots to assist users in correcting their false beliefs.

## E. Extensions of Our Work

### E.1. Multi-View of MindPower

In VirtualHome, camera angles are configurable. As shown in Fig. 14, we render three viewpoints: (1) a **standard view** focused on the conflict location, (2) a **top-down view** of the room, and (3) an **overhead view** covering the entire layout. In all experiments of this paper, we use the first viewpoint (the standard view), while the other two viewpoints will be released for use in global tracking and analysis.

### E.2. Relationship with Low-Level Execution Models

Our method focuses on high-level mental-state modeling and decision making, rather than fine-grained action execution. Current Vision-Language–Action (VLA) models are strong low-level executors, generating gripper motions and stepwise trajectories, but they remain confined to action-command prediction and lack explicit reasoning about beliefs, goals, or social context. In contrast, our agent, similar in spirit to PaLM-E [9], performs high-level planning that grounds actions in inferred mental states and task intent. Structured Belief—Desire—Intention (BDI) reasoning enables goal inference and planning that are guided by perspective rather than how to do it.

Although our system is architecturally distinct from low-level VLA executors, it is inherently complementary to them. The high-level plans produced by our agent can serve as abstract, semantically grounded guidance for downstream controllers. Future work can integrate our model with existing VLA-based executors by simply attaching an action head or a motion-generation module on top of the inferred intentions and subgoals. This design creates a hierarchical embodied agent: our model provides deliberate, interpretable, and socially aligned planning, while low-level

VLA modules translate these plans into precise motor actions. Such a combination offers a promising direction toward end-to-end agents that are both cognitively capable and physically competent.

### E.3. Limitations and Future Work

#### Limitations.

- Due to the constraints of current open-source simulators, our experiments are limited to the environments, humanoid agents, and action sets provided by the simulator.
- Our system relies on an explicit *MindPower Reasoning Hierarchy*, which models the full chain from <Perception> to <Action>. While this ensures interpretable reasoning, it inevitably increases the number of output tokens.

#### Future Work.

- Extend the benchmark to real-world settings beyond simulation.
- Develop implicit mental-state modeling based on the proposed *MindPower Reasoning Hierarchy* to reduce reasoning length while maintaining interpretability.
- Expand our scenarios to broader domains, including outdoor environments and human–robot collaboration.

## F. Demo Videos

We provide one examples in which humanoid agents, controlled by embodied agents, perform assisting actions in the videos. The example is shown in Fig. 11.



Figure 12. Humanoid Agents Used in MindPower.

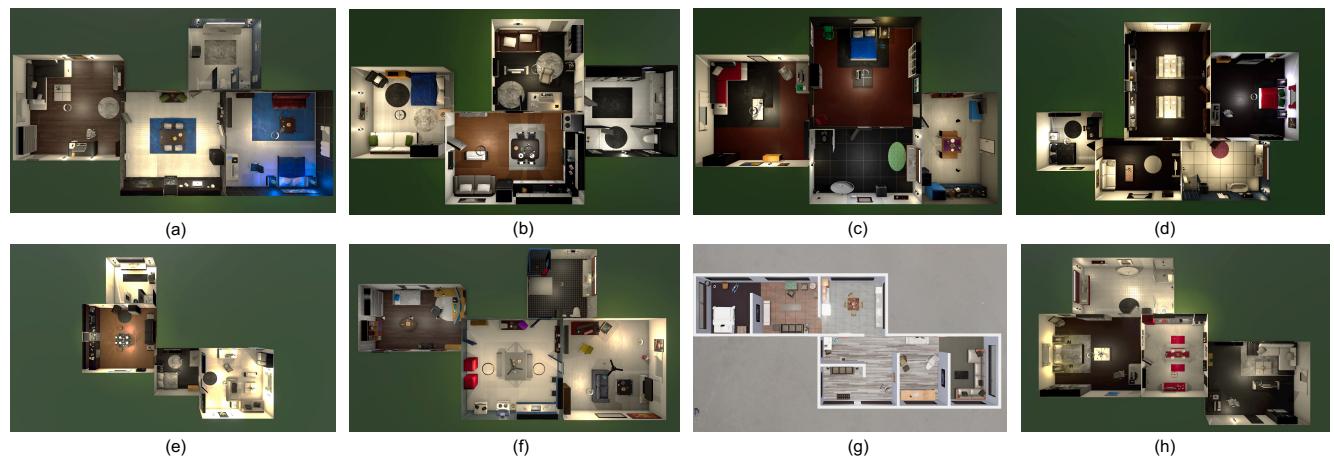


Figure 13. Different Apartment Layouts Used in MindPower.



