MMToM-QA: Multimodal Theory of Mind Question Answering

Chuanyang Jin[1], Yutong Wu[2], Jing Cao[3], Jiannan Xiang[4],
Yen-Ling Kuo[5], Zhiting Hu[4], Tomer Ullman[2], Antonio Torralba[3], Joshua Tenenbaum[3], Tianmin Shu[6]
[1]NYU, [2]Harvard, [3]MIT, [4]UCSD, [5]UVA, [6]JHU

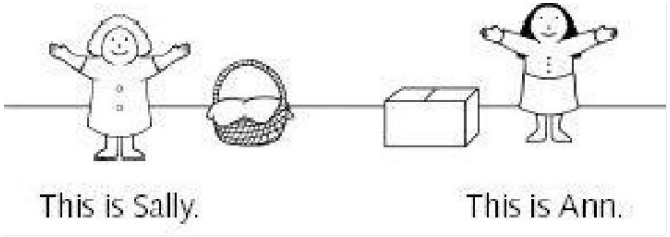
ACL 2024

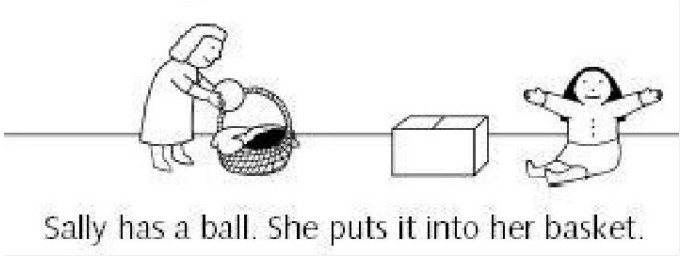
Outstanding Paper Award

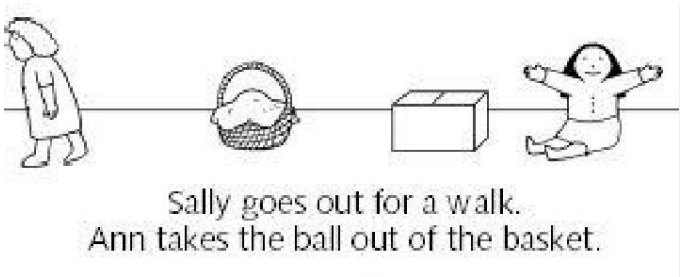
Theory of Mind (the ability to understand people's mind)

— Towards AI agents with social intelligence

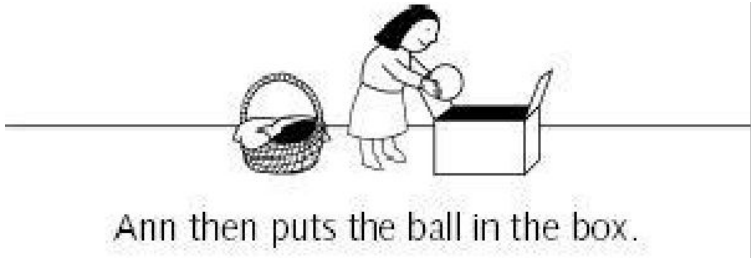
An example human ToM experiment: Sally-Anne test







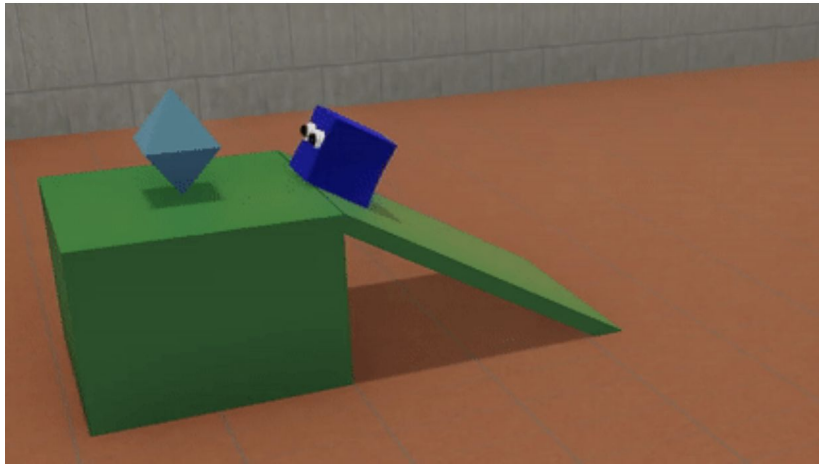
Understand this story from images and/or text



Prior Theory of Mind benchmarks are unimodal

Video benchmarks

AGENT (Shu et al., 2021)



