# Whodunnit? Inferring what happened from multimodal evidence

**Sarah A. Wu*[1], Erik Brockbank*[1], Hannah Cha[2], Jan-Philipp Fränken[1], Emily Jin[2], Zhuoyi Huang[2], Weiyu Liu[2], Ruohan Zhang[2], Jiajun Wu[2], Tobias Gerstenberg[1]**

[1]Department of Psychology, Stanford University, USA
[2]Department of Computer Science, Stanford University, USA

{sarahawu, erikbee}@stanford.edu

## Abstract

Humans are remarkably adept at inferring the causes of events in their environment; doing so often requires incorporating information from multiple sensory modalities. For instance, if a car slows down in front of us, inferences about *why* they did so are rapidly revised if we also hear sirens in the distance. Here, we investigate the ability to reconstruct others' actions and events from the past by integrating multimodal information. Participants were asked to infer which of two agents performed an action in a household setting given either visual evidence, auditory evidence, or both. We develop a computational model that makes inferences by generating multimodal simulations, and also evaluate our task on a large language model (`GPT-4`) and a large multimodal model (`GPT-4V`). We find that humans are relatively accurate overall and perform best when given multimodal evidence. `GPT-4` and `GPT-4V` performance comes close overall, but is very weakly correlated with participants across individual trials. Meanwhile, the simulation model captures the pattern of human responses well. Multimodal event reconstruction represents a challenge for current AI systems, and frameworks that draw on the cognitive processes underlying people's ability to reconstruct events offer a promising avenue forward.

**Keywords:** causal inference; mental simulation; multimodal integration; language models; multimodal models.

## Introduction

In Sir Arthur Conan Doyle's Sherlock Holmes story, "The Adventure of Silver Blaze", Sherlock is presented with the case of a murder in the Dartmoor horse stables (Doyle, 1894). He ultimately identifies the culprit as somebody close to the family based on the fact that the dog who slept in the stables did *not* bark during the murder ("the curious incident of the dog in the night-time"). Like Sherlock, we often make sense of the world around us by looking at information from multiple sense modalities. Imagine seeing crumbs in the hallway outside your roommate's bedroom; you might at first suspect that they had taken some food into their room. However, if this were accompanied by the sound of their dog chewing on the other side of the door, the most likely cause of the crumbs in the hallway changes suddenly. People have a remarkable ability to draw on evidence from different senses to make causal inferences about behaviors and events from the past. The advent of large language models (LLMs) and large multimodal models (LMMs) raises exciting questions about the potential of AI systems for multisensory reasoning as well. While these systems have shown impressive capacities for instruction following and knowledge retrieval (e.g. Bitton et al., 2023; Li et al., 2023; Z. Yang et al., 2023), little work has studied whether they possess deep causal understanding of their inputs. In this paper, we study how people integrate visual and auditory information to infer others' actions and events, and benchmark this ability in a state-of-the-art LLM (`GPT-4`) and LMM (`GPT-4V`). We compare human and large model performance to a Bayesian model that relies on multimodal simulations to reconstruct past events.

## Multimodal inference in humans

We first review findings related to people's ability to understand actions and reconstruct events from visual evidence, then highlight more recent results with multimodal evidence.

**Action understanding and event reconstruction** How people make sense of others' actions has long been a central question in psychology (Heider, 1958; Ross & Nisbett, 1991). Our *theory of mind* (ToM), the ability to interpret others' behavior as resulting from mental states such as beliefs, goals, and desires (Dennett, 1989; Gopnik & Meltzoff, 1997; Premack & Woodruff, 1978; Wellman, 2014), lies at the heart of everyday action understanding (Malle, 2004; Woodward, 1998). Prior work shows that people can work backwards from observed actions to hidden mental states by modeling others as rational planners (Baker et al., 2009, 2017; Jara-Ettinger et al., 2016, 2020).

