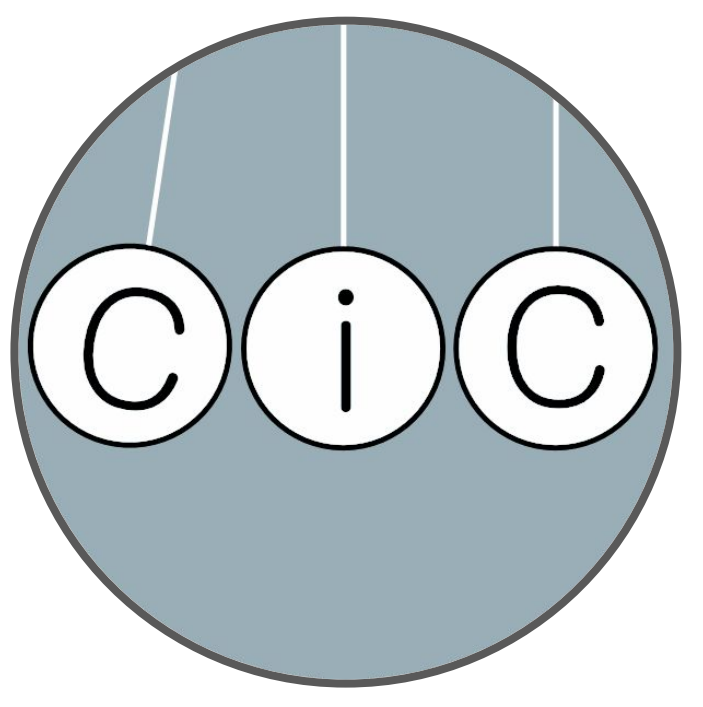




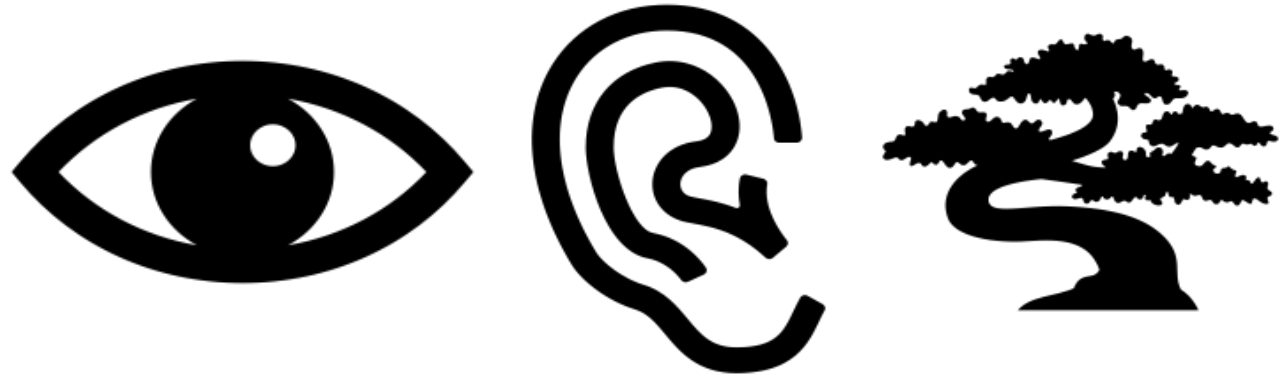
Whodunnit? Inferring what happened from multimodal evidence



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

BACKGROUND

People combine information from multiple senses to understand the environment.



