Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference under Ambiguities

Zheyuan Zhang[1*] **Fengyuan Hu**[1*] **Jayjun Lee**[1*]
Freda Shi[2,3] **Parisa Kordjamshidi**[4] **Joyce Chai**[1] **Ziqiao Ma**[1]
[1]University of Michigan [2]University of Waterloo
[3]Vector Institute, Canada CIFAR AI Chair [4]Michigan State University
https://spatial-comfort.github.io/

Abstract

Spatial expressions in situated communication can be ambiguous, as their meanings vary depending on the frames of reference (FoR) adopted by speakers and listeners. While spatial language understanding and reasoning by vision-language models (VLMs) have gained increasing attention, potential ambiguities in these models are still under-explored. To address this issue, we present the COnsistent Multilingual Frame Of Reference Test (COMFORT), an evaluation protocol to systematically assess the spatial reasoning capabilities of VLMs. We evaluate nine state-of-the-art VLMs using COMFORT. Despite showing some alignment with English conventions in resolving ambiguities, our experiments reveal significant shortcomings of VLMs: notably, the models (1) exhibit poor robustness and consistency, (2) lack the flexibility to accommodate multiple FoRs, and (3) fail to adhere to language-specific or culture-specific conventions in cross-lingual tests, as English tends to dominate other languages. With a growing effort to align vision-language models with human cognitive intuitions, we call for more attention to the ambiguous nature and cross-cultural diversity of spatial reasoning.

1 Introduction

Even a simple spatial expression like "the basketball to the right of the car" may have multiple interpretations. People may use different *frames of reference* [FoR; 34, 22, *inter alia*] to resolve ambiguity about the underlying coordinate system, as illustrated in Figure 1a. The diversity of conventions across languages and cultures further complicates this ambiguity—different languages employ different conventions in choosing one FoR among multiple competing options. As shown in Figure 1b, speakers may project themselves onto the ball or consider an imaginary listener facing them [63]. These ambiguities are not easily resolvable based solely on linguistic expressions [64, 37]. We refer to Appendix A.1 for more background and related works.

Our main research question is not new: *Do vision-language models represent space, and how?* Several benchmarks [31, 38] have been developed for this purpose, consisting of text-image pairs where objects may or may not follow certain spatial relations. However, the aforementioned spatial ambiguities remain largely under-explored when studying VLM-based spatial language understanding and reasoning. We emphasize that FoRs are crucial to studying spatial cognition across modalities, as they provide a foundational framework for understanding how spatial relationships are perceived, interpreted, and communicated [35].

[*]Authors contributed equally to this work.

Pluralistic Alignment Workshop @ 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

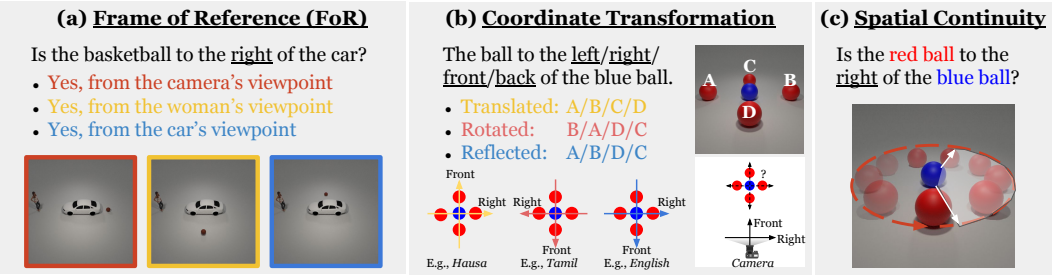


Figure 1: In situated communication, spatial language understanding and reasoning are often ambiguous, leading to varying interpretations among people from different cultural backgrounds. Specifically: (a) different frames of reference can result in different interpretations of the same spatial expression; (b) speakers of different languages may use distinct coordinate frames for non-fronted reference objects; and (c) spatial relations extend beyond exact axes to include acceptable regions.

