Egocentric Bias in Vision-Language Models

Maijunxian Wang[*‡1], **Yijiang Li**[*2], **Bingyang Wang**[3], **Tianwei Zhao**[4], **Ran Ji**[5], **Qingying Gao**[†6], **Emmy Liu**[†7], **Hokin Deng**[†8] **& Dezhi Luo**[‡9]

[1]Cognitive Science Program, University of California, Berkeley
[2]Department of Electrical and Computer Engineering, University of California San Diego
[3]School of Computer Science, Georgia Institute of Technology & Emory University
[4]Department of Computer Science, Johns Hopkins University
[5]Department of Cognitive Science, University of California San Diego
[6]Department of Computer Science & Wilmer Eye Institute, Johns Hopkins University
[7]Language Technologies Institute, Carnegie Mellon University
[8]Robotics Institute, Carnegie Mellon University
[9]Weinberg Institute for Cognitive Science, University of Michigan

Abstract

Visual perspective taking-inferring how the world appears from another's viewpoint-is foundational to social cognition. We introduce FlipSet, a diagnostic benchmark for Level-2 visual perspective taking (L2 VPT) in vision-language models. The task requires simulating 180-degree rotations of 2D character strings from another agent's perspective, isolating spatial transformation from 3D scene complexity. Evaluating 103 VLMs reveals systematic egocentric bias: the vast majority perform below chance, with roughly three-quarters of errors reproducing the camera viewpoint. Control experiments expose a compositional deficit-models achieve high theory-of-mind accuracy and above-chance mental rotation in isolation, yet fail catastrophically when integration is required. This dissociation indicates that current VLMs lack the mechanisms needed to bind social awareness to spatial operations, suggesting fundamental limitations in model-based spatial reasoning. FlipSet provides a cognitively grounded testbed for diagnosing perspective-taking capabilities in multimodal systems.

Keywords: visual perspective taking; vision-language models; egocentrism; theory of mind; mental rotation

Introduction

Visual perspective taking (VPT), the ability to reason about how a scene appears from another's point of view, is foundational to both human social cognition and situated artificial intelligence. It underlies skills such as interpreting spatial instructions, anticipating occluded views, and understanding what others can or cannot see. Cognitive science distinguishes between two levels of this ability. Level-1 (L1) VPT involves recognizing whether an object is visible from a particular viewpoint, while Level-2 (L2) VPT concerns how that object appears—for example, understanding that a "6" might look like a "9" from across the table (Zhao et al., 2016). Unlike L1, L2 VPT requires mentally transforming spatial representations, often through mental rotation or embodied simulation (Gallese & Goldman, 1998; Shepard & Metzler, 1971), and typically emerges later in human development (Edwards & Low, 2019; Moll & Meltzoff, 2011). The development of L2 VPT, described by Jean Piaget (Piaget, 1954) as "the elimination of egocentrism," is regarded as a developmental landmark in children's social cognition (Flavell, 2013).

Recent advances in Vision-Language Models (VLMs) (Alayrac et al., 2022; J. Li et al., 2023; Radford et al., 2021; Y. Zhang et al., 2024) have demonstrated remarkable capabilities in perception (Chen et al., 2025; Z. Cheng et al., 2024; Jiang et al., 2025; Team et al., 2025; P. Wang et al., 2024), reasoning (K. Cheng et al., 2024; G. Xu et al., 2024; R. Zhang et al., 2024), and visual commonsense understanding (Park et al., 2020; Zellers et al., 2019). These advancements prompt the considerations of whether VLMs are capable of situated social reasoning that necessitates taking the perspective of another agent when it conflicts with their own. In this respectm, L2 VPT poses a particularly stringent benchmark, testing whether models can go beyond line-of-sight reasoning to simulate another's visual experience—a capability essential for effective interaction in real-world environments.

In this paper, we introduce FlipSet, a diagnostic benchmark that isolates the spatial transformation component of L2 VPT while controlling for basic theory-of-mind demands. Each item shows a printed card with 2D symbols (e.g., "81") and a plush monkey on the opposite side, facing the card's back. The model is asked: "What does the monkey see on the card?" To answer correctly, models must mentally rotate the card 180°, simulating the monkey's perspective. By using 2D strings that transform under rotation rather than complex 3D scenes, we minimize confounding spatial complexity. Crucially, we introduce control conditions that explicitly separate theory-of-mind recognition (understanding that another's view differs) from mental rotation capabilities (transforming spatial representations), enabling precise diagnosis of where models fail in the L2 VPT cognitive pipeline.