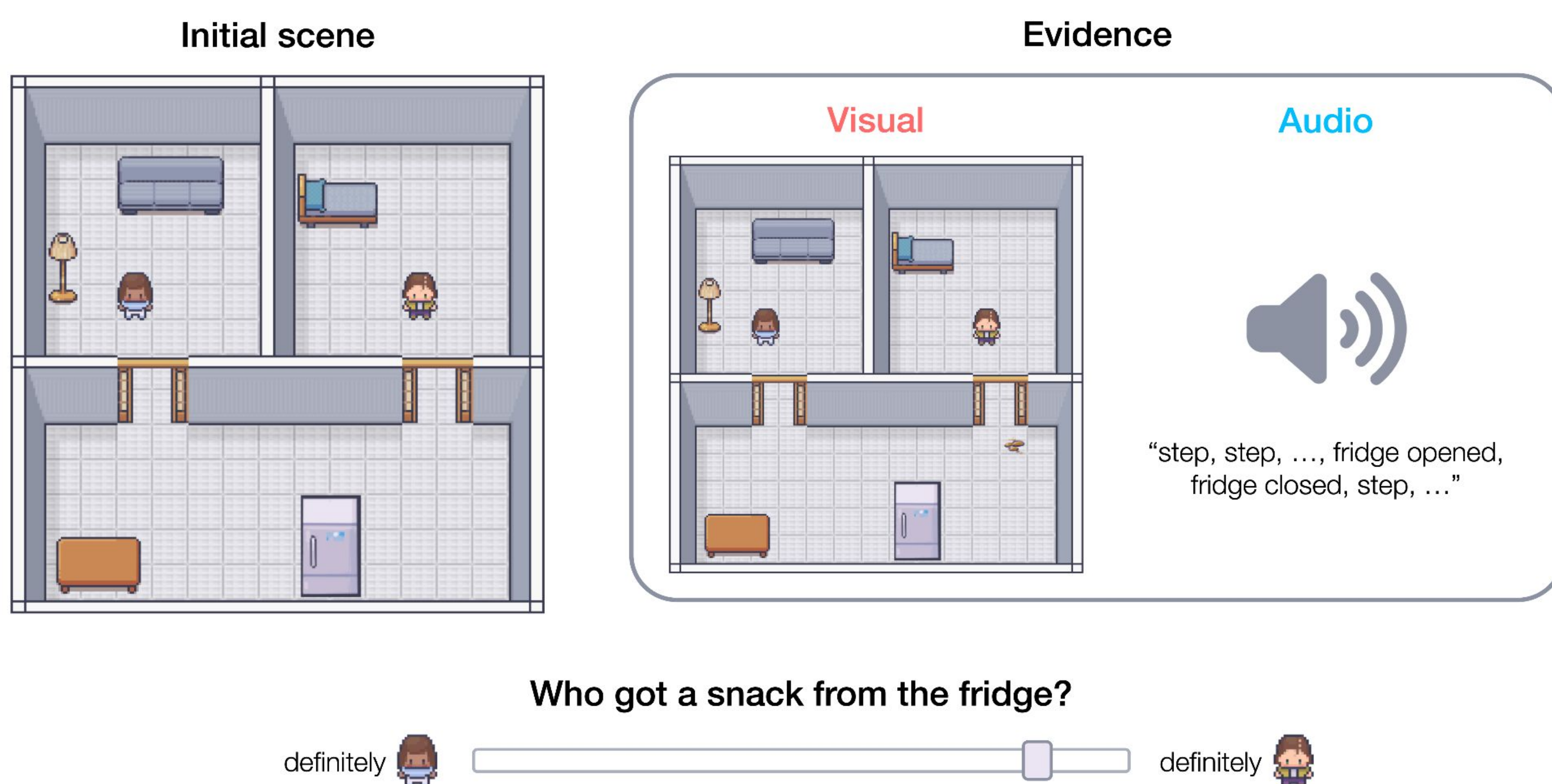
Mental models of how agents plan actions allow us reconstruct others' past behavior.



How do people combine evidence from multiple senses to understand past behavior? *And can AI systems (LLMs, MLMs)?*

