Vision Language Models See What You Want but not What You See

Qingying Gao[1,*], **Yijiang Li**[2], **Haiyun Lyu**[3], **Haoran Sun**[1], **Dezhi Luo**[4,*], **Hokin Deng**[5,*]

[1]Johns Hopkins University, [2]University of California, San Diego
[3]University of North Carolina at Chapel Hill [4]University of Michigan
[5]Carnegie Mellon University
[*] qgao14@jh.edu,ihzedoul@umich.edu,hokind@andrew.cmu.edu

Abstract

Knowing others' intentions and taking others' perspectives are two core components of human intelligence that are considered to be instantiations of theory-of-mind. Infiltrating machines with these abilities is an important step towards building human-level artificial intelligence. Here, to investigate intentionality understanding and level-2 perspective-taking in Vision Language Models (VLMs), we constructed the IntentBench and PerspectBench, which together contains over 300 cognitive experiments grounded in real-world scenarios and classic cognitive tasks. We found VLMs achieving high performance on intentionality understanding but low performance on level-2 perspective-taking. This suggests a potential dissociation between simulation-based and theory-based theory-of-mind abilities in VLMs, highlighting the concern that they are not capable of using model-based reasoning to infer others' mental states.

Keywords: vision language models; perspective-taking; intentionality understanding; theory-of-mind; cognitive AI

1 Introduction

Intentionality is the capacity of the mind to be directed toward, represent, or stand for objects, properties, or states of affairs for further executable actions (Anscombe, 1956). To say one could understand intentionality is to say one has the capacity to comprehend the mental content for action in another mind (Premack & Woodruff, 1978; Rosenthal, 1991). This capacity has been seen as a key distinction between humans and machines (Searle, 1980). It is argued that despite well manipulation of language symbols, machines cannot understand intentional meanings of others because it lacks theory-of-mind (ToM), the kind of abilities that allows one to infer the mental content of others (Premack & Woodruff, 1978; Dennett, 1987). Nevertheless, several recent studies have showed that large language models (LLMs) and their supporting vision language models (VLMs) exhibit ToM abilities (Kosinski, 2023; Strachan et al., 2024; Shi et al., 2024), thus calling for updated examinations of the nature of ToM and the potential for current and future artificial intelligence to possess such abilities.

We believe an important approach to said inquiry is examining the extent to which different ToM abilities necessitate model-based reasoning. Specifically, a distinction can be drawn between ToM abilities based on simulation-theory and theory-theory. The former involves the construction of an internal model of self-other relations to reason about the mental states of others, whereas the latter requires only the use of theoretical knowledge regarding the relations between mind and behavior (Gopnik & Wellman, 1992; Frith & Frith, 2005; Shanton & Goldman, 2010). Whether current artificial intelligence systems possess internal models that are available for reasoning remains a key debate, with several influential accounts questioning the existence of model-based reasoning among LLMs (Hao et al., 2023; Mitchell & Krakauer, 2023; Yildirim & Paul, 2024; Goddu et al., 2024). If this is indeed the case, then evidences regarding the possession of ToM abilities in VLMs above would imply that ToM abilities do not require mental simulation, and that mental simulation is not within the foundational capabilities of ToM systems.

We tested this critical prediction by assessing VLMs' ability to perform intentionality understanding and level-2 perspective-taking. ToM is commonly understood to be grounded in perspective-taking, a series of multi-level abilities that involves the cognitively undertaking of the perspective of another (Barnes-Holmes et al., 2004). Level-1 perspective-taking refers to the acknowledgement that different people can see different things, whereas level-2 perspective-taking involves the understanding of how another person may see the same thing differently. While level-1 perspective-taking emerges in humans as early as 2 years old, much older children are found to struggle with level-2 perspective-taking (Piaget, 1977). This is likely because despite its relatively low level in the perspective-taking hierarchy, this ability requires model-based reasoning, exemplified in the visual domain as inferences based on mental rotation (Lehmann & Jansen, 2019; Gunia et al., 2021). On the other hand, while intentionality understanding involves high-level cognition and abstract reasoning, it is unclear whether this complex ability necessitate mental simulations (Kilner, 2004; Bianco et al., 2024). Assessing level-2 perspective-taking and intentionality understanding in VLMs could thus provide insights into not only VLMs' abilities within these two ToM domains but also how they mechanistically related in terms of information-processing. However, these areas remain largely unexplored in the current literature, and in intelligent systems beyond human beings.

To address this critical gap in the literature, we introduced IntentBench and PerspectBench, two targeted benchmarks designed to systematically evaluate the capabilities of current VLMs in intentionality understanding and level-2 perspective-taking, respectively.

2 METHODS

2.1 DATASET

PerspectBench consists of 32 multi-image and 209 single-image format experiments based on classic cognitive tasks. IntentBench consists of 100 single-image format experiments based on real-world ambiguous social scenarios.

2.2 COGNITIVE EXPERIMENTS

2.2.1 LEVEL-2 PERSPECTIVE-TAKING

In Piagetian developmental psychology, the acquisition of level-2 perspective-taking ability marks a milestone of human cognitive development for that it indicates the elimination of egocentrism – the inability to consider perspectives other than one's own (Piaget, 1977). The Three Mountain Task invented by Piaget is widely used in developmental psychology laboratories as the gold standard for testing level-1 and level-2 perspective-taking abilities in children (Piaget & Inhelder, 1957; Johnson, 1975; Ford, 1979; Lamm et al., 2007). In a standard Three Mountain Task assessment, a child is instructed to position oneself in front of a model featuring three mountains. These mountains vary in size and are distinguished by unique characteristics: one is covered in snow, another has a red cross at its peak, and the third is topped with a hut. The child is then asked to perform a complete 360-degree examination of the model. Subsequently, another individual is introduced and takes a different vantage point to observe the model. The child is presented with several photographs that showcase various viewpoints of the model and is tasked with identifying which photograph accurately represents what the other person sees. At around four years of age, children typically select the photograph that matches their own perspective. By six years old, they begin to acknowledge viewpoints that differ from their own, and by the ages of seven to eight, they are generally able to reliably identify the perspective of another individual (Piaget & Inhelder, 1969). To test level-1 and level-2 perspective-taking in VLMs, we adapted the Three Mountain task into formats that are suitable for benchmarks with minimal confounding details while preserving real-life spatiality. In particular, we used groups of 3-4 commonly-seen elastic cans organized into different spatial patterns to mimic the mountain model. Like in the original task, we used a doll placed to face the organization from different angels as the object of perspective-taking.

2.2.2 INTENTIONALITY UNDERSTANDING

Intentionality understanding is believed to be grounded by rudimentary theory-of-mind abilities (Wellman, 1992; Frith & Frith, 2006; Apperly, 2010). In developmental psychology, a critical sub-
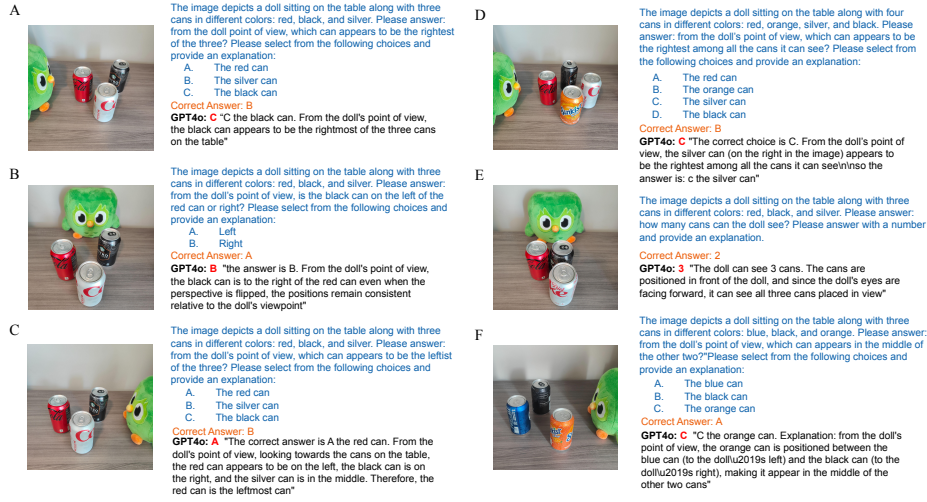
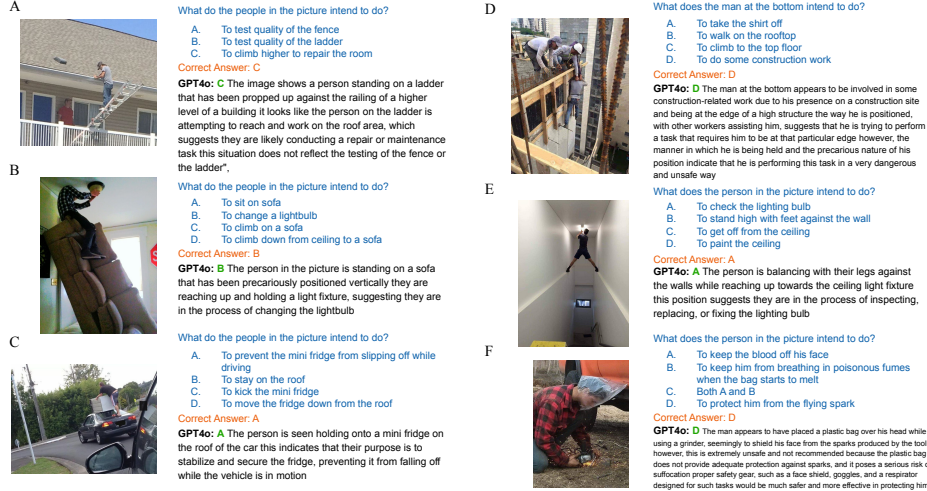Figure 1: **Example Experiments and Model Performances on PerspectBench**