We evaluate 103 publicly available VLMs under unified zero-shot conditions, analyzing not only accuracy but the distribution of error types through systematic answer choice design: correct, egocentric (camera viewpoint), confusable (visually similar), and random. We find that 91.3% of models perform below the 25% chance level, with 75.88% of errors being egocentric—models simply reproduce what the camera

*Equal Contributions
†Equal Advising
‡Correspondence: mjxwang@berkeley.edu; ihzedoul@umich.edu

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

sees. Chain-of-Thought prompting fails to mitigate and often amplifies this bias. However, control experiments on 24 models reveal a more nuanced picture: models achieve high theory-of-mind accuracy (90.4%), recognizing that agents see differently, yet perform near chance on isolated mental rotation (26.1%) and catastrophically on L2 VPT (10.3%). Critically, most models show L2 VPT performance below what their component abilities predict, revealing a compositional deficit—models possess cognitive building blocks but cannot integrate them in situated reasoning contexts.

Our contributions are threefold:

- We introduce FLIPSET, a controlled benchmark for L2 VPT using 2D rotation that isolates spatial transformation from 3D complexity and theory-of-mind recognition from mental rotation capabilities, enabling the first large-scale evaluation (103 VLMs) that can precisely diagnose component failures.
- Through systematic answer choice design distinguishing correct, egocentric, confusable, and random responses, we reveal dominant egocentric bias (75.88% of errors).
- Via controlled experiments separating theory-of-mind, mental rotation, and L2 VPT, we provide behavioral reductionistic evidence of a compositional deficit: models recognize perspective differences (ToM: 90.4%) and perform above-chance mental rotation in isolation (MR: 26.1%), yet fail catastrophically when integration is required (L2 VPT: 10.3%), illuminating fundamental limitations in current VLM architecture.

Related Work

Human Cognitive Science. VPT plays a foundational role in social cognition and spatial reasoning. Cognitive science distinguishes between L1 VPT—judging what is visible from another's viewpoint—and L2 VPT—judging how an object appears from that viewpoint. Crucially, L2 VPT is a composite ability that integrates two distinct cognitive requirements: it demands not only basic theory of mind understanding that others perceive the world differently, but also the capacity to perform structured mental operations that actively transform spatial representations (Edwards & Low, 2019; Flavell, 2013; Moll & Meltzoff, 2011). This dual requirement situates L2 VPT within Piaget's Concrete Operational Stage (Piaget, 1954, 1977), the developmental period during which children acquire the ability to mentally manipulate spatial relationships through reversible, systematic operations. The mental transformation component involves model-based simulation rather than static perception, as evidenced by classic studies showing that humans incur a linear reaction time cost proportional to the degree of rotation required (Shepard & Metzler, 1971; Tarr & Pinker, 1989). Neuroimaging studies further demonstrate that L2 perspective taking activates the intraparietal sulcus and premotor cortex—regions associated with embodied simulation of spatial transformations and another person's position (Gallese, 2007; Gunia et al., 2021; Zacks, 2008). This composite architecture—requiring both perspective awareness and spatial transformation capabilities—distinguishes L2 VPT from simpler visibility judgments and makes it a principled probe for embodied social understanding in artificial systems.

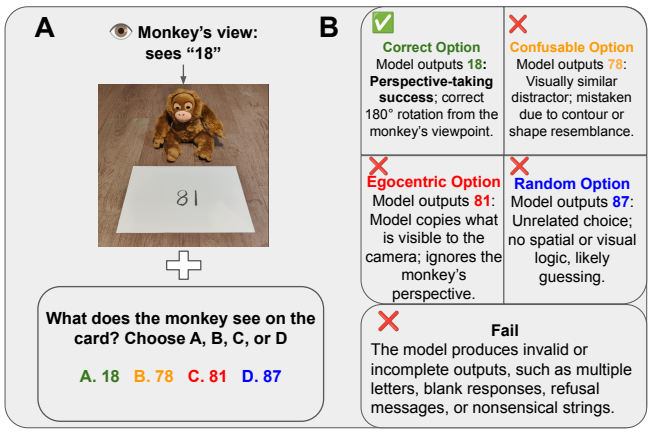
Cognitive Benchmarking in VLMs. Recent benchmarks have attempted to assess VPT in VLMs but suffer from confounding factors that prevent clear diagnostic interpretation. CLEVR-PT (Singh et al., 2023) and Spatial-VQA (X. Li et al., 2024) introduce multi-camera or 3D setups that conflate viewpoint simulation with depth perception, object occlusion, and multi-object tracking. PerspectBench (Gao et al., 2024) (part of CoreCognition (Y. Li et al., 2025)), COMFORT (Z. Zhang et al., 2024), and Omni-Perspective (B. Wang et al., 2025) assess L2 VPT using controlled and real-life 3D spatial arrangements that are primarily inspired by Piaget's three-mountain task (Piaget, 1977), where children must describe how a scene of three mountains appears from different viewpoints. However, precisely because L2 VPT comprises both theory of mind understanding and spatial transformation abilities, benchmarks relying on complex 3D configurations introduce a critical diagnostic ambiguity: when models fail, we cannot determine whether the deficit stems from inability to recognize that others see differently (ToM), inability to mentally transform spatial representations (mental rotation), or both. To address this limitation, we design FLIPSET to isolate the spatial transformation component while controlling for basic ToM demands. Our benchmark uses simple two-dimensional symbols that appear different under 180-degree rotation, minimizing spatial complexity as a confounding factor, and introduces control conditions that explicitly separate ToM recognition from mental rotation capabilities, enabling precise diagnosis of where models fail in the L2 VPT cognitive pipeline.

