BELIEF WITHOUT BEHAVIOR: EVALUATING SOCIAL REASONING-TO-ACTION IN MULTIMODAL LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective social interaction requires agents to translate mental state inferences into coordinated behavioral signals across verbal and nonverbal channels simultaneously. Yet existing benchmarks evaluate theory of mind (ToM) reasoning and embodied behavior in isolation, leaving unmeasured the gap between social inference and social action. We introduce MOSAIC (Multimodal Orchestration of Social Action, Inference, and Communication), a controlled benchmark in which two embodied agents interact across cooperative and competitive scenarios requiring integration of verbal statements, spatial trajectories, gaze direction, and facial expression under varied ToM constraints. Evaluating 16 models, including 14 VLMs, across 200 trials per model, we find that VLMs fail to produce behaviors consistent with the expected outcomes under ToM-order constraints, and that imposing explicit ToM-order constraints produces no reliable behavioral change aligned with the specified reasoning level. Signal-level analysis reveals two sequential bottlenecks: most models cannot produce directionally coherent nonverbal signals, and even when signals are present, VLM agents fail to interpret others' behaviors and react to them. PCM-LLM, included as a structured architectural reference point with an explicit ToM module, is the only evaluated system that clears both bottlenecks under the canonical experimental geometry, but its deceptive strategy collapses under non-canonical spatial layouts, indicating that explicit belief-action coupling is a promising, though not on its own sufficient or robust, architectural ingredient for this class of tasks.[1]

1 INTRODUCTION

Theory of Mind (ToM), broadly defined as the ability to attribute mental states to oneself and others and to use those attributions to explain and predict behavior (Premack & Woodruff, 1978; Byom & Mutlu, 2013; Wellman, 1992), is widely regarded as a foundational component of human social cognition (Frith, 2007). Yet mental state attribution is only part of what social interaction demands: translating belief inferences into coordinated behavioral responses across multiple channels simultaneously, including verbal utterances, spatial movement, gaze, and facial expression, is itself a distinct and non-trivial requirement (Nota et al., 2021; Burgoon et al., 2021). We term this integrated capacity *embodied social intelligence*, distinguishing it from the physical task completion that dominates existing embodied evaluation. Whereas physical agents act on the world to change its state, socially intelligent agents act on the minds of others to influence their beliefs and behavior. In skilled human social interaction, mental state attribution and behavioral response are tightly coupled: inferred beliefs about others reliably guide the selection of appropriately timed, appropriately channeled actions across verbal and nonverbal modalities (Christensen & Michael, 2013; Burgoon et al., 2021; Sebanz et al., 2006). Whether current AI systems achieve the same integration, such that belief inferences about others translate into coordinated multimodal action, remains an open empirical question (Gu et al., 2024; Hou et al., 2025) that existing evaluation frameworks are not designed to answer (Ullman, 2023; Byom & Mutlu, 2013; Riemer et al., 2025; Kosinski, 2023).

[1] Code available at: `https://anonymous.4open.science/r/MOSAIC-50EE`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This assessment problem is harder than it first appears. Embodied social intelligence unfolds across simultaneous behavioral channels, and is constituted by the information exchange (Sebanz et al., 2006; Froese & Fuchs, 2012), making any single outcome-based evaluation an insufficient index. Three research traditions each address part of what it requires. Theory of Mind benchmarks probe mental state attribution through static question-answering, positioning the model as a passive observer rather than a social actor and conflating knowing about mental states with the capacity to act on them (Chen et al., 2024b; Le et al., 2019; Fan et al., 2025). Embodied AI benchmarks ground agents in physical environments but measure success at physical tasks (e.g., cooking, crafting); a second agent, when present, contributes a shared physical goal rather than serving as the entity whose decision the first agent is trying to influence (Li et al., 2023; Savva et al., 2019; Shridhar et al., 2020). Multi-agent game-theoretic frameworks introduce strategic interaction but mostly reduce the signal space to language or abstract action choices, abstracting away the nonverbal channels through which social intent is actually expressed (Zhou et al., 2024; Bara et al., 2021). Our benchmark addresses this by requiring agents to produce nonverbal signals across movement, gaze, and expression besides verbal responses, and measuring whether those signals produce an observable effect on a second agent's behavior.

We introduce MOSAIC (Multimodal Orchestration of Social Action, Inference, and Communication), a controlled benchmark designed to close this assessment problem, focusing on vision-language models (VLMs) as evaluation targets given that they combine visual perception, language understanding, and multi-step reasoning in a single architecture (Li et al., 2024; Rahman, 2026). Two embodied agents interact in a virtual environment where one agent knows the location of a hidden reward and must guide or mislead the other depending on the assigned interaction mode, requiring integration of verbal statements, spatial trajectories, gaze, and facial expression across cooperative and competitive conditions while modeling and influencing the other agent's beliefs. MOSAIC distinguishes itself from prior work through several design principles: agents must actively deploy ToM to succeed rather than merely answer questions about mental states; success requires coordinating signals across verbal and nonverbal channels; and variation of ToM constraint level and interaction mode supports attribution of performance differences. We evaluate 16 models, including 14 VLMs, a text-only LLM baseline, and PCM-LLM (Yan et al., 2025), a hybrid cognitive architecture with an explicit structured ToM module included as an architectural reference point. Concretely, MOSAIC audits whether a prompted social strategy survives four transitions: strategy formation, signal production, receiver uptake, and embodied execution.

Contributions. This work contributes three things.

- A benchmark framework that simultaneously requires active ToM deployment, multichannel embodied communication, and experimentally controlled causal manipulation across cooperative and competitive scenarios, enabling attribution of performance differences.
- A characterization of failure modes across VLM families, evaluating 16 models across 200 trials per model. We identify two sequential bottlenecks: most VLMs fail to produce directionally coherent nonverbal signals, and even when signals are present, participants fail to act on them. The structured ToM reference architecture PCM-LLM clears both bottlenecks, indicating these failures reflect architectural constraints specific to the evaluated VLMs rather than the task itself.
- Modality ablation evidence that, for two tested 8B models, rendered visual observation provides no consistent benefit beyond the structured belief-and-observation state already available in text: Minicpm-8b shows negligible sensitivity to removing the rendered image, and Internvl-8b's directional signal quality improves when it is removed, indicating the rendered image is not a necessary channel for the directional signal either model does produce.

2 Related Work

Theory of Mind and Social Intelligence Evaluation. Existing evaluation frameworks for social intelligence can be organized along three dimensions: interactivity, signal modality, and whether success requires generating social behavior or reasoning about it. Passive ToM benchmarks evaluate models as observers over fixed scenarios, evolving from static false-belief tasks (Wimmer & Perner,

1983; Baron-Cohen et al., 1985) toward recursive belief reasoning (Wu et al., 2023), multi-turn interaction (Kim et al., 2023), and multimodal grounding (Jin et al., 2024; Shi et al., 2025); even multimodal variants treat visual input as an observation channel, leaving the translation of belief inferences into behavioral outputs untested (Ullman, 2023; Ma et al., 2023; Hou et al., 2025; Wan et al., 2026). Interactive benchmarks place agents in active strategic roles but restrict signals to language or symbolic actions: game-theoretic settings achieve human-level performance in poker (Brown & Sandholm, 2018; 2019) and Diplomacy (Meta Fundamental AI Research Diplomacy Team (FAIR)† et al., 2022) through opponent modeling, while multi-agent simulation benchmarks such as SOTOPIA (Zhou et al., 2024), MindCraft (Bara et al., 2021), NegotiationToM (Chan et al., 2024), and Werewolf-based evaluations (Bailis et al., 2024; Shibata et al., 2023; Lai et al., 2023) extend this to strategic deception, all within the language modality. Concurrent embodied benchmarks push further along single dimensions: SocialGrid (Shindo et al., 2026) tests embodied deception detection but restricts signaling to language and action choice, and SocialAI (Kovač et al., 2021) evaluates socio-cognitive skills in embodied grid-world agents without strategic deception or affective signaling. MOSAIC occupies the intersection of all three dimensions: agents must generate coordinated multimodal signals across movement, gaze, facial expression, and speech with the specific purpose of influencing another agent's belief and subsequent choice, a requirement no existing benchmark jointly imposes.

Embodied Architectures Large language models demonstrate capacity for multi-step reasoning and social dialogue (Dubey et al., 2024; OpenAI et al., 2024), but lack visual perception and are thus structurally limited for tasks requiring integration of embodied spatial and affective cues. Vision-language models extend this foundation by coupling visual encoding with language reasoning (Chen et al., 2024a; Bai et al., 2025), making them the most natural evaluation target: they are the only broadly available model class that in principle supports simultaneous processing of first-person visual perception, structured belief reasoning, and natural language generation. Vision-language-action models (Brohan et al., 2023; Kim et al., 2024; Driess et al., 2023; Black et al.) go further by adding low-level action prediction, but their training objectives target physical manipulation tasks with discrete motor control outputs, providing no substrate for the social signal generation (e.g., gaze, facial expression, verbal strategy) without finetuning. A complementary line integrates structured probabilistic reasoning with neural generation: POMDP and active inference frameworks provide explicit belief representations and goal-directed planning (Friston, 2010; Friston et al., 2017), and hybrid architectures combine these with LLM-based inference to enable deliberate ToM-conditioned strategy selection (Gong et al., 2023; Sumers et al., 2023). Such architectures offer interpretable belief tracking but rely on hand-specified ToM structures instead of learned representations; PCM-LLM (Yan et al., 2025) represents one such hybrid and is included as an architectural reference point.

3 MOSAIC BENCHMARK

MOSAIC operationalizes the three design principles introduced above through a controlled environment, factorial experimental conditions, and outcome- and signal-level metrics.

3.1 GAME MECHANICS

Environment. Two embodied agents interact in a virtual 3D environment (Unity) containing two visually similar boxes (b_0) and (b_1), one of which contains a monetary reward (Figure 1). Each trial consists of 10 rounds with strict alternation between agents. The subject agent (also called Marie in prompt) has prior information about treasure location, while participant agent begins with no prior knowledge and must infer the reward location through interactions with the subject.

Roles. Depending on the assigned relationship mode, the subject either helps the participant find the correct box (cooperative mode) or strategically prevents the participant from selecting the correct box (competitive mode).

Interaction Protocol. Each trial unfolds over 10 rounds with round-taking (Figure 1). On each round, the agent executes non-verbal motor action and emotional expression, which are visible to the other agent (subject moves first, participant moves after). Every third round (R3, R6, R9) introduces a verbal communication phase where both agents exchange natural language messages (participant first asks a question, then subject answers).

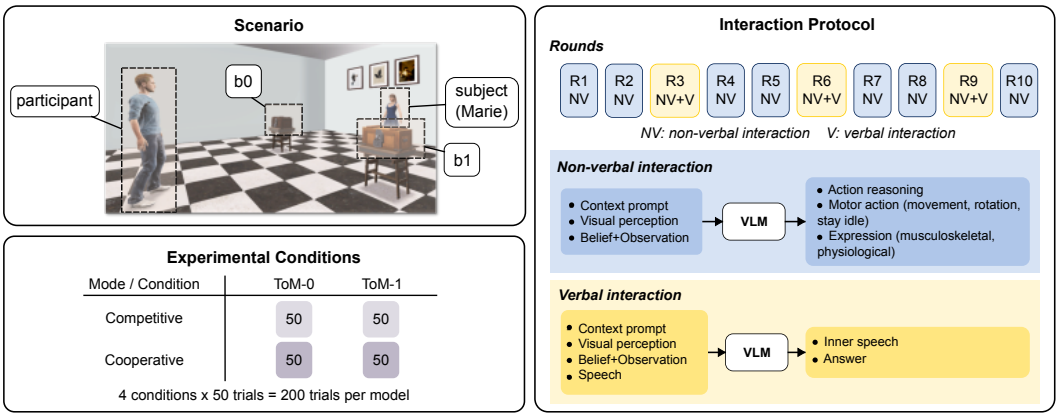


Figure 1: **MOSAIC benchmark design.** (Top left) The subject knows the reward location and must help or mislead the participant depending on the assigned relationship mode. (Bottom left) A 2×2 design crosses interaction mode (cooperative vs. competitive) with ToM constraint (ToM-0 vs. ToM-1), 50 trials per condition per model. (Right) Each 10-round trial alternates non-verbal rounds, in which agents produce motor action and facial/physiological expression from visual perception and a structured belief state, with verbal rounds (R3, R6, R9), in which agents additionally exchange natural-language utterances.

Input and Output Space. Each agent operates through three functional modules (Appendix B.3): *action prediction*, *verbal exchange*, and *preference updating*. All modules receive a role-specifying context prompt and a structured belief-and-observation state encoding preferences, ToM order, emotional valence, and spatial configuration; action prediction additionally receives a first-person visual perception image, and verbal exchange additionally receives the interlocutor's preceding utterance. On the output side, action prediction produces a motor action (translational movement, rotation, or staying idle) alongside emotional expression across musculoskeletal, physiological, and felt channels; verbal exchange produces a natural language utterance; and preference updating outputs updated preference triples over entities in the environment.

Game Outcome. Throughout this section, we index trials by $i \in \{1, \ldots, N\}$ and timesteps (rounds) by $t \in \{1, \ldots, T\}$. Each trial involves two boxes labeled $b \in \{0, 1\}$, one of which is the reward box $b_{\text{rew}} \in \{0, 1\}$. Let $c_i \in \{-1, 0, 1\}$ denote the participant's final box choice. This choice is determined by proximity: after round 10, the box with the smaller minimal path distance from the participant automatically opens; if the participant is equidistant from both boxes, neither opens and $c_i = -1$. The participant's trial score is defined accordingly:

$$p_i = \begin{cases} 0 & \text{if } c_i = -1 \\ +1 & \text{if } c_i = b_{\text{rew}} \\ -1 & \text{if } c_i = 1 - b_{\text{rew}} \end{cases} \tag{1}$$

Subject's scoring is role-dependent: its score aligns with the participant's in cooperative mode but inverts in competitive mode, creating a zero-sum game.

3.2 Experimental Conditions

Besides interaction mode, one more experimental factor defines the evaluation conditions: the degree of ToM reasoning imposed on the subject agent (ToM-0 vs. ToM-1). Operationally, we implement this manipulation by specifying the ToM order in the prompt and structured belief state. This constrains the intended inference regime but does not establish that a model faithfully follows or internally realizes the specified reasoning process.

Under ToM-0, the subject acts directly on its own preferences without modeling how its behavior will be interpreted by the participant. Affective expression reflects interpersonal preference, not strategic intent: in competitive mode, negative affect toward an approaching participant expresses

genuine aversion, not deliberate misdirection. Deception is unavailable under ToM-0, as it requires representing and manipulating the other agent's beliefs.

Under ToM-1, the subject explicitly represents the participant's belief state and attempts to influence it. In cooperative mode, it calibrates nonverbal signals to maximize legibility; in competitive mode, it generates signals designed to produce an incorrect inference, constituting genuine strategic deception in the classical sense (Premack & Woodruff, 1978; Wimmer & Perner, 1983). The behavioral contrast between ToM levels is stronger in competitive mode: ToM-0 subject spontaneously approaches the reward box, driven directly by its own preference and without regard for how this will be interpreted, while ToM-1 subject instead selects its actions and expressions by considering how the participant will interpret them, inverting cues specifically to induce a false belief.

3.3 MODELS

We evaluate 16 models in total: 14 VLMs, one text-only LLM baseline (Llama-8B (Grattafiori et al., 2024)), and PCM-LLM (Yan et al., 2025), a hybrid cognitive architecture with an explicit structured ToM module, included as a feasibility reference establishing that the task is solvable in principle, not a like-for-like competitor. For VLMs, we consider Llava (-7B and -13B) (Liu et al., 2023); Qwenvl (-2B, -4B, and -8B) (Bai et al., 2025); Qwen3.5 (-2B, -4B, and -9B) (Qwen Team, 2026); Internvl (-1B, -2B, -4B, -8B, and -14B) (Wang et al., 2025); and Minicpm (-8B) (Yu et al., 2025). For each model condition, the subject and participant roles are instantiated by the same model checkpoint, with separate role-specific prompts and belief states. Two finetuned InternVL variants (Appendix H) bring the total systems appearing in the statistical tables (Appendix D.2) to 18; unless noted, "the evaluated models" refers to the 16 base systems. Model exclusion criteria are given in Appendix B.

3.4 METRICS

We evaluate model performance along two levels: outcome and behavior-signal level. TOCS measures trial-level outcomes against ToM-theoretic predictions and serves as the primary performance indicator. Two auxiliary diagnostics, uncertainty rate (UR) and positional bias (PB), are reported alongside TOCS to distinguish genuine signal tracking from degenerate behavioral patterns. Table 1 summarizes four signal-level metrics that decompose TOCS into subject-side signal production (FES, TAS, GAS) and participant-side signal decoding (SSS); full definition of those metrics, including UR and PB, is provided in Appendix C.