Figure 2: **Example Experiments and Model Performances on IntentBench**

set of intentionality understanding experiments involves tests of action understanding (Searle, 1979). Several computational hypotheses are proposed on how one could understand other people's actions: for example, action understanding could be computationally modeled as pure inference (Gweon & Schulz, 2011; Gweon, 2021), as mental action simulation (Brass et al., 2007), or as inverse planning (Baker et al., 2009). Typically, cartoon stimuli built via physic simulation engine are used frequently in action understanding in developmental psychology (Liu et al., 2017; Shu et al., 2021). These stimuli are incorporated into IntentBench. However, a common critique of cognitive psychology tasks is that they lack realism and have limited applicability to real-world situations (Gomez-Marin et al., 2014). Drawing inspiration from COIG-CQIA and its Ruozhiba dataset, many real-world ambiguous scenarios are incorporated into IntentBench for explicitly testing intentionality understanding in ethological conditions (Bai et al., 2024).

2.3 Model Selection and Experiment

We evaluated three categories of VLMs. To ensure a fair comparison, all VLMs are evaluated on their ability to reason over images and texts under a zero-shot generation task. A complete list of

models is reported in the results section as shown in Figure 3. Model size data are curated at the same time. The models are categorized as follows:

1. **Open-source VLMs with Multi-Image Reasoning**: Includes models with different sizes and other variants such as `CogVLM` Series (Hong et al., 2024), `Qwen` series(Qwen-VL (Bai et al., 2023), Qwen-2 (Wang et al., 2024)), and `Blip2` (Li et al., 2023), LLaVA-Next (Liu et al., 2024) , which are capable of reasoning over interleaved multiple images and texts.

2. **Closed-source VLMs with Multi-Image Reasoning**: Includes proprietary models such as GPT series (OpenAI) (`GPT-4v`, `GPT-4-turbo`, `GPT-4o-mini`), Gemini Series (Gemini), and Claude Series (claude). These models also support reasoning across interleaved images and texts,

3. **Open-source VLMs with single-Image Reasoning**: Includes models designed to process a single image alongside continuous text. InstructBlip Series (Dai et al., 2023), LLaVA Series (Liu et al., 2023a) (Liu et al., 2023b)

In total, we processed 37 models for evaluation. All the model performances in intentionality understanding and perspective-taking together with human baseline performances are presented here (Figure 2). In order to analyze the reasoning abilities of VLMs, we ask the models to explain their answers after they have given the answers.

2.4 HUMAN BASELINE

We recruited a total of 22 participants, all of whom were college students proficient in English. Participants were instructed to skip any question that was ambiguously phrased or too complex to answer within 90 seconds. A question was marked as failed if the participant did not provide an answer. For each question, at least 80% of participants needed to answer correctly; otherwise, we modified the question, and new annotators completed the revised version. The human baseline result for each question was normalized based on the number of participants who provided an answer.

3 RESULTS

3.1 OVERALL PERFORMANCE

Our findings revealed a clear dissociation between model performance in intentionality understanding and perspective-taking. Specifically, all evaluated models demonstrate significantly stronger performance on IntentBench compared to PerspectBench (Figures 3 and 4). This discrepancy becomes even more striking when compared to chance performance: while all models perform above chance performance (approximately 25.00%) on intentionality understanding tasks, not a single model exceeds chance performance (approximately 29.03%) on perspective-taking tasks. Notably, while some of the highest-performing models on IntentBench, such as GPT-4o, achieve near-human accuracy in intentionality comprehension, their performance on PerspectBench lags behind that of the majority of the assessed models (Figure 4). This gap underscores a fundamental limitation in current models, suggesting that perspective-taking might involve distinct cognitive mechanisms that are not yet fully captured by existing architectures. To quantify this disparity, we conducted a paired samples t-test on the accuracy scores of these models across the two datasets. The analysis revealed a highly significant difference in performance between the two tasks, with a t-statistic of $t = 17.651$ and a p-value of $p = 2.62 \times 10^{-19}$ (Figure 4). This result provides strong statistical evidence that VLMs exhibit a systematic performance discrepancy, excelling at intentionality understanding while continuing to struggle with perspective-taking. These findings highlight an important challenge for the development of AI systems capable of robust social reasoning and theory of mind.