Text benchmarks

ToMi (Le et al., 2019)

	Type	Story	Question		Answers
(a)	FACT	Sophia entered the study. Noah entered the study. The dress is in the treasure chest. Noah exited the study. Hannah entered the garden. Sophia moved the dress to the box.	Where is the dress really?	✓🤖	box treasure chest
(b)	M-1-FB	Noah entered the garden. Nathan entered the garden. Evelyn likes the pumpkin. The banana is in the basket. Nathan exited the garden. Noah moved the banana to the suitcase.	Where will Nathan look for the banana?	✓🤖	basket suitcase
(c)	M-2-TB	Lily entered the patio. Aiden is in the patio. Mila entered the patio. Mila hates the radish. The coat is in the box. Aiden moved the coat to the crate. Mila exited the patio.	Where does Aiden think that Mila searches for the coat?	✓🤖	crate box
(d)	M-1-TB	Elizabeth entered the cellar. Carter entered the cellar. The slippers is in the crate. Elizabeth moved the slippers to the container. Carter exited the cellar.	Where will Carter look for the slippers?	✓ 🤖	container crate
(e)	M-1-FB	Evelyn entered the living room. Jackson entered the playroom. James entered the playroom. The beans are in the treasure chest. James exited the playroom. Jackson moved the beans to the pantry. Jackson exited the playroom. James entered the living room.	Where will James look for the beans?	✓ 🤖	treasure chest pantry
(f)	M-2-FB	Isla likes the potato. Ella entered the laundry. Oliver entered the laundry. The slippers are in the box. Ella exited the laundry. Oliver moved the slippers to the basket. Isla entered the office.	Where does Ella think that Oliver searches for the slippers?	🤖 ✓	basket box

Multimodal Theory of Mind Question Answering (MMToM-QA)

Video of a person's household activities

Text description of the environment and the actions

Question about the person's mind

VIDEO INPUT

TEXT INPUT

What's inside the apartment: … The kitchen is equipped with a microwave, eight cabinets, … Inside the microwave, there is a cupcake. There is a wine glass and an apple on one of the kitchen tables. There are water glasses, a bottle wine, a condiment bottle, and a bag of chips in inside the cabinets. …
Actions taken by Emily: Emily is initially in the bathroom. She then walks to the kitchen, goes to the sixth cabinet, opens it, subsequently closes it, and then goes towards the fourth cabinet.

QUESTION

Which one of the following statements is more likely to be true?

(a) Emily has been trying to get a cupcake. ✔ (b) Emily has been trying to get a wine glass. ✖

Diverse Scenarios for Belief and Goal Inference

Belief Inference

Type 1.1: True belief, short-term

Scene: … Inside the bridge, you'll find a bottle of wine…
Actions: … Finally, she moves towards the fridge, preparing to open it.

Question: If Elizabeth has been trying to get a bottle of wine, which one of the following statements is more likely to be true?
(a) Elizabeth thinks that there is a bottle of wine inside the fridge.
(b) Elizabeth thinks that there isn't any bottle of wine inside the fridge.

Type 1.2: False belief, short-term

Scene: … The living room features a cabinet… The cabinet is filled with a bag of chips, a remote controller, a bottle of wine, and a water glass.
Actions: Jennifer is situated in the living room. She heads towards the cabinet and is about to open it.

Question: If Jennifer has been trying to get a cupcake, which one of the following statements is more likely to be true?
(a) Jennifer thinks that there isn't a cupcake inside the cabinet.
(b) Jennifer thinks that there is a cupcake inside the cabinet.

Type 1.3: Belief tracking, long-term

Scene: … The kitchen is equipped with a fridge, sofa, dishwasher, eight cabinets, a stove, a microwave, and a kitchen table…
Actions: … He walks to the seventh kitchen cabinet, opens and closes it. He repeats the same action with the sixth kitchen cabinet. Subsequently, he moves towards the dishwasher.

Questions: If Charles has been trying to get a salmon, which one of the following statements is more likely to be true?
(a) Charles thinks that there is a salmon inside the fridge.
(b) Charles thinks that there isn't any salmon inside the fridge.

Goal Inference

Type 2.1: Goal given true belief

Scene: … The living room is furnished with a cabinet, … The cabinet is filled with two apples, …, and a bottle of wine. … Inside the fridge, there are two apples.
Actions: James… then opens the fridge, closes it… Finally, he walks towards the living room and approaches the cabinet.

Question: Which one of the following statements is more likely to be true?
(a) James has been trying to get a bottle of wine.
(b) James has been trying to get an apple.