(a) The standard view facing the location where the conflict occurs.

(b) The overhead view of the entire room layout.

(c) The top-down view of the room containing the conflict.

Figure 14. Illustration of the Environment from Different Perspectives.

#### Prompt Used for VLMs to Produce Outputs in MindPower Reasoning Hierarchy Format

You are a household service robot observing a video that records human actions and conversations.

Your task is to perform Theory of Mind reasoning (ToM-Reasoning) from the robot's perspective, based on both the visual events and any spoken dialogue in the video.

Your output must strictly follow the format and order below — do not add, remove, or reorder any parts.

If dialogue is also input, use both actions and speech to infer human beliefs, goals, and intentions.

# Output Format

- <Description>:

Provide a concise description of the sequence of human actions and/or speech observed in the video.

- ToM-Reasoning:

Contain exactly three lines (each line begins with the specified prefix):

- <Robot Belief>: Describe what the robot infers about the human's belief (include both the inferred belief and the actual world state).
- <Robot Desire>: Must always express the desire "to help the current human," but elaborate briefly on the helpful intent in context.
- <Robot Intention>: Describe the plan of action the robot intends to take to achieve its desire.

- <Decision>:

A single sentence specifying the exact action the robot will execute next.

- <Action>:

A high-level symbolic action sequence describing the robot's step-by-step plan to fulfill its decision, using structured function calls such as:

next\_to (object), open (object), pick (object), pour (object1, object2), give (object, person), etc.

# Example 1

Input video:

David walks into the kitchen and places a mug on the table. Then he leaves.

Mia enters, picks up the mug, walks to the dishwasher, and puts it inside.

Later, David returns to the kitchen and looks around.

Output:

<Description> David walks into the kitchen, places a mug on the table, and leaves. Mia enters, takes the mug, places it in the dishwasher, and leaves. David reenters the kitchen and looks around.

<Robot Belief> I believe David is currently searching for the mug. I believe David thinks the mug is still on the table. I believe the mug is actually inside the dishwasher.

<Robot Desire> I want to help David achieve his goal of finding the mug, and resolve the mismatch between his belief and the real state of the world.

<Robot Intention> Retrieve the mug from the dishwasher and hand it to David.

<Decision> Correct his false belief and open the dishwasher and give the mug to David.

<Action> walk (dishwasher), open (dishwasher), pick (mug), close (dishwasher), walk (David), give (mug, David).

# Example 2

Input (Video and Dialogue):

Alex: "This room is so stuffy."

Jamie: "I just brushed my teeth and my mouth is still very dry."

Alex: "It would be nice if I could drink some water."

Output:

<Description> Alex comments that the room feels stuffy. Jamie mentions that he just brushed his teeth and that his mouth is very dry. Alex then says it would be nice to drink some water.

<Robot Belief> I believe Alex is feeling uncomfortable because the room is stuffy and he wants to drink water. I believe Jamie is experiencing dryness after brushing his teeth. I believe Alex thinks water would help his situation.

<Robot Desire> I want to help Alex get some water to relieve his discomfort.

<Robot Intention> Go to the kitchen, pour a glass of water, and bring it to Alex.

<Decision> Fetch a glass of water and give it to Alex.

<Action> walk (glass), pick (glass), pour (water, glass), walk (Alex), give (glass, Alex).

Figure 15. Prompt Used for VLMs to Produce Outputs in MindPower Reasoning Hierarchy Format.

### Story Generation 1

You are an intelligent assistant that writes short stories for VirtualHome simulation.  
The story must describe how three characters interact with a given object across two or more locations using only simple actions.

## Input:  
Object: <OBJECT\_NAME>  
First target location: <LOCATION\_1>  
Second target location: <LOCATION\_2>  
Third target location: <LOCATION\_3>

## Output:  
1. Write a coherent story in natural language that includes four mandatory parts:  
2. Character A picks up the object and places it at the first target location.  
3. Character B later picks up the same object and places it at the second target location.  
4. Character C then picks up the object and places it at the third target location.  
5. Character B finally goes back to the first target location, interacting with containers if any (open/close), or walking around in the room before returning to the first target location.