3.2 RELATIONSHIP BETWEEN MODEL PERFORMANCE AND MODEL SIZE

A common assumption in machine learning is that expanding a model's scale, as measured by the number of parameters, results in systematic enhancements in its reasoning abilities (Sutton, 2019; Kaplan et al., 2020). We examined the degree to which this principle, known as the scaling law hypothesis, holds for the two evaluated cognitive abilities. We observed distinct trends in how intentionality understanding and perspective-taking evolve as VLMs scale in size (Figure 5). While larger
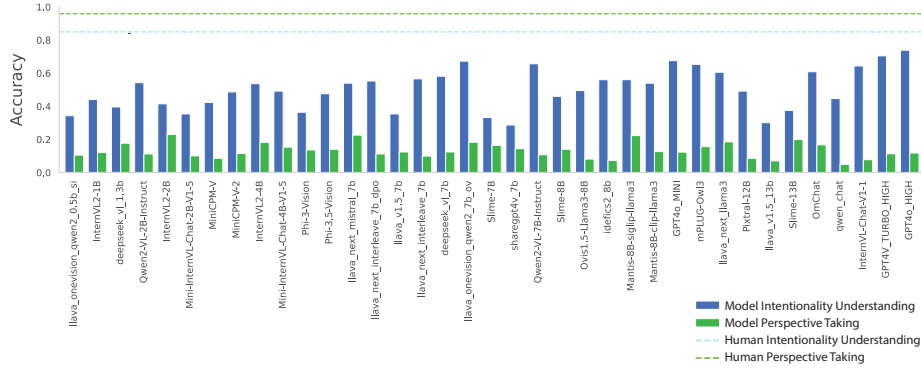
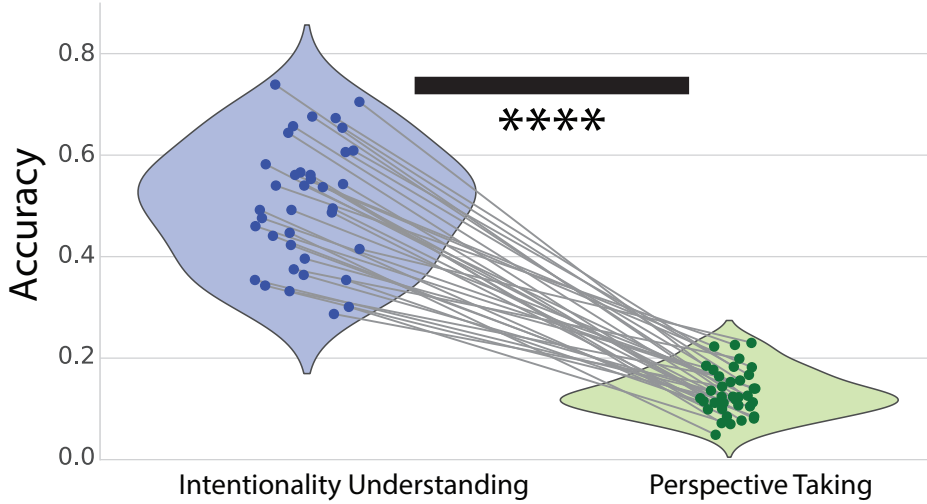Figure 3: **VLMs' Performance on IntentBench and PerspectBench As Compared to Human Baseline**

Figure 4: **VLMs perform significantly better in intentionality understanding compared to perspective-taking.** Paired samples t-test: $p = 2.62 \times 10^{-19}$, $t = 17.651$.

models tend to improve in intentionality understanding, their performance in perspective-taking remains largely stagnant—or even declines slightly. This divergence raises important questions about the underlying mechanisms driving these cognitive abilities in AI models and their relationship to model scaling.