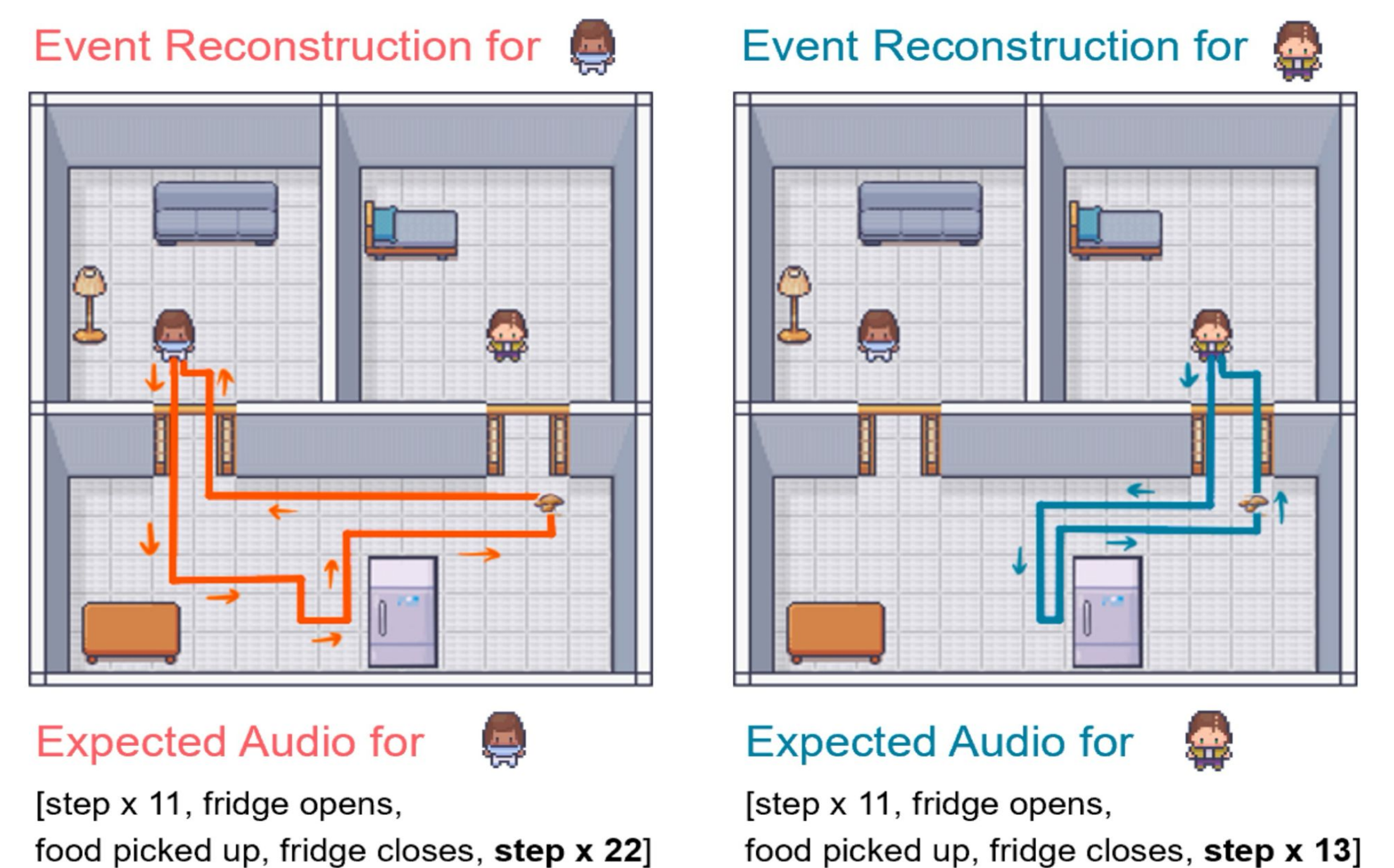
EXPERIMENT

Agents get a snack from the kitchen or watch TV in the living room, leaving audio and/or visual traces.