Type 2.2: Goal given false belief

Scene: … There is a water glass inside the seventh cabinet… The fridge stores two cupcakes…
Actions: Mark… advances towards the seventh kitchen cabinet.

Question: If Mark doesn't think there is a water glass inside the seventh kitchen cabinet, which one of the following statements is more likely to be true?
(a) Mark has been trying to get a water glass.
(b) Mark has been trying to get a cupcake.

Type 2.3: Goal given updated belief

Scene: … The first cabinet, from left to right, contains a bag of chips.
Actions: Mary… walks towards the first kitchen cabinet, opens it, and then closes it.

Question: Which one of the following statements is more likely to be true?
(a) Mary has been trying to get a bag of chips.
(b) Mark has been trying to get a condiment bottle.

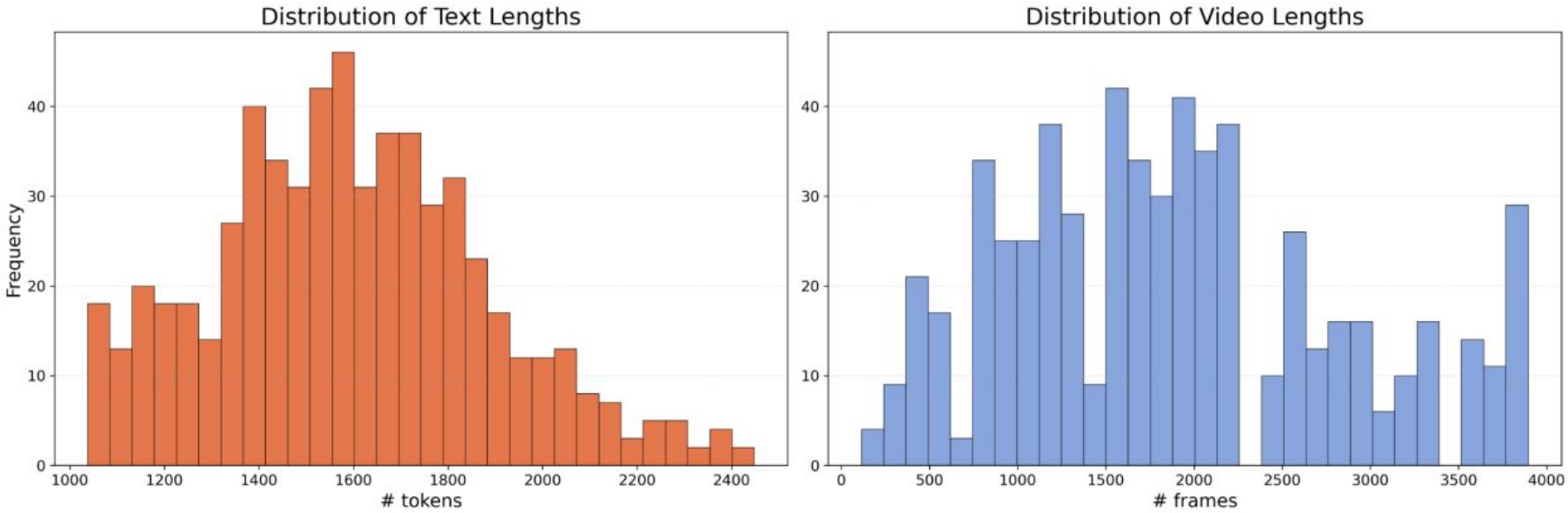
Type 2.4: Goal given future actions

Scene: … The dishwasher holds a dish bowl… The first cabinet from the left holds a bag of chips and a wine glass… The fifth cabinet has an apple…
Actions: Williams… advances towards the first kitchen cabinet, opens it, and then shuts it. He then moves towards the fifth kitchen cabinet.

