

# Appendix

## 9 Comparison of ToM Benchmarks

Table 2 provides a comparison between our MuMA-ToM benchmark and prior ToM benchmarks, highlighting key features such as the size of the test set, input modalities, and evaluation metrics. Our benchmark stands out as the only benchmark with multi-modal inputs and multi-agent interactions. It simultaneously evaluates multi-agent social interactions with belief, goal, and belief of other agents’ goals, as well as the ability to infer mental states from multi-modal inputs.

## 10 MuMA-ToM Benchmark Details

### 10.1 More Quantitative Results

The results of all experiments conducted in our study are shown in Table 3.

**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.

**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.

**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%.

**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.

Benchmark	Agent number	Tested concepts	Size	Modality	Comm	Evaluation
<b>Triangle COPA Gordon (2016a)</b>	Single agent	Social Interaction	100	Text	No	Multiple choice Q&A
<b>ToMi Le et al. (2019)</b>	Multi agents	First & Second Order belief	400	Text	No	Multiple choice Q&A
<b>Phase Ne-tanyahu et al. (2021a)</b>	Multi agents	Goals and Social relationships	500	Video	No	Multiple choice recognition
<b>Agent Shu et al. (2021)</b>	Single agent	Goal Preferences, Action Efficiency, Unobserved Constraints, and Cost-Reward Trade-offs	960	Video	No	Surprise rating
<b>Epistemic reasoning Cohen (2021)</b>	Multi agents	Knowledge and Belief	2000	Text	No	True or false judgements
<b>BIB Gandhi et al. (2021)</b>	Single & Multi agents	Goal Preferences, rational actions, constraints	5000	Video	No	Surprise rating
<b>Adv-CSFB Kosinski (2023)</b>	Single agent	False belief	183	Text	No	Multiple choice filling in the blanks
<b>Hi-ToM He et al. (2023)</b>	Multi agents	High-order beliefs	600	Text	Yes	Multiple choice Q&A
<b>FANToM Kim et al. (2023)</b>	Multi agents	Belief & information tracking	4807	Text	Yes	Question answering
<b>BigToM Gandhi et al. (2024)</b>	Single agent	Belief	5000	Text	No	Question answering
<b>MMTOM-QA Jin et al. (2024)</b>	Single agent	Belief & Goal	600	Text & Video	No	Multiple choice Q&A
<b>TomBench Chen et al. (2024b)</b>	Multi agents	Emotion, desire, intention, knowledge, belief, non-literal communication	5330	Text	Yes	Multiple choice Q&A
<b>OpenToM Xu et al. (2024)</b>	Multi agents	Second-order belief, attitude	696	Text	No	Question answering
<b>Negotiation ToM Chan et al. (2024)</b>	Multi agents	Belief, desire, intention	13800	Text	Yes	Question answering
<b>Infant Cognition Benchmark Li et al. (2024a)</b>	Multi agents	False belief, social goal	2000	Video	No	Surprise rating
<b>Common-ToM Soubki et al. (2024)</b>	Multi agents	High order belief	2104	Text	Yes	True of false judgements
<b>EmoBench Sabour et al. (2024)</b>	Multi agents	Complex emotions, personal beliefs & experiences, emotional cues, perspective taking	200	Text	Yes	Multiple choice Q&A
<b>Our MuMA-ToM benchmark</b>	Multi agents	Belief, social goal and belief of other's goal	900	Text & Video	Yes	Multiple choice Q&A

Table 2: Comparison between MuMA-ToM and prior ToM Benchmarks

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 3: 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.

## 10.2 Qualitative Results

We provide two examples where Gemini 1.5 Pro, the best-performing LMM on the MuMA-ToM benchmark, fails while LIMP succeeds, highlighting the challenges state-of-the-art LMMs face on our benchmark. We also provide an example where hallucinations lead to LIMP also failing to solve the problem.

### Example of Gemini’s due to failure to understand diverse social interactions



Question: Given the above interaction, assuming that Jessica knows what is inside the fridge, which of the following statements is LEAST likely?

