# Human-AI Cooperation: An Inclusive Embodied Social Intelligence Challenge

NeurIPS 2024 Poster

https://vis-www.cs.umass.edu/CHAIC/

Weihua Du* · Qiushi Lyu* · Jiaming Shan · Zhenting Qi · Hongxin Zhang · Sunli Chen
Andi Peng · Tianmin Shu · Kwonjoon Lee · Behzad Dariush · Chuang Gan


* *Equal Contribution*

# Humans can help others with constraints

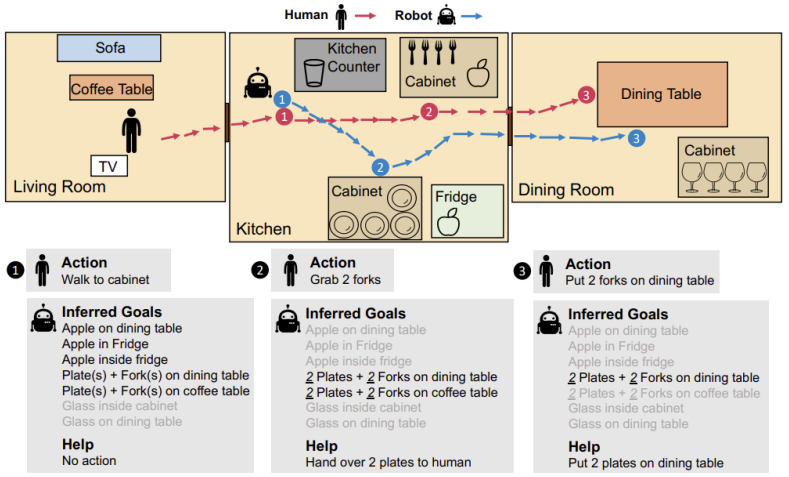
- Humans can figure out who needs help without communication, e.g.:

- Finding out people in need is an essential commonsense, while AI is hard to do it.
  - Most AI agents need clear instruction about what to do next, while real life lacks such instruction.

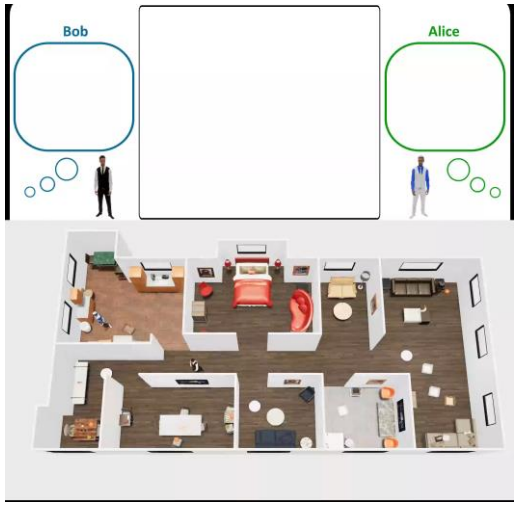
# Existing benchmarks neglect social intelligence

| Challenge Name | Accessibility Setting | Multi-Agent Support | Goal Inference | Observation Type | Outdoor Scenes | Emergent Event |
|---|---|---|---|---|---|---|
| Watch-and-Help (Puig et al., 2021) | × | ✓ | × | Symbolic | × | × |
| NOPA (Puig et al., 2023a) | × | ✓ | ✓ | Symbolic | × | × |
| Social Rearrangement (Szot et al., 2023) | × | ✓ | × | Visual* | × | × |
| TDW-MAT (Zhang et al., 2023) | × | ✓ | × | Visual | × | × |
| Hazard (Zhou et al., 2024) | × | × | × | Visual | ✓ | × |
| Smart Help (Cao et al., 2024) | ✓ | ✓ | ✓ | Symbolic** | × | × |
| CHAIC (Ours) | ✓ | ✓ | ✓ | Visual | ✓ | ✓ |

e.g.




