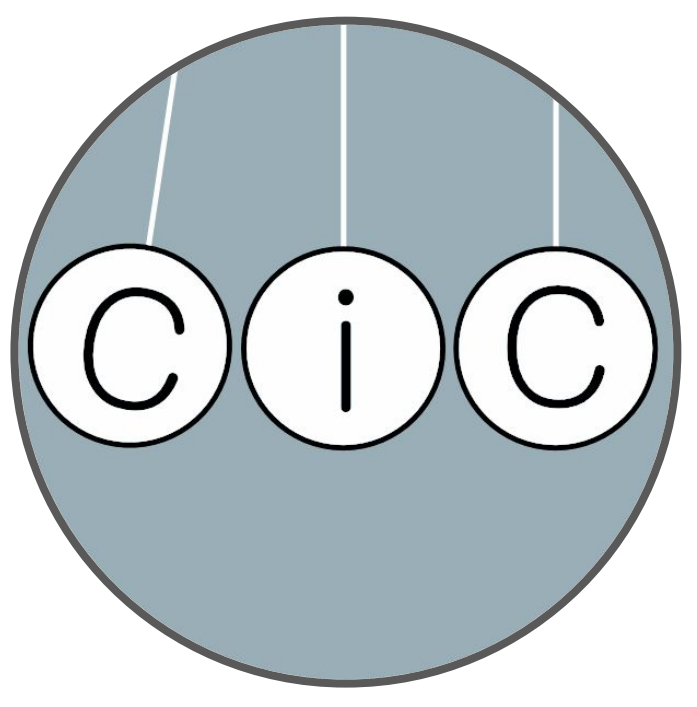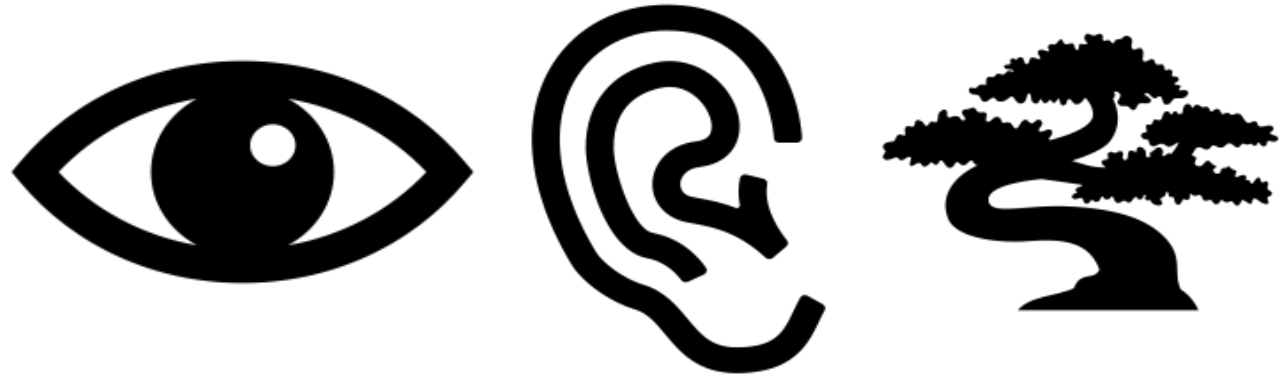# Whodunnit? Inferring what happened from multimodal evidence

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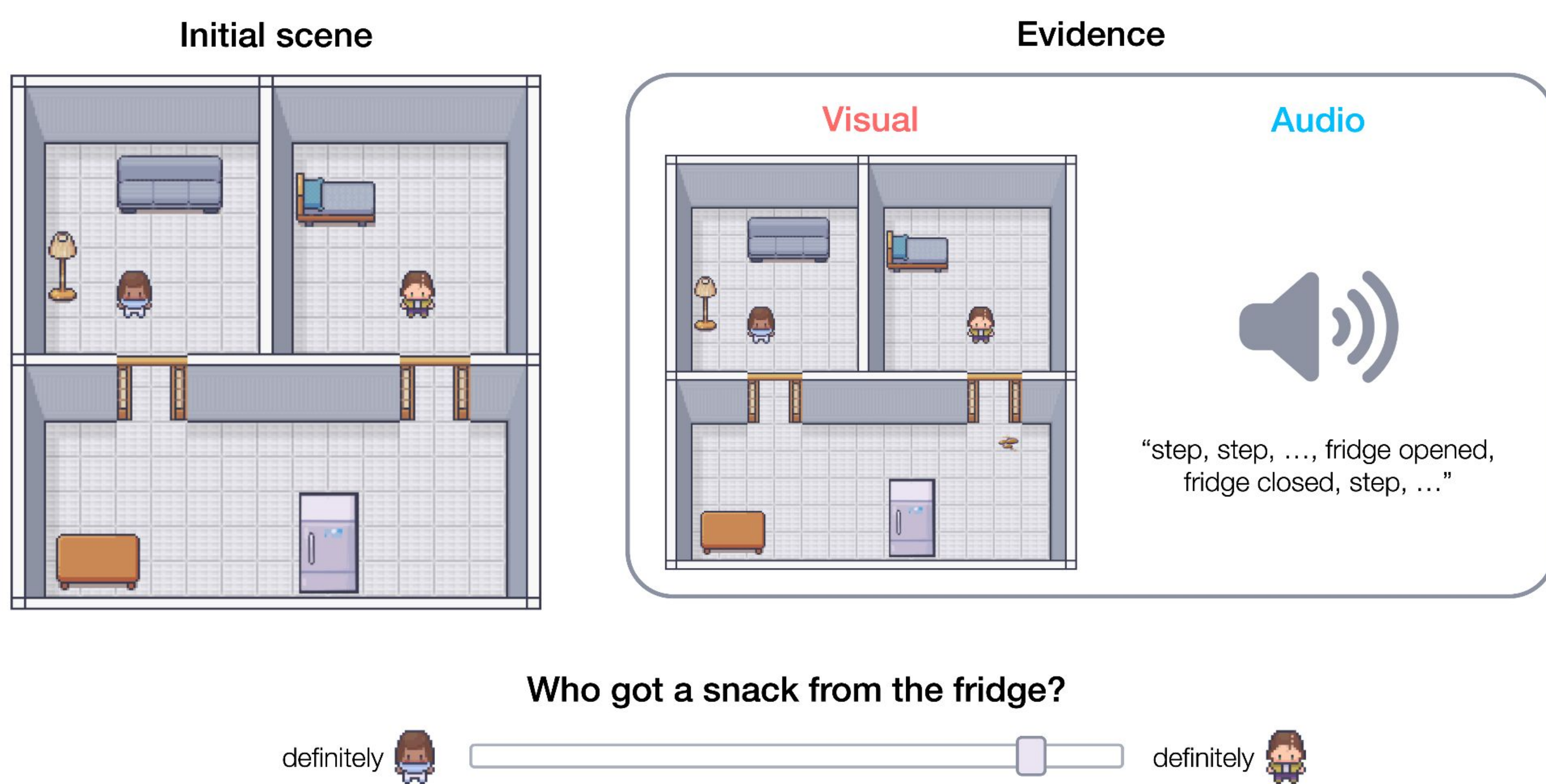
Mental models of how agents plan actions allow us reconstruct others' past behavior.