Methods

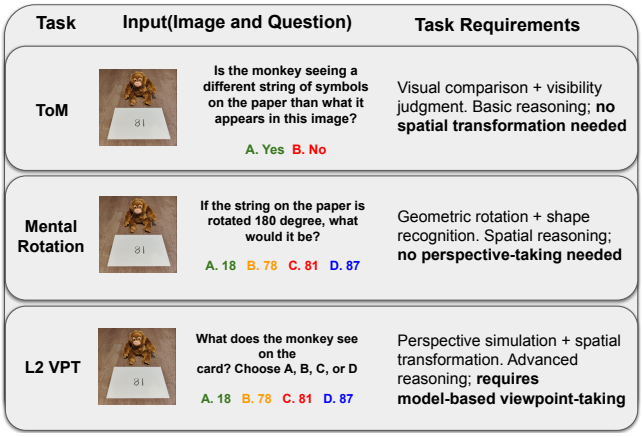
Main Experiments

We introduce FLIPSET, a diagnostic benchmark designed to evaluate L2 VPT in VLMs. Building on the seminal 6→9 rotation paradigm introduced by Zhao et al. (2016), we expand this approach into a systematic benchmark that isolates perspective taking from confounding spatial complexity. Unlike the three-mountain paradigm and its computational variants that rely on complex 3D spatial arrangements (Gao et al., 2024; Piaget, 1977), we use 2D strings that transform under 180° rotation (e.g., "81"→"18", "pond"→"puod"). This design reduces cognitive load from depth perception, occlusion, and object tracking, enabling us to strictly isolate the mental rotation component of L2 VPT from broader 3D spatial understanding.

Each FLIPSET item consists of an image paired with a multiple-choice question, as illustrated in Figure 1a. The image shows a white A4 card (21×30 cm) with centrally printed black Arial characters, placed upright on a wooden floor. The camera faces the card front-on, while a plush monkey sits on the opposite side, facing the card's back. To answer correctly, models must mentally rotate the card 180° to simulate the monkey's viewpoint, requiring genuine perspective transformation rather than surface pattern recognition.



(a) Main Experiments



(b) Control Experiment

Figure 1: *FlipSet benchmark design and evaluation approach.* **(a) Prompt type and Error types in model responses across cognitive tasks.** Each FlipSet item asks the model what a monkey sees on the back of a card—requiring a 180° mental rotation from the monkey's viewpoint. Answer options correspond to distinct reasoning outcomes: Correct (successful perspective taking), Egocentric (front-view repetition), Confusable (visually similar distractor), Random (unrelated guess), and Fail (invalid or empty output). **(b) Three cognitive tasks comparison.** L1 VPT requires simple visibility judgment, MR involves pure geometric transformation, and L2 VPT demands complex perspective simulation. All tasks use identical visual stimuli.

Category	Example	#Items
Two Digits	81→18	2
Two Digits (mixed)	89→68	2
Three Digits (mixed)	168→891	4
Single Letter	d→p	4
Two Letters	dd→pp	2
Two Letters (mixed)	do→op	4
Three Letters (mixed)	nod→pou	6
Four Letters (mixed)	pond→puod	2
Four Combo (mixed)	W819→618M	2

Table 1: FlipSet question classes ordered by difficulty (28 items total).

FlipSet contains 28 baseline items organized by difficulty, ranging from simple digit reversals to complex alphanumeric transformations (Table 1). Items progress from symmetric digit pairs (e.g., "81"→"18") to letter mirroring (e.g., "nod"→"pou") and mixed alphanumeric rearrangement (e.g., "W819"→"618M"). All strings use Arabic numerals and English alphabet letters—a constraint that ensures most VLMs have substantial training exposure to these characters while enabling tight experimental control. All visual elements—font size, lighting, camera angle, and background—are standardized to minimize perceptual artifacts.