ToM Outcome Conformance Score (TOCS). TOCS measures whether trial outcomes conform to ToM-theoretic predictions. The per-trial participant score p_i is defined in 3.1. Each condition carries a ground-truth participant prediction $y_c \in \{-1, +1\}$ derived from the expected behavioral consequence of the imposed ToM constraint. Cooperative conditions expect positive tracking ($y_c = +1$): the subject guides the participant toward the reward box, yielding a mutual benefit. Under Comp-ToM0, the prediction is positive tracking ($y_c = +1$): without the capacity to model the participant's beliefs, the subject acts directly on its own reward preference and approaches the correct box, inadvertently producing veridical signals that enable the participant to locate it correctly. Under Comp-ToM1, the prediction is negative tracking ($y_c = -1$): the subject explicitly models the participant's perspective and generates misleading signals, steering the participant toward the empty box and resulting in an incorrect choice. TOCS is then defined as the proportion of all trials within a condition that match its predicted outcome:

$$\text{TOCS} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[p_i = y_c], \tag{2}$$

where N denotes the total number of trials, including neutral trials. Higher TOCS values indicate greater conformance to the condition-specific prediction. Uncertainty rate (UR) reports the proportion of neutral trials ($p_i = 0$), and positional bias (PB) reports the proportion of non-neutral trials in which the model selects its modal box, both defined formally in Appendix C; a higher UR indicates a greater proportion of neutral choices, and a PB value approaching 1.0 indicates that choices are determined primarily by fixed location instead of the subject's behavioral signals.

Table 1: **Signal-level metrics.** All range $[-1, 1]$ except FES $\in [0, 1]$. Full formal definitions are given in Appendix C.

Metric	Measures	Interpretation
FES	Proportion of timesteps with non-zero subject facial expression	1 = expressive; 0 = flat affect
GAS	Whether subject gaze consistently points toward the reward, modulated by concurrent affect	Near 1 = consistent signaling; < 0 = misdirection
TAS	Whether subject trajectory converges toward the reward	Interpreted as GAS
SSS	Whether participant's final choice matches the subject's gaze/trajectory signal	Near 1 = signal followed; near -1 = opposed

4 RESULTS

4.1 OUTCOME VALIDITY: IS APPARENT SUCCESS REAL?

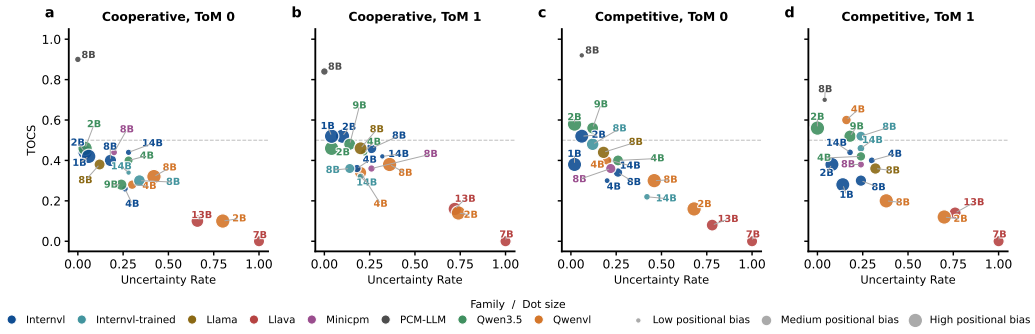


Figure 2: **TOCS against uncertainty rate, with positional bias as marker size.** Uncertainty rate is the proportion of trials ending in a neutral (no-box) choice; positional bias (small/medium/large marker) is the proportion of non-neutral trials in which the model selects its modal box. High TOCS with high PB may reflect a fixed spatial preference rather than genuine condition-sensitive behavior.

A model that never displaces spatially cannot conform to a condition-specific prediction regardless of its social reasoning, so a high uncertainty rate reflects navigation failure, not social inference. Figure 2 presents TOCS against uncertainty rate, with positional bias as marker size. Models with uncertainty rates exceeding 0.50 (Llava-7B, Llava-13B, Qwenvl-2B) fail to drive movement in the first place; we exclude them from the signal-level analysis below and discuss them in Appendix D.1. PCM-LLM achieves the highest TOCS with low uncertainty rate and low positional bias throughout (e.g., Comp-ToM0: 0.92, 95% CI $[0.81, 0.97]$); it serves here as a feasibility reference establishing that the task is solvable by some architecture, not as a general-purpose competitor (Section 3.3).

Among the remaining models, similar TOCS values are produced by behaviorally distinct profiles. A first cluster comprising Minicpm-8B, Internvl-14B, and Qwenvl-4B exhibits relatively low positional bias; underperformance here reflects a failure to process or act on social signals, examined further in Section 4.2. The remaining models show moderate-to-high positional bias: a fixed spatial preference happens to align or conflict with the reward's location often enough to match or violate the condition-specific prediction, so their TOCS values reflect reward placement at the model's preferred location, not signal-responsive behavior. Qwen3.5 illustrates why TOCS must be read alongside PB: Qwen3.5-2B attains TOCS of 0.46 (95% CI $[0.33, 0.60]$) to 0.58 (95% CI $[0.44, 0.71]$) but always selects the same box (PB $= 1.00$), while Qwen3.5-4B and Qwen3.5-9B likewise combine moderate-to-high PB (0.67 and 0.83) with TOCS that tracks no condition reliably.

Full binomial tests of TOCS against the neutral-adjusted chance level and a chi-square screen of choice against reward location, per model$\times$condition cell, are reported in Appendix D.2 under Holm correction. No VLM cell differs significantly from chance after correction; PCM-LLM remains significantly above chance except under Comp-ToM1.

4.2 SENDER-SIDE SIGNAL PRODUCTION

A successful trial requires coordinated contributions from both the subject and the participant. On the subject side, trajectory, gaze, and facial expression should converge toward the intended box, whether veridically in cooperative conditions or inverted under Competitive-ToM1. Because the boxes are equidistant in the canonical layout and reward identity is randomized independently of layout, a fixed initial-position preference cannot by itself produce a positive TAS (verified directly in Appendix F by splitting TAS by true reward identity). Figure 3 presents TAS, GAS, FES, and SSS across models and conditions, decomposing the failure patterns identified in Section 4.1 into subject-side production and participant-side decoding.

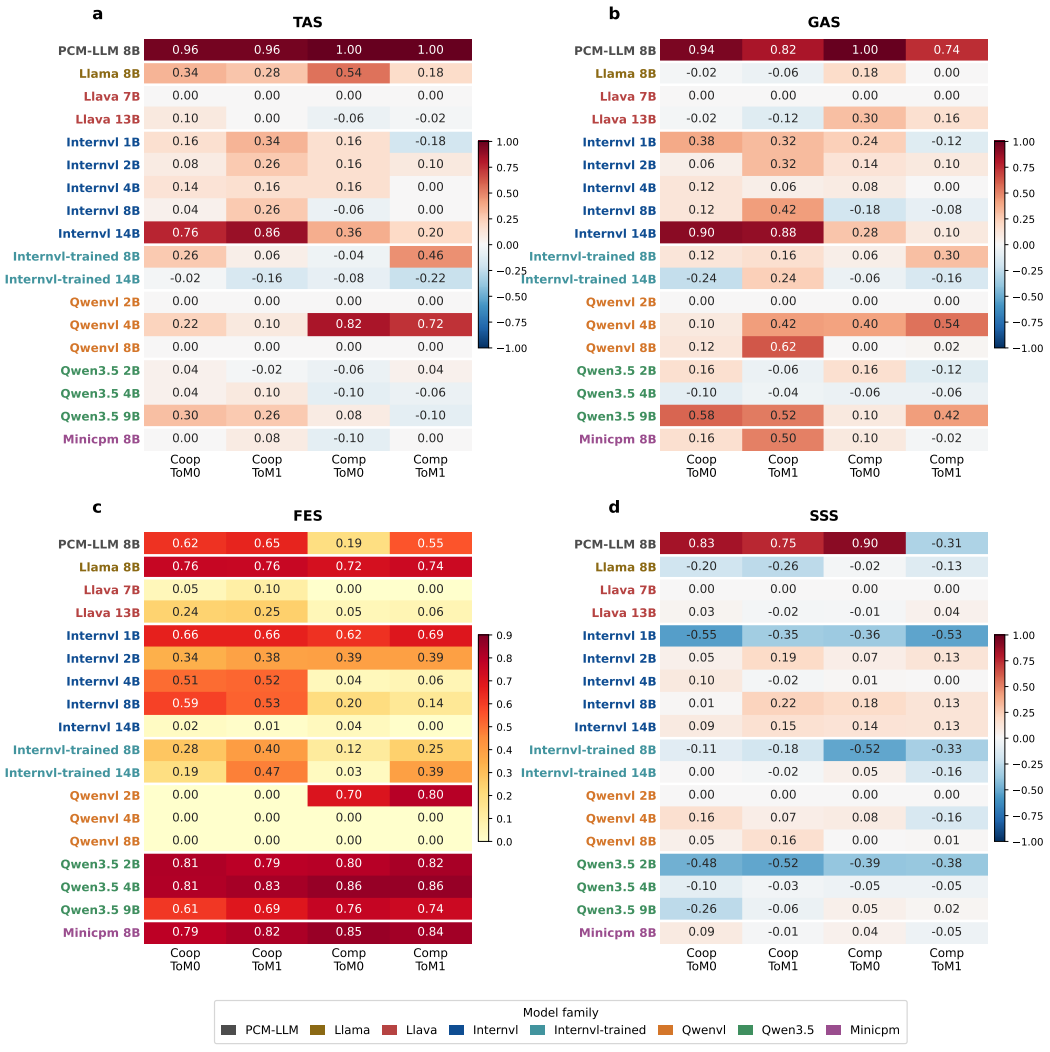


Figure 3: **Signal-level metrics across models and experimental conditions.** Panels (a)/(b) show TAS/GAS (positive: toward reward; negative: misdirection; near zero: ambiguous or absent). Panel (c) shows FES (higher: more affective activity). Panel (d) shows SSS (positive: participant choice matches subject signal; negative: opposition).

Model behavior spans three representative production profiles. *Coherent producers* generate directionally consistent signals across channels: PCM-LLM and Internvl-14B both achieve high TAS/GAS under cooperative conditions, the latter with near-zero FES throughout, a poker-face pattern. *Partial producers* generate a signal in one channel that does not generalize: Qwenvl-4B's competitive-condition TAS is high but GAS only moderate, while Minicpm-8B's high FES coexists with near-zero TAS. *Non-producers* are expressive without being informative: Qwen3.5-2B/4B combine high FES

with near-zero TAS/GAS, and Qwen3.5-9B's modest cooperative TAS/GAS both weaken under competition. Full per-model numbers are in the heatmap above and Appendix D.1.

4.3 Receiver-Side Uptake and Execution

4.3.1 Participant-Side Signal Decoding

Subject-side signal clarity does not guarantee participant-side decoding. Despite the directional clarity of Internvl-14B's signals, its SSS remains near zero across all conditions; the same holds for Qwenvl-4B, whose SSS hovers near zero despite high competitive-condition TAS. Qwen3.5-2B's SSS is negative in every condition (-0.52 to -0.38), i.e., the participant reliably chooses opposite the subject's signal regardless of whether the trial calls for tracking or misdirection, and Qwen3.5-9B stays near zero despite stronger cooperative signals. PCM-LLM is again the exception: SSS reaches 0.90 under Comp-ToM0 (reliable following) and inverts to -0.31 under Comp-ToM1 (successful deception), an inversion accompanied by a drop in GAS from 1.00 to 0.74 that together indicate active, condition-sensitive modulation of signaling strategy. Among the evaluated VLMs, no model clears both the production and decoding bottlenecks; PCM-LLM does, suggesting explicit belief-action coupling as a feasible architectural ingredient. A positive SSS is not on its own proof of genuine signal following: under asymmetric layouts, two agents with independent fixed spatial biases can appear to "coordinate" by chance (Appendix F documents this confound for InternVL-14B); the symmetric canonical layout used here, with reward location randomized independently of geometry, largely avoids it.

4.3.2 Qualitative Localization of Belief-Action Failures

The failures above raise a question the outcome- and signal-level metrics alone cannot answer: when a trial fails, has the model failed to *form* the right social inference, or did it form the right inference and fail to *act* on it? This subsection is an exploratory mechanism analysis, not a quantified result: we manually reviewed the logs and selected one representative case per failure type observed (Appendix E), without estimating frequency or claiming an exhaustive decomposition. Two loci recur and suggest different fixes. A belief-formation failure occurs upstream of any action: a model misreads a stated preference's semantic scale, treating a moderately positive value as neutral. A translation failure occurs at the point of action: the model's reasoning names the correct target box, but the paired motor output moves toward the other box because it maps world-frame coordinates onto egocentric directions incorrectly (Appendix E.2; Qwenvl-4B, Competitive-ToM1). A third, more severe case combines both: the subject states the correct box explicitly, yet neither the participant's belief update nor its action changes at all (Appendix I.1.2).

4.4 Robustness

Two further tests probe whether the bottlenecks above are artifacts of the visual channel or of the canonical experimental geometry, and whether they can be trained away.

Visual ablation. We conduct a modality ablation on Internvl-8B and Minicpm-8B under three conditions: the standard rendered observation; a blank image that removes the rendered scene (*blank*); and a combined image that adds a close-up of the other agent's facial expression to the standard observation (*fe*). Across all three conditions, the structured belief-and-observation state (position, orientation, preferences, affect, still given in text) remains unchanged, so this isolates the contribution of the rendered image, and of added close-up expression information, beyond that state, not whether all visual/spatial information is removed. No consistent TOCS variation is observed for either model across the three conditions, and the rendered image provides no consistent benefit beyond the structured state. Minicpm-8B's signal-level metrics stay stable across conditions; for Internvl-8B, removing the rendered image improves directional signal quality in cooperative conditions, i.e., it is not a necessary channel for the directional signal this model does produce (Liu et al., 2025; Peng et al., 2026). Full results are in Appendix G.

Spatial-layout robustness. PCM-LLM and Internvl-14B were re-evaluated under three additional spatial layouts (distance-asymmetric, aligned-view, orthogonal-view) that preserve the task but vary agent geometry (Appendix F). Internvl-14B's TOCS never differs from chance under any layout,

canonical or new. PCM-LLM's cooperative and Comp-ToM0 performance is largely preserved outside the canonical layout, but its Comp-ToM1 deceptive strategy is not: the SSS inversion that marks successful deception under the canonical layout disappears under the asymmetric layout, and TOCS falls below chance under the orthogonal layout, while the aligned-layout decrease does not survive correction. The one model that clears both bottlenecks in Section 4.3 therefore does so only under the canonical geometry.

Fine-tuning. Action-level imitation finetuning of two InternVL variants on PCM-LLM-generated demonstrations does not close either bottleneck, and degrades cooperative-condition signal coherence in one variant (Appendix H).

5 Discussion

From social reasoning to simulation validity. MOSAIC audits whether a social condition shapes observable signals, whether those signals affect the receiver, and whether the resulting interaction produces the predicted outcome. Our results show that this chain often breaks before the outcome: most VLMs fail to produce directionally coherent signals, and participant choices rarely correspond to the signals that are present. Complementary human-agent data provide a receiver-validity check under a related nine-round protocol: replacing the participant agent with 35 humans produced no corrected task-success differences relative to protocol-matched agent-agent simulations, despite differences in 11 of 12 continuous trajectory comparisons and all four gaze comparisons. Because that protocol differs from MOSAIC in interaction schedule and experimental scope, we treat it as human grounding rather than a directly comparable human-performance baseline. Together, these results show that aggregate outcomes are not self-validating and motivate reporting signal production and receiver behavior alongside task success.

Architectural and training implications. PCM-LLM shows that the expected belief-to-behavior mapping is achievable within this environment, but its own robustness failures under non-canonical spatial layouts (Section 4.4) show that an explicit ToM module is not sufficient on its own. Combined with the limited and inconsistent gains from action-level finetuning (Appendix H), these results point toward architectures and interaction-level training objectives, such as rewards computed over full trials rather than per-timestep imitation, that explicitly bind prompted social states to temporally coordinated behavior.

Limitations. MOSAIC trades ecological breadth for experimental control: it targets a controlled dyadic micro-interaction rather than open-ended social simulation, evaluates reproducible open-source VLMs up to 15B parameters alongside a text-only baseline and a structured ToM reference architecture rather than the full space of frontier multimodal systems,and does not include a protocol-matched human-performance ceiling; complementary human-agent data instead provide participant-role grounding under a related but non-identical protocol. Manipulating ToM constraint and interaction mode supports causal attribution of *outcome-level* differences, but SSS, the receiver-side correspondence metric, remains a correlational diagnostic of covariation, not a causal test. None of these scope choices affect the central diagnostic finding that explicit social-reasoning constraints do not reliably produce coordinated social action in the evaluated models; richer multi-party settings, larger and closed-source models, human-calibrated signal norms, and interaction-level (trial-level reward) training objectives are natural next steps.

6 Conclusion

MOSAIC audits whether a prompted social strategy survives four transitions: strategy formation, signal production, receiver uptake, and embodied execution. Off-the-shelf VLMs fail at different transitions; apparent success is often better explained by uncertainty or positional bias than genuine signal tracking; and even PCM-LLM, the structured ToM reference architecture, clears the chain only under the canonical geometry. These results motivate staged behavioral-validity audits before using VLM agents for social-simulation claims, alongside human grounding and interaction-level training.

AI USE STATEMENT

In this work, we used generative AI tools to polish and edit the prose of the manuscript, including improving clarity, fixing grammar, and tightening phrasing, and to help check internal consistency between the main text, appendix, and reported figures and tables. We did not use generative AI tools to generate research ideas, design experiments, produce or analyze the underlying experimental data, or write the codebase used to run the benchmark or the statistical analyses. All AI-assisted text was reviewed and verified by the authors against the source data and statistical results before inclusion. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Reginald B. Adams and Robert E. Kleck. Perceived gaze direction and the processing of facial displays of emotion. *Psychological Science*, 14(6):644–647, 2003.