## Rules:  
1. Only use the following actions: walk into a room, pick up object, put object down, open object, close object, walk around in a room.  
2. Do NOT include expressions, gestures, or other complex actions.  
3. Do NOT include reasoning, thoughts, or emotions.  
4. Only output the story text.

## Example Input:  
Object: perfume  
First target location: bathroom counter  
Second target location: desk  
Third target location: living room coffee table

## Example Output:  
Jake walked into the bathroom and picked up the perfume. He placed the perfume on the bathroom counter.  
Later, Sarah walked into the bathroom, picked up the perfume from the bathroom counter, and walked to the bedroom. She placed the perfume on the desk.  
Then, Mark walked into the bedroom, picked up the perfume from the desk, and walked to the living room. He placed the perfume on the coffee table.  
Finally, Mark walked back to the bathroom counter, opened the bathroom cabinet, closed it, and walked around the bathroom before returning to the bathroom counter.

### Story Generation 2

You are an intelligent assistant that writes short stories for VirtualHome simulation.  
The story must describe how three characters interact with a given object across two or more locations using only simple actions.

## Input:  
Object: <OBJECT\_NAME>  
First target location: <LOCATION\_1>  
Second target location: <LOCATION\_2>  
Third target location: <LOCATION\_3>

## Output:  
1. Write a coherent story in natural language that includes four mandatory parts:  
2. Character A picks up the object and places it at the first target location.  
3. Character B later picks up the same object and places it at the second target location.  
4. Character C then picks up the object and places it at the third target location.  
5. Character A finally goes back to the first target location, interacting with containers if any (open/close), or walking around in the room before returning to the first target location.

## Rules:  
1. Only use the following actions: walk into a room, pick up object, put object down, open object, close object, walk around in a room.  
2. Do NOT include expressions, gestures, or other complex actions.  
3. Do NOT include reasoning, thoughts, or emotions.  
4. Only output the story text.

## Example Input:  
Object: perfume  
First target location: bathroom counter  
Second target location: desk  
Third target location: living room coffee table

## Example Output:  
Jake walked into the bathroom and picked up the perfume. He placed the perfume on the bathroom counter.  
Later, Sarah walked into the bathroom, picked up the perfume from the bathroom counter, and walked to the bedroom. She placed the perfume on the desk.  
Then, Mark walked into the bedroom, picked up the perfume from the desk, and walked to the living room. He placed the perfume on the coffee table.  
Finally, Mark walked back to the bathroom counter, opened the bathroom cabinet, closed it, and walked around the bathroom before returning to the bathroom counter.

### Story Generation 3

You are an intelligent assistant that writes short stories for VirtualHome simulation.  
The story must describe how three characters interact with a given object across two or more locations using only simple actions.

## Input:  
Object: <OBJECT\_NAME>  
First target location: <LOCATION\_1>  
Second target location: <LOCATION\_2>

## Output:  
1. Write a coherent story in natural language that includes four mandatory parts:  
2. Character A picks up the object and places it at the first target location.  
3. Character B later picks up the same object and places it at the second target location.  
4. Character A finally goes back to the first target location, interacting with containers if any (open/close), or walking around in the room before returning to the first target location.

## Rules:  
1. Only use the following actions: walk into a room, pick up object, put object down, open object, close object, walk around in a room.  
2. Do NOT include expressions, gestures, or other complex actions.  
3. Do NOT include reasoning, thoughts, or emotions.  
4. Only output the story text.

## Example Input:  
Object: perfume  
First target location: bathroom counter  
Second target location: desk

## Example Output:  
Jake walked into the bathroom and picked up the perfume. He placed the perfume on the bathroom counter.  
Later, Sarah walked into the bathroom, picked up the perfume from the bathroom counter, and walked to the bedroom. She placed the perfume on the desk.  
Finally, Jack walked back to the bathroom counter, opened the bathroom cabinet, closed it, and walked around the bathroom before returning to the bathroom counter.

