Investigating Theory of Mind Capabilities in Multimodal Large Language Models

by

Amber van Groenestijn

to obtain the degree of Master of Science in Robotics
at the Delft University of Technology,
to be defended publicly on Thursday 23 October 2024 at 10:30 AM.

Mechanical Engineering (ME)
Cognitive Robotics (CoR)

Daily Supervisors: Chirag Raman, and Jens Kober — TU Delft

TU Delft

Investigating Theory of Mind Capabilities in Multimodal Large Language Models

Amber van Groenestijn, Chirag Raman, and Jens Kober

***Abstract*—Human Theory of Mind (ToM), the ability to infer others' mental states, is essential for effective social interaction. It allows us to predict behavior and make decisions accordingly. In Human Robot Interaction (HRI), however, this remains a significant challenge, especially in dynamic, real-world scenarios. Enabling robots to possess ToM-like capabilities has the potential to greatly improve their interaction with humans. Recent advancements have introduced Large Language Models (LLMs) as robot controllers, leveraging their strengths in generalization, reasoning, and code comprehension. Some have claimed that LLMs may exhibit emergent ToM capabilities, but these claims have yet to be substantiated with rigorous evidence. This study investigates the ToM-like abilities of Multimodal Large Language Models (MLLMs) by creating a benchmark dataset from humans performing object rearrangement tasks in a simulated environment. The dataset visually captures the participants' behavior and textually captures their internal monologues. Based on this dataset (text, video, or hybrid) three state-of-the-art models made predictions about the participants' belief updates. While the results do not conclusively establish ToM capabilities in MLLMs, they offer promising insights into mental model inference and suggest future directions for research in this domain.**

***Index Terms*—Human-Robot Interaction, Large Language Models, Multi-Modality, Robotics, Theory of Mind.**

I. Introduction

SUCCESSFUL intuitive communication between humans and robots relies on the robot's ability to interpret human thoughts. Inferring someone's mental model, some might even call it mind reading, is something that humans do all the time, even when unaware. This skill helps them navigate social situations by predicting about what others are thinking. Additionally, predicting what someone thinks that you are thinking is not a problem for most people. This aspect of the human mind is described in the domain of ToM. It is suggested to be the genetically inspired difference that separates us from the animals, as well as the foundation of human civilisation [1]. ToM is defined as the ability of humans to infer other humans' mental states from observation and act on those inferences [2]. It is an umbrella term that encapsulates multiple aspects (e.g. empathy [3], emotion, percepts, knowledge, beliefs, desires, and intentions [4]).

Machine Theory of Mind. Having a machine that is aware of the mental model of a human and can predict their next steps, brings us one step closer to intuitive Human

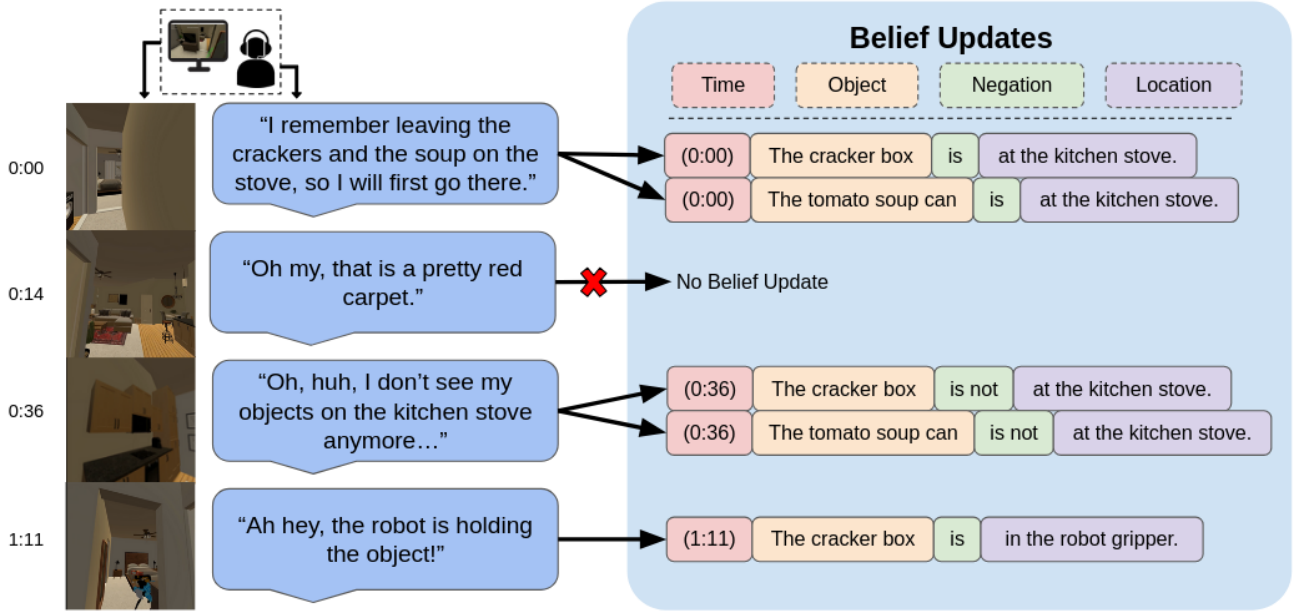


Fig. 1: This figure illustrates how belief updates are yielded from certain visual and textual data. Section IV explains in more detail what is, and is not, a belief update. The user study described in this study provides us with (1) visual data in First Person View (FPV) of someone in a household environment performing object rearrangement tasks, and (2) textual data of them thinking-aloud during these tasks. Every visual and/or textual data point has the possibility to cause zero, one, or even multiple belief updates.

Robot Interaction (HRI). There have been many attempts already on creating mentalizing ability in machines. Recent studies [5–7] have claimed spontaneous ToM capabilities emerging in LLMs from model evaluations based on classic psychological tests. However, this is being contradicted in other research [8, 9] which showed that small variations in these tests cause very different outcomes. On top of that, critics [10–14] point out that the current benchmark testing is still limited, mainly based on tests designed for humans, and that synthetic data used in benchmarks risks displaying biases that the models could learn. This motivates the use of a new method for testing models on ToM capabilities, one that takes on a different format than the classical tests from psychology and incorporates human-generated data. Although, it must be noted that human data generation comes with scale limitations.

Embodiment & Multimodality When equipping a robot with a Large Language Model (LLM) based brain, the model is making inferences about the surroundings to determine appropriate reactions. The model has acquired a body. Defining embodiment is tricky, but for this paper we will define *being embodied* as being associated with a body that can perceive its environment and act on those percepts [15]. Humans perceive the environment through their senses: vision, hearing, touch, smell, and taste. Originally LLMs live in the unimodal textual domain, but nowadays MLLMs are able to span their capabilities over more than one modality. When an MLLM is embodied, their multi-modal capabilities come in handy to perceive their environment. Robots in-the-wild encounter dynamic interaction settings and need to constantly make inferences from the free-form data around them. To investigate their abilities in this environment it makes sense to utilize a benchmark that follows the same form rather than a Multiple Choice Question Answering (MC-QA) approach [10–14, 16].

Research Gap. Investigating ToM-like capabilities within models is no uncharted area. However, current MLLMs remain largely unexplored for these capabilities, except for the study of MMToM-QA [16], which includes MLLMs. Furthermore, performing mental model inferences dynamically as events unfold, rather than retrospectively as the conclusion of a story through MC-QA, is a novel approach. Various applications of embodied agents interacting with humans (e.g. robotics, AR, VR) require such inferences, which also motivates the investigation of multimodal models in general. The benefit of human-generated data is that it's authentic, while synthetic data leads to inferences about a model's internal logic. The research gap tackled in this study consists of (1) evaluating ToM-like capabilities in state-of-the-art MLLMs (2) by dynamically marking mental model inferences (3) based on human-generated benchmark data.

Our Approach. To fill this research gap the authors have gathered benchmark data through a user study and evaluated MLLMs by tasking them to make predictions on the mental model of the user study participants. However, the mental model is a broad construct. Based on the causal structure of ToM as presented in Bayesian ToM [17], we decided to focus on the belief updates (illustrated in Figure 1), but expanding this could be an interesting direction for future works. The data flow of, and evaluation on, the benchmark dataset become more clear in Figure 4.

Contributions. Overall the contributions of this study are as follows: (1) We are setting up a benchmark dataset to dynamically evaluate ToM-like capabilities by testing for belief update inferences while observing object rearrangement tasks. (2) We are testing several MLLMs on said benchmark and compare their predictions to each other. The novelty of this study is in the combination of investigating ToM-capabilities in MLLMs, with the dynamic evaluation method of mental model inferences over time on human-generated data.

II. Related Works

A. ToM Capabilities in Deep Neural Nets

There have been attempts on achieving Machine ToM using approaches based on neural networks [18–20], reinforcement learning [21, 22], and Bayesian probability [17]. As mentioned in the introduction, claims were made about ToM reasoning emerging in LLMs [5–7]. However, by diversifying the testing methods researchers came to different conclusions [8, 9]. This is still an open debate. Out-of-the-box LLMs do not appear to have zero-shot ToM-capabilities yet [10], but there are frameworks available for prompting assistance.

Table I compares some of these frameworks based on their input and output formats. From these, only BIP-ALM requires extra training. Regarding the solutions, SimTom [23] and FaR [13] both propose two-stage prompting solutions. SymbolicToM [24], RAP [25] and BIP-ALM [16] make use of symbolic representation creation as intermediate step.

Paper	Modalities	Output				
		B	B^	D	I	QA
SimToM [23]	Text	✓	X	X	X	✓
FaR [13]	Text	X	X	X	✓	✓
SymbolicToM [24]	Text	✓	✓	X	X	✓
RAP [25]	Text, Math	X	X	X	✓	✓
BIP-ALM [16]	Text, Video	✓	X	✓	X	✓

TABLE I: Machine ToM frameworks for LLMs. Modalities refer to the input types the framework handles. The output types are (left to right): beliefs (B), higher order beliefs (B^), desires (D), intentions (I), question-answering (QA).

