MuMA-ToM: Multi-modal Multi-Agent Theory of Mind

Haojun Shi[1*], Suyu Ye[1*], Xinyu Fang[1], Chuanyang Jin[1], Leyla Isik[1], Yen-Ling Kuo[2], Tianmin Shu[1]

[1]Johns Hopkins University,
[2]University of Virginia
{hshi33, sye10, xfang21, cjin33, lisik, tianmin.shu}@jhu.edu, ylkuo@virginia.edu

Abstract

Understanding people's social interactions in complex real-world scenarios often relies on intricate mental reasoning. To truly understand how and why people interact with one another, we must infer the underlying mental states that give rise to the social interactions, i.e., Theory of Mind reasoning in multi-agent interactions. Additionally, social interactions are often multi-modal – we can watch people's actions, hear their conversations, and/or read about their past behaviors. For AI systems to successfully and safely interact with people in real-world environments, they also need to understand people's mental states as well as their inferences about each other's mental states based on multi-modal information about their interactions. For this, we introduce MuMA-ToM, a Multi-modal Multi-Agent Theory of Mind benchmark. MuMA-ToM is the first multi-modal Theory of Mind benchmark that evaluates mental reasoning in embodied multi-agent interactions. In MuMA-ToM, we provide video and text descriptions of people's multi-modal behavior in realistic household environments. Based on the context, we then ask questions about people's goals, beliefs, and beliefs about others' goals. We validated MuMA-ToM in a human experiment and provided a human baseline. We also proposed a novel multi-modal, multi-agent ToM model, LIMP (Language model-based Inverse Multi-agent Planning). Our experimental results show that LIMP significantly outperforms state-of-the-art methods, including large multi-modal models (e.g., GPT-4o, Gemini-1.5 Pro) and a recent multi-modal ToM model, BIP-ALM.

Code and data —
https://scai.cs.jhu.edu/projects/MuMA-ToM/

Introduction

Humans live in a social world; we not only engage in social interactions ourselves but can also understand other people's social interactions. Studies in Developmental Psychology have indicated that the ability to understand different kinds of social interactions develops early and is one of the bases for more sophisticated social skills developed later in life (Denham et al. 2003; Wellman, Cross, and Watson 2001; Hamlin, Wynn, and Bloom 2007). Crucially, understanding social interactions goes beyond action recognition. We often need to reason about *why* people interact with one another in a certain manner. We can achieve this by inferring people's mental states as well as how they reason about one another's mental states, i.e., multi-agent Theory of Mind (ToM) reasoning. For instance, if Alice puts away a book on Bob's desk, she may be trying to clean up or hide the book, depending on both her social goal (helping or hindering) and where she believes Bob wants the book (belief of other's goal). As an observer, it may be difficult to disambiguate between these scenarios. However, if we had heard Bob asking Alice where the book was before, we would confidently infer that Alice wanted to hinder Bob. Such multi-modal, multi-agent Theory of Mind abilities are not only crucial for humans but also for AI systems that are deployed in human living environments, such as assistive robots. Without a robust understanding of people's mental states in complex social interactions, AI systems may cause detrimental errors in their interactions with people.

Despite the recent advances in evaluating and engineering machine Theory of Mind, prior works have not adequately addressed the challenge of Theory of Mind reasoning in multi-modal social interactions. First, common Theory of Mind benchmarks (Gordon 2016; Gandhi et al. 2021; Shu et al. 2021; Kosinski 2023; Jin et al. 2024) have only focused on individuals' mental states. However, there are other important aspects of multi-agent mental reasoning, including social goals (e.g., helping, hindering) and beliefs about others' goals. Second, there has not been a multi-modal social interaction dataset designed for systematic Theory of Mind reasoning evaluation. The only prior multi-modal Theory of Mind dataset is MMToM-QA, which solely focuses on single-agent activities. Text-only benchmarks such as HiTOM (Wu et al. 2023) feature multi-agent events, but lack visual inputs. Thus, it remains unclear how we can evaluate the multi-modal multi-agent Theory of Mind capacity in machine learning models.

To address these challenges, we introduce a new Theory of Mind benchmark, MuMA-ToM (Multi-modal Multi-Agent Theory of Mind benchmark). MuMA-ToM includes a large set of question-answering trials. As summarized in Figure 1, questions in MuMA-ToM are organized into three categories: (1) belief inference, (2) social goal inference, and

* Equal contribution.

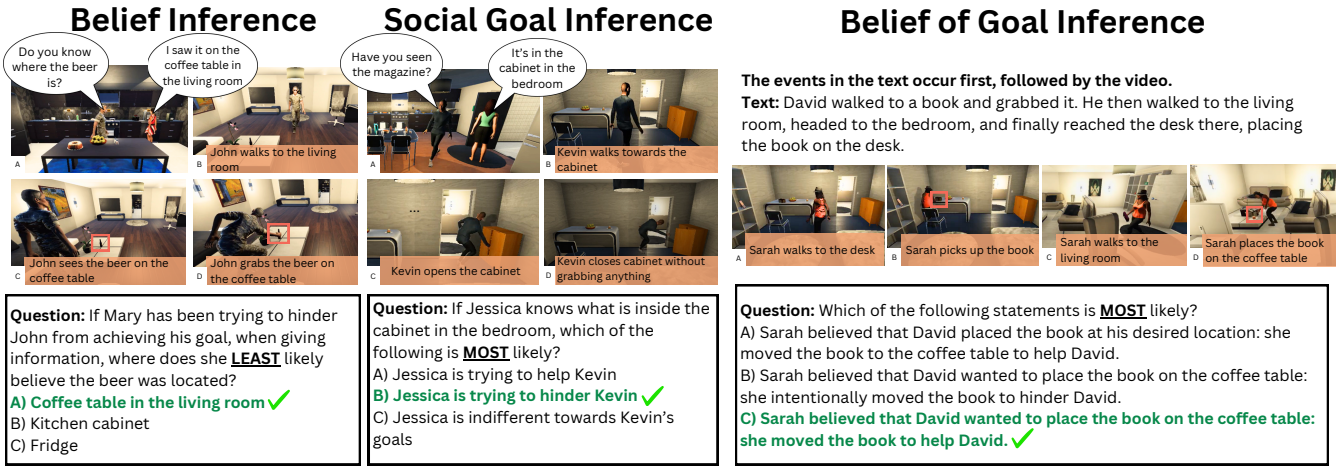

Figure 1: Example questions for each question type. We provide keyframes for the video in each example. The conversations in the chat bubbles are provided as subtitles and shown as part of the multi-modal inputs when viewing the video. Note that the captions on the bottom of the frames are for illustrative purposes only and are not shown in the videos. The checkmarks indicate the correct answers. We provide the videos and text for the examples in the supplementary material.

