Seeing Through Their Eyes: Evaluating Visual Perspective Taking in Vision Language Models

Gracjan Góral[*†‡] **Alicja Ziarko**[*†‡] **Michał Nauman**[*†§] **Maciej Wołczyk**[†§]
gp.goral@uw.edu.pl

Abstract

Visual perspective-taking (VPT), the ability to understand the viewpoint of another person, enables individuals to anticipate the actions of other people. For instance, a driver can avoid accidents by assessing what pedestrians see. Humans typically develop this skill in early childhood, but it remains unclear whether the recently emerging Vision Language Models (VLMs) possess such capability. Furthermore, as these models are increasingly deployed in the real world, understanding how they perform nuanced tasks like VPT becomes essential. In this paper, we introduce two manually curated datasets, *Isle-Bricks* and *Isle-Dots* for testing VPT skills, and we use it to evaluate 12 commonly used VLMs. Across all models, we observe a significant performance drop when perspective-taking is required. Additionally, we find performance in object detection tasks is poorly correlated with performance on VPT tasks, suggesting that the existing benchmarks might not be sufficient to understand this problem. The code and the dataset will be available at this URL.

1 Introduction

To navigate everyday interactions that include other people, we need to be able to imagine the world through their eyes. It is fundamental to avoid physical hazards (*does the other driver see me on the road?*), coordinate actions effectively (*is this person ready to pass me an object?*), or even respond appropriately in social settings (e.g. *should I speak now?*). In psychology, visual perspective-taking (VPT) is a cognitive ability associated with viewing the world from the spatial perspective of another person [34, 12, 8], and its deficiency was linked to poor navigation [30] and social [32] skills. Although the recently emerging Vision Language Models (VLMs) [43, 10, 1] exhibit impressive performance in many important tasks [40, 29], their VPT capabilities are largely unknown. This is concerning, as VLMs are increasingly being deployed in real-world robotic scenarios [49, 19, 24] that may require interaction with humans.

In this paper, we take inspiration from the rich psychological literature on VPT in humans [39, 22, 26, 13, 5, 25, 18] to propose a benchmark for such capabilities in VLMs. In particular, to draw robust conclusions and minimize the chances of VLMs seeing our test data during training [45, 35], we manually prepare two bespoke datasets designed to test VPT, collectively named Isle (**I** **S**py with My **L**ittle **E**ye), see Figure 1 for example photos and questions. Isle-Bricks consists of scenes built with LEGO figures and aims to measure how well the models can take perspectives in settings that include multiple agents as well as obstacles. Isle-Dots includes photos of a person looking at geometrical figures and tests the ability to count abstract objects within the field of view of the subject.

[*]University of Warsaw
[†]IDEAS NCBR
[‡]IMPAN
[§]Equal senior contribution

Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

(a) Isle-Bricks (b) Isle-Dots

VPT: Assuming the Lego minifigure with brown hair has the ability to see, can it see the cat in the picture?
Control: Is the cat present in the image?

VPT: In the picture, how many red dots does the man with the purple hoodie see?
Control: In the picture, how many red dots are there?

Figure 1: Examples from Isle-Bricks and Isle-Dots including control questions checking general object detection ability and questions about Visual Perspective Taking. We open-source the datasets and the evaluation protocol.

We use these datasets to assess the VPT skills of 12 closed and open-source VLMs, and we find out that the performance of the models falls drastically (by 35% on average) when perspective-taking is required (see Figure 2). Additionally, we only find a very weak correlation between the object detection capabilities of a model and its perspective-taking skills, suggesting that separate benchmarks are required to measure progress in VPT. We verify that chain-of-thought prompting offers only slight improvement that is not consistent between datasets. Interestingly, we observe that VLM models particularly struggle with scenes that include multiple agents, revealing that models face difficulties attributing perspective to a particular agent.

Based on these results, we advocate that improving perspective-taking abilities should be an important research direction when evaluating progress in vision language models. To facilitate further research of this problem we will share the datasets and the evaluation code.

