Building Cooperative Embodied Agents Modularly with Large Language Models

Minglai Yang

(Ming)

In Dr. Josh Tenenbaum’s talk, the gap between current function approximation/pattern recognition to the world intelligence (WI)

- Explain and understand what we see
- Imagine things we could see but haven’t yet
- Plan actions and solve problems to make these things real => Planning
- Build new models when we learn more about the world

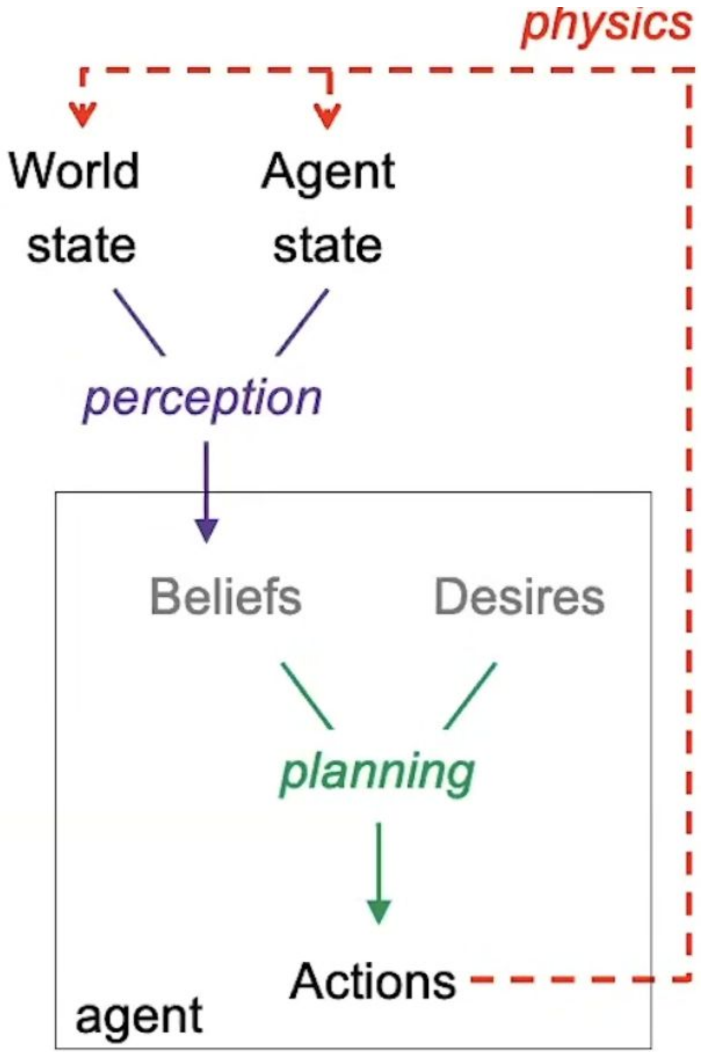
Theory of Mind

How you think your brain is working…

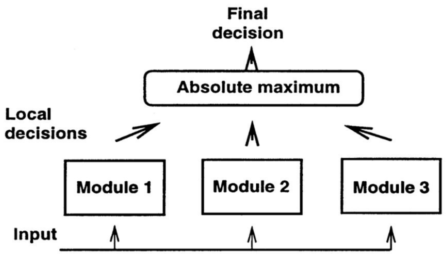
ToMCAT. A collaboration between the Information School (INFO), Computer Science (CS), and Family Studies and Human Development (FSHD) has been awarded a large grant to develop a **theory of mind-based** cognitive architecture for teams (ToMCAT). The grant ($7.5M, for 48 months) is part of the DARPA Artificial Social Intelligence for Successful Teams (ASIST) program.

The goal of the project is to build artificially intelligent agents that understand both the social and goal-oriented aspects of teams in mission-like scenarios (e.g., search-and-rescue missions), and are able to reason about possible interventions. **The agent, ToMCAT, needs to model human players' affect and beliefs about the situation and about each other's affect and beliefs (theory of mind).** We will ground this work in extensive measurements of humans interacting in small teams, that will include audio, video, eye tracking, electrocardiography (EKG), electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), and self report. The participants will execute missions within a Minecraft environment with one, two, three, or four human players interacting with the ToMCAT agent.

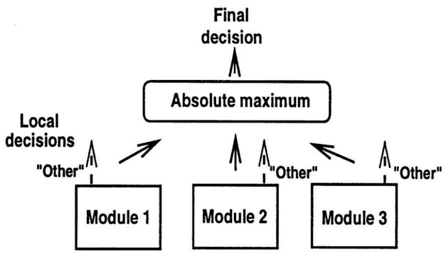
(Text by Prof. Kobus Barnard, image by Prof. Joshua Tenenbaum)



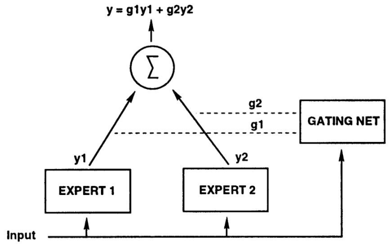
Modular Neural Network (MNN) [Audo, 1999]



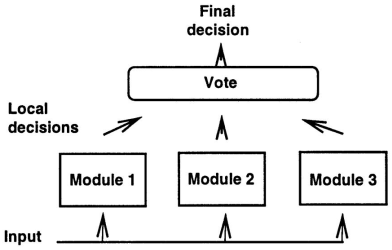
Decoupled modules



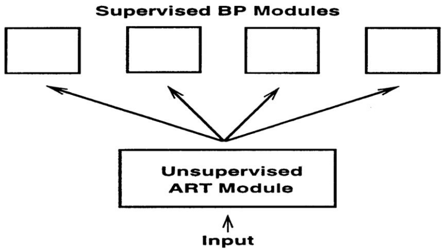
Other-o/p model



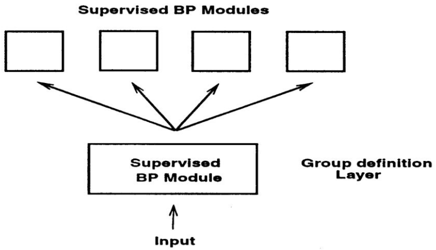
Multiple experts



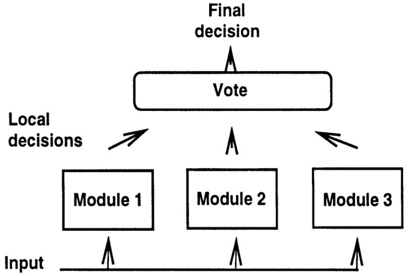
Ensemble (majority vote)



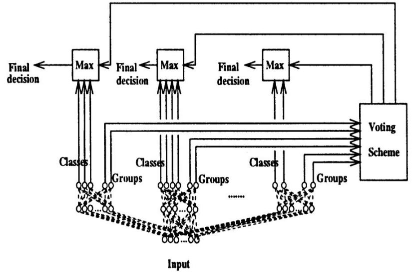
ART-BP model



Hierarchical modules

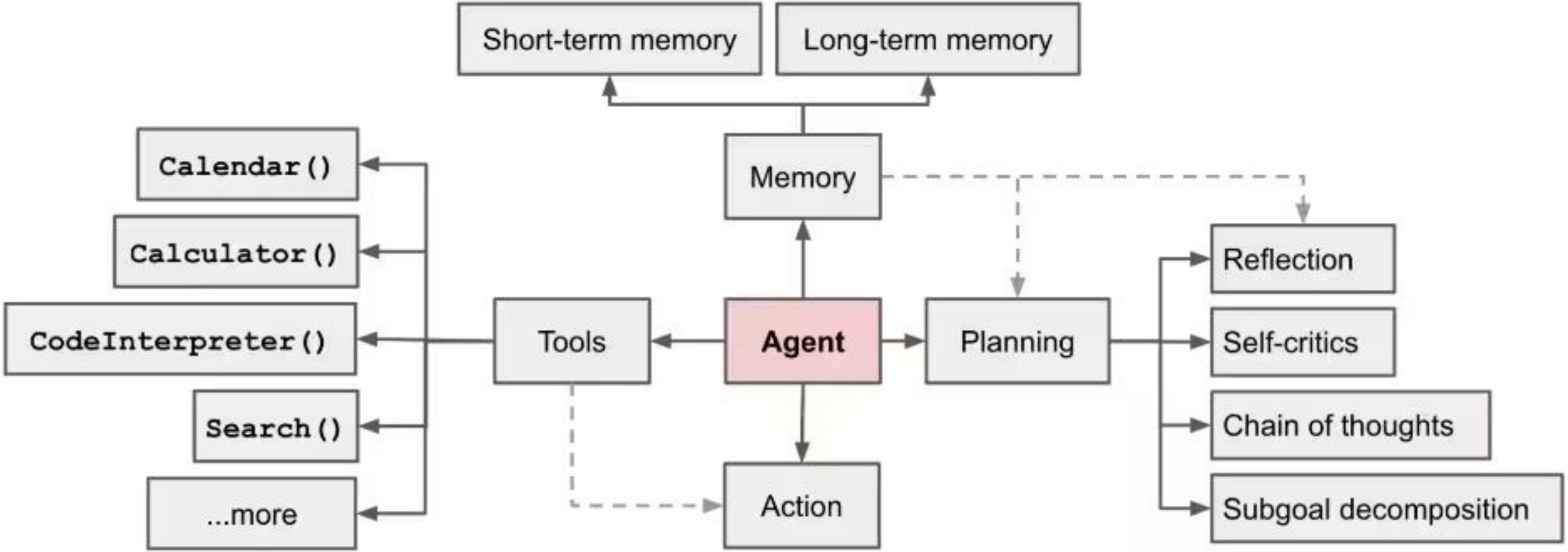


Ensemble (average vote)



Cooperative MNN

Multi Module decision-making


Short-term memory
Long-term memory
Memory
Calendar()
Calculator()
CodeInterpreter()
Search()
...more
Tools
Agent
Planning
Reflection
Self-critics
Chain of thoughts
Subgoal decomposition
Action

Building Cooperative Embodied Agents Modularly with Large Language Models

Hongxin Zhang[1*] , Weihua Du[2*] , Jiaming Shan[3] , Qinhong Zhou[1] , Yilun Du[4] , Joshua B. Tenenbaum[4] , Tianmin Shu[4] , Chuang Gan[1,5]

[1] University of Massachusetts Amherst [2] Tsinghua University [3] Shanghai Jiao Tong University [4] MIT [5] MIT-IBM Watson AI Lab

ICLR 2024

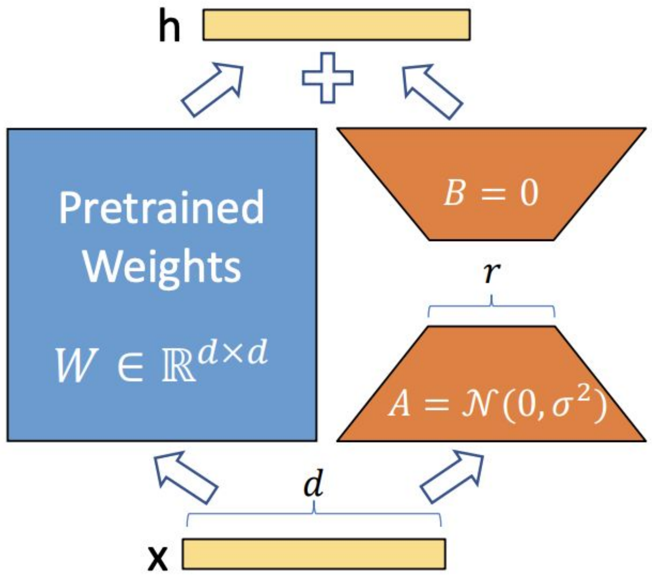
Paper   <> Code

They use both prompting and fine-tuning techniques in different stages of their framework.

Prompting with GPT-4: use the strong reasoning, language comprehension, and generation capabilities of GPT-4 to drive their cooperative agent (CoELA). This approach allows the agent to generate plans and communicate naturally, making it effective in cooperative tasks.

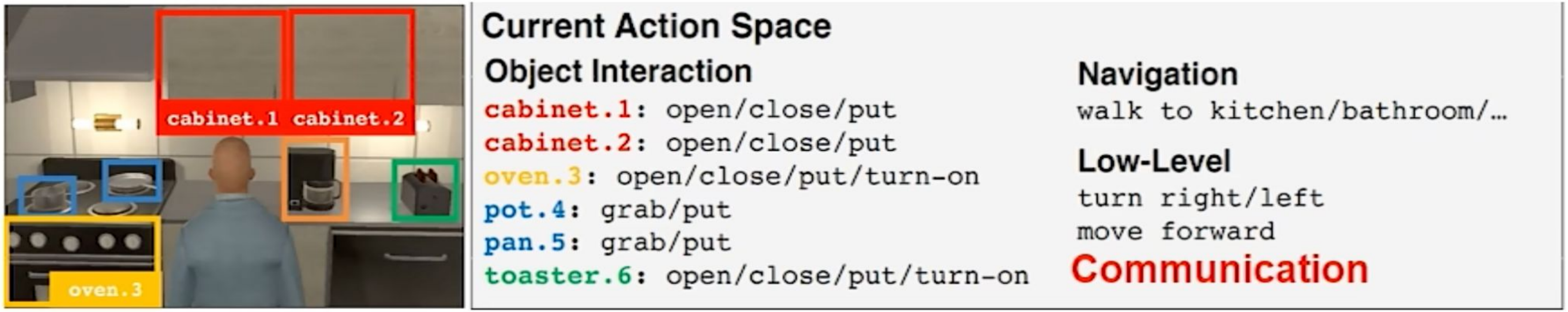
Fine-tuning with LoRA:
They also fine-tune an open-source model, LLAMA-2, using LoRA.

The fine-tuned model, referred to as CoLLAMA, was trained on a small set of task-specific data collected by their agents. This fine-tuning improved the model's ability to perform in specific multi-agent tasks (subtasks in this paper).



Cooperative Planning Under DEC-POMDP-COM

- Decentralized Partially Observable Markov Decision Process (DEC - POMDP)



Testing Environments And Defining Tasks (Goals)

3D World - Platform (3D Multi Agent Transport)

Watch And Help - A CHALLENGE FOR SOCIAL PERCEPTION AND HUMAN-AI COLLABORATION

3D World - Platform (3D Multi Agent Transport)

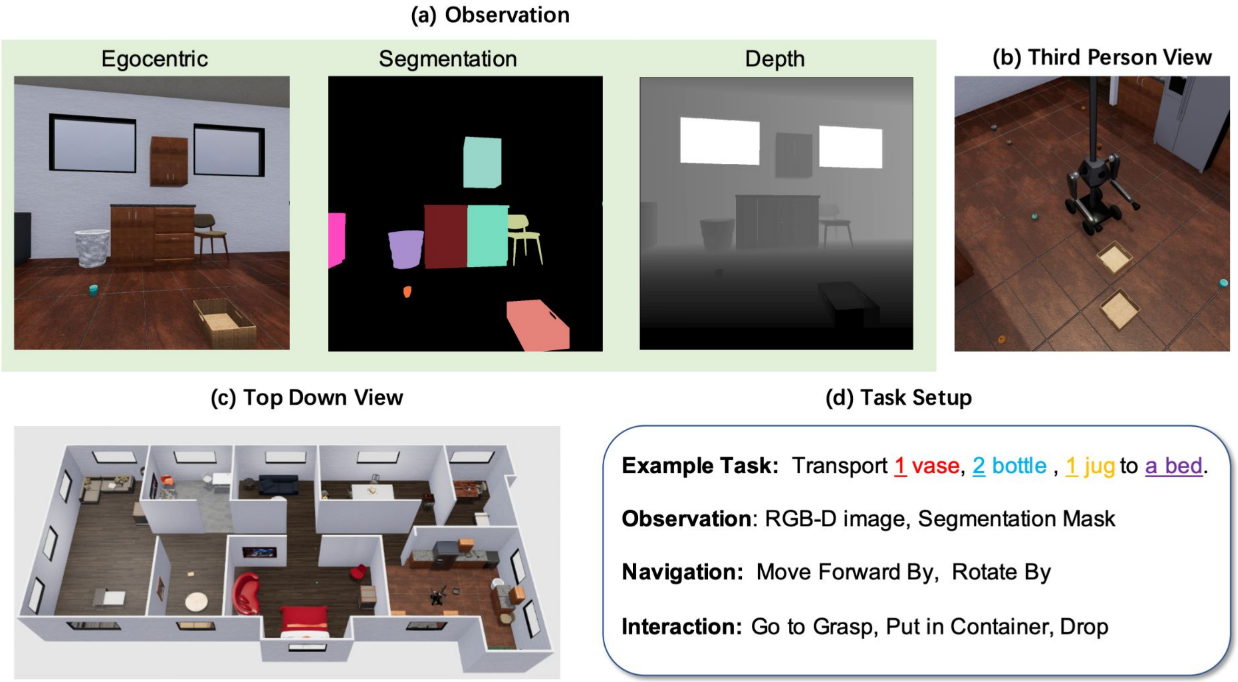


Figure 2: The details of ThreeDWorld Transport Challenge. (a) The observation state includes first-person view RGB image, Depth image, and semantic segmentation mask; (b) and (c) are Third-person view, and top-down view of the environment respectively; (d) Outline of the task and action space.

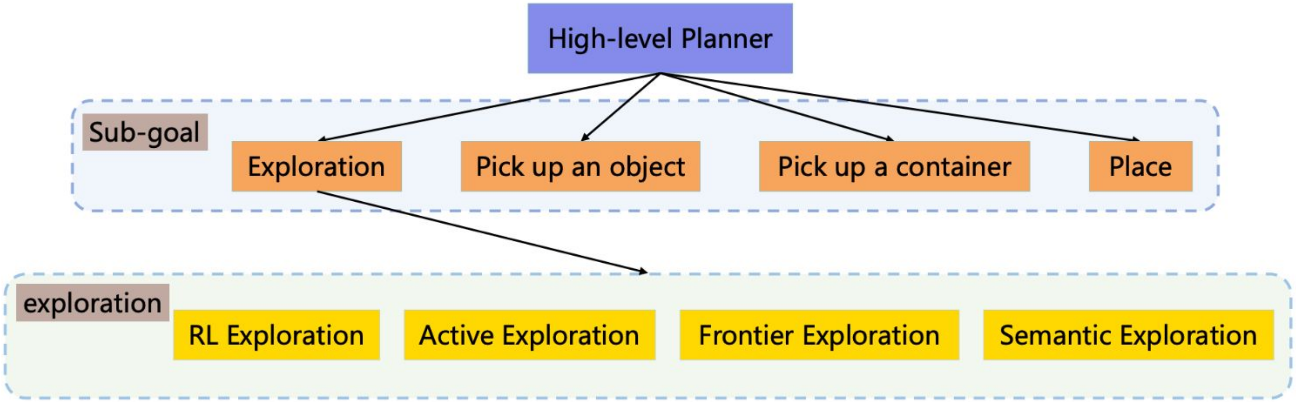


Figure 4: The flowchart of high-level and low-level planners.

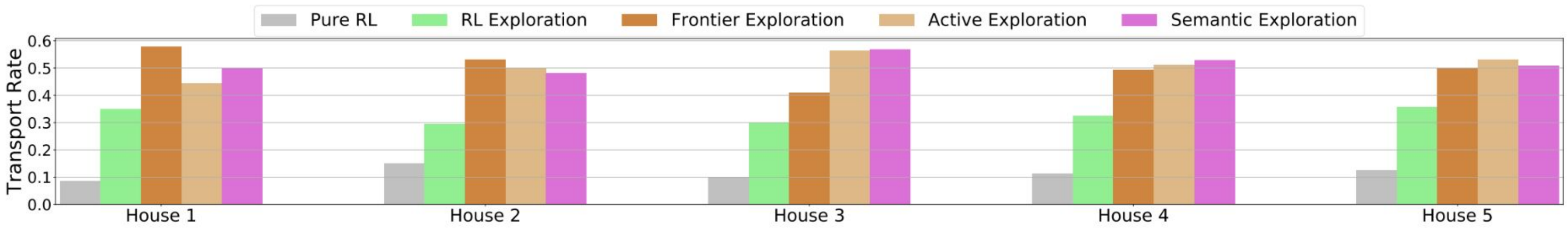


Figure 5: Comparisons of transport rates in each unseen room.

In this challenge, an embodied agent with two articulated arms and equipped with RGB-D vision is randomly placed in a virtual house.

The task for the agent is to search for specific target objects scattered around multiple rooms and transport them to a designated location, such as a bed, using realistic physics.

Containers are available within the environment that the agent can use to carry multiple objects at once, enhancing efficiency.

• Synergy between navigation and interaction. The agent cannot move to grasp an object if this object is not in the partial view, or if the direct path to it is obstructed (e.g. by a table).

• Physics-aware Interaction. Grasping might fail if the agent’s arm cannot reach an object.

• Physics-aware navigation. Collision with obstacles might cause objects to be dropped and significantly impede the transport efficiency.

• Reasoning about tool usage. While the containers help the agent transport more than two items, it also takes some time to find them. The agent thus has to reason about a case-by-case optimal plan.

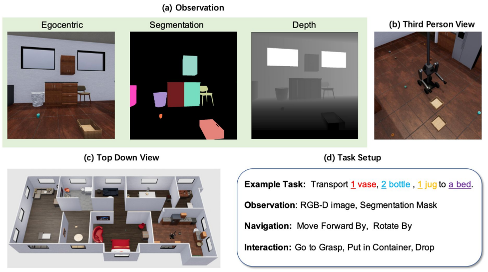


Figure 2: The details of ThreeDWorld Transport Challenge. (a) The observation state includes first-person view RGB image, Depth image, and semantic segmentation mask; (b) and (c) are Third-person view, and top-down view of the environment respectively; (d) Outline of the task and action space.

3D-LLM: Injecting the 3D World into Large Language Models

Yining Hong
University of California, Los Angeles

Haoyu Zhen
Shanghai Jiao Tong University

Peihao Chen
South China University of Technology

Shuhong Zheng
University of Illinois Urbana-Champaign

Yilun Du
Massachusetts Institute of Technology

Zhenfang Chen
MIT-IBM Watson AI Lab

Chuang Gan
UMass Amherst and MIT-IBM Watson AI Lab

Figure 1: **Examples from our generated 3D-language data, which covers multiple 3D-related tasks.**

Key features of 3D World challenge include:

- **Physics-aware Interaction**: Agents must interact with objects under realistic physical constraints (e.g., objects may be dropped if the agent collides with obstacles).
- **Navigation and Planning**: The challenge involves planning efficient routes for object retrieval and transportation, taking into account obstructions, object placement, and tool use (like containers).
- **High-Level Action**: The environment supports high-level action commands for interaction and movement, but the agent must manage complex physics-based constraints in real time.

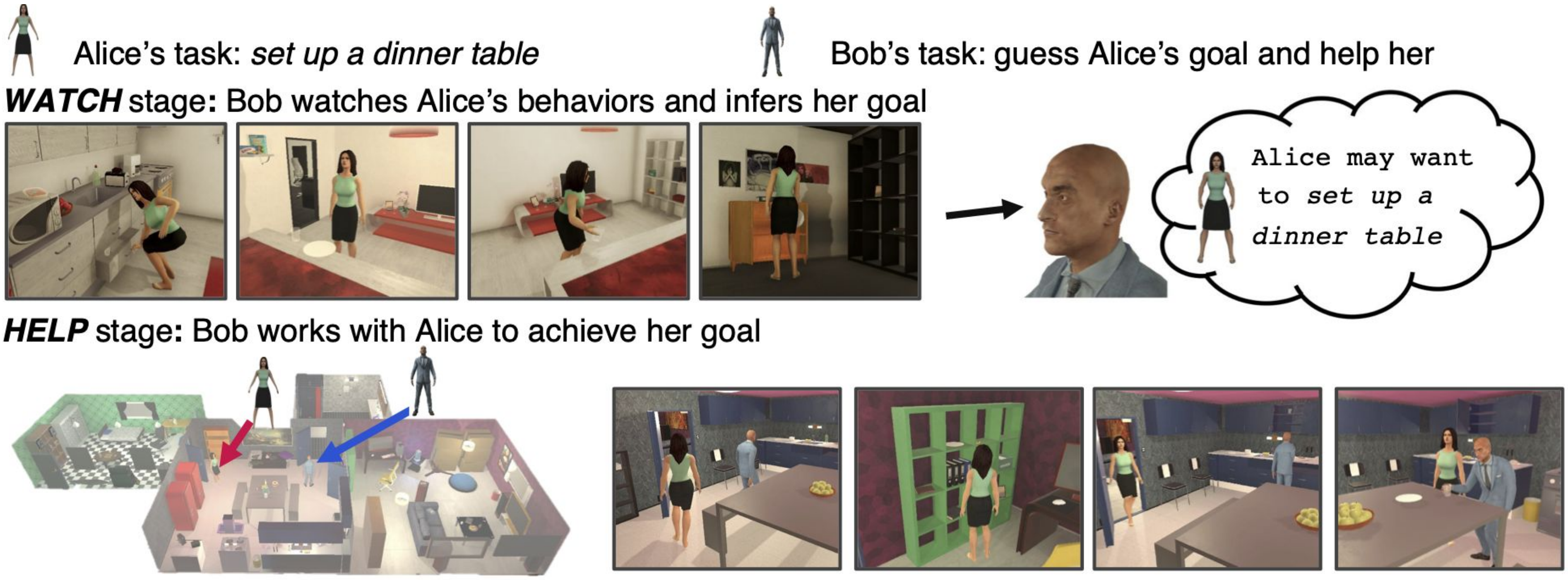
Watch And Help!

An AI agent needs to help a human-like agent perform a complex household task efficiently.

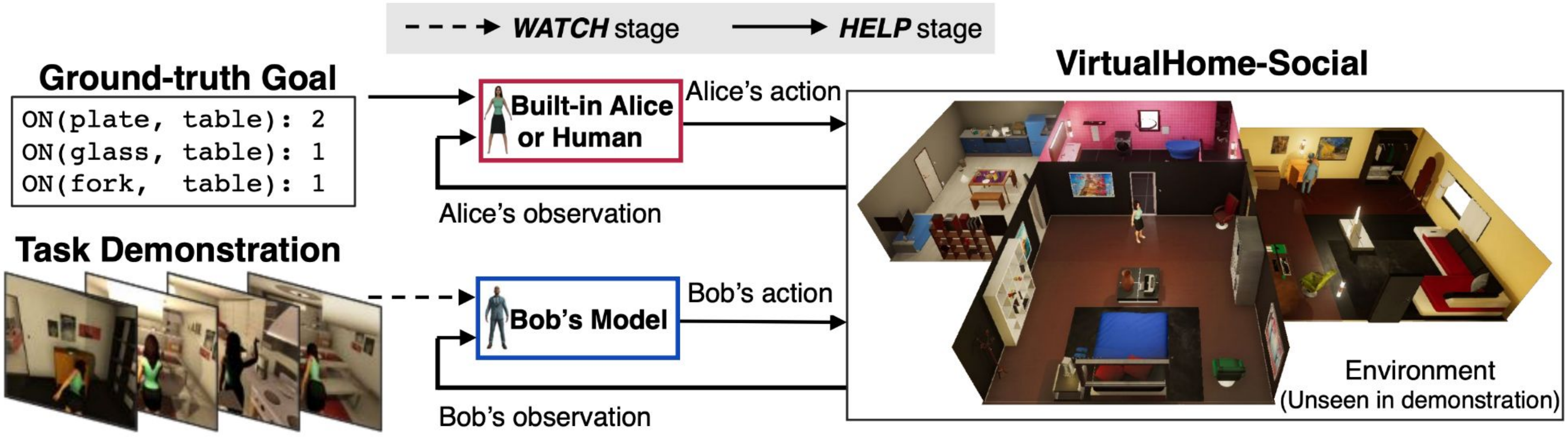
To succeed, the AI agent needs to

i) understand the underlying goal of the task by watching a single demonstration of the human-like agent performing the same task (social perception)

ii) coordinate with the human-like agent to solve the task in an unseen environment as fast as possible (human-AI collaboration).



Puig and Shu (2021).
https://github.com/xavierpuigf/watch_and_help



- Understanding of teammates' goal
- Partial observation of the environment
- Adapt to Alice's plan

MCTS - Monte Carlo Tree Search

Selection: Start from the root of the tree and select a node using a strategy (like maximizing an upper confidence bound) until you reach a leaf node.

Expansion: If the leaf node has possible unexplored actions, expand the tree by adding a new node (i.e., simulate one of the actions).

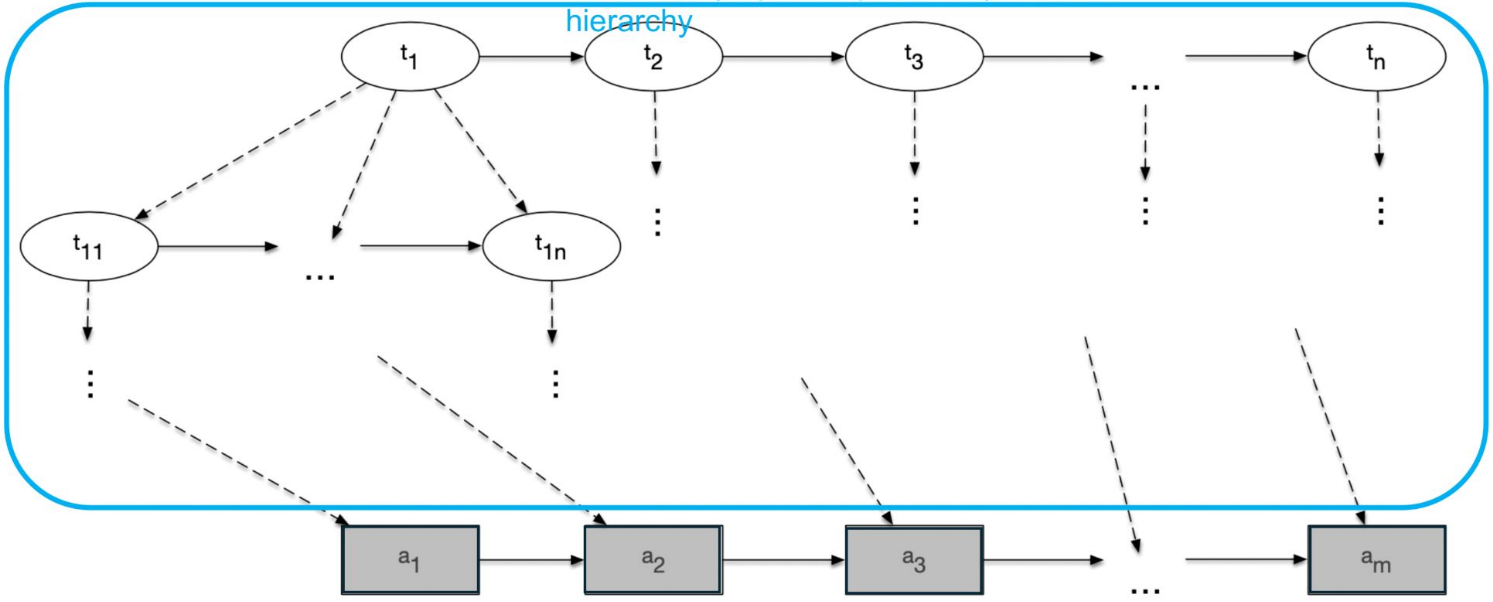
Simulation: Simulate random actions starting from the newly added node until reaching a possible end state.

Backpropagation: Update the values of all the nodes in the path based on the result of the simulation.

Hierarchical Task Networks - Monte Carlo Tree Search

HTN Plan Recognition

A solution (or plan explanation) is a task hierarchy



By Loren Rieffer-Champlin

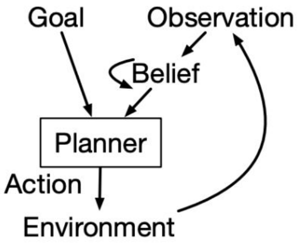


Figure 3: Overview of the human-like agent.

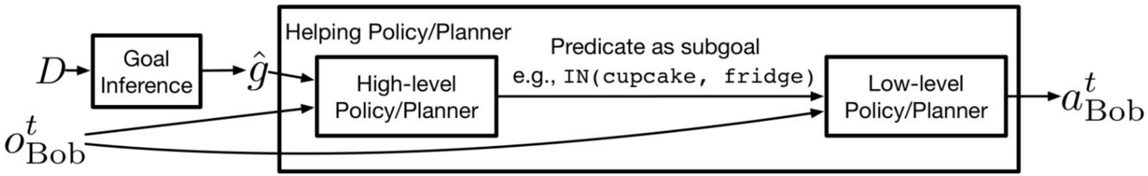


Figure 4: The overall design of the baseline models. A goal inference model infers the goal from a demonstration D and feeds it to a helping policy (for learning-based baselines) or to a planner to generate Bob's action. We adopt a hierarchical approach for all baselines.

Puig and Shu (2021)

1) a belief of object locations in the environment

2) a hierarchical planner

Table 2: Predicate sets used for defining the goal of Alice in five types of activities.

Set up a dinner table	ON(plate,dinnertable),ON(fork,dinnertable), ON(waterglass,dinnertable),ON(wineglass,dinnertable)
Put groceries	IN(cupcake,fridge),IN(pancake,fridge),IN(poundcake,fridge), IN(pudding,fridge),IN(apple,fridge), IN(juice,fridge),IN(wine,fridge)
Prepare a meal	ON(coffeepot,dinnertable),ON(cupcake,dinnertable), ON(pancake,dinnertable),ON(poundcake,dinnertable), ON(pudding,dinnertable),ON(apple,dinnertable), ON(juice,dinnertable),ON(wine,dinnertable)
Wash dishes	IN(plate,dishwasher),IN(fork,dishwasher), IN(waterglass,dishwasher),IN(wineglass,dishwasher)
Read a book	HOLD(Alice,book),SIT(Alice,sofa),ON(cupcake,coffeetable), ON(pudding,coffeetable),ON(apple,coffeetable), ON(juice,coffeetable),ON(wine,coffeetable)

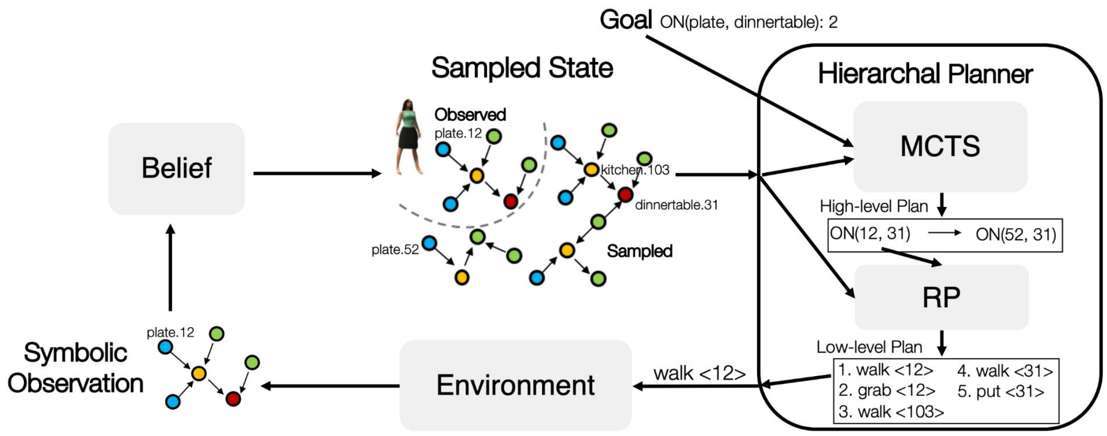


Figure 12: Schematic of the human-like agent. Based on the state graph sampled from the belief, the hierarchical planner searches for a high-level plan over subgoals using MCTS; then RP searches for a low-level plan over actions for each subgoal. The first action of each plan is sent back to the environment for execution.

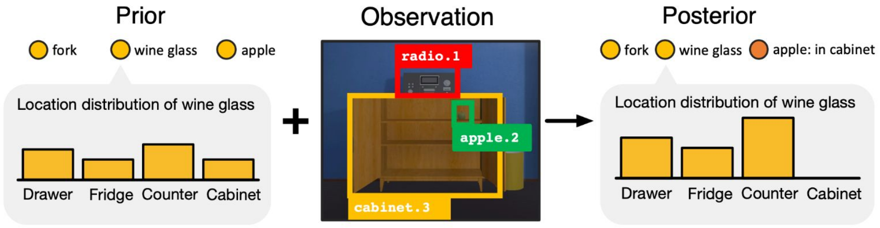


Figure 13: The agent's belief is represented as the location distribution of objects, and is updated at each step based on the previous belief and the latest observation. In the example, the open cabinet reveals that the wine glass can not be in there, and that there is an apple inside, updating the belief accordingly.

HTN Plan Recognition

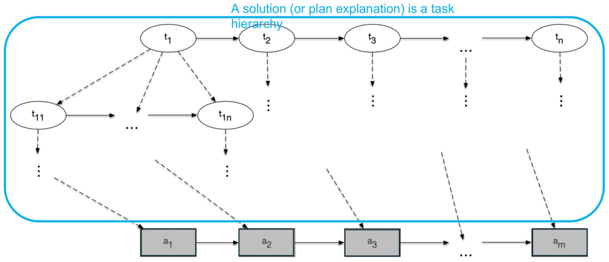


Figure 1: A challenging multi-agent cooperation problem with decentralized control, raw sensory observations, costly communication, and long-horizon multi-objective tasks.

Two AI Agents cooperate to transport multiple target objects in a multi-room house

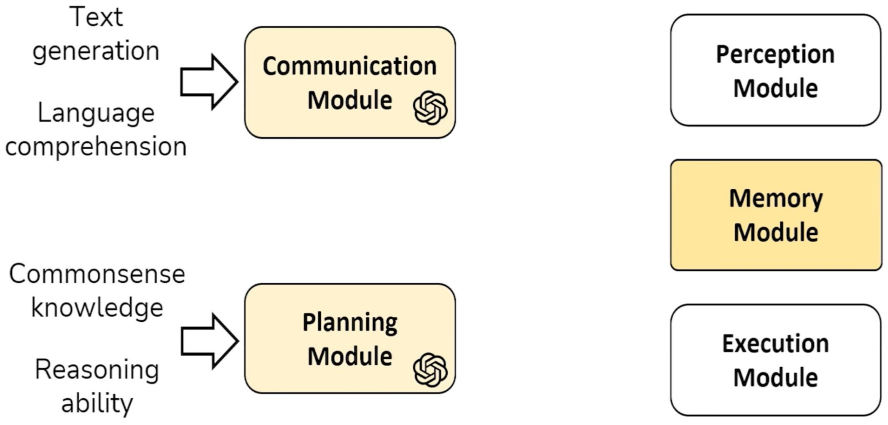
Common Sense Knowledge

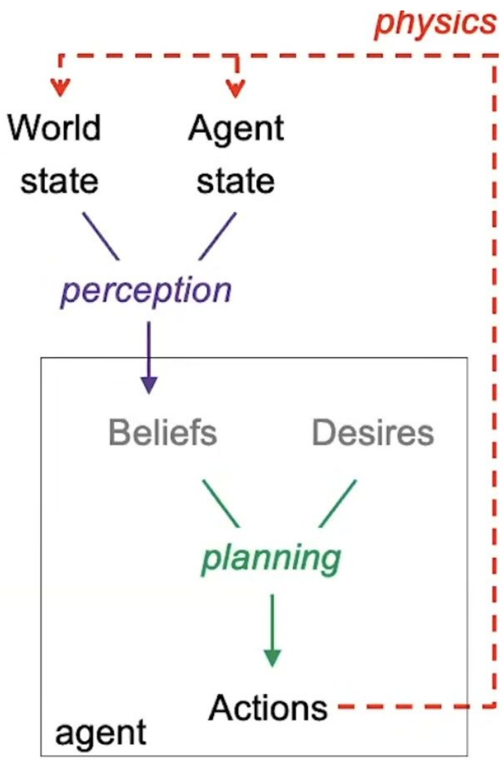
Reasoning Ability

Text Generation (Communication)

Language Comprehension

Cognitive Science -> LLM + Modular Cognitive Framework





Cooperative Embodied Language Agent

(CoELA)

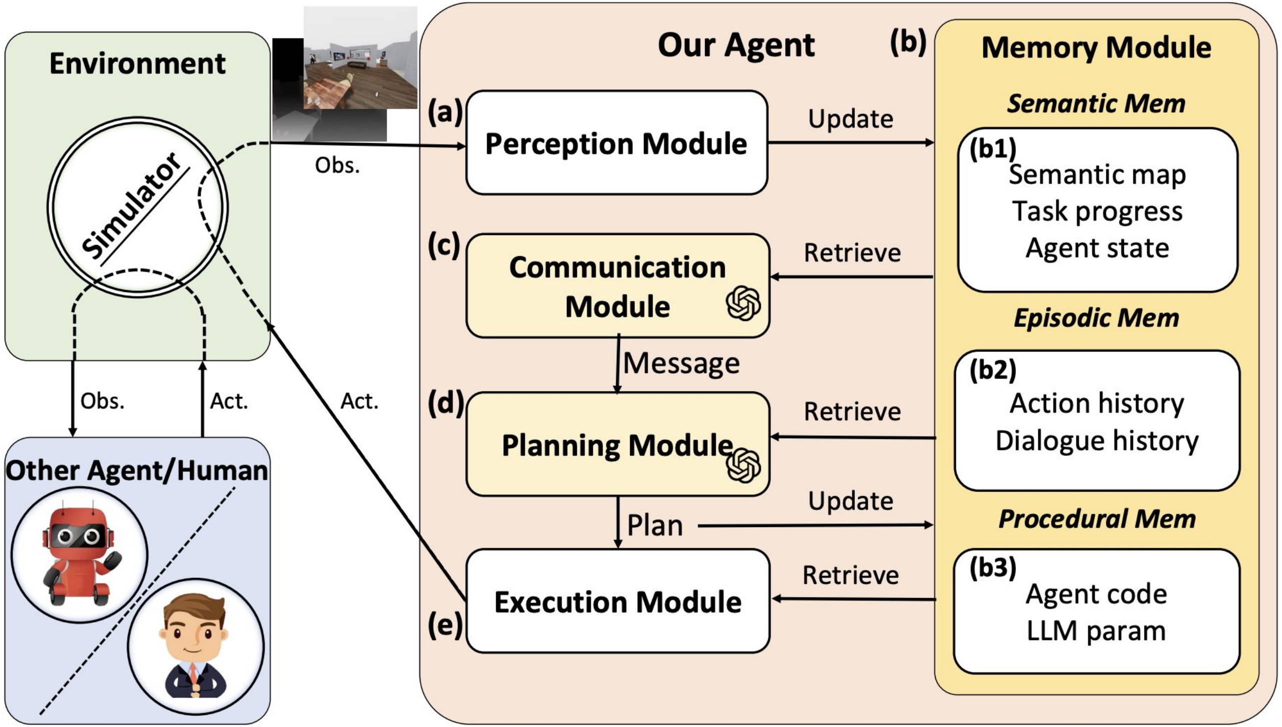
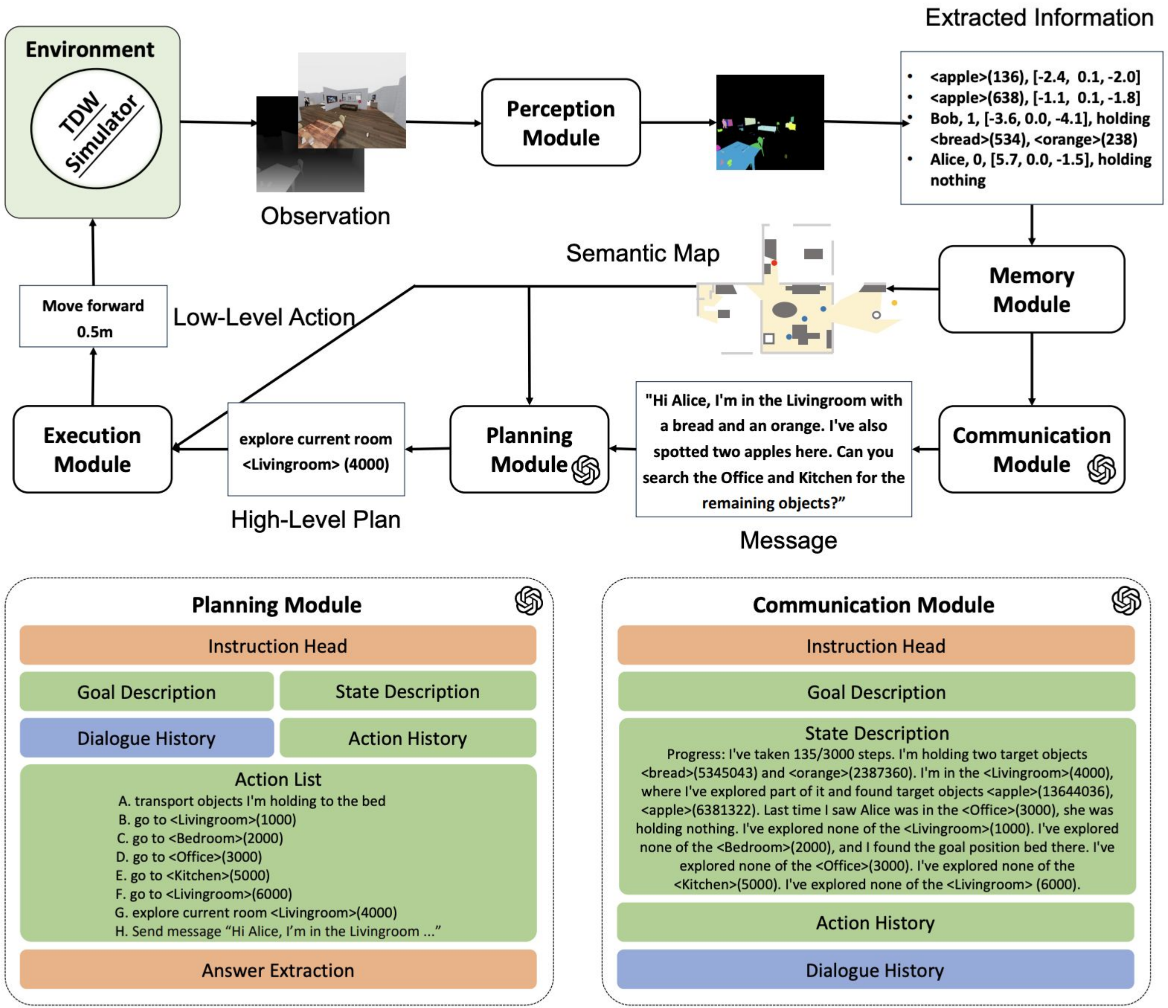


Figure 2: An overview of *CoELA*. There are five key modules in our framework: (c) The Communication Module and (d) the Planning Module leverage LLMs to generate messages and make plans, (b) The Memory Module stores the agent's knowledge and experience about the world and others in semantic, episodic and procedural memory respectively, (a) The Perception Module and (e) the Execution Module interact directly with the external environment by perceiving raw observations and generating primitive actions. More design details can be found in Appendix A.


Extracted Information
Environment
TDW Simulator
Observation
Perception Module
<apple>(136), [-2.4, 0.1, -2.0]
<apple>(638), [-1.1, 0.1, -1.8]
Bob, 1, [-3.6, 0.0, -4.1], holding <bread>(534), <orange>(238)
Alice, 0, [5.7, 0.0, -1.5], holding nothing
Semantic Map
Memory Module
Move forward 0.5m
Low-Level Action
"Hi Alice, I'm in the Livingroom with a bread and an orange. I've also spotted two apples here. Can you search the Office and Kitchen for the remaining objects?"
Execution Module
explore current room <Livingroom> (4000)
Planning Module
Communication Module
High-Level Plan
Message
Planning Module
Instruction Head
Goal Description
State Description
Dialogue History
Action History
Action List
A. transport objects I'm holding to the bed
B. go to <Livingroom>(1000)
C. go to <Bedroom>(2000)
D. go to <Office>(3000)
E. go to <Kitchen>(5000)
F. go to <Livingroom>(6000)
G. explore current room <Livingroom>(4000)
H. Send message "Hi Alice, I'm in the Livingroom ..."
Answer Extraction
Communication Module
Instruction Head
Goal Description
State Description
Progress: I've taken 135/3000 steps. I'm holding two target objects <bread>(5345043) and <orange>(2387360). I'm in the <Livingroom>(4000), where I've explored part of it and found target objects <apple>(13644036), <apple>(6381322). Last time I saw Alice was in the <Office>(3000), she was holding nothing. I've explored none of the <Livingroom>(1000). I've explored none of the <Bedroom>(2000), and I found the goal position bed there. I've explored none of the <Office>(3000). I've explored none of the <Kitchen>(5000). I've explored none of the <Livingroom> (6000).
Action History
Dialogue History

ThreeDWorld Multi-Agent Transport (TDW-MAT)
ThreeDWorld Transport Challenge (Gan et al., 2022)

- Transport Rate (TR)
  - Subtasks complement rate
- Average steps L:
  - C-WAH (communicative watch and help) tasks average steps
- Efficiency Improvement (EI)
  - delta (M) / M_{0}

Experiment

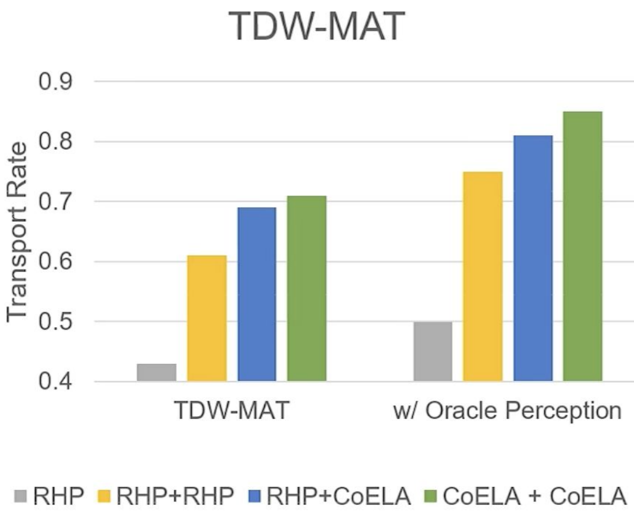
3D World - Platform (3D Multi Agent Transport) (TDW-MAT, Page 6)

	Symbolic Obs	**Visual Obs**
MHP	111	141
MHP + MHP	75(↑33%)	103(↑26%)
MHP + *CoELA*	59(↑45%)	94(↑**34%**)
CoELA* + *CoELA	**57(↑49%)**	**92(↑34%)**

Table 2: **Quantitative results on C-WAH.** We report the average steps(Efficiency Improvement) here over 5 runs for MHP and 1 run for *CoELA* due to cost constraints. The best performance is achieved when cooperating with *CoELA*.

Evaluations

Metrics: Transport Rate ⬆



Metrics: Average Steps ⬇

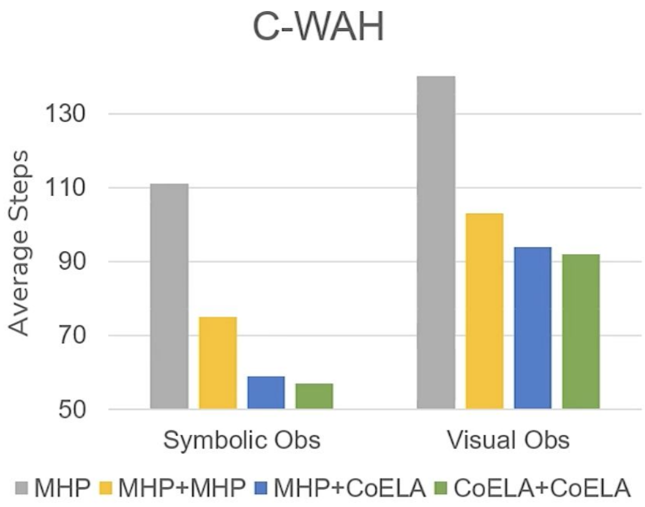
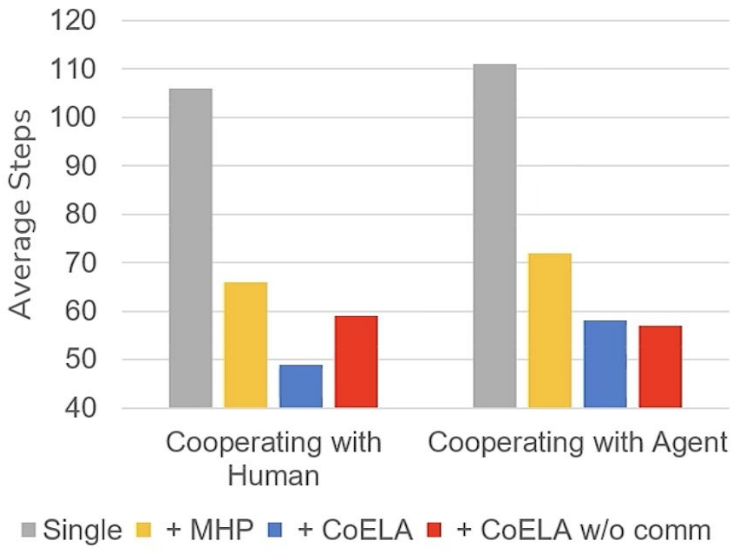
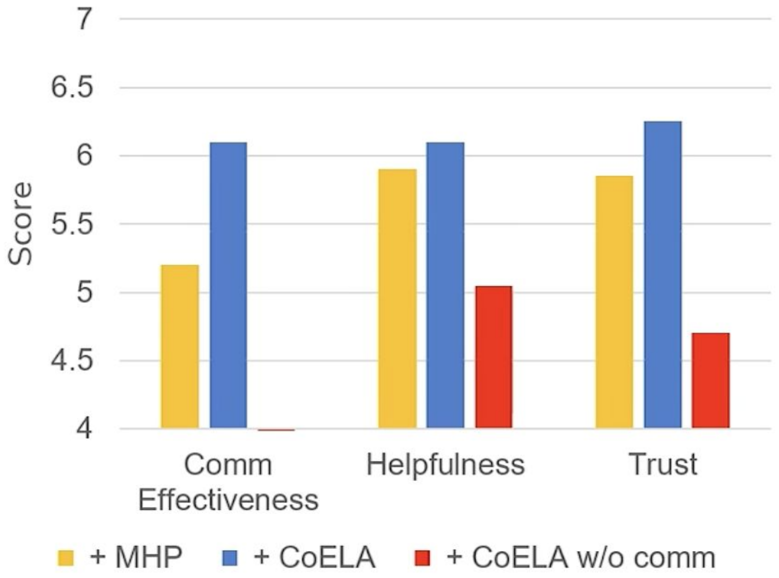


Figure 3: **Example cooperative behaviors** demonstrating *CoELA* can communicate effectively and are good cooperators.

Cooperate with Humans (8 HUMANS)





Human Collaboration

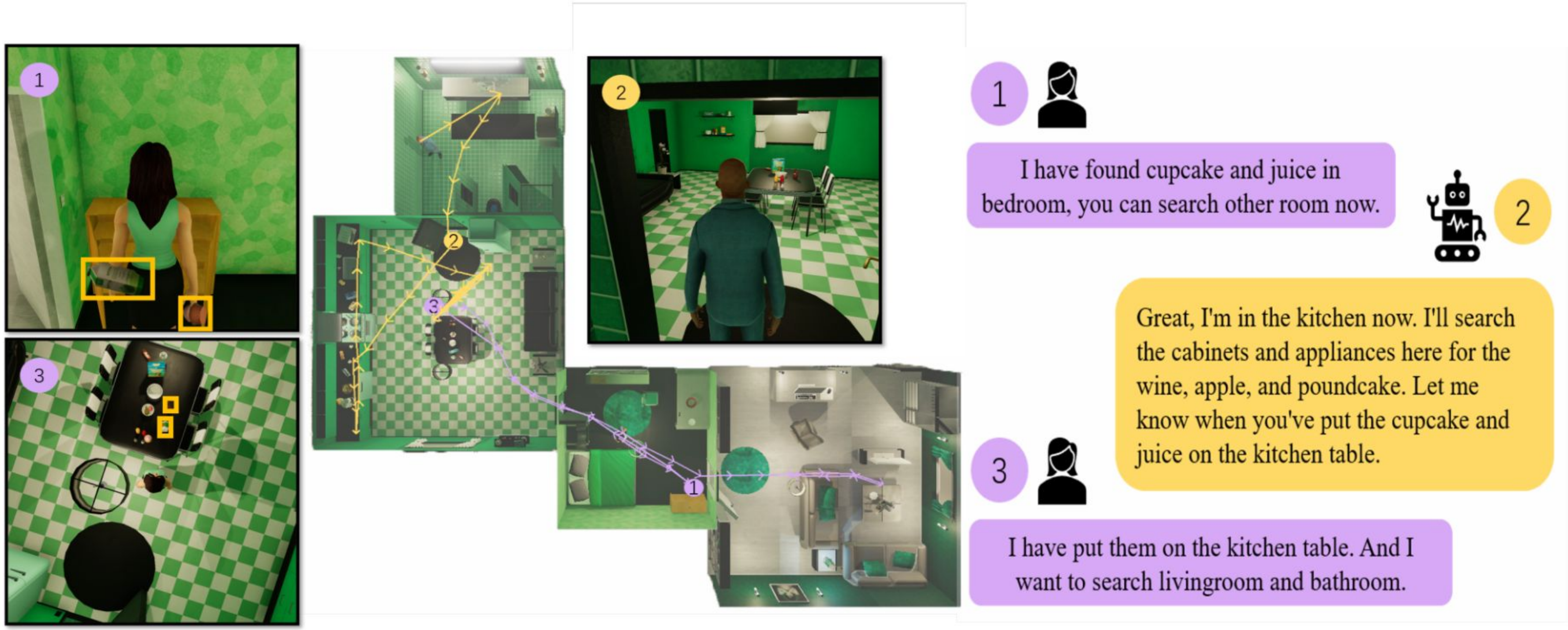
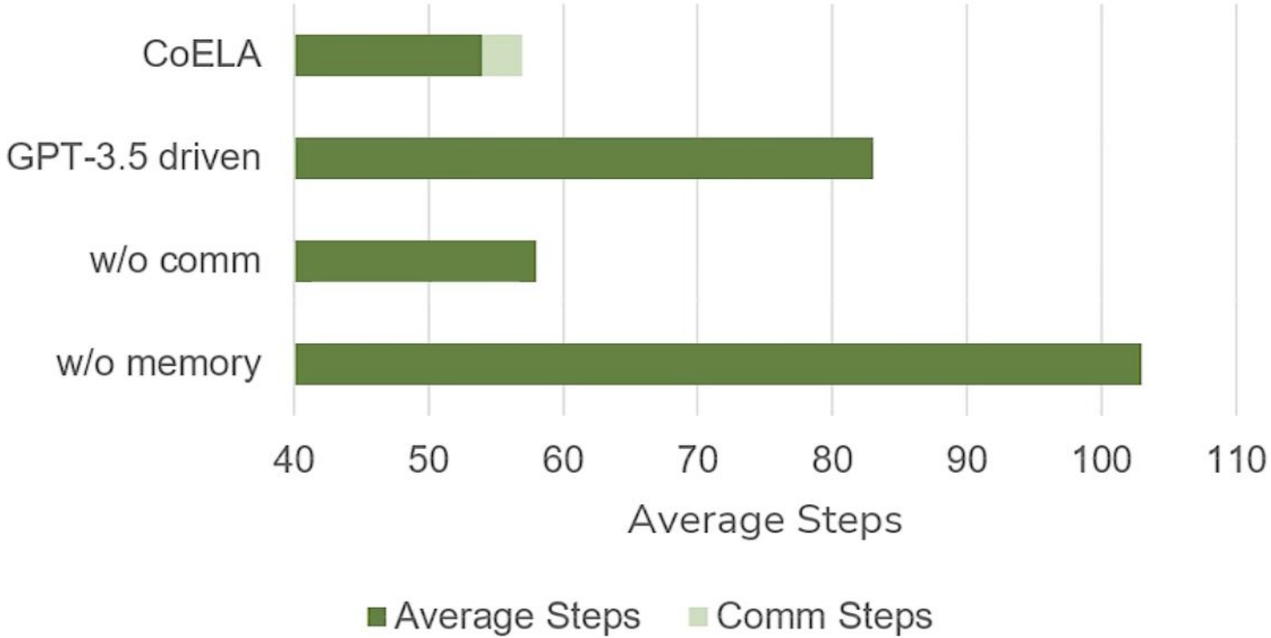


Figure 10: A qualitative example in Human + *CoELA* experiments, showcasing *CoELA* can communicate with Humans well and end up with a perfect division of the exploration trajectory.

Strong LLM is required



Limitation

- No usage of 3D spatial information. We can use 3D models to improve performance

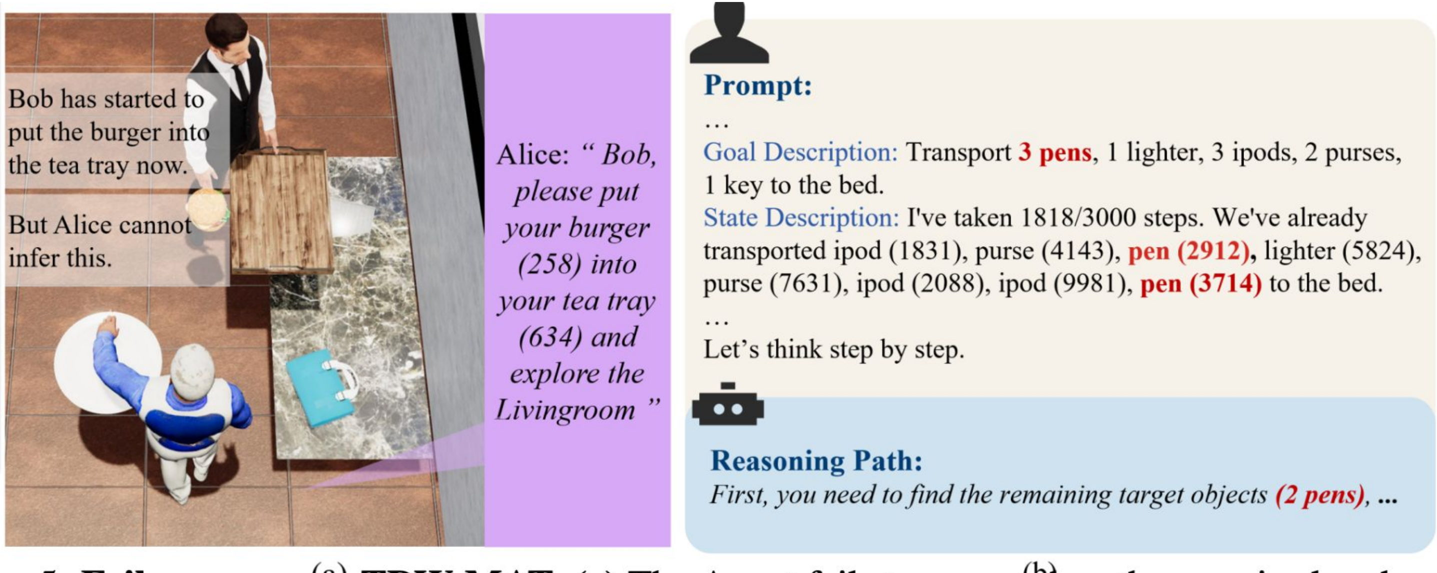


Figure 5: **Failure cases on TDW-MAT**. (a) The Agent fails to reason the other one is already putting the burger into the container. (b) The LLM counts the number of the remaining target objects wrong.

Limitation

- Unstable performance on complex reasoning

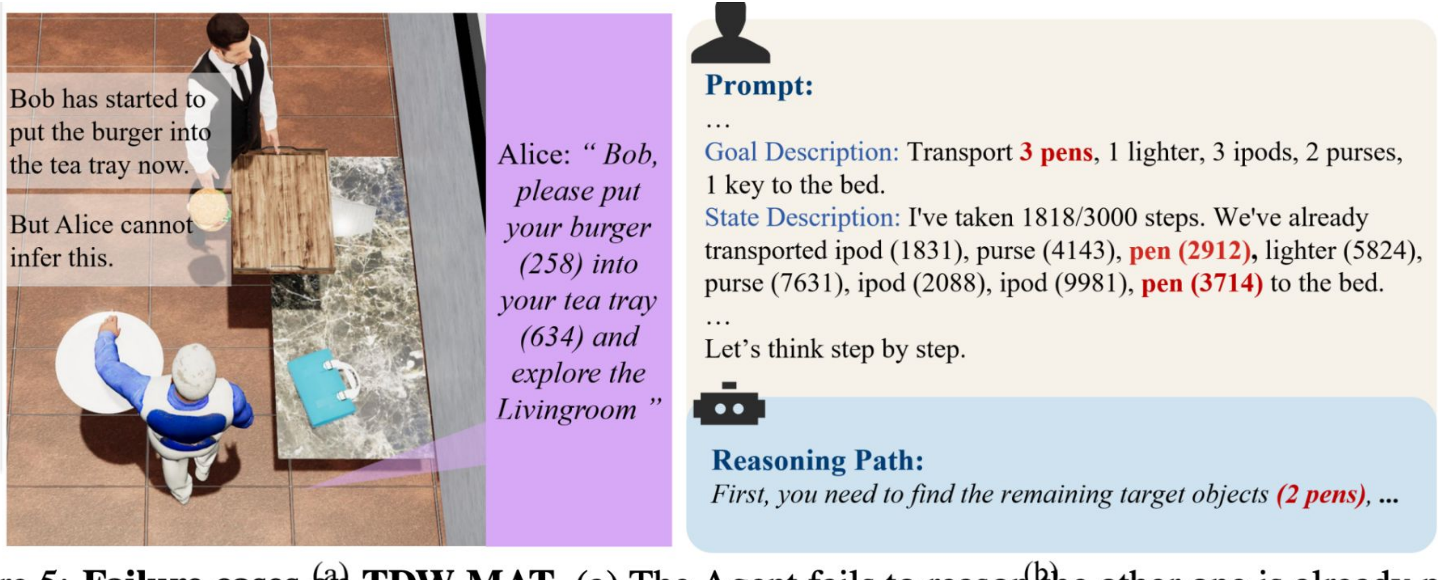


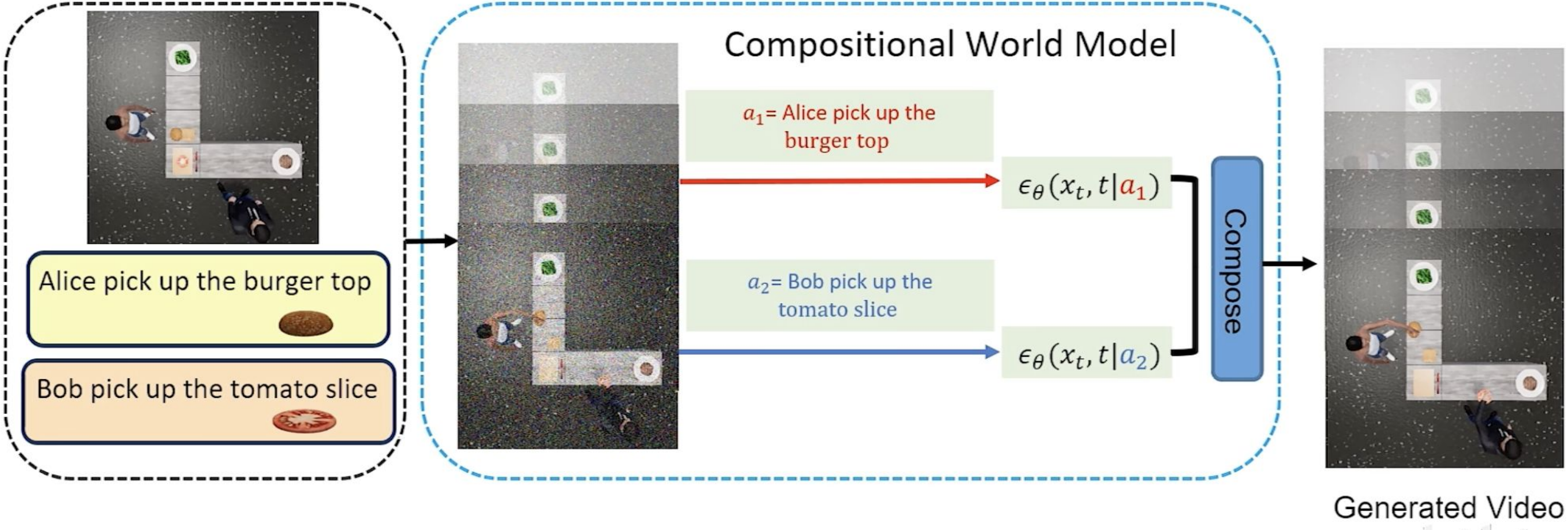
Figure 5: **Failure cases on TDW-MAT**. (a) The Agent fails to reason the other one is already putting the burger into the container. (b) The LLM counts the number of the remaining target objects wrong.

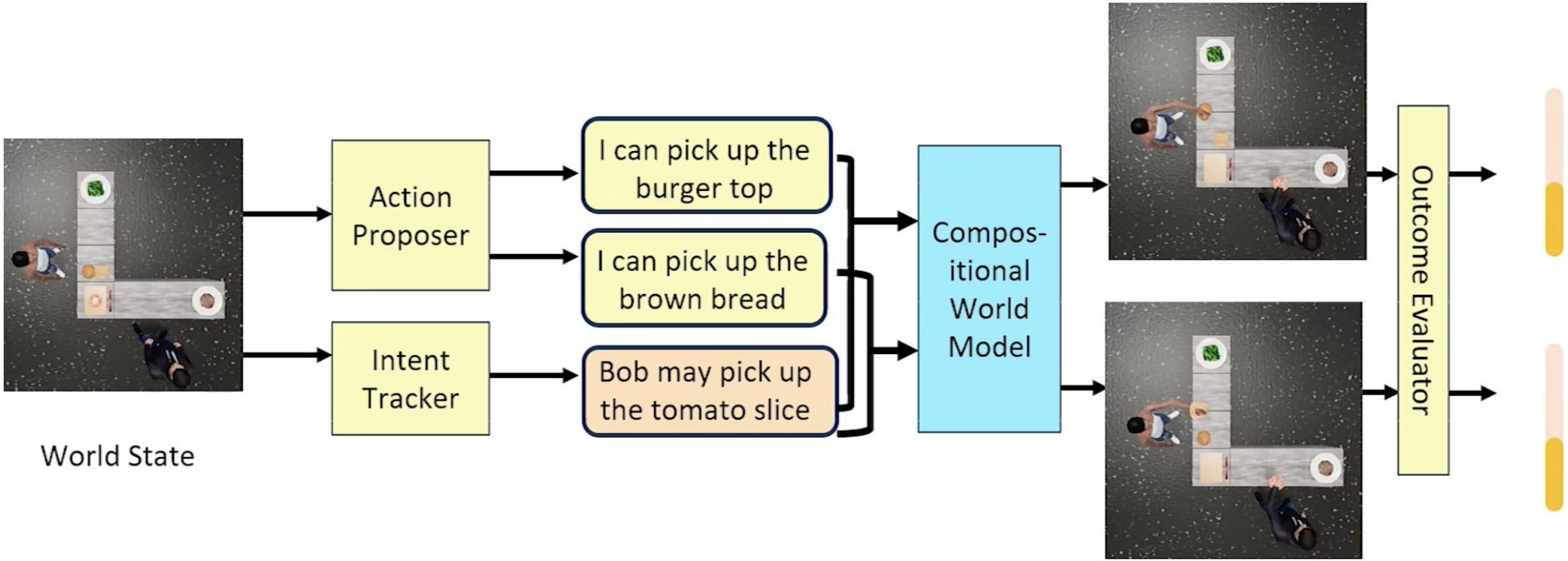
Questions/Problems

1. How does CoELA handle the scenario where no consensus is reached? For example, Alice wants Bob to goto A and Bob wants Alice to goto B.
2. Previous works using LLM for implementing embodied agents?
3. Cognitive architecture is not a language based approach. While it seeks to integrate Large Language Models (LLMs) into a cognitive-inspired modular framework for cooperative embodied agents, it overlooks a key principle of cognitive architecture: such architectures are typically not language-based but instead rely on symbolic or neural representations for reasoning and perception.

COMBO, a compositional world model with video diffusion models. A planning framework combining VLMs to imagine the world changes in the long run for better multi-agent cooperation.

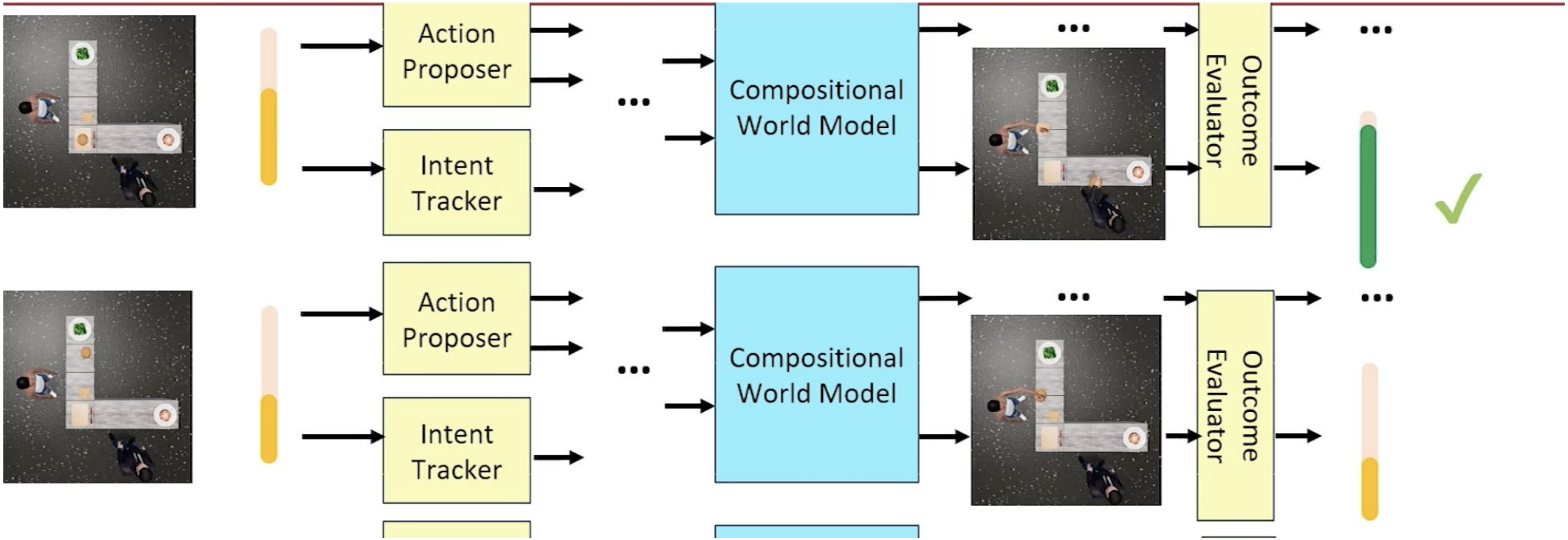
Using diffusion model to Monte Carlo Sampling





Action Proposer
Intent Tracker
I can pick up the burger top
I can pick up the brown bread
Bob may pick up the tomato slice
Compos-itional World Model
Outcome Evaluator
World State

Find best path and execute it



Average Steps Evaluation (Fewer is better)

Metrics: Average steps ⬇