(3) belief of goal inference. In each trial, there is a multi-agent event in a household environment depicted by video and text. As shown in Figure 1, in some trials, text may show a conversation between two agents; in other trials, text may describe a part of an event that is not depicted in the video. Based on the multi-modal inputs, there will be a question about agents' mental states. We evaluated both humans and state-of-the-art multi-modal models on MuMA-ToM. While humans can achieve near-perfect performance, baselines all fail to robustly infer the mental states based on the multi-modal context.

To bridge the gap between human ToM and machine ToM, we propose a novel multi-modal multi-agent Theory of Mind method – LIMP (Language model-based Inverse Multi-modal Planning). Inspired by a recent method, BIP-ALM, proposed by (Jin et al. 2024), LIMP incorporates language models as components for inverse planning. Unlike BIP-ALM, LIMP (1) introduces multi-agent planning with two-level reasoning, (2) eliminates the need for manually defined symbolic representations for a better generality, and (3) can leverage any pretrained LLMs whereas BIP-ALM requires LLMs finetuned on symbolic representations. Experimental results demonstrate that LIMP significantly outperforms baselines.

In sum, our contribution includes (1) the first benchmark on multi-modal multi-agent Theory of Mind reasoning, (2) a human experiment validating the benchmark and providing a human baseline, (3) a systematic evaluation of state-of-the-art large multi-modal models (LMMs), and (4) a novel multi-modal multi-agent ToM method combing inverse multi-agent planning and language models.

Related Works

Single-agent ToM benchmarks (Gordon 2016; Gandhi et al. 2021; Shu et al. 2021; Kosinski 2023; Jin et al. 2024) have extensively tested concepts like belief, goal, preferences, constraints, and rationality. Multi-agent benchmarks are typically built based on the classic Sally-Anne test (Baron-Cohen, Leslie, and Frith 1985) for false beliefs and higher-order beliefs (Le, Boureau, and Nickel 2019; He et al. 2023; Xu et al. 2024; Soubki et al. 2024). There have also been multi-agent benchmarks that focus on a single agent's beliefs & intentions in complex conversations or interactions (Kim et al. 2023; Chen et al. 2024; Chan et al. 2024; Sabour et al. 2024). In these benchmarks, other agents are usually present to add context or complexity, but there are no questions about inter-agent relationships. Prior works on testing social relationship understanding (Netanyahu et al. 2021; Li et al. 2024) rely on simple animations, which lack the realism of embodied human interactions. Most existing ToM benchmarks have only either text or video. The only exception is MMToM-QA (Jin et al. 2024), which has single-agent activities depicted in video and text. Our MuMA-ToM benchmark features two agents interacting in an embodied household environment, with both text and video as multi-modal inputs, and includes questions that test the agents' social intentions and their reasoning about each other's mental states.

Multi-Modal Benchmarks. Given the recent advances in LLMs, there has been increasing interest in developing multi-modal QA benchmarks. Most of these benchmarks focus on models' ability to fuse information from multiple modalities, where answers are directly retrievable without complex reasoning (Li et al. 2023b; Sanders et al. 2023; Li et al. 2023a; Ying et al. 2024a; Tang et al. 2024; Pandya et al. 2024). A recent benchmark, Perception Test (Patraucean et al. 2024), evaluates physical reasoning such as predicting world states and explaining counterfactual facts. But it differs from ToM reasoning. Pipelines for generating multi-modal datasets, SEED-story (Yang et al. 2024) and TaskMeAnything (Zhang et al. 2024), also do not evaluate ToM reasoning. MMToM-QA (Jin et al. 2024), a

recent multi-modal ToM benchmark, evaluates ToM with multi-modal inputs about single-agent behaviors. Unlike MMToM-QA, our benchmark includes multi-agent interactions and evaluates models' understanding of mental state reasoning in multi-modal social interactions.

Machine Theory of Mind. Traditional approaches to Theory of Mind reasoning fall into two categories: end-to-end training (Rabinowitz et al. 2018; Han and Gmytrasiewicz 2019) and Bayesian Inverse Planning (Baker et al. 2017; Zhi-Xuan et al. 2020; Stacy et al. 2024). There have been works on neural amortized inference that combine these two methods for efficient and robust ToM inference in visual domains (Jha et al. 2024; Puig et al. 2023). Recently, LLMs demonstrated some ToM reasoning capabilities (Kosinski 2023; Bubeck et al. 2023), but their ToM reasoning is still brittle (Verma, Bhambri, and Kambhampati 2024; Amirizaniani et al. 2024; Ullman 2023; Sclar et al. 2023a; Ivanova et al. 2024). Approaches using prompt engineering have been proposed to enhance the ToM capacities in LLMs for text-based QAs (Wilf et al. 2023; Sclar et al. 2023b). Jin et al. (2024) proposed, BIP-ALM, for multi-modal ToM. BIP-ALM first extracts and fuses symbolic representations from multi-modal inputs and then combines a language model and Bayesian inverse planning to conduct ToM reasoning based on the symbolic representations. While achieving promising results on MMToM-QA, BIP-ALM lacks multi-agent reasoning capacity and requires finetuning a language model on hand-designed symbols. Our LIMP model builds on BIP-ALM and introduces key improvements including multi-agent planning and general, domain-invariant representations.

MuMA-ToM Benchmark

General Structure

The benchmark consists of 225 multi-modal social interactions between two agents. There are 900 multi-choice questions based on these social interactions. Each question depicts a social interaction in video and text jointly. As shown in Figure 1, the text may show a conversation between the agents or a part of the event, and the video shows the complementary part of the event. Given the multi-modal inputs, the questions are designed to assess the understanding of agents' mental states during these interactions, probing three main concepts: (1) beliefs, (2) social goals, and (3) beliefs of others' goals. Each concept has 300 questions. We also created a training set consisting of 1,030 videos annotated with the agents' actions and goals. The training set does not provide example questions. It is intended for a model to learn about typical multi-agent household activities.