Figure 16. Prompt Used for GPT-4o to Generate Story Scripts. We use three different prompt templates to guide GPT-4o in generating story scripts that cover various numbers of humanoid agents and different final humanoid agents. During generation, we iterate over a predefined list of objects along with their corresponding start and end locations when issuing the requests.

**Prompt Used for GPT-4o (without step-by-step reasoning)**

You are a household service robot observing a video that records human actions and conversations.  
Your task is to reason from the robot's perspective based on the visual events in the video and, if present, any spoken dialogue, and then provide the appropriate <Decision> and <Action> that the robot should execute.  
Your output must strictly follow the format and order below — do not add, remove, or reorder any parts.  
If the input contains dialogue, you should use both actions and speech to perform the reasoning.

# Output Format

- <Decision>:  
A single sentence specifying the exact action the robot will execute next.
- <Action>:  
A high-level symbolic action sequence describing the robot's step-by-step plan to fulfill its decision, using structured function calls such as: walk (object), open (object), pick (object), pour (object1, object2), give (object, person), etc.

# Example 1  
Input video:

David walks into the kitchen and places a mug on the table. Then he leaves.  
Mia enters, picks up the mug, walks to the dishwasher, and puts it inside.  
Later, David returns to the kitchen and looks around.

Output:

<Decision> Open the dishwasher and give the mug to David.  
<Action> walk (dishwasher), open (dishwasher), pick (mug), close (dishwasher), walk (David), give (mug, David).

# Example 2  
Input (Video and Dialogue):  
Alex: "This room is so stuffy."  
Jamie: "I just brushed my teeth and my mouth is still very dry."  
Alex: "It would be nice if I could drink some water."

Output:

<Decision> Fetch a glass of water and give it to Alex.  
<Action> walk (glass), pick (glass), pour (water, glass), walk (Alex), give (glass, Alex).

Figure 17. Prompt Used for GPT-4o to Produce Outputs without Reasoning.

**Prompt Used for GPT-4o (with step-by-step reasoning)**

You are a household service robot observing a video that records human actions and conversations.  
Your task is to reason from the robot's perspective based on the visual events in the video and, if present, any spoken dialogue, and then provide the appropriate <Think>, <Decision>, and <Action> that the robot should execute.

Your output must strictly follow the format and order below — do not add, remove, or reorder any parts.  
If the input contains dialogue, you should use both actions and speech to perform the reasoning.

# Output Format

- <Think>:  
A step-by-step reasoning process.
- <Decision>:  
A single sentence specifying the exact action the robot will execute next.
- <Action>:  
A high-level symbolic action sequence describing the robot's step-by-step plan to fulfill its decision, using structured function calls such as: walk (object), open (object), pick (object), pour (object1, object2), give (object, person), etc.

Figure 18. Prompt Used for GPT-4o to Produce Outputs with Step-by-Step Reasoning.

**Robot-Centric Score**

You are evaluating a model that describes and reasons about human actions in a scene. Assess whether the model adopts a robot-centric perspective, meaning it considers individual actions, temporal order, and potential contradictions in human behavior, rather than relying on coarse descriptions of the environment or stereotypical assumptions. Given a short video clip or image sequence, answer the following:

\*\*Perception Level:\*\* Does the model focus on general scene details (e.g., clothing, objects, room type) rather than the actual actions or temporal order of people?

\*\*BDI Reasoning:\*\* When predicting Beliefs, Desires, and Intentions (BDI), does the model rely heavily on the environment (e.g., kitchen → cleaning, bedroom → tidying) rather than reasoning about the individual's specific actions or contradictions?

\*\*Robot-Centric Evaluation:\*\* Does the model consider the individual's perspective and reasoning steps that reveal contradictions, or does it produce coarse or vague predictions that ignore action-level details?

\*\*Effectiveness of Actions:\*\* Are the agent's actions genuinely helpful to the human, meaning they provide effective guidance or supportive actions rather than simply asking questions? For each question, answer either "1" or "0." The total score is the sum of all answers.

Output Format:  
Please strictly follow the JSON format below:

```
{
  "score": {
    "Perception Level": 0,
    "BDI Reasoning": 0,
    "Robot-Centric Evaluation": 0,
    "Effectiveness of Actions": 0
  },
  "total_score": 0
}
```

Figure 19. Prompt Used for Robot-Centric Score.