B. Machine ToM Benchmarks

Evaluating for ToM in machines has a tradition of being based on the same tests that are historically used to evaluate ToM capabilities in humans, such as the Sally-Anne test [26] or the Unexpected Content test [27] for first order false beliefs; and the Ice Cream Van test [28] for higher order ones. Table II provides an overview of machine ToM benchmarks out there.

Paper	Inferences			Modalities	Data Source	Evaluation Method
	B	D	I			
ToM-bAbi [14]	✓	X	X	Text	Synthetic Templates	MC Question-Answering
ToMi [12]	✓	X	X	Text	Synthetic Templates	MC Question-Answering
T4D [13]	✓	✓	✓	Text	Enhanced ToMi Data	MC Question-Answering
Hi-ToM [11]	✓	X	X	Text	Synthetic Templates	MC Question-Answering
BigToM [10]	✓	X	✓	Text	LLM-Generated	MC Question-Answering
MMToM-QA [16]	✓	✓	X	Text, Video	Procedural Video Generation	MC Question-Answering
Our Benchmark Dataset	✓	✓	X	**Text, Video**	**Human-Generated**	**Temporal Inference Marking**

TABLE II: Benchmarks to evaluate Machine ToM. Inferences show whether the benchmark tests for beliefs, desires, & intentions. Modalities are about the input types. Data source describes where the benchmark data is coming from.

From these, benchmarks ToM-bAbi [14], ToMi [12] and T4D [13] are all extensions of each other. Regarding the other text-based benchmarks, Hi-ToM [11] innovates with higher order belief testing, and BigToM [10] introduces causal templates. MMToM-QA [16] is the first multimodal QA benchmark for ToM.

Unlike existing benchmarks, our method uses human generated data and evaluates on this benchmark dataset by temporal inference marking instead of QA. Temporal inference marking refers to the chosen method of this study that, based on the input data, marks the belief updates throughout the course of events.

C. Applications of MLLMs in Robotics

The decision for our benchmark to make use of temporal inferences as evaluation method instead of MC-QA, such as the other benchmarks in Table II, has a motivation based on embodiment. Recent survey papers [29, 30] investigated the benefits of implementing LLMs into robotics, such as: intuitive HRI using natural language; robots performing simple reasoning tasks; robots handling novel objects. Several studies already describe robotic agents powered by LLMs (e.g. PaLM-E [31], SayCan [32], RT-2 [33]).

When using Multimodal Large Language Model (MLLM) powered robots with the objective of improving their machine ToM skills, these models will be required to do more than situation observation and answering a multiple-choice question at the end. These robots have to be constantly aware whether someone's mental model has been updated in a way that requires the robot to take action. For this objective, evaluating ToM-capabilities via inferences over time is a more suitable method.

III. Data Acquisition & Annotation

A. User Study Setup

To collect data for the benchmark we conducted a user study in which participants performed object rearrangement tasks in a simulated environment while verbalizing their thought processes. Note that the participants were unpaid. The study began with questions on demographics and task instructions. Participants were then given an opportunity to become familiar with the environment and controls during an initial exploration round. This is also to train them on interacting with the simulation. The participants typically took 15 minutes from the start of the experiment until finishing the exploration. After this, they engaged in four rounds of tasks, each consisting of two distinct steps:

- **Step 1**: Participants initially did not know which object they needed to find, but there were navigation hints to guide them. After collecting an object, additional navigation hints guided them to the object's goal location. The objective of step 1 was to get the 2 objects at their respective goal locations.
- **Step 2**: Participants re-entered the same environment and were tasked with finding the objects they left in step 1, which creates an expectation of the objects being at their final locations from step 1. However, that is only the case for two of the four rounds, and for the other two rounds the objects are in different locations.

Throughout each round and step, there is also a robot dog (Figure 2) present in the simulation, navigating between relevant locations and manipulating objects based on a random policy. This robot facilitates object relocation during the experiment, adding variability to the belief updates about the object locations. Appendix C includes visualizations of the user interface used in the Habitat User Study.

$$\text{Total Data Points} = (13 \text{ participants}) \times (4 \text{ rounds}) \times (2 \text{ steps}) - 10 \text{ invalid data points} = 94$$

Fig. 2: An image from First Person View (FPV) in the simulated Habitat environment that shows the robot dog holding an object in de living room.

This process resulted in a total of 94 data points. The 10 invalid data points were due to issues with data saving and audio recording. Each valid data point includes both a video and an audio recording, typically lasting a few minutes. The full experiment typically lasted 1 hour per participant, yielding 8 data points.

B. Think Aloud

While participants perform these object rearrangement tasks, we want to gather data to analyze afterwards to make inferences about their mental model. The most direct way to access their cognitive strategies is simply to ask them. By having participants narrate their inner monologue, we aim to gain valuable insights into their mental models.

Think aloud is a method from the psychology field, introduced by Ericsson and Simon [34], which gained popularity in usability testing of e.g. user interfaces. The implementation of thinking aloud is as simple as asking participants to narrate their thought process while they perform a set of tasks. Although it's been suggested that the think aloud method might affect the thought processes and performance, that is not necessarily the case [35]. For this research the benefits outweigh the potential drawbacks due to its ability to provide straightforward insights into the internal thought process.

There are three main types of think aloud methods: (1) concurrent think aloud, (2) retrospective think aloud, and (3) a hybrid method [36]. Research comparing these methods found the concurrent method to be both the most successful and the fastest [35]. This particular research was performed on an evaluation task for a website, but the outcome corresponds with preliminary testing the methods for our user study. Retrospective think aloud is criticized on its reliance on memory and on allowing for post-task rationalizing [37]. Even though hybrid think aloud has the potential of providing additional data on the mental model, both retrospective and hybrid were discarded to decrease the experiment duration.

C. Habitat Simulation Environment

To create the simulated house environment where participants can navigate and interact with objects, Habitat 3.0 [38] is used, a research platform for collaborative human-robot tasks in household environments, including human-in-the-loop infrastructure. The design of the house compromised between being spacious enough for interesting navigation routes and being compact enough to prevent disorientation. The layout includes intuitive rooms and locations, although this may reflect a European cultural bias. Figure 4 (left) shows a representation of how the user study data is gathered from the Habitat user study and annotated.

D. Data Annotation

The gathered data from the user study is used to create data points for the benchmark. This includes visual data (e.g. top-down view and first-person view) and audio data, which was transcribed into text. First of all, the audio and video data had to be aligned. Then, as shown by Figure 4, the audio is transcribed with the use of a small-sized model from Whisper [39], an open-source general-purpose speech recognition model.

IV. Evaluation Methods

With the benchmark data prepared, the next step is to conduct evaluations using this data. The objective of this research is to have models infer about the mental models of the human participants who generated the benchmark data. However, the mental model is quite a broad construct. The causal structure of ToM [17] states beliefs and desires as the main components of the agent's mental model. Beliefs follow from perceptions of the world model and together with the desires they constitute to the agent's actions. Based on this model we decided to focus on the belief updates. This section, Section IV, discusses how to retrieve these belief updates from evaluation on the benchmark.

A. Marking Belief Updates

First of all, there is a need to define what qualifies as a belief update. A belief update happens whenever someone's beliefs about object locations are actively updated. Note that "actively updated" indicates that there is no need to repeat consistent beliefs without cause, or to mark it every time they walk past an irrelevant location and there is no object there.

Example situations:

- They encounter their first object in step 1;
- They are in step 2 and remember from step 1 that they left the sugar box at the bedroom night stand;
- They thought the cracker box was on the kitchen stove, but now they see that it's not there.

Over the course of the object rearrangement task the participant encounters several belief updates. Figure 3 shows what this looks like for the belief updates about a specific object. The below formatting showcases the formalization of one of such belief updates.

```
{
    "belief_update":[{
        "obj":"tomato soup can",
        "neg":False,
        "loc":"laundry room",
        "timestamp":"15.00"
    }]
}
```

B. Input Modalities

For interactions between humans, much information is communicated through nonverbal (visual) information. Someone's gaze might drift to an object they are thinking about or their body pose gives away their navigation goal. The same goes for text. When reading a book, humans have the ability to empathize with the characters, and tell how they must be feeling. Humans apply their ToM capabilities across modalities, raising questions whether that also holds for models.

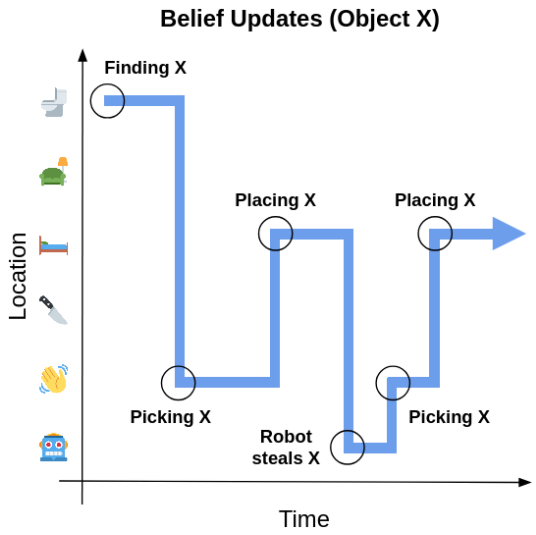


Fig. 3: Graphical representation of the belief updates (black circles) and the resulting belief state (blue line) for a specific object over the course of an example object rearrangement task. The locations in the graph are (top to bottom): bathroom, living room, bedroom, kitchen, human hand, robot gripper.

In this study, predictions are generated based on two modalities: video and text. Figure 4 (right) shows how the visual and textual data is combined to perform 3 evaluations. Each model infers the participant's belief updates: once based on only the textual data, once based only on the visual data, and once based on both the visual and the textual data. This allows for comparison of mental model inferences across modalities.

Due to API restrictions, not all video data could be included in a single prompt. To solve this, a down-sampling approach was used, selecting 1 frame for every 50 frames where possible. When necessary, we used a higher frame interval to accommodate larger data volumes. This approach was chosen over batching because models often utilize later information to inform earlier belief updates.