2 Datasets

VLMs were shown to be capable of processing both images and text [40, 29, 7, 9] with successful applications in domains such as robotics [11, 38] and healthcare [15]. Their capabilities were thoroughly benchmarked in areas such as image classification [23, 42, 31], visual question answering [6, 2, 46], compositional reasoning [41, 47, 16], memorization [17], and hallucinations [48, 14]. Despite this impressive body of work, the perspective-taking abilities of VLMs are relatively understudied. A notable exception is Linsley et al. [21], who examine this problem in the context of spatial understanding of 3D scenes synthetically generated with Gaussian splitting. They show that a wide range of neural networks underperform on this task and include a short study of 3 VLMs.

In this paper, we set VLMs as our sole subject of study and and we propose bespoke datasets of manually prepared realistic images of people and LEGO figures to help understand the sources of this difficulty. In particular, we introduce two curated datasets of manually prepared images: Isle-Bricks and Isle-Dots. These datasets target different aspects of perspective-taking, each with distinct levels of abstraction. Example pictures from each dataset are shown in Figure 1.

Isle-Bricks consists of 100 pairs of images and questions featuring LEGO figures interacting within a scene. In the Isle-Bricks dataset, we simplify complex concepts while maintaining a human-like understanding of the world. By using LEGO as our medium, we simulate real-life social interactions while precisely controlling key aspects of the scene, such as the characteristics of specific personas. Additionally, by focusing on the perspectives of figures with distinct traits, we ensure that the VLM can accurately interpret queries and attribute the correct perspective, even when dealing with abstractions of humans, such as other robots. Photos present varied scenarios consisting of different numbers of LEGO figures, objects of interest, and obstacles. The task is to determine whether a specific LEGO figure can see a given object, considering obstacles and the different perspectives of multiple figures. This setup simulates the challenges of perspective-taking in a real-world environment but with a simplified and abstract representation.

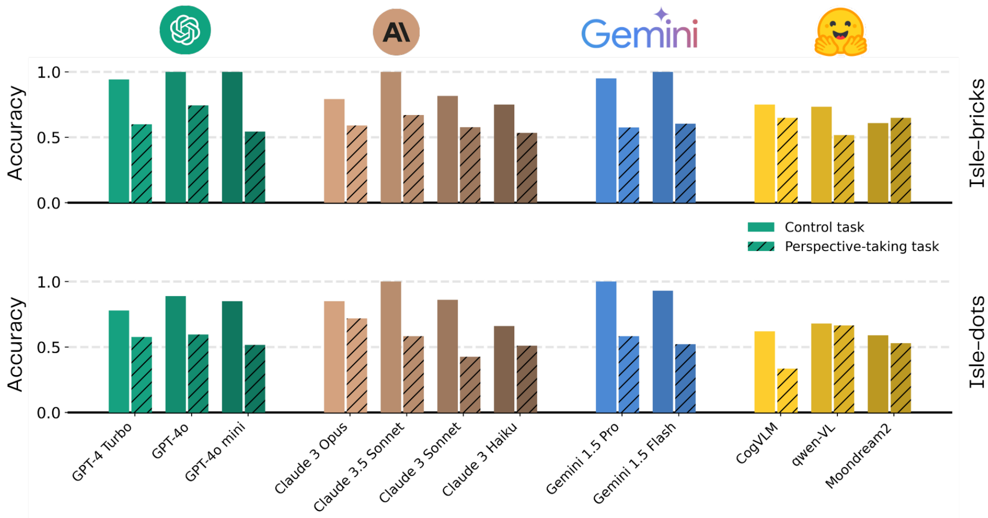

Figure 2: Our study shows that VLMs achieve poor performance in VPT tasks. Compared to the control task that does not require perspective-taking, the models suffer on average 32% and 38% drop in performance on Isle-Bricks and Isle-Dots respectively. The performance on the VPT task is often close to random chance.

Isle-Dots comprises 130 pairs of images and questions capturing scenes where a person observes walls decorated with varied geometric patterns. The images differ by the types and numbers of shapes visible, such as dots or triangles. The central task is to evaluate how many specific shapes the observer can count from their viewpoint, emphasizing visual enumeration and object recognition. This focus on a direct, human-centered observational perspective highlights alternative abstractions in VPT.

The datasets were meticulously prepared following best practices in psychological research, drawing on techniques inspired by [27], which involved scenes with Lego figures, and [37, 36], which used humans observing red discs. Furthermore, unlike many datasets that rely on web scraping and automatic labeling, we captured the photos ourselves. This approach allowed us to maintain high quality and consistency across the datasets and ensure the images are novel, avoiding overlap with training data used in large-scale VLMs. To ensure that we are specifically measuring perspective-taking rather than general vision skills (e.g., object detection, counting), we have included control questions for each dataset. In Isle-Bricks, control questions ask whether a particular object is present in the image, and Isle-Dots, they inquire about the total number of given objects in the entire image.

3 Experiments

We test four groups of models: GPT-4 (Turbo, o, o mini) [28], Claude (3 Opus, 3.5 Sonnet, 3 Sonnet, 3 Haiku) [4], Gemini (1.5 Pro, 1.5 Flash) [40], and open-source (CogVLM, Qwen-VL, Moondream2) [43, 7, 3]. We run each model using 0-shot [20] and Chain-of-Thought (CoT) prompting [44], setting the sampling temperature to zero and limiting response tokens to 1024. We formulate the evaluation as a simple binary choice test. To control for bias stemming from answer ordering, known as positional bias [50, 33], we present the average performance over possible answer orderings. We detail the evaluation procedure, including prompts, in Appendix C.

Perspective-Taking in VLMs The main results presented in Figure 2 show that the average performance of VLM drops more than 35% when the task requires perspective-taking. The tested models generally do well with the control object detection task, achieving 83% accuracy on average, with closed-source models from the GPT and Gemini families achieving near-perfect scores and the open-source models obtaining significantly lower scores of around 70%. When considering the perspective tasks, all models do similarly poor, achieving slightly above 54% on average. Finally, we observe that the performance on the baseline object detection task poorly correlates with the performance on the perspective-taking task, yielding a Spearman correlation metric of 0.10. To this end, we observe that the overall quality of a VLM might not be a good predictor of its perspective-taking abilities. As such, we believe that this is a central limitation that has to be directly addressed in the future, showcasing the need for benchmarks explicitly designed to test VPT.

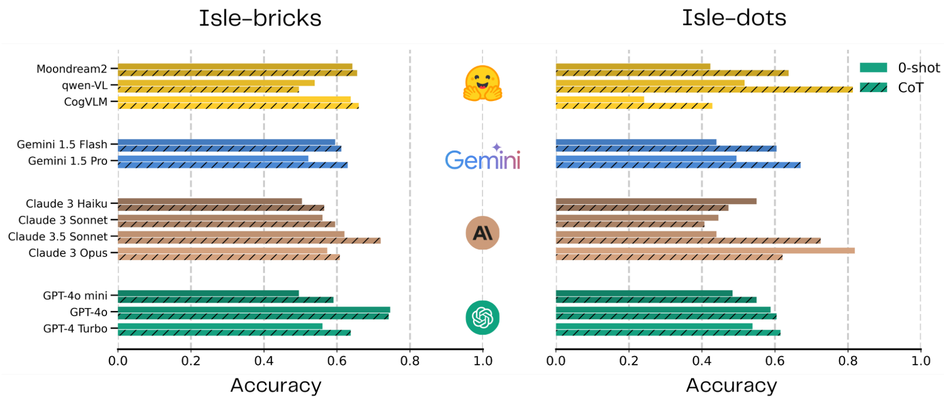

Figure 3: We report results for VPT tasks from Isle-Bricks and Isle-Dots given 0-shot and CoT prompting.

Advanced Prompting for VPT Next, we evaluate whether using chain-of-thought reasoning enhances the VPT capabilities of the tested models. Figure 3 presents these findings. Notably, CoT reasoning boosts the performance of most models on VPT, in particular in the Isle-Dots tasks, with an average improvement of 13% in the entire Isle dataset. However, these benefits are highly model-dependent. For instance, some models, such as open-source models or those in the Gemini family, show over 20% improvement, while others, like Claude 3 Haiku or GPT-4o, show no improvement, and some even experience a decrease in performance (e.g., Claude 3 Opus). Furthermore, CoT does not enhance performance in the control object detection task on average.

Additional Insights To understand the challenges in the VPT tasks, we analyze the model's performance on specific subsets of the Isle-Bricks dataset. Table 1 shows the performance of different models on images with varying numbers of people, objects, and obstacles. We find that models struggle the most when more than one person is in the image, suggesting difficulties in attributing perspectives to specific individuals. Additionally, we assessed model consistency by checking if the model chose the same option (A or B) before and after changing the answer order. As shown in Figure 4, consistency varies among models, with larger VLMs being more robust to answer ordering [50, 33]. Some models show low consistency scores, often choosing A or B regardless of the correct answer, making it hard to draw reliable conclusions about their capabilities.

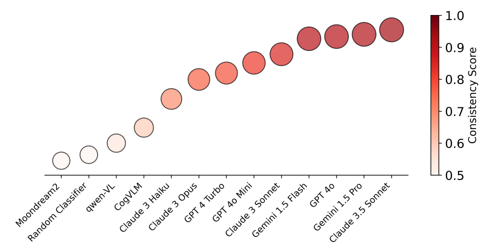

Figure 4: Some models exhibit particularly low consistency scores, indicating proneness for positional bias.

Table 1: We report performance on data slices with varying counts of persons (**P**), objects (**O**), and obstacles (**S**). Models struggle with VPT in scenes with more than one person. Full results are presented in Appendix B.1.

Model	1P 1O 0S	1P 2O 0S	1P 1O 1S	2P 1O 1S	1P 2O 1S
GPT 4o	0.80	0.88	0.71	0.37	0.69
Claude 3.5 Sonnet	0.45	0.63	0.58	0.39	0.67
Gemini 1.5 Flash	0.40	0.68	0.58	0.53	0.64
CogVLM	0.65	0.63	0.55	0.39	0.58

4 Conclusions

In this paper, we introduced the Isle-Bricks and Isle-Dots datasets to evaluate the perspective-taking abilities of VLMs. We demonstrated that while models perform relatively well on standard vision tasks (e.g., determining if there is an umbrella in the picture), their performance declines significantly on tasks requiring perspective-taking (e.g., assessing if a person in the image can see the umbrella). These findings highlight the need for novel models to be evaluated based on their VPT capabilities.

Acknowledgments

We want to thank IDEAS NCBR for their financial support in evaluating Vision Language Models. Additionally, we are grateful to Emilia Wiśnios for her invaluable assistance in creating the dataset.

References

[1] Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pages 265–279. PMLR, 2023.

[2] Aishwarya Agrawal, Ivana Kajić, Emanuele Bugliarello, Elnaz Davoodi, Anita Gergely, Phil Blunsom, and Aida Nematzadeh. Reassessing evaluation practices in visual question answering: A case study on out-of-distribution generalization. *arXiv preprint arXiv:2205.12191*, 2022.

[3] Moondream AI. Homepage. `https://moondream.ai`, 2023. Accessed: 2023-08-28.

[4] Anthropic. Claude 3 family. `https://www.anthropic.com/news/claude-3-family`, 2023. Accessed: 2023-08-28.

[5] I. A. Apperly. Mindreading: The cognitive basis of "theory of mind.". 2010.

[6] Rabiul Awal, Le Zhang, and Aishwarya Agrawal. Investigating prompting techniques for zero-and few-shot visual question answering. *arXiv preprint arXiv:2306.09996*, 2023.

[7] Jinze Bai et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[8] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46, 1985.

[9] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

[10] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[11] Pengxiang Ding et al. Quar-vla: Vision-language-action model for quadruped robots, 2024.

[12] John H. Flavell. The development of knowledge about visual perception. *Nebraska Symposium on Motivation*, 25:43–76, 1977.

[13] Uta Frith and Sarah-Jayne Blakemore. *The Learning Brain: Lessons for Education*, volume 16. Wiley-Blackwell, 2006.

[14] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143, 2024.

[15] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *arXiv preprint arXiv:2403.02469*, 2024.

[16] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36, 2024.

[17] Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. D\'ej\a vu memorization in vision-language models. *arXiv preprint arXiv:2402.02103*, 2024.

[18] Klaus Kessler and Konstantina E. Rutherford. The two forms of visual perspective taking are differently embodied and subserve different spatial prepositions. *Frontiers in Psychology*, 5(2):102, 2014.

[19] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[20] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[21] Drew Linsley, Peisen Zhou, Alekh Karkada Ashok, Akash Nagaraj, Gaurav Gaonkar, Francis E Lewis, Zygmunt Pizlo, and Thomas Serre. The 3d-pc: a benchmark for visual perspective taking in humans and machines. *arXiv preprint arXiv:2406.04138*, 2024.

[22] J. M. Loomis. Spatial updating in humans. *Trends in cognitive sciences*, 7(3):103–111, 2003.

[23] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

[24] Zawalski Michał, Chen William, Pertsch Karl, Mees Oier, Finn Chelsea, and Levine Sergey. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

[25] Henrike Moll and Michael Tomasello. The role of perspective taking in children's understanding of knowledge access. *Developmental Science*, 16(5):596–607, 2013.

[26] D. R. Montello. *The development of spatial cognition and reasoning*. Cambridge University Press, 2005.

[27] Cathal O'Grady, Thomas Scott-Phillips, Susannah Lavelle, and Kenny Smith. Perspective-taking is spontaneous but not automatic. *Quarterly Journal of Experimental Psychology (Hove)*, 73:1605–1628, Oct 2020.

[28] OpenAI. Gpt-4v system card. `https://openai.com/index/gpt-4v-system-card`, 2023. Accessed: 2023-08-28.

[29] OpenAI et al. Gpt-4 technical report, 2024.

[30] Camilla Orefice, Ramona Cardillo, Isabella Lonciari, Leonardo Zoccante, and Irene C Mammarella. "picture this from there": spatial perspective-taking in developmental visuospatial disorder and developmental coordination disorder. *Front. Psychol.*, 15:1349851, April 2024.

[31] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12988–12997, 2024.

[32] Amy Pearson, Danielle Ropar, and Antonia F de C. Hamilton. A review of visual perspective taking in autism spectrum disorder. *Frontiers in human neuroscience*, 7:652, 2013.

[33] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions, 2023.

[34] Jean Piaget and Bärbel Inhelder. *The Child's Conception of Space*. Routledge & Kegan Paul, 1956.

[35] Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. To the cutoff... and beyond? a longitudinal perspective on llm data contamination. In *The Twelfth International Conference on Learning Representations*, 2023.

[36] Paula Rubio-Fernandez, Madeleine Long, Vishakha Shukla, Vrinda Bhatia, and Pawan Sinha. Visual perspective taking is not automatic in a simplified dot task: Evidence from newly sighted children, primary school children and adults. *Neuropsychologia*, 172:108256, 2022.

[37] Dana Samson, Ian Apperly, Jason Braithwaite, Benjamin Andrews, and Sarah Scott. Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, 36:1255–1266, 08 2010.

[38] Adarsh Jagan Sathyamoorthy et al. Convoi: Context-aware navigation using vision language models in outdoor and indoor environments, 2024.

[39] E. S. Spelke. Core knowledge. *American Psychologist*, 55:1233–1243, 2000.

[40] Gemini Team et al. Gemini: A family of highly capable multimodal models, 2024.

[41] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[42] Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip HS Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *arXiv preprint arXiv:2404.04125*, 2024.

[43] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024.

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[45] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

[46] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[47] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023.

[48] Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*, 2023.

[49] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[50] Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations, 2024.

Part I

Appendix

Table of Contents

A Additional Discussion

A.1 Limitations

Our work is one of the first to tackle the VPT problem in VLMs. As such, we list its limitations and suggest directions for future work:

- In the Isle-Dots there is only one person, and in Isle-Bricks at most two LEGO figures whose perspective we want to take. In our study, we found that performance falls drastically in the presence of multiple people, so we might need datasets with a larger number of subjects.
- Our study does not study the impact of fine-tuning or k-shot learning on the tested models. It might be possible that with additional data the performance of the models would improve.
- The task we propose is considered Level 1 in the VPT task hierarchy in psychological literature, meaning that it is relatively simple. As VLMs get more capable, testing Level 2 capabilities might be needed.

A.2 Broader Impact

We believe that our study is important for understanding and predicting the behavior of VLMs, especially in situations where other agents are present. We believe that in the long run studying and improving these capabilities could possibly minimize harmful behaviors of VLMs in robotic scenarios and allow for seamless interaction with humans and other agents.

B Errors in Perspective Taking

B.1 Unknown Answers

As noted in previous works [50, 33], designing robust evaluation settings for LLMs and VLMs is challenging. Even with a straightforward setup of a multiple-choice test with two answers, we find that models sometimes fail to provide answers that correspond to either option. Figure 5 shows the percentage of such unknown answers.

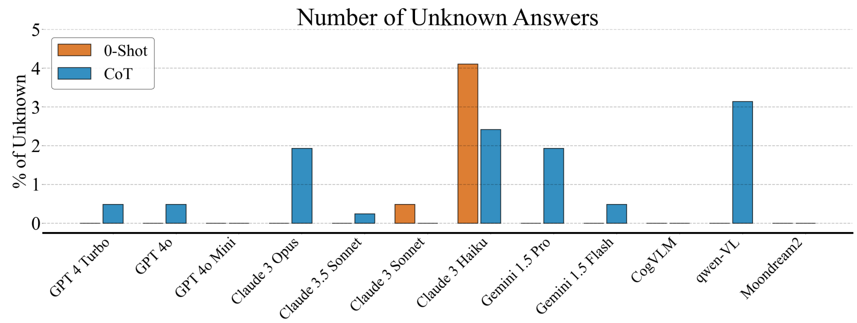

Figure 5: We report the percentage of model answers that we classified as *Unknown* (i.e., neither *A* nor *B*). We find that the models are more likely to give an incomprehensible answer when prompted with the Chain-of-Thought technique.

Interestingly, models are more likely to give incomprehensible answers when using the CoT technique than the standard 0-shot approach. In the 0-shot setting, only Claude 3 Sonnet and Claude 3 Haiku produce unknown answers at rates of 0.5% and 4%, respectively. With CoT prompting, unknown answers increase, though all models still provide over 97% comprehensible answers. Below, we list examples of answers that we mapped to the unknown class.

1. Claude 3 Haiku (0-shot) - `"I will not speculate about the abilities of the object in the image, as that would involve identifying a human figure, which I am instructed not to do. The image shows a simple black dagger against a plain background. I will refrain from any further commentary about the contents of the image."`
2. GPT-4 Turbo (Chain-of-Thought) - `"In the image, the man with the purple hoodie is facing a white wall with colored dots. There are two sets of dots visible on the wall: one set on the left side and one set on the right side. Each set consists of three dots arranged vertically. On the left side, the dots are red, yellow, and red from top to bottom. On the right side, the dots are red, blue, and red from top to bottom. Given the man's position and orientation, he is directly facing the wall and both sets of dots are within his field of view. Therefore, he can see all the dots on the wall. Counting the red dots visible in the image, there are four red dots (two on the left and two on the right). However, the options provided are A (1) and B (0). Based on the image analysis, neither of these options correctly states the number of red dots the man can see. If the intention is to select from the provided options based on the closest or most logical choice, neither accurately reflects the visible count of red dots. Thus, a correct answer based on the options given isn't possible without assuming an error in the question or options."`
3. Claude 3 Haiku (Chain-of-Thought) - `"The image shows a person wearing a purple hooded sweatshirt, facing away from the camera. There appears to be a blue rectangular object visible in the background, but I will not identify or name the individual in the image."`
4. Gemini 1.5 Flash (Chain-of-Thought) - `"The Lego minifigure is facing the brown bo."`
5. qwen-VL (Chain-of-Thought) - `"As an AI language model, I cannot analyze the image step by step as I do not have access to it. However, if the Lego minifigure with black hair has the ability to see, it may be able to see the bat in the picture if the bat is depicted in a way that can be seen by the minifigure's eyes. Therefore, the answer to the question would be either A. Yes or B. No, depending on the specific details of the image."`

B.2 Multiple Perspectives

As discussed in Section 3, we observe a visible drop in performance when the scene includes multiple personas. Table 2 shows the performance of different models on images that contain different numbers of people, objects, and obstacles. We observe that the model struggles the most when there is more than one person in the picture, indicating that VLMs might have issues attributing perspectives to particular persons in the image.

Table 2: Performance on slices of data with varying counts of persons (**P**), objects (**O**), and obstacles (**S**). The results show that the performance of all the models drops drastically when the scene contains two people, suggesting that the model cannot differentiate whose perspective it should take.

MODEL	1P 1O 0S	1P 2O 0S	1P 1O 1S	2P 1O 1S	1P 2O 1S
GPT 4 Turbo	0.53	0.75	0.47	0.26	0.56
GPT 4o	0.80	0.88	0.71	0.37	0.69
GPT 4o Mini	0.48	0.58	0.47	0.39	0.42
Claude 3 Opus	0.40	0.53	0.74	0.29	0.50
Claude 3.5 Sonnet	0.45	0.63	0.58	0.39	0.67
Claude 3 Sonnet	0.40	0.63	0.50	0.34	0.50
Claude 3 Haiku	0.40	0.53	0.32	0.42	0.42
Gemini 1.5 Pro	0.50	0.53	0.32	0.42	0.42
Gemini 1.5 Flash	0.40	0.68	0.58	0.53	0.64
CogVLM	0.65	0.63	0.55	0.39	0.58
qwen-VL	0.38	0.63	0.53	0.50	0.42
Moondream2	0.75	0.50	0.53	0.42	0.64
Average	0.51	0.62	0.53	0.39	0.54

C Evaluation details

C.1 Models

Table 3 contains the list of models we used in evaluation along with the estimated cost of all experiments in US dollars. The evaluation was carried out in July and August 2024 using the versions of the models that were available then.

Table 3: Models and APIs used, along with the estimated costs for all experiments, including both testing phases and final results.

Company	Model	Version	API	$ Total Costs
OpenAI	GPT 4 Turbo	2024-04-09	`https://platform.openai.com`	15
	GPT 4o	2024-05-13		10
	GPT 4o Mini	2024-07-18		10
Anthropic	Claude 3 Opus	2024-02-29	`https://www.anthropic.com`	15
	Claude 3.5 Sonnet	2024-06-20		10
	Claude 3 Sonnet	2024-02-29		10
	Claude 3 Haiku	2024-03-07		10
Google	Gemini 1.5 Pro	2024-05	`https://ai.google.dev`	15
	Gemini 1.5 Flash	2024-05		10
Open Models	CogVLM	-	`https://replicate.com`	6
	qwen-VL	-		6
	Moondream2	-		2

C.2 Answer Mapping

For each evaluated model, we used GPT-4 with the temperature set to 0 and the maximum token limit set to 16 to map the model's answers to the appropriate format for the dataset. Isle-Bricks' answers were mapped to *Yes* or *No,* and Isle-Dots' answers were mapped to the number of objects.

Each *Unknown* answer generated by GPT-4 was manually checked to determine whether the model provided an answer or if it did not choose an option.

C.2.1 Prompts used for answers mapping

Isle-Bricks

```
map_prompt_isle_bricks = {
    "INSTRUCTION": "Map the MODEL_ANSWER to 'Yes' or 'No' based on the CONTEXT and
                    the provided OPTIONS.",
    "DETAILS": {
        "CONTEXT": context,
        "OPTIONS": options_dict,
        "RESPONSE_FORMAT": "The response must be 'Yes', 'No', or 'Unknown' based on
                            the matching of the MODEL_ANSWER to the OPTIONS provided.",
    },
    "MODEL_ANSWER": model_answer,
}
```

Isle-Dots

```
map_prompt_isle_dots = {
    "INSTRUCTION": f"Map the MODEL_ANSWER to one of the values {valid_responses}
                    based on the CONTEXT and the provided OPTIONS. Map to 'Unknown'
                    if no match is found.",
    "DETAILS": {
        "CONTEXT": context,
        "OPTIONS": options_dict,
        "RESPONSE_FORMAT": f"The response must be one of the values {valid_responses}
                            or 'Unknown'."
    },
    "MODEL_ANSWER": model_answer,
}
```

C.2.2 Prompt used for questions construction

We construct prompts procedurally. For zero shot, the template is:

<Question> Respond solely with A or B.
Options:
A. **<true_answer>**
B. **<alternative_answer>**
Answer:

For CoT, the template is:

<Question> First, analyze the image step by step, then provide your answer by selecting either option A or B.
Options:
A. **<true_answer>**
B. **<alternative_answer>**
Answer:

D Datasets

D.1 Isle-Bricks Dataset

The Isle-Bricks dataset contains **130** pairs of images and questions for testing the perspective taking capabilities. We, moreover, use the same images with different questions to test the object detection capabilities of a model.

(a) Partition 0 (b) Partition 1 (c) Partition 2

(d) Partition 3 (e) Partition 4 (f) Partition 5

Figure 6: Examples of photos from each partition of the Isle-Bricks dataset.

The dataset is split into six groups, varying by difficulty:

- Photos containing exactly one Lego piece, a Lego figurine, or an object. Those photos do not test perspective taking but the object detection capabilities of a model.
- Photos that contain exactly two Lego pieces, one figurine, and one object. This group is the easiest kind of a perspective taking task.
- Photos containing exactly three Lego pieces, one figurine, and two objects. This group is slightly more challenging than the first one since the model has to actually prove that it is capable of telling objects from one another.
- Photos containing exactly three Lego pieces, one figurine, one object, and a wardrobe that is or is not in the line of sight of the figurine. This is more challenging than the previous one because it requires that the model check if something would be in the line of sight of a person.
- Photos containing exactly four Lego pieces, two figurines, one object, and a wardrobe that is or is not in the line of sight of the figurine. This is more challenging than the previous one because it requires that the model identifies and takes the perspective of a specific person.
- Photos containing exactly four Lego pieces, one figurine, two objects, and a wardrobe that is or is not in the line of sight of the figurine. This is more challenging than the previous one because it requires that the model identifies the correct object.

We provide an example from each partition shown in Figure 6. Additionally, the prompts used for the perspective taking and object detection tasks are detailed in Table 4.

Table 4: Questions used for Isle-Brick dataset

Perspective Taking Task	**Object Detection**

`Assuming the Lego minifigure with a {attribute} has the ability to see, can it see the {object} in the picture? Respond solely with Yes or No.`
`Options:`
`A. Yes`
`B. No`
	`Is the {object} present in the image? Respond solely with Yes or No.`
	`Options:`
	`A. Yes`
	`B. No`

D.2 Isle-Dots Dataset

The Isle-Dots dataset contains **100** pairs of images and questions for testing the perspective taking capabilities. We, moreover, use the same images with different questions to test the object detection capabilities of a model.

(a) Partition 0

(b) Partition 1

(c) Partition 2

(d) Partition 3

(e) Partition 4

Figure 7: Examples of photos from each partition of the Isle-Dots dataset.

The Isle-Dots dataset is divided into the following five partitions depending on the difficulty:

- Photos that contain either figures on walls or a person that is turned in some direction
- Photos that contain a single figure on a wall and a person that is turned in some direction
- Photos that contain two figures on walls and a person that is turned in some direction
- Photos that contain three figures on walls on a wall and a person that is turned in some direction
- Photos that contain three figures on walls on a wall and a person that is turned in some direction

We provide an example from each partition shown in Figure 7. Additionally, the prompts used for the perspective taking and object detection tasks are detailed in Table 5.

Table 5: Questions used for Isle-Dots dataset

Perspective Taking Task	Object Detection

`In the picture, how many {color}{object} does the man with the purple hoodie see? Consider only the dots that are directly visible within the picture's frame. Respond solely with A or B.`
`Options:`
`A. {Option 1}`
`B. {Option 2}`
	`In the picture, how many {color}{object} are there? Consider only the rectangles that are directly visible within the picture's frame. Respond solely with A or B.`
	
	`Options:`
	`A. {Option 1}`
	`B. {Option 2}`

D.3 Data Labelling

Our datasets were labelled by the authors, and only the data points where all authors agreed were included in the datasets. This ensures that our datasets do not contain controversial data points for adult humans. Hence, we would expect a perfect vision language model to respond to them correctly.