However, these results typically rely on agents' present actions as evidence for their underlying goals or beliefs. Making Sherlock-style inferences about who, why, or how someone did something in the past using only clues they left behind requires an additional understanding of how behavior leaves detectable *traces* in the environment. Adults and children as young as 12 months old make inferences about agents from the appearance of the environments they were in or objects they interacted with (Gosling et al., 2002; Jacobs et al., 2021; Jara-Ettinger & Schachner, 2024; Lopez-Brau et al., 2022; Newman et al., 2010; Pelz et al., 2020). This ability to reconstruct events after the fact from "behavioral residues", combined with ToM, enables people to draw inferences about actions and their actors even without seeing what originally happened (Lopez-Brau et al., 2022).

**Integrating multiple senses** The ability to reconstruct events from visual clues speaks to the richness of human reasoning about the behavior of those around us. One limitation



In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* 


Figure 1: **Experiment interface.** A sample trial in which participants are presented with visual and auditory evidence. Note that text was never shown for the auditory evidence; an audio clip played out loud instead. Participants first saw the initial scene, clicked a button to reveal the evidence, and then responded to the inference question. Depending on the condition, the visual or auditory evidence was occluded. For `GPT-4`, the prompt contained scene graph representations of the scenes as text input, instead of the images.

of existing work is capturing the complexity of such inferences from more than merely visual evidence. Integrating information from multiple senses allows people to achieve a more complete causal understanding of their surroundings (Jin et al., 2024). Sometimes, like Sherlock's adventures show, multimodal reasoning is even necessary. However, this poses a challenge in its own right: how do people extract and combine relevant information from these multiple sources to form a single cohesive narrative?

Early research on cue integration found that basic perception involves rapidly incorporating information from distinct sensory inputs (Ernst & Banks, 2002; Körding et al., 2007). In fact, people use multiple modalities to draw complex inferences about the structure of the environment and agents' behavior within it (Agrawal & Schachner, 2023; Gerstenberg et al., 2021; Schachner & Kim, 2018; Siegel et al., 2021). However, young children struggle with abstract multimodal inferences (Agrawal & Schachner, 2023; Gori et al., 2008; Outa et al., 2022) and models of adults have been limited to reasoning about physical events (e.g., Gerstenberg et al., 2021). Causal event reconstruction with multimodal evidence in situations involving agents has yet to be studied.

## Multimodal inference in AI systems

Solving mysteries is no easy feat – scaling up structured models of event reconstruction (e.g., Lopez-Brau et al., 2022) and multimodal inference (e.g., Gerstenberg et al., 2021) to realistic environments poses a computational challenge for human reasoners. The growing competence of LLMs and LMMs on a wide range of tasks previously considered out of reach for AI has sparked active discussion about the general reasoning capacities of these models (e.g. Bubeck et al., 2023). `GPT-4V` and other LMMs are proficient at image captioning, visual question answering, and broad knowledge retrieval (Bitton et al., 2023; Li et al., 2023; OpenAI, 2023; Z. Yang et al., 2023). But they still lag behind humans in benchmarks involving ToM reasoning and contextual knowledge (Gandhi et al., 2024; Jin et al., 2024; Li et al., 2023), and have not been evaluated on any task involving causal reconstruction of others' actions and past events.

# Experiment

The current work seeks to fill a gap in prior research on multimodal reasoning in humans and AI systems. We evaluate the ability to reconstruct the cause of real-world behaviors using visual and auditory evidence. To do so, we designed a task where subjects were presented with evidence of agents engaging in everyday actions like fixing a snack and watching TV. Subjects had to evaluate which of two agents was most likely to have performed the action given the available evidence. We designed three versions of each scenario – one in which only visual evidence was available, one in which there was only auditory evidence, and one in which both visual and auditory evidence were available. By comparing judgments across these different modalities, we can better understand how humans and AI systems are able to integrate multiple sources of information to reconstruct what happened. We tested our scenarios on humans, `GPT-4`, and `GPT-4V`, and also present a simulation-based model of multimodal event reconstruction that attempts to computationally capture people'sreasoning.