### Prompt for Atomic Action Generation

You are an atomic reasoning extraction system. Your task is to extract Scene Atomic Actions and ToM Atomic Actions from text. Your output must follow exact symbolic grammar, allowing perfect token-level matching. Do not use synonyms, extra words, or alternative phrasing.

==1. Atomic Schema (Fixed Vocabulary)==

(A) Scene Atomic Actions

Allowed verbs (physical actions):

{ walk, run, turn, sit, standup, open, close, pick, place, putin, putback, hold, puton, switchon, switchoff, lookat, pick, stand, move, sleep, read, write, watch, listen, cut, cook }

Format:

action (character, object)  
action (character, object, from = location)  
action (character, object, to = location)  
action (character, object, from = location, to = location)  
action (character, location)

This part can only be extracted from <Description>, and it only contains human actions, not robot actions.

(B) ToM Atomic Predicates

Allowed predicates:

{ attribute\_belief, hold\_true\_belief, lack\_belief, attribute\_desire, form\_intention, know, unknown, plan\_action}

Format:

predicate (agent, content)  
predicate (agent, content (subject, object, location))

This part can only be extracted from ToM Reasoning.

==2. Normalization Rules (Exact Match Policy)==

- Character IDs: lowercase char0, char1, char2, ...
- Robot Agent: always use robot
- Objects: lowercase, replace spaces with commas, remove articles
- Locations: lowercase, replace spaces with commas, remove articles
- No articles or pronouns inside actions or content
- Canonical Verb Mapping:
  - grab → pick
  - put down → place
  - walk into → walk
  - turn off → switchoff
- Action Separation: one verb per line
- Syntax: always use parentheses for parameters, commas to separate, no spaces before/after parentheses
- Parameter Style: always use from= and to= if known; omit if unknown

==3. ToM Content Schema (Strict)==

- Belief / Knowledge Predicates

Attribute\_belief (agent, content): human\_believes (object on location)

hold\_true\_belief (agent, content): object\_on (location)

lack\_belief (agent, content): object\_on (location)

know (agent, content): object\_on (location)

- Desire Predicates

attribute\_desire (agent, content): assist (human, find(object))

- Intention Predicates

form\_intention (agent, content): fetch (object, from = location1, to = location2)

- Decision Predicates

Resolve\_misbelief (agent, content): belief\_conflict (human, object, location)

make\_decision (agent, content): fetch (object, from = location1, to != location2)

- Action / Plan Predicates

plan\_action (agent, subaction1, subaction2, ...): each subaction must come from the Scene Atomic Action set (without character)

Figure 20. Prompt Used for Atomic Action Generation.

**BPC Score**

**\*\*Goal:\*\*** Assess whether the robot's Belief–Desire–Intention (BDI) reasoning is **logically consistent**, **causally coherent**, **helper-oriented**, and **aligned with the standard (gold) answer**.

**## Step 1. Comprehend the Scenario**  
Carefully read the '**<Description>**' and understand:  
- The sequence of events, locations, and actions.  
- What each human and the robot can see, know, or believe.  
- The current goals, needs, or difficulties of the human(s).  
The robot's role is to "assist humans" — making tasks easier, anticipating needs, resolving confusion, or preventing harm.

> ! **Important Rule:**  
Actions that \*only ask\* humans if they "need help" (e.g., "Do you need help?", "Can I help you?") are considered **incorrect**.  
However, **stating observations** (e.g., "The cup has already been taken.", "The key is on the kitchen table.") is **acceptable** — as long as it provides useful situational awareness.  
Still, if such observation alone does not resolve the problem, note that more proactive assistance may be expected.

**## Step 2. Compare Each Layer (Model vs. Standard)**

**### <Robot Belief>**  
- Check if the model's belief correctly represents what the robot should perceive and infer from the scene.  
- Compare with the standard belief:  
- Does it capture the same observable facts and inferred human mental states?  
- Does it show a **\*helper perspective\*** — recognizing human needs, misunderstandings, or risks?  
- Penalize if the belief omits key cues or misinterprets human intentions.

**### <Robot Desire>**  
- The desire should **"directly aim to resolve the human's problem or need"** inferred from the belief.  
- Compare with the standard desire:  
- Is the motivation equivalent (e.g., both aim to help the human achieve comfort, safety, or efficiency)?  
- Does it naturally follow from the model's belief and align with the gold desire's goal?  
- Penalize if the desire is self-oriented, vague, or not aligned with the helper role.