Crucially, we extend beyond prior work through systematic answer choice design that enables diagnostic error analysis. Each item includes four answer choices reflecting distinct cognitive strategies: correct (perspective-transformed answer), egocentric (camera viewpoint), confusable (visually similar distractor), and random (unrelated option). For example, when the card shows "81" from the camera view, the choices are "18" (correct), "81" (egocentric), "78" (confusable), and "87" (random), as shown in Figure 1a. To eliminate positional bias, we systematically permute these four choices across 12 counterbalanced layouts per item, yielding 28 items × 12 layouts = 336 evaluation instances per model. Analysis confirms minimal positional effects across layouts. We classify all model responses into five categories that reveal distinct failure modes beyond binary accuracy:

- **Correct** — Outputs "18," successfully simulating the rotated viewpoint through mental transformation.
- **Egocentric** — Outputs "81," reproducing the camera view and failing to adopt the monkey's perspective.
- **Confusable** — Selects "78," showing partial visual reasoning based on shape similarity to the correct answer without completing the perspective transformation.
- **Random** — Outputs "87," an unrelated option with no plausible geometric relation, indicating arbitrary guessing.
- **Fail** — Produces invalid, empty, or nonsensical outputs (e.g., "ABC" or "I don't know"), reflecting instruction-following failures rather than visual reasoning.

Control Experiments

To systematically dissociate the cognitive mechanisms underlying L2 VPT, we designed control experiments that isolate three distinct components: Theory of Mind (ToM), Mental Rotation (MR), and L2 VPT itself. As illustrated in Figure 1b, these tasks employ identical visual stimuli but vary cognitive demands through targeted prompts.

ToM isolates basic social awareness—recognizing that another agent's view differs from one's own—through a simple visibility judgment requiring only visual comparison without spatial transformation. We use "ToM" here as a simplified designation rather than implying full human theory of mind reasoning. **MR** isolates pure geometric transformation, requiring

	Llama	LLaVA	BLIP	InternVL	Video	DeepSeek	InternLM	Molmo	VILA	Gemma	Janus	Qwen
# Models	15	12	8	6	5	5	5	4	4	4	4	3
Params	7-90B	2-72B	3-13B	1-38B	7-13B	3-28B	7-8B	1-72B	3-40B	3-27B	1-7B	3-72B
Acc (%)	4.8	9.3	13.8	12.0	10.6	4.5	1.4	6.1	8.6	13.2	4.5	17.3
Ego (%)	88.4	72.6	62.2	70.8	57.0	93.9	97.3	79.2	80.7	75.5	91.6	53.9
	Parrot	**Moondream**	**Yi-VL**	**ChatUniVi**	**EMU**	**MiniGPT**	**MiniMonkey**	**XVERSE**	**WeMM**	**VisualGLM**	**TransCore**	**Falcon**
# Models	2	2	2	2	2	1	1	1	1	1	1	1
Params	7-14B	2B	6-34B	8-13B	8-37B	13B	2B	13B	7B	6B	13B	11B
Acc (%)	1.5	3.9	0.4	17.6	2.2	19.0	0.9	22.9	9.5	6.0	33.0	25.0
Ego (%)	91.7	87.5	98.5	60.7	90.3	41.4	97.3	24.2	87.5	34.8	33.3	50.0
	Pixtral	**Phi**	**Fuyu**	**Other**	**OpenFlam.**	**OmniLMM**	**NVLM**	**GLM**	**Monkey**	**MGM**	**360VL**	
# Models	1	1	1	1	1	1	1	1	1	1	1	
Params	12B	6B	8B	25B	9B	12B	72B	9B	7B	7B	70B	
Acc (%)	6.0	5.4	16.1	7.7	25.6	25.0	31.0	0.6	0.3	1.2	6.8	
Ego (%)	86.0	87.5	39.6	82.7	17.6	50.0	24.7	99.4	99.4	97.0	76.2	

Table 2: Summary of 35 model families evaluated on FlipSet (N=103 models). For each family: number of models evaluated, parameter range, average accuracy, and average egocentric response rate. Models sorted by count within each family.

spatial reasoning without any perspective taking component. **L2 VPT** integrates both: models must adopt the monkey's perspective and mentally transform spatial representations, constituting the composite ability of genuine L2 perspective taking. By systematically varying cognitive demands while holding visual stimuli constant, we isolate each component's contribution and examine their interactions.

Model Evaluations

For the main experiment, we evaluate 103 publicly available VLMs spanning a wide range of model families, parameter sizes (1B to 90B), and architectures. For the control experiments, we evaluate a subset of 24 models across all three tasks (ToM, MR, and L2 VPT) to enable systematic comparison of cognitive components.