## Participants

The experiment was preregistered[1] and posted as a task on Prolific.[2] 90 participants (*age*: M = 39, SD = 12; *gender*: 44

[1] https://osf.io/fzxre/?view_only=8463d5f64aa5413a8b0c8befe27db627

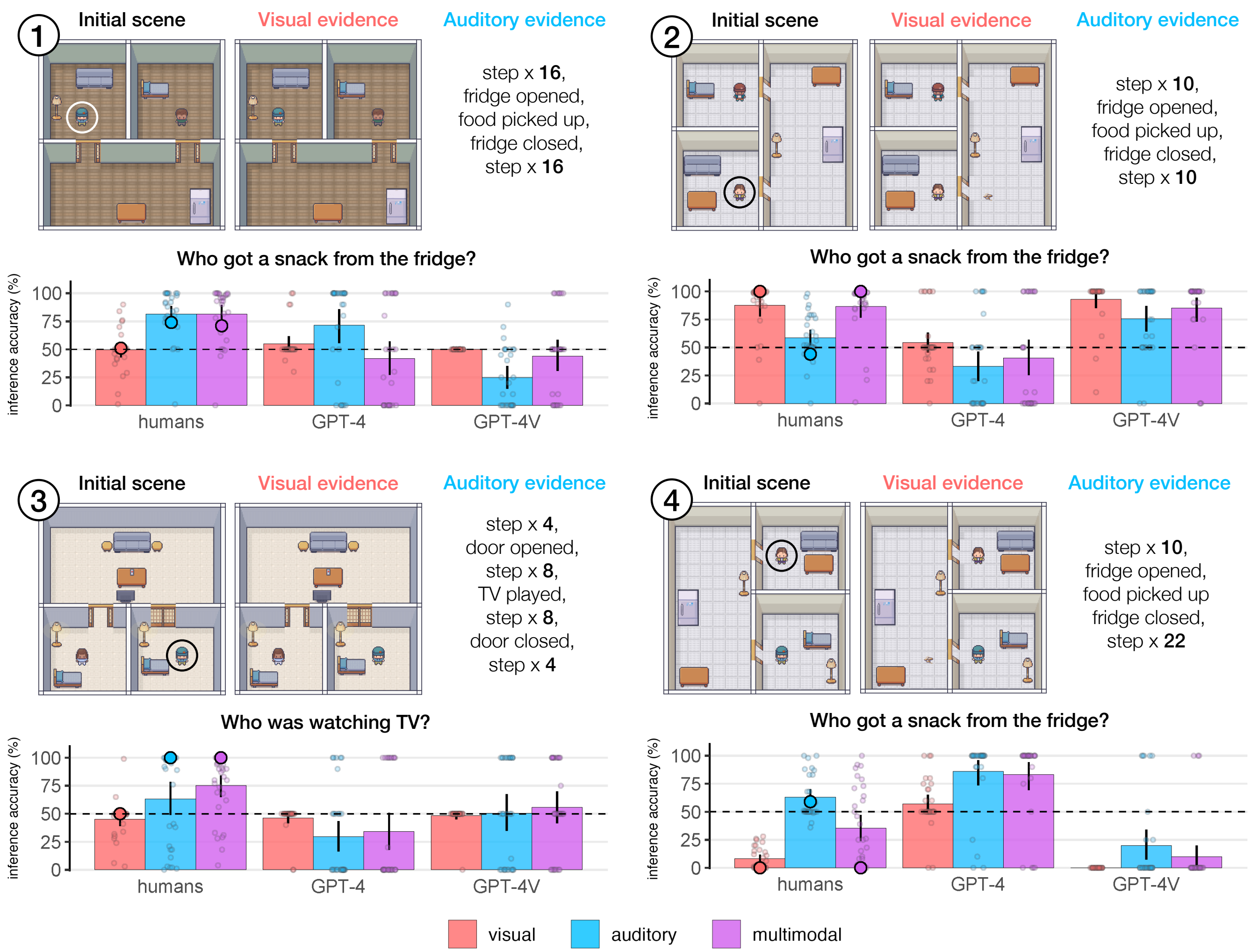


Figure 2: **Inference accuracy on select scenarios.** Each diagram illustrates the initial scene (correct agent is circled here, but was not shown in the experiment), final scene, and transcription of the audio. For participants, an audio clip was simply played out loud (no text was displayed). Bars show mean accuracy, small points show individual responses, large point shows simulation model predictions, and error bars are bootstrapped 95% confidence intervals.

female, 40 male, 3 non-binary, 3 undisclosed) were recruited and compensated $12/hour. Each participant was shown only one version (*visual*, *auditory*, or *multimodal*) for each scenario. All participants received a roughly equal distribution of the three versions across the entire experiment; we obtained $n = 30$ responses for each version of each scenario.