- A) When giving information, Jessica was indifferent towards Michael's goals.
- B) When giving information, Jessica has been trying to prevent Michael from finding the beer.
- C) When giving information, Jessica has been trying to help Michael locate the beer.

Correct Answer: C

Gemini Answer: B

The correct answer to this question is 'helping' because it is assumed that Jessica knows there is no beer in the fridge, yet still directs Michael there, suggesting deliberate misleading, so helping is least likely. However, Gemini fails to correctly understand the concept of hindering. It assumes that as long as the agent provides information when asked, it is trying to help, regardless of whether the information is accurate. Since Gemini cannot interpret the act of providing misleading information as hindering, it fails to answer this question correctly. Gemini's explanation for its answer is shown below.

#### Gemini's answer and explanation:

The correct answer is B) When giving information, Jessica has been trying to prevent Michael from finding the beer.

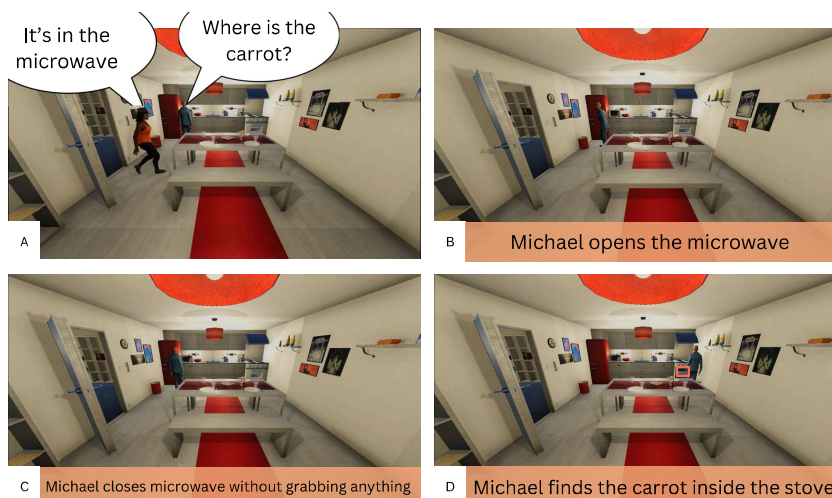
Reason: We can see in the video, after Michael asks where the beer is, Jessica goes to the living room and waits there. When Michael goes to the kitchen, Jessica makes no attempt to stop him, this suggests that she is okay with him finding the beer.

Let's look at why the other answers are incorrect:

A) When giving information, Jessica was indifferent towards Michael's goals - This could be true as Jessica doesn't explicitly state that she wants him to find the beer nor does she show any disappointment when he finds it.

C) When giving information, Jessica has been trying to help Michael locate the beer - This is the most likely answer. Jessica provided the location of the beer to Michael after he asked, making it easier for him to find it.

#### Example of Gemini's failure due to visual recognition failure



Question: Given the above interaction, assuming that Mary knows what is inside the microwave, which of the following statements is MOST likely?

- A) When giving information, Mary was indifferent towards Michael's goals
  - B) When giving information, Mary has been trying to help Michael locate the carrot
  - C) When giving information, Mary has been trying to prevent Michael from finding the carrot
- Correct Answer: C  
Gemini Answer: B

**Gemini's answer and explanation:**

The answer is B) When giving information, Mary has been trying to help Michael locate the carrot. Reasoning: The dialogue clearly shows that Michael is seeking information about the carrot's location. Mary responded directly, indicating she understood his request and provided effective information. Even though Michael's further seek of carrot is not shown, this suggests she was trying to assist him. There's no evidence to support options A or C. Mary's direct and relevant response implies she's not being indifferent or obstructive but rather cooperative.

The correct answer to this question is 'hindering' because it is assumed that Mary knows there is no carrot in the microwave, yet still directs Michael there, suggesting deliberate misleading. Gemini fails to recognize this because it cannot identify the carrot in the video. Without this recognition, it is unable to verify whether Michael actually found the carrot in the microwave and mistakenly interprets Mary's instructions as an attempt to help.