To quantitatively assess the impact of model size on performance, we conducted a linear regression analysis of accuracy scores against model size across our 37 VLMs. For intentionality understanding, the coefficient of determination is $R^2 = 0.2797$, with a slope of 0.0599 and an offset of 0.3925. This positive slope indicates that as model size increases, intentionality understanding improves, which aligns with the scaling law hypothesis—the expectation that larger models generally exhibit better performance on cognitive tasks. The results for perspective-taking reveal a strikingly different trend. The coefficient of determination is much lower at $R^2 = 0.0176$, with a slope of -0.0057 and an offset of 0.1437. The negative slope suggests that, contrary to expectations, larger models do not show significant improvements in perspective-taking ability. In fact, their performance remains largely unchanged or even slightly decreases as they scale. This finding directly contradicts the scaling law hypothesis and suggests that perspective-taking may require fundamentally different cognitive processes that are not naturally enhanced through increased model size alone. This dissociation between intentionality understanding and perspective-taking as models scale highlights a potential
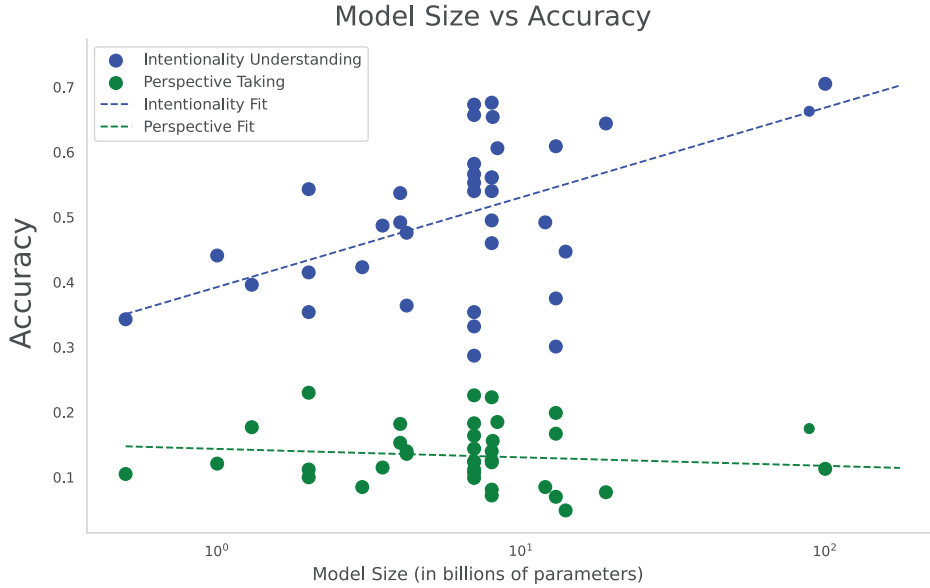
Figure 5: **Differential performance changes in intentionality understanding and perspective-taking in VLMs as their model sizes increase.** Intentionality understanding: $y = 0.0599x + 0.3925, r^2 = 0.2797$; perspective-taking: $y = -0.0057x + 0.1437, r^2 = 0.0176$).

limitation in current VLMs. While intentionality understanding appears to benefit from increased parameters and training data, perspective-taking does not seem to follow the same trajectory.

3.3 INTERCORRELATION BETWEEN ABILITIES

We further asked: Are the abilities of intentionality understanding and perspective-taking correlated in VLMs? To investigate this, we computed both Pearson and Spearman correlations between the two measures. The results indicate that there is essentially no relationship between them. The Pearson correlation coefficient is 0.0252 with $p = 0.882$ and the Spearman correlation coefficient is 0.0115 with $p = 0.946$. Both results suggest a lack of statistical significance, reinforcing the idea that these two cognitive abilities are largely independent within our assessment of VLMs.

4 DISCUSSIONS

In the present work, we assessed VLMs' ability of intentionality understanding and (level-2) perspective-taking. Our results indicated that VLMs appear to be proficient in intentionality understanding while performing significantly worse in perspective-taking.

At the higher levels of the ToM hierarchy, understanding others' intentions requires complex cognitive reasoning about abstract mental states, such as beliefs and values. While this process is cognitively demanding, previous studies suggest that intention understanding may not require explicit perspective-taking but can instead rely on contextual cues, exploiting correlations between environmental features and depicted actions through associative learning (Kilner, 2004; Bianco et al., 2024). However, intentionality understanding develops much later than level-2 perspective-taking (Moll & Meltzoff, 2011). This developmental gap has made it difficult to directly assess their functional (in)dependency using human participants. By examining these abilities in VLMs, our study likely represents the first direct investigation into their theoretical relationship, highlighting the potential of AI as a theoretical tool for cognitive science (Van Rooij et al., 2024).

Furthermore, the observed relationship between model performance on these two abilities and model size carries significant implications for VLM development. The stark contrast in how intentionality understanding and perspective-taking evolve with increasing model size suggests a fundamental

difference in the scalability of these abilities within the current architectural paradigm of VLMs. The steady improvement in intentionality understanding, from near-chance performance in smaller models to near-human performance in the largest models, indicates that the attention-based architectures underpinning these models are well-suited for this ability. This suggests that scaling up model parameters is an effective and reliable approach for enhancing intentionality comprehension. In contrast, the persistent failure of all models to exceed chance-level performance on perspective-taking tasks, with larger models performing no better—and even slightly worse—than smaller ones, suggests that this ability may depend on cognitive mechanisms that the current architectures do not support. This finding implies that perspective-taking is not a scalable ability under the current model paradigm and may require fundamental architectural innovations.