## Procedure

Participants were first guided through instructions with example animations and audio clips to familiarize them with the stimuli. They were then required to answer three comprehension questions correctly before proceeding to the main portion of the experiment.

Participants were presented 20 scenarios in a randomized order. In each scenario, they were shown an image of an initial scene and clicked a button to reveal an image of the final scene (visual evidence), an audio clip that played in the browser (auditory evidence), or both (multimodal; see Figure 1). On audio and multimodal trials, participants were required to listen to the entire clip at least once before responding, and could replay it as many times as they liked. Participants were asked which agent performed the action (e.g., "Who got a snack from the fridge?" or "Who was watching TV?") and answered on a slider with endpoints labeled "definitely [agent 1]" and "definitely [agent 2]". The experiment took an average of 14 minutes ($SD = 4$) to complete.

## Design

In 13 of the 20 scenarios, an agent walked into the kitchen and got a snack from the fridge then walked back to their starting location, sometimes leaving behind crumbs or open/closed doors as visual evidence. In the other seven scenarios, an

[2]Code for all models, analyses, and experiments can be found at: https://github.com/cicl-stanford/whodunnit_multimodal_inference.

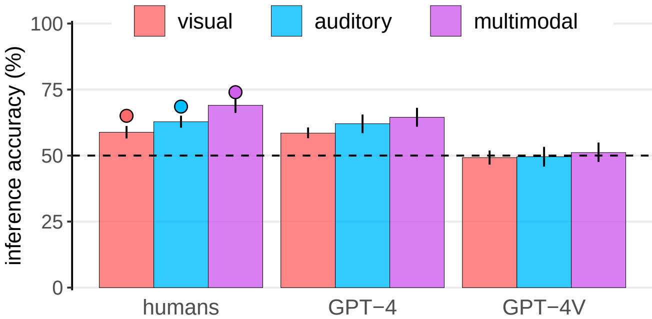


Figure 3: **Overall inference accuracy.** Accuracy across all scenarios for humans, `GPT-4`, and `GPT-4V` ($n = 30$ per condition). Bars show mean accuracy, error bars are bootstrapped 95% confidence intervals, points are simulation model accuracy, and the dashed line represents maximum uncertainty.

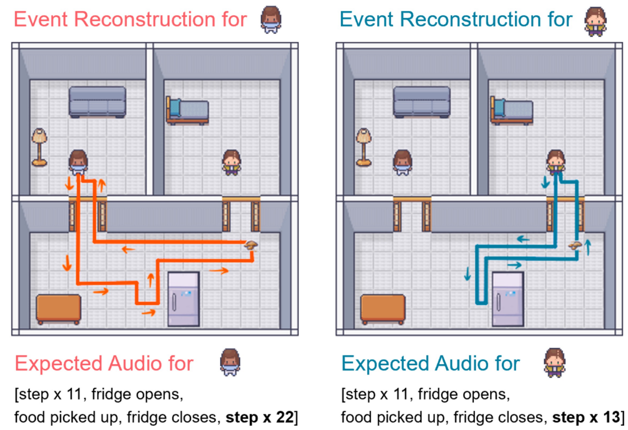


Figure 4: **Multimodal event simulation.** The simulation model generates possible paths each agent could have taken to produce the given visual evidence (crumbs near the right door) and auditory evidence (sequence of sounds heard).

agent walked into the living room and watched TV for a short period, sometimes moving the TV remote or leaving behind open/closed doors before walking back. The audio clip accompanying each scenario included sounds associated with all actions except when the agent dropped crumbs. We reserved the crumbs as visual evidence only because in real life, visual changes can often be silent, and this allowed us to better differentiate the three versions of each trial. Across scenarios we manipulated the agents' starting locations and the locations of the fridge, crumbs, TV, and remote.