**Example of LIMP's failure due to hallucination**



Question: Given the above interaction, assuming that Sarah knows what is inside the bathroom cabinet, which of the following statements is LEAST likely?

- A) When giving information, Sarah was indifferent towards Michael's goals
  - B) When giving information, Sarah has been trying to prevent Michael from finding the wine
  - C) When giving information, Sarah has been trying to help Michael locate the wine
- Correct Answer: C  
LIMP's answer: B

The question is similar to the previous two, with Sarah providing misleading information. Gemini 1.5 Pro here mistakenly extracts Michael grabbing the wine from the bathroom cabinet while Michael actually does not. As the initial state retrieval process in LIMP relies on the action, LIMP will

mistakenly identify the location of wine as inside the bathroom cabinet and mistakenly interpret Sarah’s misleading words as helping.

### 10.3 Benchmark Statistics

There are 225 interactive scenarios in our MuMA-ToM benchmark, 150 of which have language communication, and 75 of which do not have communication. The episodes in the benchmark are generated to be factually correct, concise, and human readable. Each interactive scenario happens in one of the four apartment with 10+ containers, 10+ surfaces, and 300+ objects in total. There are 17 relevant objects for the agents’ goal in total, distributed among 11 initial locations. Figure 5 shows the distribution of objects’ initial location.