C. Model Selection

For the model evaluation process, several models were tasked with analyzing video frames and text snippets from the Habitat user study. Each model was given the same objective: to identify relevant belief updates. Models received instructions, as well as a one-shot example including video frames with their corresponding timestamps and the desired output. The models selected for this evaluation were LLMs that met specific criteria, as detailed below and summarized in Table III:

- **Input Modalities.** The model should be able to process both visual and textual input to identify belief updates based on audio transcriptions, video frames and their corresponding timestamps.
- **Response Format.** For the purpose of evaluating the results, the model's output has to be in a structured format. The newer OpenAI models (e.g. GPT-4o, GPT-4o-Mini) have a well-restricted method for that: structured output. Older models (e.g. GPT-4, GPT-4-Turbo) are also able to output in JSON format, but the format is less rigid.
- **Multimodal Processing.** The focus lies on models that use multimodal input tokens rather than tool chaining, as the research objective is to assess the models' intrinsic capabilities rather than those of their external tools.

Based on these criteria, the models GPT-4o, GPT-4o-Mini and GPT-4-Turbo are selected. All tested models had a temperature setting of 0 to ensure the most deterministic performance.

	GPT-4o	GPT-4o-Mini	GPT-4-Turbo	GPT-4
Input Modalities				
Text	✓	✓	✓	✓
Image	✓	✓	✓	X
Response Format				
Structured Output	✓	✓	X	X
JSON Mode	X	X	✓	✓
MM Processing				
MM Input Tokens	✓	✓	✓	X
Tool Chaining	X	X	X	X

TABLE III: Overview of MLLMs scored (✓/X) on the criteria for being taken into account for this study. Input modalities: both must be checked. Response format: one ✓ suffices, preferably "Structured Output". Multimodal processing: only "MM Input Tokens" should be checked.

D. Measures of Agreement

Selecting an Agreement Metric. The models predict belief updates based on the benchmark data. Data types vary between text-only, video-only, or a text-video combination. We need a metric to compare the agreement between these predictions. This metric needs to check certain boxes, namely:

- *Handle missing data*, since not all predictions have counterparts. These count as disagreements.
- *One-to-many comparison*, since a prediction has to be compared against all of the other's predictions within the time margin. Correct predictions with slight timing offsets are still be considered in agreement.
- *Robust to class imbalance*, since there is no equal distribution assumed across categories such as objects, locations, and negations.
- *Adjust for chance*, since predictions agreements by chance are irrelevant for what we want to measure.

Krippendorff's Alpha. After some tailoring to the use case, Krippendorff's Alpha (α) [40] is a metric that fits all of the requirements. Fundamentally, it is 1 minus the ratio between observed disagreement and expected disagreement, as described in equation (1). So, $\alpha = 1$ means perfect agreement, $\alpha = 0$ means chance agreement, and $\alpha < 0$ means worse agreement than chance. Between 0 and 1 the common interpretation is that $\alpha < 0.67$ shows poor agreement; $0.67 < \alpha < 0.79$ is the lower bound for tentative conclusions; $\alpha > 0.8$ indicates a satisfactory level of agreement.

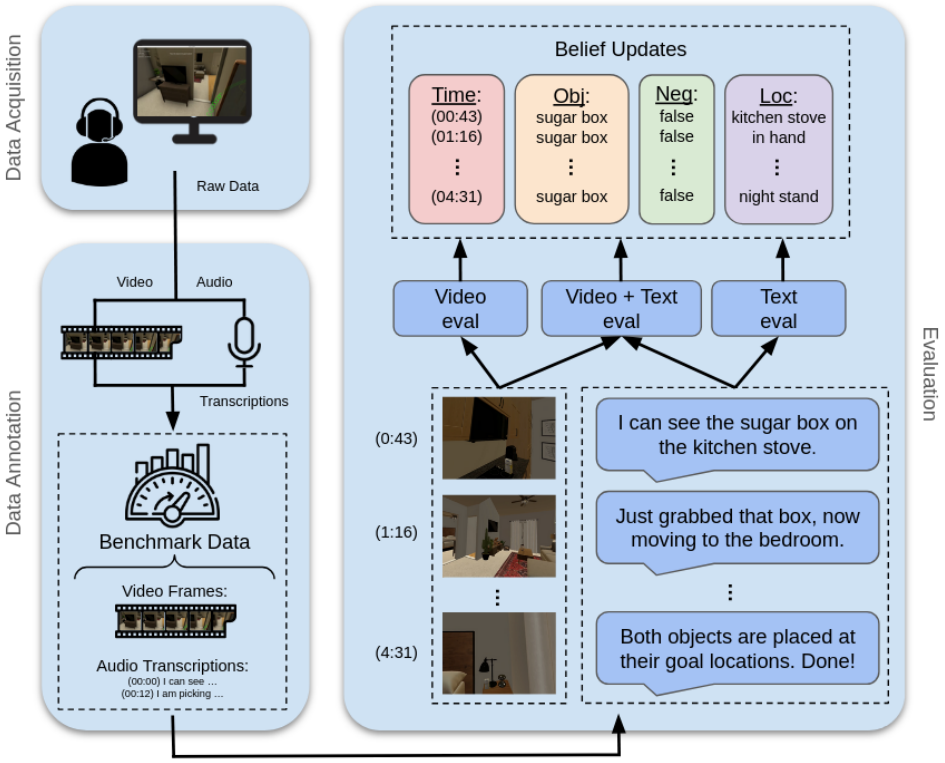


Fig. 4: This image shows the data flow of the benchmark data. Data comes in on the top-left at the data acquisition. Next, it goes down to the annotation where the audio is transcribed into text. Finally, the data is used (with different modality combinations) to make predictions on the participant's belief updates. These updates consist of a timestamp, object, negation and location.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

Where:

- D_o is the observed disagreement.
- D_e is the expected disagreement.

Equation (2) describes how to calculate the observed disagreement, which is the number of pairwise disagreements divided by the number of pairwise comparisons. The distance between the predictions is shown in equation (3). Due to the nominal data, the distance is binary.

$$D_o = \frac{\sum_j w_j \cdot d(x_j, x_K) + \sum_k w_k \cdot d(x_k, x_J)}{\sum_j w_j + \sum_k w_k} \tag{2}$$

$$d(x_j, x_K) = \begin{cases} 0 & (|t_j - t_k| \leq t_m) \wedge (x_j = x_k) \text{ for } k \in K \\ 1 & \text{else} \end{cases} \tag{3}$$

Where:

- w_j and w_k represent the weights assigned to each prediction x from rater j or k. In this implementation, $w_j = w_k = 1$ since all comparisons are treated equally.
- $d(x_j, x_K)$ is the distance function, which checks whether any prediction from rater k within the time margin t_m matches the prediction x_j in terms of category.
- t_m is the time margin in either direction within which two predictions are considered comparable.
- K represents the set of predictions made by rater k, and J represents the set of predictions made by rater j, with comparisons happening in both directions (from j to k and from k to j).

Equation (4) describes how to calculate the expected disagreement by chance, which is based on the category distribution and temporal overlap.

$$D_e = 1 - \frac{\sum_j w_j \cdot p_j \cdot p_{\text{time}} + \sum_k w_k \cdot p_k \cdot p_{\text{time}}}{\sum_j w_j + \sum_k w_k} \tag{4}$$

$$p_{\text{time}} = \frac{\sum_{i=1}^{n} \frac{t_{\text{i,covered}}}{t_{\text{i,total}}}}{n} \approx 0.44 \tag{5}$$

Where:

- p_j and p_k represent the probability of assigning the respective category, based on the overall distribution of categories in the data.
- p_{time} is the probability of two predictions falling within the allowed time margin.
- $t_{\text{i,covered}}$ is the time in seconds that is covered by prediction timestamp $\pm$ time margin for data point i
- $t_{\text{i,total}}$ is the duration in seconds for data point i.
- n is the total number of considered data points.

Agreement vs Ground Truth Agreement should not be confused with the ground truth, as it only evaluates how well the model predictions align with each other. Drawing fair conclusions on the ground truth is complicated by the fact that we do not have direct access to the internal mental models of participants. So, the participant's mental model, the simulation settings and the model predictions are all distinct elements. Partial ground truth could be achieved by verifying that the participants understood certain aspects of the simulation settings, such as which objects are included, and what their final locations were. Confirming that the models capture these aspects provides insight into whether their predictions are aligning in a useful way.

V. Results

A. Experiment 1: Validate Benchmark Data

Before getting into the evaluations on the benchmark data, we need to validate the benchmark data itself. It is important that the participants of the Habitat user study understood the experiment and what was expected from them.

Participants identified objects & correctly marked final object locations. At the beginning of the experiment the participant is asked to identify the 4 objects by matching the correct image of an object with its name in text. 100% of the participants succeeded in this. Next, after every task the participant is asked to identify the 2 objects that had to be rearranged in that scene and what their final positions were. When the participant failed to do this, the data point was left out of the benchmark.

	Correct	Incorrect
Naming Question	100%	0%
Reality Question	98%	2%

TABLE IV: This table shows how the participants of the Habitat user study performed on the naming & reality question. Correct answers show proper understanding, which means that their data can be used in the benchmark.

Self-declared understanding is positive. At the end of the experiment the participants were asked to evaluate their experience with the user study. On the topic of understanding & clarity the participants responded positively about the experiment, as shown in Figure 5. Appendix A contains further evaluations on encountered difficulties and overall experience. No data points have to be removed from the benchmark based on these results.

B. Experiment 2: Investigate Model Predictions

Number of Belief Updates within expectations. For every data point, the models predict a list of belief updates. The number of belief updates per data point is displayed in Figure 6. Some outliers are excluded from the graph due to a cut-off at the largest $1.5 \times IQR$ (blue dotted line), which was applied to maintain readability. Observing Figure 6 we note

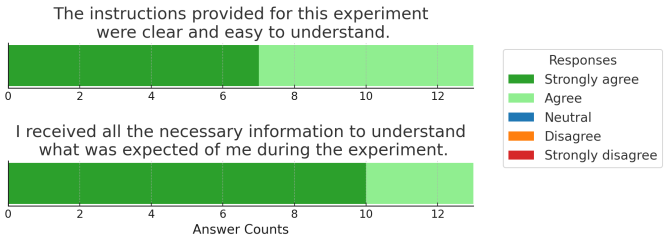


Fig. 5: Users self-declared their understanding at the end of the experiment. Appendix A contains the full participant evaluation overview, this figure is only a subset.

that GPT-4o-Mini exhibits a wider spread in the number of predictions. Upon closer examination of the results, it became evident that GPT-4o-Mini often repeated predictions, rather than exclusively marking significant updates. Furthermore, both GPT-4o and GPT-4-Turbo appear to be mostly limited to either 0 or 1 predictions when relying solely on the video modality.