Together, the visual and auditory evidence afford different levels of diagnosticity about which agent performed the action. In some scenarios, the visual information alone was uninformative, while the audio information was revealing. For example, in Figure 2 scenario 1, there is no visual evidence, but the audio reveals a very long path to and from the fridge, which is suggestive of the agent initially father away on the left. Other scenarios have the opposite pattern. In scenario 2, the audio is equally plausible for both agents, but the crumbs implicate the agent on the bottom. We also designed a few scenarios in which the visual and auditory evidence were ambiguous or seemingly conflicting. For example, in scenario 4, crumbs have been left closer to the agent on the bottom, but the audio reveals a long path back from the fridge, making it possible that either agent meandered back to their room.

## Models

We compared participants' performance to two models. The first is a simulation-based model that combines prior work on visual event reconstruction (Lopez-Brau et al., 2022) and multimodal inference (Gerstenberg et al., 2021). It attempts to capture the cognitive processes underlying human multimodal inference. Alongside this, we also compare responses to `GPT-4` and `GPT-4V`. These models have been successful at solving a range of visual and social reasoning problems, allowing us to explore whether their more domain-general capacities support multimodal reasoning in the current context.

### Multimodal event simulation model

The event simulation model is given a scene graph representation of the initial state in each trial, as well as "evidence" in the form of the final state (visual), a transcript of sounds (auditory), or both (multimodal). For an agent $a \in \{\text{agent 1}, \text{agent 2}\}$, the probability of the agent performing the action given the evidence $x \in \{\text{visual}, \text{auditory}, \text{multimodal}\}$ can be written as:

$$p(a|x) = \frac{p(x|a)p(a)}{p(x)}$$

The prior probability of each agent performing any given action $p(a)$ is assumed to be uniform (the identity of the agents was not a useful predictor in the task), as was the probability of receiving evidence in any particular modality $p(x)$. Thus, determining which of the two agents was most likely to have caused the evidence is a function of the relative likelihoods $p(x|a = \text{agent 1})$ and $p(x|a = \text{agent 2})$.

The model estimates these likelihoods using simulations of the agents' behavior to produce a distribution of visual and auditory evidence for each agent (Figure 4).[3] The model then estimates the probability of the evidence observed (or heard) under the simulated evidence distribution for each agent using rejection sampling. After determining the probability of each agent completing the event in a given trial $p(a|x)$, the model produces a slider value $\hat{y}$ to match participants' responses by normalizing these probabilities: $\hat{y} = \frac{p(a|x)}{\sum_i p(a_i|x)}$.

### Large language and multimodal models

The same scenarios shown to humans were also tested on `GPT-4` and `GPT-4V`. For `GPT-4`, we provided scene graph representations of each room, including positions and dimen-

[3] Code and implementation details for the model can be found in the public github repository linked above.

sions of all agents, furniture, and objects (see Figure 1). For `GPT-4V`, we used the same images shown to participants. The auditory evidence was presented as a transcribed list of the sounds heard at each time step (e.g., "step", "door opened"). In each trial, models were provided the initial scene (as either a scene graph or an image) and then either visual evidence (scene graph or image of the final scene), auditory evidence (transcribed audio), or both. The task prompt encouraged models to analyze state changes, positions of objects, and any new elements in the visual clues, and to focus on the number of steps and different sounds in the audio clues. Both models were prompted with a temperature of 0.7 and zero-shot chain-of-thought prompt optimization (i.e. "Take a deep breath. Let's think step-by-step."; Kojima et al., 2023; C. Yang et al., 2024). We queried each model $n = 30$ times. The models were prompted to produce a continuous numerical response on the same scale as the slider shown to participants.

## Results

For each trial, the correct answer was coded as 0 if the evidence was generated by the agent presented on the left end of the slider (see Figure 1), and 100 if it was the agent on the right. We computed accuracy as the absolute difference between the correct answer and participants' slider responses. Note that in some trials, the evidence could have been generated by either agent. In those cases, we would not expect participants or models to achieve full accuracy, especially in the single modality conditions.

