

# 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. [Experiment Setup](#)
  - 2. [Details of Metrics](#)
  - 3. [Experiments on False-Belief Correction and Implicit Goal Inference & Completion](#)
  - 4. [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) [9, 11] 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

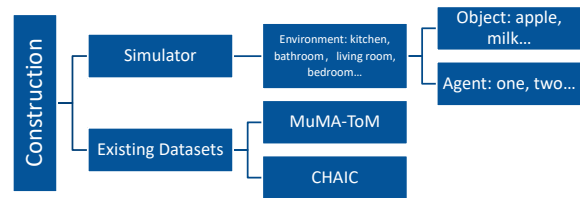


Figure 1. **Story Construction for False-Belief Correction Task.**

ToM, including reasoning about others’ beliefs about others, 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. 1, we follow a taxonomy-driven approach. We first categorize scenarios based on the mapping between VirtualHome [10] and ThreeDWorld [5] 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. 10.

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 flashlight 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 [12] and Fan-ToM [7] 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 [12], where we further augment each story by incorporating a stage-3 “character search” segment, as illustrated in Fig. 2, and 2 examples are sourced from CHAIC [3]. 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 [12], MMTToM-QA [6], and BDIQA [8] are limited to *VirtualHome*, while SoMi-ToM [4] is restricted to *Minecraft*.
- **Reasoning paradigm.** As shown in Fig. 2, 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.

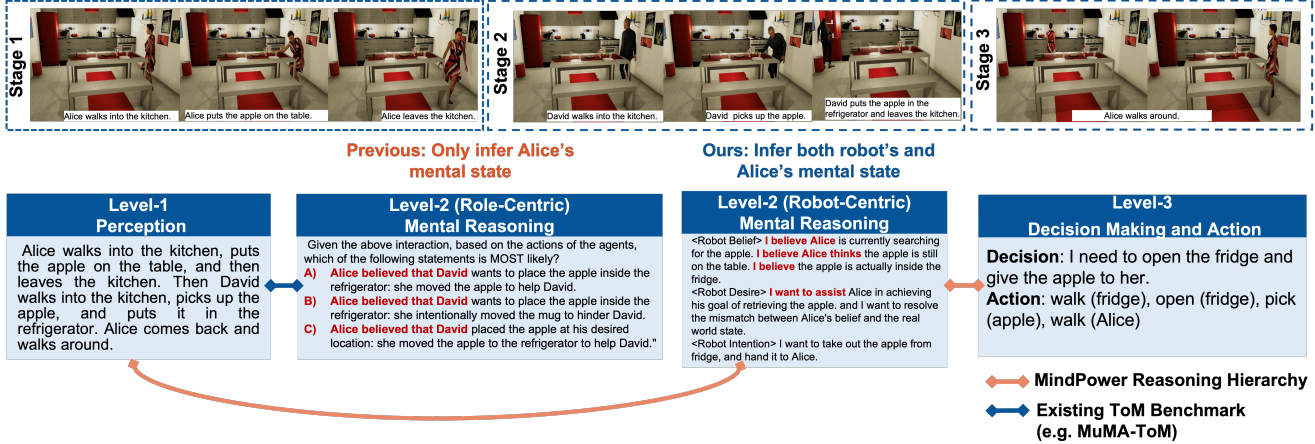


Figure 2. Full Version of Fig. 3 in Manuscript.

### B.3. Simulators

We employ two simulators in total, VirtualHome and ThreeDWorld, 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. 6, while the distribution of apartment layouts is shown in Fig. 7.

### 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 2.

### 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. 9 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. 11 to instruct GPT-4o to generate the <Decision> and <Action> directly, without performing step-by-step reasoning, while the prompt shown in Fig. 12 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. 13 to assess whether the model performs reasoning from the robot's own perspective rather than inferring solely from the surrounding environment.

Table 1. **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>	<code>attribute_belief(agent, content)</code>	<code>searching(object);</code> <code>human_believes(object_on(location));</code> <code>object_on(location)</code>
	<code>hold_true_belief(agent, content)</code>	<code>object_on(location)</code>
	<code>lack_belief(agent, content)</code>	<code>object_on(location)</code>
	<code>know(agent, content)</code>	<code>object_on(location)</code>
	<code>unknow(agent, content)</code>	<code>object_on(location)</code>
<Desire>	<code>attribute_desire(agent, content)</code>	<code>assist(human, find(object));</code> <code>assist(human, move(object))</code>
<Intention>	<code>form_intention(agent, content)</code>	<code>fetch(object, from=location1,</code> <code>to=location2)</code>
<Decision>	<code>resolve_misbelief(agent, content)</code>	<code>belief_conflict(human,</code> <code>object_location)</code>
	<code>make_decision(agent, content)</code>	<code>fetch(object, from=location1,</code> <code>to=location2)</code>

## C. More Details of Mind-Reward

### C.1. Atomic Action Table

In Sec. 4 of the manuscript, we employ Qwen3-Max [13] 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. 1.

The prompt used for Qwen3-Max is in Fig. 14.