Question Types

As identified in prior works in cognitive science (Ullman et al. 2009; Shu et al. 2020) and multi-agent planning (Gmytrasiewicz and Doshi 2005; Tejwani et al. 2021), there are three mental variables that are crucial to ToM reasoning in multi-agent interactions: an agent's belief of the physical state, its social goal, and its belief of other agents' goals. Therefore, we design three types of questions in our benchmark corresponding to the three mental variables: belief inference, social goal inference, and belief of goal inference. Each type of question asks about the corresponding mental variable of one of the agents. Among the three options, we make sure that there is always one option that is clearly the most likely to be correct.

One of the challenges in designing these three types of questions is that given an interaction, multiple combinations of these mental variables could be equally possible. For instance, if we see that Alice's actions prevent Bob from reaching his goal, it could be because Alice is hindering Bob, knowing Bob's true intent; or she may try to help Bob but has a false belief of Bob's goal and ends up accidentally hindering Bob. To address the challenge of large hypothesis space, we always ask a question about a mental variable conditioned on explicitly provided assumptions about the other two mental variables. For instance, as shown in the example question of the belief inference in Figure 1, the goal of John can be inferred from his question about where the beer is (the goal is getting beer), and Mary should be aware of this as she answered John's question; the social goal of Mary is unclear, therefore the question provides a hypothetical social goal, hindering, as the condition. The remaining mental variable of Mary, her belief of the physical state can be clearly inferred given her social goal and her belief of John's goal.

We explain the design of each question type as follows.

Belief Inference. These questions focus on inferring a person's belief about the physical state based on their utterance and social goal. The person may have a *true belief* or *false belief* about the location of the object, which can be inferred when we constrain their social goal to be helping or hindering. In the example depicted in Figure 1, John asks Mary where he can find the beer. Mary suggests the coffee table, which turns out to be the correct location, as John successfully finds the beer there. This could be interpreted in two ways: (1) Mary helps John, genuinely believing the beer is on the coffee table, or (2) Mary accidentally helps John while intending to mislead him, mistakenly believing that the beer isn't on the coffee table. To answer correctly, a model needs to understand: (1) Mary knows John's goal (from their conversation), (2) John follows Mary's directions (from their conversation and his actions afterward in the video), and (3) John achieves his goal by following Mary's directions (as shown in the video). We balance true and false beliefs in the ground-truth answers.

Social Goal Inference. In these questions, we ask about a person's social goal. Specifically, we consider helping, hindering, or acting independently as the three possible social goal categories, which are also the common social goal types in physically grounded social interaction reasoning studied by prior works in cognitive science (Hamlin, Wynn, and Bloom 2007; Ullman et al. 2009; Shu et al. 2020; Malik and Isik 2023). The example in Figure 1 shows an interaction similar to the one in the example for belief inference questions. In this particular example, Jessica misleads Kevin to the cabinet where there is no magazine inside. In the question, we assume that Jessica does indeed know the true state, and therefore, one should infer that Jessica is trying to hin-

der Kevin. To achieve this correct inference, a model needs to focus on (1) how Jessica infers Kevin's goal (from the conversation), (2) how Kevin searches the room after the conversation (from both the conversation and the video following the conversation), and (3) whether Kevin can find his goal object at the location suggested by Jessica (from the video). We balance cooperative and adversarial behaviors for the ground-truth answers.

Belief of Goal Inference. Belief of goal inference asks a model of how one person thinks about another person's goal given the context. In each option for a question of this type, we always pair the belief of another person's goal with the corresponding social goal to minimize ambiguity. For instance, in the interaction for the example question of belief of goal inference in Figure 1, Sarah moves the book to the coffee table after David places it on the desk. However, it is unclear whether Sarah is aware that David places the book there and whether Sarah thinks that David wants to keep the book on the desk. If Sarah were trying to help David, as assumed in the correct option, she would have believed that David wanted the book on the coffee table instead. In this case, as a third-person observer, we may not be certain of David's true intent, but we can still infer Sarah's belief of David's goal given that her social goal is helping him. For this type, half of the questions have a true belief of goal as the correct answer, and the other half have a false belief of goal as the correct answer.

Multi-modal Information

In MMToM-QA, the only prior multi-modal ToM QA benchmark, each modality covered *all* aspects of the human activity and environment, making it difficult to discern what information could be extracted from one modality but not the other. Our benchmark aims to provide clearly separate information accessible only through one modality, allowing us to understand precisely how a model needs to fuse multi-modal inputs to answer each question.

As illustrated in Figure 1, there are two main ways in which multi-modal information must be integrated. First, if there are conversations between two agents, the model must understand the exchanged information and how it impacts each person's mental state, including any changes in their beliefs about each other. The model must also observe actions and outcomes, connecting them to the conversation to reason further about mental states. Note that conversations can occur at any point in the video. Second, for interactions without verbal communication, we provide part of the event in text and the remaining part in video. Specifically, we either describe the first half in text and show the second part in video or show the first part in video and describe the second half in text. These two designs are randomly sampled to describe interactions jointly in video and text.

Procedural Generation

We use a multi-agent household simulator, VirtualHome (Puig et al. 2018, 2020), to procedurally synthesize social interactions between two agents. For each interaction, we sample an environment and goals for the agents. We consider three general social scenarios: an agent is trying to help another agent, an agent is trying to hinder another agent, and two agents are acting independently. Agents only have partial observations and do not know each others' goals. They can optionally talk to each other. We leverage a recent method proposed by Ying et al. (2024b)—Goal-Oriented Mental Alignment (GOMA)—-to generate action plans as well as verbal communication. GOMA combines hierarchical planning, goal inference, and large language models (LLMs) to generate multi-modal interactions between embodied agents. Prior work (Puig et al. 2020) has demonstrated that activities synthesized in VirtualHome indeed resemble real-world human activities. We provide more details on the procedural generation in the supplementary material.