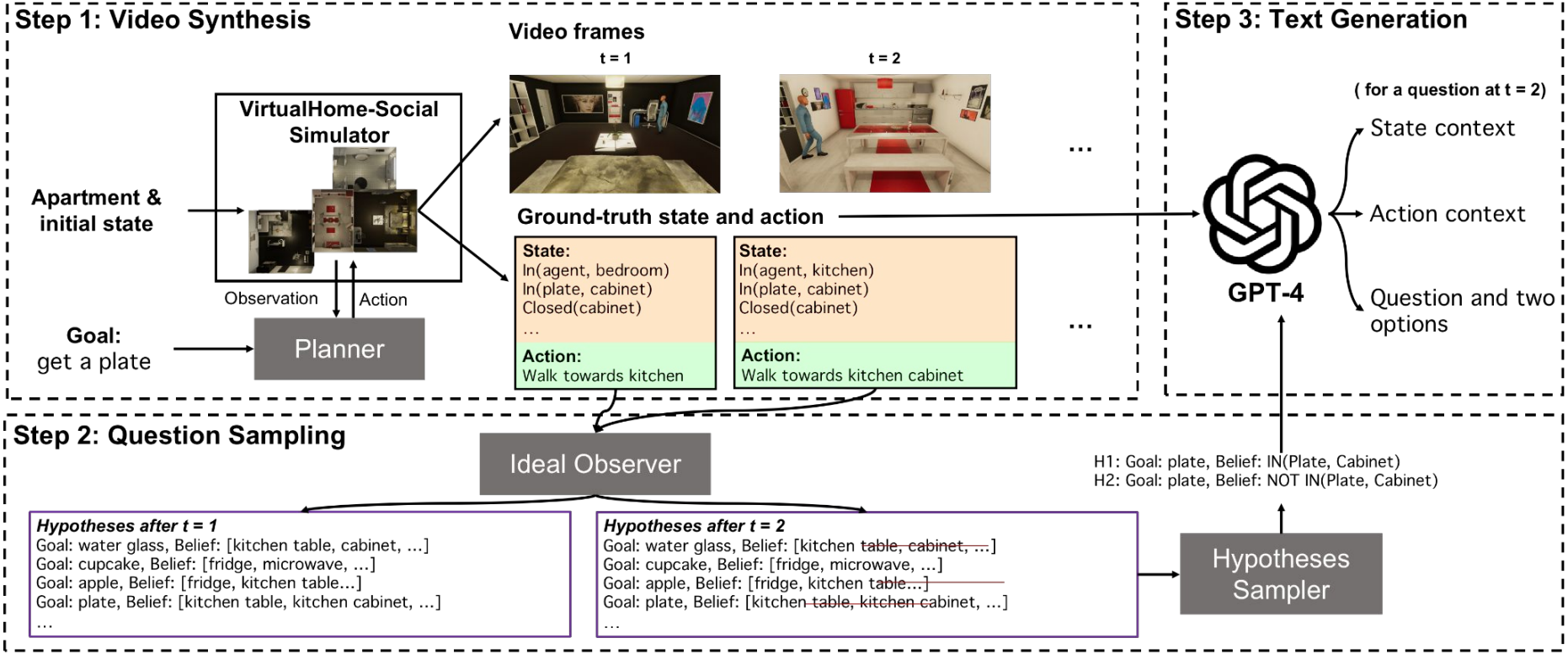
Question: Which one of the following statements is more likely to be true?
(a) William has been trying to get a wine glass.
(b) William has been trying to get a dish bowl.

Longer text and video context

Prior benchmarks: < 100 tokens, < 500 frames



Procedural generation of QAs



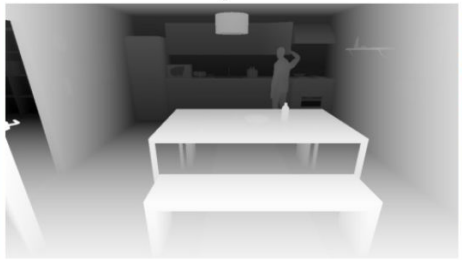
Training Data

1000 synthesized videos of human household activities (no training QAs)