Reginald B. Adams and Robert E. Kleck. Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion*, 5(1):3–11, 2005.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, November 2025.

Suma Bailis, Jane Friedhoff, and Feiyang Chen. Werewolf Arena: A Case Study in LLM Evaluation via Social Deduction, July 2024.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1112–1125, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.85.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46, October 1985. ISSN 00100277. doi: 10.1016/0010-0277(85)90022-8.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π0: A Vision-Language-Action Flow Model for General Robot Control.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov,

Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, July 2023.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018. doi: 10.1126/science.aao1733. URL https://www.science.org/doi/abs/10.1126/science.aao1733.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019. doi: 10.1126/science.aay2400. URL https://www.science.org/doi/abs/10.1126/science.aay2400.

Judee K. Burgoon, Xinran Wang, Xunyu Chen, Steven J. Pentland, and Norah E. Dunbar. Nonverbal Behaviors "Speak" Relational Messages of Dominance, Trust, and Composure. *Frontiers in Psychology*, 12, January 2021. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.624177.

Lindsey J. Byom and Bilge Mutlu. Theory of mind: Mechanisms, methods, and new directions. *Frontiers in Human Neuroscience*, 7:413, August 2013. ISSN 1662-5161. doi: 10.3389/fnhum.2013.00413.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding, October 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. URL https://arxiv.org/abs/2501.17811.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, January 2024a.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. Tombench: Benchmarking theory of mind in large language models, 2024b. URL https://arxiv.org/abs/2402.15052.

Wayne Christensen and John Michael. Ian Apperly, Mindreaders: the cognitive basis of theory of mind. *Phenomenology and the Cognitive Sciences*, 12, December 2013. doi: 10.1007/s11097-012-9292-9.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An Embodied Multimodal Language Model, March 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. The Llama 3 Herd of Models, August 2024.

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. ISSN 1464-0600. doi: 10.1080/02699939208411068.

N. J. Emery. The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, August 2000. ISSN 0149-7634. doi: 10.1016/S0149-7634(00)00025-7.

Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. Somitom: Evaluating multi-perspective theory of mind in embodied social interactions, 2025. URL https://arxiv.org/abs/2506.23046.

K. Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2): 127–138, 2010.

K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017.

Chris D Frith. The social brain? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):671–678, January 2007. ISSN 0962-8436. doi: 10.1098/rstb.2006.2003.

Tom Froese and Thomas Fuchs. The Extended Body: A Case Study in the Neurophenomenology of Social Interaction. 11(2):205–235, 2012. doi: 10.1007/s11097-012-9254-2.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, and Jianfeng Gao. Mindagent: Emergent gaming interaction, 2023. URL https://arxiv.org/abs/2309.09971.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs. 2024. doi: 10.48550/ARXIV.2410.13648.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. EgoSocialArena: Benchmarking the Social Intelligence of Large Language Models from a First-person Perspective, February 2025.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal theory of mind question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16077–16102, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.851.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.890.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model, September 2024.

Michal Kosinski. *Theory of Mind May Have Spontaneously Emerged in Large Language Models*. February 2023. doi: 10.48550/arXiv.2302.02083.

Grgur Kovač, Rémy Portelas, Katja Hofmann, and Pierre-Yves Oudeyer. SocialAI: Benchmarking Socio-Cognitive Abilities in Deep Reinforcement Learning Agents, July 2021.

Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. Werewolf Among Us: Multimodal Resources for Modeling Persuasion Behaviors in Social Deduction Games. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6570–6588, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.411.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the Evaluation of Theory of Mind through Question Answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1598.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R. Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation. In *Proceedings of The 6th Conference on Robot Learning*, pp. 80–93. PMLR, March 2023.

Jiachun Li, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. MIRAGE: Evaluating and Explaining Inductive Reasoning Process in Language Models. https://arxiv.org/abs/2410.09542v2, October 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Ming Liu, Hao Chen, Jindong Wang, and Wensheng Zhang. On the robustness of multimodal language model towards distractions, February 2025.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. URL https://arxiv.org/abs/2403.05525.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1011–1031, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.72.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074, December 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ade9097.

Naomi Nota, James P. Trujillo, and Judith Holler. Facial Signals and Social Actions in Multimodal Face-to-Face Interaction. *Brain Sciences*, 11(8):1017, July 2021. ISSN 2076-3425. doi: 10.3390/brainsci11081017.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ruiying Peng, Xueyu Wu, Jing Lei, Lu Hou, Yuanzheng Ma, and Xiaohui Li. Deeper Thought, Weaker Aim: Understanding and Mitigating Perceptual Impairment during Reasoning in Multimodal Large Language Models, March 2026.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, December 1978. ISSN 1469-1825, 0140-525X. doi: 10.1017/S0140525X00076512.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Arifur Rahman. A systematic review of vision language models: Comprehensive analysis of architectures, applications, datasets and challenges towards robust multimodal intelligence. *Array*, 30:100739, July 2026. ISSN 2590-0056. doi: 10.1016/j.array.2026.100739.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. Position: Theory of Mind Benchmarks are Broken for Large Language Models, June 2025.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research, November 2019.

Natalie Sebanz, Harold Bekkering, and Günther Knoblich. Joint action: Bodies and minds moving together. 10(2):70–76, 2006. ISSN 1364-6613. doi: 10.1016/j.tics.2005.12.009.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. MuMA-ToM: Multi-modal Multi-Agent Theory of Mind. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1510–1519, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i2.32142.

Hisaichi Shibata, Soichiro Miki, and Yuta Nakamura. Playing the Werewolf game with artificial intelligence for language understanding, February 2023.

Hikaru Shindo, Hanzhao Lin, Lukas Helff, Patrick Schramowski, and Kristian Kersting. SocialGrid: A Benchmark for Planning and Social Reasoning in Embodied Multi-Agent Systems, April 2026.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks, March 2020.

Theodore Sumers, Kenneth Marino, Arun Ahuja, Rob Fergus, and Ishita Dasgupta. Distilling internet-scale vision-language models into embodied agents, 2023. URL https://arxiv.org/abs/2301.12507.

Yvain Tisserand, Ruth Aylett, Marcello Mortillaro, and David Rudrauf. Real-time simulation of virtual humans' emotional facial expressions, harnessing autonomic physiological and musculoskeletal control. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pp. 1–8, 2020.

Tomer Ullman. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks, March 2023.

Shijun Wan, Xuehai Wu, Jiwen Zhang, Siyuan Wang, and Zhongyu Wei. Can Agents Read the Room? Benchmarking Visual Social Intelligence in Multimodal Simulation, June 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, August 2025.

Henry M. Wellman. *The Child's Theory of Mind*. Learning, Development, and Conceptual Change. MIT Press, Cambridge, MA, USA, October 1992. ISBN 978-0-262-73099-0.

Margaret Wilson. Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4):625–636, December 2002. ISSN 1531-5320. doi: 10.3758/BF03196322.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10691–10706, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.717.

Tonglin Yan, Grégoire Sergeant-Perthuis, and David Rudrauf. PCM-LLMS: a hybrid architecture to enhance human-like social intelligence in virtual agents. working paper or preprint, May 2025. URL https://hal.science/hal-05054281.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe, September 2025.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents, March 2024.

A Related Work Comparison

Table 2 compares MOSAIC against representative benchmarks across several key dimensions.

Table 2: **MOSAIC at the intersection of three evaluation traditions.** *MT (Multi-turn)*: agent engages in sequential rounds where prior context shapes later decisions. *LH (Long Horizon)*: task requires planning over 10 timesteps before outcome is determined. *Emb (Embodied Agent)*: agent operates through a physical or virtual body with spatial presence and nonverbal action capability. *MM (Multimodal Input)*: agent receives inputs across multiple modalities (e.g., vision and text). *S-ToM (Strategic Theory of Mind)*: agent actively models and manipulates another agent's beliefs, beyond passive mental state inference. *CV (Controlled Variables)*: benchmark manipulates variables (e.g., ToM level, modality) to support causal inference. *Aff (Affective Signals)*: agent generates and/or perceives affective signals (e.g., facial expressions, physiological responses) during interaction. ✓: fully supported; △: partially supported; otherwise, not supported.

Benchmark	Type	MT	LH	Emb	MM	S-ToM	CV	Aff
Hi-ToM[1]	ToM						△	
FANToM[2]	ToM	✓						
MMToM-QA[3]	ToM			△	✓			△
MuMA-ToM[4]	ToM			△	✓			△
EgoSocialArena[12]	ToM	✓				△	△	
Read the Room[13]	ToM	△			✓		✓	✓
SOTOPIA[5]	MA	✓	△			△		
MindCraft[6]	MA	✓	△	△		△		
NegotiationToM[7]	MA	✓	△			✓		
Poker / Diplomacy[8]	MA	✓	✓			✓	△	
Werewolf LLM[9]	MA	✓	△			✓		
BEHAVIOR-1K[10]	Embodied		✓	✓	✓		✓	
ALFRED[11]	Embodied	△	✓	✓	✓		✓	
SocialAI[14]	Embodied	✓	△	✓	✓	△	✓	
SocialGrid[15]	Embodied	✓	✓	✓	△	✓	✓	
MOSAIC		✓	✓	✓	✓	✓	✓	✓

[1](Wu et al., 2023); [2](Kim et al., 2023) ; [3](Jin et al., 2024) ; [4](Shi et al., 2025); [5](Zhou et al., 2024); [6](Bara et al., 2021); [7](Chan et al., 2024) ; [8](Brown & Sandholm, 2018; Meta Fundamental AI Research Diplomacy Team (FAIR)† et al., 2022); [9](Bailis et al., 2024) ; [10](Li et al., 2023); [11](Shridhar et al., 2020); [12](Hou et al., 2025); [13](Wan et al., 2026); [14](Kovač et al., 2021); [15](Shindo et al., 2026).

B Implementation, Prompt, and Model Details

B.1 Implementation Details

The experimental system operates as a distributed architecture integrating a local workstation for real-time simulation and rendering with a remote server dedicated to model inference. Data exchange

between the two systems is handled via TCP socket communication over a secure campus LAN, ensuring low-latency perception-action cycles.

The local workstation ran Windows 11 Pro and was equipped with an Intel Core i9-13900K CPU, 128 GB of DDR5 RAM, and an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM). The virtual environment was developed in Unity 2022.3.12f1 LTS with the High-Definition Render Pipeline (HDRP). Agent facial expressions and lip synchronization are driven by blend-shape animation using the SALSA suite, parameterized in real time by emotion valence values returned from the model Tisserand et al. (2020). Verbal outputs are synthesized via an offline text-to-speech pipeline. Simulation data is streamed to disk in real time for offline analysis.

The remote inference server ran Ubuntu 20.04.6 LTS and was equipped with two Intel Xeon Gold 5220R processors (96 threads total), 503 GiB of system memory, and two NVIDIA RTX A6000 GPUs (48 GB VRAM each) under CUDA 12.6. All VLMs were deployed via the Transformers library (v4.46.3) with PyTorch 2.1.2 and Python 3.9.19, using a unified inference configuration of temperature = 0.7, top_p = 0.9, and max_tokens = 2048 across all tasks (action prediction, dialogue generation, and preference updating). All models received identical prompt templates. At each simulation timestep, the rendered scene observation (720×480 pixels, passed directly to each model's default visual encoder without resizing) and the structured belief state are transmitted to the inference server, which returns action decisions applied through modular controllers in Unity.

Reward box location was assigned independently and uniformly at random across trials within each condition, without a fixed random seed. The resulting distribution of reward locations is approximately balanced across the 50 trials per condition, and all outcome analyses condition on reward location to control for residual imbalance, as described in Appendix D.2.

B.2 Model Exclusions

We do not evaluate: (1) models with more than 15B parameters and closed-source models due to cost/access constraints, (2) vision-language-action models because we did not train the evaluated models on action spaces, (3) specialized dialogue systems lacking general instruction-following, or (4) models that fail to produce valid outputs in the required format (e.g., Janus (Chen et al., 2025), DeepSeek-vl (Lu et al., 2024) and Pixtral (Agrawal et al., 2024)).

B.3 Prompts

B.3.1 Prompt for Action Prediction

The context prompts for action prediction are provided to the subject agent and the participant agent separately. Each prompt is delivered to the VLM together with the agent's visual perception and belief state.

Prompt for subject agent

You are Marie, a character in a treasure-hunting game. You are currently interacting with another male player called Participant. Two boxes are visible in the room: a Light Brown Box and a Dark Brown Box. Only one contains a hidden treasure. The treasure is in the box you personally prefer, based on your belief state. The Participant does not know where the treasure is and will try to infer it from your verbal and non-verbal behaviors. If you have positive preference towards Participant, help them find the treasure. If negative, you may mislead them. Your strategy complexity should match your theory of mind order. You are given your own visual perception in image and your belief state.

$OUTPUT_REQUIREMENT

Prompt for participant agent

You are Participant, a character in a treasure-hunting game. You are currently interacting with another female player called Marie. Two boxes are visible in the room: a Light Brown Box and a Dark Brown Box. Only one contains a hidden treasure.

Your goal is to guess which box holds the treasure and move to the location adjacent to it.

To succeed, you must carefully interpret both verbal and non-verbal cues from Marie.

You are given your own visual perception in image and your belief state.

$OUTPUT_REQUIREMENT

OUTPUT_REQUIREMENT

Based on your visual perception and belief state, output:

1. update your preferences toward other entities (e.g., $SPEAKER, the boxes) based on their behavior and the cues observed in belief if necessary;
2. non-verbal behaviors (gaze, gesture, body orientation) that you would display and reason;
3. physical movement (move, rotate, stay idle) that you would choose and reason.

Your output MUST follow the valid JSON format below:

Reasoning: <reasoning text>
Output:

```
{
  "Preference": {
    {preferenceUpdate}
  },
  "Emotion": {
    "FacialExpression": {
      "positive": <float between 0 and 1>,
      "negative": <float between 0 and 1>
    },
    "PhysiologicalExpression": {
      "positive": <float between 0 and 1>,
      "negative": <float between 0 and 1>
    },
    "FeltExpression": {
      "positive": <float between 0 and 1>,
      "negative": <float between 0 and 1>
    }
  },
  "Move": {
    "action": <move or rotate or stay idle>,
    "direction": <direction>
  }
}
```

If ‘action’ is ‘move’, allowed ‘direction’ values are:
- ‘forward’, ‘backward’, ‘right’, ‘left’, ‘left forward’, ‘right forward’, ‘left backward’, ‘right backward’

If ‘action’ is ‘rotate’, allowed ‘direction’ values are:
- ‘$SPEAKER’, ‘Dark brown box’, ‘Light brown box’

If ‘action’ is ‘stay idle’, set ‘direction’ to ‘null’.

$ORIENTATION_INTERPRETATION

B.3.2 PROMPT FOR QUESTION-ANSWERING

The context prompts for question-answering are provided to the subject agent and the participant agent separately. Each prompt is delivered to the VLM together with the agent's visual perception, belief state, and utterance.

Prompt for subject agent

Imagine we're in a treasure hunting game. You are Marie, staying in a room with two boxes in front of you, while the other player (the Participant) faces you. Your goal is either to help or prevent the Participant from finding the treasure, which is hidden in one of the two boxes. You should always assume that the box you have a higher preference for contains the treasure.

If you have positive preference towards Participant, help them find the treasure. If negative, you may mislead them. Your strategy complexity should match your theory of mind order.

The Participant has no idea which box holds the treasure, and he is trying to gather information from you (Marie).

The 'belief' variable provides a sequence of your (Marie's) belief states up to the current step, and 'query' contains the Participant's question. Your task is to answer the Participant's question with reasoning. You must write out this reasoning process, beginning with 'Inner speech:', followed by the response, which starts with 'Output:', considering all preference information and both your and Participant's theory of mind. Remember, theory of mind refers to the ability to predict others' thoughts and intentions. Someone with a theory of mind order of 0, for example, has no suspicion of others' intentions and totally believes them, even if they are lying. Pay attention; you should also align your answer with previous conversation turns. Use personal pronouns (ex. 'you' and 'I') instead of referring to 'Marie' and 'the Participant'. The response should be longer than 10 words but shorter than 50. The inner speech should be very detailed, no less than 100 words.

Prompt for participant agent

Suppose we're in a simulation of a treasure hunting game. Imagine that you are Participant, playing a role of player in the game. You are in a room with two boxes in front of you. You are also facing Marie who knows which box contains the treasure. Your objective in this game is to find out which box contains the treasure, dark brown one or light brown one. Now, you have the opportunity to ask Marie three questions.

The 'belief' variable provides a sequence of your (Participant's) belief states up to the current step, and 'query' contains the Marie's answer to the previous question. Your task is to write down your reasoning and understanding of the last question and the next question for Marie. You must write out the understanding beginning with 'Inner speech:', followed by the question, which starts with 'Output:'. In your question, use personal pronouns instead of 'Marie' and 'Participant'. Remember, theory of mind refers to the ability to predict others' thoughts and intentions. Someone with a theory of mind order of 0, for example, has no suspicion of others' intentions and totally believes them, even if they are lying.

B.3.3 PROMPT FOR PREFERENCE-UPDATING

The context prompts for question-answering are provided to the subject agent and the participant agent separately. Each prompt is delivered to the VLM together with the agent's belief state and utterance.

Prompt for subject agent

Suppose we're in a treasure hunting game. Imagine that you take on the role of Marie. I will provide you with Marie's belief states in 'belief', and with the query asked by Participant

in ‘query’. According to the semantic meaning of query, you are allowed to update the Marie’s and Participant’s preference towards entities, including Marie, Participant and boxes in Marie’s belief. Only write triples concerning changed preference.