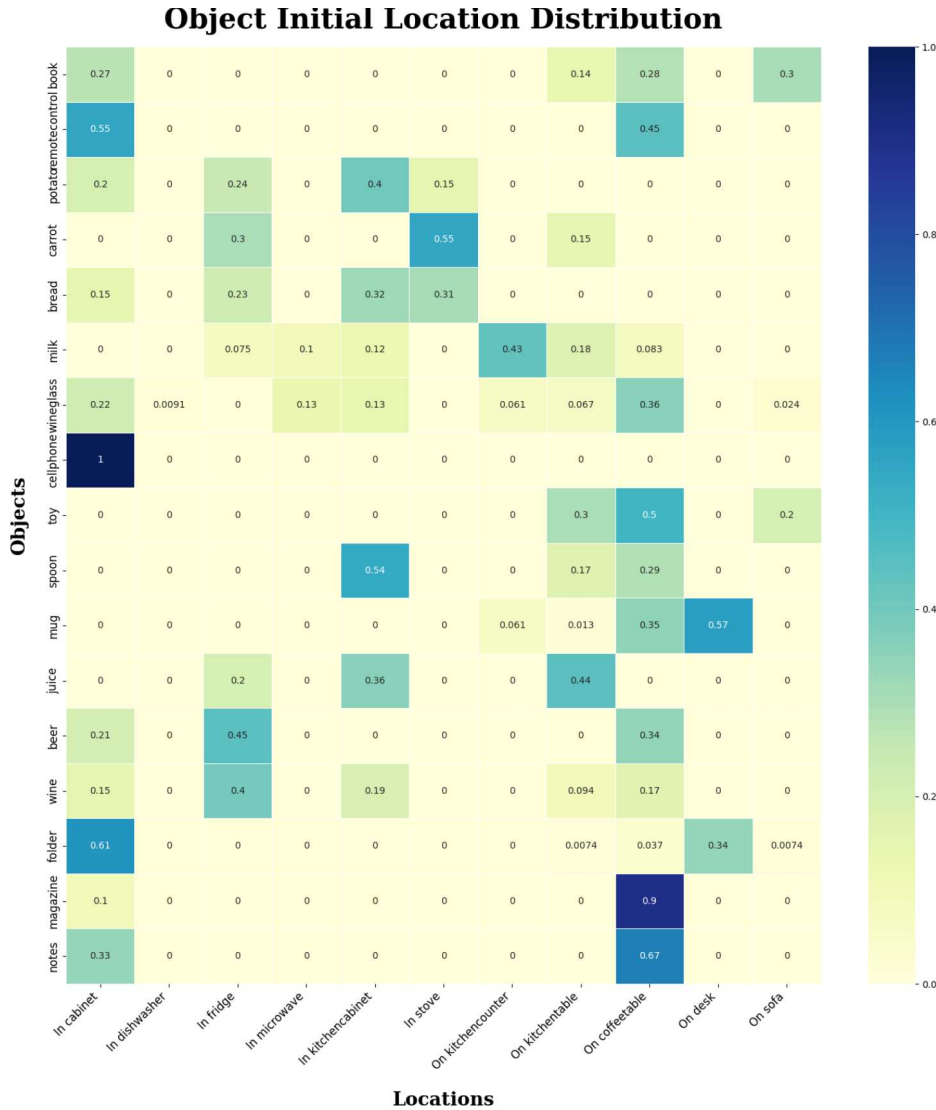


Figure 5: Objects’ initial locations in MuMA-ToM.

Figure 6 shows the distribution of text and video length over all the scenarios. On average, the videos have 364.8 frames (approximately 36 seconds long), and the text inputs contain 136 tokens (many of which are just conversations). The relatively short context length reduces the need for the model to retrieve valid information from a large context, allowing us to focus on testing models’ ToM capability without long-context tracking.

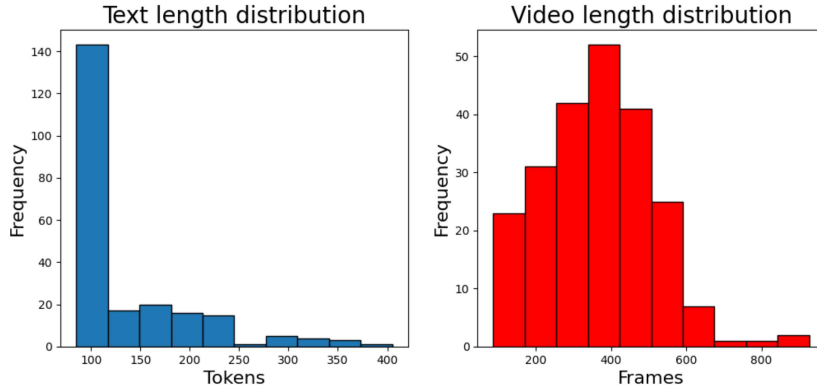


Figure 6: MuMA-ToM context length distribution. The texts and videos are designed to be as concise as possible, allowing us to focus on testing models’ ToM capability without long-context tracking. The videos are rendered at 10 frames per second.

#### 10.4 Available Data

We also provide depth images, instance segmentation, ground-truth actions, states, and camera data for our benchmark in addition to RGB videos and text. Even though our LIMP model does not rely on any of this information to make inferences, this information can be helpful for testing models’ capability of solving ToM problems with some additional information: for example, ground-truth actions and object locations (from instance segmentation).