GT goals, beliefs
+

RGB

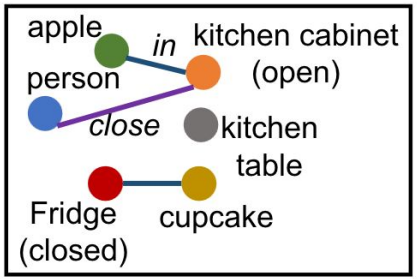
Depth



Instance Segmentation

3D Human Pose

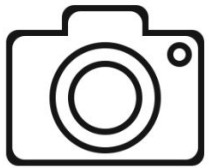
Scene Graph



Action

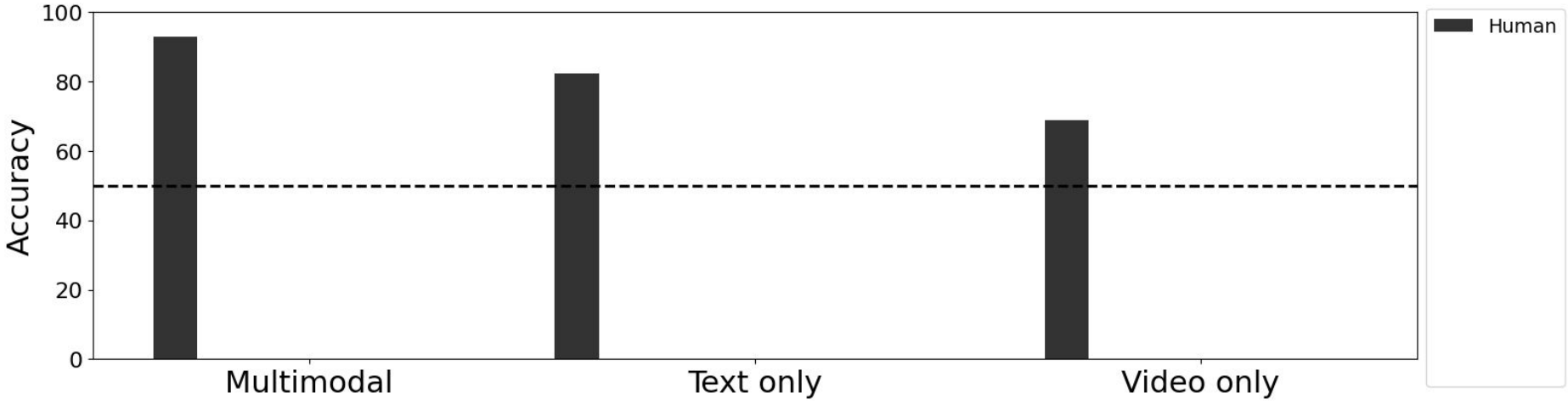
`open kitchen cabinet`

Camera Data



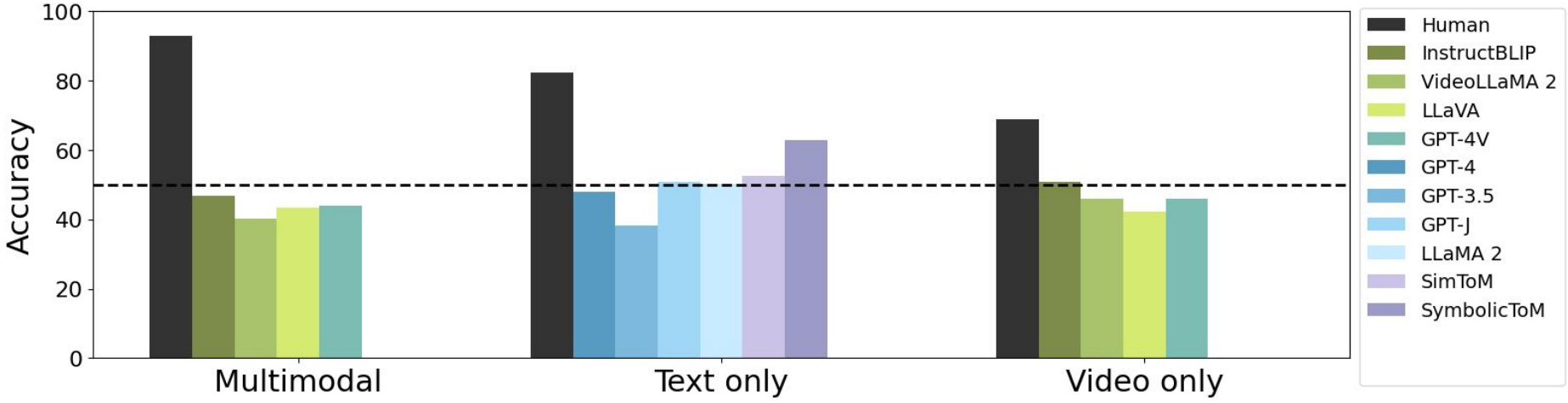
Results on MMToM-QA

- Humans perform the best in the multimodal condition



Results on MMToM-QA

- LLMs and LMMs perform poorly compared to humans 😕



A failure example of GPT-4V

Scene: The microwave holds two cupcakes … The cabinet is filled with a bag of chips …
Actions: Jennifer heads towards the cabinet and is about to open it.

Question: If Jennifer has been trying to get a cupcake, which one of the following statements is more likely to be true?
(a) Jennifer thinks that there isn't a cupcake inside the cabinet.
(b) Jennifer thinks that there is a cupcake inside the cabinet.