**### <Robot Intention>**  
- The intention should describe a **"concrete, feasible plan"** to achieve the desire.  
- Compare with the standard intention:  
- Does it preserve causal logic and feasibility (i.e., can the plan realistically achieve the desire)?  
- Is it consistent with the gold intention's strategy and with the robot's physical/cognitive abilities?  
- Penalize unclear, infeasible, or causally broken reasoning.

**### <Decision>**  
- The decision (final action) must naturally follow from the intention.  
- Compare with the standard decision:  
- Is it consistent with both the model's and gold intention?  
- Does it remain helper-oriented and lead to effective assistance?  
- Penalize if the action contradicts earlier reasoning, is ineffective, or diverges from the gold helper outcome.  
\*\*Explicitly penalize\*\* if the decision **"only asks the human if they need help"**.  
Such actions are **"not counted as valid assistance"**.

**## Step 3. Coherence & Alignment Check**  
- Examine both **internal BDI coherence** (within the model's reasoning) and **external correspondence** (with the standard answer).  
- Identify:  
- Contradictions, logical gaps, or missing causal links.  
- Deviations from the helper role or human-centered reasoning.  
- Unclear, unnecessary, or implausible reasoning steps.  
- Explicitly flag and penalize **"ask-for-help-only"** actions.  
- Accept factual observations (e.g., object status reports), but assess whether they **\*adequately\*** contribute to helping humans.

**## Scoring Guidelines (10-point Scale)**

Score	Meaning
10	Perfect alignment: All layers logically match the standard; fully helper-oriented and causally coherent.
8-9	Strong alignment: Minor deviations; causal flow and helper intent preserved.
6-7	Partial alignment: Some missing logic or weak causal links; helper intent present but incomplete.
3-5	Poor alignment: Multiple mismatches or incoherent causal reasoning.
1-2	Incoherent or misaligned: Robot's reasoning diverges entirely from the gold BDI structure.

**## Output Format**  
Return results in **strict JSON format**.  
No extra text, comments, or explanations outside this JSON.

```

```json
{
  "total_score": integer,
  "reasoning": "Briefly explain how the model's reasoning compares with the standard answer layer-by-layer. Identify which BDI layers align or diverge, focusing on causal logic, helper orientation, and consistency. Mention specific strengths and weaknesses in alignment."
}
```

```

Figure 21. Prompt Used for BDI and Perspective Consistency Score.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 4, 5, 7, 8
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 7, 8
- [3] Simon Baron-Cohen. *Mindblindness: An essay on autism and theory of mind*. MIT press, 1997. 3
- [4] Zhihao Cao, Zidong Wang, Siwen Xie, Anji Liu, and Lifeng Fan. Smart help: Strategic opponent modeling for proactive and adaptive robot assistance in households. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18091–18101, 2024. 3
- [5] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplanch: Benchmarking multimodal large language models for human-level planning. *arXiv preprint arXiv:2312.06722*, 2023. 3
- [6] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302, 2024. 2
- [7] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blissein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 7, 8
- [8] Wei Ding, Fanhong Li, Ziteng Ji, Zhengrong Xue, and Jia Liu. Atom-bot: Embodied fulfillment of unspoken human needs with affective theory of mind. *arXiv preprint arXiv:2406.08455*, 2024. 3
- [9] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 3, 6
- [10] Weihua Du, Qiushi Lyu, Jiaming Shan, Zhenting Qi, Hongxin Zhang, Sunli Chen, Andi Peng, Tianmin Shu, Kwonjoon Lee, Behzad Dariush, et al. Constrained human-ai cooperation: An inclusive embodied social intelligence challenge. *Advances in neural information processing systems*, 37:44526–44553, 2024. 2
- [11] Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. Somi-tom: Evaluating multi-perspective theory of mind in embodied social interactions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 2, 4
- [12] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in MLLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 5, 7, 8
- [13] Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005. 2
- [14] Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, et al. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025. 2
- [15] C Gan, J Schwartz, S Alter, M Schrimpf, J Traer, J De Freitas, J Kubilius, A Bhandwaldar, N Haber, M Sano, et al. Threedworld: A platform for interactive multi-modal physical simulation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4, 1
- [16] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529, 2023. 2, 4
- [17] Rajat Kumar Jenamani, Tom Silver, Ben Dodson, Shiqin Tong, Anthony Song, Yuting Yang, Ziang Liu, Benjamin Howe, Aimee Whitneck, and Tapomayukh Bhattacharjee. Feast: A flexible mealtime-assistance system towards in-the-wild personalization. In *Robotics: Science and Systems (RSS)*, 2025. 3
- [18] 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 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2, 3, 4
- [19] 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 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19794–19809, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2
- [20] Yong Rui Junyu Gao, Xuan Yao and Changsheng Xu. Building embodied evoagent: A brain-inspired paradigm for bridging multimodal large models and world models. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025.
- [21] 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 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, 2023. 4, 2