#### 10.5 Procedural Generation Details

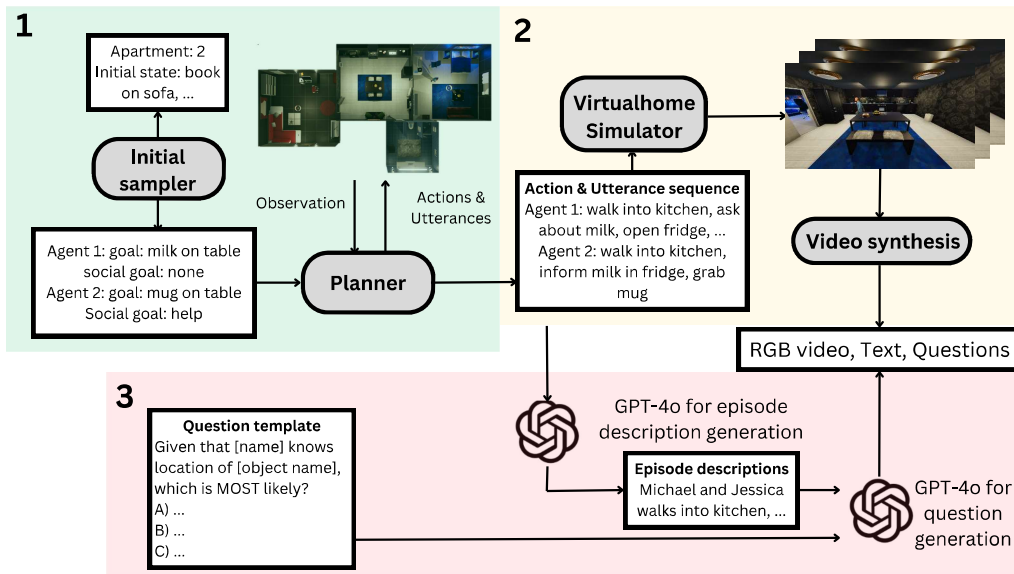


Figure 7: Overview of the Procedural generation process. This method ensures that the episodes and ground truth answers are factually correct, while maintaining realistic conversations and scenarios.

Figure 7 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.

Step 1 in Figure 7 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. The prompts we use are shown below.

**Prompt for adding variety for inquiry**

Objective: Generate natural language from a language template.

User Input: Questions with a basic templated format in the form of "Where is X? Where is Y?"

Instructions: Convert this question into natural conversational language. Make it seem like everyday conversation. If the user asks about multiple objects, combine the objects into a single question.

**Prompt for adding variety to response**

Objective: Generate natural language from language template.

User Input: The locations of an object with a basic templated format, with entries separated by ;. For instance, apple on table 121 livingroom; apple inside fridge 240 kitchen; apple null; banana on counter 101 kitchen; banana null means that there is an apple on the table in the livingroom, an apple inside the fridge in the kitchen, and the location of the third apple is unknown. There is a banana on the kitchen counter, and the location of the other banana is unknown.

Instructions: Convert this statement into natural conversational language. If the multiple locations are provided for the same object, and some of them are null, ignore the null objects in the final description. In the above example, ignore apple null and banana null since the locations of the other apples and banana are known.

Step 2 in Figure 7 shows the visual generation procedure. After generating raw action & utterance sequence, we use the Virtualhome simulator to render RGB, depth, and instance segmentation frames as well as supporting data like camera data or scene graphs. Then, raw frames are combined together into a video. For interactive scenarios with language, agent names', avatars and communication are overlaid as captions on the video frames.

Step 3 in Figure 7 shows the text generation procedure for input text and questions. With two agents' actions and utterance sequence, we prompt GPT-4o to generate a description of the two agents' actions in a story-like way while maintaining chronological sequence. Portions of these descriptions are then used as textual input for the questions: for language scenarios, the conversation serves as the text input; for no-language scenarios, one agent's action is provided as the text input while the other agent's action is shown in the video. We then prompt GPT-4o to generate questions based on our pre-designed templates and the full descriptions. Essentially, GPT-4o fills in the blanks of the question templates using the information from the complete description.



**Prompt for generating episode descriptions** Objective: Create a description of a two-agent interaction scenario based on the provided language template.

User Input: A list of actions by each agent, Verbal communication between the agents.

Structure: Actions: A list of actions taken by agent 0 and agent 1, Language: Verbal communication between the agents in a list format.

Instructions:

1. Synchronization guidelines: Synchronize actions and language, the first entry in the "language" list corresponds to the first action step, the second entry in the "language" list corresponds to the second action and so on. If a language entry is null, there is no communication at that timestep. Synchronize descriptions of actions and language strictly by timesteps.

2. Agent names: Choose from a predefined list of common names.

3. Description guidelines: Describe the actions and language of both agents together, step by step. Avoid adjectives and excessive descriptions. Do not skip any action or language steps.

4. After establishing the timeline, make the description shorter, more concise and flow a lot like a story. Do not skip any actions

5. Place more emphasis on the events immediately following the language conversation (if any)

6. When describing a action involved with grabbing objects, make sure to also include the original place of the object

[Example input and generated scenario]

### **Prompt for generating question templates**

Objective: Generate questions from the language template by filling in the blank.

User input: A description of the episode.

Expected output: Questions following this templated format, filling in the blanks, denoted by [] where necessary.