NOPA [1]




TDW-MAT [2]

[1] Puig, Xavier, et al. "NOPA: Neurally-guided Online Probabilistic Assistance for Building Socially Intelligent Home Assistants"
[2] Zhang, Hongxin, et al. "Building Cooperative Embodied Agents Modularly with Large Language Models."






# Constrained Human-AI Collaboration Challenge (CHAIC)

Constrained Human-AI Cooperation Challenge with Diverse Agents and Rich Scenes and Elements

## Diverse Agents

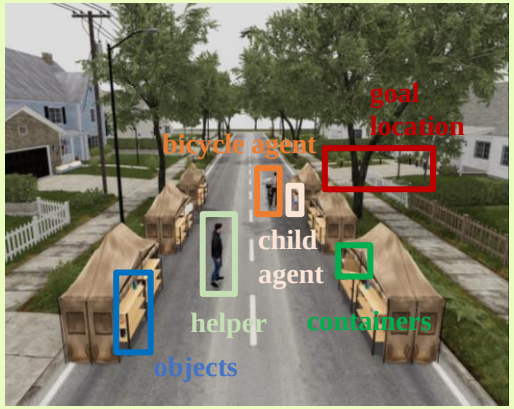

## Rich Scenes

### Indoor Examples


Obstacle Task


High Container Task

### Outdoor Examples




Shopping Task


Moving House Task

## Elements

objects

containers

goal location

# Constrained Human-AI Collaboration Challenge

- CHAIC features various constraints for agents:
- Common constraint:
  - **Height constraint**: an action is more likely to fail when the objects is higher / lower;
  - **Weight constraint**: an action is more likely to fail when the objects is heavier;
- The probabilities vary between different types of agents:
  - For example:
    - **Little girl**: struggle to reach objects in high places;
    - **Man in wheelchair**: struggle to reach low places, cannot go through narrow places.

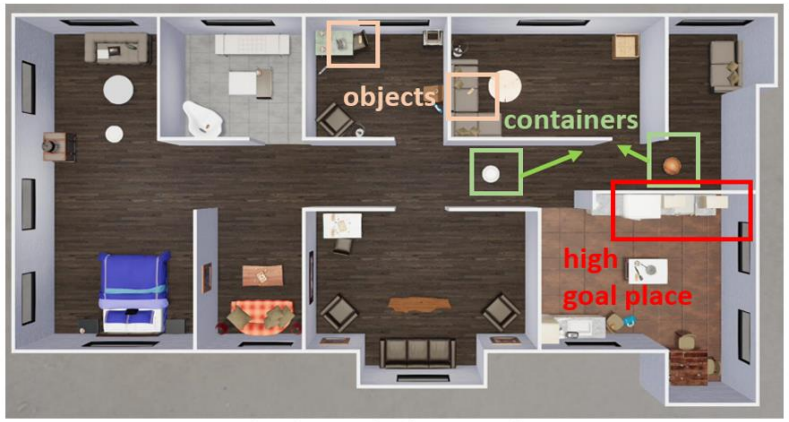
# Indoor Scene

- Various scenes: both indoor and outdoor
- Indoor:


high container task

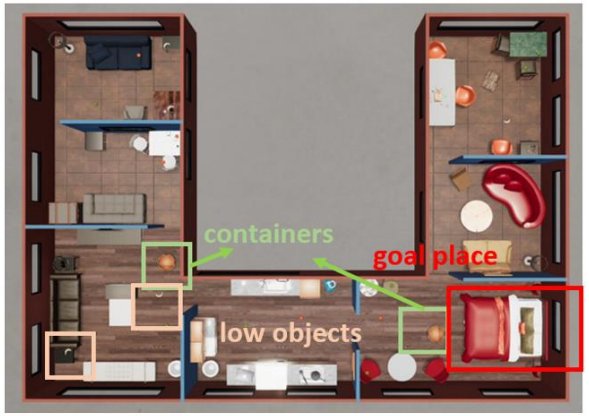

wheelchaired task


low container task



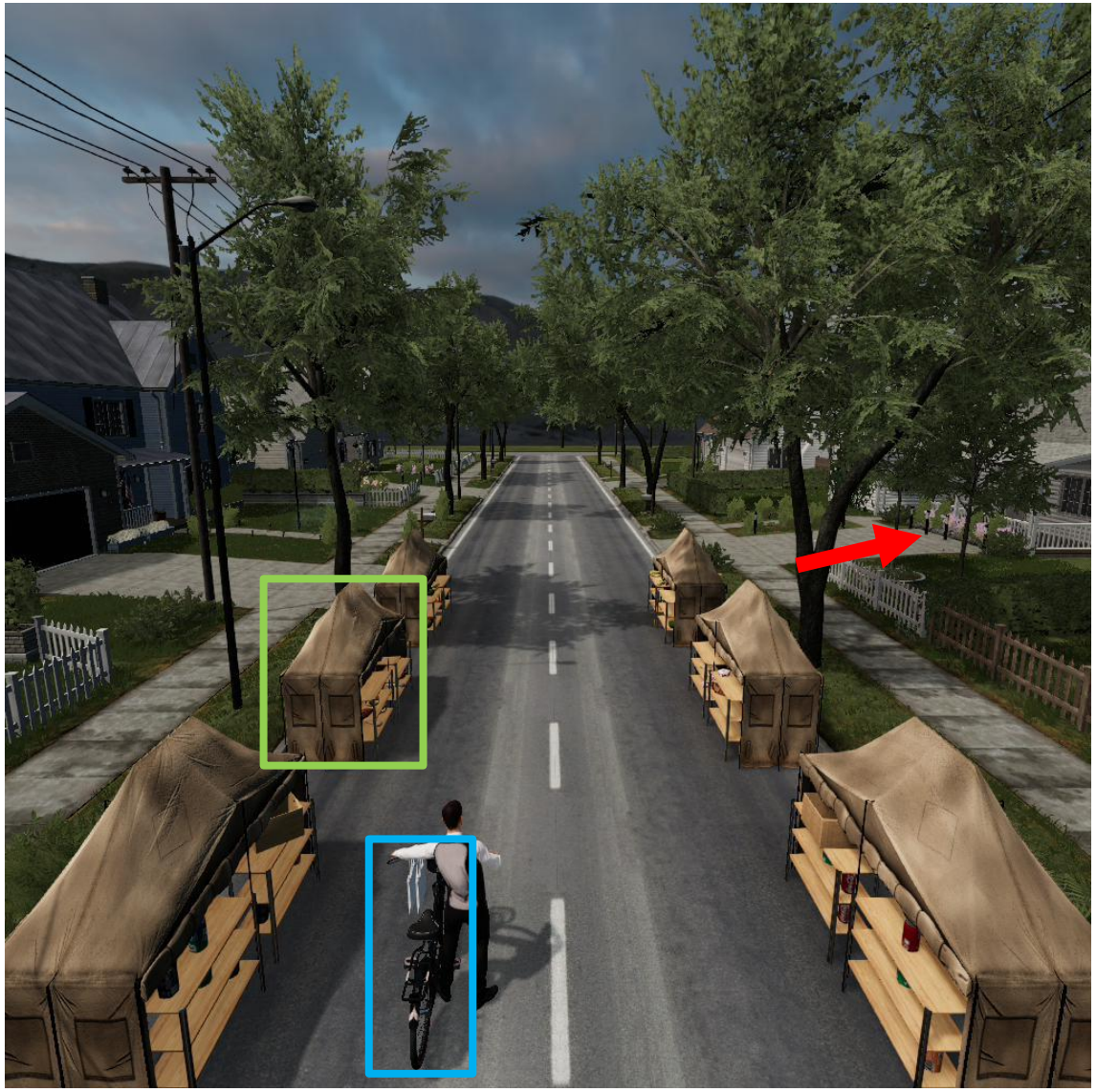

high goal place task


high target task




low target task

# Outdoor Scene (shopping)



The agent is riding a bike and he needs to buy something from the store and carry them to home.

