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

1 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.

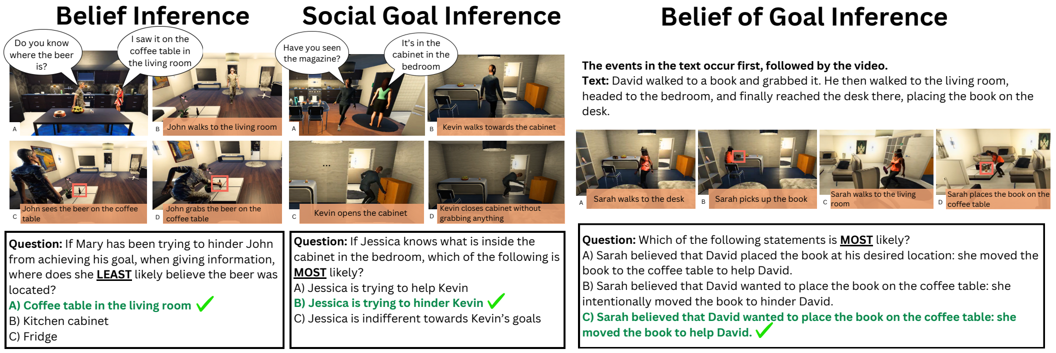
2 Introduction

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 et al., 2001; Hamlin et al., 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. The multi-agent Theory of Mind abilities are not only crucial for humans but also for AI systems. Without a robust understanding of people's mental states in social interactions, AI systems may cause detrimental errors in their interactions with people.

To address the challenges of multi-agent Theory of Mind reasoning, 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

*Denotes equal contribution
Code and data are available at: `https://github.com/SCAI-JHU/MuMA-ToM`

NeurIPS 2024 Workshop on Behavioral Machine Learning.


Belief Inference
Do you know where the beer is?
I saw it on the coffee table in the living room
A
B
John walks to the living room
C
John sees the beer on the coffee table
D
John grabs the beer on the coffee table
Social Goal Inference
Have you seen the magazine?
It's in the cabinet in the bedroom
A
B
Kevin walks towards the cabinet
...
C
Kevin opens the cabinet
D
Kevin closes cabinet without grabbing anything
Belief of Goal Inference
A
Sarah walks to the desk
B
Sarah picks up the book
C
Sarah walks to the living room
D
Sarah places the book on the coffee table

Question: If Mary has been trying to hinder John from achieving his goal, when giving information, where does she **LEAST** likely believe the beer was located?

A) Coffee table in the living room ✓

B) Kitchen cabinet

C) Fridge

Question: If Jessica knows what is inside the cabinet in the bedroom, which of the following is **MOST** likely?

A) Jessica is trying to help Kevin

B) Jessica is trying to hinder Kevin ✓

C) Jessica is indifferent towards Kevin's goals

The events in the text occur first, followed by the video.

Text: David walked to a book and grabbed it. He then walked to the living room, headed to the bedroom, and finally reached the desk there, placing the book on the desk.

Question: Which of the following statements is **MOST** likely?

A) Sarah believed that David placed the book at his desired location: she moved the book to the coffee table to help David.

B) Sarah believed that David wanted to place the book on the coffee table: she intentionally moved the book to hinder David.

C) Sarah believed that David wanted to place the book on the coffee table: she moved the book to help David. ✓

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.

(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. 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. 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 combining inverse multi-agent planning and language models.

3 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.

