

Beyond Recognition: Evaluating Visual Perspective Taking in Vision Language Models

Gracjan Goral^{1,2,5*} Alicja Ziarko^{1,2,5} Piotr Milos^{1,2,5}

Michał Nauman⁴ Maciej Wolczyk⁵ Michał Kosinski³

¹Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, S. Banacha 2, 02-097 Warsaw, PL

²Institute of Mathematics, Polish Academy of Sciences, J. & J. Śniadeckich 8, 00-656 Warsaw, PL

³Graduate School of Business, Stanford University, Stanford, CA 94305, USA

⁴Robot Learning Lab, University of California, Berkeley, CA 94720, USA

⁵IDEAS NCBR, Chmielna 69, 00-801 Warsaw, PL

*Corresponding author: gp.goral@uw.edu.pl

We investigate the ability of Vision Language Models (VLMs) to perform visual perspective taking using a novel set of visual tasks inspired by established human tests. Our approach leverages carefully controlled scenes, in which a single humanoid minifigure is paired with a single object. By systematically varying spatial configurations – such as object position relative to the humanoid minifigure and the humanoid minifigure’s orientation – and using both bird’s-eye and surface-level views, we created 144 unique visual tasks. Each visual task is paired with a series of 7 diagnostic questions designed to assess three levels of visual cognition: *scene understanding*, *spatial reasoning*, and *visual perspective taking*. Our evaluation of several state-of-the-art models, including GPT-4-Turbo, GPT-4o, Llama-3.2-11B-Vision-Instruct, and variants of Claude Sonnet, reveals that while they excel in scene understanding, the performance declines significantly on spatial reasoning and further deteriorates on perspective-taking. Our analysis suggests a gap between surface-level object recognition and the deeper spatial and perspective reasoning required for complex visual tasks, pointing to the need for integrating explicit geometric representations and tailored training protocols in future VLM development.

1 Introduction

Recent advances in artificial intelligence have led to the development of Vision Language Models (VLMs) capable of jointly processing visual and textual information (1, 2, 3). These models hold the promise of revolutionizing fields ranging from robotics and autonomous driving (4, 5) to healthcare (6, 7, 8). Yet many of these applications require more than mere visual recognition and language comprehension - they demand the capacity to infer how others visually perceive the world. For example, an autonomous vehicle must consider what a nearby driver does or does not see to anticipate potential hazards, while surgical and industrial robots need to assess whether a human coworker can see an object before handing it over.

In humans, the capacity to adopt another’s vantage point is known as visual perspective taking (VPT) (9, 10) and is a fundamental component of theory of mind (11). VPT supports a wide range of cognitive and social functions, including spatial navigation and coordination of joint actions. Its importance is underscored by evidence that deficiencies in VPT are associated with poor navigation and reduced social skills (12, 13). Within the psychological literature, VPT is often conceptualized along two hierarchically organized levels (14): Level 1 focuses on understanding what others can or cannot see (e.g. *Do they see the object?*) while Level 2 involves mentally adopting another’s vantage point to determine how objects appear from that perspective (e.g. *From their viewpoint, is the object to the right or left?*).

Previous work has shown that large language models can differentiate between the beliefs of different story characters, effectively seeing the world through their *minds’ eyes* in the language space (15, 16, 17). In this work, we build on the rich psychological literature on human VPT (18, 19, 20) to study VLMs’ ability to take others’ *visual* perspectives, that is, to see the world through another’s *actual eyes*. Our results, recapped in Figure 1, show that while current VLMs perform well on scene understanding questions, they struggle with questions requiring spatial reasoning and mostly fail in understanding visual perspectives.

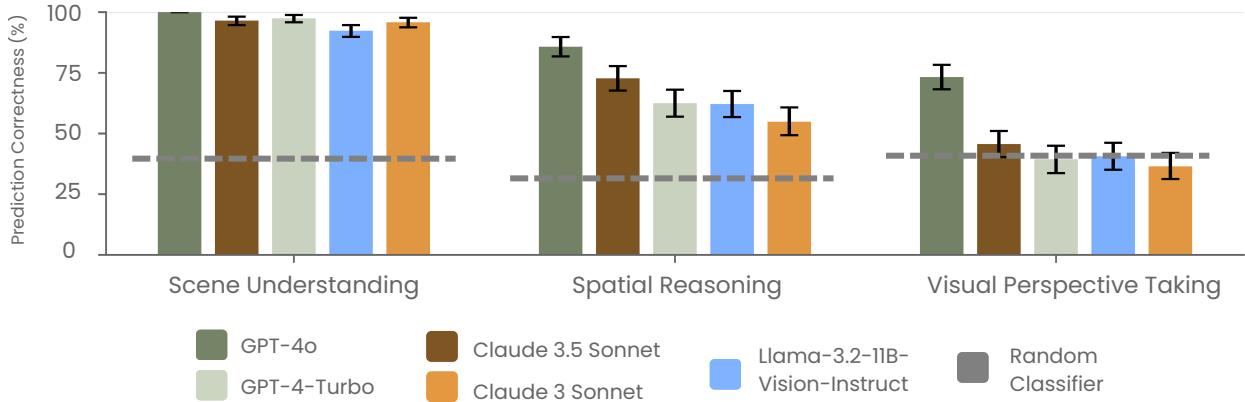


Figure 1: Prediction correctness across three categories of growing difficulty: *scene understanding*, *spatial reasoning*, and *visual perspective taking*. Error bars represent 95% confidence intervals (estimated using bootstrapping (10,000 iterations)). The random classifier is a baseline choosing an answer uniformly at random, see [Appendix A.3](#).

2 Methods

2.1 Visual Tasks

We developed 144 visual tasks inspired by the established human VPT tests ([21, 22](#)). The visual tasks represented a humanoid minifigure and an object placed on the same surface. We used nine distinct humanoid minifigures (varying in hairstyle, clothing, and accessories) and nine distinct objects: a plant, a wardrobe, a cat, a dog, a goblet, a chair, a desk, a bat, and a computer monitor. For each humanoid minifigure-object pair, we systematically varied their spatial positions, see [Figure 2](#). These variations resulted in a total of 144 tasks (9 pairs \times 4 spatial positions \times 2 orientations \times 2 viewpoints), which can be downloaded from [Isle-Brick-V2](#).

Novel visual tasks were developed to ensure the models had not encountered them during training and could not simply recall memorized solutions.

The use of LEGO elements further enabled precise control over image and scene properties without post hoc manipulation. Thus, our approach complemented previous work relying on web-scraped images that may be part of training data ([23](#)) or used abstract visual elements such as dots and arrows overlaid on real images ([24](#)).