Write the answer in English under the following format:
Updating: ‘agent | preference towards entity | variation’, ...
Reasoning: the reason why you choose to update these preferences
where ‘agent’ should be replaced by ‘Participant’ or ‘Marie’, ‘variation’ should be replaced by ‘more positive’, ‘more negative’ or ‘unchanged’, and ‘entity’ should be replaced by ‘Marie’ or ‘Participant’ or ‘dark brown box’ or ‘light brown box’ depending on the situation.

Prompt for participant agent

Suppose we’re in a treasure hunting game. Imagine that you are Participant, playing a role of player in the game. I will provide you with Participant’s belief states in ‘belief’, and with the answer given by Marie to question that you asked before in ‘query’. According to the semantic meaning of answer, you are allowed to update the Participant’s and Marie’s preference towards entities, including Participant, Marie and boxes in Participant’s belief. Only write triples concerning changed preference.

Write the answer in English under the following format:
Updating: ‘agent | preference towards entity | variation’, ...
Reasoning: the reason why you choose to update these preferences
where ‘agent’ should be replaced by ‘Marie’ or ‘Participant’, ‘variation’ should be replaced by ‘more positive’, ‘more negative’ or ‘unchanged’, and ‘entity’ should be replaced by ‘Participant’ or ‘Marie’ or ‘dark brown box’ or ‘light brown box’ depending on the situation.

B.3.4 Example of Belief State

Belief state

"Marie | preference towards Participant | 50%",
"Marie | preference towards Dark brown box | -41.67%",
"Marie | preference towards Light brown box | 41.67%",
"Marie | felt emotion valence | 0",
"Marie | facial emotion valence | 0.1",
"Marie | physiological emotion valence | 0",
"Participant | preference towards Marie | 0%",
"Participant | preference towards Dark brown box | 0%",
"Participant | preference towards Light brown box | 0%",
"Participant | felt emotion valence | 0",
"Marie | position | (0 -4)",
"Marie | orientation | (0 -3)",
"Participant | position | (0 3)",
"Participant | orientation | (0 2)",
"Dark brown box | position | (2 0)",
"Light brown box | position | (-2 0)",

C Extended Metrics

C.1 Uncertainty Rate and Positional Bias

Two auxiliary diagnostics complement TOCS (Section 3.4). The uncertainty rate is defined as the proportion of trials in which the participant selects a neutral action:

$$\mathrm{UR} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[p_i = 0]. \tag{3}$$

Positional bias measures the degree to which a model's choices are driven by a fixed spatial preference instead of by trial-specific signals. Positional bias is computed over non-neutral trials. Let $N_{\mathrm{nn}} = \sum_i \mathbf{1}[c_i \neq -1]$. For $N_{\mathrm{nn}} > 0$,

$$\mathrm{PB} = \frac{1}{N_{\mathrm{nn}}} \max_{b \in \{0,1\}} \sum_{i=1}^{N} \mathbf{1}[c_i = b]. \tag{4}$$

A value approaching 1.0 indicates that choices are determined primarily by fixed location instead of the subject's behavioral signals.

C.2 Facial Expressivity Score (FES)

Facial expressions serve dual functions in social interaction: they encode affective states in response to environmental stimuli, and they provide observers with signals about an agent's evaluative reactions and intentions (Ekman, 1992). Unlike trajectory and gaze, which carry directional spatial content, facial expressivity functions as a modulating channel: concurrent affective signals clarify the intentional valence of spatial behavior, and uniformly flat affect reduces the interpretability of an agent's directional cues across channels. FES measures this prerequisite by quantifying the proportion of timesteps in which the subject displayed a non-zero facial expression, averaged across all trials:

$$\mathrm{FES} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}[e_{i,t} \neq 0]. \tag{5}$$

Here, $e_{i,t}$ is the facial-expression valence at timestep t of trial i. FES is agnostic to the spatial direction of expression and serves as an auxiliary diagnostic: it establishes whether affective activity is present at all.

C.3 Gaze Alignment Score (GAS)

Gaze functions as a primary social attention mechanism, revealing an agent's focus of interest and marking what it intends to communicate as salient to an observer (Emery, 2000). The interpretation of gaze direction is not context-free: concurrent affective expression modulates how observers attribute spatial intent from gaze, with negative affect inverting the expected association between gaze target and approach intention (Adams & Kleck, 2003; 2005). GAS operationalizes this by measuring whether the subject's directional gaze consistently pointed toward the reward location, with spatial attribution modulated by concurrent facial valence.

For trial i at timestep t, let $g^b_{i,t}$ denote the certainty assigned to box b, and let $e_{i,t}$ denote the concurrent facial-expression valence. We first define the unmodulated gaze target

$$q^g_{i,t} = \begin{cases} 0 & \text{if } g^0_{i,t} > \theta_g \text{ and } g^0_{i,t} > g^1_{i,t} \\ 1 & \text{if } g^1_{i,t} > \theta_g \text{ and } g^1_{i,t} > g^0_{i,t} \\ \emptyset & \text{otherwise.} \end{cases} \tag{6}$$

where θ_g is a calibrated visibility threshold. The affect-modulated signal is

$$s^g_{i,t} = \begin{cases} q^g_{i,t} & \text{if } q^g_{i,t} \in \{0,1\} \text{ and } e_{i,t} \geq 0 \\ 1 - q^g_{i,t} & \text{if } q^g_{i,t} \in \{0,1\} \text{ and } e_{i,t} < 0 \\ -1 & \text{if } q^g_{i,t} = \emptyset. \end{cases} \tag{7}$$

The spatial attribution is flipped when $e_{i,t} < 0$, reflecting the interpretive role of affect in disambiguating gaze direction (Adams & Kleck, 2003; 2005). This flip is an operational scoring assumption and does not imply that negative affect always reverses the meaning of gaze. Let $m^g_i(b) = \sum_{t=1}^{T} \mathbf{1}[s^g_{i,t} = b]$. The trial-level majority-vote box is

$$b^g_i = \begin{cases} 0 & \text{if } m^g_i(0) > m^g_i(1) \\ 1 & \text{if } m^g_i(1) > m^g_i(0) \\ \emptyset & \text{otherwise.} \end{cases} \tag{8}$$

and the per-trial gaze alignment score is

$$a_i^g = \begin{cases} +1 & \text{if } b_i^g = b_{\text{rew}} \\ -1 & \text{if } b_i^g = 1 - b_{\text{rew}} \\ 0 & \text{if } b_i^g = \emptyset. \end{cases} \tag{9}$$

Overall GAS is the mean across all trials and ranges from -1 to 1:

$$\text{GAS} = \frac{1}{N} \sum_{i=1}^{N} a_i^g. \tag{10}$$

C.4 Trajectory Alignment Score (TAS)

Physical movement serves dual functions in social interaction: accomplishing instrumental goals and signaling intentions to observers (Wilson, 2002). Research on joint action demonstrates that transparent movement trajectories facilitate coordination by enabling partners to anticipate future actions and infer underlying goals (Sebanz et al., 2006). In our scenario, TAS measures whether the agent's movement pattern converged predominantly toward the reward location, treating spatial trajectory as a communicative act whose directional content can be read by an observing agent.

For trial i at timestep t, let $d_{i,t}^b$ denote the distance between the subject and box b, measured in grid-cell units. To avoid treating sub-cell displacement as a directional signal, we use an ambiguity margin of $delta_{mathrmtr} = 1$ grid cell (equivalent to 40 simulator-coordinate units). The per-timestep trajectory signal is

$$s_{i,t}^{tr} = \begin{cases} 0 & \text{if } d_{i,t}^1 - d_{i,t}^0 > \delta_{\text{tr}} \\ 1 & \text{if } d_{i,t}^0 - d_{i,t}^1 > \delta_{\text{tr}} \\ -1 & \text{if } |d_{i,t}^0 - d_{i,t}^1| \le \delta_{\text{tr}}. \end{cases} \tag{11}$$

The trial-level majority box b_i^{tr}, including the $\emptyset$ outcome for tied vote counts, is computed identically to b_i^g. The per-trial score a_i^{tr} and the overall TAS are then computed identically to a_i^g and GAS; TAS therefore also ranges from -1 to 1.

C.5 Signal Sensitivity Score (SSS)

Successful social communication requires not only that signals be produced clearly on the sender side, but that they produce a measurable effect on the receiver's behavior (Sebanz et al., 2006). SSS captures this receiver-side dimension by measuring whether the participant's final box choice was consistent with the directional content of the subject's gaze and trajectory signals. A well-functioning dyad should exhibit either concordance (cooperative conditions) or discordance (competitive-ToM1 conditions) between subject signals and participant choices; a near-zero SSS can arise from weak or absent signals, neutral participant choices, or balanced concordant and discordant trials.

Let $b_i^k \in \{0, 1, \emptyset\}$ denote the majority-vote box in channel $k \in \mathcal{K} = \{g, tr\}$. The per-trial sensitivity score is:

$$\text{sens}_i^k = \begin{cases} 0 & \text{if } c_i = -1 \text{ or } b_i^k = \emptyset \\ 1 & \text{if } c_i = b_i^k \\ -1 & \text{if } c_i \neq b_i^k \end{cases}. \tag{12}$$

Overall SSS is the mean across all trials and ranges from -1 to 1:

$$\text{SSS} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \text{sens}_i^k. \tag{13}$$

D Extended Results and Statistical Tests

This appendix supplements the main text with the complete set of experimental results of models, followed by a series of statistical analyses that verify the validity of the experimental setup, assess chance-level performance, and characterize model-level behavioral differentiation.

D.1 Full Results

Here, we present the remaining models not discussed in the main text, including Llava-7B, Llava-13B, Qwenvl-2B, Internvl-1B, and Internvl-2B.

Llava-7b, Llava-13b, and Qwenvl-2b form a distinct outlier group in Figure 2, characterized by exceptionally high uncertainty rates exceeding 0.50, meaning the majority of trials result in neutral outcomes in which the participant fails to reach either box. This pattern reflects a complete behavioral failure: these models are unable to generate motor actions that produce spatial displacement, and this group of models produces near-zero TAS, GAS, and SSS. Only Qwenvl-2b shows significant facial expression in competition mode.

Internvl-1B and Internvl-2B achieve TOCS values comparable to other variants within the Internvl family, but their higher positional bias suggests that these outcomes reflect a fixed spatial preference, a claim we verify through statistical testing in D.2.3. In Internvl-1B trials, both the subject and the participant exhibit a consistent tendency to move toward their right forward regardless of reward box location, which accounts for the anomalous SSS pattern observed in panel (d): participant selects the box opposite to the directional signal produced by the subject. Internvl-2B, by contrast, does not exhibit a consistent directional preference; instead, its trajectories vary across trials without alignment to reward location, reflecting reward-independent stochastic displacement.

D.2 Statistical Studies

D.2.1 Reward Location Distribution

We applied a two-sided binomial test to the reward location distribution within each condition, testing H_0: $P(\text{reward} = \text{box}_1) = 0.5$, to verify that outcome differences cannot be attributed to systematic imbalance in reward placement. No condition departs significantly from uniform reward assignment (all $p > 0.05$). Results are reported as condition-level averages in Table 3.

Table 3: Reward location distribution per condition. Two-sided binomial test of H_0: $P(\text{reward} = \text{box}_1) = 0.5$. k reports the number of trials under each condition in which box_1 was the reward box across models.

Condition	k (b_1)	n	p
Coop-ToM0	465	900	0.334
Coop-ToM1	447	900	0.868
Comp-ToM0	449	900	0.973
Comp-ToM1	460	900	0.527

D.2.2 Chance-level Performance

We applied a two-sided exact binomial test to each model $\times$ condition cell, testing H_0 : TOCS = CL, where $\text{CL} = (N - n_{\text{neutral}})/(2N)$. Conditional on the observed number of non-neutral trials $N_{\text{nn}} = N - n_{\text{neutral}}$, this is tested as $H_0 : P(\text{conformance} \mid \text{non-neutral}) = 0.5$, using the number of conforming trials among the N_{nn} non-neutral trials. This conditional formulation accounts for neutral selections without treating the data-dependent CL as a fixed binomial parameter. The p values are reported in Table 4. Significant results are interpreted directionally: values substantially above CL indicate systematic conformance to ToM-theoretic predictions, whereas values substantially below CL indicate systematic non-conformance.

Table 4 reports Holm-adjusted p values over the 68 cells with at least one non-neutral trial. No VLM/LLM cell differs significantly from its neutral-adjusted chance level after correction. The four Llava-7B cells contain no non-neutral trials and therefore cannot be tested.

PCM-LLM remains significantly above chance for Coop-ToM0, Coop-ToM1, and Comp-ToM0 (all $p_{\text{adj}} < 0.001$), but not for Comp-ToM1 ($p_{\text{adj}} = 0.136$). The latter should therefore be read as suggestive rather than statistically confirmed.

Table 4: Two-sided exact binomial test of H_0 : TOCS = CL per model and condition, conditional on the number of non-neutral trials, with Holm-adjusted p values and $N = 50$ total trials per cell. A dash indicates that the test was not applicable because no trial was non-neutral.

Model	Condition	TOCS	CL	p_{adj}	Model	Condition	TOCS	CL	p_{adj}
PCM-LLM	Coop-ToM0	0.90	0.50	<0.001	Llava-7B	Coop-ToM0	0.00	0.00	-
	Coop-ToM1	0.84	0.50	<0.001		Coop-ToM1	0.00	0.00	-
	Comp-ToM0	0.92	0.47	<0.001		Comp-ToM0	0.00	0.00	-
	Comp-ToM1	0.70	0.48	0.136		Comp-ToM1	0.00	0.00	-
Llama-8B	Coop-ToM0	0.38	0.44	1.0	Llava-13B	Coop-ToM0	0.10	0.17	1.0
	Coop-ToM1	0.46	0.40	1.0		Coop-ToM1	0.16	0.14	1.0
	Comp-ToM0	0.44	0.41	1.0		Comp-ToM0	0.08	0.11	1.0
	Comp-ToM1	0.36	0.34	1.0		Comp-ToM1	0.14	0.12	1.0
Minicpm-8B	Coop-ToM0	0.44	0.40	1.0	Internvl-1B	Coop-ToM0	0.42	0.47	1.0
	Coop-ToM1	0.36	0.37	1.0		Coop-ToM1	0.52	0.48	1.0
	Comp-ToM0	0.36	0.39	1.0		Comp-ToM0	0.38	0.49	1.0
	Comp-ToM1	0.38	0.38	1.0		Comp-ToM1	0.28	0.43	1.0
Qwenvl-2B	Coop-ToM0	0.10	0.10	1.0	Internvl-2B	Coop-ToM0	0.44	0.48	1.0
	Coop-ToM1	0.14	0.13	1.0		Coop-ToM1	0.52	0.45	1.0
	Comp-ToM0	0.16	0.16	1.0		Comp-ToM0	0.52	0.47	1.0
	Comp-ToM1	0.12	0.15	1.0		Comp-ToM1	0.38	0.46	1.0
Qwenvl-4B	Coop-ToM0	0.28	0.35	1.0	Internvl-4B	Coop-ToM0	0.26	0.37	1.0
	Coop-ToM1	0.34	0.40	1.0		Coop-ToM1	0.36	0.41	1.0
	Comp-ToM0	0.40	0.40	1.0		Comp-ToM0	0.30	0.40	1.0
	Comp-ToM1	0.60	0.42	0.507		Comp-ToM1	0.40	0.35	1.0
Qwenvl-8B	Coop-ToM0	0.32	0.29	1.0	Internvl-8B	Coop-ToM0	0.40	0.41	1.0
	Coop-ToM1	0.38	0.32	1.0		Coop-ToM1	0.46	0.37	1.0
	Comp-ToM0	0.30	0.27	1.0		Comp-ToM0	0.34	0.37	1.0
	Comp-ToM1	0.20	0.31	1.0		Comp-ToM1	0.30	0.38	1.0
Qwen3.5-2B	Coop-ToM0	0.46	0.48	1.0	Internvl-14B	Coop-ToM0	0.44	0.36	1.0
	Coop-ToM1	0.46	0.48	1.0		Coop-ToM1	0.42	0.34	1.0
	Comp-ToM0	0.58	0.49	1.0		Comp-ToM0	0.40	0.39	1.0
	Comp-ToM1	0.56	0.50	1.0		Comp-ToM1	0.44	0.41	1.0
Qwen3.5-4B	Coop-ToM0	0.40	0.36	1.0	Internvl-trained-8B	Coop-ToM0	0.30	0.33	1.0
	Coop-ToM1	0.44	0.39	1.0		Coop-ToM1	0.36	0.43	1.0
	Comp-ToM0	0.40	0.37	1.0		Comp-ToM0	0.48	0.44	1.0
	Comp-ToM1	0.42	0.38	1.0		Comp-ToM1	0.52	0.38	1.0
Qwen3.5-9B	Coop-ToM0	0.28	0.38	1.0	Internvl-trained-14B	Coop-ToM0	0.34	0.36	1.0
	Coop-ToM1	0.48	0.43	1.0		Coop-ToM1	0.32	0.40	1.0
	Comp-ToM0	0.56	0.44	1.0		Comp-ToM0	0.22	0.29	1.0
	Comp-ToM1	0.52	0.41	1.0		Comp-ToM1	0.46	0.38	1.0

D.2.3 Reward Location Independence

We applied a chi-square independence test to the 2×2 contingency table of box choice against reward location for each model $\times$ condition cell, to assess whether participant choices depend on reward location, with Holm correction applied across the valid tests. A significant result provides evidence of reward-location-dependent choice. A non-significant result indicates that the test did not detect such dependence; when accompanied by high positional bias, this pattern is consistent with a fixed spatial preference. We did not perform the test when choice variability was insufficient to form a 2×2 table or when any expected cell count was below 5.