All models are evaluated under zero-shot conditions with no fine-tuning or in-context demonstrations. We invoke each model using its native API or open-source implementation, preserving default inference behavior. To standardize evaluation across heterogeneous model types, we developed a lightweight toolkit that formats prompts, invokes inference, parses responses, and maps model outputs to answer categories (Correct, Egocentric, Confusable, Random, Fail) based on ground-truth 180° rotations. All models receive identical image-text inputs. Our evaluation encompasses both direct-response models and those with Chain-of-Thought capabilities, enabling comprehensive behavioral analysis across different reasoning modes while maintaining experimental control.

Results

Main Results: Egocentric Bias

We begin by evaluating overall model accuracy on FlipSet. Each item is formatted as a four-way multiple-choice question, yielding a 25% chance-level baseline. However, as illustrated in Figure 2, 91.3% of models perform substantially below this threshold. Average accuracy across all 103 models is 8.96%, with the best-performing model achieving only 33.9% and median performance at 5.36%—well below chance. These results reveal not merely poor performance but systematic failure: current VLMs are not solving the task via structured spatial reasoning from the agent's viewpoint, but instead rely on superficial heuristics that bypass spatial transformation altogether.

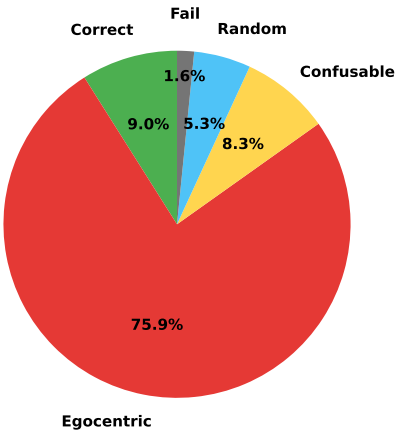


Figure 2: Error type distribution across 103 models on L2 visual perspective taking tasks. Categories include Correct, Egocentric, Confusable, Random, and Fail responses. Complete results by model families are presented in Table 2.

The error distribution reveals the mechanism underlying this failure. Egocentric responses—where models reproduce the string visible from the camera's viewpoint while disregarding the monkey's perspective—account for 75.88% of all answers. In contrast, correct answers comprise only 8.96%, while confusable (8.29%), random (5.30%), and fail (1.57%) responses remain comparatively rare. This pattern provides diagnostic evidence: if models were attempting mental rotation but failing at visual recognition, we would expect higher rates of confusable errors. Instead, the overwhelming dominance of egocentric responses indicates that models default to their own visual reference frame without engaging in spatial transformation. This "**egocentric bias**" persists regardless of model architecture or training approach.