GPT-4V

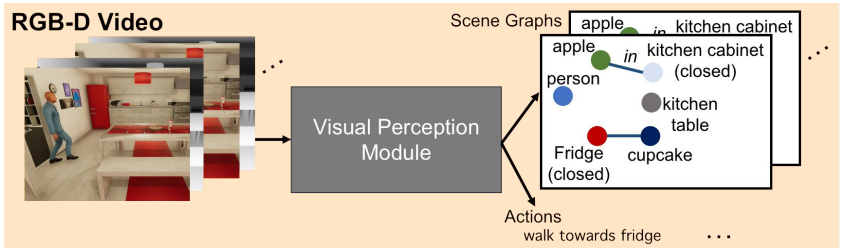
(a) … Since Jennifer is heading towards the cabinet which is said to contain a bag of chips, but no mention of cupcakes, it suggests that Jennifer does not think there is a cupcake inside that cabinet.

GPT-4V cannot distinguish belief from the true world state

How can we bridge the gap between model and human performance?

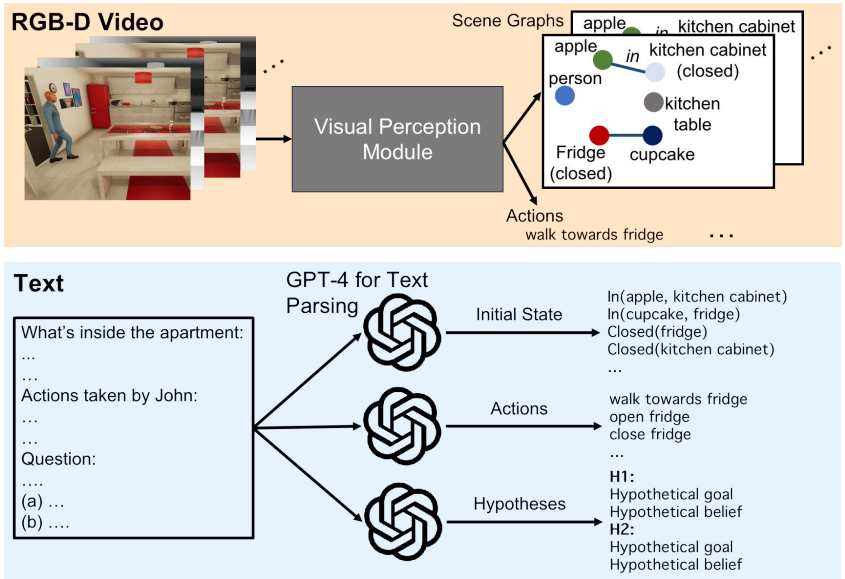
BIP-ALM: Bayesian **I**nverse **P**lanning **A**ccelerated by **L**anguage **M**odels

- Extracts symbolic representations from video



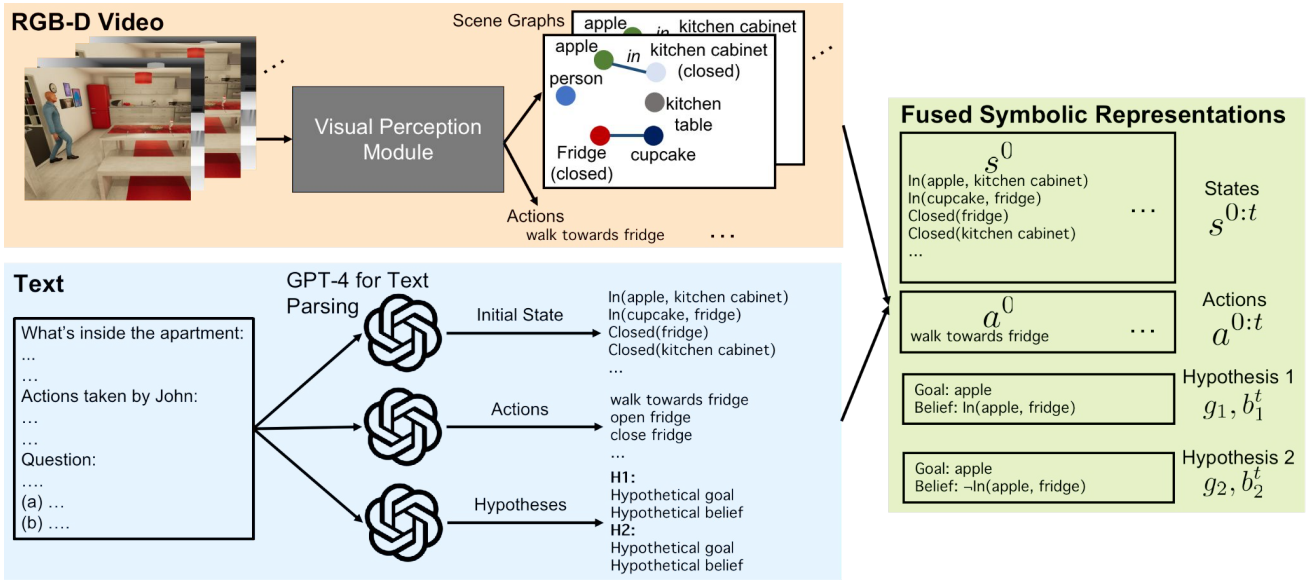
BIP-ALM: Bayesian Inverse Planning Accelerated by Language Models

- Extracts symbolic representations from video and text



BIP-ALM: Bayesian Inverse Planning Accelerated by Language Models

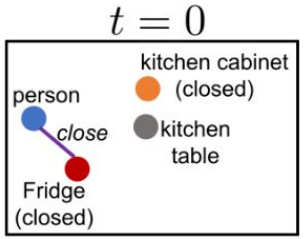
- Extracts symbolic representations from video and text
- Aligns and fuses the representations

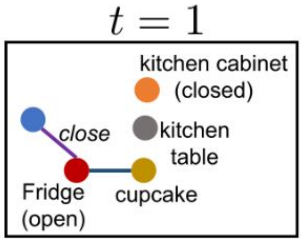


Fusion

Dynamic states

From Video






```
Close(agent, fridge)
Closed(fridge)
Closed(kitchen cabinet)
…
```

```
Close(agent, fridge)
IN(cupcake, fridge)
Open(fridge)
Closed(kitchen cabinet)
…
```

From Text

Initial state

```
IN(apple, kitchen cabinet)
IN(cupcake, fridge)
Closed(fridge)
Closed(kitchen cabinet)
…
```

Fused Symbolic Representations

s^0

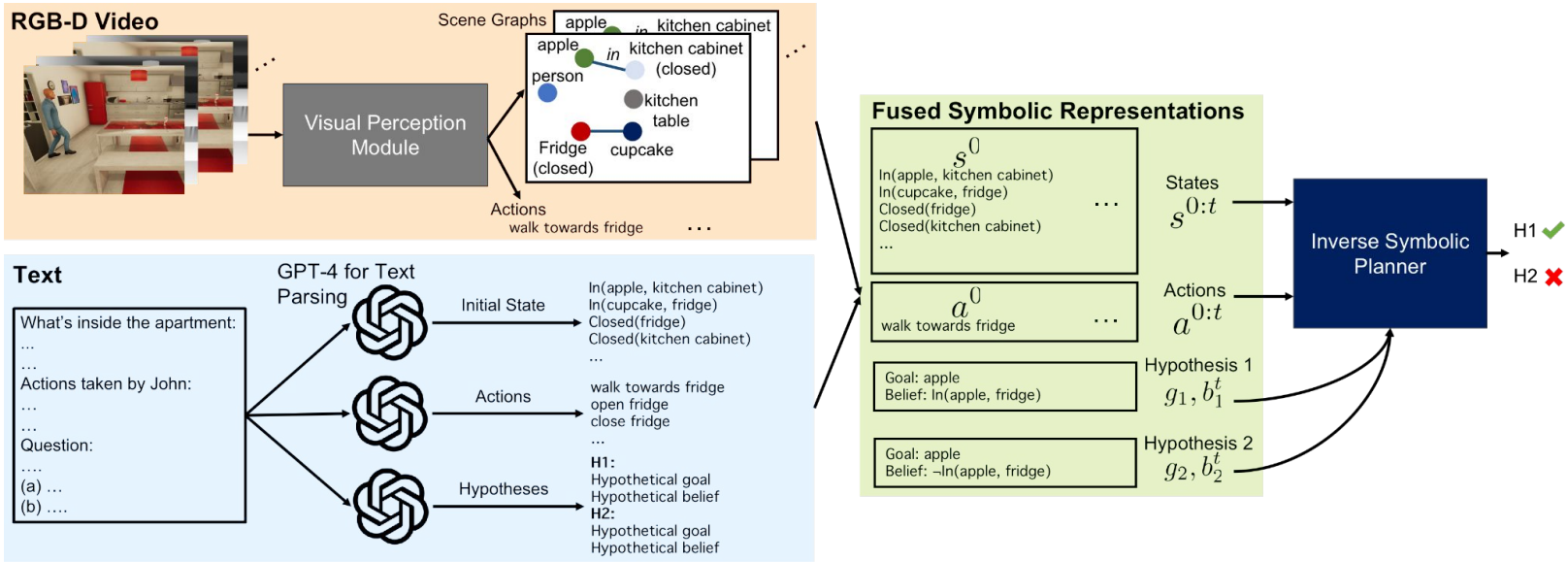
```
Close(Agent, Fridge)
IN(apple, kitchen cabinet)
IN(cupcake, fridge)
Closed(fridge)
Closed(kitchen cabinet)
…
```

s^1

```
Close(Agent, Fridge)
IN(apple, kitchen cabinet)
IN(cupcake, fridge)
Open(fridge)
Closed(kitchen cabinet)
…
```