To fill this gap, we present COnsistent Multilingual Frame Of Reference Test (COMFORT), a framework that systematically evaluates the spatial reasoning capabilities of VLMs, emphasizing consistency in understanding ambiguous and disambiguated spatial expressions. COMFORTintroduces (1) a set of spatial reasoning tasks instantiated by synthetic 3D images and corresponding text describing spatial relations and (2) metrics to evaluate the robustness and consistency of the model responses. We extend the setup to multilingual settings by evaluating models in 109 languages across 170 regions worldwide. We find that VLMs show alignment with English conventions in spatial language understanding when resolving ambiguities. However, they (1) are still far from achieving robustness and consistency, (2) lack the flexibility to accommodate multiple FoRs, and (3) fail to adhere to linguistic and cultural conventions in cross-lingual tests, as English tends to dominate other languages. With a growing effort to align vision-language models with human cognition, we highlight the ambiguous nature of spatial language and call for increased attention to cross-cultural diversity in spatial reasoning.

2 Consistent Multilingual Frame of Reference Test (COMFORT)

We introduce the COnsistent Multilingual Frame Of Reference Test (COMFORT), a new evaluation protocol with dataset, tasks, and comprehensive metrics, to study VLM behaviors in spatial language reasoning with FoR-related ambiguity. This protocol accommodates spatial continuity and various ambiguities, drawing insights from several well-defined metrics to assess performance and prediction consistency. Given our primary focus on the analytical inquiry of models' linguistic competence (i.e., spatial knowledge encoded in the latent representations) rather than performance (i.e., behavioral evaluation) only [68, 60],[2] we additionally develop better evaluation and consistency metrics to deepen our understanding of model capabilities.

- **COMFORT-CAR**: When the relatum is fronted, as examples in Figure 1a, multiple FoRs are possible to interpret the reference system.
- **COMFORT-BALL**: When the relatum is non-fronted, as examples in Figure 1b, we focus on the ambiguity of conventions to determine its coordinate transformation for egocentric relative FoR.

A language prompt (Table 4) queries whether a spatial relation $r \in \mathcal{R}$ is satisfied by a referent-relatum pair in the image under FoR $f \in \mathcal{F}$ (Figure 3) in language $\ell \in \mathcal{L}$. This work also examines models using queries with no FoR specified; therefore, a test case in COMFORTis defined as a 4-tuple in $\mathcal{S} \times \mathcal{R} \times (\mathcal{F} \cup \{\emptyset\}) \times \mathcal{L}$. While there are many spatial relations in daily languages, we primarily focus on four canonical directions; that is, the considered relation set $\mathcal{R} = \{\textit{to the left of}, \textit{to the right of}, \textit{in front of}, \textit{behind}\}$. COMFORTcovers $|\mathcal{L}| = 109$ languages worldwide; however, we use English as an example to describe the data synthesis and evaluation processes for simplicity and clarity, and refer readers to Appendix A.2 for more details.

3 Empirical Experiments and Main Findings

In principle, COMFORT can be applied to all VLMs, whether multilingual or monolingual. We note that most existing open-source VLMs are English-based language models; therefore, we begin our

[2]Here, we use the terms *competence* and *performance* analogically to Chomsky [15].