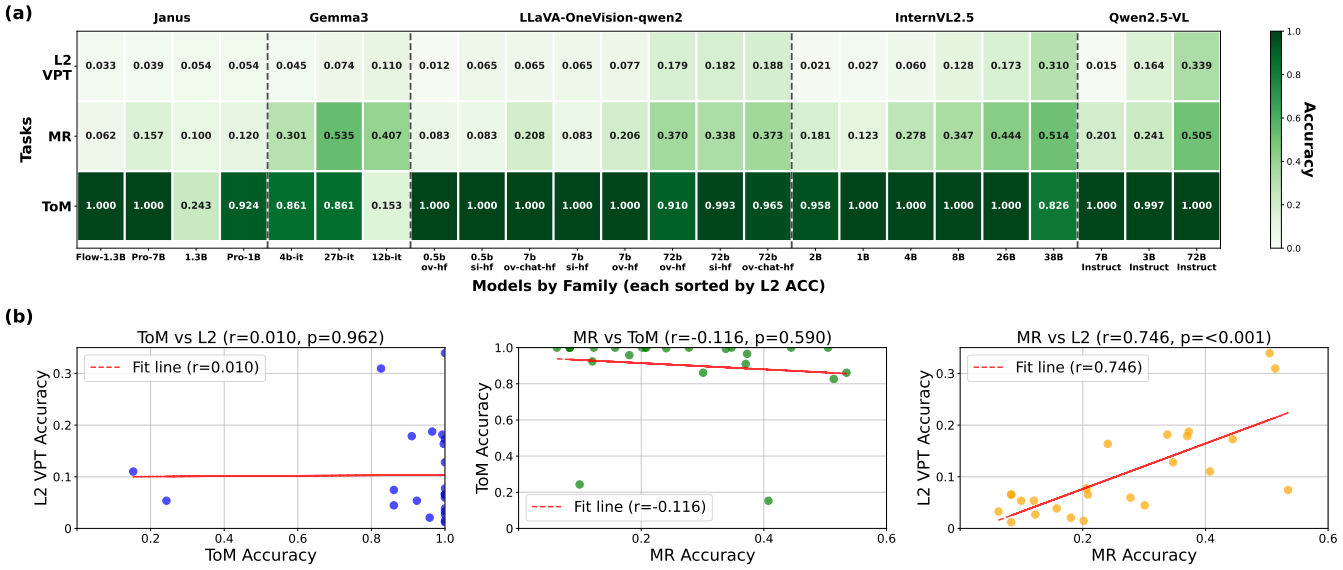


Figure 3: **Control Experiment Results: Model Performance and Task Correlations. (a)** Performance comparison of 24 models on three cognitive tasks: ToM (theory of mind), L2 (Level 2 perspective taking), and MR (mental rotation). Models are grouped by family and sorted by L2 accuracy within each family. **(b)** Correlation analysis between the three cognitive tasks, showing the relationships between ToM, L2 VPT, and MR performance across the 24 models.

To examine whether this bias extends to models' reasoning processes, we analyzed representative Chain-of-Thought traces from models that generate explicit reasoning steps. Qualitative analysis reveals that even when models articulate reasoning, they frequently invoke egocentric assumptions or superficial visual heuristics rather than accurate mental rotation. These findings demonstrate that egocentric bias persists even under structured reasoning approaches.

Control Results

The control experiment reveals the cognitive mechanisms underlying models' egocentric bias. We evaluated 24 models across three tasks (ToM, MR, L2 VPT) to dissociate component abilities. As shown in Figure 3a, most models exhibit a consistent performance hierarchy: ToM > MR > L2 VPT.

Models demonstrate excellent ToM performance (mean: 90.4%), indicating they recognize that agents positioned differently see different things. However, MR performance remains modest (mean: 26.1%, barely above the 25% chance baseline), and L2 VPT performance is markedly lower (mean: 10.3%). Notably, architectural specialization emerges: LLaVA-OneVision models excel at ToM (mean: 98.4%) but struggle with MR (mean: 21.8%), while Gemma models show weaker ToM (mean: 62.5%) but stronger MR (mean: 41.4%).

Correlation analysis across the 24 models (Figure 3b) reveals critical cognitive dissociations. ToM and L2 VPT show no correlation ($r = 0.010$, $p = 0.962$), as do MR and ToM ($r = -0.116$, $p = 0.590$), indicating these are independent processes. However, MR and L2 VPT correlate strongly ($r = 0.746$, $p < 0.001$), confirming that mental rotation is necessary for L2 perspective taking.

Most critically, we observe a systematic **compositional deficit**: L2 VPT performance consistently falls below what would be expected from models' component abilities. If models could successfully integrate ToM and MR, L2 VPT accuracy should approximate their product (ToM × MR). However, examining individual models reveals this is not the case. For instance, Qwen2.5-VL-72B-Instruct achieves perfect ToM (1.000) and above-chance MR (0.505), yet its L2 VPT performance (0.339) falls substantially below the expected 0.505. Similarly, InternVL2.5-38B demonstrates ToM = 1.000 and MR = 0.514, but achieves only 0.310 on L2 VPT—a 40% deficit from the expected 0.514. This pattern holds across model families: 22 of 24 models (91.7%) show L2 VPT performance below the factorial prediction, with a median deficit of 52.3% relative to expected performance. Strikingly, 16 models (66.7%) achieve above-chance MR performance (>0.25) while remaining at or below chance on L2 VPT (≤0.25), despite near-perfect ToM scores.

This dissociation reveals that models' L2 VPT failures stem not merely from deficient mental rotation or social awareness in isolation, but from a fundamental inability to **integrate these capacities within a situated reasoning context**. Models possess the constituent abilities—recognizing different viewpoints (ToM) and performing geometric transformations (MR)—yet systematically fail to deploy them jointly when the task demands coordinated perspective taking and spatial transformation.

Discussion

Our findings reveal systematic egocentric bias in VLMs' L2 visual perspective taking: models overwhelmingly reproduce the camera viewpoint rather than simulating the agent's perspective. Our systematic answer choice design—distinguishing cor-

rect, egocentric, confusable, and random responses—enables fine-grained behavioral analysis. Through controlled experiments, we demonstrate this reflects not only failures in mental rotation but also a compositional deficit: models fail to integrate social awareness with spatial reasoning capabilities in situated contexts. This behavioral reductionistic approach, inspired by developmental psychology (Flavell, 2013; Piaget, 1977; Zhao et al., 2016), dissociates component processes to expose the cognitive architecture underlying failures.