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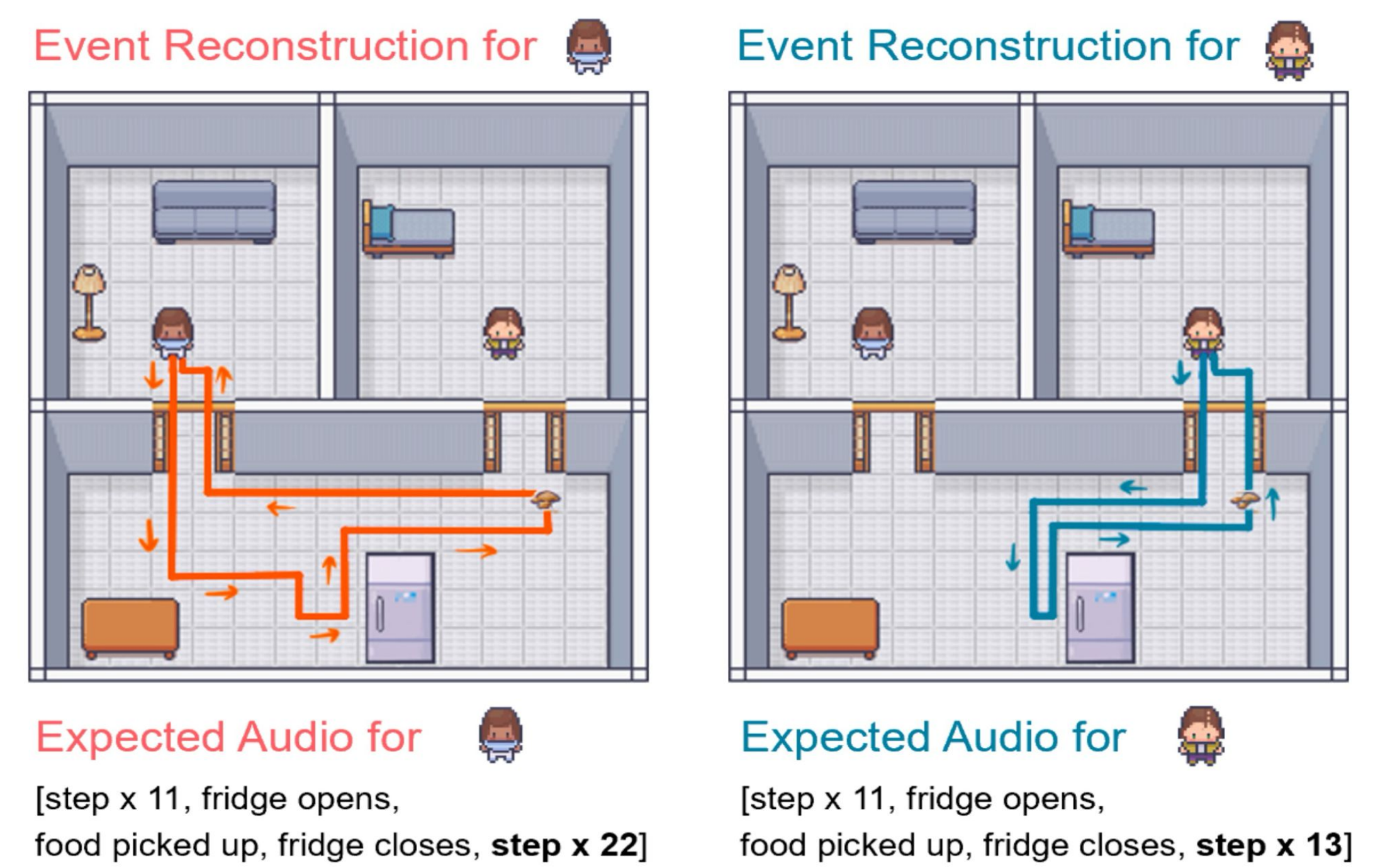
EXPERIMENT

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

Initial scene

Evidence

Visual   **Audio**

"step, step, ..., fridge opened, fridge closed, step, ..."

Who got a snack from the fridge?