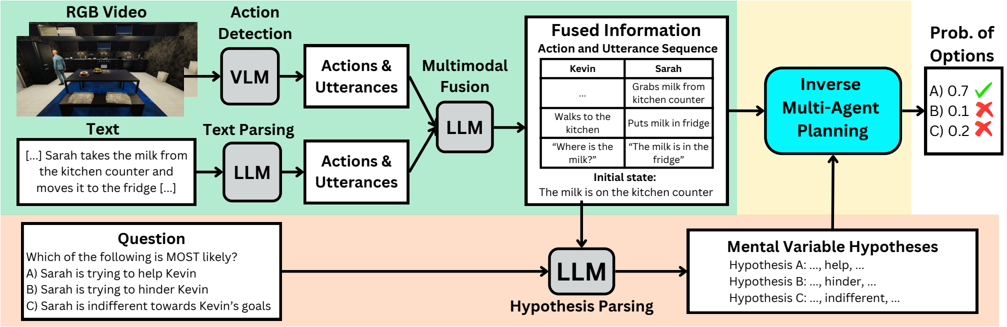

RGB Video
Action Detection
VLM
Actions & Utterances
Text
[...] Sarah takes the milk from the kitchen counter and moves it to the fridge [...]
Text Parsing
LLM
Actions & Utterances
Multimodal Fusion
LLM
Fused Information
Action and Utterance Sequence
Kevin
Sarah
...
Grabs milk from kitchen counter
Walks to the kitchen
Puts milk in fridge
"Where is the milk?"
"The milk is in the fridge"
Initial state:
The milk is on the kitchen counter
Inverse Multi-Agent Planning
Prob. of Options
A) 0.7 ✓
B) 0.1 ✗
C) 0.2 ✗
Question
Which of the following is MOST likely?
A) Sarah is trying to help Kevin
B) Sarah is trying to hinder Kevin
C) Sarah is indifferent towards Kevin's goals
LLM
Hypothesis Parsing
Mental Variable Hypotheses
Hypothesis A: ..., help, ...
Hypothesis B: ..., hinder, ...
Hypothesis C: ..., indifferent, ...

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.

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. More details can be found in section 8 of the appendix.

Multi-modal Information Unlike MMToM-QA, the only prior multi-modal ToM QA benchmark, our benchmark has completely separate information in different modalities. 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., 2018a, 2020), to 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 in section 11 of the appendix.

4 Our Model

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, LIMP identifies three mental variables crucial to understanding multi-agent interactions—belief, social goal, and belief of goal. Second, 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,

4.1 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. 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. 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 sequences, we can infer the state at each step.

There is often missing information in the visual perception results. Sometimes the VLM may fail to recognize the object a person has picked up, resulting in an ambiguous description like "grabs some object." This is also a challenge for human observers, as the object picked up by the person is often occluded. However, humans can still infer the most likely object 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. An example of this is shown in figure 3 in the appendix.

In this work, we use Gemini 1.5 Pro for the VLM and GPT-4o for the LLM as they produce the best results.

4.2 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, $H = \langle b(s), g_i, b(g_j) \rangle$.

4.3 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:

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

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

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.

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. We find that GPT-4o can accurately estimate the action and utterance policies based on the given condition.

5 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** with finetuned Llama 2 (Jin et al., 2024), the best-performing model on a prior multi-modal ToM benchmark, MMToM-QA.

Results We report the human and model performance in Table 2. 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. The best-performing LMM baseline is Gemini 1.5 Pro, but its overall accuracy is only 56.4%. 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. Notably, BIP-ALM had an accuracy of 33.9%, indicating its inability to understand multi-agent interactions. Our LIMP model significantly outperforms all state-of-the-art models on our benchmark, with an overall accuracy of 76.6%. There is still a gap between the best model performance and human performance, highlighting the need for further studies.

6 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

Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses. *arXiv preprint arXiv:2406.05659*, 2024.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding. *arXiv preprint arXiv:2404.13627*, 2024.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. Tombench: Benchmarking theory of mind in large language models, 2024. URL `https://arxiv.org/abs/2402.15052`.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. URL `https://arxiv.org/abs/2406.07476`.

Susanne A. Denham, Kimberly A. Blair, Elizabeth DeMulder, Jennifer Levitas, Katherine Sawyer, Sharon Auerbach-Major, and Patrick Queenan. Preschool emotional competence: Pathway to social competence? *Child Development*, 74(1):238–256, 2003. doi: 10.1111/1467-8624.00533. URL `https://doi.org/10.1111/1467-8624.00533`.

P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, July 2005. ISSN 1076-9757. doi: 10.1613/jair.1579. URL `http://dx.doi.org/10.1613/jair.1579`.

J Kiley Hamlin, Karen Wynn, and Paul Bloom. Social evaluation by preverbal infants. *Nature*, 450 (7169):557–559, 2007.