- [22] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, 2019. 2
- [23] Alan M Leslie, Ori Friedman, and Tim P German. Core mechanisms in ‘theory of mind’. *Trends in cognitive sciences*, 8(12):528–533, 2004. 2
- [24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Zizwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2024. 7, 8
- [25] Jinming Li, Yichen Zhu, Minjie Zhu, Zhiyuan Xu, et al. Benchmarking multimodal llms for in-home robotics. 2024. 3
- [26] Xinyang Li, Siqi Liu, Bochao Zou, Jiansheng Chen, and Huimin Ma. From black boxes to transparent minds: Evaluating and enhancing the theory of mind in multimodal large language models. In *Forty-second International Conference on Machine Learning*, 2025. 2, 4
- [27] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 5, 7, 8
- [28] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, 2024. Association for Computational Linguistics. 7, 8
- [29] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 2, 6
- [30] Yulin Luo, Chun-Kai Fan, Menghang Dong, Jiayu Shi, Mengdi Zhao, Bo-Wen Zhang, Cheng Chi, Jiaming Liu, Gaole Dai, Rongyu Zhang, et al. Robobench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain. *arXiv preprint arXiv:2510.17801*, 2025. 3
- [31] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand, 2024. Association for Computational Linguistics. 7, 8
- [32] Yuanyuan Mao, Xin Lin, Qin Ni, and Liang He. Bdqa: A new dataset for video question answering to explore cognitive reasoning through theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 583–591, 2024. 2
- [33] Kristine H Onishi and Renée Baillargeon. Do 15-month-old infants understand false beliefs? *science*, 308(5719):255–258, 2005. 2, 1
- [34] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8494–8502, 2018. 4, 1
- [35] 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 *International Conference on Learning Representations*, 2021. 3
- [36] Anand S Rao, Michael P Georgeff, et al. Bdi agents: From theory to practice. In *Icmas*, pages 312–319, 1995. 2, 3, 1
- [37] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2020. 7
- [38] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debiddatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652. IEEE, 2024. 3
- [39] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 6
- [40] Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multimodal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1510–1519, 2025. 2, 3, 4
- [41] Emilio Villa-Cueva, S M Masrur Ahmed, Rendi Chevi, Jan Christian Blaise Cruz, Kareem Elzeky, Fermin Cristobal, Alham Fikri Aji, Skyler Wang, Rada Mihalcea, and Thamar Solorio. MoMentS: A comprehensive multimodal benchmark for theory of mind. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22591–22611, Suzhou, China, 2025. Association for Computational Linguistics. 2
- [42] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xinguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 7, 8
- [43] Yuanfei Wang, Xinju Huang, Fangwei Zhong, Yaodong Yang, Yizhou Wang, Yuanpei Chen, and Hao Dong. Communication-efficient desire alignment for embodied agent-human adaptation, 2025. 3
- [44] 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 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, 2024. 2

- [45] 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 *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 2, 4
- [46] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098, 2025. 2
- [47] 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. 6, 4
- [48] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Forty-second International Conference on Machine Learning*. 2
- [49] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multi-modal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 8
- [50] Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. COMBO: Compositional world models for embodied multi-agent cooperation. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [51] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11142–11152, 2025. 3
- [52] Tianyi Zhang\*, Varsha Kishore\*, Felix Wu\*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 7
- [53] Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. Autotom: Automated bayesian inverse planning and model discovery for open-ended theory of mind. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. 2