definitely   definitely
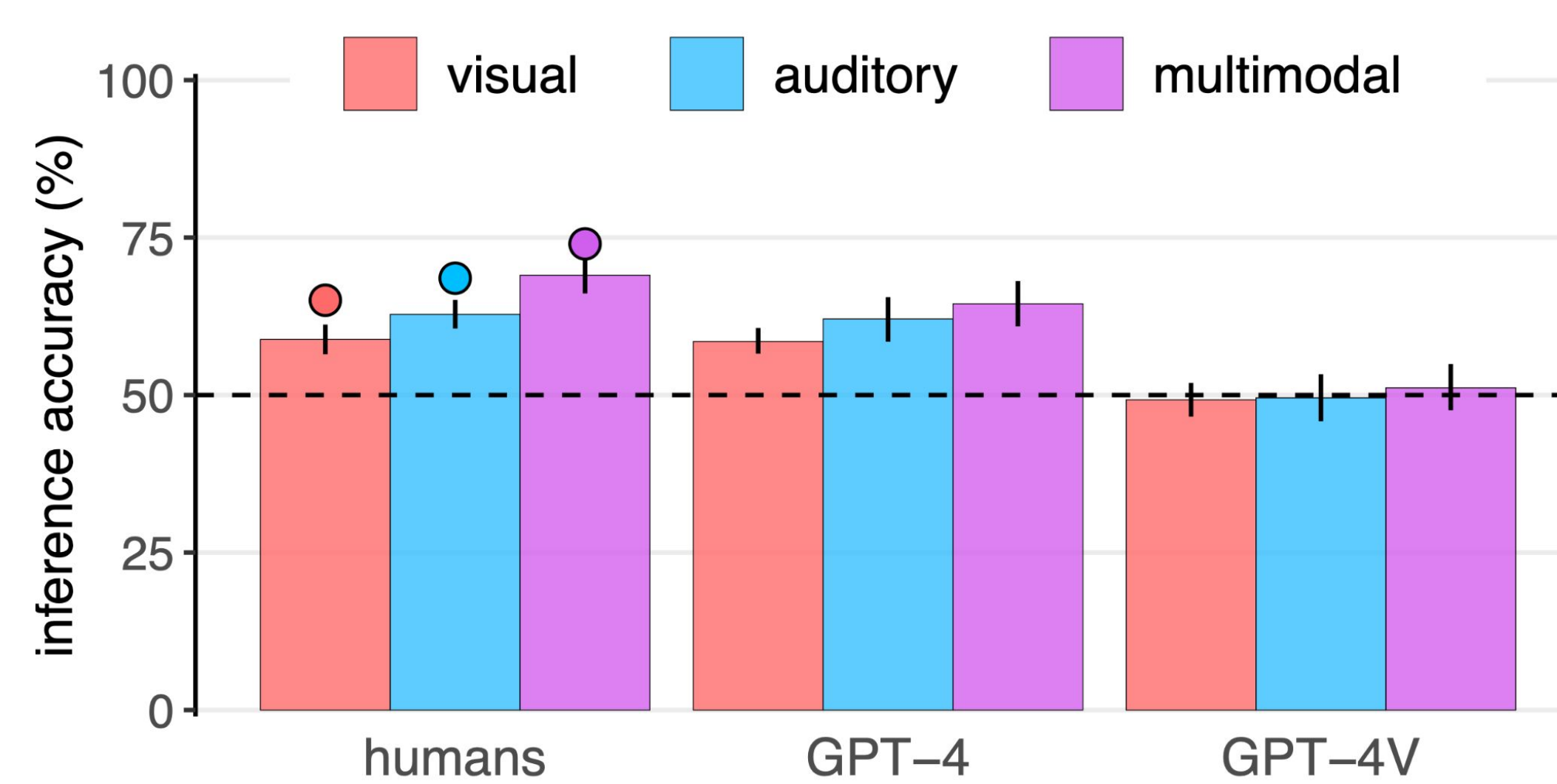
MODELS

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

Event Reconstruction for   **Event Reconstruction for**

Expected Audio for
[step x 11, fridge opens, food picked up, fridge closes, **step x 22**]