### Accuracy across modalities

We hypothesized that participants' inferences would be more accurate in the multimodal condition for each scenario than with either modality alone. Figure 3 shows inference accuracy for participants, `GPT-4`, and `GPT-4V` in each condition across all scenarios as well as predictions of our multimodal event simulation model. As predicted, humans performed best when presented with multimodal evidence. Meanwhile, `GPT-4` exhibits a qualitatively similar pattern, while `GPT-4V` accuracy is similar across all modalities.

To quantify these comparisons, we fit Bayesian linear mixed effects models to predict accuracy values for human and GPT responses in each modality condition. The models include fixed intercepts and fixed effects of the modality (dummy coded as either 0 = "single" or 1 = "multimodal"). For participants, it also includes a random intercept for each participant. Table 1 shows that modality was a credible positive predictor for both humans and `GPT-4`, but not `GPT-4V`.[4]

Table 1: **Effects of modality on accuracy.** 'Intercept' and 'Modality' show the posterior means of each predictor along with 95% highest density intervals (HDIs) in brackets. The model was given by the formula `accuracy ~ 1 + modality + (1 | participant)` for humans and without the random intercept for `GPT-4` and `GPT-4V`. Modality was dummy coded as either 0 = "single" or 1 = "multimodal".

| Subject | Intercept | Modality |
|---|---|---|
| Humans | 60.75 [58.61, 62.96] | 8.38 [5.44, 11.36] |
| `GPT-4` | 60.29 [58.07, 62.52] | 4.17 [0.27, 7.97] |
| `GPT-4V` | 49.40 [47.08, 51.76] | 1.77 [−2.11, 5.82] |

Finally, to understand how participants integrated visual and auditory evidence, we measured correlation in mean accuracy across the three conditions. Accuracy was most correlated between the visual and multimodal conditions, ($r = 0.75$, $p < .001$), moderately correlated between auditory and multimodal ($r = 0.48$, $p = .04$), and not correlated between visual and auditory ($r = -0.04$, $p = .87$). We fit a Bayesian linear mixed effects model to predict multimodal accuracy using z-scored visual and auditory accuracy as fixed effects, and further compared the difference in the posteriors. We found a credible positive effect of both predictors (posterior on visual: 15.09 [10.67, 19.42], posterior on auditory: 9.83 [5.31, 14.30]), and there was a credibly larger effect of the visual predictor compared to the auditory predictor on multimodal accuracy (difference in posteriors: 5.26 [5.17, 5.36]).

### Accuracy across trials

The trials in the current experiment were designed to elicit inferences that varied substantially across modalities and when combining them; we compare human and model performance at the level of *individual trials* to isolate the accuracy of these inferences. Figure 2 shows results for a subset of scenarios that illustrate how visual and auditory evidence can be differentially diagnostic. In scenario 1, visual evidence alone was uninformative but participants became more accurate with the addition of auditory information. `GPT-4` and `GPT-4V` did not extract relevant information from the audio evidence particularly in the multimodal condition. In scenario 2, the audio evidence was uninformative, but visual evidence improved participants' accuracy in the visual and multimodal conditions. `GPT-4V` also performed well in those conditions, while `GPT-4` did not. In scenarios 3 and 4, the evidence was ambiguous or conflicting. Scenario 3 shows successful multimodal integration from participants, but not models. In contrast, scenario 4 illustrates a case of poor multimodal integration by participants (notably failing to incorporate audio evidence in the multimodal condition) while `GPT-4` was far more accurate.

To extend this comparison to the full range of trials, we calculate the correlation in trial accuracy between participants and `GPT-4` (Figure 3A), `GPT-4V` (3B), and our event simulation model (3C). Despite the similarities in *overall* accuracy across modalities, participant and `GPT-4` responses are only weakly correlated for individual trials ($r = 0.37$, $p = 0.003$). Human and `GPT-4V` accuracy exhibits almost no relationship ($r = 0.17$, $p = 0.19$). In contrast, the simulation model strongly captures participants' responses ($r = 0.87$, $p < 0.001$). Broken down by modality, the simulation model

[4] We adopt the convention of calling an effect *credible* if the 95% HDI of the estimated parameter in the Bayesian model excludes 0.

