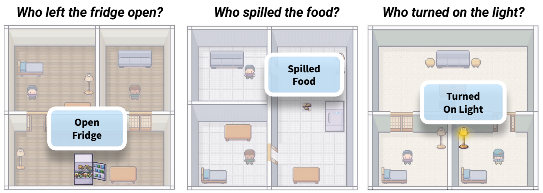
MARPLE: A Benchmark for Long-Horizon Inference

Emily Jin*, Zhuoyi Huang*, Jan-Philipp Fränken, Weiyu Liu, Hannah Cha,
Erik Brockbank, Sarah A. Wu, Ruohan Zhang, Jiajun Wu, Tobias Gerstenberg

Project Page

Motivation

Everyday, we solve a number of **"whodunit" problems** that require long horizon inferences.

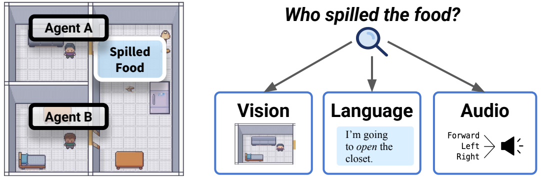


Humans draw on their understanding of the physical world, human behaviors, and multimodal cues (vision, language, audio).

Inference Scenarios

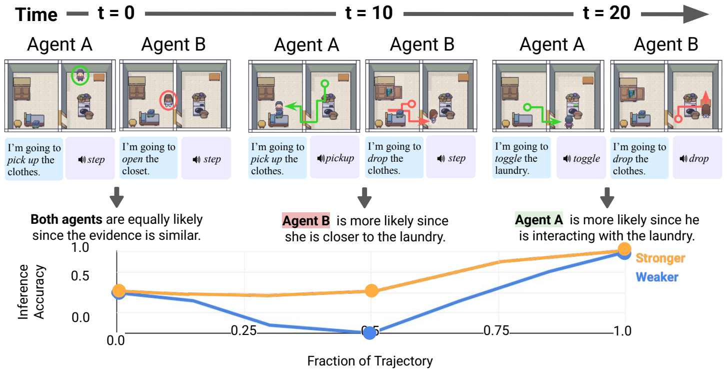
MARPLE challenges models to **figure out who did it by**:

- Reasoning over **long time horizons**.
- Reasoning at a high-level about **complex scenarios.**
- Integrating evidence from **multiple modalities**.



Problem Setup

Which agent is more likely to have turned on the laundry? Answer: Agent A


Time
t = 0
t = 10
t = 20
Agent A
Agent B
Agent A
Agent B
Agent A
Agent B
I'm going to *pick up* the clothes.
step
I'm going to *open* the closet.
step
I'm going to *pick up* the clothes.
pickup
I'm going to *drop* the clothes.
step
I'm going to *toggle* the laundry.
toggle
I'm going to *drop* the clothes.
drop
Both agents are equally likely since the evidence is similar.
Agent B is more likely since she is closer to the laundry.
Agent A is more likely since he is interacting with the laundry.
Inference Accuracy
1.0
0.5
0.0
0.0
0.25
0.5
0.75
1.0
Stronger
Weaker
Fraction of Trajectory

Contributions

Multimodal Simulator

- **Diverse** agent behaviors of **semantically rich** activities
- Within **procedurally generated** households
- Multimodal evidence (**vision, language, audio**)

Inference Scenarios

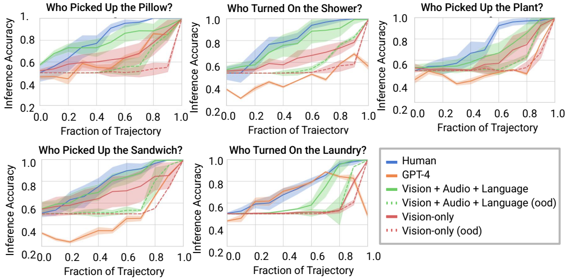
- **5** inference scenarios, **varying in difficulty**
- **Pre-collected datasets** for training & evaluation
- **Evaluation Metric:** when method achieves high inference accuracy

Benchmark Experiments

- Performance of ML baselines (**simulation w/ learned agent models**, **GPT-4**)
- **Behavioral study** with human participants

Results

Performance Across All Inference Scenarios



- **Humans** outperform all AI baselines.
- **GPT-4** fails to converge for two scenarios.
- Simulation models (**V+A+L**, **V**) converge but struggle to generalize.

Generalization of Simulation Baselines

Table: Evidence (fraction of trajectory) to achieve 0.8 accuracy.

	Human	Vision + Audio + Language	Vision + Language	Vision + Audio	Vision Only
ID⬇	0.48	0.58	0.64	0.80	0.85
OOD⬇	0.48	0.81	0.85	0.91	0.92

Humans make accurate predictions earlier, especially OOD.
For even the best simulation baseline, performance drops by 23%.