Expected Audio for
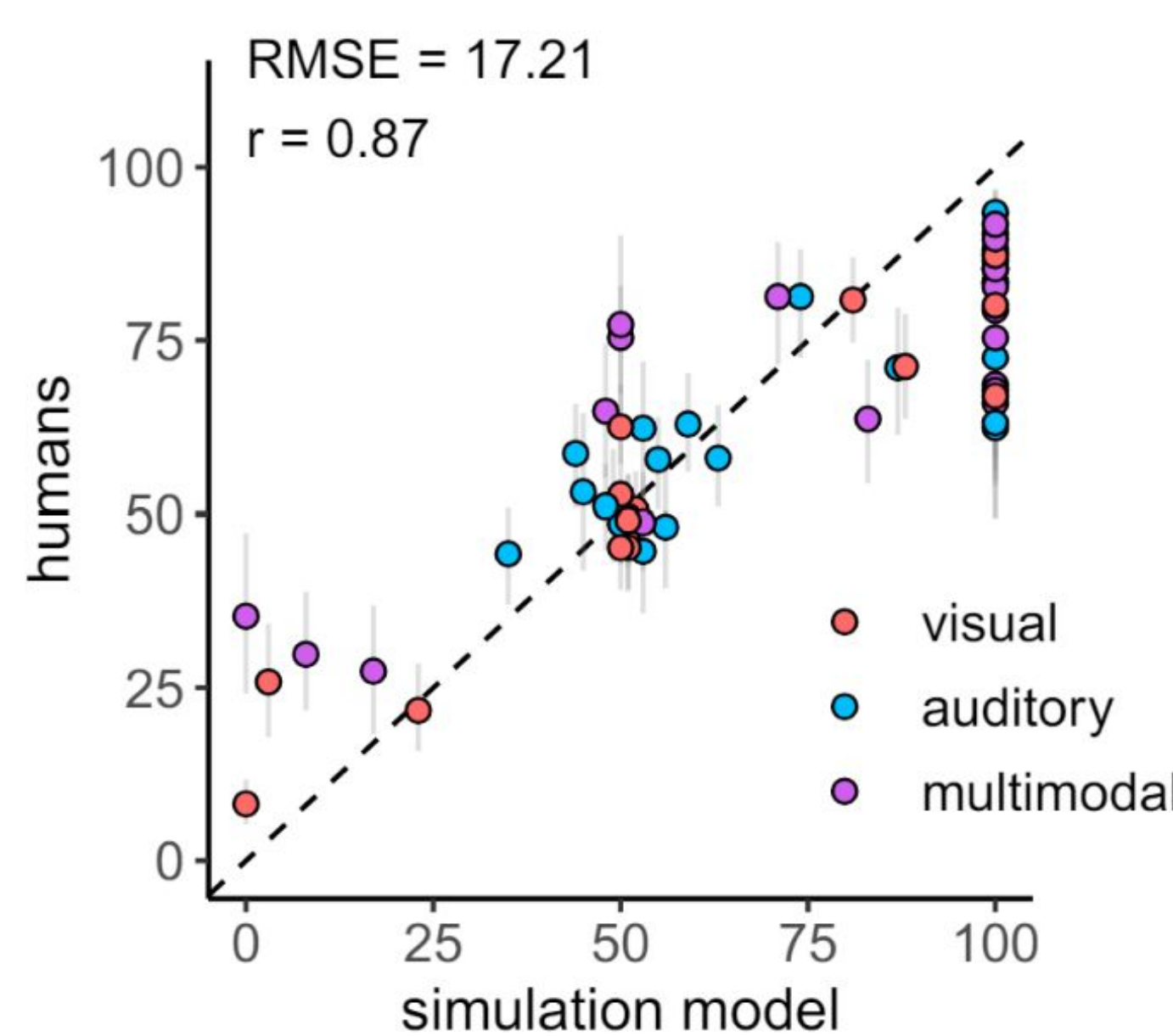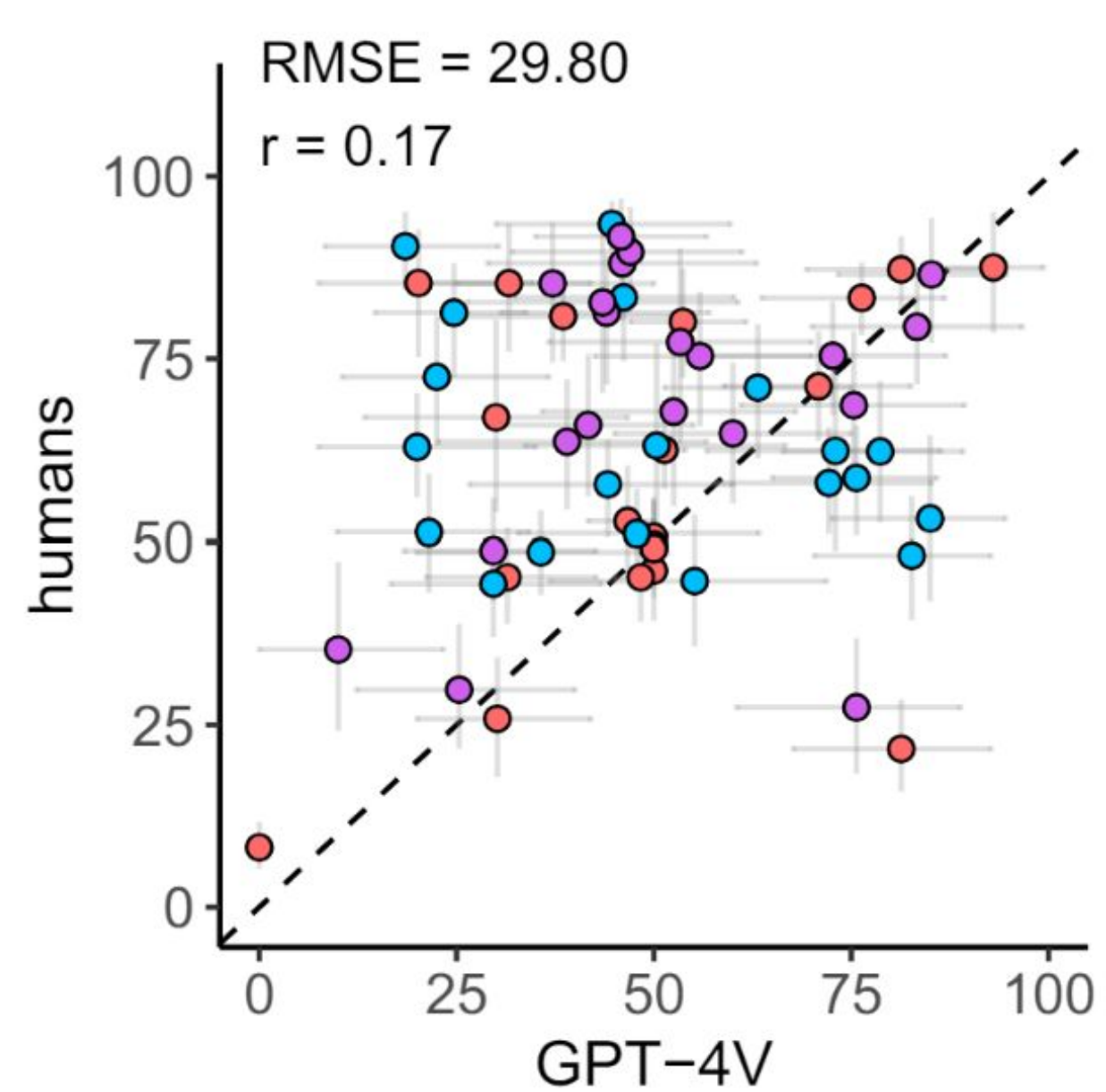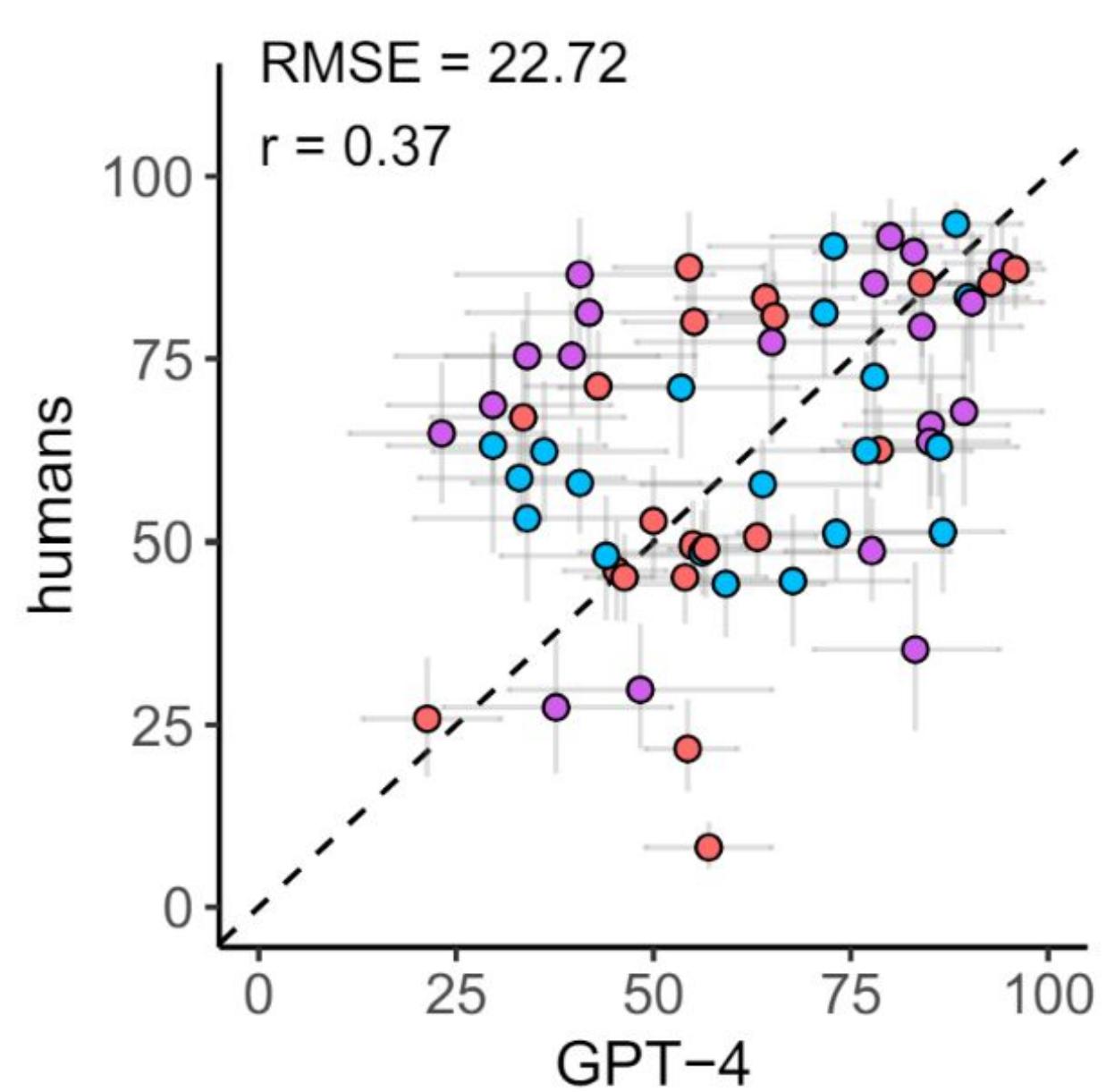