The observed pattern echoes aspects of Piaget's account of childhood egocentrism as a signature of the preoperational stage (Piaget, 1954): an inability to coordinate one's own perspective with another's through structured mental operations. Our control experiments reveal an analogous limitation in VLMs—models recognize that agents see differently (high ToM accuracy) yet fail when this awareness must be operationalized through spatial transformation (low L2 VPT despite above-chance MR). Most models show L2 VPT performance below what their component abilities predict, suggesting they possess cognitive building blocks but lack integrative mechanisms to deploy them in context. This illuminates a fundamental question for AI development: does complex perspective taking emerge compositionally from scaling simpler capacities? Our findings indicate that current VLMs, in addition to improvements in spatial reasoning, may require targeted mechanisms for binding social awareness to spatial operations.

Our results illuminate broader concerns regarding the architectural constraints of VLMs. Humans perform L2 VPT through model-based spatial reasoning supported by mental simulations—constructing internal scene models and mentally simulating transformations (Gunia et al., 2021; Shepard & Metzler, 1971; Zacks, 2008). Recent work argues that VLMs rely on coarse-grained visual encodings that may be mechanistically unsuitable for fine-grained spatial reasoning (Fu et al., 2025; Luo, Gao, & Deng, 2025; J. Zhang et al., 2024). Our findings align with this view: the compositional deficit we observe—where models fail to integrate ToM and MR despite possessing both—may reflect VLMs' reliance on learned visual-linguistic associations rather than structured spatial representations that support systematic transformations. This explains why Chain-of-Thought prompting fails to improve L2 VPT (Y. Li et al., 2025; B. Wang et al., 2025): linguistic reasoning operates disconnected from spatial structure, producing fluent but spatially invalid rationalizations that reinforce egocentric biases. These limitations resonate with broader concerns about foundation models' inability to perform structured physical reasoning (Gao et al., 2025; Sun et al., 2024, 2025; R. Wang et al., 2024; Z. Xu et al., 2024).

Addressing these limitations calls interventions at multiple levels: targeted trainings on multi-view or egocentric-to-allocentric data to encourage perspective-invariant representations (Luo, Li, & Deng, 2025; Yin et al., 2025); systems supporting model-based simulation beyond purely pattern-based retrieval; and even novel architectures that enable fine-grained visual encoding or explicit 3D scene representations to provide structured spatial substrates. Diagnostic benchmarks like FlipSet that isolate cognitive components will be essential for tracking progress.

Conclusion

We introduced FlipSet, a diagnostic benchmark for Level-2 visual perspective taking that isolates 180° rotation from another agent's viewpoint. Evaluating 103 VLMs reveals systematic egocentric bias: models overwhelmingly reproduce the camera viewpoint, with Chain-of-Thought reasoning often amplifying rather than mitigating this failure. Our controlled experiments expose the cognitive architecture underlying this bias. Models demonstrate high social awareness (ToM: 90.4%) yet perform near chance on mental rotation (MR: 26.1%) and catastrophically on L2 VPT (10.3%). Critically, L2 VPT performance falls below what component abilities predict, revealing a compositional deficit—models possess cognitive building blocks but cannot integrate them in situated contexts. This echoes findings from developmental psychology that visual perspective taking requires coordinating awareness with structured spatial operations, highlighting that current VLMs need architectural innovations supporting model-based spatial reasoning rather than purely pattern-matching mechanisms.

Acknowledgments

This work was supported by a MiraclePlus Compute Grant to Growing AI Like A Child (https://growing-ai-like-a-child.github.io/).

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: A visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716–23736.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., . . . Wang, W. (2025). Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. https://arxiv.org/abs/2412.05271

Cheng, K., Li, Y., Xu, F., Zhang, J., Zhou, H., & Liu, Y. (2024). Vision-language models can self-improve reasoning via reflection. *arXiv preprint arXiv:2411.00855*.

Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., & Bing, L. (2024). Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. https://arxiv.org/abs/2406.07476

Edwards, K., & Low, J. (2019). Level 2 perspective-taking distinguishes automatic and non-automatic belief-tracking. *Cognition*, *193*, 104017.

Flavell, J. H. (2013). Perspectives on perspective taking. In *Piaget's theory* (pp. 107–139). Psychology Press.

Fu, S., Bonnen, T., Guillory, D., & Darrell, T. (2025). Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*.

Gallese, V. (2007). Before and below 'theory of mind': Embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B*, *362*(1480), 659–669.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501.

Gao, Q., Li, Y., Lyu, H., Sun, H., Luo, D., & Deng, H. (2024). Vision language models see what you want but not what you see. *arXiv preprint arXiv:2410.00324*.