Template:

Question: Given the above interaction, assuming that [second agent's name] knows what is [inside

on a location that second agent mentioned in the communication], which of the following statement is MOST likely?

A) When giving information, [second agent's name] has been trying to help [first agent's name] locate [the object that second agent mentioned]

B) When giving information, [second agent's name] has been trying to prevent [first agent's name] from finding [the object that second agent mentioned]

C) When giving information, [second agent's name] was indifferent towards [first agent's name]'s goals

Answer: B) When giving information, [second agent's name] has been trying to prevent [first agent's name] from finding [the object that second agent mentioned]

## **11 LIMP Implementation Details**

### **11.1 Visual Perception**

A previous multi-modal ToM model, BIP-ALMJin et al. (2024), relies on instance segmentation to extract ground truth object information and spacial relationships from visual inputs. However, this level of detailed information is not available to other baselines or human participants, and is rarely available in real-world datasets. In RGB videos, detecting the exact object is challenging for both models and humans, as the objects might be small or obscured from view. To maintain generality, we use Gemini 1.5 ProReid et al. (2024) as our visual perception model. We upload the RGB videos to the Gemini web version in google AI studio without any extra information about possible locations and objects, and asked it to extract the action and utterance sequences of the agents. The prompt we use is shown below.

**Prompt for Gemini 1.5 Pro visual extraction**

Task: You will watch a video depicting two agents performing some actions. Your goal is to infer and describe the actions in chronological order. For [agent name], provide details about his/her actions, including what objects she handled, where she obtained them from, and where she placed them. Formulate all actions into a single line. Do not include any newline characters. Note that an agent moving their arm probably indicates opening a container or picking up an item. If you cannot decipher the location that [agent name] grabs from, make your best guess based on all the context in the video. If you cannot effectively identify the object, just leave it as grab some object without trying to guess the exact one.

## 11.2 Text Parsing & Multi-modal Fusion

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 we use 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.

The prompts we use for text parsing and multi-modal fusion are shown below.

**Prompt for text parsing**

You will read a piece of text describing actions of some number of people with distinctive names. You will also have a name, which is the name of the person whom you should pay attention to. Summarize the person's actions and utterance separately in a chronological order. Only include the actions and utterance directly taken by the person in the text, and exclude any previous actions mentioned indirectly. If you cannot find either utterance or actions of the person in the text, leave the corresponding section blank. When reading words like "it", replace it with inferred object or location to make actions clearer. Do not include agent's communication as part of it. Organize your answer in this form: Actions: ["action one", "action two", "action three", ...] ... Utterance: ["utterance one", "utterance two", "utterance three", ...] ...

Text: text input

Name: name of the agent

**Prompt for error recovery**

You will read some text describing a person's action. The name of the person is given. Summarize and reorganize the person's actions. Possible actions include walk towards somewhere, grab something from somewhere, open some container, close some container, put something somewhere. Only summarize these actions and their synonyms in this form

and abandon mismatch actions. Omit person's name. When mentioning location name, try to infer room the location is inside and include it in the action. Check objects mentioned in the Additional Information section. Replace any object mentioned in action with the object appeared in that section. Formulate your final answer in the following form. Actions: ["action1", "action2", ...]

Input text: raw output of Gemini  
Additional information: context  
Person's name: agent name

#### **Prompt for initial state retrieval**

You will read one or two person's actions in a list like form. From the actions taken, extract the initial state of the environment before any people act. Check each grab action or synonyms. Describe it in the form "There is a [object grabbed] [on/inside location of grabbing]. Only include environment states statements. Do not include any other information or extra contents. Actions: all agents' actions

### **11.3 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. The prompt is shown below.

#### **Prompt for latent variable extraction**

You will read a question about agents' mind and ideas, and the initial state of the environment from which agents' are interacting in. Agents' knowledge & belief are about this initial state, but not necessarily changed state after some actions. For each choice, extract one set of second person's belief (make sure to turn it into some statement about the environment state), second person's social goal toward first person's actions (help, hinder or some similar words of independent), and second person's believed first person's physical goal (some arrangement of objects). Organize the answer in this way: A: Belief: contents; Social goal: contents; Believed Goal: contents. B: Belief: contents; Social goal: contents; Believed Goal: contents. C: Belief: contents; Social goal: contents; Believed Goal: contents. Do not include any other information or extra contents. Make sure your answer follow the format requirement, use ";" to separate variables within each choice and end response with ".". Separate contents of "A", "B" and "C" with "."