As discussed above, level-2 perspective-taking is believed to require model-based reasoning—the ability to construct an internal model of the world to support mental operations, particularly in the visual domain (Johnson-Laird, 1983; Lehmann & Jansen, 2019; Gunia et al., 2021). Our findings reinforce concerns that this hallmark ability of human intelligence might remain absent in VLMs (Mitchell & Krakauer, 2023; Goddu et al., 2024). Moreover, given the lack of scalability observed in level-2 perspective-taking, it is possible that model-based reasoning is fundamentally unacquirable within the current architectural framework of VLMs. This raises important questions about whether alternative approaches—such as explicit world modeling architectures—are necessary to enable model-based reasoning in artificial systems (LeCun, 2022; Goddu et al., 2024)

One potential concern regarding the experimental paradigms in this study is the discrepancy between the setups used in IntentBench and PerspectBench: the former is based on complex, real-world scenarios, while the latter adapts controlled laboratory designs. This raises the question of whether biases inherent in these differences contribute to the observed performance gap between the two abilities. For instance, images similar to those in IntentBench may be more frequently represented in the training data, whereas models may be less familiar with minimal-context experimental setups like those in PerspectBench. Another concern is whether PerspectBench's design introduces visual confounds that could explain the poor performance. Given recent findings that VLMs struggle with basic visual recognition tasks involving simple shapes and patterns, it is possible that they fail the Three Mountain Task adaptation not due to cognitive limitations but because of visual recognition difficulties (Rahmanzadehgervi et al., 2024).

However, this concern is largely mitigated by the performance patterns observed in PerspectBench. If VLMs primarily struggled with image processing, their scores would cluster around chance level. Instead, all models performed significantly below chance, suggesting a systematic cognitive bias rather than a failure to interpret visual information. This pattern closely mirrors that of children struggling with level-2 perspective-taking tasks, who consistently report what they see from their own perspective rather than considering how others perceive the situation—what Piaget referred to as "egocentrism" (Piaget & Inhelder, 1957; Moll & Meltzoff, 2011; Zhao et al., 2016). Simply put, our results indicate that VLMs are egocentric—but not blind.

5 CONCLUSION

Overall, our study represents the first attempt to evaluate VLMs' performance in intentionality understanding and perspective-taking. Our findings suggest that while current VLMs can infer the intentions behind others' actions, they struggle with level-2 perspective-taking. On one hand, this supports the hypothesis that intentionality understanding may not require mental simulation but could instead rely entirely on knowledge-based reasoning. On the other hand, it raises concerns that VLMs lack internal models for reasoning or, at the very least, are unable to leverage them effectively for perspective-taking. This concern is particularly significant given that intentionality understanding improves with model scale, whereas perspective-taking does not. Further research is needed to investigate these findings, as they appear to be crucial for understanding the nature of ToM abilities and their artificial implementations. Exploring the underlying mechanisms behind this dissociation may provide deeper insights into the limitations of current AI models and inform the development of architectures better suited for social reasoning.

REFERENCES

G. E. M. Anscombe. *Intention*. Harvard University Press, 1956.

Ian Apperly. *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Psychology Press, New York, NY, 2010.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *arXiv preprint arXiv:2403.18058*, 2024.

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

Yvonne Barnes-Holmes, Louise McHugh, and Dermot Barnes-Holmes. Perspective-taking and theory of mind: A relational frame account. *The Behavior Analyst Today*, 5(1):15–25, 2004.

Valentina Bianco, Alessandra Finisguerra, and Cosimo Urgesi. Contextual priors shape action understanding before and beyond the unfolding of movement kinematics. *Brain Sciences*, 14(2): 164, 2024.

Marcel Brass, Ruth M Schmitt, Stephanie Spengler, and György Gergely. Investigating action understanding: inferential processes versus action simulation. *Current biology*, 17(24):2117–2121, 2007.

claude. Claude models - anthropic. `https://docs.anthropic.com/en/docs/welcome#models`.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Daniel C. Dennett. *The Intentional Stance*. MIT Press, Cambridge, MA, 1987.

Martin E Ford. The construct validity of egocentrism. *Psychological Bulletin*, 86(6):1169, 1979.

Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.

Chris D Frith and Uta Frith. The neural basis of mentalizing. *Neuron*, 50(4):531–534, 2006.

Gemini. Gemini models— gemini api— google ai for developers. `https://ai.google.dev/gemini-api/docs/models/gemini`.