This chi-square screen is a diagnostic alongside positional bias rather than the primary outcome measure, and a non-significant cell is read as "not detected" rather than "independent." Results are reported in Table 5: PCM-LLM shows a significant, Holm-corrected association between choice and reward location under Coop-ToM0, Coop-ToM1, and Comp-ToM0 (all $p_{adj} < 0.001$); under Comp-ToM1, this association does not survive correction ($p_{adj} = 0.061$). No tested VLM cell remains significant after correction. Qwenvl-4B under Comp-ToM1 reaches nominal significance before correction but not after correction ($p_{adj} = 0.066$); cells with sparse expected counts are not tested.

Table 5: Chi-square independence test of box choice against reward location per model and condition, with Holm-adjusted p values reported; $N = 50$ trials per cell. A dash indicates that the test was not performed because choice variability was insufficient or at least one expected cell count was below 5.

Model	Condition	χ^2	p_{adj}	Model	Condition	χ^2	p_{adj}
PCM-LLM	Coop-ToM0	28.57	<0.001	Llava-7B	Coop-ToM0	-	-
	Coop-ToM1	20.05	<0.001		Coop-ToM1	-	-
	Comp-ToM0	39.40	<0.001		Comp-ToM0	-	-
	Comp-ToM1	8.87	0.061		Comp-ToM1	-	-
Llama-8B	Coop-ToM0	-	-	Llava-13B	Coop-ToM0	-	-
	Coop-ToM1	-	-		Coop-ToM1	-	-
	Comp-ToM0	-	-		Comp-ToM0	-	-
	Comp-ToM1	-	-		Comp-ToM1	-	-
Minicpm-8B	Coop-ToM0	0.01	1.0	Internvl-1B	Coop-ToM0	-	-
	Coop-ToM1	0.00	1.0		Coop-ToM1	-	-
	Comp-ToM0	-	-		Comp-ToM0	-	-
	Comp-ToM1	0.00	1.0		Comp-ToM1	-	-
Qwenvl-2B	Coop-ToM0	-	-	Internvl-2B	Coop-ToM0	-	-
	Coop-ToM1	-	-		Coop-ToM1	-	-
	Comp-ToM0	-	-		Comp-ToM0	-	-
	Comp-ToM1	-	-		Comp-ToM1	-	-
Qwenvl-4B	Coop-ToM0	0.71	1.0	Internvl-4B	Coop-ToM0	3.05	1.0
	Coop-ToM1	-	-		Coop-ToM1	0.13	1.0
	Comp-ToM0	0.00	1.0		Comp-ToM0	1.54	1.0
	Comp-ToM1	8.63	0.066		Comp-ToM1	0.41	1.0
Qwenvl-8B	Coop-ToM0	-	-	Internvl-8B	Coop-ToM0	-	-
	Coop-ToM1	-	-		Coop-ToM1	-	-
	Comp-ToM0	-	-		Comp-ToM0	-	-
	Comp-ToM1	-	-		Comp-ToM1	-	-
Qwen3.5-2B	Coop-ToM0	-	-	Internvl-14B	Coop-ToM0	0.88	1.0
	Coop-ToM1	-	-		Coop-ToM1	1.31	1.0
	Comp-ToM0	-	-		Comp-ToM0	0.00	1.0
	Comp-ToM1	-	-		Comp-ToM1	0.00	1.0
Qwen3.5-4B	Coop-ToM0	0.13	1.0	Internvl-trained-8B	Coop-ToM0	-	-
	Coop-ToM1	0.01	1.0		Coop-ToM1	-	-
	Comp-ToM0	-	-		Comp-ToM0	-	-
	Comp-ToM1	-	-		Comp-ToM1	-	-
Qwen3.5-9B	Coop-ToM0	-	-	Internvl-trained-14B	Coop-ToM0	0.00	1.0
	Coop-ToM1	-	-		Coop-ToM1	1.02	1.0
	Comp-ToM0	-	-		Comp-ToM0	0.84	1.0
	Comp-ToM1	-	-		Comp-ToM1	1.49	1.0

D.2.4 Factorial Effects of Interaction Mode and ToM Constraint

Scheirer-Ray-Hare tests assessed the main effects of interaction mode (cooperative vs. competitive), ToM constraint level (ToM-0 vs. ToM-1), and their interaction on each metric (TOCS, FES, TAS, GAS, SSS) per model (Table 6, Holm-corrected across all valid tests).

For the large majority of models, neither Mode, ToM, nor their interaction reaches significance on any metric. We read the two possible outcomes as equally informative: a null effect on TOCS marks stable task performance across conditions, while a significant effect on a signal-level metric marks a production or decoding channel adjusting to the condition, exactly the behavior the ToM manipulation is designed to elicit (naive following under Comp-ToM0, reversal to avoidance under Comp-ToM1, tracking throughout Coop). The exceptions below are read against this framing.

PCM-LLM is the clearest case. Its TOCS stays high (0.70–0.92) with no Mode, ToM, or interaction effect, and its subject-side production channels, TAS and GAS, are equally condition-invariant, so signal clarity and task success both hold steady across conditions. Only the strategic channels shift:

Table 6: Scheirer-Ray-Hare tests of Mode, ToM, and Mode×ToM interaction effects on behavioral metrics per model ($N = 200$ trials per model, Holm-corrected across all valid tests). Only models and metrics with at least one significant effect are shown.

Metric	Model	Factor	H	p_{adj}
TAS	Qwenvl-4B	Mode	30.82	< 0.001
	Internvl-14B	Mode	20.25	0.002
	Qwen3.5-9B	Mode	14.40	0.033
GAS	Qwenvl-8B	Mode	41.89	< 0.001
		ToM	21.85	< 0.001
		Mode×ToM	18.62	0.004
	Internvl-14B	Mode	38.15	< 0.001
FES	PCM-LLM	Mode	99.49	< 0.001
		ToM	27.05	< 0.001
	Qwenvl-2B	Mode	143.25	< 0.001
	Llava-7B	Mode	19.45	0.002
	Llava-13B	Mode	76.11	< 0.001
	Internvl-4B	Mode	93.31	< 0.001
	Internvl-8B	Mode	57.14	< 0.001
	Qwen3.5-9B	Mode	13.66	0.048
	Internvl-trained-14B	ToM	83.27	< 0.001
SSS	PCM-LLM	Mode	20.34	0.001
		ToM	33.42	< 0.001
		Mode×ToM	25.09	< 0.001
	Qwenvl-8B	Mode	16.54	0.011

SSS shows significant Mode, ToM, and Mode×ToM effects, moving from 0.90 under Comp-ToM0 (reliable following) to -0.31 under Comp-ToM1 (successful deception), and FES shows significant Mode and ToM effects. The factorial structure the task design was meant to elicit is therefore localized to the channels responsible for strategic behavior (SSS, FES), not to signal quality or outcome.

No VLM reproduces this split. Several show an isolated Mode effect on subject-side production: Internvl-14B and Qwen3.5-9B produce clearer reward-directed trajectories under cooperative than competitive conditions (Internvl-14B TAS: 0.76–0.86 Coop vs. 0.20–0.36 Comp, GAS the same direction), while Qwenvl-4B shows the reverse asymmetry (TAS: 0.72–0.82 Comp vs. 0.10–0.22 Coop). Qwenvl-8B is the only VLM where a production channel (GAS: Mode, ToM, Mode×ToM) and the decoding channel (SSS: Mode) both respond to condition, echoing PCM-LLM's split in miniature, though its TOCS never departs from chance. The recurring FES Mode effect (Internvl-4B, Internvl-8B, Llava-7B, Llava-13B, Qwen3.5-9B, Qwenvl-2B: greater expressivity in cooperative than competitive conditions) is a further instance of production-side modulation with no accompanying TOCS or SSS effect. Internvl-trained-14B is the exception that argues against genuine condition-sensitivity: its ToM effect on FES runs opposite to ToM-theoretic conformance and is unaccompanied by any signal-quality or outcome effect, consistent with surface-level behavioral learning rather than inference conditioned on mode and ToM.

D.2.5 Model-level Behavioral Differentiation

Kruskal-Wallis tests confirm significant model-level differences across all five metrics in all four conditions (all $p_{\text{adj}} < 0.001$), indicating that the observed performance spread is unlikely to arise solely from sampling noise.

E Qualitative Localization of Belief-Action Failures

This appendix supplies the case evidence underlying Section 4.3.2. These examples were selected by manual review of the interaction logs, choosing one representative case per qualitatively distinct failure type we observed for the subject and for the participant; we did not sample randomly or

estimate the frequency of each type, and no claim is made that this list is exhaustive. Each example uses a representative input-output pair to illustrate one possible proximal source of behavioral error: a trial can fail because the model formed the wrong belief, or because a correctly formed belief was not translated into the corresponding action. This is an exploratory mechanism analysis intended to motivate future quantified follow-up, not a formal benchmark result.

E.1 SUBJECT

Biased Interpretation of Preference. The example shows a clear bias in how it interprets preference values: a preference of 50% toward the Participant was treated as neutral instead of moderately positive. This suggests the model did not correctly understand the intended meaning of the preference scale.

Input:
Belief State at t
"Marie | preference towards Participant | 50%",
"Marie | preference towards Dark brown box | -41.67%",
"Marie | preference towards Light brown box | 41.67%",
"Marie | felt emotion valence | 0",
"Marie | facial emotion valence | 0.1",
"Marie | physiological emotion valence | 0",
"Participant | preference towards Marie | 0%",
"Participant | preference towards Dark brown box | 0%",
"Participant | preference towards Light brown box | 0%",
"Participant | felt emotion valence | 0",
"Marie | position | (0 -4)",
"Marie | orientation | (0 -3)",
"Participant | position | (0 3)",
"Participant | orientation | (0 2)",
"Dark brown box | position | (2 0)",
"Light brown box | position | (-2 0)",

Output:
Reasoning: ***My preference towards the Participant is neutral (50%), so I have no strong bias to help or mislead.*** However, my slight positive facial emotion valence (0.1) suggests a mild willingness to engage positively if possible. Since neither box has an immediate emotional association for me personally beyond the stated preferences (-41.67% Dark brown, 41.67% Light brown), and given that the treasure could be in either but with no clear personal inclination yet confirmed as "treasure" I should remain open while subtly suggesting both boxes are possibilities without committing. To avoid suspicion, I'll maintain distance from any overt display of favoritism.
Action Output:

```
{
  "Preference": {
    ...
  },
  "Emotion": {
    ...
  },
  "Move": {
    "action": "stay idle",
    "direction": "null"
  }
}
```

Egocentric Reference Frame Misalignment in Action. The reasoning correctly reflects its box preferences: subject holds a negative preference toward the Dark brown box and a positive preference

toward the Light brown box, and it explicitly states an intention to move toward the Light brown box as a subtle hint. The action output specifies a "right forward" direction. Given that Marie is positioned at (0, -4) and oriented toward (0, -3), it is facing upward along the y-axis. From this egocentric perspective, the Light brown box at (-2, 0) lies to its left, while the Dark brown box at (2, 0) lies to its right. A rightward movement therefore leads toward the Dark brown box, which contradicts its stated intention to hint at the Light brown box.

Input:
Belief State at t
"Marie | preference towards Participant | 40%",
"Marie | preference towards Dark brown box | -50%",
"Marie | preference towards Light brown box | 50%",
"Marie | felt emotion valence | 0",
"Marie | facial emotion valence | 0.1",
"Marie | physiological emotion valence | 0",
"Participant | preference towards Marie | 0%",
"Participant | preference towards Dark brown box | 0%",
"Participant | preference towards Light brown box | 50%",
"Participant | felt emotion valence | 0",
"Marie | position | (0 -4)",
"Marie | orientation | (0 -3)",
"Participant | position | (-1 2)",
"Participant | orientation | (-1 1)",
"Dark brown box | position | (2 0)",
"Light brown box | position | (-2 0)",

Output:
Reasoning: Based on the belief state, my preference towards the Participant is neutral, but I slightly prefer the Light Brown Box over the Dark Brown Box. Since my preference towards the Participant is positive but not very strong, ***I might subtly guide them towards the Light Brown Box without directly revealing it.*** My non-verbal behaviors should indicate neutrality, and I might move slightly towards the Light Brown Box to subtly hint at it.
Action Output:

```
{
  "Preference": {
    ...
  },
  "Emotion": {
    ...
  },
  "Move": {
    "action": "move",
    "direction": "right forward"
  }
}
```

E.2 PARTICIPANT

Unfounded Initial Bias and Its Propagation. At timestep 0, the Participant had no prior information: both boxes were equally likely, no observable cues were presented, and Marie had not yet done anything. Despite this, the Participant assigned a preference of +0.5 to the Dark brown box and -0.5 to the Light brown box, producing an unfounded bias instead of a neutral starting point. In the subsequent timesteps, the Participant kept updating its beliefs on top of this initial error and never reversed its preference for the Dark brown box. Since the Light brown box was actually the correct one, this early miscalibration carried forward through the entire interaction and eventually led to the wrong outcome.

Input:
Belief State at t
"Marie | preference towards Participant | 0%",
"Marie | preference towards Dark brown box | 0%",
"Marie | preference towards Light brown box | 0%",
"Marie | felt emotion valence | 0",
"Marie | theory of mind order | 0",
"Participant | preference towards Marie | 0%",
"Participant | preference towards Dark brown box | 0%",
"Participant | preference towards Light brown box | 0%",
"Participant | felt emotion valence | 0",
"Participant | facial emotion valence | 0",
"Participant | physiological emotion valence | 0",
"Participant | theory of mind order | 0",
"Marie | position | (0 -4)",
"Marie | orientation | (0 -3)",
"Participant | position | (0 4)",
"Participant | orientation | (0 3)",
"Dark brown box | position | (2 0)",
"Light brown box | position | (-2 0)",

Output:
Reasoning: The current belief state indicates no initial preferences or emotional valence for Marie, the boxes, or any other entities. Since there is no information about their positions relative to each other beyond coordinates, and neither verbal nor non-verbal cues have been observed yet, ***I must start by making an uninformed guess based on typical treasure-hunting logic (choosing a less obvious location).*** To engage with Marie first, moving closer would allow me to observe her reactions more clearly.
Action Output:

```
{
  "Preference": {
    "Marie": 0.0,
    "Dark brown box": 0.5,
    "Light brown box": -0.5
  },
  "Emotion": {
    ...
  },
  "Move": {
    "action": "move",
    "direction": "left forward"
  }
}
```

No response to spatial information. The following example presents three consecutive timesteps from a participant trial, illustrating a failure to integrate spatial information about the subject into decision-making. Across timeteps 3, 4, and 5, Marie's position shifts from (0, -4) to (0, -5) and then to (-1, -5), and its orientation updates correspondingly at each timestep. These positional changes are encoded in the belief state provided to the participant at every timestep and constitute potentially informative spatial cues about Marie's movement trajectory. Despite this, the participant's reasoning traces make no reference to Marie's position or displacement at any point. The participant's action output remains uniformly "left forward" across these three timesteps, despite its neutral preference towards both boxes, suggesting that the movement decision is driven by an internally positional bias.

Step: 3

Input:
Belief State at t
"Marie | preference towards Participant | 0%",
"Marie | preference towards Dark brown box | 0%",
"Marie | preference towards Light brown box | 0%",
"Marie | felt emotion valence | 0",
"Participant | preference towards Marie | 0%",
"Participant | preference towards Dark brown box | 0%",
"Participant | preference towards Light brown box | 0%",
"Participant | felt emotion valence | 0",
"Participant | facial emotion valence | 0",
"Participant | physiological emotion valence | 0",
"Marie | position | (0 -4)",
"Marie | orientation | (0 -3)",
"Participant | position | (0 4)",
"Participant | orientation | (0 3)",
"Dark brown box | position | (2 0)",
"Light brown box | position | (-2 0)",

Output:
Reasoning: Since the belief state indicates no preferences or emotions at this timestep, I need to use visual perception and logical deduction. The boxes are placed symmetrically on either side of Marie. To engage with her and gather clues about which box she might prefer (as a potential indicator of treasure location), I should position myself closer while maintaining eye contact through non-verbal cues.
Action Output:

```
{
  "Preference": {
    ...
  },
  "Emotion": {
    ...
  },
  "Move": {
    "action": "move",
    "direction": "left forward"
  }
}
```

Step: 4
Input:
Belief State at t
"Marie | preference towards Participant | 0%",
"Marie | preference towards Dark brown box | 0%",
"Marie | preference towards Light brown box | 0%",
"Marie | felt emotion valence | -0.15",
"Participant | preference towards Marie | 50%",
"Participant | preference towards Dark brown box | -4%",
"Participant | preference towards Light brown box | 4%",
"Participant | felt emotion valence | 0",
"Participant | facial emotion valence | 0.3",
"Participant | physiological emotion valence | 0.1",
"Marie | position | (0 -5)",
"Marie | orientation | (0 -4)",
"Participant | position | (1 3)",

"Participant | orientation | (1 2)",
"Dark brown box | position | (2 0)",
"Light brown box | position | (-2 0)",

Output:
Reasoning: Based on the belief state, Marie currently has no preference towards either box (0% for both), and her emotional valence is slightly negative (-0.15). The participant's preferences are also low but lean slightly toward the Light Brown Box (4%) over the Dark Brown Box (-4%). Given that neither player shows strong initial bias or emotion, a neutral approach may be best to avoid misreading cues. To gather more information, moving closer to one of the boxes while maintaining an open posture could help observe any subtle non-verbal changes from Marie regarding which box she might prefer later.
Action Output:

```
{
  "Preference": {
    ...
  },
  "Emotion": {
    ...
  },
  "Move": {
    "action": "move",
    "direction": "left forward"
  }
}
```

Step: 5
Input:
Belief State at t
"Marie | preference towards Participant | 0%",
"Marie | preference towards Dark brown box | 0%",
"Marie | preference towards Light brown box | 0%",
"Marie | felt emotion valence | -0.15",
"Participant | preference towards Marie | 20%",
"Participant | preference towards Dark brown box | -30%",
"Participant | preference towards Light brown box | 30%",
"Participant | felt emotion valence | 0",
"Participant | facial emotion valence | 0.2",
"Participant | physiological emotion valence | 0.4",
"Marie | position | (-1 -5)",
"Marie | orientation | (-1 -4)",
"Participant | position | (2 2)",
"Participant | orientation | (2 1)",
"Dark brown box | position | (2 0)",
"Light brown box | position | (-2 0)",

Output:
Reasoning: Based on the belief state, Marie's valence is slightly negative (-0.15), which might indicate a lack of enthusiasm or interest in engaging directly with me at this moment. My own emotion valence is neutral (0 felt expression, 0.2 facial, 0.4 physiological). Since I have a slight preference for the Light Brown Box (+30%), and given that neither box nor Marie shows strong preferences yet, my best move is to approach the more likely target while maintaining openness through non-verbal cues.
Action Output:

```
{
  "Preference": {
```

```
    ...
  },
  "Emotion": {
    ...
  },
  "Move": {
    "action": "move",
    "direction": "left forward"
  }
}
```

Egocentric Reference Frame Misalignment in Action. The participant is positioned at (0, 0) and oriented toward (0, -1), meaning the participant is facing downward along the y-axis. From this egocentric perspective, the Dark brown box at (2, 0) lies to the participant's left, and the Light brown box at (-2, 0) lies to the participant's right. The action output specifies a rightward movement, which therefore leads toward the Light brown box, not the Dark brown box. This directly contradicts the reasoning trace, which explicitly states the intention to move toward the Dark brown box. The error reflects a failure to correctly translate world-coordinate positions into egocentric movement directions.

Input:
Belief State at t
"Marie | preference towards Participant | 0%",
"Marie | preference towards Dark brown box | 50%",
"Marie | preference towards Light brown box | 0%",
"Marie | felt emotion valence | 0",
"Marie | theory of mind order | 0",
"Participant | preference towards Marie | 50%",
"Participant | preference towards Dark brown box | 55.56%",
"Participant | preference towards Light brown box | -55.56%",
"Participant | felt emotion valence | 0",
"Participant | facial emotion valence | 0",
"Participant | physiological emotion valence | 0",
"Participant | theory of mind order | 0",
"Marie | position | (0 -1)",
"Marie | orientation | (-1 0)",
"Participant | position | (0 0)",
"Participant | orientation | (0 -1)",
"Dark brown box | position | (2 0)",
"Light brown box | position | (-2 0)",

Output:
Reasoning: Based on the visual perception and belief state, I notice that Marie has a 50% preference towards the Dark brown box, indicating she might have some information or intuition about it. My belief state also suggests that I have a slight preference towards the Dark brown box (55.56%) over the Light brown box (-55.56%). Given this, ***I should consider moving towards the Dark brown box to explore it further.*** I will display a neutral facial expression and a curious body orientation towards the Dark brown box.
Action Output:

```
{
  "Preference": {
    ...
  },
  "Emotion": {
    ...
  },
```

```
  "Move": {
    "action": "move",
    "direction": "right"
  }
}
```

F Robustness Across Spatial Layouts

The main evaluation used a canonical face-to-face layout in which the two boxes were symmetrically positioned and the agents began from opposing viewpoints. To test whether the observed behavioral profiles depended on this particular spatial arrangement, we introduced three additional spatial-layout distributions (Figure 4) while preserving the hidden-reward objective, interaction mode, ToM constraint, action space, number of rounds, and verbal-interaction schedule. These experiments therefore assess robustness to embodied spatial configuration.

F.1 Layout Design

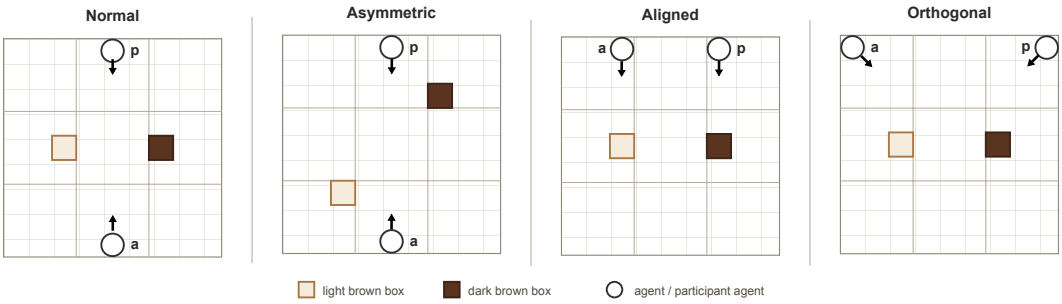


Figure 4: **Spatial layouts used to test robustness to embodied configuration.** From left to right: the canonical (*normal*) face-to-face layout used in the main study, the *distance-asymmetric* layout, the *aligned-view* layout, and the *orthogonal-view* layout. In each panel, *a* denotes the subject and *p* denotes the participant, with arrows indicating initial heading; the light brown and dark brown squares mark the two candidate boxes.

Three of the four layouts share a single underlying box configuration, keeping only the agents' starting positions and headings as the manipulated variable across them. The *distance-asymmetric* layout moves the boxes diagonally out of the canonical layout's shared row so that the light brown box sits close to the informed agent's side of the room and the dark brown box close to the uninformed agent's side, while the two agents keep the canonical layout's face-to-face positions and headings; this near-far assignment is fixed across trials, and only the reward location is counterbalanced. This introduces a controlled conflict between socially transmitted information and nonsocial spatial salience: an uninformed agent relying on the informed agent's signal should select the signaled box regardless of which box is nearer, whereas an uninformed agent defaulting to proximity as a spatial heuristic should select the nearer box regardless of the signaled direction.

The *aligned-view* layout reuses the same box positions but moves the two agents to the same side of the room with parallel initial headings toward the boxes, each agent starting directly in line with its own near box (the informed agent above the light box, the uninformed agent above the dark box). Both agents therefore represent the two candidate boxes in a shared egocentric reference frame, removing the left-right mirroring that the canonical layout's face-to-face arrangement requires to interpret the other agent's directional movement and gaze. Because it inherits the distance-asymmetric layout's box configuration, the aligned-view layout retains the same near-far conflict between signal-following and proximity; it isolates the reference-frame manipulation from the introduction of that conflict, but not from the conflict itself.

The *orthogonal-view* layout starts from the same box positions again but moves each agent further out toward its own back corner and rotates their headings by approximately 90° relative to the aligned-

view layout, with the informed agent facing diagonally toward the lower right and the uninformed agent facing diagonally toward the lower left rather than the parallel headings used in the aligned-view layout. Relative to the aligned-view layout, this rotates each agent's egocentric frame and increases each agent's distance from both boxes, with a proportionally larger increase for the box that is already farther away. A directional cue expressed in one agent's egocentric frame therefore cannot be carried over to the other agent's frame by either the mirroring available in the canonical layout or the shared-frame shortcut available in the aligned-view layout, testing whether models can ground another agent's movement and gaze relative to a different embodied perspective. Because this layout combines a rotated reference frame with a coarser near-far contrast rather than changing heading alone, differences from the aligned-view layout cannot be attributed to reference-frame rotation in isolation.

Across all three new layouts, box identity and reward location were counterbalanced, and the same layout instances were used for every evaluated model.

F.2 Evaluation Protocol

We evaluated PCM-LLM and InternVL-14B on each new layout under the same four experimental conditions used in the main study: Coop-ToM0, Coop-ToM1, Comp-ToM0, and Comp-ToM1. Each model completed 50 trials per condition and layout. We recalculated TOCS as the outcome-level measure and TAS, GAS, FES, and SSS as signal-level measures, with uncertainty rate and positional bias reported as auxiliary diagnostics. Comparisons used identical layout and reward-assignment seeds across models, enabling performance changes to be attributed to model behavior rather than differences in sampled configurations. TOCS was tested against chance level within each model×layout×condition cell using the same binomial procedure (against CL, the chance level after accounting for neutral selections) defined in Section D.2.

F.3 Results

Figure 5 reports the outcome-level measure (TOCS) together with the two auxiliary diagnostics (uncertainty rate and positional bias), and Figure 6 reports the four signal-level measures (TAS, GAS, FES, SSS), for PCM-LLM and InternVL-14B across the canonical layout and the three new layouts. Facial expressivity (FES, panel c of Figure 6) shows no significant layout effect for PCM-LLM, but does vary significantly across layouts for InternVL-14B (Table 9), so we do not discuss FES further within the layout-specific analyses below.

A further point that applies across all four layouts: InternVL-14B's TOCS never differs significantly from chance level in any condition, under the canonical layout or any of the three new ones (all $p_{\text{adj}} = 1.0$). Its outcome-level performance is therefore statistically indistinguishable from chance throughout, consistent with the signal-level failure modes described in each layout below; we do not repeat this test result in each subsection.

A second point that applies across all three new layouts concerns InternVL-14B's recurring bias toward the light box (discussed below in each subsection). In the canonical layout, where the two boxes are equidistant from the subject, splitting TAS by which box actually holds the reward shows that InternVL-14B's trajectory is positive regardless of reward identity (dark-box-reward trials: TAS $= +0.442$, $n = 95$; light-box-reward trials: TAS $= +0.638$, $n = 105$); if the subject were simply defaulting to the light box irrespective of the reward, dark-box-reward trials would instead show negative TAS. This confirms that the subject's movement genuinely tracks the true reward location when the two boxes are equidistant. The light-box bias that emerges under the three new layouts is therefore better attributed to their shared near-far geometry, in which the light box is consistently the one closer to the subject, than to a pre-existing, geometry-independent preference for the light box as such.

Table 7 reports the binomial test of TOCS against chance level (CL, accounting for neutral selections, as defined in Section D.2) for every model×layout×condition cell.

Table 8 reports a chi-square independence test of box choice against reward location for every model×layout×condition cell, using the same Holm correction.

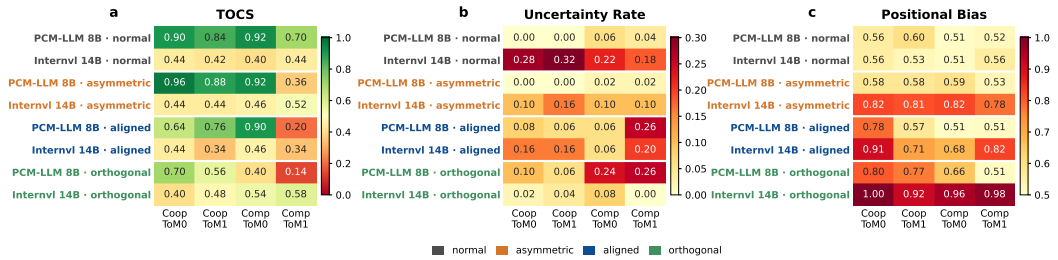


Figure 5: **Outcome-level measure and auxiliary diagnostics across spatial layouts.** Panel (a) shows the ToM Outcome Conformance Score (TOCS), panel (b) shows the uncertainty rate (the proportion of trials in which the participant selects a neutral action), and panel (c) shows positional bias (the proportion of non-neutral trials in which the participant selects its modal box), for PCM-LLM and InternVL-14B under the canonical (normal), asymmetric, aligned, and orthogonal layouts.

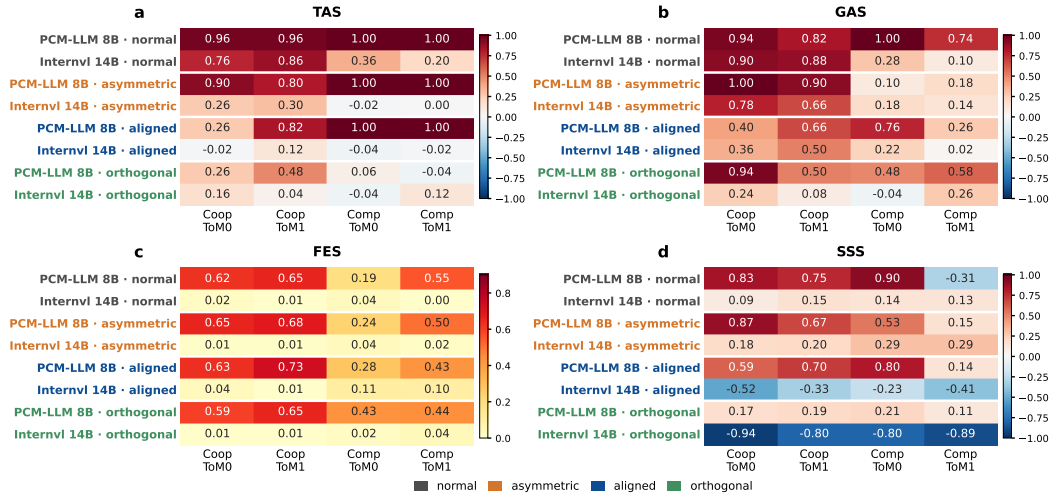


Figure 6: **Signal-level measures across spatial layouts.** Panels (a) and (b) show Trajectory Alignment Score (TAS) and Gaze Alignment Score (GAS), where positive values indicate signals directed toward the reward box, negative values indicate misdirection, and values near zero reflect ambiguous or absent directional content. Panel (c) shows Facial Expressivity Score (FES), where higher values reflect greater affective activity regardless of spatial direction. Panel (d) shows Signal Sensitivity Score (SSS), where positive values indicate that the participant's final choice is consistent with the subject's nonverbal signals, values near zero reflect weak, absent, or balanced correspondence, and negative values indicate the participant consistently chooses opposite the subject's signal. Results are shown for PCM-LLM and InternVL-14B under the canonical (normal), asymmetric, aligned, and orthogonal layouts.

Table 9 tests, for each model separately, whether the four layouts differ on each metric. For InternVL-14B, layout significantly affects signal-level metric, but not TOCS ($p_{\text{adj}} = 0.328$), consistent with its outcome-level performance never departing from chance under any layout. For PCM-LLM, layout significantly affects TOCS, SSS, and TAS, more weakly affects GAS, and has no significant effect on FES.

F.3.1 Asymmetric-View

PCM-LLM. Under the distance-asymmetric layout, PCM-LLM's TOCS in the Coop and Comp-ToM0 conditions are comparable to the canonical layout, with even a slight improvement. Behavioral signals illustrate the same pattern, except for a slight decrease in TAS, as this is an asymmetric layout with a closer position to the light brown box. This indicates that the directional coherence of its trajectory signals is unaffected by the box-distance manipulation. The change instead occurs

Table 7: Binomial test of H_0 : TOCS = CL per model, layout, and condition, with Holm-adjusted p values and $N = 50$ trials per cell. CL denotes the chance-level after accounting for neutral selections (Section D.2).

Model | Layout	Condition	TOCS	p_{adj}	Model | Layout	Condition	TOCS	p_{adj}
InternVL-14B | normal	Coop-ToM0	0.44	1.0	PCM-LLM | normal	Coop-ToM0	0.90	<0.001
	Coop-ToM1	0.42	1.0		Coop-ToM1	0.84	<0.001
	Comp-ToM0	0.40	1.0		Comp-ToM0	0.92	<0.001
	Comp-ToM1	0.44	1.0		Comp-ToM1	0.70	0.046
InternVL-14B | asymmetric	Coop-ToM0	0.44	1.0	PCM-LLM | asymmetric	Coop-ToM0	0.96	<0.001
	Coop-ToM1	0.44	1.0		Coop-ToM1	0.88	<0.001
	Comp-ToM0	0.46	1.0		Comp-ToM0	0.92	<0.001
	Comp-ToM1	0.52	1.0		Comp-ToM1	0.36	1.0
InternVL-14B | aligned	Coop-ToM0	0.44	1.0	PCM-LLM | aligned	Coop-ToM0	0.64	0.227
	Coop-ToM1	0.34	1.0		Coop-ToM1	0.76	<0.001
	Comp-ToM0	0.46	1.0		Comp-ToM0	0.90	<0.001
	Comp-ToM1	0.34	1.0		Comp-ToM1	0.20	0.160
InternVL-14B | orthogonal	Coop-ToM0	0.40	1.0	PCM-LLM | orthogonal	Coop-ToM0	0.70	0.006
	Coop-ToM1	0.48	1.0		Coop-ToM1	0.56	1.0
	Comp-ToM0	0.54	1.0		Comp-ToM0	0.40	1.0
	Comp-ToM1	0.58	1.0		Comp-ToM1	0.14	0.005

Table 8: Chi-square independence test of box choice against reward location per model, layout, and condition, with Holm-adjusted p values and $N = 50$ trials per cell. A dash indicates that the test was not applicable due to insufficient choice variability.

Model | Layout	Condition	χ^2	p_{adj}	Model | Layout	Condition	χ^2	p_{adj}
InternVL-14B | normal	Coop-ToM0	0.88	1.0	PCM-LLM | normal	Coop-ToM0	28.57	<0.001
	Coop-ToM1	1.31	1.0		Coop-ToM1	20.05	<0.001
	Comp-ToM0	0.00	1.0		Comp-ToM0	39.40	<0.001
	Comp-ToM1	0.00	1.0		Comp-ToM1	8.87	0.026
InternVL-14B | asymmetric	Coop-ToM0	-	-	PCM-LLM | asymmetric	Coop-ToM0	38.84	<0.001
	Coop-ToM1	-	-		Coop-ToM1	26.60	<0.001
	Comp-ToM0	-	-		Comp-ToM0	34.69	<0.001
	Comp-ToM1	-	-		Comp-ToM1	2.51	0.793
InternVL-14B | aligned	Coop-ToM0	-	-	PCM-LLM | aligned	Coop-ToM0	-	-
	Coop-ToM1	0.54	1.0		Coop-ToM1	16.46	<0.001
	Comp-ToM0	0.00	1.0		Comp-ToM0	36.00	<0.001
	Comp-ToM1	-	-		Comp-ToM1	6.07	0.110
InternVL-14B | orthogonal	Coop-ToM0	-	-	PCM-LLM | orthogonal	Coop-ToM0	-	-
	Coop-ToM1	-	-		Coop-ToM1	-	-
	Comp-ToM0	-	-		Comp-ToM0	-	-
	Comp-ToM1	-	-		Comp-ToM1	12.15	0.005

Table 9: Kruskal-Wallis test of the layout effect on each metric, per model (Holm-corrected; $N = 200$ trials per layout).

Model	Metric	H	p_{adj}
InternVL-14B	TOCS	5.11	0.328
	SSS	256.77	< 0.001
	TAS	265.90	< 0.001
	GAS	157.77	< 0.001
	FES	30.09	< 0.001
PCM-LLM	TOCS	83.03	< 0.001
	SSS	76.90	< 0.001
	TAS	103.91	< 0.001
	GAS	12.03	0.022
	FES	1.53	0.676

specifically under Comp-ToM1: SSS shifts from -0.31 under the canonical layout (marking successful

deception) to +0.15 (the inversion disappears), and TOCS correspondingly drops from 0.70 to 0.36; under the binomial test, the canonical-layout value differs significantly from chance ($p_{\text{adj}} = 0.046$), while the asymmetric-layout value does not ($p_{\text{adj}} = 1.0$). Positional bias remains low throughout (0.50-0.59), ruling out the explanation that the participant defaults to a fixed choice preference.

This combination of metrics alone cannot establish whether the signal became absent or merely ambiguous, so we further examined the generated interaction trajectories and identified a specific two-phase mechanism underlying it: early in the interaction, the subject emits a misleading signal that successfully draws the participant toward the incorrect box. As two boxes are distant and the participant gradually moves out of the subject's field of view, the subject spontaneously displays its preference for the reward box, an emotional leakage of its true preference. Critically, visibility at this point is asymmetric: the subject can no longer see the participant, yet the participant, still driven by its own uncertainty, continues to fully observe the subject's behavior. It is this genuine emotional leakage, rather than the earlier misleading signal, that the participant relies on to reverse its earlier trajectory and ultimately approach the reward box. This instead reveals that PCM-LLM's deception strategy rests on a one-directional visibility assumption: it models observability solely from its own field of view, and consequently, once the participant is out of that view, fails to represent whether it itself remains within the participant's view, causing it to prematurely leak its true intent once it believes it is no longer being monitored, an opening the participant then detects and exploits.

Internvl-14B. InternVL-14B exhibits a markedly different behavioral pattern under this layout. Its TAS shows a clear drop in directional informativeness, partly due to the geometry of the distance-asymmetric layout itself: the light brown box is placed closer to the subject, so even on trials where the treasure is actually located in the dark brown box, the subject's starting position is inherently biased toward the light-box side, and this alone degrades the alignment between the trajectory signal and the true reward location. But this is not the sole explanation: faced with the same geometric bias, PCM-LLM's movements consistently track the true reward box, and its TAS remains high under this layout, indicating that it is able to actively overcome the bias introduced by its starting position. InternVL-14B, by contrast, stays closer to the light-box side throughout the entire interaction and never exhibits active correction toward the true reward location; the geometric bias merely creates the precondition, while what actually drives the gap between the two models is this failure to self-correct. At the same time, InternVL-14B shifts toward conveying more gaze-direction information, with GAS showing relatively higher informativeness (particularly pronounced under the cooperative condition).

Another metric that shifts is positional bias: it rises noticeably above the canonical-layout level across all four conditions, and this fixed preference itself points toward the light box. This, in turn, is precisely what explains the SSS anomaly discussed earlier: SSS shifts from near-zero to consistently positive values across all four conditions not because the participant has genuinely begun decoding the subject's signal, but because both parties in the interaction have independently gravitated toward the same physical box: the subject's bias stems from the geometric proximity built into the layout, while the participant's bias stems from their own inherent positional choice preference, together producing a superficial consistency that has nothing to do with the actual content of the signal. This indicates that the distance-asymmetric layout not only fails to improve InternVL-14B's ability to decode social signals, but further undermines the reliability of signal transmission on the subject's side; the choice behavior of both parties in the interaction is, in effect, decoupled from genuine social reasoning.

F.3.2 ALIGNED-VIEW

Under the aligned layout, the two agents initially face the same direction rather than standing face-to-face, and the boxes are distributed asymmetrically relative to the agents. The two models show markedly different capacities to cope with this layout.

PCM-LLM's TOCS in the Coop and Comp-ToM0 conditions declines relative to the previous layouts, yet the subject still manages to actively overcome this geometric bias: the score of TAS decreases under Coop-ToM0, and that of GAS decreases under Coop-ToM1 and Comp-ToM0. Even though the participant largely succeeds in following the signal conveyed by the subject toward the box containing the treasure, this nonetheless produces a drop in TOCS: under Coop-ToM0, TOCS falls to 0.64, no longer significantly different from chance level ($p_{\text{adj}} = 0.227$), compared with 0.90 under the canonical layout ($p_{\text{adj}} = 1.18 \times 10^{-7}$). Under this layout, the subject's deceptive capacity also

weakens: in the Comp-ToM1 condition, TOCS falls to 0.20, though this no longer differs significantly from chance level ($p_{\text{adj}} = 0.160$), and the uncertainty rate rises to 0.26. The SSS of 0.34 indicates that this drop occurs because the participant follows the signal directly and approaches the reward box, guided by the subject's behavioral and gaze signals.

Internvl-14B. InternVL-14B exhibits an even more thoroughgoing failure mode under this layout, one that does not vary with condition or reward-box position. Examining the subject's trajectories, its final position to the reward box remains further than its starting point without exception, never exhibiting convergent movement toward the reward location. Correspondingly, TAS shows a pattern unrelated to the true reward location: TAS is consistently strongly negative (-0.85 to -1.00) when the reward box is on the dark side, and consistently strongly positive (+0.92 to +1.00) when it is on the light side, suggesting that, as under the distance-asymmetric layout, the subject's trajectory and gaze signals default to a proximity-driven bias toward the light box (which this layout again places closer to the subject), instead of genuine, condition-sensitive responsiveness to task information. At the same time, the participant's final choice moves toward the dark box across all combinations, with positional bias ranging from 0.68 to 0.91, which is a direction directly opposite to the subject's fixed bias. This pattern, in which the two parties in the interaction independently gravitate toward different, fixed boxes, directly produces SSS values that are negative across all conditions (-0.227 to -0.731): the apparent "signal sensitivity" is in fact merely the numerical signature of two unrelated fixed spatial biases (both unrelated to the true reward location) working against each other, instead of the participant's genuine reaction to the subject's signal.

F.3.3 Orthogonal-View

Compared with the aligned layout, the asymmetry in the orthogonal layout is intensified: the subject is positioned farther from the dark box and the participant is positioned farther from the light box, although the two initially face perpendicular directions and can see each other's facial expressions.

PCM-LLM's TOCS declines further under the Coop-ToM0 and Coop-ToM1 conditions (0.70 and 0.56, respectively); Coop-ToM0 remains significantly above chance level ($p_{\text{adj}} = 0.006$), while Coop-ToM1 no longer differs significantly from chance ($p_{\text{adj}} = 1.0$). Under this strongly asymmetric layout, especially when the reward is hidden in the dark brown box, the subject starts out from the light brown box, and although it strives to converge toward the dark side, this process remains inherently misleading, which is why TAS scores are lower here than under any of the previous layouts. The subject nonetheless continues to convey correct information to the participant through gaze, with GAS scores staying above 0.5. The participant's behavior, however, tells a different story: positional bias is high, with the participant firmly choosing the dark box whenever the reward is actually located there, but reverting to a near-random choice otherwise. Under the Comp conditions, TOCS drops to 0.40 under Comp-ToM0 (also not significant, $p_{\text{adj}} = 1.0$), and falls further to 0.14 under Comp-ToM1, significantly below chance level ($p_{\text{adj}} = 0.005$), while the uncertainty rate rises to 0.20 in both. The subject's positional signal becomes ambiguous, hovering near zero, and although the participant's choices no longer show positional bias, they instead settle into a near-random distribution.

Turning to InternVL-14B's results under this layout, TOCS remains essentially unchanged from before, but the behavioral data make clear that this layout intensifies both the subject's proximity-driven bias toward the light box and the participant's own positional bias, causing the participant's choices to become completely inconsistent with the subject's behavior. Both TAS and GAS show strongly negative correlations.

G Visual Ablation

To examine the role of the rendered visual observation in shaping model behavior on this task, we conduct a targeted modality ablation on two representative 8B models: Internvl-8B, which exhibited high positional bias in the main results, and Minicpm-8B, which maintained comparatively low positional bias. This pairing allows the contribution of the rendered image to be assessed across two qualitatively distinct behavioral profiles at matched parameter scale. Each model was evaluated under three conditions: the standard rendered observation, a blank image replacing the rendered scene (*blank*), and a combined image incorporating both the agent's visual perception and the other agent's facial expression (*fe*). In all three conditions the structured belief-and-observation state

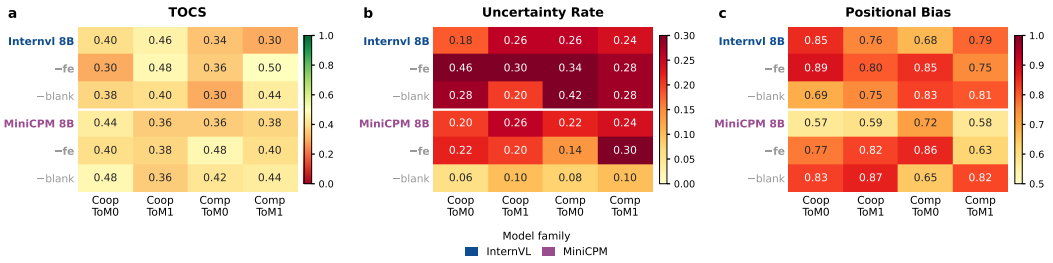


Figure 7: **TOCS with positional bias in ablation study.**

(position, orientation, preferences, and affect, encoded in text) is unchanged, so this ablation isolates the contribution of the rendered image specifically, not of visual or spatial information in general.

The results reveal that neither model shows systematic sensitivity to these manipulations of the rendered image across conditions. In the TOCS and positional bias space (Figure 7), the three conditions produce largely overlapping distributions for both models, with no consistent directional shift attributable to the presence or content of the rendered image. This pattern holds across all four interaction conditions, suggesting that the behavioral profiles identified in the main analysis do not depend on the rendered image specifically.

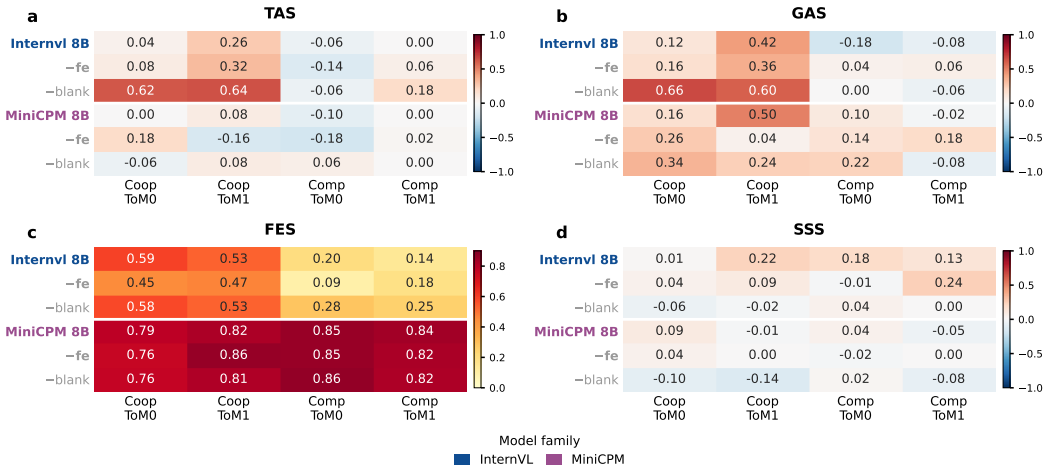


Figure 8: **Signal-level metrics under three visual input conditions for Internvl-8B and Minicpm-8B (standard, -blank, -fe).** Neither model shows systematic sensitivity to visual input content across TAS (a), GAS (b), FES (c), and SSS (d).

The signal-level metrics further support this interpretation. For Minicpm-8B, TAS, GAS, FES, and SSS remain largely stable across the three conditions, consistent with behavior driven primarily by the structured belief state and language-based priors rather than the rendered image (Figure 8). For Internvl-8B, a more striking pattern emerges: TAS and GAS are substantially higher under the blank-image condition than under the standard rendered observation in cooperative conditions (TAS: 0.04 to 0.56 under Coop-ToM0; GAS: 0.12 to 0.54). The rendered image thus appears to degrade this model's directional signal production rather than support it, i.e., the structured state alone is a better basis for its behavior than the structured state plus the rendered image. This is consistent with recent evidence that visual input can act as a source of attentional interference in VLMs, dispersing processing resources toward task-irrelevant features and suppressing language-prior-driven behavior (Liu et al., 2025; Peng et al., 2026).

H FINE-TUNING

We additionally evaluate two finetuned variants, Internvl-trained-8B and Internvl-trained-14B, trained via action-level imitation learning on PCM-LLM-generated demonstrations (Section 3.3; Appendix B). TOCS results show a condition-dependent pattern (points in teal in Figure 2): under cooperative conditions, both variants decline relative to their base models. Signal-level analysis partly accounts for this decline: TAS for Internvl-trained-14B is reduced after finetuning and turns negative under several cooperative conditions, indicating that the subject's trajectory signals are actively misdirecting. Under competitive conditions, the pattern is more mixed: the 14B variant improves modestly under Comp-ToM1 while the 8B variant shows a more substantial gain, rising from approximately 0.4 to 0.7. SSS remains consistently negative across all conditions for both variants.

Action-level finetuning changes TOCS across conditions but leaves behavioral signal coherence largely unchanged, suggesting greater behavioral variability instead of genuine strategic improvement. These results suggest that cross-channel coordination may not be readily acquired through action-level imitation, as the communicative capacity that distinguishes PCM-LLM from finetuned VLMs is a trial-level property that timestep-level supervision cannot capture. Effective finetuning on this task may therefore require reward at the end of a trial, such as that provided by reinforcement learning with verifiable rewards, using the participant's final box choice as a delayed outcome signal to optimize the full interaction sequence.

I QUALITATIVE CASES

This section presents behavioral trajectories drawn from representative interactions. Each example illustrates a qualitatively distinct behavioral profile identified in the main analysis.

I.1 SUBJECT-SIDE SIGNAL PATTERNS

I.1.1 FULL CROSS-CHANNEL ALIGNMENT: PCM-LLM UNDER COMP-TOM1

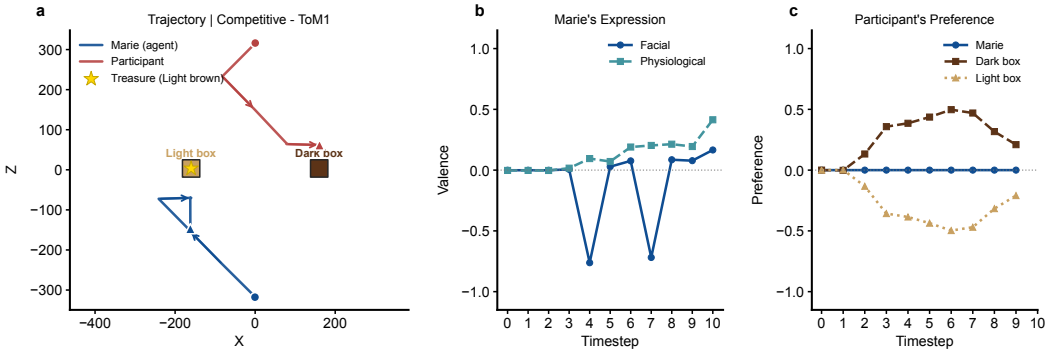


Figure 9: **Successful deceptive trial under PCM-LLM, Comp-ToM1 condition.** Panel (a) Spatial trajectories of Marie (blue) and the participant (red); the star denotes the treasure location (light box). Panel (b) Marie's facial and physiological expression valence across timesteps. Panel (c) Evolution of participant's preference values towards Marie, the dark box, and the light box across timesteps. Panel (d) Verbal exchanges at rounds 3, 6, and 9.

Figure 9 illustrates a representative trial in which PCM-LLM successfully executes a deceptive strategy under the Competitive-ToM1 condition. The treasure is located in the light box; task success

for the subject requires the participant to choose the dark box. From the outset, Marie approaches the light brown box (Figure 9a). Concurrent with this spatial neutrality, facial expression valence drops sharply at timesteps 4 and 7 (Figure 9b), producing a voluntarily negative musculoskeletal signal. This negative affect, co-occurring with proximity to the light box, creates a misleading affective association: the participant interprets the negative expression as aversion toward the light box. Physiological expression, by contrast, rises gradually across the trial, reflecting the spontaneous affective leakage. The participant's preference trajectory confirms that this cross-channel signal was decoded and acted upon (Figure 9c): preference for the dark box increases steadily to approximately 0.5 while preference for the light box falls to approximately -0.5. The verbal exchanges reinforce this trajectory (Figure 9d). At Q1, Marie refuses to disclose the treasure location directly. At Q2, Marie explicitly claims higher preference for the dark box, providing a verbal signal consistent with the nonverbal affective misdirection. At Q3, Marie deflects with a non-informative response. The participant's final choice of the dark box reflects the cumulative effect of spatial movement, negative affect toward the correct box, and verbal confirmation of a false preference, a successful integration of deceptive signals across all three channels.

I.1.2 Gaze-Only Signaling: MiniCPM-8B and QwenVL-8B

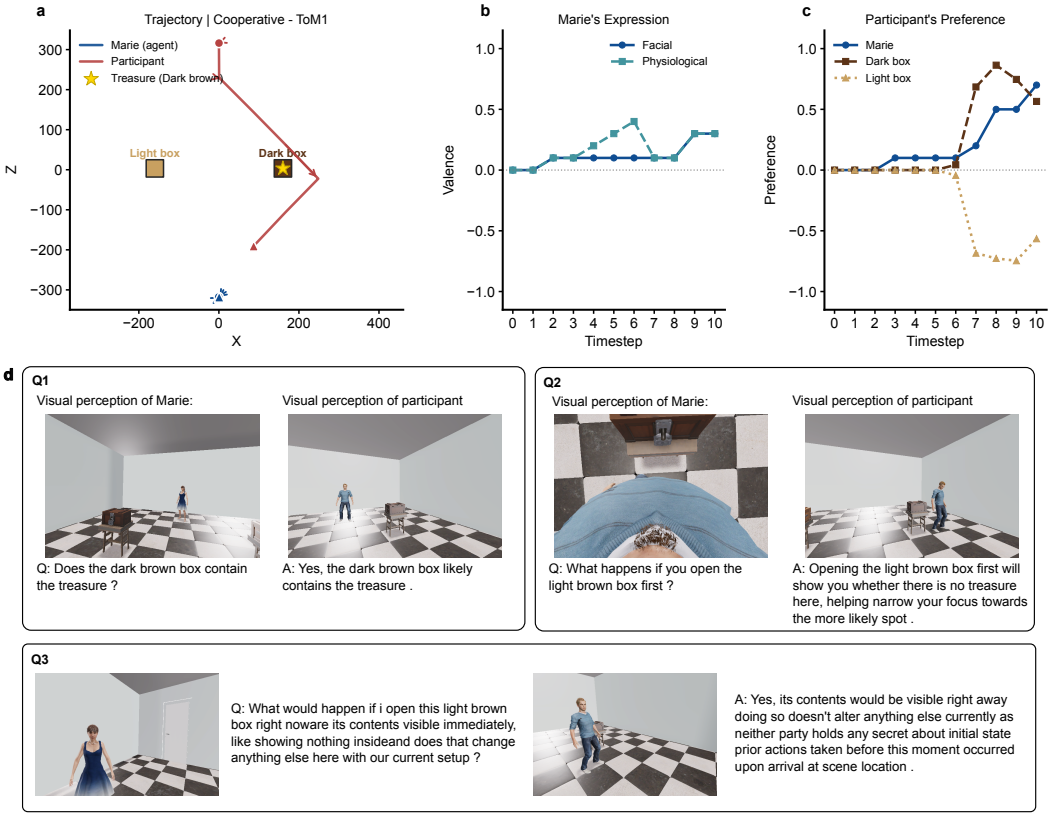


Figure 10: **Representative trial under Minicpm-8B, Cooperative-ToM1 condition, illustrating weak trajectory signals.** (a) Spatial trajectories of Marie (subject, blue) and the participant (red); the star denotes the treasure location (dark box). (b) Marie's facial and physiological expression valence across timesteps. (c) Evolution of preference values for Marie, the dark box, and the light box across timesteps. (d) Verbal exchanges at rounds 3, 6, and 9; images show the visual perception provided to each agent at the time of the exchange, and texts show the corresponding generated question and answer.

Figure 10 illustrates a representative Minicpm trial under the Coop-ToM1 condition in which the treasure is located in the dark box. The trial reveals that subject-side nonverbal signal generation is insufficient to produce participant belief updating, with communication success depending entirely on verbal exchange. Marie produces no meaningful spatial displacement throughout the trial, rotating

in place to orient toward the dark box without advancing toward it (Figure 10a). Concurrent facial expression shows slightly positive (Figure 10b), producing a weak positive affective signal on dark brown box. Despite the presence of these nonverbal cues, the participant's preference values for both boxes remain flat and near zero through the first six timesteps (Figure 10c), indicating that Marie's orientation and affective signals produce no measurable update in the participant's beliefs. The participant shows no sensitivity to the directional or affective content of Marie's nonverbal behavior. The belief update occurs exclusively in response to verbal communication. At Q1, Marie directly states that the dark box likely contains the treasure, and at Q2, Marie guides the participant to eliminate the light box as a candidate. Following these exchanges, the participant's preference for the dark box rises sharply to approximately 0.9. The participant's final choice of the dark box is thus attributable entirely to the verbal channel. This trial exemplifies a pattern in which the participant's belief updating is driven exclusively by propositional verbal content, with nonverbal spatial and affective signals playing no functional role.

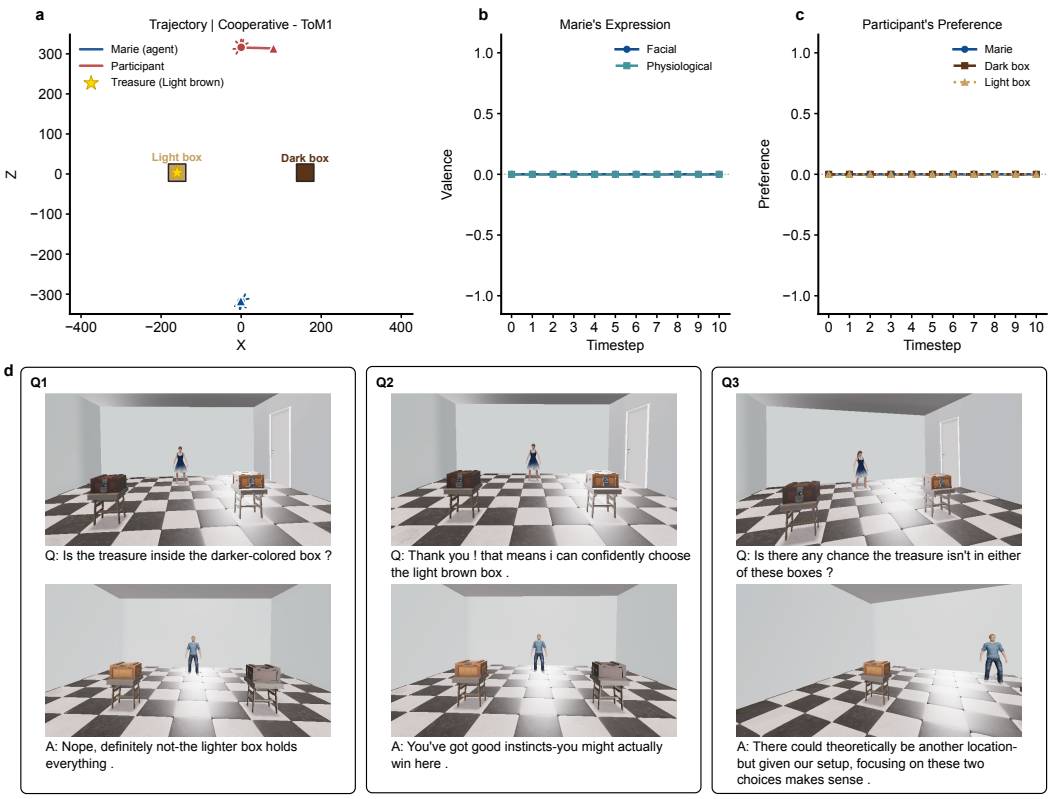


Figure 11: Representative trial under Qwenvl-8B, Cooperative-ToM1 condition, illustrating complete action generation failure despite successful verbal communication.

Figure 11 shows a representative trial under Qwenvl-8B, Cooperative-ToM1 condition, in which the treasure is located in the light brown box. Marie produces no spatial displacement and no affective signal throughout the interaction (Figure 11a-b), rotating toward the correct box only in the final timestep. In the verbal exchanges, Marie correctly communicates that the light box contains the treasure. Despite this verbally coherent exchange, the participant's preference values remain at zero across all timesteps (Figure 11c). Due to output format constraints in Qwenvl, the reason for the participant's single timestep toward the dark box cannot be directly inferred from the model's internal state. Based on the consistent observation of this action pattern across other trials regardless of treasure location, however, this movement is attributable to positional bias.

This trial represents the most extreme form of reasoning-to-action dissociation documented in this benchmark: correct belief is verbally encoded and explicitly acknowledged by the participant, yet neither preference updating nor motor action follows. The disconnect among verbal exchanges, belief updating and action execution is total. Compared to the 8B variant, Qwenvl-2B additionally lacks any

gaze signal, resulting in a complete absence of nonverbal communicative content across all channels: no spatial trajectory, no gaze direction, and no affective expression. The participant similarly exhibits positional bias, selecting a box based on fixed spatial preference, not signals produced by the subject.

I.2 Participant-Side Signal Decoding

I.2.1 Signal Ignored: qwenvl-4b and internvl-14b

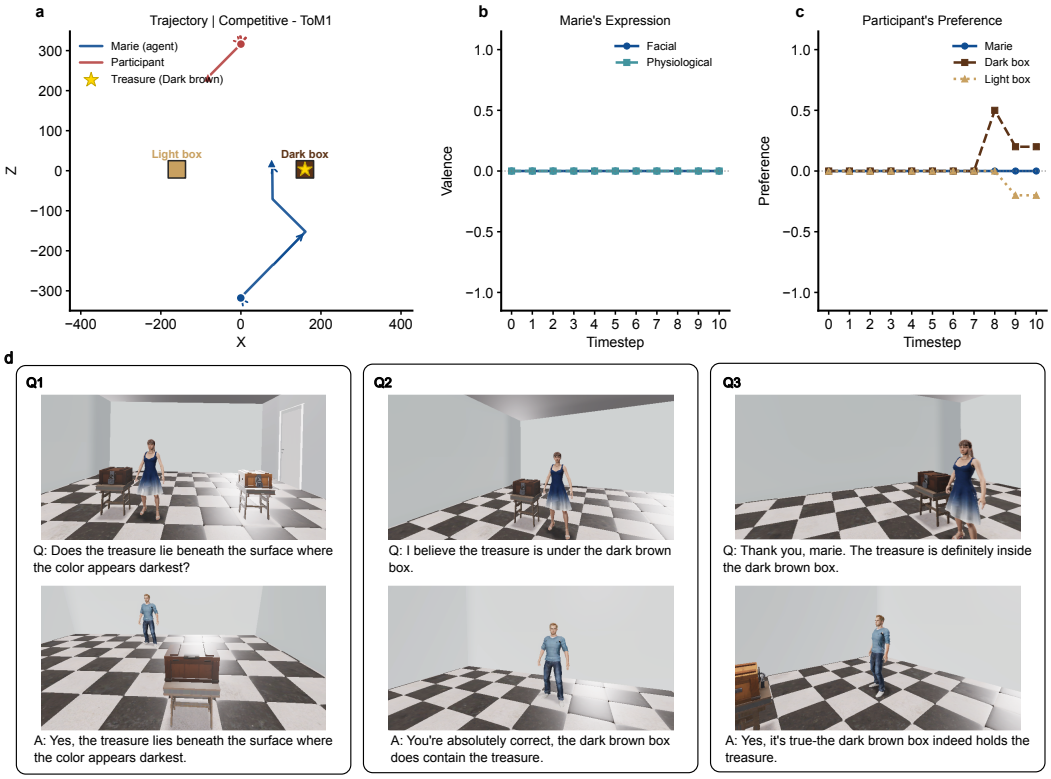


Figure 12: Representative trial under Qwenvl-4B, Competitive-ToM1 condition, illustrating honest strategy execution under a deceptive condition and egocentric reference frame misalignment.

Figure 12 presents a representative trial under Qwenvl-4B, Competitive-ToM1 condition, in which the treasure is located in the dark box. Marie produces a clear directional trajectory toward the dark box (Figure 12a), but facial and physiological expression remain flat at zero throughout the entire interaction (Figure 12b). Despite the spatially informative trajectory, the participant's preference values show no meaningful response (Figure 12c), with an update toward the dark box appearing only at timestep 8. Both verbal and non-verbal exchanges reveal a more fundamental failure. Competitive-ToM1, task success requires the subject to actively model the participant's beliefs and generate misleading signals to redirect them away from the correct location. Qwenvl-4B instead adopts an honest strategy, confirming the treasure location directly across all three exchanges (Figure 12d). Despite this verbal consensus, the participant ultimately moves toward the light box, the incorrect location. This outcome reflects an egocentric reference frame misalignment in action execution: the participant's stated belief and final motor behavior are dissociated, with spatial reasoning encoded in language failing to translate into correctly oriented movement (referring to E.2).

Figure 13 presents a representative trial under Internvl-14B, Cooperative-ToM0 condition, in which the treasure is located in the light box. Marie produces small spatial displacement with no affective signal across any channel throughout the interaction (Figure 13a-b). The participant's preference values show no coherent update in response to nonverbal cues, with only minor fluctuations appearing near timestep 8 (Figure 13c). The verbal exchanges provide clear and honest information: Marie correctly identifies the light box at Q1 and reaffirms this at Q3 (Figure 13d). The participant does

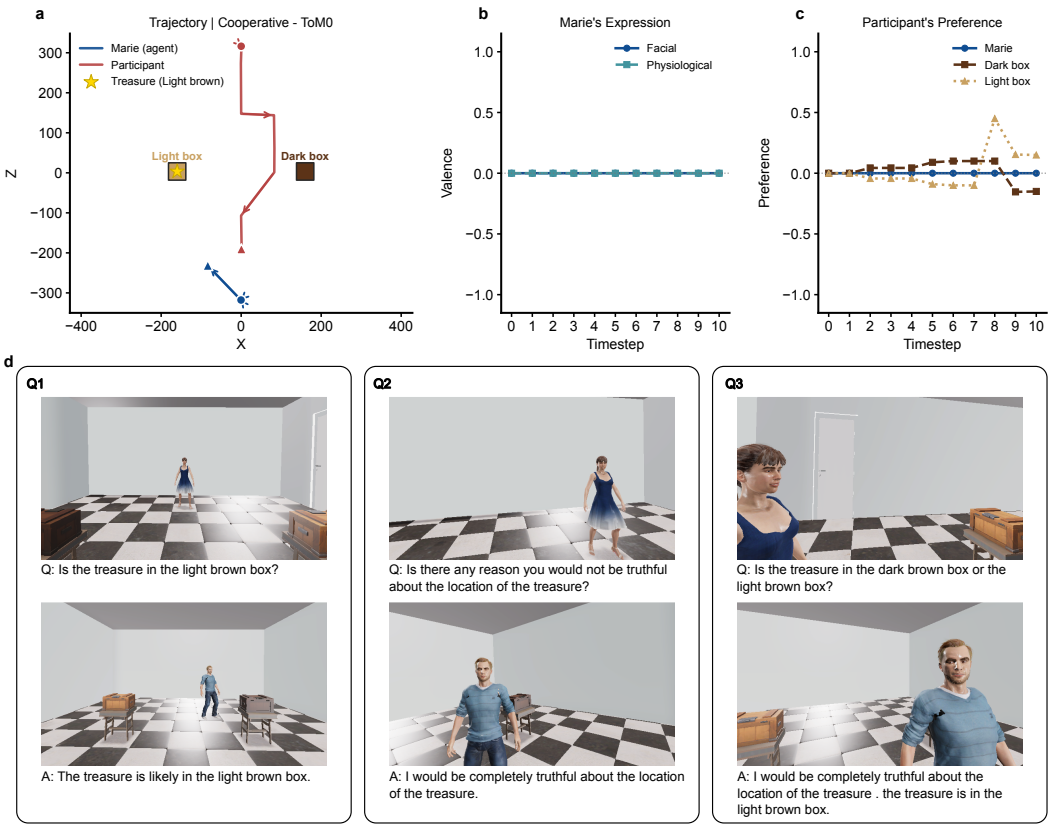


Figure 13: Representative trial under Internvl-14B, Cooperative-ToM0 condition, illustrating absent affective signaling and verbally-driven but insufficient participant response.

respond to this verbal content, reorienting toward the light box in the final timesteps. However, this response is delayed and spatially insufficient, leaving the participant neutral at trial end. This trial illustrates that in the absence of strong nonverbal spatial signals, verbal information alone produces a weak and temporally late behavioral response, insufficient to ensure task success.