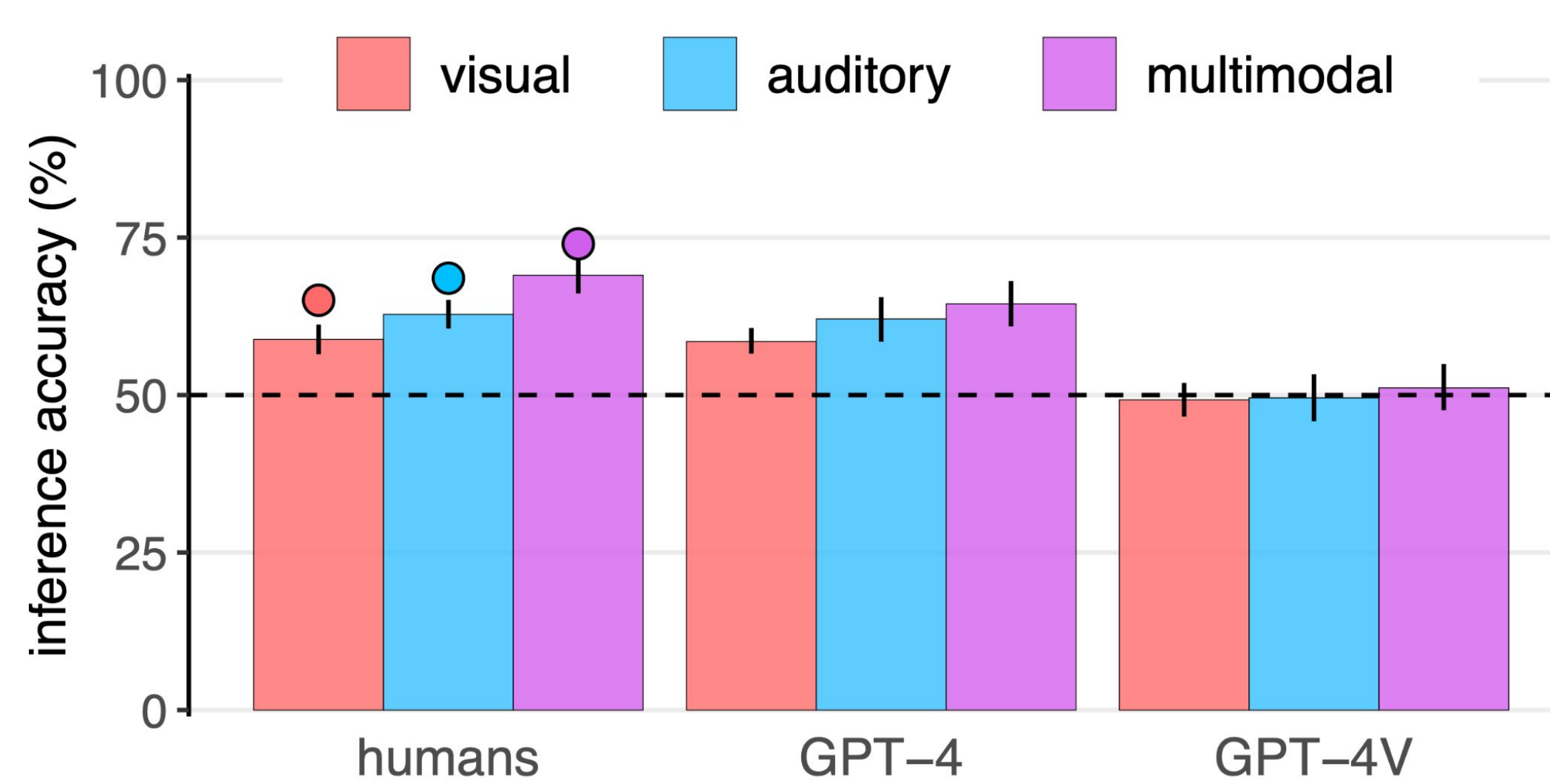
MODELS

1. **GPT-4V** - images & audio transcript
2. **GPT-4V** - scene graphs & audio transcript
3. Multimodal event simulation model