# Outdoor Scene (shopping)

Each store sells a unique category of items.

And there may be a box inside some of the stores.

Agent can use the box as a container to collect more things.

# Outdoor Scene (furniture)

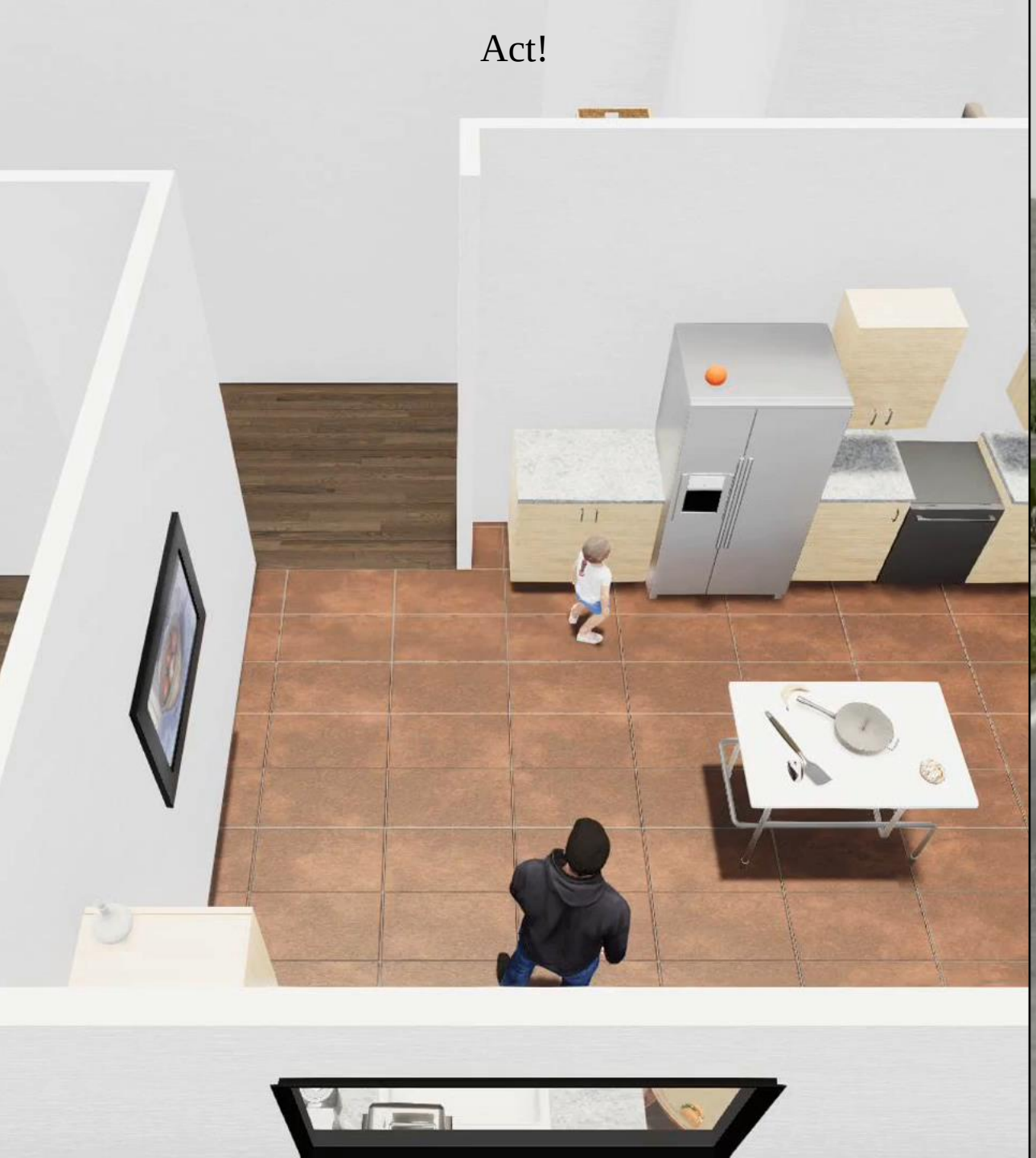
The agent needs to carry the furniture to the truck.