Yanlin Han and Piotr Gmytrasiewicz. 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, Jul. 2019. doi: 10.1609/aaai.v33i01.33016062. URL `https://ojs.aaai.org/index.php/AAAI/article/view/4562`.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind. *arXiv preprint arXiv:2407.01455*, 2024.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models, 2024. URL `https://arxiv.org/abs/2405.09605`.

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

Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Yuan Yuan, Zhuoqun Hao, Xinyi Bai, Weijie J. Su, Camillo J. Taylor, and Tanwi Mallick. Towards rationality in language and multimodal agents: A survey, 2024. URL https://arxiv.org/abs/2406.00252.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*, 2024.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022. URL https://api.semanticscholar.org/CorpusID:249017743.

Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. M^{3}it: A large-scale dataset towards multi-modal multilingual instruction tuning, 2023b. URL https://arxiv.org/abs/2306.04387.

Wenjie Li, Shannon C Yasuda, Moira Rose Dillon, and Brenden Lake. 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, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

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

Aviv Netanyahu, Tianmin Shu, Boris Katz, Andrei Barbu, and Joshua B Tenenbaum. Phase: Physically-grounded abstract social events for machine social perception. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, pages 845–853, 2021.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Pranshu Pandya, Agney S Talwarr, Vatsal Gupta, Tushar Kataria, Vivek Gupta, and Dan Roth. Ntsebench: Cognitive reasoning benchmark for vision language models, 2024. URL https://arxiv.org/abs/2407.10380.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018a.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs, 2018b. URL https://arxiv.org/abs/1806.07011.

Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Joshua B. Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration, 2020.

Xavier Puig, Tianmin Shu, Joshua B Tenenbaum, and Antonio Torralba. Nopa: Neurally-guided online probabilistic assistance for building socially intelligent home assistants. *arXiv preprint arXiv:2301.05223*, 2023.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind, 2018. URL https://arxiv.org/abs/1802.07740.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Juanzi Li, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. Emobench: Evaluating the emotional intelligence of large language models, 2024. URL https://arxiv.org/abs/2402.12071.

Kate Sanders, David Etter, Reno Kriz, and Benjamin Van Durme. Multivent: Multilingual videos of events with aligned natural text. *arXiv preprint arXiv:2307.03153*, 2023.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada, July 2023a. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.780.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker, 2023b.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker, 2023c. URL https://arxiv.org/abs/2306.00924.

Tianmin Shu, Marta Kryven, Tomer D Ullman, and Josh Tenenbaum. Adventures in flatland: Perceiving social interactions under physical dynamics. In *CogSci*, 2020.

Stephanie Stacy, Siyi Gong, Aishni Parab, Minglu Zhao, Kaiwen Jiang, and Tao Gao. A bayesian theory of mind approach to modeling cooperation and communication. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(1):e1631, 2024.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024.

Ravi Tejwani, Yen-Ling Kuo, Tianmin Shu, Boris Katz, and Andrei Barbu. Social interactions as recursive mdps. In *Conference on Robot Learning*, pages 949–958. PMLR, 2021.

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

Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua Tenenbaum. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22, 2009.

Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 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*, pages 36–45, 2024.

H M. Wellman, D Cross, and J Watson. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3):655–684, 2001. doi: 10.1111/1467-8624.00304. URL https://doi.org/10.1111/1467-8624.00304.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities, 2023. URL https://arxiv.org/abs/2311.10227.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*, 2024. URL https://arxiv.org/abs/2407.08683.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024a.

Lance Ying, Kunal Jha, Shivam Aarya, Joshua B. Tenenbaum, Antonio Torralba, and Tianmin Shu. Goma: Proactive embodied cooperative communication via goal-oriented mental alignment, 2024b. URL `https://arxiv.org/abs/2403.11075`.

Dong Zhang, Zhaowei Li, Pengyu Wang, Xin Zhang, Yaqian Zhou, and Xipeng Qiu. Speechagents: Human-communication simulation with multi-modal multi-agent systems, 2024a. URL `https://arxiv.org/abs/2401.03945`.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. URL `https://arxiv.org/abs/2306.02858`.

Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task me anything. *arXiv preprint arXiv:2406.11775*, 2024b.

Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, and Vikash Mansinghka. Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, 33:19238–19250, 2020.

Appendix

7 Related Works

Theory of Mind Benchmarks Most multi-agent benchmarks focus on single-agent beliefs and intentions, without exploring inter-agent relationships. (Kim et al., 2023; Chen et al., 2024; Chan et al., 2024; Sabour et al., 2024). Prior works on testing social relationship understanding use simple animations and lack realism (Netanyahu et al., 2021; Li et al., 2024). Moreover, existing benchmarks generally have only 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 Most multi-modal QA 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., 2024b), also do not evaluate ToM reasoning.

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 (Kosinski, 2023; Bubeck et al., 2023) and social reasoning Jiang et al. (2024); Zhang et al. (2024a) capabilities in complicated social tasks, but their ToM reasoning is still brittle (Verma et al., 2024; Amirizaniani et al., 2024; Ullman, 2023; Sclar et al., 2023b; Ivanova et al., 2024; Jiang et al., 2024), suffering from reasoning errors and rationality grounding. 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., 2023a). (Jin et al., 2024) proposed, BIP-ALM, for multi-modal ToM. 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.

8 Question Type Details

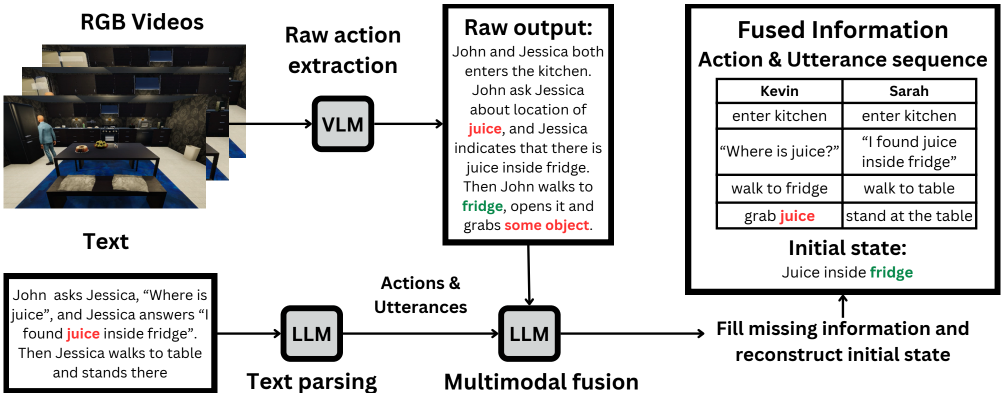
In our MuMA-ToM benchmark, we have three types of questions: belief inference, social goal inference, belief of goal inference. Three question types are explained as follows.

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 et al. (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 hinder 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.

9 LIMP Details



John and Jessica both enters the kitchen. John ask Jessica about location of juice, and Jessica indicates that there is juice inside fridge. Then John walks to fridge, opens it and grabs some object.

Kevin	Sarah
enter kitchen	enter kitchen
"Where is juice?"	"I found juice inside fridge"
walk to fridge	walk to table
grab juice	stand at the table

Juice inside fridge

John asks Jessica, "Where is juice", and Jessica answers "I found juice inside fridge". Then Jessica walks to table and stands there

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.

9.1 Multi-modal Information Fusion

Figure 3 shows a detailed example of the text parsing and multimodal fusion process. In this example, the VLM was unable to see the occluded object that John grabs from the fridge. However, the LLM was able to figure out that John grabs the juice based on context from the question.

For processing textual information, we directly use GPT-4o to parse the actions and utterances of each agent separately, in chronological order. Then, this parsed text information, along with the raw visual outputs from text input as well as raw visual outputs from Gemini, is provided to GPT-4o for information fusion.

A key step in our multi-modal fusion process is filling in missing information from the visual output based on the context. In the prompt given to Gemini, we instruct the model to leave blanks for exact object names, as accurately recognizing small or obscured objects is often impossible and could lead to unreasonable results. The raw visual output, along with text input that provides necessary context, is then used by GPT-4o to fill in these blanks with the correct object names mentioned in the context. This method reduces the model's reliance on recognizing small objects directly, and takes a more human-like approach to the problem.