Our Model

Formulation

To model social interactions between two agents, i and j, and the recursive mental reasoning between them, we adopt an Interactive Partially Observable Markov Decision Processes (I-POMDP) formulation (Gmytrasiewicz and Doshi 2005). We define s^t as the state, a_i^t and a_j^t as agents' actions, and u_i^t and u_j^t as agents' utterances at time t. Each agent maintains its own beliefs b_i^t and b_j^t, as well as goals g_i and g_j. To capture recursive reasoning, we define interactive states for the agents, denoted as $is_{i,\ell}$ and $is_{j,\ell}$ at level ℓ. From the perspective of agent i, its interactive state at each level is defined as follows (we consider the first two levels in this work):

- **Level 0:** $is_{i,0} = s$
- **Level 1:** $is_{i,1} = (s, b_{j,0}, g_j)$ (where $b_{j,0}$ is a distribution over agent j's level 0 interactive state, $is_{j,0}$)
- ...

Given the belief of interactive state $b(is_{i,1})$, an agent's action policy will be $\pi(a_i|is_{i,1}, g_i)$, and its utterance policy will be $\pi(u_i|is_{i,1}, g_i)$.

Overview

Previous works on Inverse Multi-agent Planning (IMP) (Ullman et al. 2009; Netanyahu et al. 2021) have demonstrated that IMP can robustly infer agents' mental states in social interactions. However, these methods rely on manually crafted planners and are limited to simple visual scenarios, such as 2D grid worlds. Jin et al. (2024) introduced the BIP-ALM model, which leverages language models for inverse planning to achieve single-agent Theory of Mind reasoning in complex, realistic settings. Inspired by BIP-ALM, we propose a novel method, Language model-based Inverse Multi-agent Planning (LIMP), to combine IMP and language models for robust multi-agent Theory of Mind reasoning based on multi-modal inputs.

As illustrated in Figure 2, LIMP consists of three key components: multi-modal information fusion, hypothesis parsing, and inverse multi-agent planning. Compared to BIP-ALM, our approach offers several improvements. First, while BIP-ALM is limited to single-agent scenarios, LIMP identifies three mental variables crucial to understanding multi-agent interactions—belief, social goal, and belief of

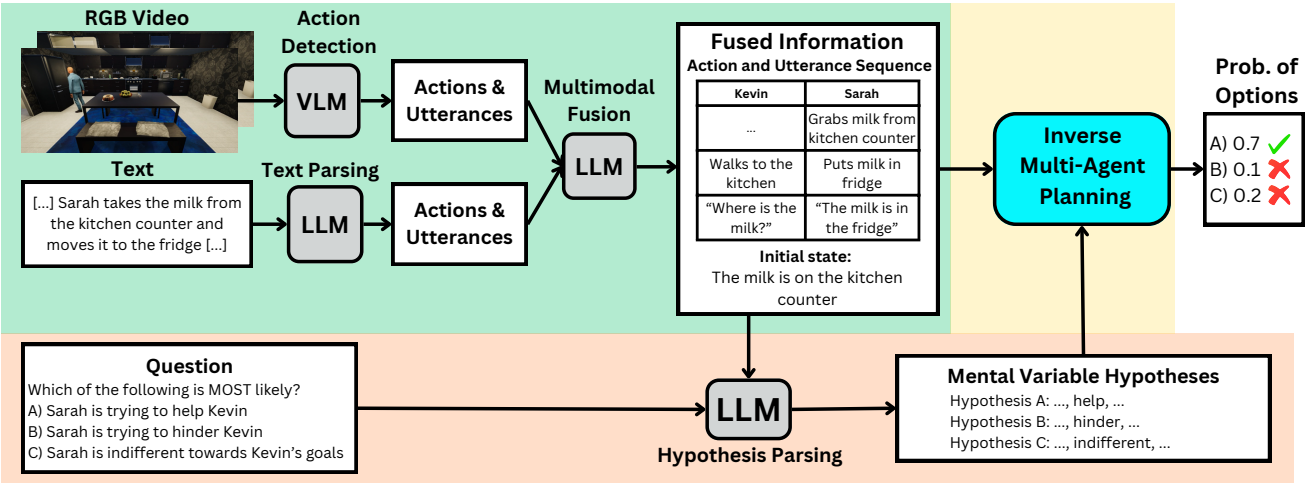

Figure 2: Overview of LIMP. LIMP has three components: (1) the multi-modal information fusion module extracts and fuses information from vision and text; (2) the hypothesis parsing module generates hypothetical values for the three mental variables given the question and the fused information; and (3) the inverse multi-agent planning module assesses the probabilities of each option given the hypothetical mental variables and the multi-modal agent behavior described in the fused information.

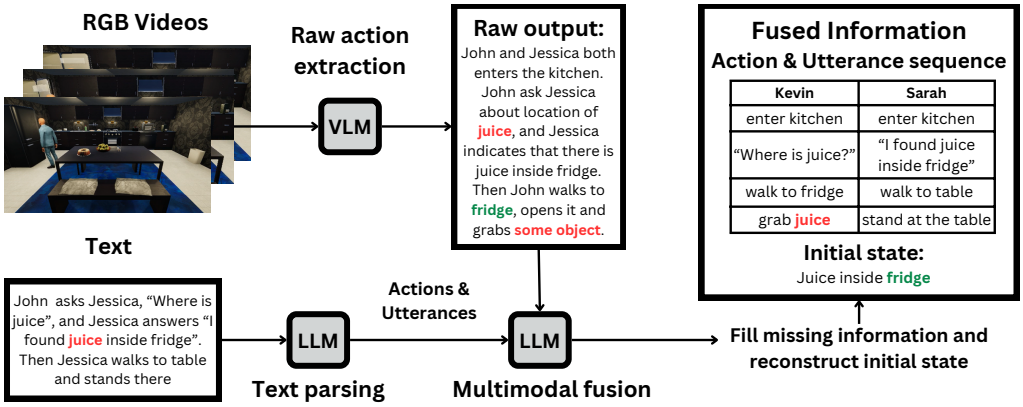

Figure 3: Illustration of the multi-modal information fusion in LIMP. It fills in missing information based on the context and recovers the initial state from agents' actions.

goal. We then implement multi-agent planning based on these variables to reason about multi-modal social interactions. Second, BIP-ALM relies on hand-designed symbolic representations and requires finetuning language models on these representations. In contrast, LIMP uses natural language to represent states, actions, and utterances, eliminating the need for finetuning and enhancing generalizability across domains. Finally, LIMP's multi-modal information fusion module can fill in missing information from visual perception using contextual cues from text or action sequences (Figure 3), a capability absent in BIP-ALM. We discuss the details of each component in the remaining section.