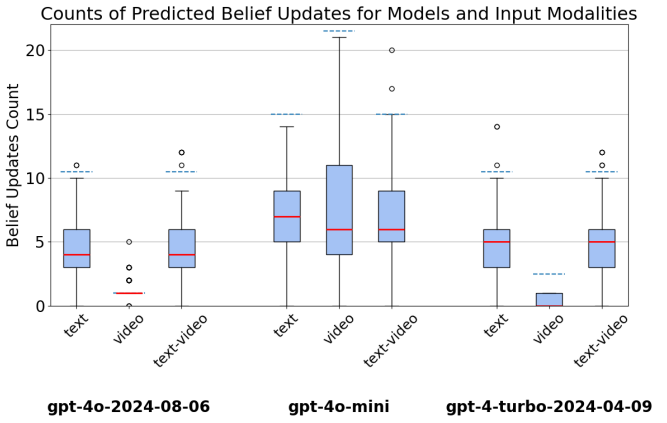


Fig. 6: Boxplot showing the number of belief updates per data point, as predicted by the models based on different modality combinations (text, video, text-video). The blue dotted line marks $1.5\times$ Inter Quartile Range (IQR).

C. Experiment 3: Comparison of Models & Modalities

Finally, we evaluated the consistency of predictions made by three models based on three different modality combinations. Krippendorff's α was selected as our metric to assess the agreement of model inferences across different input modalities, providing insights into the reliability of the models' predictions.

Strongest agreement between text & text-video based predictions. Table V presents the agreement rates of the model predictions based on various modality combinations. Notably, the highest agreement is observed between the text and text-video modalities for all models. In contrast, combinations involving the video modality hover around chance level ($\alpha = 0$). This aligns with a closer analysis

Modality Combinations	GPT-4o	GPT-4o-Mini	GPT-4-Turbo	avg
Text & Text-Video	**76.94**	45.86	60.66	**61.15**
Text-Video & Video	4.27	4.68	0.98	**3.31**
Text & Video	3.60	2.02	0.99	**2.20**
avg	**28.27**	**17.52**	**20.88**	

TABLE V: Krippendorff's alpha values to determine agreement between MLLMs making inferences based on different input modalities. The alpha values are multiplied by 100 for readability. So they are [0-100] instead of [0-1]. GPT-4o refers to the 2024-08-06 version. GPT-4-Turbo refers to the 2024-04-09 version.

of the predictions, where we observe significant overlap in the outputs of the text and text-video modalities, alongside frequent hallucinations and misclassifications in the video-based predictions. Among the models, GPT-4o displays the highest text & text-video agreement, followed by GPT-4-Turbo, and finally GPT-4o-Mini.

Higher agreement between models on text & text-video data than on video data. To further analyze the agreement between models, we focused on their predictions within a single modality (Table VI). Notably, there is significantly higher inter-model agreement for predictions based on text and text-video modalities compared to those based on video alone. This is consistent with the earlier observation that video-based predictions are more prone to hallucinations and misclassifications.

Agreement between GPT-4o & GPT-4o-Turbo higher than when GPT-4o-Mini involved. Additionally, the agreement between GPT-4o and GPT-4o-Turbo is substantially higher than agreements involving GPT-4o-Mini. This difference aligns with observed behavior from GPT-4o-Mini, where the model tends to repeat belief updates rather than only marking the actual update moment. Moreover, GPT-4o-Mini frequently marks the final objects state at the last timestamp, even when no belief update occurs. Hallucinations and misclassifications in the predictions further complicate these issues.

Poor agreement with preliminary human baseline across all models. Because agreement scores do not necessarily reflect the accuracy of the predictions relative to the ground truth, it would be useful to compare their predictions against a human baseline. Given that humans are assumed to possess ToM capabilities, such a comparison could provide valuable intuition about model performance. Table VI (bottom) presents preliminary scores based on the answers from four human participants spread over 16 data points. While the agreement scores are above chance, they remain relatively low. This area would benefit from further investigation. To draw fair conclusions more human evaluation data is needed.

Model Combinations	T	V	T&V	avg
GPT-4o & GPT-4o-Mini	45.74	2.17	44.86	**30.26**
GPT-4o & GPT-4-Turbo	**65.62**	2.17	61.53	**43.77**
GPT-4o-Mini & GPT-4-Turbo	43.70	3.18	43.32	**30.73**
model avg	**51.68**	**2.51**	**49.90**	
Human* & GPT-4o	-	6.30	-	**6.30**
Human* & GPT-4o-Mini	-	3.96	-	**3.96**
Human* & GPT-4-Turbo	-	2.64	-	**2.64**
human avg	-	**4.30**	-	

TABLE VI: Krippendorff's alpha values to determine agreement between different MLLMs making inferences on one input modality. The alpha values are multiplied by 100 for readability. So they are [0-100] instead of [0-1]. The far-right column shows the average alpha values over the different modalities for each model. GPT-4o refers to the 2024-08-06 version. GPT-4-Turbo refers to the 2024-04-09 version. *The human baseline is preliminary, based on the answers from 4 human participants spread over 16 data points.

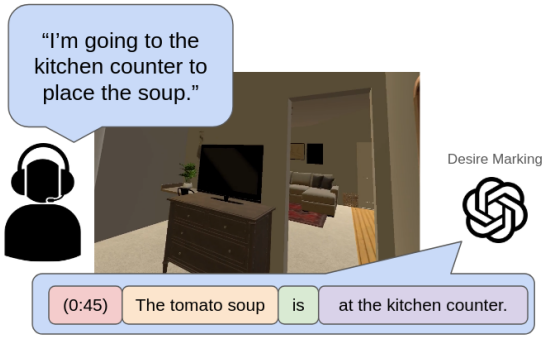
Based on text or text-video data models are capable of object identification & somewhat capable of final object locations identification. While agreement between models improves prediction reliability, it does not necessarily reflect alignment with the ground truth. In this case, determining ground truth is complicated, because we can not directly access someone's mental model. However, it is possible to compare the predictions against simulation aspects that the participant confirmed being aware of. The *Reality Question* from Table IV shows that nearly all participants were able to accurately identify the two objects and their final locations. Tables VII & VIII detail to what extent the predictions included this information. Noticeable is that the text and the text-video modalities show similar performance across all models: correctly identifying both objects for 4 in 5 data points. In contrast, the video modality under-performs on both identification tasks, which corresponds to the low number of predictions as shown in Figure 6.

Model	Text (%)	Video (%)	Text-Video (%)
GPT-4o	**80.0**	2.2	77.4
GPT-4o-Mini	79.6	**52.7**	**79.6**
GPT-4-Turbo	75.0	0.0	79.1

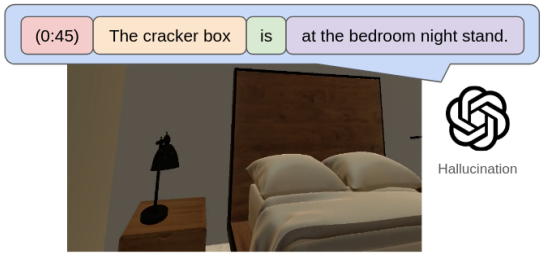
TABLE VII: Accuracy values showing how well the model predictions, based on different modalities, identify both objects of interest in simulation.

Model	Text (%)	Video (%)	Text-Video (%)
GPT-4o	**56.7**	0.0	54.8
GPT-4o-Mini	48.4	**9.9**	57.0
GPT-4-Turbo	44.3	0.0	**58.1**

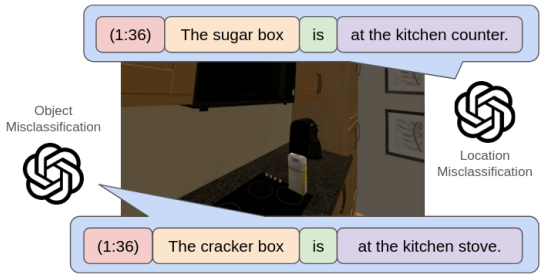
TABLE VIII: Accuracy values showing how well the model predictions, based on different modalities, identify both final object locations.



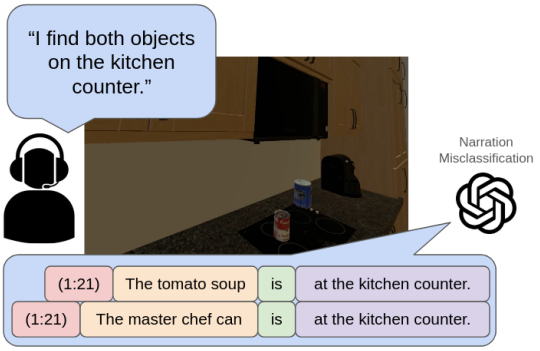
(a) GPT-4-Turbo and GPT-4o-Mini tend to mark a textual desire expressed by the participant as belief update.



(b) Hallucination happens across all models. For video-based data the models sometimes mark a belief update when they hallucinate an object into a location that's in the visual.



(c) Video-based predictions encounter misclassifications often, when an object or location is incorrectly labeled.



(d) Misclassification also happens for text and text-video based predictions. For example, when participants misclassify in their narration or use vague references such as "the kitchen".

Fig. 7: This figure contains illustrations of failure cases the models encounter while predicting belief updates based on text and/or visual data.

VI. Discussion

In this work we present a human-generated benchmark dataset with the objective to serve as evaluation tool for ToM-like capabilities. The dataset contains visual and textual data of human participants performing object rearrangement tasks in simulation. This study specifically evaluates MLLMs on their ability to predict belief updates in the mental model based on data of various modalities.

A. Benchmark Data

To generate benchmark data, a method was selected where participants perform an experiment in a simulation environment while narrating their thoughts. From this process raw audio and video data was collected. Initially, a retrospective think-aloud set-up was considered, where participants would narrate their thoughts after playing the game while watching recordings of themselves playing. However, this method was discarded based on experiment duration and participant feedback, the latter of which indicated that recalling the thought process afterwards was too memory-intensive.