Another important step in the multi-modal fusion process is initial state retrieval. The initial state of the environment is crucial for the planning process, as the agents' beliefs are based on the initial state instead of the changed state, unless they observe other agent moving things around directly. Since we do not use instance segmentation, it is challenging for the model to directly identify object locations or generate scene graphs from visual input. Instead, we use the agents' actions to infer the initial state of the environment. This reduces uncertainty for the model and allows it to focus on relevant objects to the interaction while ignoring unrelated ones.

9.2 Hypothesis Parsing

We identify the three latent variables: belief, social goal and the belief of goal for understanding social interactions. The questions are designed in a way that for each option, there will be a set of these three latent variables corresponding to it. In the latent variable extraction stage, GPT-4 is prompted to extract the three sets. Initial state and actions of agents are also given as context as there are descriptions like "knows the location of the object" or "has put the object at desired location" requiring checking action & initial state to figure out the exact location of the object.

9.3 Inverse Multi-Agent Planning

Unlike open-source models, GPT-4o does not provide the log probability for any given completion, so the exact probability of the utterance or action cannot be calculated. However, GPT-4o does offer the log probabilities for the top 5 responses it generates. To address this, we implement a method that asks GPT-4o to assess the likelihood of a given utterance or action and restricts its most likely responses to two choices: A) Likely, or B) Unlikely. We then calculate the probability of the completion by using the log probability of the token 'A'.

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).

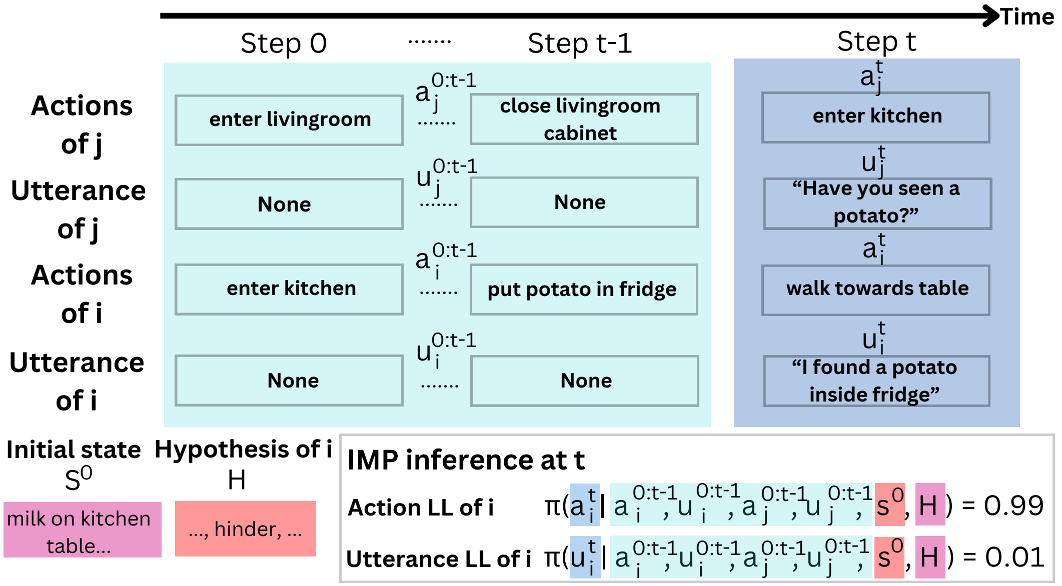


Figure 4: Illustration for inverse multi-agent planning. We estimate the action and utterance likelihood of agent i at each step t given the past actions and utterances of both agents from step 0 to step $t-1$, the initial state s^0, and the hypothesis H. LL in the figure stands for likelihood.

10 More Experiment Results

We show the additional experiment results in Table **??**

10.1 Chain of Thought Prompting.

We evaluate state-of-the-art models' performance on our dataset with zero-shot chain of thought (CoT) prompting, as introduced by Kojima et al. (2022). We add the phrase "Let's think step by step" after the question prompt but before the list of options.

For all models tested, using CoT prompting showed no significant improvement in performance. In fact, for many models, using CoT prompting caused a decrease in performance. While there are instances where CoT led to some improvement, such as in belief inference for InternVL 2 26B, the overall impact effect was negligible on more challenging social goal and belief of goal inference questions. These results further highlight the current limitations of state-of-the-art LMMs. Even with CoT guidance, they struggle to effectively understand social interactions.