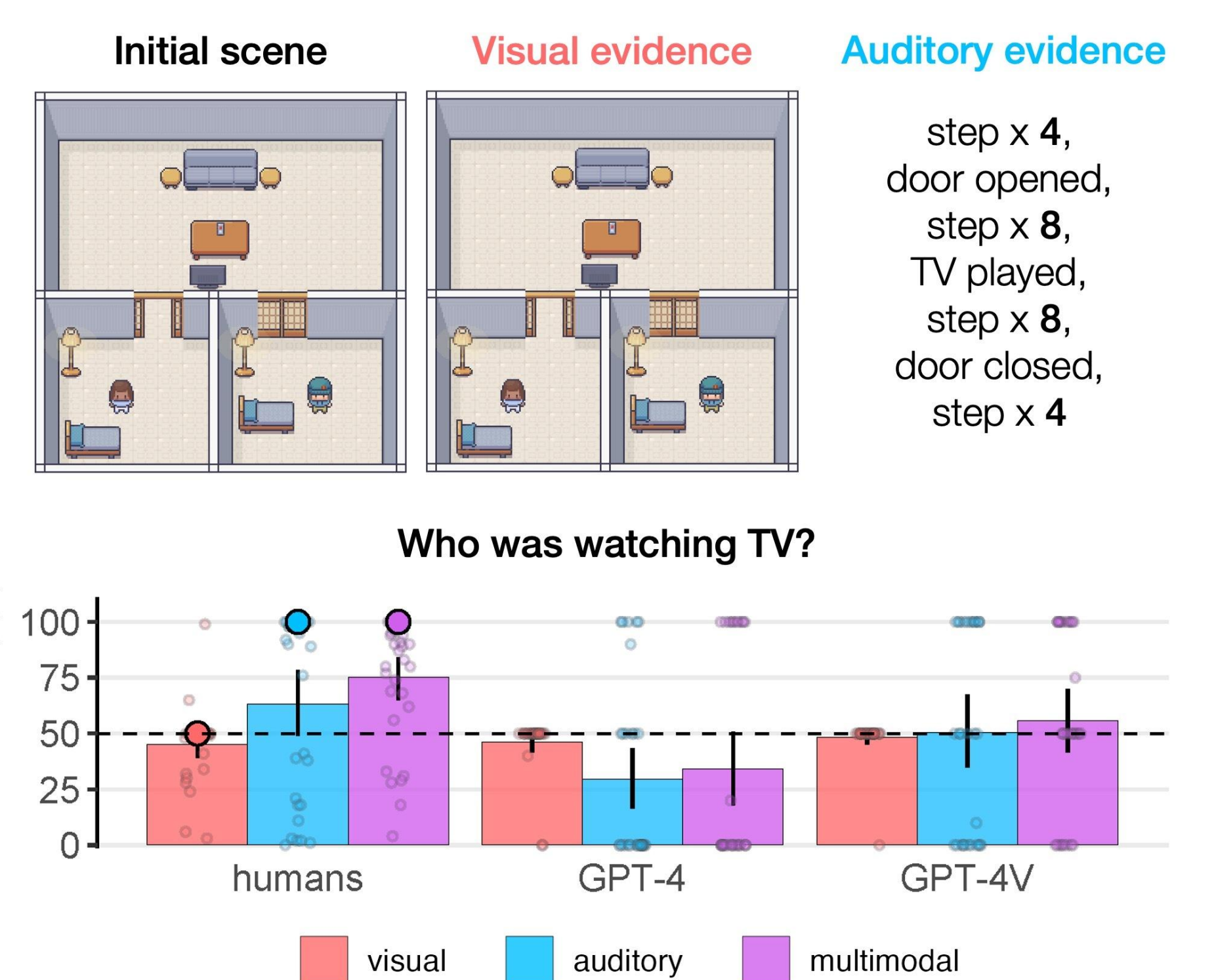[step x 11, fridge opens, food picked up, fridge closes, **step x 13**]