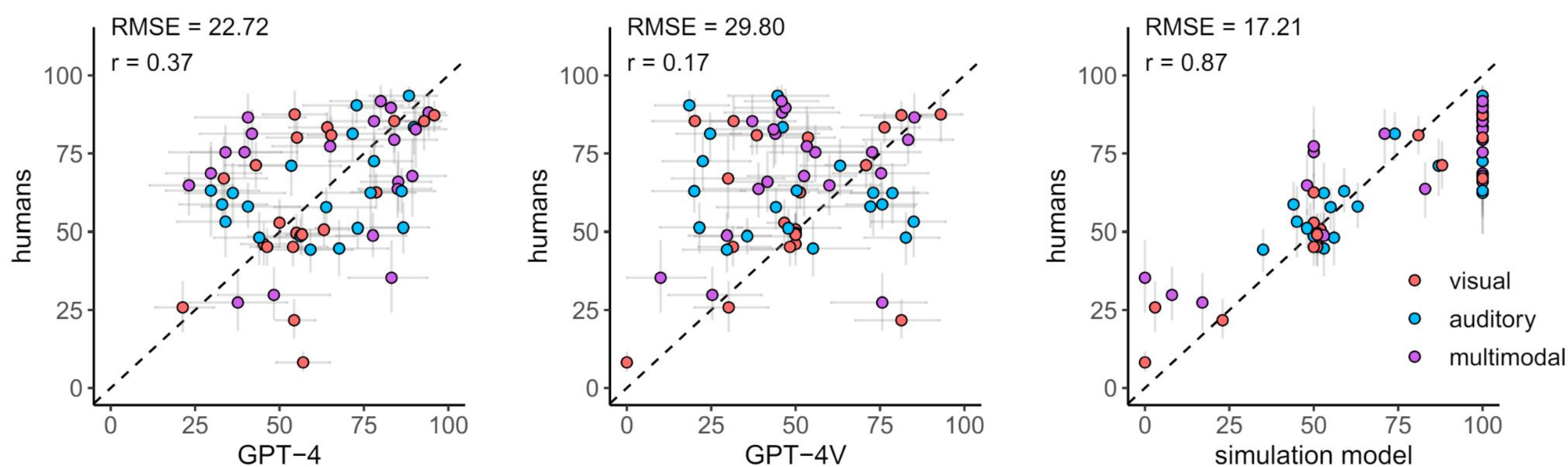
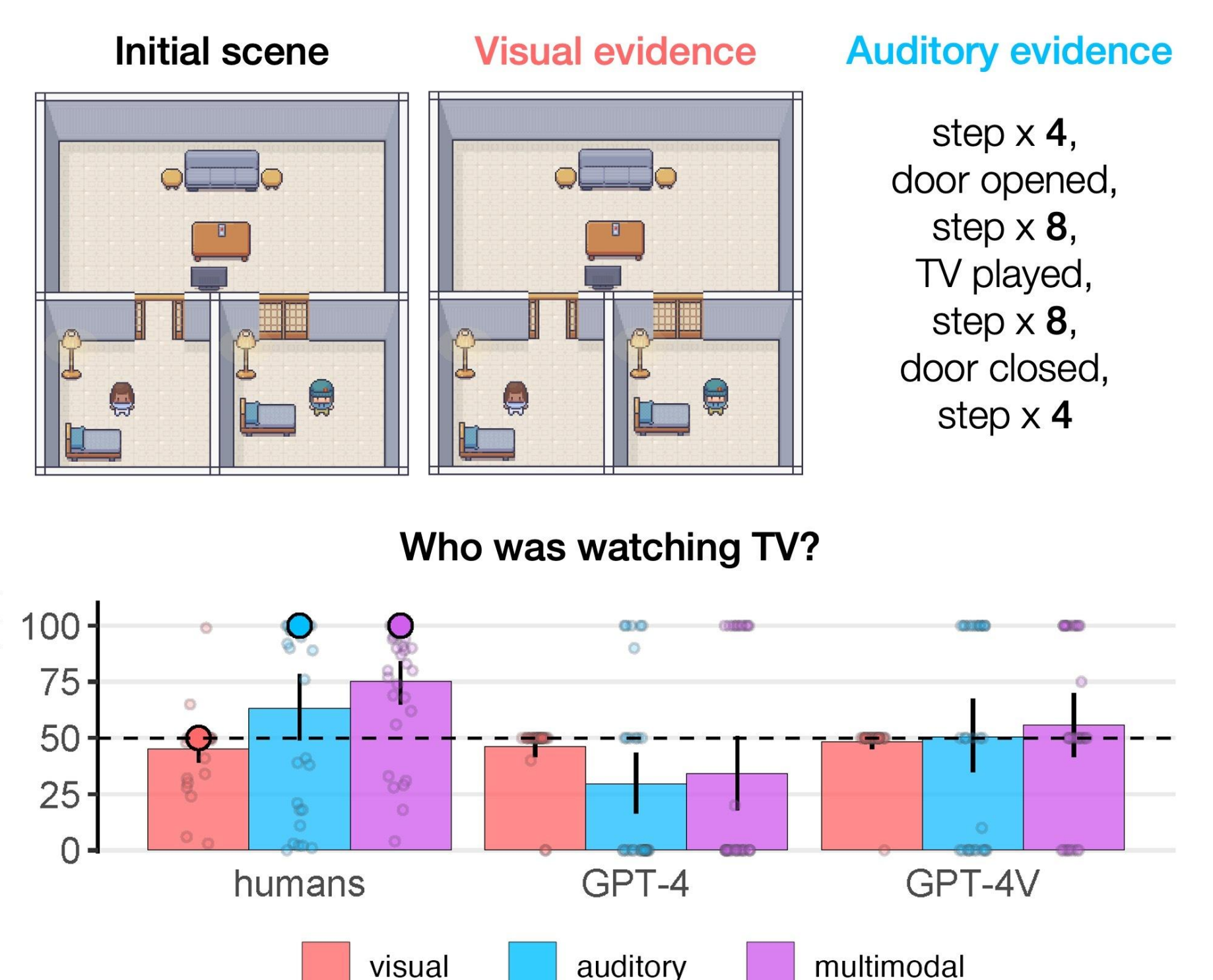


RESULTS

- ★ Humans and **GPT-4** more accurate with multimodal evidence, but not **GPT-4V**
- ★ Participants' visual and auditory accuracy can both predict multimodal accuracy



Example trial



- ★ *Simulation model captures indiv. human responses*
- ★ **GPT-4** and **GPT-4V** do not match human responses

TAKEAWAYS

- ★ People use evidence from multiple senses to reconstruct others' behavior
- ★ This remains a challenge for **GPT-4** and **GPT-4V**. Our simulation model captures humans