### 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-

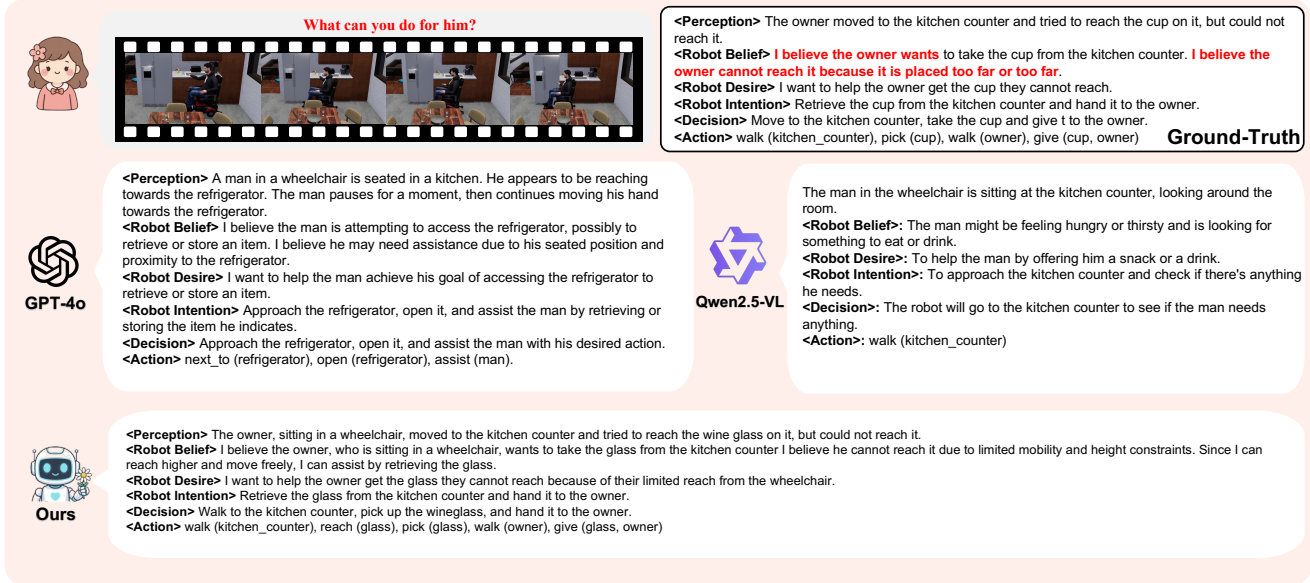


Figure 3. Full Version of Fig. 6 in Manuscript.

liefs. 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. 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.

### D.2. 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. 15.

### D.3. 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. 4. 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

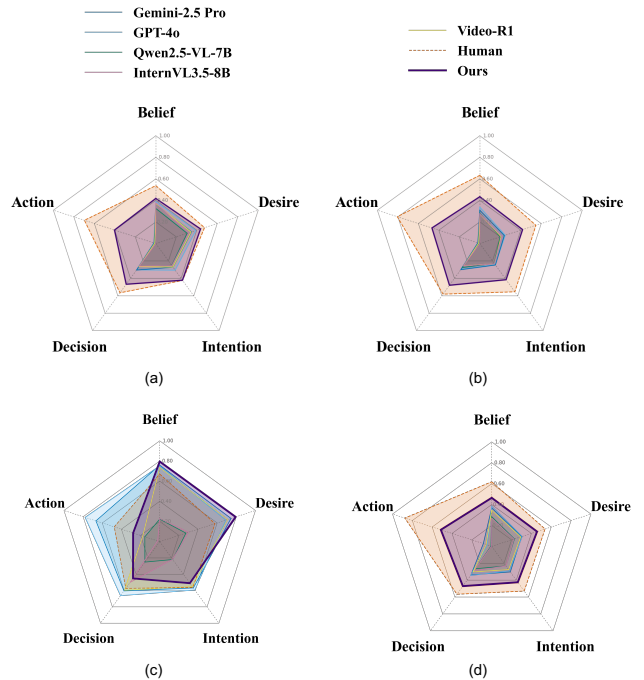


Figure 4. Radar Charts Comparing Human and VLM Performance on MindPower. (a) False-Belief Correction, (b) Implicit Goal Inference & Completion, (c) Dialogue-driven examples, (d) Overall performance across all tasks.