Multi-modal Information Fusion

We use a vision-language model (VLM) to extract the actions and utterances of each person depicted in the video. Given text, we use an LLM to extract the actions and utterances of each person. We then fuse the extracted information to form the initial state and the complete sequences of actions and utterances using an LLM as follows.

Unlike MMToM-QA, our benchmark does not provide a text description of the full state, as such descriptions are rarely provided in real-world applications. However, as objects may be occluded or too small to detect even for humans, inferring the state directly from the RGB videos could be difficult. Instead, we prompt an LLM with the inferred actions and utterances of both agents to infer the part of the initial state relevant to the activity. For example, if Alice grabs a carrot from the fridge, and moves it to the kitchen table, we can infer that the carrot was originally in the fridge. Using this method, the reconstructed initial state will only consider objects relevant to human actions and utterances. This simplifies the context and can consequently improve the accuracy of the inference. Given the initial state and the action

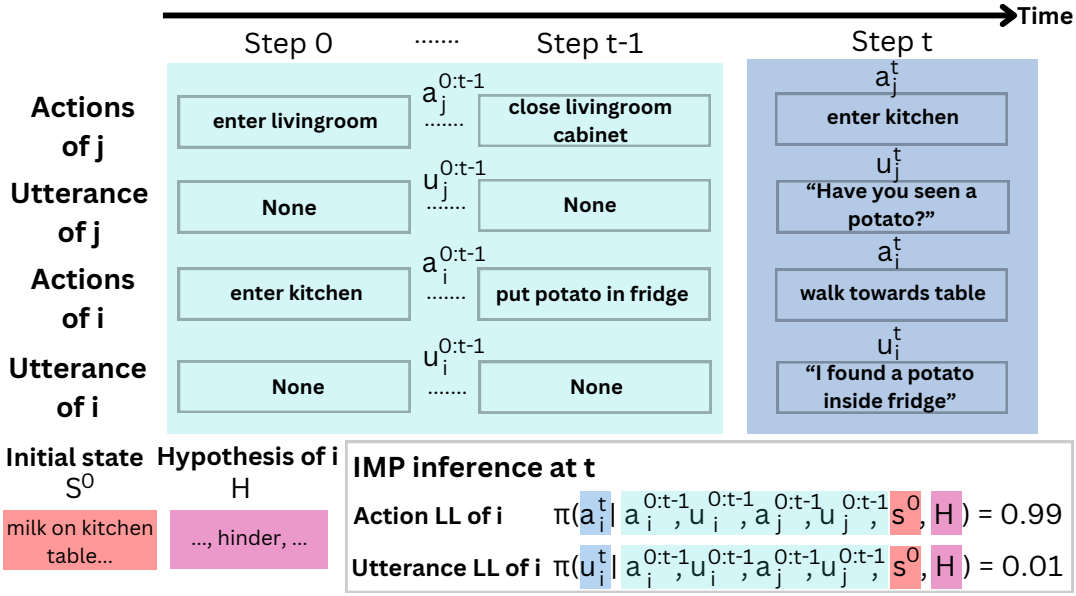

Figure 4: Illustration for inverse multi-agent planning. We estimate the action and utterance likelihood of agent i at each step t given the past actions and utterances of both agents from step 0 to step $t-1$, the initial state s^0, and the hypothesis H. LL in the figure stands for likelihood.

sequences, we can infer the state at each step.

There is often missing information in the visual perception results. For instance, as shown in Figure 3, the VLM did not recognize the object the person grabbed and produced an ambiguous action – "grabs some object." This is also common to people, as the object picked up by the person is often occluded. However, we can still infer that the object is likely juice based on the context provided in the text. To emulate such ability, we leverage an LLM to fuse information extracted from video and text, which infers the information missing from visual perception based on the complementary information described in the text. In this work, we use Gemini 1.5 Pro for the VLM and GPT-4o for the LLM as they produce the best results.

Hypothesis Parsing

To answer the question about a person's mental state in a social interaction, LIMP will parse relevant hypotheses of all mental variables of that person (agent i) – belief of state $b(s)$, social goal g_i, and belief of other agent's goal $b(g_j)$. For this, we prompt GPT-4o with the initial state and question text to generate a reasonable hypothesis of the three mental variables for each option, $H = \langle b(s), g_i, b(g_j) \rangle$.

Inverse Multi-Agent Planning

Given the fused information from multi-modal inputs and the parsed hypotheses, inverse multi-agent planning conducts Bayesian inference over a person's mental state by evaluating the likelihood of actions and utterances given each hypothesis. Following the I-POMDP formulation, we define this probabilistic inference as follows:

$$
\begin{aligned}
&P(H \mid a_i^{0:T}, u_i^{0:T}, a_j^{0:T}, u_j^{0:T}, s^0) \\
&\propto P(H) \prod_{t=1}^{T} \pi(a_i^t \mid a_i^{0:t-1}, u_i^{0:t-1}, a_j^{0:t-1}, u_j^{0:t-1}, s^0, H) \\
&\cdot \prod_{t=1}^{T} \pi(u_i^t \mid a_i^{0:t-1}, u_i^{0:t-1}, a_j^{0:t-1}, u_j^{0:t-1}, s^0, H), \qquad (1)
\end{aligned}
$$

where the action policy and the utterance policy can be estimated by the log probabilities of the prompt completion by a language model for each time step t. Note that in the standard policy definitions in I-POMDP, we need agent i's belief of agent j's belief of the state at each step. This, however, is difficult to explicitly estimate. Instead, in this work, we consider past actions and utterances of all agents as part of the condition of the policies to avoid the explicit belief of belief inference. We prompt an LLM with the hypothesis, the initial state, and the previous actions and utterances of both agents to estimate the action and utterance policies. In this way, we do not need to implement domain-specific planning, which can be extremely challenging and slow for multi-agent interactions with both physical actions and verbal communication. In this work, we use GPT-4o for the LLM. We find that GPT-4o can accurately estimate the action and utterance policies based on the given condition.

Figure 4 illustrates how IMP evaluates the action and utterance likelihood at one time step. Given the condition, the LLM estimates that it is likely that agent i will take the observed action ("walk towards table") but is unlikely to say "I found a potato inside fridge" as it is inconsistent with the social goal of hindering agent j (agent i had just put a potato in the fridge before the conversation).

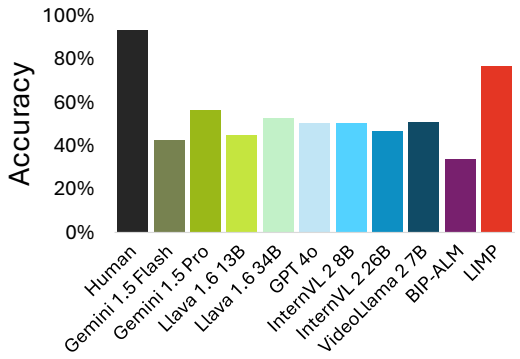

Figure 5: Human and model performance on MuMA-ToM.

Method	Belief	Social Goal	Belief of Goal	All
Human	98.9	94.4	87.1	93.5
Gemini 1.5 Flash	53.9	33.0	41.4	42.7
Gemini 1.5 Pro	78.9	43.9	46.9	56.4
Llava 1.6 13B	70.2	43.2	17.9	43.7
Llava 1.6 34B	**93.6**	37.2	27.5	52.8
GPT-4o	67.9	39.6	44.4	50.6
InternVL 2 8B	62.2	44.6	45.1	50.6
InternVL 2 26B	59.3	44.9	35.5	46.6
VideoLlama 2 7B	70.1	45.6	37.7	51.1
BIP-ALM	41.2	34.1	30.6	33.9
LIMP	93.4	**67.7**	**68.7**	**76.6**

Table 1: Human and model performance for different question types as well as for all questions.

Experiments

Human Experiment

We recruited 18 participants (mean age = 36.0; 10 female) from Prolific to answer 90 questions randomly sampled from the benchmark. Each question received responses from 3 participants. The experiment was approved by an institutional review board.

Baselines

We evaluated our benchmark on state-of-the-art LMMs. For models capable of processing video input, the entire video was provided. For models without video input capabilities, we uniformly sample one frame every 20 frames from the video episode as input. We evaluated **GPT-4o** (OpenAI 2023), **Llava 1.6** (Liu et al. 2023), **Gemini 1.5** (Reid et al. 2024), **InternVL2** (Chen et al. 2023) and **VideoLlama 2** (Cheng et al. 2024). We evaluated the latest version of each LMM at the time of submission. For **LIMP**, we use Gemini 1.5 Pro as the VLM and GPT-4o as the LLM. Finally, we evaluated **BIP-ALM** using finetuned Llama 2(Jin et al. 2024), the best-performing model on a prior multi-modal ToM benchmark, MMToM-QA. For this evaluation, we used the original **BIP-ALM** model, which was fine-tuned on the MMToM-QA dataset. More details of the experiments are provided in the supplementary material.

Results

We report the human and model performance in Figure 5 and Table 1. Human participants achieved almost perfect accuracy across all questions, with 98.9% of the correct answers having majority agreement. The overall performance averaged across individual participants is 93.5%. The slightly lower performance on social goal inference (94.4%) and belief of goal inference (87.1%) indicates these questions are more challenging and require greater focus.

All LMM baselines performed poorly on MuMA-ToM, indicating a substantial gap between machine and human ToM. Among the three question types, belief inference is the easiest for LMMs. In particular, Llava 34B achieved the highest accuracy for belief inference. However, all LMMs struggle with the more challenging social goal inference and belief of goal inference questions. The best-performing LMM model overall is Gemini 1.5 Pro, with an accuracy of 56.4%. Notably, BIP-ALM had an accuracy of 33.9%, indicating its inability to understand multi-agent interactions. This is because BIP-ALM only relies on single-agent goals and beliefs for inverse planning but does not consider social goals and beliefs of others' goals. Additionally, BIP-ALM assumes certain symbolic representations, which also limits its generality. Our LIMP model significantly outperforms allstate-of-the-art models on our benchmark, with an overall accuracy of 76.6%. Critically, by using GPT-4o to estimate action and utterance likelihood for inverse multi-agent planning, LIMP can achieve much better performance than GPT-4o itself. There is still a gap between the best model performance and human performance, highlighting the need for further studies.

We finetuned VideoLlama 2 on a video instruction dataset created from our training set. The finetuned model did not perform better, suggesting that common finetuning approaches for LMMs do not improve their ToM capacities. We also evaluated the performance of LMMs with chain-of-shot prompting (Kojima et al. 2022) and found no improvement (Table 2 in the supplementary material).

Discussion

Why do LMMs perform poorly? There are two sources of systematic errors for LMMs. First, LMMs struggle with understanding multi-modal behavior in complex social situations, often failing to distinguish between deliberate hindering and failed attempts to help due to incorrect beliefs. Most models can solve belief inference tasks where helping is the assumed social goal. However, they consistently struggle in scenarios where hindering is the assumed social goal. (e.g., "If Mary is trying to hinder Jack, where does she least likely believe...?") The failure to understand adversarial behaviors is even more prominent in social goal inference and belief of goal inference. Second, LMMs often fail to correctly interpret visual inputs, such as when an object is too small or is occluded when the agent is picking it up, leading to incorrect conclusions about the agent's actions. While humans are able to use contextual clues to infer what the object might be, VLMs struggle with this task. These errors in

recognizing crucial actions likely contribute significantly to their overall poor performance on our benchmark.

Why does LIMP outperform the best LMMs? LIMP overcomes the two aforementioned weaknesses of LMMs – the inability to recognize multi-modal behavior under various social goals and the sensitivity to noisy visual perception. First, while LLMs struggle with direct ToM reasoning, they excel at the forward generation of multi-modal behavior given mental states. For example, it is much harder for an LLM to correctly infer whether an agent is hindering another by lying than it is for the model to generate a lie based on the agent's belief and social goal. Such multi-modal behavior generation ability enables LIMP to estimate the action and utterance likelihood, identifying the key actions and/or utterances that reveal the true mental state of an agent (as shown in Figure 4). Second, when a VLM fails to recognize the exact object that the agent is interacting with, LIMP can fill in this missing information with context from the text input. We evaluated the action accuracy by using semantic similarity and found that this approach increases inferred action accuracy from 54.4% to 86.6%. As a result, LIMP is able to perform inference on much more accurate information.

Further ablation studies highlight the critical role of individual LIMP components in its performance. GPT-4o provided with ground truth actions achieves an accuracy of only 53.2%, and LIMP without inverse planning achieves an accuracy of 55.3%. Both are much lower than LIMP's performance of 76.6%, showing the importance of inverse planning.

How general is LIMP? Prior inverse planning models, including BIP-ALM, all require handcrafted representations for specific domains. LIMP, however, represents all information using natural language, which enables the direct use of any pretrained LLMs and VLMs without domain-specific knowledge or finetuning. By utilizing powerful pretrained VLMs for visual perception, LIMP can directly recognize actions from RGB videos in an open-ended way, without specifying target action labels for a domain. LIMP also leverages an LLM to use contextual clues from the text to fill in missing information from visual perception, providing a general method for multi-modal information fusion. One can also easily upgrade LIMP by plugging in *any* state-of-the-art VLMs and LLMs.

What are the limitations of LIMP? Hallucinations created by the VLM can cause significant errors in LIMP. For example, an agent may only open and close the fridge, but the VLM may mistakenly think that the agent also grabs something from the fridge. Such hallucinations in action recognition can not be corrected by the textural context. As a result, LIMP will incorrectly interpret the agent's behavior. Additionally, LIMP does not explicitly infer an agent's belief of another agent's belief. It instead prompts an LLM with past actions and utterances to implicitly infer that, which can become costly for longer events. LIMP also does not perform recursive reasoning for more than two levels.

What are the limitations of our benchmark? The scenarios in our benchmark are currently limited to interactions between two agents in household settings, where there are three social goals: helping, hindering, and acting independently. Moreover, the current benchmark has synthetic human activities. These synthetic activities are realistic as verified in prior work (Puig et al. 2020) and enable large-scale testing. However sim-to-real evaluation could be valuable for future studies.

Conclusion

We present the first multi-modal Theory of Mind benchmark for multi-agent interactions in complex embodied settings. We have systematically evaluated humans and state-of-the-art LMMs on our benchmark. We have also proposed a novel multi-modal ToM model that outperforms all baselines while maintaining generality. In future work, we intend to incorporate more complex real-world scenarios beyond household environments and introduce multi-modal social interactions involving more than two agents. We also plan to create a test set with real-world videos for ToM evaluation in real-world scenarios.

References

Amirizaniani, M.; Martin, E.; Sivachenko, M.; Mashhadi, A.; and Shah, C. 2024. Do LLMs Exhibit Human-Like Reasoning? Evaluating Theory of Mind in LLMs for Open-Ended Responses. *arXiv preprint arXiv:2406.05659*.

Baker, C. L.; Jara-Ettinger, J.; Saxe, R.; and Tenenbaum, J. B. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4): 1–10.

Baron-Cohen, S.; Leslie, A. M.; and Frith, U. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1): 37–46.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Chan, C.; Jiayang, C.; Yim, Y.; Deng, Z.; Fan, W.; Li, H.; Liu, X.; Zhang, H.; Wang, W.; and Song, Y. 2024. NegotiationToM: A Benchmark for Stress-testing Machine Theory of Mind on Negotiation Surrounding. *arXiv preprint arXiv:2404.13627*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*.

Chen, Z.; Wu, J.; Zhou, J.; Wen, B.; Bi, G.; Jiang, G.; Cao, Y.; Hu, M.; Lai, Y.; Xiong, Z.; and Huang, M. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. arXiv:2402.15052.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; and Bing, L. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Denham, S. A.; Blair, K. A.; DeMulder, E.; Levitas, J.; Sawyer, K.; Auerbach-Major, S.; and Queenan, P. 2003.

Preschool emotional competence: Pathway to social competence? *Child Development*, 74(1): 238–256.

Gandhi, K.; Stojnic, G.; Lake, B. M.; and Dillon, M. R. 2021. Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others. *Advances in Neural Information Processing Systems*, 34: 9963–9976.

Gmytrasiewicz, P. J.; and Doshi, P. 2005. A Framework for Sequential Planning in Multi-Agent Settings. *Journal of Artificial Intelligence Research*, 24: 49–79.

Gordon, A. S. 2016. Commonsense Interpretation of Triangle Behavior. In *AAAI Conference on Artificial Intelligence*.

Hamlin, J. K.; Wynn, K.; and Bloom, P. 2007. Social evaluation by preverbal infants. *Nature*, 450(7169): 557–559.

Han, Y.; and Gmytrasiewicz, P. 2019. IPOMDP-Net: A Deep Neural Network for Partially Observable Multi-Agent Planning Using Interactive POMDPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 6062–6069.

He, Y.; Wu, Y.; Jia, Y.; Mihalcea, R.; Chen, Y.; and Deng, N. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*.

Ivanova, A. A.; Sathe, A.; Lipkin, B.; Kumar, U.; Radkani, S.; Clark, T. H.; Kauf, C.; Hu, J.; Pramod, R. T.; Grand, G.; Paulun, V.; Ryskina, M.; Akyürek, E.; Wilcox, E.; Rashid, N.; Choshen, L.; Levy, R.; Fedorenko, E.; Tenenbaum, J.; and Andreas, J. 2024. Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. arXiv:2405.09605.

Jha, K.; Le, T. A.; Jin, C.; Kuo, Y.-L.; Tenenbaum, J. B.; and Shu, T. 2024. Neural Amortized Inference for Nested Multi-agent Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 530–537.

Jin, C.; Wu, Y.; Cao, J.; Xiang, J.; Kuo, Y.-L.; Hu, Z.; Ullman, T.; Torralba, A.; Tenenbaum, J. B.; and Shu, T. 2024. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*.

Kim, H.; Sclar, M.; Zhou, X.; Bras, R. L.; Kim, G.; Choi, Y.; and Sap, M. 2023. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. *arXiv preprint arXiv:2310.15421*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. *ArXiv*, abs/2205.11916.

Kosinski, M. 2023. Theory of Mind May Have Spontaneously Emerged in Large Language Models. *arXiv preprint arXiv:2302.02083*.

Le, M.; Boureau, Y.-L.; and Nickel, M. 2019. 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)*, 5872–5877.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Li, L.; Yin, Y.; Li, S.; Chen, L.; Wang, P.; Ren, S.; Li, M.; Yang, Y.; Xu, J.; Sun, X.; Kong, L.; and Liu, Q. 2023b. M^3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. arXiv:2306.04387.