Gao, Q., Pi, X., Liu, K., Chen, J., Yang, R., Huang, X., Fang, X., Sun, L., Kishore, G., Ai, B., et al. (2025). Do vision-language models have internal world models? towards an atomic evaluation. *arXiv preprint arXiv:2506.21876*.

Gunia, A., Moraresku, S., & Vlček, K. (2021). Brain mechanisms of visuospatial perspective-taking in relation to object mental rotation and the theory of mind. *Behavioural Brain Research*, *407*, 113247.

Jiang, Y., Wang, Y., Zhao, R., Parag, T., Chen, Z., Liao, Z., & Unnikrishnan, J. (2025). Videop2r: Video understanding from perception to reasoning. *arXiv preprint arXiv:2511.11113*.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *CONFERENCE*.

Li, X., Wang, Y., & Feng, J. (2024). Spatial-vqa: Benchmarking spatial reasoning in vision–language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Li, Y., Gao, Q., Zhao, T., Wang, B., Sun, H., Lyu, H., Hawkins, R. D., Vasconcelos, N., Golan, T., Luo, D., et al. (2025). Core knowledge deficits in multi-modal language models. *Forty-second International Conference on Machine Learning*.

Luo, D., Gao, Q., & Deng, H. (2025). Rethinking the simulation vs. rendering dichotomy: No free lunch in spatial world modelling. *arXiv preprint arXiv:2510.20835*.

Luo, D., Li, Y., & Deng, H. (2025). The philosophical foundations of growing ai like a child. *arXiv preprint arXiv:2502.10742*.

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

Park, J. S., Bhagavatula, C., Mottaghi, R., Farhadi, A., & Choi, Y. (2020). Visualcomet: Reasoning about the dynamic context of a still image. https://arxiv.org/abs/2004.10796

Piaget, J. (1954). *The construction of reality in the child*. Routledge.

Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures*. Viking Press.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv: 2103.00020*.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Singh, Y. K., Kaur, G., & Batra, D. (2023). Rotated CLEVR: Assessing spatial reasoning of vision–language models. *Proceedings of the 17th Pacific Rim International Conference on Artificial Intelligence*.

Sun, H., Gao, Q., Lyu, H., Luo, D., Li, Y., & Deng, H. (2024). Probing mechanical reasoning in large vision language models. *arXiv preprint arXiv:2410.00318*.

Sun, H., Yu, S., Li, Y., Gao, Q., Lyu, H., Deng, H., & Luo, D. (2025). Probing perceptual constancy in large vision language models. *arXiv preprint arXiv:2502.10273*.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*(2), 233–282.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-b., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., . . . Hussenot, L. (2025). Gemma 3 technical report. https://arxiv.org/abs/2503.19786

Wang, B., Li, Y., Zhou, Q., Leong, H. Y., Zhao, T., Ye, L., Deng, H., Luo, D., & Vasconcelos, N. (2025). Do vision language models infer human intention without visual perspective-taking? towards a scalable" one-image-probe-all" dataset. *ICML 2025 Workshop on Assessing World Models*.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., & Lin, J. (2024). Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. https://arxiv.org/abs/2409.12191

Wang, R., Todd, G., Xiao, Z., Yuan, X., Côté, M.-A., Clark, P., & Jansen, P. (2024). Can language models serve as text-based world simulators? *arXiv preprint arXiv:2406.06485*.

Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., & Yuan, L. (2024). Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al. (2025). Spatial mental modeling from limited views. *Structural Priors for Vision Workshop at ICCV'25*.

Zacks, J. M. (2008). Neuroimaging studies of mental rotation: A meta-analysis and review. *Journal of Cognitive Neuroscience*, *20*(1), 1–19.

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. https://arxiv.org/abs/1811.10830

Zhang, J., Hu, J., Khayatkhoei, M., Ilievski, F., & Sun, M. (2024). Exploring perceptual limitation of multimodal large language models. *arXiv preprint arXiv:2402.07384*.

Zhang, R., Zhang, B., Li, Y., Zhang, H., Sun, Z., Gan, Z., Yang, Y., Pang, R., & Yang, Y. (2024). Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., & Li, C. (2024). Video instruction tuning with synthetic data. https://arxiv.org/abs/2410.02713

Zhang, Z., Hu, F., Lee, J., Shi, F., Kordjamshidi, P., Chai, J., & Ma, Z. (2024). Do vision–language models represent space and how? evaluating spatial frame of reference under ambiguities.

Zhao, X., Malle, B., & Gweon, H. (2016). Is it a nine, or a six? prosocial and selective perspective taking in four-year-olds. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, 1693–1698.