Initial state:

Question:

Actions:

### **11.4 Inverse Multi-agent Planning with GPT-4o**

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'. The prompt is shown below:

**Prompt for GPT-4o Inverse Planning**

Decide if agent's action is likely with the information provided, and respond with only either A or B:

agent's social goal: social goal

agent's belief: belief

agent's belief of other agent's goal: belief of goal

other agent's utterance: utterance

Initial state: initial state

Previous Actions: actions taken previously

Respond with only either A or B:

Agent's Action or Utterance:

A) Likely

B) Unlikely

## 12 Baseline Implementation Details

### 12.1 LMMs

For the large multi-modal models we tested on our benchmark, we provided the same videos for humans, LIMP, and state-of-the-art LMMs capable of handling video input. For Intern-VL, we use a provided script to sample 8-10 frames from our video, as it cannot directly handle video input. For other LMMs that only accept images as input, we uniformly sampled one frame every 20 frames (approximately every 2 seconds), resulting in 15-30 frames per video. For videos with language communication, additional context was provided regarding the timing of the communication to ensure the utterances align with the frame sequence. Specifically, for Llava 1.6 and GPT-4o, we provided frames as input. For Gemini 1.5 and VideoLlama 2, we provided the full videos. For InternVL 2, we used the officially provided script to sample 8-10 frames, adapting the video input to the model's frame-based capabilities.

### 12.2 BIP-ALM

To adapt BIP-ALM to work on our benchmark, we needed to provide additional information. Since BIP-ALM relies on instance segmentation and depth images for visual perception, we followed its method of projecting pixels to 3D and detecting object relationships based on bounding boxes & instance segmentations. We also enhanced the original object detection mechanisms (which only detected closeness, inside and open relationships) by adding the detection of 'on' and 'holding' relationships. Additionally, because BIP-ALM uses symbolic representations, its text parsing module requires a strict data format. Therefore, we provided descriptions of the full interactive scenarios to ensure BIP-ALM functions correctly. Despite receiving more information than LIMP, BIP-ALM still performed no better than random guessing on MuMA-ToM, suggesting that single-agent inverse planning models cannot be directly applied to multi-agent interactions.

## 13 Full Version of the Example Questions

### 13.1 Belief Inference

Video: <https://youtu.be/kXCUL7KB2Dc>

**Text:** John: Do you know where the beer is?

Mary: I discovered a beer on the coffee table in the living room.

**Question:** Given the above interaction, if Mary has been trying to hinder John from achieving his goal, which of the following statements is LEAST likely?

- A) When giving information, Mary believed that there was beer on the coffee table in the living room
- B) When giving information, Mary believed that there was beer inside the kitchen cabinet
- C) When giving information, Mary believed that there was beer inside the fridge

**Correct Answer:** A

### 13.2 Social Goal Inference

Video: <https://youtu.be/e4360D7yYV8>

**Text:** Kevin: Any idea where the magazine might be?  
Jessica: I discovered a magazine inside the cabinet in the bedroom.

**Question:** Given the above interaction, assuming that Jessica knows what is inside the cabinet in the bedroom, which of the following statements is MOST likely?

- A) When giving information, Jessica has been trying to help Kevin locate the magazine
- B) When giving information, Jessica has been trying to prevent Kevin from finding the magazine
- C) When giving information, Jessica was indifferent towards Kevin's goals

**Correct Answer:** B

### 13.3 Belief of Goal Inference

Video: <https://youtu.be/bpw1jpbViFE>

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

**Question:** "Given the above interaction, based on the actions of the agents, 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.

**Correct Answer:** C