RESULTS

★ Humans and **GPT-4** more accurate with multimodal evidence, but not **GPT-4V**

★ Participants' visual and auditory accuracy can both predict multimodal accuracy

visual   auditory   multimodal

inference accuracy (%)

humans   GPT-4   GPT-4V

RMSE = 22.72
r = 0.37

humans

GPT-4

RMSE = 29.80
r = 0.17

humans

GPT-4V

RMSE = 17.21
r = 0.87

humans

simulation model

visual
auditory
multimodal

Example trial

Initial scene   **Visual evidence**   **Auditory evidence**

step x 4,
door opened,
step x 8,
TV played,
step x 8,
door closed,
step x 4

Who was watching TV?

inference accuracy (%)

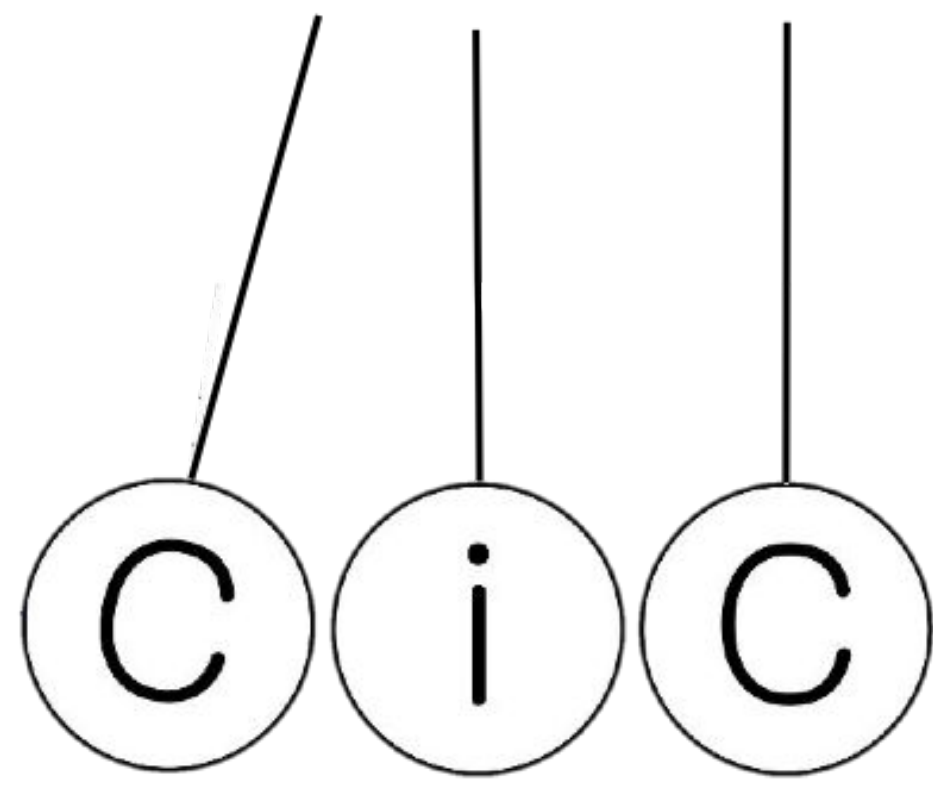humans   GPT-4   GPT-4V

visual   auditory   multimodal

★ *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

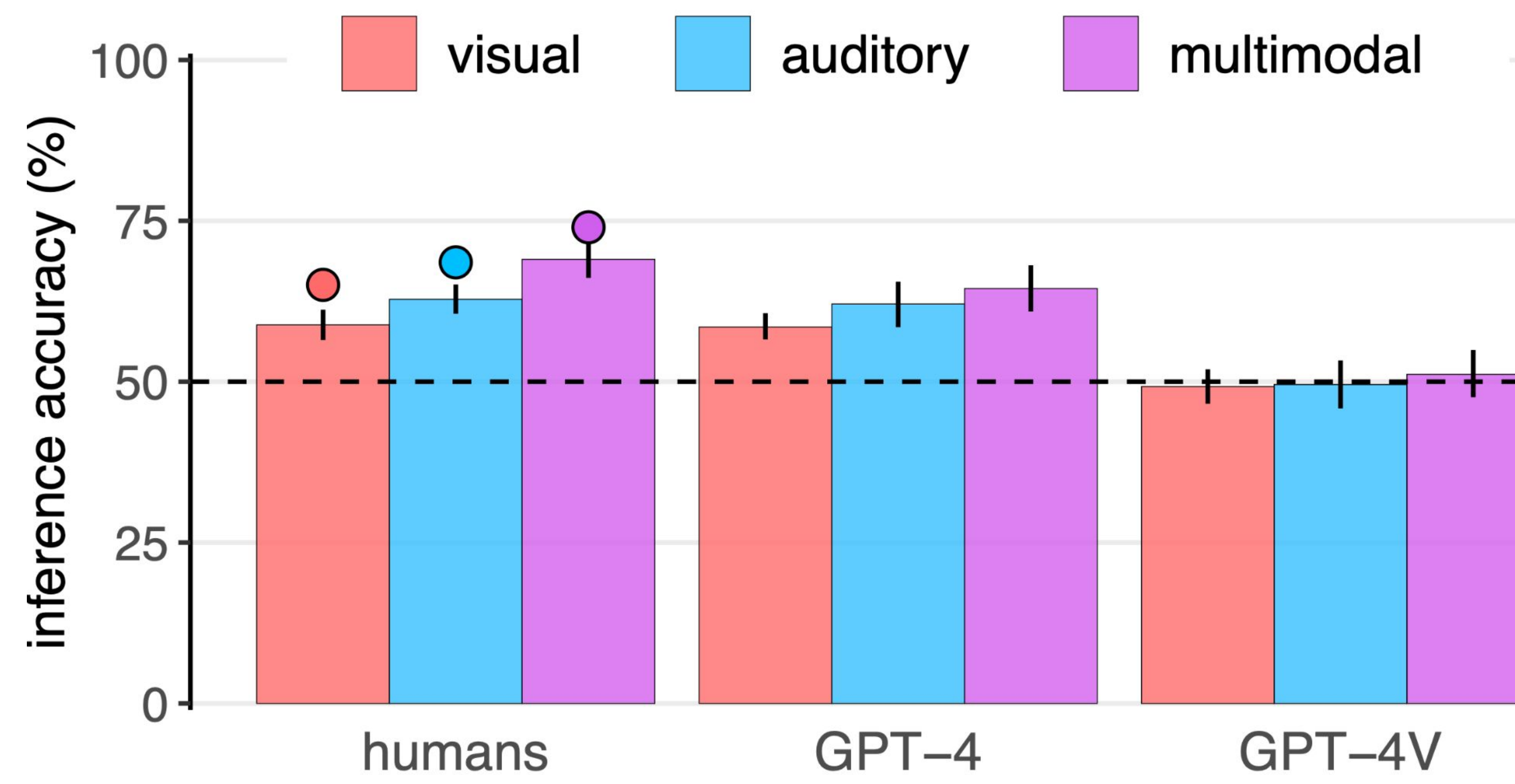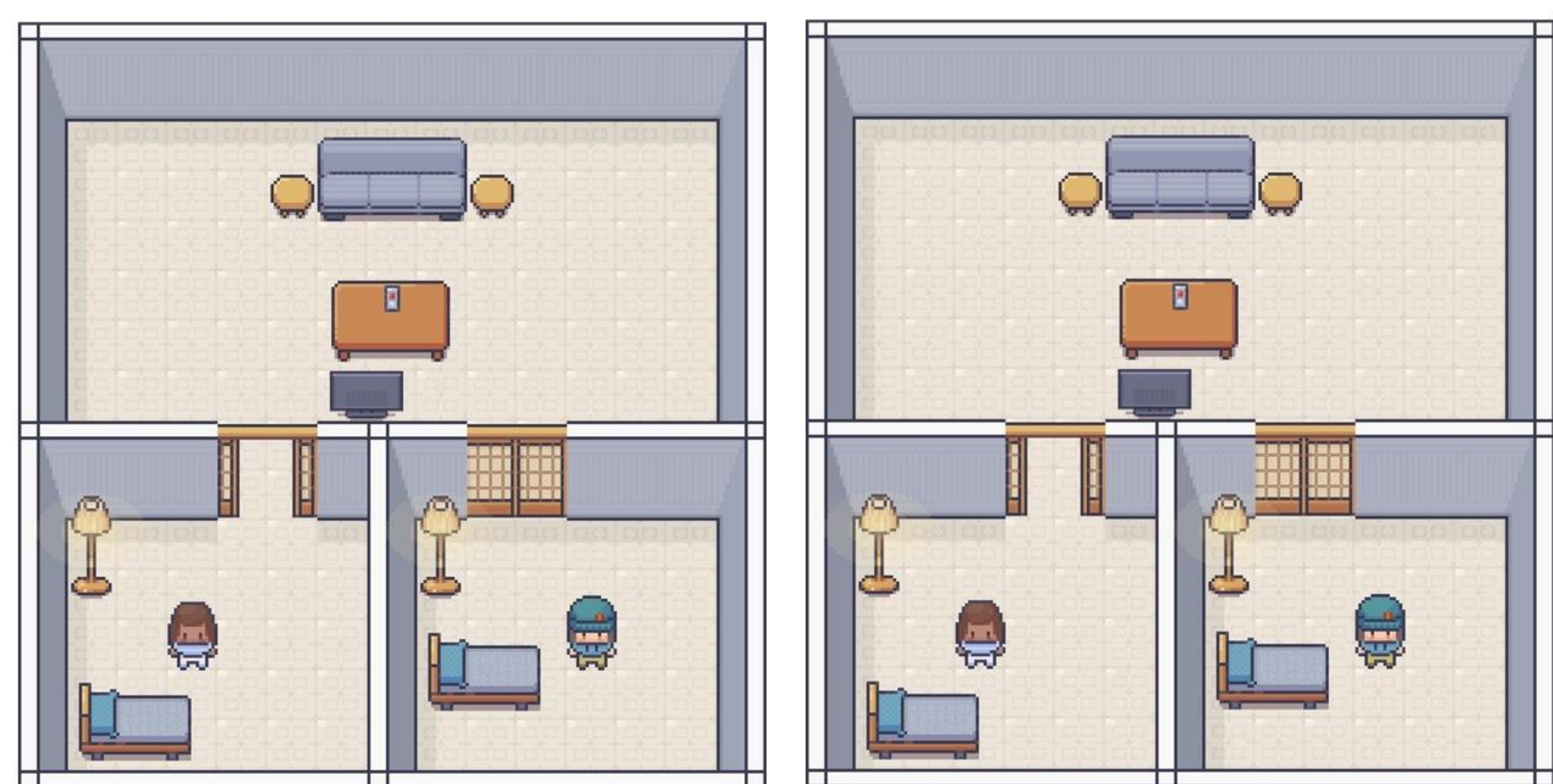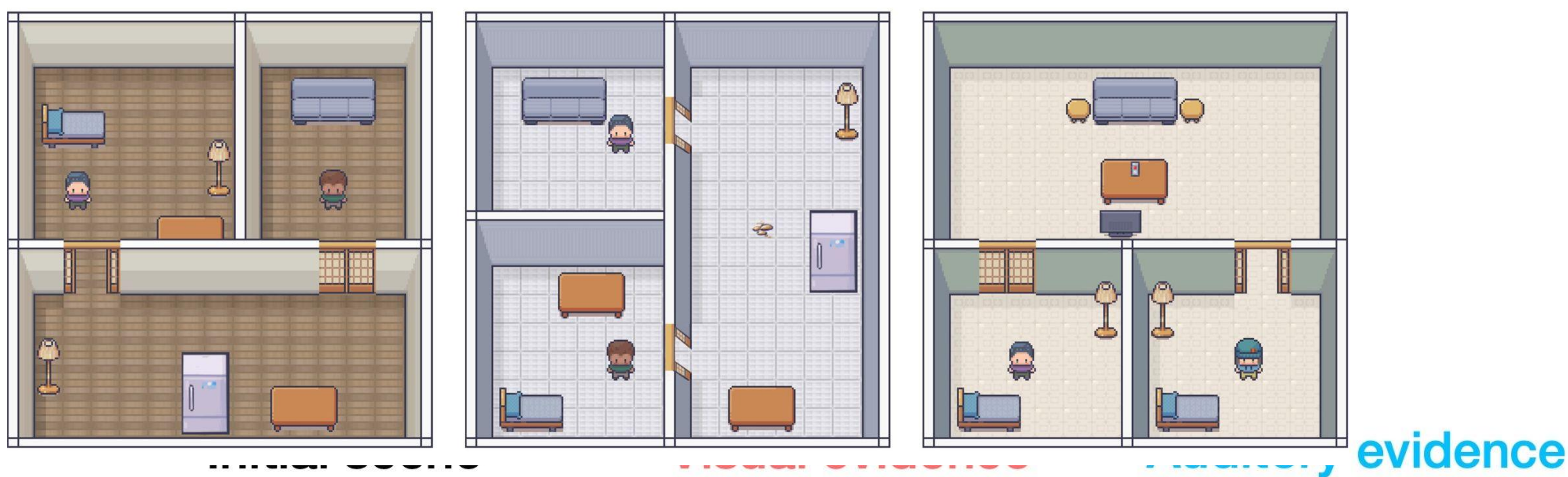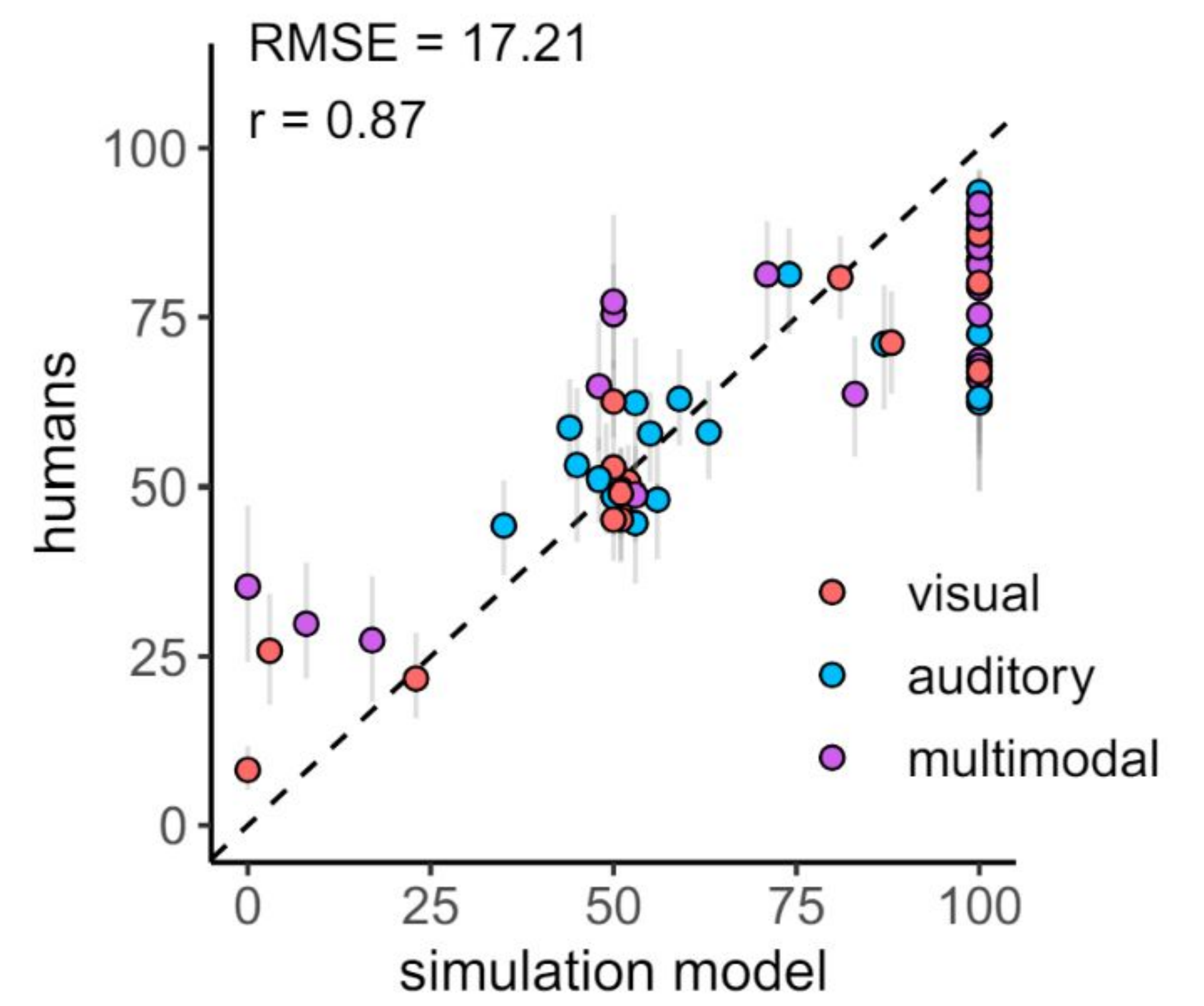★ This remains a challenge for **GPT-4** and **GPT-4V**. Our simulation model captures humans

Initial scene

Evidence

Visual   Audio

"step, step, …, fridge opened, fridge closed, step, …"

Who got a snack from the fridge?

definitely   definitely

Reconstruction for

Event Reconstruction for

RMSE = 22.72
r = 0.37

humans

GPT−4

RMSE = 29.80
r = 0.17

humans

GPT−4V

RMSE = 17.21
r = 0.87

humans

simulation model

- visual
- auditory
- multimodal

Initial scene   Visual evidence   Auditory evidence

step x **4**,
door opened,
step x **8**,
TV played,
step x **8**,
door closed,
step x **4**

inference accuracy (%)

- visual
- auditory
- multimodal

humans   GPT−4   GPT−4V

Who was watching TV?

inference accuracy (%)

humans   GPT-4   GPT-4V

- visual
- auditory
- multimodal