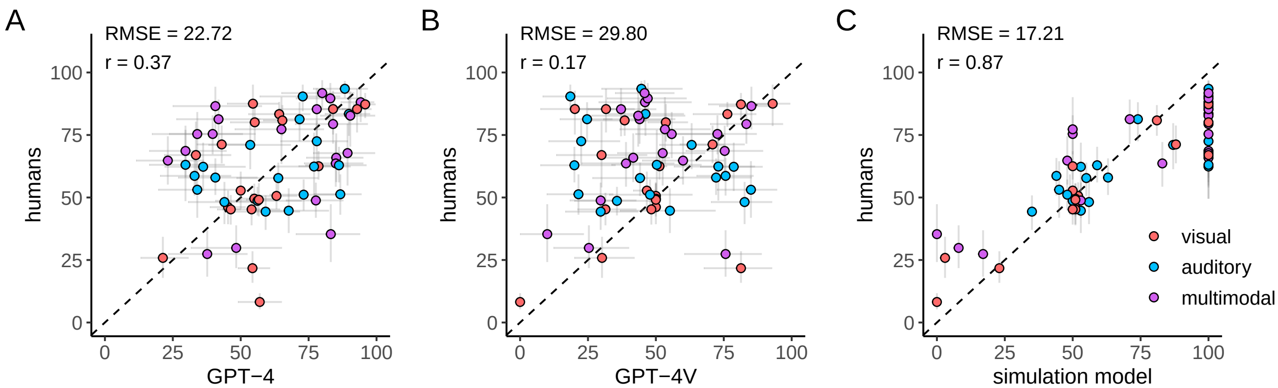


Figure 5: **(A) `GPT-4`, (B) `GPT-4V`, and (C) simulation model accuracy.** Inference accuracy for participants compared to `GPT-4` accuracy, `GPT-4V` accuracy, and simulation model predictions across all conditions.

was most strongly correlated with human accuracy on visual trials ($r = 0.96$), followed by multimodal ($r = 0.83$) and then audio ($r = 0.79$; all $ps < .05$).

## Discussion

In this paper, we investigate the ability to reconstruct events from naturalistic agent behaviors. We explore how people do this by drawing on multimodal evidence from vision and audio to acquire a rich causal understanding of the events that occurred. Prior work shows that people can use "behavioral traces" left behind in an environment to infer how an agent previously acted (e.g., Lopez-Brau et al., 2022). In physical settings like Plinko games, people can successfully integrate visual and auditory information to account for game outcomes (Gerstenberg et al., 2021; Schachner & Kim, 2018). However, such multimodal causal reasoning for more complex scenarios involving agent behavior has not yet been studied. Furthermore, the recent rise of large language models such as `GPT-4`, and large multimodal models such as `GPT-4V` equipped to process visual and textual input opens the door to exploring these same questions in modern AI systems.

### Multimodal event reconstruction

We find evidence that people incorporate multimodal reasoning in causal event reconstructions. Overall accuracy with multimodal evidence exceeded accuracy with either visual or auditory evidence alone. Our results suggest that multimodal reasoning in this context was more strongly affected by the visual evidence than by the auditory evidence. In fact, scenario 4 in Figure 2 illustrates a case where accuracy *decreased* with the addition of visual evidence (relative to audio alone), perhaps as a result of this overreliance on (sometimes misleading) visual cues in multimodal reasoning. Such a pattern might arise if the visual evidence was easier to process (compared to, for example, counting steps in the audio signal to determine the most likely agent). However, it is also possible that the visual evidence caused participants to consider a broader set of explanations or simulated behaviors than the auditory evidence, leading to greater uncertainty in their responses. In this vein, future work should consider the ways evidence across modalities constrains or supports simulation.

### A benchmark for AI

While `GPT-4`'s and `GPT-4V`'s overall accuracy was not far behind participants (Figure 3), the weak correlation across individual trials (Figures 5A and B) suggests no systematic relationship and likely different underlying mechanisms in how humans and `GPT-4` solve the inference task. In contrast, the multimodal event simulation model captured participants' responses across trials well (Figure 5C). This model combines *inverse planning*, in which observers assume those around them choose efficient paths to attain their goals, with an internal model that can perform *mental simulations* of how such actions can lead to different patterns of evidence, visual or auditory, in the environment (Gerstenberg et al., 2021; Lopez-Brau et al., 2022). Our results indicate that causal event reconstruction is still a challenging task for foundation models. However, the ability of our more structured simulation model to capture successful human reasoning in this task raises an opportunity for future work integrating such processes into large language and multimodal models.