10.2 Finetuned Baseline

We finetuned the VideoLlama 2 7B model on our training set for action captioning tasks following Zhang et al. (2023), using two A100 GPUs for 1 epoch, with a learning rate of 2e-5 and a batch size of 4. The performance of the model was lower after finetuning, suggesting that the model may have inherent limitations in ToM reasoning or action recognition. We experimented with finetuning for up to 3 epochs and found that extending finetuning beyond one epoch leads to over-fitting, and the model was unable to answer the questions with A, B, or C.

10.3 Advanced Prompting for ToM.

Recent works have leveraged language models to tackle ToM problems through multi-step reasoning approaches Wilf et al. (2023); Sclar et al. (2023c); Hou et al. (2024). Among these text-only models, we chose to evaluate SimToM, as the code for the other models was either unavailable or required extensive modifications to integrate with our benchmark. Since SimToM only accepts textual input, we adapted it to our dataset by adding Gemini 1.5 Pro's visual extraction results after the textual input as input for SimToM and tested it with GPT-4o serving as the primary language model. SimToM, which analyzes the perspective of each agent to assist the language model, achieved the highest accuracy in belief-of-goal questions among all the baselines tested. This suggests that a multi-step approach can improve a language model's capacity for ToM reasoning. However, the overall accuracy is still below 50%.

10.4 LIMP w/ Llama 3.1 8B for Inverse Multi-agent Planning

Solving ToM problems with language models usually requires some form of finetuning or few-shot prompting to equip the model with domain-specific knowledge. In contrast, LIMP leverages the forward planning capabilities of language models to address the inverse planning problem without any finetuning or additional domain knowledge. Beyond testing very large models like GPT-4o, we also explored the potential of smaller models, such as Llama 3.1 8B, as an inverse planner for LIMP. However, the results indicate that smaller models lack the ability to effectively function as inverse planners for multi-agent actions. A closer qualitative examination of Llama 8B's failure patterns shows that the model is unable to understand the concept of hindering, which leads to poor performance across all questions related to hindering.

11 Procedural Generation Details

Figure 5 summarizes the procedural generation process. We follow a recent paper GOMA Ying et al. (2024b) to generate actions & utterance sequence, use the virtualhome Puig et al. (2018b) 3D simulator to generate humanoid actions within a realistic household environment and use GPT-4o to generate texts and questions.

Method	Belief Inference	Social Goal Inference	Belief of Goal Inference	All
Llava 1.6 34B	93.6	37.2	27.5	52.8
Llava 1.6 34B CoT	93.2	46.1	19.4	52.9
Llava 1.6 13B	70.2	43.2	17.9	43.7
Llava 1.6 13B CoT	64.9	41.6	25.3	43.9
Gemini 1.5 Flash	53.9	33.0	41.4	42.7
Gemini 1.5 Flash CoT	56.7	35.6	41.4	43.6
Gemini 1.5 Pro	78.9	43.9	46.9	56.4
Gemini 1.5 Pro CoT	79.8	42.6	41.1	54.5
GPT-4o	67.9	39.6	44.4	50.6
GPT-4o CoT	62.2	33.6	39.8	45.2
InternVL 2 8B	62.2	44.6	45.1	50.6
InternVL 2 8B CoT	57.7	44.9	43.5	48.7
InternVL 2 26B	59.3	44.9	35.5	46.6
InternVL 2 26B CoT	64.1	44.9	36.1	48.4
VideoLlama 2 7B	70.1	45.6	37.7	51.1
VideoLlama 2 7B CoT	51.8	42.9	34.9	42.8
VideoLlama 2 7B (finetuned)	42.7	35.7	34.3	37.3
SimToM	54.6	43.5	44.8	47.6
LIMP with Llama 3.1 8B	35.8	23.4	37.7	33.0
BIP-ALM	41.2	34.1	30.6	33.9
LIMP with GPT-4o	93.4	67.7	68.7	76.6

Table 2: All experiment results: For models that accept video input, the full videos were provided. For models that do not, uniformly sampled frames (every 20 frames) were used instead. Since SimToM is a text-based model, we provided it with the action recognition outputs from Gemini 1.5 Pro.

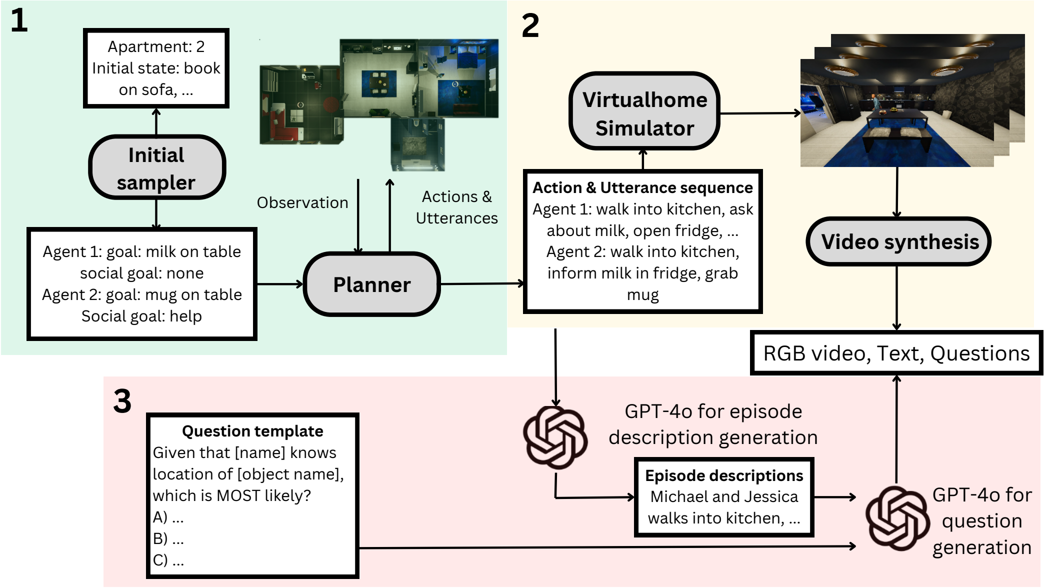

1
Apartment: 2
Initial state: book on sofa, ...
Initial sampler
Agent 1: goal: milk on table
social goal: none
Agent 2: goal: mug on table
Social goal: help
Observation
Actions & Utterances
Planner
2
Virtualhome Simulator
Action & Utterance sequence
Agent 1: walk into kitchen, ask about milk, open fridge, ...
Agent 2: walk into kitchen, inform milk in fridge, grab mug
Video synthesis
RGB video, Text, Questions
3
Question template
Given that [name] knows location of [object name], which is MOST likely?
A) ...
B) ...
C) ...
GPT-4o for episode description generation
Episode descriptions
Michael and Jessica walks into kitchen, ...
GPT-4o for question generation

Figure 5: Overview of the Procedural generation process. This method ensures that the episodes and ground truth answers are factually correct, while maintaining realistic conversations and scenarios.

Step 1 in Figure 5 shows the action & utterance sequence generation process. We use four different apartments as the base environment for two agents' interactions, sampling objects to different containers & surfaces within the apartment to generate a distinctive environment for each interactive scenario. Two agents' initial location (room location), physical goal (finding or rearranging an object), initial belief (ground-truth belief, false belief, or uniform belief), and social intentions (help, hinder, independent) are also sampled. For interactive scenarios without language, we sampled the environment and agents' goal in a way that ensures two agents' are aiming to put the same object to different locations and there is only one object of that type in the environment. In this way, agents will have to rearrange the object after the other agent has placed the object. Afterward, a Monte Carlo

Tree Search (MCTS) planner is used to compute the action sequence for each agent. The utterance is computed separately: for each step, if the two agents are in the same room and the first agent is uncertain about its goal object's location (entropy of its belief probability distribution exceeds a threshold), the first agent will send an inquiry. Upon receiving the inquiry, the second agent will answer based on its social intention (provide a contradictory answer with its belief when trying to hinder), and the first agent will update its belief accordingly. As agents' beliefs do not necessarily match the ground-truth state, the combination of intention with the ground-truth environment state is complicated: for instance, providing false information can be interpreted as trying to help but failing due to mistaken belief or deliberately trying to hinder. After the original utterance is generated, we use GPT-4o to add variety and improve the quality of language communication.