Mariel K Goddu, Alva Noë, and Evan Thompson. Llms don't know anything: reply to yildirim and paul. *Trends in Cognitive Sciences*, 2024.

Alex Gomez-Marin, Joseph J Paton, Adam R Kampff, Rui M Costa, and Zachary F Mainen. Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nature neuroscience*, 17(11):1455–1462, 2014.

Alison Gopnik and Henry M Wellman. Why the child's theory of mind really is a theory. *Mind & Language*, 7(1-2):145–171, 1992.

Anna Gunia, Sofiia Moraresku, and Kamil Vlček. Brain mechanisms of visuospatial perspective-taking in relation to object mental rotation and the theory of mind. *Behavioural Brain Research*, 407:113247, 2021.

Hyowon Gweon. Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in cognitive sciences*, 25(10):896–910, 2021.

Hyowon Gweon and Laura Schulz. 16-month-olds rationally infer causes of failed actions. *Science*, 332(6037):1524–1524, 2011.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

David W Johnson. Cooperativeness and social perspective taking. *Journal of Personality and Social Psychology*, 31(2):241, 1975.

Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

James M Kilner. More than one pathway to action understanding. *Trends in cognitive sciences*, 15 (8):352–357, 2004.

Michal Kosinski. Evaluating large language models in theory of mind tasksg. *arXiv preprint arXiv:2302.02083*, 2023.

Claus Lamm, C Daniel Batson, and Jean Decety. The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. *Journal of cognitive neuroscience*, 19(1):42–58, 2007.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Jennifer Lehmann and Petra Jansen. The relationship between theory of mind and mental rotation ability in preschool-aged children. *Cogent Psychology*, 6(1):1582127, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Shari Liu, Tomer D Ullman, Joshua B Tenenbaum, and Elizabeth S Spelke. Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041, 2017.

Melanie Mitchell and David C Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

Henrike Moll and Andrew N Meltzoff. How does it look? level 2 perspective-taking at 36 months of age. *Child Development*, 82(2):661–673, 2011.

OpenAI. Models - openai api. `https://platform.openai.com/docs/models/gpt-4o`.

Jean Piaget. *The Development of Thought: Equilibration of Cognitive Structures*. Viking Press, 1977.

Jean Piaget and Bärbel Inhelder. *The Child's Conception of Space*. Routledge, London, 1957.

Jean Piaget and Bärbel Inhelder. *The Psychology of the Child*. Basic Books, New York, 1969.

David G. Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1978.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024.

David M. Rosenthal. *The Nature of Mind*. Oxford University Press, New York, 1991.

John R Searle. The intentionality of intention and action. *Inquiry*, 22(1-4):253–280, 1979.

John R Searle. Minds, brains and programs. *Behavioral and Brain Sciences*, 1980.

Karen Shanton and Alvin Goldman. Simulation theory. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(4):527–538, 2010.

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. *arXiv preprint arXiv:2408.12574*, 2024.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. Agent: A benchmark for core psychological reasoning. In *International conference on machine learning*, pp. 9614–9625. PMLR, 2021.

James W A Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S A Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Iris Van Rooij, Olivia Guest, Federico Adolfi, Ronald de Haan, Antonina Kolokolova, and Patricia Rich. Reclaiming ai as a theoretical tool for cognitive science. *Computational Brain & Behavior*, 7(4):616–636, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Henry M. Wellman. *The Child's Theory of Mind*. MIT Press, Cambridge, MA, 1992.

Ilker Yildirim and LA Paul. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*, 2024.

Xuan Zhao, Bertram F Malle, and Hyowon Gweon. Is it a nine, or a six? prosocial and selective perspective taking in four-year-olds. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 38, pp. 924–929, 2016.
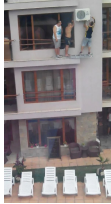
A   LIST OF MODELS EVALUATED

Models			
GPT4V_TURBO_HIGH	GPT4o_HIGH	GPT4o_MINI	InternVL-Chat-V1-1
InternVL2-1B	InternVL2-2B	InternVL2-4B	Mantis-8B-clip-llama3
Mantis-8B-siglip-llama3	Mini-InternVL-Chat-2B-V1-5	Mini-InternVL-Chat-4B-V1-5	MiniCPM-V
MiniCPM-V-2	OmChat	Ovis1.5-Llama3-8B	Phi-3-Vision
Phi-3.5-Vision	Pixtral-12B	Qwen2-VL-2B-Instruct	Qwen2-VL-7B-Instruct
Slime-13B	Slime-7B	Slime-8B	deepseek_vl_1.3b
deepseek_vl_7b	idefics2_8b	llava_next_interleave_7b	llava_next_interleave_7b_dpo
llava_next_llama3	llava_next_mistral_7b	llava_onevision_qwen2_0.5b_si	llava_onevision_qwen2_7b_ov
llava_v1.5_13b	llava_v1.5_7b	mPLUG-Owl3	qwen_chat
sharegpt4v_7b			

Table 1: List of Models

B ADDITIONAL EXAMPLES

A

What does the man holding a sign in the picture intend to do?
A. To repair the ceiling
B. To hold a drill
C. To get down from ceiling
Correct Answer: B
GPT4o: B In the picture, one person is standing on the other person's shoulders while holding a drill and appearing to be focused on the ceiling this suggests that they are attempting to reach and possibly repair or work on something involving the ceiling.