BIP-ALM: Bayesian Inverse Planning Accelerated by Language Models

- Extracts symbolic representations from video and text
- Aligns and fuses the representations
- Conducts inverse planning using language models



BIP-ALM: Bayesian Inverse Planning Accelerated by Language Models

Bayesian Inverse Planning (Baker et al., 2009, 2017)

$$P(\text{Goal}, \text{Belief}|\text{Action}, \text{State}) \propto P(\text{Action}|\text{Goal}, \text{Belief})P(\text{Belief}|\text{State})P(\text{Belief})P(\text{Goal})$$

Computational bottleneck

Planning in large state & action spaces under partial observability

Language models for estimating the action likelihood

```
goal: {hypothetical goal}
state: {state}
belief (possible locations the person believes the {hypothetical goal} could be):
{hypothetical or predicted belief}
action:
```

Language models for estimating the action likelihood

```
goal:  {hypothetical goal}
state:  {state}
belief (possible locations the person believes the {hypothetical goal} could be):
{hypothetical or predicted belief}
action:
```

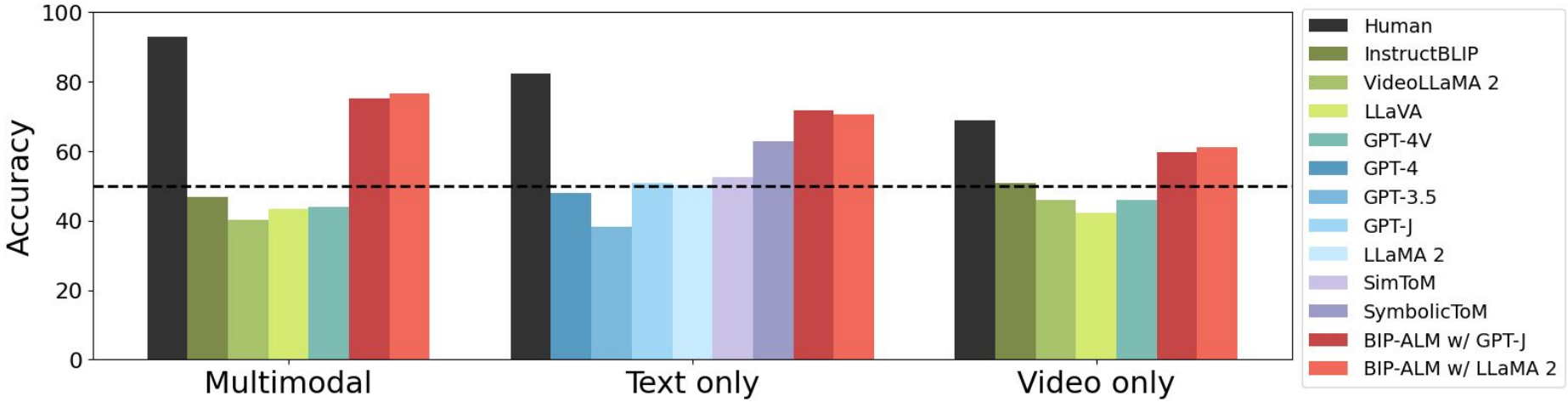
Finetune LMs on synthetic data (with LoRA):

Using the training videos

20,000 samples of “goal, state, belief -> action” transitions

Results on MMToM-QA

- LLMs and LMMs perform poorly compared to humans 😕
- Our method shows promising results 😊



Main Takeaways

Human ToM benefits from multimodal inputs

LLMs and LMMs perform poorly

BIP-ALM shows promising results, benefiting from:

(1) the modality-invariance of symbolic representations

(2) the robustness and interpretability of inverse planning

(3) the scalability and flexibility of language models

Future directions

Multimodal Theory of Mind: Question Answering $\rightarrow$ Planning

Multimodal assistive agents (e.g., perception, verbal communication)

Robots, web agents, etc., performing tasks interactively with humans

MMToM-QA

Multimodal Theory of Mind
Question Answering

https://chuanyangjin.com/mmtom-qa

Code, data &
leaderboard