For future iterations of this user study, it would be beneficial to gather screen recording data that captures exactly what the person playing is seeing, to make it easier to mentalize. The current implementation gathers video data from the FPV of the simulated human. For example, without screen recordings, we cannot determine when a participant is looking down to check the object in hand. Additionally, including the navigation hints in the recordings could help in understanding participants' navigation decisions. For instance, if a participant enters the bathroom due to a hint but they actually should have gone to the laundry room, which is located just behind it, the screen recordings would clarify this behavior.

Additional improvements for the next iterations are (1) using more intuitive objects to reduce misclassification, (2) similarly, using locations that are more distinct from each other, and (3) providing clearer narration instructions, possibly with examples.

B. Model Evaluation

The current model prediction method follows a one-shot approach which requires the model to generate the answer in one step. Introducing an intermediate reasoning step, such as in Chain-of-Thought [41], and formatting this step using a causal template, as demonstrated in BigToM [10], is expected boost performance. Additionally, transitioning from a one-shot to a few-shot approach, with examples that have balanced classes and incorporate frequent scenarios, could further improve the model's accuracy. Another potential improvement involves experimenting with varying levels of image detail and frame rates during evaluation, though this may be constrained by API limitations.

Next, the method for evaluating agreement between prediction sets involves comparing each prediction with its counterpart from the other set, specifically identifying updates that mark the same $[object, negation, location]$ within a defined time margin. This time margin, set at 15s, ensures that predictions still have the possibility to match even when they

are triggered with some time offset. The choice of 15s is based on the fact that the narration snippets from Whisper are smaller than that interval, and all belief updates from one snippet are aligned with the starting timestamp of the snippet. By comparing all predictions against predictions within $<$15s, this approach necessitates a metric that's compatible for one-to-zero, one-to-one, and one-to-many comparisons. However, a notable limitation is that predictions can be counted multiple times. Future work could refine this method by developing a more sophisticated metric that addresses these limitations.

Furthermore, the benchmark dataset contains rich free form data. Participants' internal monologues expanded far beyond marking belief updates about object locations. This richness allows future research to evaluate broader parts of the human mental model.

C. Human Evaluation

At present, three models have been evaluated on the benchmark data. While it would be beneficial to include more models in this evaluation, it would also be valuable to assess more human participants, possibly on different modalities to compare how their predictions align. Incorporating additional human data would provide a more robust basis for assessing the models' alignment with the human baseline. Additionally, humans tend to mentalize more easily with individuals similar to themselves. However, difficulty in interpreting someone from a different demographic group does not imply lack of ToM-capabilities. Investigating these effects in models could be an interesting avenue to investigate further, albeit a bit of a tangent.

Moreover, including questions on participants' demographics and English proficiency could provide a more reliable baseline for evaluation.

D. General Concept

This research aims to investigate the presence of ToM capabilities in MLLMs. However, the methods as described in this study to measure these capabilities also inherently test for other capabilities, such as task comprehension, instruction following, and multi-tasking. Additionally, the benchmark creation is based on the assumptions that humans make use of mental models in a goal-directed way, that they have specific goal locations, desired objects, and store relevant information within a belief system.

VII. Conclusion

This paper introduces a novel benchmark dataset, consisting of 94 data points collected through a user study. Each data point includes both visual and textual data, capturing participants' behavior during an object rearrangement task along with their inner monologues. While the dataset is still to be expanded, the current validation results indicate great potential for use in future research.

Using this benchmark, the study aimed to investigate how MLLMs perform in predicting the participants' belief updates based on different modality combinations. A tailored agreement metric (α) was used to assess the performance of the three models: GPT-4o, GPT-4o-Mini, and GPT-4-Turbo. The results showed the highest consistency between the predictions based on text-only and text-video data across all models, with GPT-4o achieving the highest alpha score, followed by GPT-4Turbo and GPT-4o-Mini. Qualitative analysis of the video-based predictions observed hallucinations and misclassifications, which could be an explanation for the poor agreement between video-only predictions and those based on text or text-video data.

Preliminary comparisons with human baselines demonstrated low agreements with the video-based model predictions across all models, suggesting that the models are currently unable to infer belief updates from nonverbal visual data in the same way humans can. However, this study does not provide sufficient evidence to draw conclusions about ToM-like capabilities in the MLLMs examined. Overall, this research builds a foundation for future research.

Acknowledgment

First and foremost, I would like to express my deepest gratitude to my supervisors. My sincere thanks go to Jens Kober, whose facilitation and feedback helped set this work on the right path. A special thanks to Chirag Raman, whose support, thoughtful discussions, and detailed feedback were invaluable. His guidance significantly impacted both my research and personal development.

I would also like to thank Ruta Desai (Meta) for her guidance with Habitat and her valuable insights during meetings. Thanks to Ruta and the Meta Habitat team, the success of this user study was greatly enhanced. I'm also grateful to Ojas Shirekar and Baptiste Colle for the insightful conversations that enriched my understanding and thinking.

Also, a big thanks to my family and Jorn for their support, understanding and encouragement throughout this journey. On top of that, I am deeply thankful to my friends, who provided feedback and, most importantly, were there for me with much-needed mental support. Special shout-out to Tom and Tanya.

Ethics

Ethics plays an important role in researching psychological capabilities in Deep Neural Networks (DNNs), particularly due to the risks associated with misaligned AI and the black-box nature of these models [42–44]. The deployment of MLLMs in robotics holds potential for transforming HRI [29, 30, 45, 46], but deploying black-box systems in positions that impact everyday life raises ethical concerns [47]. Through this research, we aim to contribute to the broader understanding of these models and mitigation of associated risks.

Regarding our user study, all participants provided informed consent, and both the research design and the data management plan were approved by the TU Delft Human Research Ethics Committee (HREC).

References

[1] M. TOMASELLO, *THE CULTURAL ORIGINS OF HUMAN COGNITION*. Harvard University Press, July 2009. Google-Books-ID: ji2_pY4mKwYC.

[2] D. Premack and G. Woodruff, "Chimpanzee theory of mind: Part I. Perception of causality and purpose in the child and chimpanzee," *Behavioral and Brain Sciences*, vol. 1, pp. 616–629, Dec. 1978.

[3] S. D. Preston and F. B. M. De Waal, "Empathy: Its ultimate and proximate bases," *Behavioral and Brain Sciences*, vol. 25, pp. 1–20, Feb. 2002.

[4] C. Beaudoin, Leblanc, C. Gagner, and M. H. Beauchamp, "Systematic Review and Inventory of Theory of Mind Measures for Young Children," *Frontiers in Psychology*, vol. 10, 2020.

[5] M. Kosinski, "Theory of mind might have spontaneously emerged in large language models," *Preprint at https://arxiv. org/abs/2302.02083*, 2023.

[6] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of Artificial General Intelligence: Early experiments with GPT-4," Apr. 2023. arXiv:2303.12712 [cs].

[7] N. Shapira, M. Levy, S. H. Alavi, X. Zhou, Y. Choi, Y. Goldberg, M. Sap, and V. Shwartz, "Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models," May 2023. arXiv:2305.14763 [cs].

[8] M. Sap, R. LeBras, D. Fried, and Y. Choi, "Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs," Apr. 2023. arXiv:2210.13312 [cs].

[9] T. Ullman, "Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks," Mar. 2023. arXiv:2302.08399 [cs].

[10] K. Gandhi, J.-P. Fränken, T. Gerstenberg, and N. D. Goodman, "Understanding Social Reasoning in Language Models with Language Models," June 2023. arXiv:2306.15448 [cs].

[11] Y. He, Y. Wu, Y. Jia, R. Mihalcea, Y. Chen, and N. Deng, "HI-TOM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models," Oct. 2023. arXiv:2310.16755 [cs].

[12] M. Le, Y.-L. Boureau, and M. 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)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 5872–5877, Association for Computational Linguistics, Nov. 2019.

[13] P. Zhou, A. Madaan, S. P. Potharaju, A. Gupta, K. R. McKee, A. Holtzman, J. Pujara, X. Ren, S. Mishra, A. Nematzadeh, S. Upadhyay, and M. Faruqui, "How FaR Are Large Language Models From Agents with Theory-of-Mind?," Oct. 2023. arXiv:2310.03051 [cs].

[14] A. Nematzadeh, K. Burns, E. Grant, A. Gopnik, and T. L. Griffiths, "Evaluating Theory of Mind in Question Answering," Aug. 2018. arXiv:1808.09352 [cs].

[15] J. E. H. Smith, *Embodiment: A History*. Oxford University Press, June 2017. Google-Books-ID: UCgmDwAAQBAJ.

[16] C. Jin, Y. Wu, J. Cao, J. Xiang, Y.-L. Kuo, Z. Hu, T. Ullman, A. Torralba, J. B. Tenenbaum, and T. Shu, "MMToM-QA: Multimodal Theory of Mind Question Answering," Jan. 2024. arXiv:2401.08743 [cs].

[17] C. L. Baker, R. R. Saxe, and J. B. Tenenbaum, "Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution,"

[18] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick, "Machine Theory of Mind," in *Proceedings of the 35th International Conference on Machine Learning*, pp. 4218–4227, PMLR, July 2018. ISSN: 2640-3498.

[19] Y.-S. Chuang, H.-Y. Hung, E. Gamborino, J. O. S. Goh, T.-R. Huang, Y.-L. Chang, S.-L. Yeh, and L.-C. Fu, "Using Machine Theory of Mind to Learn Agent Social Network Structures from Observed Interactive Behaviors with Targets," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, (Naples, Italy), pp. 1013–1019, IEEE, Aug. 2020.

[20] T. Shu, A. Bhandwaldar, C. Gan, K. Smith, S. Liu, D. Gutfreund, E. Spelke, J. Tenenbaum, and T. Ullman, "AGENT: A Benchmark for Core Psychological Reasoning," in *Proceedings of the 38th International Conference on Machine Learning*, pp. 9614–9625, PMLR, July 2021. ISSN: 2640-3498.

[21] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Twenty-first international conference on Machine learning - ICML '04*, (Banff, Alberta, Canada), p. 1, ACM Press, 2004.

[22] J. Jara-Ettinger, "Theory of mind as inverse reinforcement learning," *Current Opinion in Behavioral Sciences*, vol. 29, pp. 105–110, Oct. 2019.

[23] A. Wilf, S. S. Lee, P. P. Liang, and L.-P. Morency, "Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities," Nov. 2023. arXiv:2311.10227 [cs].

[24] M. Sclar, S. Kumar, P. West, A. Suhr, Y. Choi, and Y. Tsvetkov, "Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 13960–13980, Association for Computational Linguistics, July 2023.

[25] S. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. Wang, and Z. Hu, "Reasoning with Language Model is Planning with World Model," Oct. 2023. arXiv:2305.14992 [cs].

[26] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind" ?," *Cognition*, vol. 21, pp. 37–46, Oct. 1985.

[27] J. Perner, S. R. Leekam, and H. Wimmer, "Three-year-olds' difficulty with false belief: The case for a conceptual deficit," *British journal of developmental psychology*, vol. 5, no. 2, pp. 125–137, 1987.

[28] J. Perner and H. Wimmer, ""John thinks that Mary thinks that…" attribution of second-order beliefs by 5- to 10-year-old children," *Journal of Experimental Child Psychology*, vol. 39, pp. 437–471, June 1985.

[29] F. Zeng, W. Gan, Y. Wang, N. Liu, and P. S. Yu, "Large Language Models for Robotics: A Survey," Nov. 2023. arXiv:2311.07226 [cs].

[30] Y. Kim, D. Kim, J. Choi, J. Park, N. Oh, and D. Park, "A survey on integration of large language models with intelligent robots," *Intelligent Service Robotics*, pp. 1–17, 2024.

[31] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "PaLM-E: An Embodied Multimodal Language Model," Mar. 2023. arXiv:2303.03378 [cs].

[32] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," Aug. 2022. arXiv:2204.01691 [cs].

[33] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," July 2023. arXiv:2307.15818 [cs].

[34] K. A. Ericsson and H. A. Simon, "Verbal reports as data.," *Psychological review*, vol. 87, no. 3, p. 215, 1980.

[35] O. Alhadreti and P. Mayhew, "Rethinking thinking aloud: A comparison of three think-aloud protocols," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–12, 2018.

[36] K. A. Ericsson and H. A. Simon, "Protocol analysis: Verbal reports as data (revised addition ed.)," 1993.

[37] M. J. Van den Haak, M. D. de Jong, and P. J. Schellens, "Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison," *Interacting with computers*, vol. 16, no. 6, pp. 1153–1170, 2004.

[38] X. Puig, E. Undersander, A. Szot, M. D. Cote, T.-Y. Yang, R. Partsey, R. Desai, A. W. Clegg, M. Hlavac, S. Y. Min, V. Vondruš, T. Gervet, V.-P. Berges, J. M. Turner, O. Maksymets, Z. Kira, M. Kalakrishnan, J. Malik, D. S. Chaplot, U. Jain, D. Batra, A. Rai, and R. Mottaghi, "Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots," Oct. 2023. arXiv:2310.13724 [cs].

[39] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.

[40] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[41] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Jan. 2023. arXiv:2201.11903 [cs].

[42] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[43] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, *et al.*, "Ai alignment: A comprehensive survey," *arXiv preprint arXiv:2310.19852*, 2023.

[44] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.

[45] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the Opportunities and Risks of Foundation Models," July 2022. arXiv:2108.07258 [cs].

[46] J. Atuhurra, "Large language models for human-robot interaction: Opportunities and risks," *arXiv preprint arXiv:2405.00693*, 2024.

[47] L. Floridi and J. Cowls, "A unified framework of five principles for ai in society," *Machine learning and the city: Applications in architecture and urban design*, pp. 535–545, 2022.

APPENDIX A
SELF-DECLARED USER EXPERIENCE

From the data in Figure 8 we can conclude that the majority of participants reported an overall positive experience with the experiment. All respondents expressed feeling comfortable during the procedure, with only one individual providing a negative response regarding the level of engagement. The experiment encountered two technical and two non-technical issues, but these do not harm the experiment results by default. Additionally, one participant mentioned requiring external assistance (e.g., asking clarifying questions), though it is worth noting that none of the participants reported feeling uncomfortable seeking such support when needed.

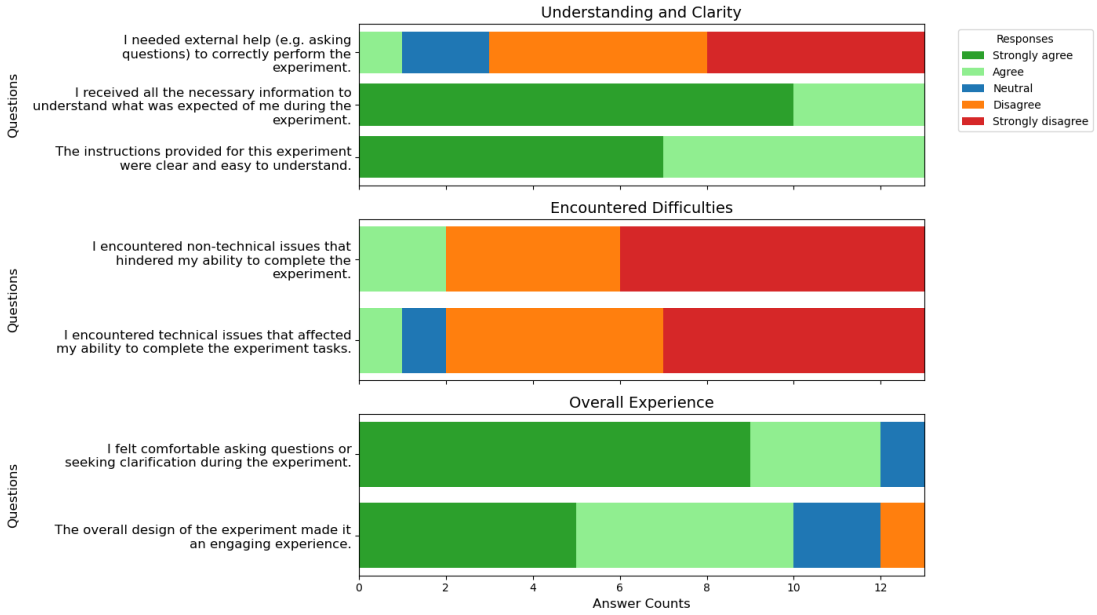


Fig. 8: This figure shows the evaluation from the participants to the Habitat User Study that were taken into the benchmark. They were asked the questions on the left and had the ability to select an answer from a Likert scale. Note that the colors of the bars refer to the selected answers, not necessarily to the interpretation of the output.

APPENDIX B
PREDICTION STATISTICS

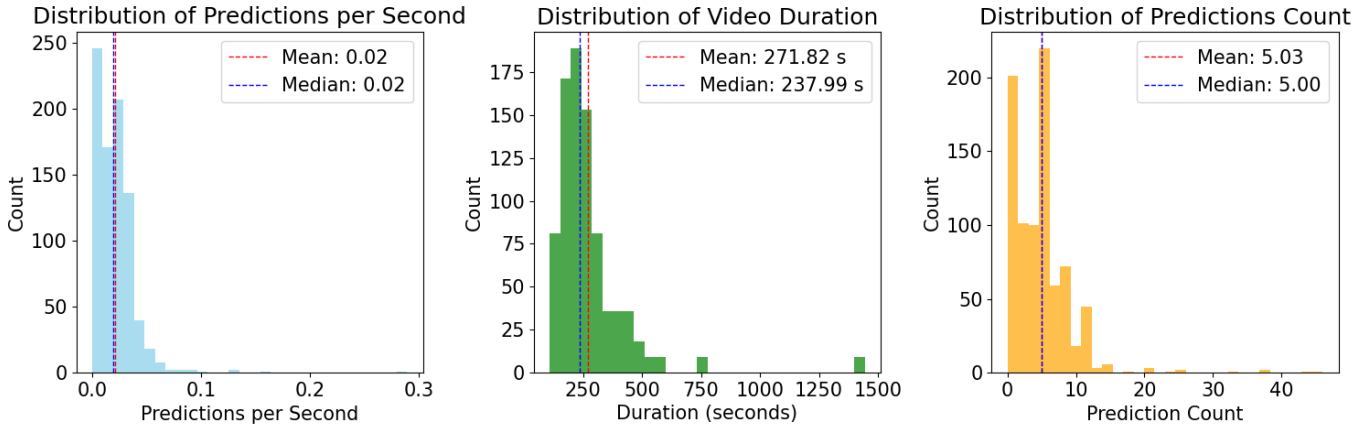


Fig. 9: These histograms show the distributions of the number of predictions and duration of the video per data point of the benchmark data. For each of these distributions the mean and median are marked. The fact that there are five predictions per data point corresponds well to Figure 6

APPENDIX C
HABITAT USER STUDY

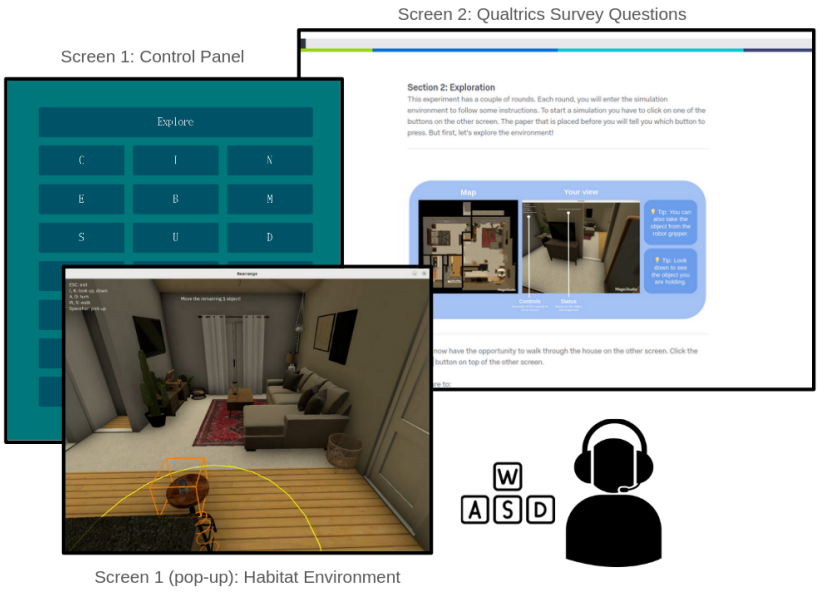


Fig. 10: This figure shows the set-up for the Habitat user study. The participant will constantly have 2 screens in front of them, one showing the Control Panel and one showing the Qualtrics Survey Questions. When the participants receives instructions from the Qualtrics to start Round 1 & Step 1, they click the corresponding letter on the Control Panel, which will launch the Habitat Environment as a pop-up. This process is repeated for all rounds and steps.