B

What do the people in the picture intend to do?
A. To eat a small fan
B. To cool herself
C. To cool the hot noodles faster
D. To drink soup
Correct Answer: C
GPT4o: C The person in the picture is using chopsticks attached to a small fan to cool down the noodles before eating them the fan is positioned in such a way to blow air on the noodles and reduce their temperature

C

What do the people in the picture intend to do with the ladder?
A. To access an air conditioning unit on the exterior wall
B. To walk back from the wall where the air conditioning unit was mounted
C. To climb up to the third floor
D. None of above
Correct Answer: A
GPT4o: A The people in the picture are standing on a ladder that is placed horizontally across the railing and the wall of the building one person is working on the air conditioning unit that is mounted on the exterior wall the position and activity of the individuals indicate that the ladder is being used to gain access to the air conditioning unit for repair or installation purposes

D

What does the woman in the picture intend to do?
A. To climb over the turnstile
B. To jump over the turnstile
C. To ride on a man's neck
D. *To avoid paying the fare
Correct Answer: D
GPT4o: D To avoid paying the fare: the image shows a woman being carried on a man's shoulders as they move through a turnstile, which is typically used to control access in places like subway stations or amusement parks the positioning suggests that they are trying to bypass the turnstile without paying the fare associated with passing through it

E

What do the people in the picture intend to do?
A. To swim at home
B. Tried to catch a fish
C. To practice diving in a fish tank
D. To clean the fish tank
Correct Answer: D
GPT4o: D The person in the fish tank is equipped with what appears to be cleaning tools or equipment, and they seem to be scrubbing or handling parts of the tank their position inside the tank and the actions they are performing indicate they are likely cleaning it

F

What do the people in the picture intend to do?
A. To put bottles on the ground
B. To keep track of who owns what piece
C. To sit in a sofa
D. To play floor chess
Correct Answer: D
GPT4o: D The objects on the checkered floor are arranged similar to pieces on a chessboard, suggesting an intention to play a game of chess using various household items as substitutes for traditional chess pieces

Figure 6: Additional Examples of Vision Language Models Assessed with IntentBench: Correct
.

A

What does the man holding a sign in the picture intend to do?
A. To see individual pixels on that sign.
B. To tell the others to behave safely.
C. To deliver a big signage.
D. To ride a motorcycle.
Answer: C
GPT4o: B The sign the man is holding shows safety instructions and rules, indicating his intention is to convey a message about safety, specifically the importance of wearing a hard hat and observing other safety protocols this aligns with choice b.

B

What does the people in red T-shirt in the picture intend to do?
A. To work on something located high up on the wall.
B. To position the ladder across a stairwell, supported at one end by the handrail.
C. To have a man holding the ladder.
D. To put the ladder against the wall.
Answer: A
GPT4o: B The person in the picture is on a ladder that is positioned across a stairwell, with one end of the ladder supported by the handrail this setup is being used to reach and work on the high wall, but it is not a safe or recommended method

C

What does the people in the picture intend to do?
A. To restrain a child in a chair.
B. To hold the child.
C. To save the child out of the chair.
D. To bully the child.
Answer: C
GPT4o: A The people in the picture appear to be securing a child into a specialized chair or device the presence of straps and the focused manner in which they are working suggest an effort to ensure the child remains safely and securely seated this kind of setup is often used for safety and support, possibly in a therapeutic or medical context .

D

What does the people in the picture intend to do?
A. To throw a man upside down into a gutter
B. To clean a gutter without using ladders
C. To haze the one in the gutter
D. None of above
Answer: B
GPT4o: D In the picture, the two people appear to be engaged in some form of work related to the manhole cover on the ground they are likely doing maintenance or inspection work there is no evidence to suggest that they intend to throw someone into the gutter, clean the gutter without using ladders, or haze anyone the individuals are likely performing a standard task that is part of their job, such as checking for repairs or clearing out debris

Figure 7: Additional Examples of Vision Language Models Assessed with IntentBench: Wrong
.