Li, W.; Yasuda, S. C.; Dillon, M. R.; and Lake, B. 2024. An Infant-Cognition Inspired Machine Benchmark for Identifying Agency, Affiliation, Belief, and Intention. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.

Malik, M.; and Isik, L. 2023. Relational visual representations underlie human social interaction recognition. *Nature Communications*, 14(1): 7317.

Netanyahu, A.; Shu, T.; Katz, B.; Barbu, A.; and Tenenbaum, J. B. 2021. Phase: Physically-grounded abstract social events for machine social perception. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, 845–853.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Pandya, P.; Talwarr, A. S.; Gupta, V.; Kataria, T.; Gupta, V.; and Roth, D. 2024. NTSEBENCH: Cognitive Reasoning Benchmark for Vision Language Models. arXiv:2407.10380.

Patraucean, V.; Smaira, L.; Gupta, A.; Recasens, A.; Markeeva, L.; Banarse, D.; Koppula, S.; Malinowski, M.; Yang, Y.; Doersch, C.; et al. 2024. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36.

Puig, X.; Ra, K.; Boben, M.; Li, J.; Wang, T.; Fidler, S.; and Torralba, A. 2018. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8494–8502.

Puig, X.; Shu, T.; Li, S.; Wang, Z.; Tenenbaum, J. B.; Fidler, S.; and Torralba, A. 2020. Watch-And-Help: A Challenge for Social Perception and Human-AI Collaboration. arXiv:2010.09890.

Puig, X.; Shu, T.; Tenenbaum, J. B.; and Torralba, A. 2023. NOPA: Neurally-guided Online Probabilistic Assistance for Building Socially Intelligent Home Assistants. *arXiv preprint arXiv:2301.05223*.

Rabinowitz, N. C.; Perbet, F.; Song, H. F.; Zhang, C.; Eslami, S. M. A.; and Botvinick, M. 2018. Machine Theory of Mind. arXiv:1802.07740.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Sabour, S.; Liu, S.; Zhang, Z.; Liu, J. M.; Zhou, J.; Sunaryo, A. S.; Li, J.; Lee, T. M. C.; Mihalcea, R.; and Huang, M. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. arXiv:2402.12071.

Sanders, K.; Etter, D.; Kriz, R.; and Van Durme, B. 2023. MultiVENT: Multilingual Videos of Events with Aligned Natural Text. *arXiv preprint arXiv:2307.03153*.

Sclar, M.; Kumar, S.; West, P.; Suhr, A.; Choi, Y.; and Tsvetkov, Y. 2023a. Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. arXiv:2306.00924.

Sclar, M.; Kumar, S.; West, P.; Suhr, A.; Choi, Y.; and Tsvetkov, Y. 2023b. 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)*, 13960–13980. Toronto, Canada: Association for Computational Linguistics.

Shu, T.; Bhandwaldar, A.; Gan, C.; Smith, K.; Liu, S.; Gutfreund, D.; Spelke, E.; Tenenbaum, J.; and Ullman, T. 2021. Agent: A benchmark for core psychological reasoning. In *International Conference on Machine Learning*, 9614–9625. PMLR.

Shu, T.; Kryven, M.; Ullman, T. D.; and Tenenbaum, J. 2020. Adventures in Flatland: Perceiving Social Interactions Under Physical Dynamics. In *CogSci*.

Soubki, A.; Murzaku, J.; Jordehi, A. Y.; Zeng, P.; Markowska, M.; Mirroshandel, S. A.; and Rambow, O. 2024. Views Are My Own, But Also Yours: Benchmarking Theory of Mind using Common Ground. *arXiv preprint arXiv:2403.02451*.

Stacy, S.; Gong, S.; Parab, A.; Zhao, M.; Jiang, K.; and Gao, T. 2024. A Bayesian theory of mind approach to modeling cooperation and communication. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(1): e1631.

Tang, J.; Liu, Q.; Ye, Y.; Lu, J.; Wei, S.; Lin, C.; Li, W.; Mahmood, M. F. F. B.; Feng, H.; Zhao, Z.; et al. 2024. MTVQA: Benchmarking Multilingual Text-Centric Visual Question Answering. *arXiv preprint arXiv:2405.11985*.

Tejwani, R.; Kuo, Y.-L.; Shu, T.; Katz, B.; and Barbu, A. 2021. Social interactions as recursive mdps. In *Conference on Robot Learning*, 949–958. PMLR.

Ullman, T. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv:2302.08399.

Ullman, T.; Baker, C.; Macindoe, O.; Evans, O.; Goodman, N.; and Tenenbaum, J. 2009. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22.

Verma, M.; Bhambri, S.; and Kambhampati, S. 2024. Theory of Mind abilities of Large Language Models in Human-Robot Interaction: An Illusion? In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 36–45.

Wellman, H. M.; Cross, D.; and Watson, J. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3): 655–684.

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

Wu, Y.; He, Y.; Jia, Y.; Mihalcea, R.; Chen, Y.; and Deng, N. 2023. Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10691–10706. Singapore: Association for Computational Linguistics.

Xu, H.; Zhao, R.; Zhu, L.; Du, J.; and He, Y. 2024. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. *arXiv preprint arXiv:2402.06044*.

Yang, S.; Ge, Y.; Li, Y.; Chen, Y.; Ge, Y.; Shan, Y.; and Chen, Y. 2024. SEED-Story: Multimodal Long Story Generation with Large Language Model. *arXiv preprint arXiv:2407.08683*.

Ying, K.; Meng, F.; Wang, J.; Li, Z.; Lin, H.; Yang, Y.; Zhang, H.; Zhang, W.; Lin, Y.; Liu, S.; et al. 2024a. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

Ying, L.; Jha, K.; Aarya, S.; Tenenbaum, J. B.; Torralba, A.; and Shu, T. 2024b. GOMA: Proactive Embodied Cooperative Communication via Goal-Oriented Mental Alignment. arXiv:2403.11075.

Zhang, J.; Huang, W.; Ma, Z.; Michel, O.; He, D.; Gupta, T.; Ma, W.-C.; Farhadi, A.; Kembhavi, A.; and Krishna, R. 2024. Task Me Anything. *arXiv preprint arXiv:2406.11775*.

Zhi-Xuan, T.; Mann, J.; Silver, T.; Tenenbaum, J.; and Mansinghka, V. 2020. Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, 33: 19238–19250.