## Conclusion

Most of us aren't detectives like Sherlock Holmes, but we all reason about the world around us by drawing on information from different senses. In this paper, we studied the ability to reconstruct actions and events from the past by integrating visual and auditory evidence. We found that humans were successful at making causal inferences, followed closely by `GPT-4` and then `GPT-4V`. Participants were most accurate when given multimodal evidence compared to either visual or auditory evidence alone. Across individual trials, their accuracy was weakly correlated with `GPT-4` and `GPT-4V`, but captured well by a computational model that draws on cognitive processes of inverse planning and mental simulation. Multimodal event reconstruction presents an outstanding challenge for current AI systems; computational models inspired by human multimodal reasoning may offer a way to improve them.

## References


Agrawal, T., & Schachner, A. (2023). Hearing water temperature: Characterizing the development of nuanced perception of sound sources. *Developmental Science*, *26*(3), e13321.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bitton, Y., Bansal, H., Hessel, J., Shao, R., Zhu, W., Awadalla, A., Gardner, J., Taori, R., & Schmidt, L. (2023). VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by Real-World Use.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Doyle, S. A. C. (1894). The Adventure of Silver Blaze. In *The Memoirs of Sherlock Holmes*. G. Newnes Ltd.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. (2024). Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, *36*.

Gerstenberg, T., Siegel, M., & Tenenbaum, J. (2021). What happened? Reconstructing the past through vision and sound. *PsyArXiv*.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Gori, M., Del Viva, M., Sandini, G., & Burr, D. (2008). Young children do not integrate visual and haptic information. *Nature Precedings*, 1–1.

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of personality and social psychology*, *82*(3), 379.

Heider, F. (1958). The naive analysis of action.

Jacobs, C., Lopez-Brau, M., & Jara-Ettinger, J. (2021). What happened here? Children integrate physical reasoning to infer actions from indirect evidence. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Jara-Ettinger, J., & Schachner, A. (2024). Traces of our past: The social representation of the physical world.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The Naïve Utility Calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Jin, C., Wu, Y., Cao, J., Xiang, J., Kuo, Y.-L., Hu, Z., Ullman, T., Torralba, A., Tenenbaum, J. B., & Shu, T. (2024). MMToM-QA: Multimodal Theory of Mind Question Answering.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023, January). Large Language Models are Zero-Shot Reasoners.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS one*, *2*(9), e943.

Li, Y., Wang, L., Hu, B., Chen, X., Zhong, W., Lyu, C., & Zhang, M. (2023). VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by Real-World Use.

Lopez-Brau, M., Kwon, J., & Jara-Ettinger, J. (2022). Social inferences from physical evidence via bayesian event reconstruction. *Journal of Experimental Psychology: General*, *151*(9).

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press.

Newman, G. E., Keil, F. C., Kuhlmeier, V. A., & Wynn, K. (2010). Early understandings of the link between agents and order. *Proceedings of the National Academy of Sciences*, *107*(40), 17140–17145.

OpenAI. (2023). GPT-4V(ision) System Card.

Outa, J., Zhou, X. J., Gweon, H., & Gerstenberg, T. (2022). Stop, children what's that sound? multi-modal inference through mental simulation. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Pelz, M., Schulz, L., & Jara-Ettinger, J. (2020). The signature of all things: Children infer knowledge states from static images.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers.

Schachner, A., & Kim, M. (2018). Entropy, order and agency: The cognitive basis of the link between agents and order. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Siegel, M. H., Magid, R. W., Pelz, M., Tenenbaum, J. B., & Schulz, L. E. (2021). Children's exploratory play tracks the discriminability of hypotheses. *Nature communications*, *12*(1), 3598.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. (2024, April). Large Language Models as Optimizers.

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., & Wang, L. (2023). The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision).