

## Evaluating Multi-Perspective Theory of Mind in Embodied Social Interactions

Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, Maarten Sap











**First-Person Evaluation** 

**Third-Person Evaluation** 

What we build

**SoMi**, an embodied multi-agent social interaction environment in Minecraft.

**SoMi-ToM**, a novel embodied ToM benchmark for evaluating multi-perspective ToM in complex multiagent social interactions within Minecraft.

(1) LVLMs perform significantly below human level on SoMi-ToM. In the 1st-person evaluation, the average accuracy gap between humans and LVLMs is **40.1%**. In the **3rd-person** evaluation, the average accuracy gap between humans and LVLMs is 26.4%. (2) The main reasons for the poor performance of LVLMs include: ignoring or inaccurately tracking resource consumption, insufficient reliance on system feedback, being misled by initial intentions rather than actual behavior, overgeneralization or inappropriate associations, failure to identify hierarchical goal structures, entity recognition confusion, and detailed errors.

Benchmark & Code: github.com/XianzheFan/SoMi-ToM Data & Dataset Card: huggingface.co/datasets/SoMi-ToM/SoMi-ToM

## **Embodied Interactions Materials and Tools Theory of Mind Evaluation** First-person screenshots Jack Perspective Alright, let's split up the tasks! I'll collect the logs and start crafting. !collectBlocks("oak\_log", 3) You are Jack. What materials or tools do you believe Jane State currently has? **Crafting Goals** A. 12 oak planks and 1 crafting table B. 12 oak planks and Inference I've got 6 oak logs now! I'll 6 oak logs C. 1 crafting table and 1 oak door turn them into planks and craft the crafting table. !craftRecipe("oak\_planks", 6) Third-person video with captions **Third-Person** Perspective ..... I've got the doors. Let's find **Social Relationships** the perfect spot to place What do you believe is the final task goal of the people Goal !placeHere("oak\_door") in the video? Inference A. Collect logs B. Craft a crafting table C. Craft 3 doors. **Behavior** What do you believe John did in the video? Social Relationship: Collaboration A. Craft a crafting table B. Place a door C. Collect birch logs. Inference Obstruction Collaboration **Crafting Goal: 3 Doors**

What we find

(1) 1st-person (1050 questions): inferring the *state* during *real-time* interaction

Method	Self-ToM	Reasoning	Others' ToM	I Reasoning	Weighted Average		
Human	91.9		89	0.0	90.0		
	w/o CoT	w/CoT	w/o CoT	w/CoT	w/o CoT	w/ CoT	
Gemini 1.5 Pro	55.1	60.9	38.0	55.1	43.7	57.0	
Gemini 2.0 Flash	48.9	56.0	43.1	54.3	45.0	54.9	
GPT-4o	40.3	70.3	33.3	54.1	35.6	59.5	
InternVL2.5 78B	48.9	52.0	44.1	40.8	45.7	44.6	
Qwen2.5-VL	41.1	52.6	42.4	46.6	42.0	48.6	
LLaVA 1.6 13B	32.3	36.0	31.2	34.4	31.6	35.0	
Average (LVLMs)	$44.4 \pm 8.1$	$54.6 \pm 11.4$	$38.7 \pm 5.4$	$47.6 \pm 8.5$	$40.6 \pm 5.7$	$49.9 \pm 9.2$	

(2) 3rd-person (175 questions): inferring the *goal* and *behavior* based on the **complete video** 

Method	Input	Input   Goal Inference   Behavior Inference		Behavior Inference 2			Average
	1			Jack	Jane	John	l
Human	Video	100.0	95.2	95.2	90.5	85.7	$93.3 \pm 5.4$
Gemini 1.5 Pro	Video	94.3	74.3	80.0	71.4	57.1	$75.4 \pm 13.5$
Gemini 1.5 Pro CoT	Video	97.1	60.0	74.3	71.4	54.3	$71.4 \pm 16.5$
Gemini 2.0 Flash	Video	94.3	88.6	80.0	65.7	62.9	$78.3 \pm 13.8$
Gemini 2.0 Flash CoT	Video	100.0	82.9	68.6	77.1	68.6	$79.4 \pm 13.0$
GPT-4o	Images	71.4	57.1	77.1	80.0	62.9	$69.7 \pm 9.6$
GPT-40 CoT	Images	91.4	80.0	85.7	82.9	74.3	$82.9 \pm 6.4$
InternVL2.5 78B	Images	85.7	71.4	82.9	71.4	68.6	$76.0 \pm 7.7$
InternVL2.5 78B CoT	Images	85.7	57.1	82.9	68.6	65.7	$72.0 \pm 12.0$
Qwen2.5-VL	Images	100.0	74.3	85.7	77.1	54.3	$78.3 \pm 16.7$
Qwen2.5-VL CoT	Images	97.1	82.9	85.7	71.4	65.7	$80.6 \pm 12.4$
VideoLLaMA 3 7B	Video	54.3	37.1	34.3	37.1	25.7	$37.7 \pm 10.4$
VideoLLaMA 3 7B CoT	Video	57.1	37.1	25.7	22.9	37.1	$36.0 \pm 13.5$
LLaVA-Video 7B	Video	77.1	34.3	54.3	51.4	34.3	$50.3 \pm 17.7$
LLaVA-Video 7B CoT	Video	74.3	37.1	51.4	48.6	31.4	$48.6 \pm 16.6$
Average (LVLMs)	/	$84.3 \pm 15.3$	$62.4 \pm 19.6$	$69.2 \pm 19.9$	$64.1 \pm 17.6$	$54.5 \pm 15.8$	$66.9 \pm 16.4$
Question Counts	/	35	35	35	35	35	/