Some of the furniture is so heavy that the constrained agent could only lift them with a small chance.

The agents can carry one piece of heavy furniture together if it can’t be carried by only one person.


Indoor Scene
(egocentric view)
The girl cannot pick up the orange because of height
I can pick up the orange for her.
Outdoor Scene
(egocentric view)
The woman cannot pick up the sofa because of weight
I can lift the sofa with her together.

Act!

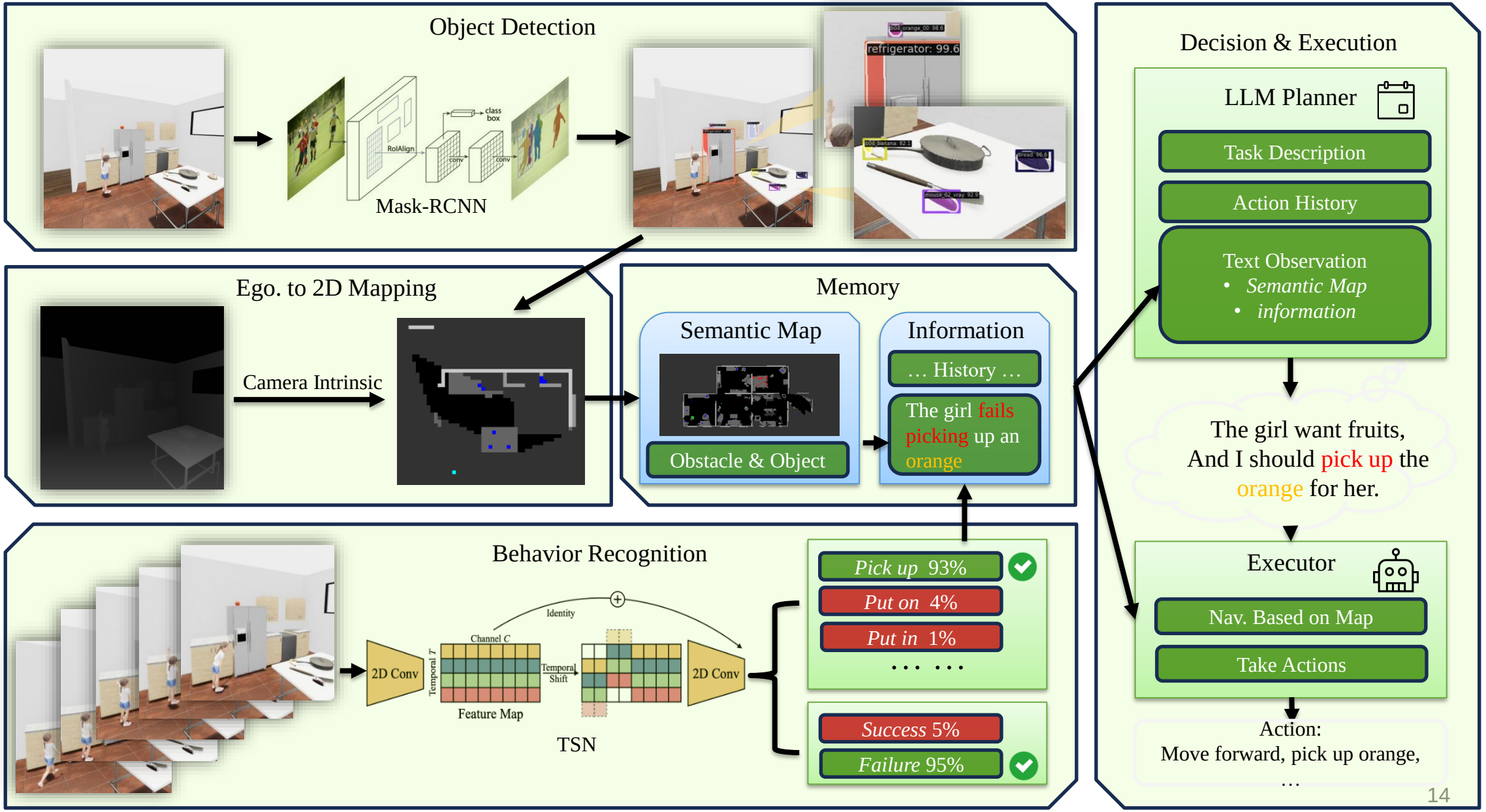


# Baselines

- Planning-Based
  - Random
  - Rule-Based Planner
  - End-to-End VLM
  - LLM + behavior modeling (BM) ★
- Learning-Based
  - Reinforcement Learning
  - Smart Help [3]
- Oracle
  - The helper knows everything, as a comparison

[3] Cao, Zhihao, et al. "Smart Help: Strategic Opponent Modeling for Proactive and Adaptive Robot Assistance in Households."

# LLM + BM Helper Baseline



# Results

| Indoor | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Helper Agent | No Constraint | | High Target | | High Container | | High Goalplace | |
| | TR(EI)↑ | IA↑ | TR(EI)↑ | IA↑ | TR(EI)↑ | IA↑ | TR(EI)↑ | IA↑ |
| w/o | 0.53 | / | 0.30 | / | 0.37 | / | 0.28 | / |
| Random | 0.52(-0.02) | 0.24 | 0.27(-0.05) | 0.29 | 0.36(0.00) | 0.25 | 0.33(0.10) | 0.14 |
| RHP | 0.64(0.15) | 0.15 | 0.35(0.11) | 0.29 | 0.45(0.19) | 0.21 | 0.35(0.18) | 0.21 |