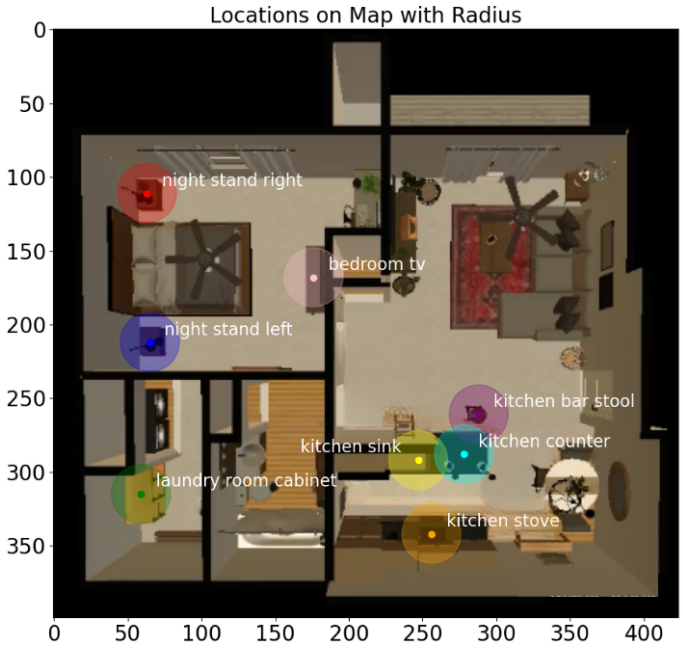


Fig. 11: This map is a top-down view of the Habitat simulation environment that the user study takes place in. The colored dots mark the potential locations for object placement. Each locations has a radius of 20 pixels around it. During the user study, the participants have to mark the final object locations. When this markings are within the radius of the correct locations, they pass the Reality Question (Table IV).

APPENDIX D
MODEL EVALUATION ON BENCHMARK DATASET

Prompt: initial message (video)

```
We have instructed someone to perform object rearrangement in a household simulation. The participant is able to hold 1 object at
a time. There is also a robot present which is able to hold 1 object at a time and navigate around the house. There are 2 objects
and they both have 1 goal location. When someone picks an object, its new location is 'in hand'. When someone places an object, it
is then located at the new location. An object can only be at 1 location at a time. I will now present to you video frames with
the top-down view of the house and the first person perspective.

Limitations of this view is that while playing the person was able to look down and see the object in  their hand, but the video
frames keep looking straight ahead. You  are a skilled assistant that is able to infer what the person playing was thinking. You
have to fill out the Belief Update template for all the moments when the participant is updating their beliefs about the location
of an object.

Examples for when something is a Belief Update:
- They don't know what object they're searching for, but then they see the tomato soup standing on the kitchen stove.
- They remember at the start that you left the sugar box on the night stand.
- They thought that the cracker box was in the kitchen, but now they see that it's not there.

Examples for when something is not a Belief Update:
- They realize that the tomato soup needs to be placed on the bedroom dresser.
- They walk past one of the locations and it is empty, but that location is currently not of interest to them.

Objects: [tomato soup can, master chef can, cracker box, sugar box]
Locations: [laundry room, bedroom night stand, bedroom dresser, kitchen sink, kitchen stove, kitchen counter, kitchen bar stool,
robot gripper, in hand]
If there are any locations unclear, try to think which location from the list is the closest. The bedroom dresser has the
television on it.

Belief Updates happen whenever the person's beliefs about the current location of an object are updated. So not the desired
location, but only the actual location. This can also be a negation (e.g. the tomato soup can is not at the night stand). Be sure
the use the 'in hand' location after they pick an object. The picking motion is not visualized in the video frames, but when
something disappears you can consider it to be picked. Same holds for the placing motion, but when something suddenly appears you
can consider it to be placed. Please state the timestamps, the objects, their locations, and whether it's a negation or not.
```

Fig. 12: This is the initial prompt that will be fed into the model to explain the task at hand. This prompt differs depending on the modality on which the model will make its predictions, such as “video” for this specific prompt.

Prompt: initial message (text-video)

```
We have instructed someone to perform object rearrangement in a household simulation. The participant is able to hold 1 object at
a time. There is also a robot present which is able to hold 1 object at a time and navigate around the house. There are 2 objects
and they both have 1 goal location. When someone picks an object, its new location is 'in hand'. When someone places an object, it
is then located at the new location. An object can only be at 1 location at a time. During the task the people were instructed to
think out loud. I will now present you the audio transcription of this thinking out loud line-by-line accompanied by the video
frame at that moment of speaking.

You have to fill out the Belief Update template for all the moments when the participant is updating their beliefs about the
location of an object. Please use all of the information.

Examples for when something is a Belief Update:
- They don't know what object they're searching for, but then they see the tomato soup standing on the kitchen stove.
- They remember at the start that you left the sugar box on the night stand.
- They thought that the cracker box was in the kitchen, but now they see that it's not there.

Examples for when something is not a Belief Update:
- They realize that the tomato soup needs to be placed on the bedroom dresser.

Objects: [tomato soup can, master chef can, cracker box, sugar box]
Locations: [laundry room, bedroom night stand, bedroom dresser, kitchen sink, kitchen stove, kitchen counter, kitchen bar stool,
robot gripper, in hand]

If there are any objects or locations mentioned which are not in this list, try to think which object or location from the list
they are probably referring to. The master chef can contains coffee. The cracker box is sometimes called cheeze-it box or cheese
box. The soup can is the tomato soup can. The bedroom dresser has the television on it.

Belief Updates happen whenever the person's beliefs about the current location of an object are updated. So not the desired
location, but only the actual location. This can also be a negation (e.g. the tomato soup can is not at the night stand). Be sure
the use the 'in hand' location after they pick an object. Please state the timestamps, the objects, their locations, and whether
it's a negation or not.
```

Fig. 13: This is the initial prompt that will be fed into the model to explain the task at hand. This prompt differs depending on the modality on which the model will make its predictions, such as “text-video” for this specific prompt.

Prompt: initial message (text)

We have instructed someone to perform object rearrangement in a household simulation. The participant is able to hold 1 object at a time. There is also a robot present which is able to hold 1 object at a time and navigate around the house. There are 2 objects and they both have 1 goal location. When someone picks an object, its new location is 'in hand'. When someone places an object, it is then located at the new location. An object can only be at 1 location at a time. I will now present to you the transcriptions of the participant thinking aloud during the task.

You are a skilled assistant that is able to infer what the person playing was thinking. You have to fill out the Belief Update template for all the moments when the participant is updating their beliefs about the location of an object.

Examples for when something is a Belief Update:
- They don't know what object they're searching for, but then they see the tomato soup standing on the kitchen stove.
- They remember at the start that you left the sugar box on the night stand.
- They thought that the cracker box was in the kitchen, but now they see that it's not there.

Examples for when something is not a Belief Update:
- They realize that the tomato soup needs to be placed on the bedroom dresser.
- They walk past one of the locations and it is empty, but that location is currently not of interest to them.

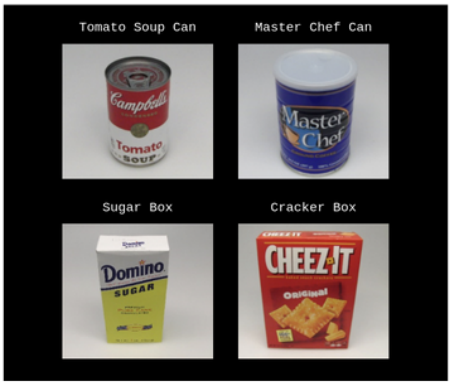
Objects: [tomato soup can, master chef can, cracker box, sugar box]
Locations: [laundry room, bedroom night stand, bedroom dresser, kitchen sink, kitchen stove, kitchen counter, kitchen bar stool, robot gripper, in hand]
If there are any locations unclear, try to think which location from the list is the closest. The bedroom dresser has the television on it.

Belief Updates happen whenever the person's beliefs about the current location of an object are updated. So not the desired location, but only the actual location. This can also be a negation (e.g. the tomato soup can is not at the night stand). Be sure the use the 'in hand' location after they pick an object. Please state the timestamps, the objects, their locations, and whether it's a negation or not.

Fig. 14: This is the initial prompt that will be fed into the model to explain the task at hand. This prompt differs depending on the modality on which the model will make its predictions, such as “text” for this specific prompt.

Prompt: introduce objects (video & text-video)

Attached is an image with the 4 objects: [tomato soup can, master chef can, cracker box, sugar box]



Prompt: introduce locations (video & text-video)

Attached is an image with the top-down view of the room with the locations: [laundry room, bedroom night stand, bedroom dresser, kitchen sink, kitchen stove, kitchen counter, kitchen bar stool, robot gripper, in hand]

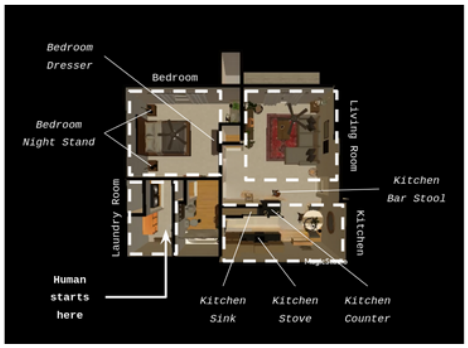


Fig. 15: These prompts introduce the objects and object locations that are available for the model to use in the predictions. These prompts will only be included for the model evaluations based on “video” and “text-video” data.

Prompt: examples (text-video)

(timestamp: 0:33) Ah I already see an object on the nightstand, let's go grab it.