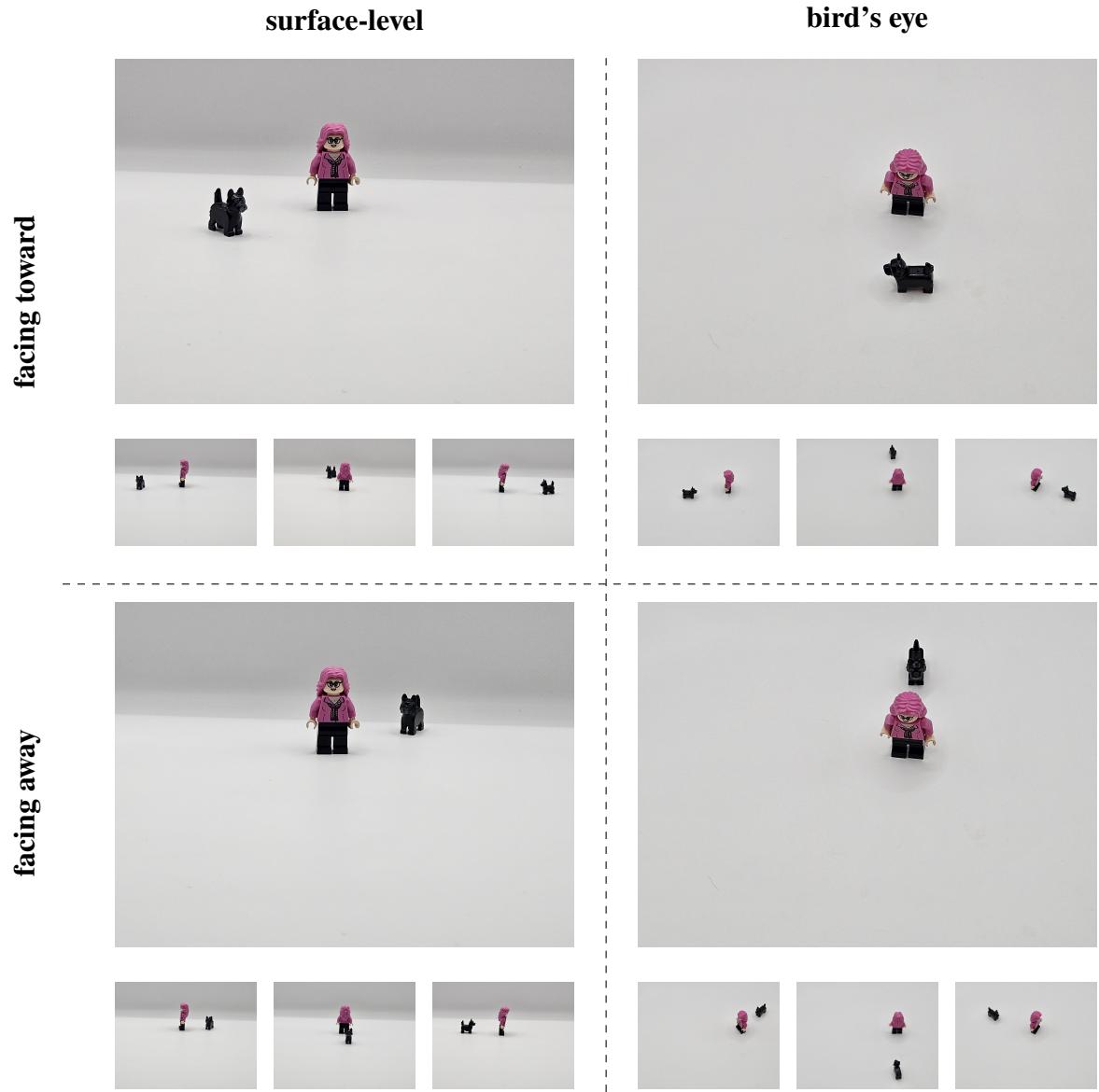


Figure 2: Sixteen tasks involving a single humanoid minifigure–object pair. Tasks vary by the object’s placement (left, right, front, back); the orientation of the humanoid minifigure (facing toward or away from the object); and camera angle (surface-level an bird’s-eye views). All images had the same dimensions, but some are enlarged here for presentation purposes.

2.2 Diagnostic Questions

Question	Ability	GPT-4-Turbo's Answer
Scene Understanding		
Q1: List and count all objects in the image that are not humanoid minifigures.	Object recognition, counting (25)	In the image, there is one object that is not a humanoid minifigure: 1. A small black dog figure. So, there is a total of 1 non-humanoid object in the image.
Q2: How many humanoid minifigures are in the picture?	Object recognition, counting (25)	There is one humanoid minifigure in the picture.
Q3: Are the humanoid minifigure and the object on the same surface?	Object placement (26)	Yes, both the humanoid minifigure and the dog are on the same flat, white surface.
Spatial Reasoning		
Q4: Assuming the top of the image is north, in which cardinal direction (i.e., north, west, east, or south) is the object located relative to the humanoid minifigure?	Spatial reasoning (27)	The dog is located to the west of the humanoid minifigure, assuming the top of the image is north.
Q5: Assuming the top of the image is north, which cardinal direction (i.e., north, west, east, or south) is the humanoid minifigure facing?	Mental rotation, spatial reasoning (28)	The humanoid minifigure is facing east.
Visual Perspective Taking		
Q6: Assuming that the humanoid minifigure can see and its eyes are open, does it see the object?	Understanding what others see (14)	Yes, assuming the humanoid minifigure can see and its eyes are open, it appears to be facing the dog, so it would be able to see the dog.
Q7: From the perspective of the humanoid minifigure, where is the object located relative to it? Please use terms such as front, left, right, or back.	Adopting others' perspective (29)	From the perspective of the humanoid minifigure, the dog is located to its left side.

Table 1: List of the diagnostic questions along an example answer by GPT-4-Turbo's to the task displayed in the top-left corner of Figure 2.

Question Design Each visual task was paired with a set of seven questions, see Table 1. Questions Q1, Q2, and Q3 focused on *scene understanding*: Q1 asked for the total number of objects, Q2 asked for the number of humanoid minifigures, and Q3 inquired whether the humanoid minifigure and objects shared the same surface. Questions Q4, and Q5 tested *spatial reasoning*: Q4 required

identifying the object’s location relative to the humanoid minifigure, and Q5 queried the humanoid minifigure’s facing direction. Finally, Q6 and Q7 evaluated *visual perspective taking*: Q6 checked if the humanoid minifigure could see the object, while Q7 asked for the object’s location from the humanoid minifigure’s perspective.

Open-Ended Format To reduce the influence of guessing and avoid common biases in multiple-choice tasks (e.g., positional biases (30)), all questions were presented in an open-ended format. Each was answered independently in a zero-shot manner (with the context cleared between questions), using a temperature parameter set to zero to minimize variance. The maximum response length was capped at 128 tokens.

Gold-Standard Answers To create the gold-standard answers for each of the 144 visual tasks paired with each of the 7 questions, three research assistants independently responded, with the initial agreement rate exceeding 99%. The rare disagreements were resolved through discussion. Multiple gold-standard answers were allowed for certain questions, e.g., in Q5, where multiple cardinal directions could be correct (north and east for the humanoid minifigure facing northeast). The distribution of gold-standard answers for each question is presented in [Appendix A.1](#).

Evaluation Procedure Two research assistants processed the model’s responses, extracting the key answer components. This included details like object counts (Q1, Q2), spatial relationship confirmations (Q3), line-of-sight judgments (Q6), egocentric locations (Q7), and the cardinal directions (Q4, Q5). Rare disagreements, less than 1%, were resolved by a third person. Combined answers (like northeast) were transformed into sets of basic directions (north, east). We use averaged *prediction correctness* as our performance metrics,¹ which for single answers is equivalent to standard accuracy and can capture partial correctness for combined answers. Details about our evaluation procedure are available in [Appendix A.2](#).

¹For a single task, question pair, its prediction correctness is precision: $P(M, G) = \frac{|M \cap G|}{|M|}$, where M is the set of model-predicted answers and G is the set of gold-standard answers. For instance, if the model predicts *north* ($M = \{\text{north}\}$) and the gold answer is *north* ($G = \{\text{north}\}$), we have $P(M, G) = \frac{|\{\text{north}\} \cap \{\text{north}\}|}{|\{\text{north}\}|} = \frac{1}{2}$, indicating partial correctness.

3 Results and Discussion

Models Evaluated We tested five popular models, including open source model – Llama-3.2-11B-Vision-Instruct (11 December 2024) (31) – and four closed source models – GPT-4-Turbo (9 April 2024), GPT-4o (6 August 2024) (32), Claude 3 Sonnet (29 February 2024), and Claude 3.5 Sonnet (20 June 2024) (33). Models’ performance for each question is presented in Table 2.

Figure 1 presents the performance averaged by categories. All error bars represent 95% CIs calculated using bootstrapping (10,000 iterations).

Question	GPT-4o	Claude 3.5	GPT-4	Llama 3.2-11B	Claude 3
	Sonnet	Turbo	Vision-Instruct	Sonnet	
Scene Understanding					
Q1	100.0% _{-0.0} ^{+0.0}	95.8% _{-3.5} ^{+2.8}	97.2% _{-2.8} ^{+2.1}	98.6% _{-2.1} ^{+1.4}	96.5% _{-3.5} ^{+2.8}
Q2	100.0% _{-0.0} ^{+0.0}	97.9% _{-2.8} ^{+2.1}	95.1% _{-3.5} ^{+3.5}	95.8% _{-3.5} ^{+2.8}	94.4% _{-4.2} ^{+3.5}
Q3	100.0% _{-0.0} ^{+0.0}	95.8% _{-3.5} ^{+2.8}	100.0% _{-0.0} ^{+0.0}	82.6% _{-6.2} ^{+6.2}	96.5% _{-3.5} ^{+2.8}
Spatial Reasoning					
Q4	98.6% _{-2.1} ^{+1.4}	89.9% _{-4.9} ^{+4.5}	83.3% _{-6.2} ^{+5.6}	85.4% _{-5.2} ^{+4.5}	79.9% _{-6.9} ^{+6.2}
Q5	72.9% _{-7.6} ^{+6.9}	55.6% _{-8.3} ^{+8.3}	41.7% _{-7.6} ^{+8.3}	38.9% _{-8.3} ^{+8.3}	29.9% _{-6.9} ^{+7.6}
Visual Perspective Taking					
Q6	87.5% _{-5.6} ^{+4.9}	56.2% _{-8.3} ^{+7.6}	48.6% _{-8.3} ^{+8.3}	49.3% _{-8.3} ^{+8.3}	38.9% _{-8.3} ^{+7.6}
Q7	59.0% _{-7.6} ^{+7.6}	35.1% _{-6.9} ^{+6.9}	30.2% _{-6.9} ^{+7.3}	31.9% _{-7.6} ^{+7.6}	34.0% _{-7.6} ^{+7.6}

Table 2: Prediction correctnesses on diagnostic questions.

Random Classifier Figure 1 also plots a random classifier baseline, i.e. the performance achieved by selecting answers uniformly at random from the permissible answer pool, for more details, see Appendix A.3.

Scene Understanding All five models performed strongly on scene understanding tasks, reflecting their ability to recognize objects and count humanoid minifigures. GPT-4o achieved perfect performance at 100.0%_{-0.0}^{+0.0} prediction correctness, closely followed by GPT-4-Turbo (97.5%_{-1.6}^{+1.4}), Claude-3.5-Sonnet (96.5%_{-1.9}^{+1.6}), and Claude 3 Sonnet (95.8%_{-2.1}^{+1.9}). Llama-3.2-11B-Vision-Instruct also demonstrated high performance (92.4%_{-2.5}^{+2.3}). This suggests that identifying *what* is in the scene – in this instance, how many humanoid minifigures or objects are present, and whether they share the same surface – has become a relatively routine task for modern VLMs.

Spatial Reasoning Models fare considerably weaker on spatial reasoning tasks. Although they performed relatively well when localizing objects relative to the humanoid minifigure (Q4), the

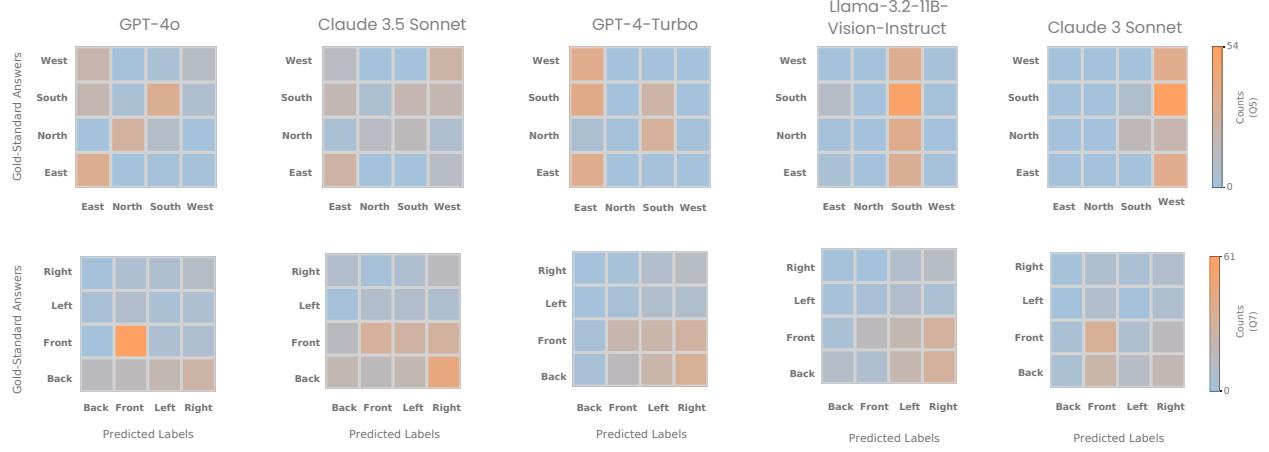


Figure 3: Comparison of co-occurrence matrix for models (columns) across questions Q5 (top row) and Q7 (bottom row). For more details see [Appendix A.5](#).

prediction correctness dropped significantly when the models had to determine the humanoid minifigure’s facing direction (Q5). We hypothesize this discrepancy arises because Q4 involved an extrinsic reference frame, where objects are localized relative to the humanoid minifigure using the fixed cardinal directions established by the image orientation. Q5, however, required understanding an intrinsic reference frame based on the humanoid minifigure’s own orientation, demanding that models interpret body posture cues to determine facing direction.

For Q5, we observed that GPT-4-Turbo, Claude 3 Sonnet, and Llama-3.2-11B-Vision-Instruct were susceptible to *directional bias*, see Figure 3. Namely, they favored certain cardinal directions, for example, GPT-4-Turbo focused on east and south, completely omitting other directions. For this model, we ran additional detailed experiments. Namely, we systematically tested variations such as removing secondary objects, zooming in on humanoid minifigures, explicitly labeling cardinal directions (N, S, E, W) in the visual tasks, and even replacing humanoid minifigures with human faces, see Figure 4. None of these was able to significantly mitigate the GPT-4-Turbo’s directional bias. This suggests that some models may rely on linguistic priors or memorized defaults rather than genuinely engaging in spatial reasoning. See also [Appendix B](#) for more details and an extended discussion.

Visual Perspective Taking This task assessed models on two levels: determining if the humanoid minifigure saw an object (Q6) and identifying the object’s relative position from the humanoid

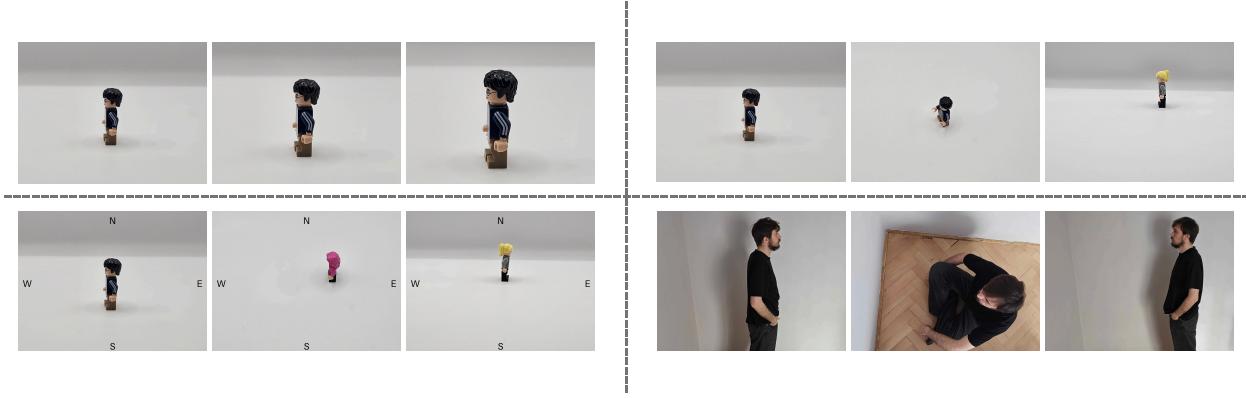


Figure 4: Experiments to investigate GPT-4-Turbo’s persistent east bias in spatial reasoning (Q5). Top-left: Zoomed-in images of humanoid minifigures, testing the impact of increased visual detail. Bottom-left: Images with explicit cardinal direction labels (N, S, E, W) added. Top-right: Images with only a humanoid minifigure (no secondary objects). Bottom-right: Human figures replacing humanoid minifigures. Additionally, we permuted the cardinal directions in the prompt, testing all 24 possible orders of north, south, east, and west.

minifigure’s perspective (Q7). GPT-4o performed well in Q6 ($87.5\%^{+4.9}_{-5.6}$), making fewer errors than its peers. By contrast, Claude 3 Sonnet often failed Q6 by rejecting the question’s premise – insisting a humanoid minifigure *cannot really see* – a recurring error pattern we analyze in detail in [Appendix A.4](#). However, for Q7, all models had difficulties, frequently misclassifying objects located behind the humanoid minifigure as being to its left or right, as indicated on co-occurrence matrices in [Figure 5](#).

This discrepancy between detecting scene content and simulating the humanoid minifigure’s true perspective highlights a critical shortfall in current VLMs. Recognizing objects does not necessarily equate to robust geometric reasoning or an inferential grasp of spatial relationships – cognitive skills in humans linked to mental rotation and perspective taking. Understanding these shortcomings requires further studies. In this work, we performed a simple analysis starting from the observation that both Q6 and Q7 can be perceived as a combination of Q4 and Q5 and very simple reasoning (e.g. for Q6, the answer is positive iff the object is located in the same direction as the humanoid minifigure is facing, these are determined answering Q4 and Q5 resp.). Following that, one could hypothesize that poor performance on Q5 is the root cause of problems with Q6 and Q7. To test it, we added the ground truth Q5 answer (i.e. the humanoid minifigure’s facing direction) to the prompt for Q6. It turned out that this results in modest improvements, suggesting

that the aforementioned decomposition is not entirely accurate. We provide more discussion in [Appendix C](#).

3.1 Conclusions

The gap between VLMs' scene recognition capabilities and their struggles with spatial reasoning, especially visual perspective taking, signals a fundamental limitation beyond mere object identification. This may suggest that current models lack robust mechanisms for geometric transformation and reasoning within an embodied frame of reference, treating scenes more as static inventories than as dynamic, navigable spaces. The nature of the errors, such as confusing relative directions from the humanoid minifigure's viewpoint and the limited impact of providing explicit orientation cues, underscores that the deficit may lie in the ability to compute and utilize perspective-dependent spatial relations. This inability to reliably infer or simulate another's visual perspective critically constrains VLM applicability in collaborative and safety-critical domains, highlighting the need for future architectures capable of constructing and manipulating internal spatial models rather than relying solely on pattern recognition.

References and Notes

1. S. Bubeck, *et al.*, Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR* **abs/2303.12712** (2023), doi:10.48550/ARXIV.2303.12712, <https://doi.org/10.48550/arXiv.2303.12712>.
2. A. Aghajanyan, *et al.*, Scaling laws for generative mixed-modal language models, in *International Conference on Machine Learning* (PMLR) (2023), pp. 265–279.
3. J. Bai, *et al.*, Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond (2023), <https://arxiv.org/abs/2308.12966>.
4. P. Ding, *et al.*, QUAR-VLA: Vision-Language-Action Model for Quadruped Robots, in *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part V*, A. Leonardis, *et al.*, Eds. (Springer), vol. 15063 of *Lecture Notes in Computer Science* (2024), pp. 352–367, doi:10.1007/978-3-031-72652-1_21, https://doi.org/10.1007/978-3-031-72652-1_21.
5. A. J. Sathyamoorthy, *et al.*, CoNVOI: Context-aware Navigation using Vision Language Models in Outdoor and Indoor Environments. *CoRR* **abs/2403.15637** (2024), doi:10.48550/ARXIV.2403.15637, <https://doi.org/10.48550/arXiv.2403.15637>.
6. I. Hartsock, G. Rasool, Vision-Language Models for Medical Report Generation and Visual Question Answering: A Review. *CoRR* **abs/2403.02469** (2024), doi:10.48550/ARXIV.2403.02469, <https://doi.org/10.48550/arXiv.2403.02469>.
7. D. Ferber, G. Wöllein, I. C. Wiest, *et al.*, In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications* **15**, 10104 (2024), doi:10.1038/s41467-024-51465-9.
8. R. Tanno, D. G. Barrett, A. Sellergren, *et al.*, Collaboration between clinicians and vision-language models in radiology report generation. *Nature Medicine* (2024), doi:10.1038/s41591-024-03302-1.
9. J. Piaget, B. Inhelder, *The Child's Conception of Space* (Routledge & Kegan Paul) (1956).

10. S. Baron-Cohen, A. M. Leslie, U. Frith, Does the autistic child have a “theory of mind”? *Cognition* **21** (1), 37–46 (1985).
11. C. M. Heyes, C. D. Frith, The cultural evolution of mind reading. *Science* **344** (6190), 1243091 (2014).
12. C. Orefice, R. Cardillo, I. Lonciari, L. Zoccante, I. C. Mammarella, “Picture this from there”: spatial perspective-taking in developmental visuospatial disorder and developmental coordination disorder. *Front. Psychol.* **15**, 1349851 (2024).
13. A. Pearson, D. Ropar, A. F. de C. Hamilton, A review of visual perspective taking in autism spectrum disorder. *Frontiers in human neuroscience* **7**, 652 (2013).
14. D. Samson, I. A. Apperly, J. J. Braithwaite, B. J. Andrews, S. E. Bodley Scott, Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance* **36** (5), 1255–1266 (2010), doi:10.1037/a0018729.
15. M. Kosinski, Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences* **121** (45), e2405460121 (2024).
16. K. Gandhi, J.-P. Fränken, T. Gerstenberg, N. D. Goodman, Understanding Social Reasoning in Language Models with Language Models (2023), <https://arxiv.org/abs/2306.15448>.
17. J. W. A. Strachan, D. Albergo, G. Borghini, *et al.*, Testing theory of mind in large language models and humans. *Nature Human Behaviour* **8**, 1285–1295 (2024), doi:10.1038/s41562-024-01882-z, <https://doi.org/10.1038/s41562-024-01882-z>.
18. J. M. Loomis, Spatial updating in humans. *Trends in cognitive sciences* **7** (3), 103–111 (2003).
19. D. R. Montello, *The development of spatial cognition and reasoning* (Cambridge University Press) (2005).
20. K. Kessler, K. E. Rutherford, The two forms of visual perspective taking are differently embodied and subserve different spatial prepositions. *Frontiers in Psychology* **5** (2), 102 (2014).

21. C. O’Grady, T. Scott-Phillips, S. Lavelle, K. Smith, Perspective-taking is Spontaneous but Not Automatic. *Quarterly Journal of Experimental Psychology (Hove)* **73**, 1605–1628 (2020), doi:10.1177/1747021820942479.
22. I. Lukosiunaite, A. Kovacs, N. Sebanz, The influence of another’s actions and presence on perspective taking. *Scientific Reports* **14**, 4971 (2024), doi:10.1038/s41598-024-55200-8.
23. S. Tafasca, A. Gupta, V. Bros, J.-M. Odobez, Toward Semantic Gaze Target Detection, in *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024).
24. D. Linsley, *et al.*, The 3D-PC: a benchmark for visual perspective taking in humans and machines. *CoRR* **abs/2406.04138** (2024), doi:10.48550/ARXIV.2406.04138, <https://doi.org/10.48550/arXiv.2406.04138>.
25. E. S. Spelke, K. D. Kinzler, Core knowledge. *Developmental Science* **10** (1), 89–96 (2007), doi:10.1111/j.1467-7687.2007.00569.x.
26. E. S. Spelke, Principles of object perception. *Cognitive Science* **14** (1), 29–56 (1990), doi:10.1207/s15516709cog1401_3.
27. D. H. Uttal, *et al.*, The malleability of spatial skills: a meta-analysis of training studies. *Psychological Bulletin* **139** (2), 352–402 (2013), doi:10.1037/a0028446.
28. R. N. Shepard, J. Metzler, Mental rotation of three-dimensional objects. *Science* **171** (3972), 701–703 (1971), doi:10.1126/science.171.3972.701.
29. A. D. Surtees, S. A. Butterfill, I. A. Apperly, Direct and indirect measures of Level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology* **30** (1), 75–86 (2012), doi:10.1111/j.2044-835X.2011.02063.x.
30. P. Pezeshkpour, E. Hruschka, Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions, in *Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (2023), <https://api.semanticscholar.org/CorpusID:261064970>.

31. Meta AI, Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models (2024), <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>.
32. OpenAI, GPT-4V System Card, <https://openai.com/index/gpt-4v-system-card> (2023), accessed: 2023-08-28.
33. Anthropic, Claude 3 Family, <https://www.anthropic.com/news/clause-3-family> (2023), accessed: 2023-08-28.

Appendix A Data

Appendix A.1 Gold-Standard Answer

In Table 3 we list the gold-standard answers distribution.

Question	Gold-Standard Answers (% of items)
Q1	1 (100.0)
Q2	1 (100.0)
Q3	yes (100.0)
Q4	east (15.3), east, north (10.4), east, south (8.3), north (14.6), north, west (10.4), south (18.1), south, west (6.9), west (16.0)
Q5	east (15.3), east, south (9.7), north (25.0), south (25.0), south, west (7.6), west (17.4)
Q6	no (50.0), yes (50.0)
Q7	back (31.9), back, left (4.9), back, right (13.2), front (39.6), front, left (7.6), front, right (2.8)

Table 3: Gold–standard answer distribution for each question type (Q1–Q7).

Appendix A.2 Prediction Correctnesses

Models sometimes generated compound answers – for instance, *northeast* in Q4 and Q5, or *back and slightly to the left* in Q7 (see Figure 5 and Table 4). Because these responses contained multiple components, our evaluation needed to acknowledge partial as well as fully correct answers.

To tackle this, we employed a precision-based metric that rewarded models for each correctly identified component while tolerating omissions. Assume that we evaluate R responses for a given diagnostic questions and a model, like in Table 2. The score P reported in the table is called *prediction correctness* and is defined as the mean precision across all R responses:

$$P = \frac{1}{R} \sum_{i=1}^R P_i, \quad \text{where } P_i = P(M_i, G_i) = \frac{|M_i \cap G_i|}{|M_i|},$$

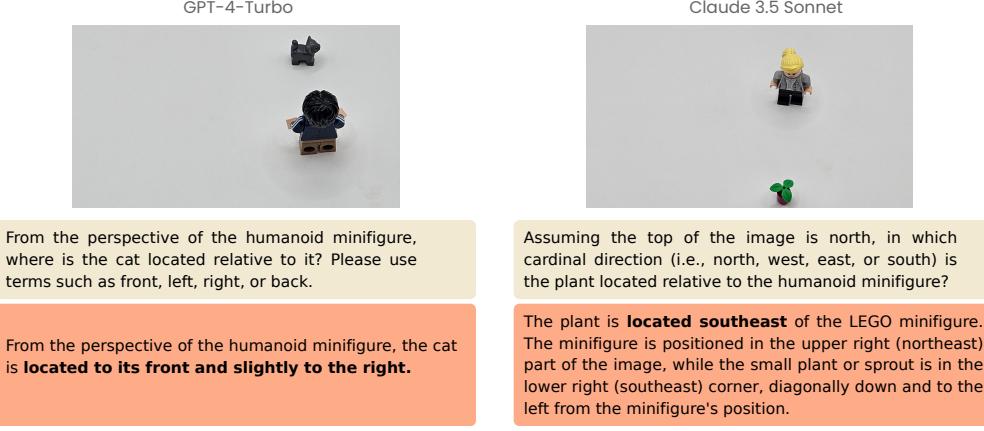


Figure 5: Example model responses to Q5 and Q7 questions. Left: GPT-4 Turbo answers *front and slightly to the right* (gold-standard answer: front). Right: Claude 3.5 Sonnet answers *southeast* (gold-standard answer: south).

where M_i (resp. G_i) is the set of components in the model’s prediction (resp. gold-standard answer) for the i response.

The value of P ranges from 0 to 1. We note that for the questions with single answers, like e.g., Q6, this metric is equivalent to standard accuracy. Moreover, we experimented with several other metrics that take into account partial correctness (e.g., the Jaccard index), and they all yielded similar results.

Model	Q4			Q5			Q7		
	Comp.	Sing.	Disc.	Comp.	Sing.	Disc.	Comp.	Sing.	Disc.
GPT-4-o	21	123	0	0	144	0	10	134	0
GPT-4-Turbo	0	144	0	0	144	0	16	127	1
Claude 3.5 Sonnet	21	117	6	0	144	0	68	76	0
Claude 3 Sonnet	0	127	17	0	144	0	1	121	22
Llama-3.2-11B-Vision-Instruct	33	111	0	1	143	0	2	139	3

Table 4: Frequency of answer types for Q4, Q5, and Q7 across models. *Comp.* (compound) denotes compound answers (e.g., *northwest* to Q5); *Sing.* (Single) denotes single-word answers (e.g., *back* to Q7); *Disc.* (Disclaimer) marks instances in which the model failed to provide a relevant answer (e.g., claiming no object is present when one is).

Appendix A.3 Random Baseline

To establish a chance performance benchmark, we defined a random baseline classifier selecting one answer uniformly at random from the N_q distinct choices available for question type q . For a given instance of question type q , the probability of this random choice being evaluated as correct is $\text{Prob}(\text{success}_q) = \frac{k_q}{N_q}$, where k_q is the number ($k_q \geq 1$) of acceptable answers specified in the gold-standard answers for that instance out of N_q total options. This formulation accurately quantifies the expected success rate of uninformed random guessing under our evaluation protocol accommodating multiple correct answers, based on the empirical distribution of k_q values observed in our gold-standard answers dataset (detailed in [Appendix A.1](#), Table 3). For instance, for Q4 ($N_4 = 4$), where 64% of instances have $k_4 = 1$ ($\text{Prob} = \frac{1}{4}$) and 36% have $k_4 = 2$ ($\text{Prob} = \frac{1}{2}$), the weighted average yields a chance level for this question type of $(0.64 \times \frac{1}{4}) + (0.36 \times \frac{1}{2}) = 0.16 + 0.18 = 0.34$.

Category-level chance performance was computed by averaging the chance levels of the constituent question types. For example, the scene understanding category comprises Q1 (chance = $\frac{1}{3}$), Q2 (chance = $\frac{1}{3}$), and Q3 (chance = $\frac{1}{2}$), resulting in an average category chance level of $(\frac{1}{3} + \frac{1}{3} + \frac{1}{2})/3 = \frac{7}{18} \approx 0.389$. Following this methodology across all categories yields the following random baseline classifier levels: scene understanding (0.389), spatial reasoning (0.317), and visual perspective taking (0.411).

Appendix A.4 Analysis of Q6 Models' Answers

Claude 3 Sonnet showed systematic errors on Q6 in 45/144 instances, unlike other evaluated VLMs. These issues, largely fixed in Claude 3.5 Sonnet, primarily involved rejecting the premise or misidentifying objects. For example, common premise rejection errors included statements like ... *inanimate LEGO toy, it does not possess actual vision....* Similarly, object misclassification occurred, such as when the model stated *There is no dog visible; the black piece appears to be a weapon....*

Appendix A.5 Co-ccurrence Matrix

Our co-occurrence matrixes presented in Section 3 shows how the model's predictions line up with each gold-standard answer. To build it, we take all questions whose gold-standard answers include a

particular label—for example, *north* in Q5—and within that subset simply count how often the model produced *north*, *east*, *south*, or *west*. Those four counts become the row for *north*. Because a single question can have several gold answers and the model may mention several answers at once (such as *northeast* in Q5 or *back and slightly to the left* in Q7), one question can be counted in more than one row or column, so the values in a row may exceed the total number of questions. When every question has exactly one gold label and the model also outputs exactly one label, this co-occurrence table collapses to the ordinary single-label confusion matrix, with each row summing to the number of items.

Appendix B Directional Bias

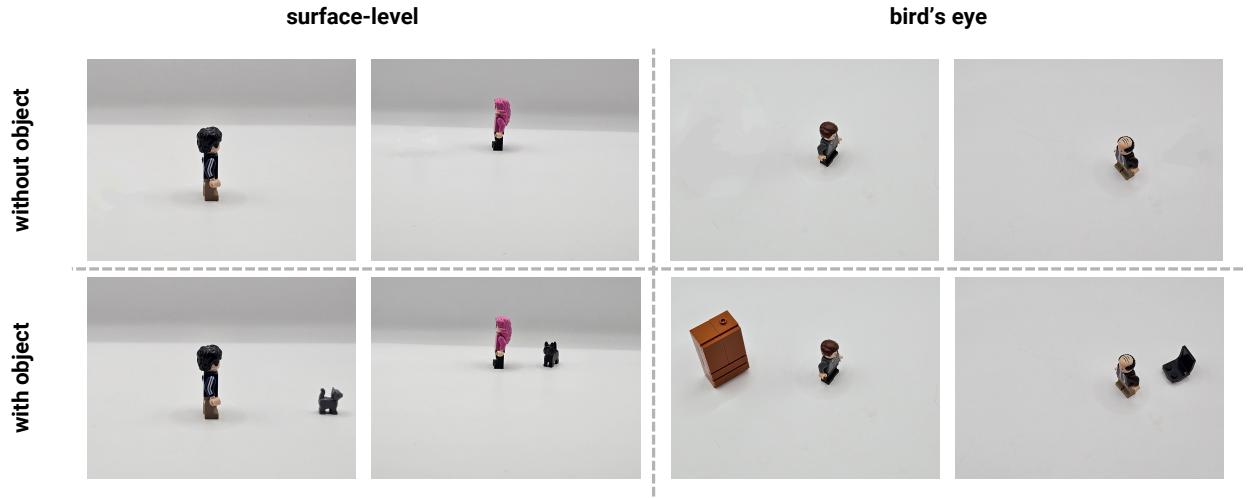


Figure 6: Illustration of the object removal process, used in investigating if the presence of contextual objects impacts the model’s orientation predictions (Q5). Top: visual tasks with objects. Bottom: corresponding visual tasks without objects. Left/Right: surface-level/bird’s-eye.

In Section 3, we described how models struggle with spatial judgments, particularly when determining the facing direction of a humanoid minifigure in Q5. We observed that some VLMs, such as GPT-4-Turbo, suffer from *directional bias* e.g. favoring directions like east or south. We aim to determine the source of the observed directional biases: do they originate primarily from challenges in visual perception, from how the prompt language is interpreted or inherently biased, or from more fundamental issues within the model’s spatial reasoning capabilities?

To this end, we are systematically investigating how specific interventions influence its judgments. We explored two paths: modifying the visual input and manipulating the textual prompts.

Removing Objects We hypothesized that extraneous objects might contribute to GPT-4-Turbo’s directional bias. To test this, we removed items like cats, chairs, etc., from the test images, see Figure 6, and re-ran the 36 visual tasks assessing humanoid minifigure orientation (e.g., *surface-towards-left, birds-eye-towards-right*). Although removing objects caused pointwise changes in 5 predictions and slightly increased the frequency of *south* answers (from 1 to 5), *east* remained the vastly predominant response (31 times). Therefore, object removal did not substantially reduce the

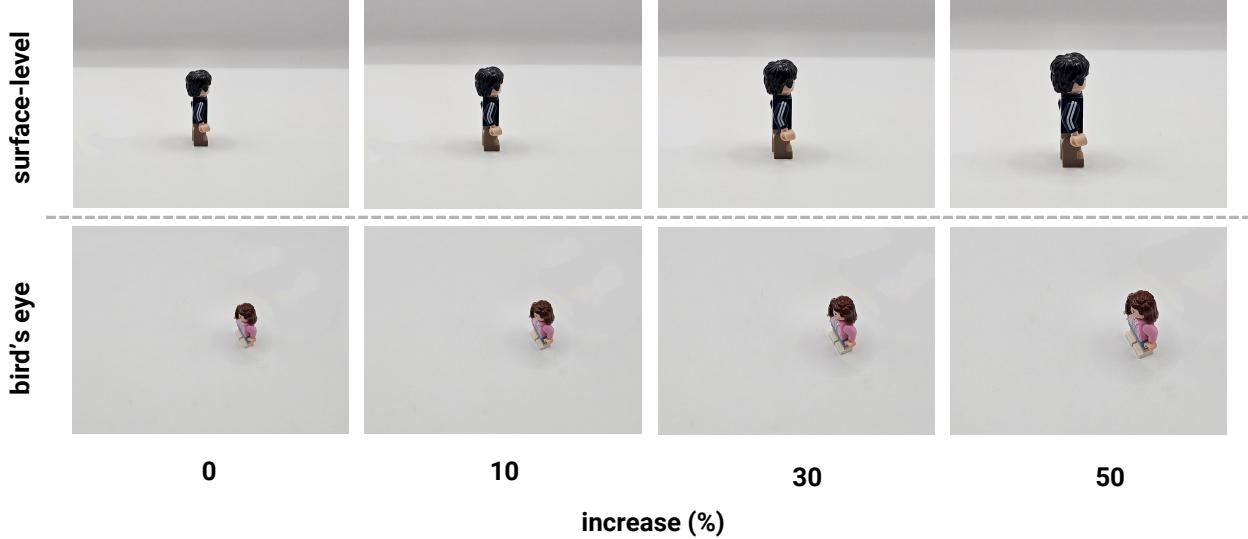


Figure 7: Examples of humanoid minifigure orientation tasks (Q5) at different zoom levels. Left to right: original image with 10% zoom, 30% zoom, and 50% zoom.

model’s tendency towards *east*, indicating this bias is robust and not merely an artifact of contextual items.

Zoom To test if the bias was related to perceiving details, we conducted an experiment gradually zooming in (10%, 30%, 50%) on the visual task without object, see Figure 7. Performance on the 36 tasks remained poor, with prediction correctness consistently around $44.4\%^{+16.7}_{-16.7}$ (e.g., $44.4\%^{+16.7}_{-16.7}/41.7\%^{+16.7}_{-16.7}/47.2\%^{+16.7}_{-16.7}$ for 10%/30%/50% zoom). Importantly, this low accuracy was accompanied by the same persistent directional bias: GPT-4-Turbo continued to predominantly output *east*, irrespective of zoom level. This persistence, even when orientation details were magnified, suggests the bias is not merely a perceptual limitation regarding fine details, but potentially points to deeper issues in the model’s spatial reasoning.

On-Visual Task Cardinal Hints To isolate reference-frame ambiguity, we overlaid *N E S W* markers on each visual task, as illustrated in Figure 8. Even with these cues, prediction correctness remained low at $34.3\%^{+14.3}_{-17.1}$, and the model still selected *east* in 27 of the 36 trials. Because every stimulus explicitly provided both the scene geometry and its coordinate system, any residual error could have arisen from the model’s internal mapping between visual layouts and directional tokens



Figure 8: Surface-level (left) and bird’s-eye (right) views of the visual task, showing *N S E W* marks on the image used for the on-visual task cardinal hints experiment.

rather than from mis-perceiving *north*. In this light, the persisting bias may reflect a hard-wired prior embedded somewhere in the model’s spatial-reasoning pipeline.

Influence of Subject Type To investigate if the observed directional bias was specific to the subject type (i.e., a plastic humanoid minifigure versus a real person), we conducted a small-scale experiment. We used 8 images of a real person, consisting of 4 surface-level and 4 bird’s-eye view shots, as shown in Figure 9. Accordingly, we modified the prompt to ask about the direction the *person* was facing, replacing *humanoid minifigure*. Notably, for all 8 images, GPT-4-Turbo responded that the person was facing east (e.g., answering *The person in the image is facing east*). This unanimous east prediction mirrored the strong bias observed with the humanoid minifigure images. Consequently, this persistence suggests that the directional bias is not solely attributable to the specific nature or structure of the humanoid minifigure itself, but may stem from a more general aspect of the model’s processing.

Influence of Prompt Order (Permutations) To assess sensitivity to the order of options, we tested the model on 69 visual tasks using 24 prompts featuring unique permutations of the cardinal directions. We analyzed the model’s predictive probabilities, calculating the total probability for

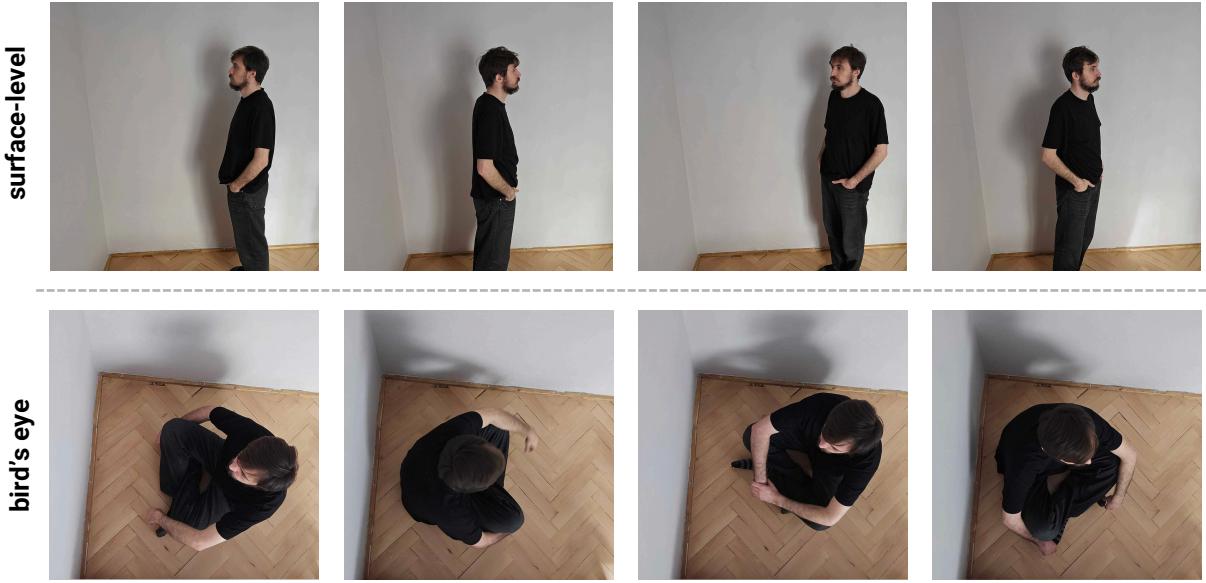


Figure 9: Surface-level (top row) and bird’s-eye (bottom row) views of a real person used in the influence of subject type experiment.

each direction by summing probabilities of likely output tokens containing direction words (e.g., *east*, *eastward*), averaged across examples for each permutation condition.

The resulting distributions, shown in Figure 10, show east having the highest probability in 18 of 24 permutations, confirming a robust underlying bias. However, a clear pattern emerged in the remaining 6 cases: permutations starting with the east (e.g., ...*facing east, north, west, or south?*) resulted in high, similar probabilities assigned to both south and east. This finding indicates that the strong default east preference interacts with prompt structure; specifically, listing east first also significantly boosts the probability of south, revealing sensitivity beyond the inherent bias.

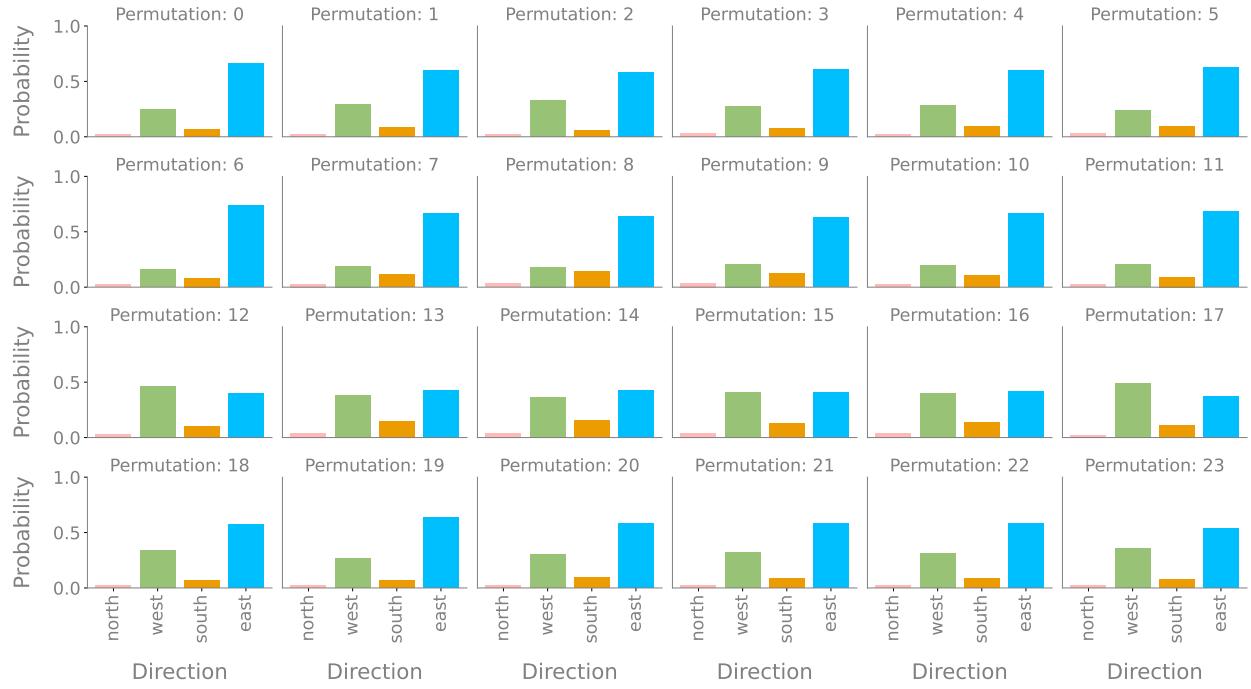


Figure 10: Probability distributions for cardinal directions averaged over 69 images for each of 24 prompt permutations ordering the options. *East* dominates in 18/24 cases, confirming a strong bias. However, permutations starting with *east* show high, similar probabilities for both *east* and *south*, revealing sensitivity to prompt structure alongside the baseline preference.

Appendix C Orientation vs Perspective Test

We started by looking at how models handle questions about what a humanoid minifigure in an image might be seeing (Q6). It seemed logical that to answer this correctly, a model might first need to figure out which direction the humanoid minifigure is facing (Q5). For example, if an object is to the north, and the humanoid minifigure is facing north, then the humanoid minifigure likely sees the object. This led us to wonder: *are the difficulties models have with visual perspective taking (Q6) mainly caused by problems identifying the humanoid minifigure’s direction (Q5)?*

To investigate this, we ran an experiment across 144 visual tasks. We specifically tested the VLMs on Q6, but with a helpful modification to the prompt: we explicitly told the model which cardinal direction the humanoid minifigure was facing, using the gold-standard answer from Q5 as a hint. Our reasoning was that if the struggle with Q5 was the primary bottleneck for Q6, providing this directional information should lead to a significant performance increase. However, as shown in Figure 11, adding this hint resulted in only marginal improvements. This suggests that the models’ difficulties with visual perspective taking might be more complex than simply getting the orientation wrong, and solving Q5 alone does not automatically lead to correctly answering Q6.

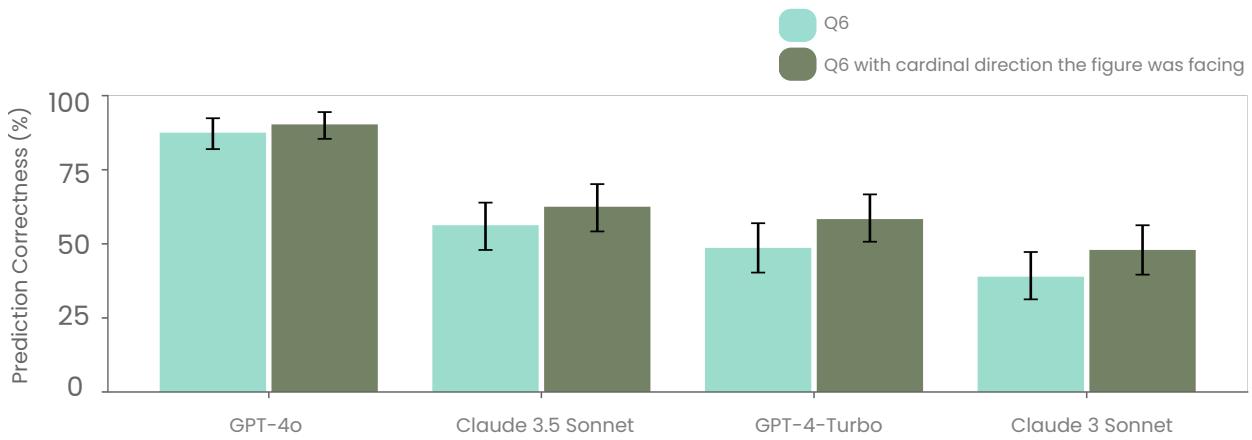


Figure 11: Comparison of VLM performance on visual perspective taking (Q6) with and without an explicit orientation hint (gold-standard Q5 answer), showing only marginal prediction correctness improvement.