| VLM (GPT-4o) | 0.63(0.14) | 0.24 | 0.33(0.06) | **0.32** | 0.43(0.12) | **0.40** | 0.26(-0.20) | 0.33 |
| LLM (GPT-4) + BM | **0.65(0.17)** | **0.25** | **0.38(0.19)** | 0.29 | **0.49(0.24)** | 0.30 | **0.36(0.23)** | **0.35** |
| Oracle | 0.77(0.31) | 0.88 | 0.49(0.37) | 0.91 | 0.69(0.47) | 0.91 | 0.61(0.56) | 0.90 |

| Indoor | | | | | Outdoor | | | |
|---|---|---|---|---|---|---|---|---|
| Helper Agent | Low Target | | Obstacle | | Shopping | | | Furniture |
| | TR(EI)↑ | IA↑ | TR(EI)↑ | IA↑ | TR(EI)↑ | IA↑ | ER↓ | TR(EI)↑ |
| w/o | 0.51 | / | 0.07 | / | 0.37 | / | / | 0.17 |
| Random | 0.50(-0.01) | 0.31 | 0.21(0.56) | 0.24 | 0.39(0.05) | 0.34 | 0.32 | 0.48(0.68) |
| RHP | 0.66(0.23) | 0.28 | **0.44**(0.77) | 0.17 | 0.49(0.22) | 0.44 | **0.30** | 0.65(0.72) |
| VLM (GPT-4o) | 0.69(0.26) | **0.46** | 0.40(0.86) | 0.35 | 0.50(0.25) | 0.72 | 0.39 | **0.70(0.78)** |
| LLM (GPT-4) + BM | **0.70(0.27)** | 0.43 | 0.42(**0.89**) | **0.47** | **0.58(0.33)** | **0.74** | 0.38 | 0.69(0.77) |
| Oracle | 0.82(0.38) | 0.91 | 0.60(0.87) | 0.82 | 0.61(0.39) | 0.87 | 0.17 | 0.76(0.80) |

- Transport Rate (TR), Efficiency Improvement (EI), and Goal Inference Accuracy (IA)
- LLM+BM Helper emerges as a strong baseline.
- Compared with Oracle, all the methods have improvement space (especially IA).