(timestamp: 1:11) I am picking up the sugar box.

(timestamp: 2:37) Let's put down the object on the kitchen stove.

(timestamp: 5:31) Oh huh! The robot is holding my sugar!

(timestamp: 5:37) Give it back to me! Ah yes I have the object back.

(timestamp: 6:11) We will put the sugar box back on the stove.

(timestamp: 7:39) I think the tomato soup is in the laundry room.

(timestamp: 7:45) I picked up the soup can.

(timestamp: 8:33) The soup is placed at the night stand next to the bed.

(timestamp: 9:00) The robot is holding the sugar again!

(timestamp: 9:05) So, yes, I grab it back from the robot.

(timestamp: 9:59) I will put the sugar back on the stove once more.

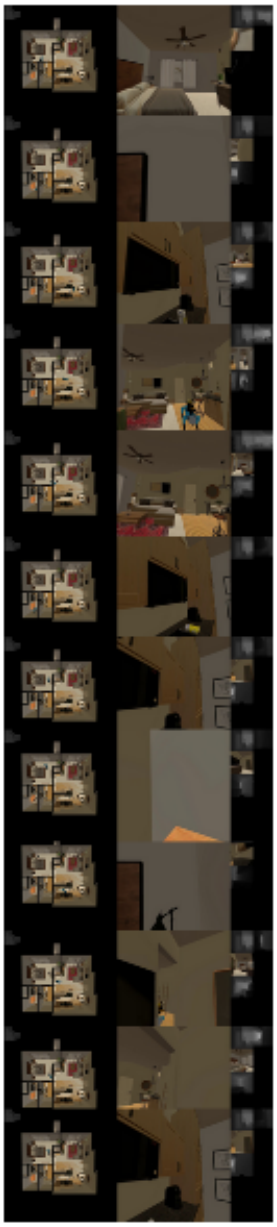


Fig. 16: This prompt provides a one-shot example to the model with both text and video data. Each connected text item and video frame are send in pair.

Prompt: examples (text)

(timestamp: 0:33) Ah I already see an object on the nightstand, let's go grab it.

(timestamp: 1:11) I am picking up the sugar box.

(timestamp: 2:37) Let's put down the object on the kitchen stove.

(timestamp: 5:31) Oh huh! The robot is holding my sugar!

(timestamp: 5:37) Give it back to me! Ah yes I have the object back.

(timestamp: 6:11) We will put the sugar box back on the stove.

(timestamp: 7:39) I think the tomato soup is in the laundry room.

(timestamp: 7:45) I picked up the soup can.

(timestamp: 8:33) The soup is placed at the night stand next to the bed.

(timestamp: 9:00) The robot is holding the sugar again!

(timestamp: 9:05) So, yes, I grab it back from the robot.

(timestamp: 9:59) I will put the sugar back on the stove once more.

Prompt: examples (video)

(timestamp: 0:33)

(timestamp: 1:11)

(timestamp: 2:37)

(timestamp: 5:31)

(timestamp: 5:37)

(timestamp: 6:11)

(timestamp: 7:39)

(timestamp: 7:45)

(timestamp: 8:33)

(timestamp: 9:00)

(timestamp: 9:05)

(timestamp: 9:59)

Fig. 17: These prompts provide the one-shot example to the model for the text data (left) and the video data (right). For the video data, the frames are send to the model in combination with their timestamp.

Prompt: examples answer


```
"belief_update":[
                    {"obj":"sugar box","neg":False,"loc":"bedroom night stand","timestamp":"33.00"},
                    {"obj":"sugar box","neg":False,"loc":"in hand","timestamp":"71.00"},
                    {"obj":"sugar box","neg":False,"loc":"kitchen stove","timestamp":"157.00"},
                    {"obj":"sugar box","neg":False,"loc":"robot gripper","timestamp":"331.00"},
                    {"obj":"sugar box","neg":False,"loc":"in hand","timestamp":"337.00"},
                    {"obj":"sugar box","neg":False,"loc":"kitchen stove","timestamp":"371.00"},
                    {"obj":"tomato soup can","neg":False,"loc":"laundry room","timestamp":"459.00"},
                    {"obj":"tomato soup can","neg":False,"loc":"in hand","timestamp":"465.00"},
                    {"obj":"tomato soup can","neg":False,"loc":"bedroom night stand","timestamp":"513.00"},
                    {"obj":"sugar box","neg":False,"loc":"robot gripper","timestamp":"513.00"},
                    {"obj":"sugar box","neg":False,"loc":"in hand","timestamp":"540.00"},
                    {"obj":"sugar box","neg":False,"loc":"kitchen stove","timestamp":"559.00"},
]}
```


Fig. 18: This prompt provides the answer for the example prompt.

APPENDIX E
HUMAN EVALUATION ON BENCHMARK DATASET

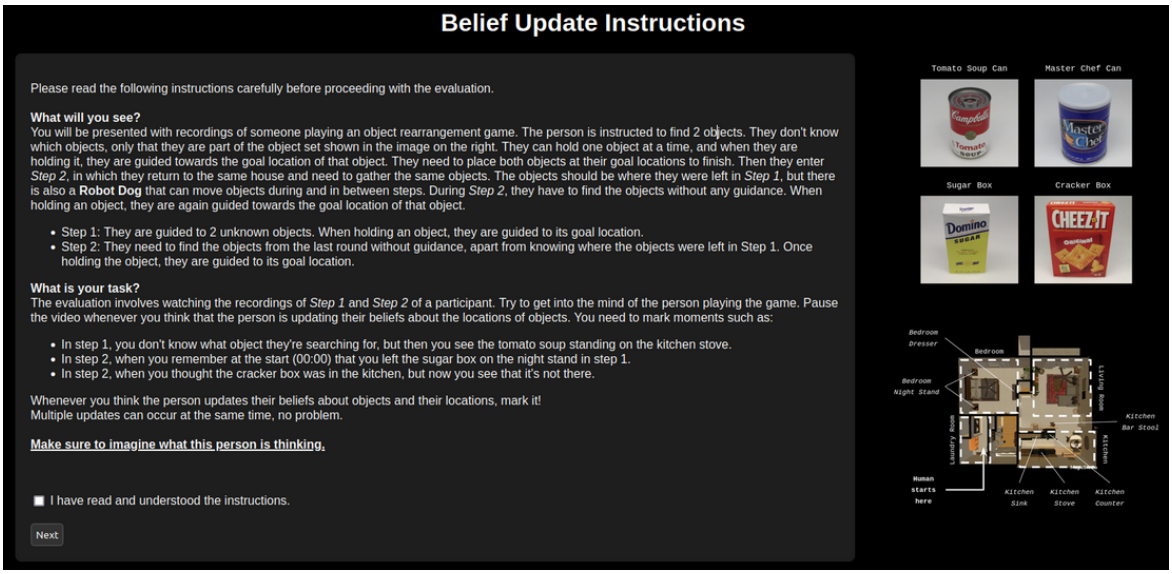


Fig. 19: This UI shows the instructions for human evaluation on the benchmark dataset. On this screen they will learn what is expected of them and get to know the relevant objects and locations. When the participants are ready, the check the box and click on Next.

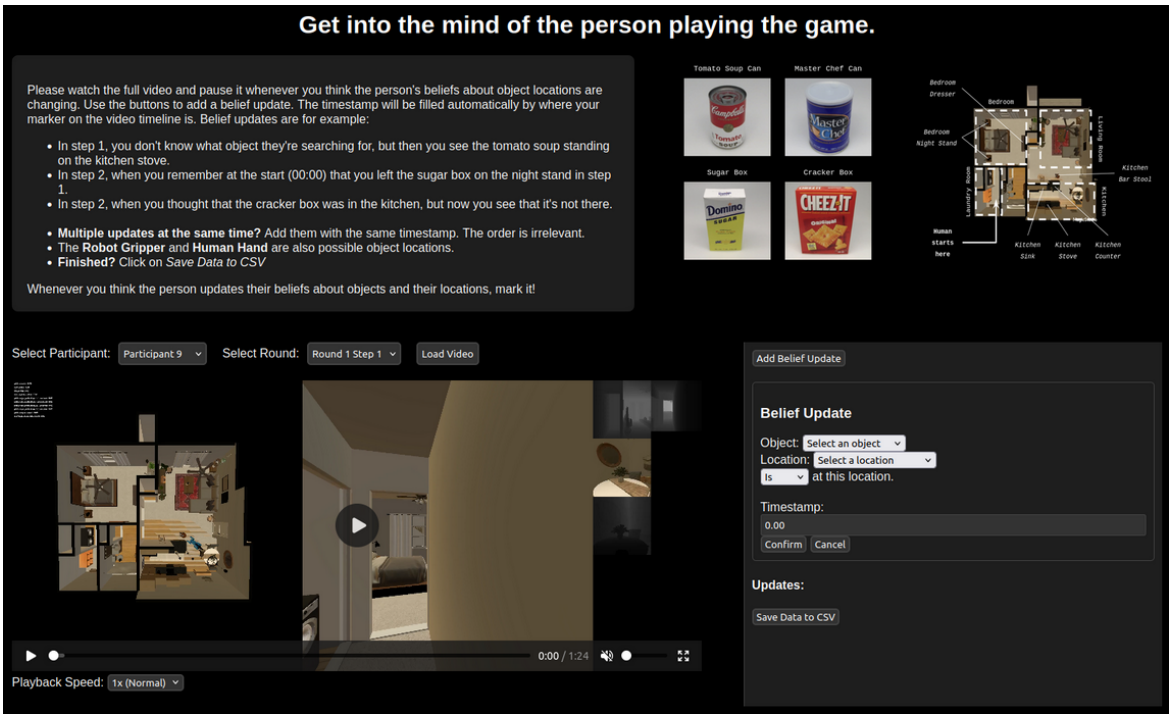


Fig. 20: After the instructions, the human participants will get to this UI. Top-left there are still some instruction pointers for reference. Top-right they will find the objects and location map, which they can click to expand. Bottom-left they are able to load the video they will be evaluating and adjust the speed if preferred. Bottom-right they are able to add a belief update (object, negation, location). The timestamp will be auto-filled by the timestamp they are at in the video (bottom-left).