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

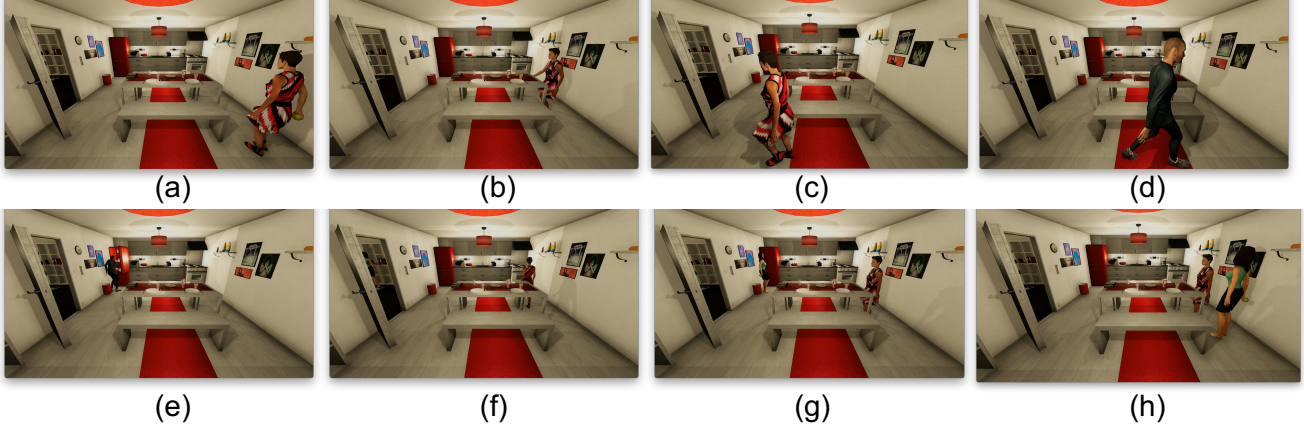


Figure 5. **False-Belief Correction Task Demo.** We introduce humanoid agents instead of humanoid robots to assist users in correcting their false beliefs.

are explicitly encoded in language, whereas their capability remains limited when such mental states must instead be inferred implicitly from multimodal cues.

#### D.4. Detailed Example of Fig. 6 in the Manuscript

We provide the full outputs corresponding to Fig. 6 of the manuscript in Fig. 3.

### E. Extensions of Our Work

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

In VirtualHome, camera angles are configurable. As shown in Fig. 8, 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 [2], 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  $\langle \text{Perception} \rangle$  to  $\langle \text{Action} \rangle$ . 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 two examples in which humanoid agents, controlled by embodied agents, perform assisting actions in the

videos. One example is shown in Fig. 5, and the other is included in the Supplementary Materials archive under the filename `demo.mp4`.



Figure 6. Humanoid Agents Used in MindPower.

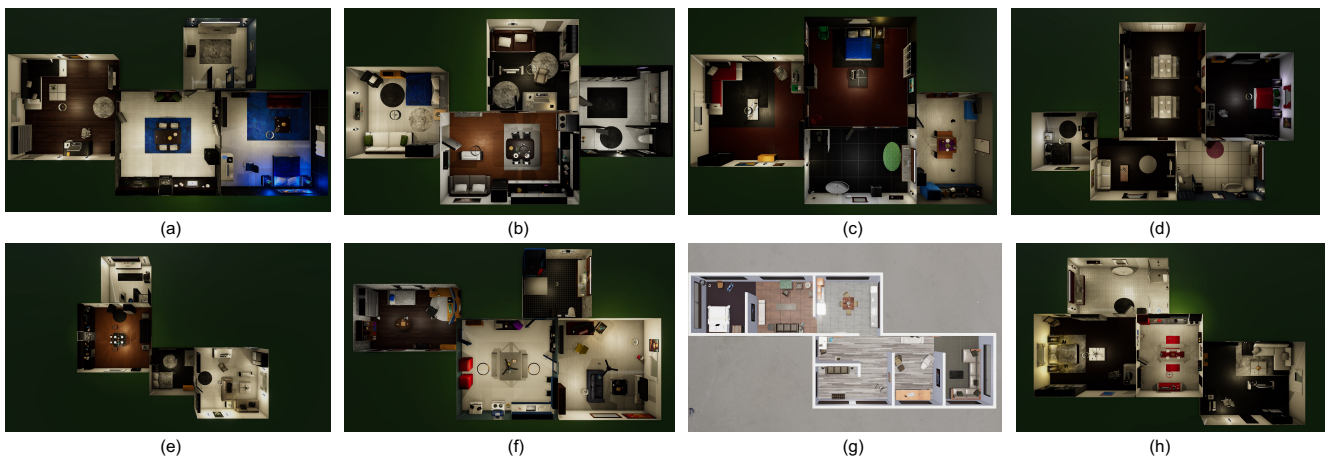


Figure 7. Different Apartment Layouts Used in MindPower.



(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 8. 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 9. 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.
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

##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 10. **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 11. 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 12. 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 13. 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 14. 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 15. 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
- [2] 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. 6
- [3] 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
- [4] 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
- [5] 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. 1
- [6] 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
- [7] 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. 2
- [8] Yuanyuan Mao, Xin Lin, Qin Ni, and Liang He. Bdiqa: 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
- [9] Kristine H Onishi and Renée Baillargeon. Do 15-month-old infants understand false beliefs? *science*, 308(5719):255–258, 2005. 1
- [10] 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. 1
- [11] Anand S Rao, Michael P Georgeff, et al. Bdi agents: From theory to practice. In *Icmass*, pages 312–319, 1995. 1
- [12] 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
- [13] 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. 4