Model	Egocentric		Intrinsic		Addressee		Aggregated	
	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$(↓)	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$(↓)	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$(↓)	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$(↓)
InstructBLIP-7B	$47.2_{(+0.0)}$	$43.5_{(+1.7)}$	$47.2_{(+0.0)}$	$42.3_{(+0.0)}$	$47.2_{(+0.0)}$	$43.6_{(+1.3)}$	$47.2_{(+0.0)}$	$43.1_{(+1.0)}$
InstructBLIP-13B	$47.2_{(+0.0)}$	$43.8_{(+0.3)}$	$47.2_{(+0.0)}$	$43.2_{(+1.5)}$	$47.2_{(+0.0)}$	$42.9_{(+1.1)}$	$47.2_{(+0.0)}$	$43.3_{(+1.0)}$
mBLIP-BLOOMZ	$51.9_{(-0.9)}$	$55.4_{(+7.7)}$	$49.8_{(-3.0)}$	$54.2_{(+5.8)}$	$49.6_{(-3.2)}$	$55.8_{(+8.7)}$	$50.4_{(-2.4)}$	$55.1_{(+7.4)}$
GLaMM	$47.2_{(-10.6)}$	$23.3_{(-0.7)}$	$47.2_{(+0.8)}$	$\mathbf{44.2}_{(-6.9)}$	$47.2_{(-2.8)}$	$42.8_{(-6.1)}$	$47.2_{(-4.2)}$	$36.8_{(-4.6)}$
LLaVA-1.5-7B	$55.2_{(-2.6)}$	$\mathbf{18.4}_{(-3.0)}$	$48.3_{(+4.7)}$	$45.7_{(-4.1)}$	$48.2_{(-5.0)}$	$43.4_{(-1.0)}$	$50.6_{(-1.0)}$	$\mathbf{35.8}_{(-2.7)}$
LLaVA-1.5-13B	$51.6_{(-15.0)}$	$23.9_{(+3.1)}$	$47.3_{(+0.8)}$	$45.0_{(-0.7)}$	$47.5_{(-3.8)}$	$\mathbf{38.9}_{(-4.2)}$	$48.8_{(-6.0)}$	$35.9_{(-0.6)}$
XComposer2	$\mathbf{85.6}_{(-7.0)}$	$18.8_{(+3.0)}$	$51.0_{(+0.5)}$	$51.0_{(-3.3)}$	$\mathbf{53.2}_{(-0.6)}$	$49.8_{(-1.6)}$	$\mathbf{63.3}_{(-2.4)}$	$39.9_{(-0.6)}$
MiniCPM-V	$72.4_{(-4.8)}$	$24.6_{(-2.1)}$	$49.9_{(-2.6)}$	$47.8_{(-3.7)}$	$52.9_{(-0.5)}$	$45.1_{(-6.2)}$	$58.4_{(-2.6)}$	$39.2_{(-4.0)}$
GPT-4o	$78.3_{(+4.6)}$	$28.1_{(-7.0)}$	$\mathbf{53.4}_{(-1.9)}$	$44.6_{(-6.3)}$	$49.1_{(-5.7)}$	$44.9_{(-6.4)}$	$60.3_{(-1.0)}$	$39.2_{(-6.6)}$

Table 1: The accuracy and cosine region parsing errors of VLMs when explicitly prompted to follow each frame of reference are provided (`cam`/`rel`/`add`). The values in parentheses indicate the performance change relative to the scenario with no perspective (`nop`) prompting.

Model	Obj F1 (↑)		Acc% (↑)		$\varepsilon^{\cos}_{\times 10^2}$ (↓)		$\varepsilon^{\text{hemi}}_{\times 10^2}$ (↓)		$\sigma_{\times 10^2}$ (↓)		$\eta_{\times 10^2}$ (↓)		$c^{\text{sym}}_{\times 10^2}$ (↓)		$c^{\text{opp}}_{\times 10^2}$ (↓)	
	BALL	CAR	BALL	CAR	BALL	CAR	BALL	CAR	BALL	CAR	BALL	CAR	BALL	CAR	BALL	CAR
InstructBLIP-7B	66.7	66.7	47.2	47.2	43.9	43.5	57.8	56.4	26.7	30.5	48.4	43.4	17.2	16.9	16.6	22.6
InstructBLIP-13B	67.3	50.3	47.2	47.2	43.0	43.8	55.5	55.9	27.1	36.8	48.2	46.4	17.3	17.0	21.0	21.9
mBLIP-BLOOMZ	99.1	33.3	47.5	51.9	52.1	55.4	62.1	65.6	43.7	48.6	54.1	60.7	29.1	30.1	33.8	42.0
GLaMM	100.0	99.8	47.2	47.2	33.0	23.3	45.2	37.6	29.9	23.4	45.0	28.4	10.1	9.4	13.7	14.6
LLaVA-1.5-7B	100.0	88.6	63.2	55.2	20.7	**18.4**	33.7	32.5	25.2	20.0	23.5	**21.8**	**5.8**	**5.4**	**8.3**	**10.7**
LLaVA-1.5-13B	100.0	98.6	55.3	51.6	25.7	23.8	37.6	37.1	19.3	20.8	24.9	29.9	7.0	5.8	9.3	10.8
XComposer2	100.0	95.3	**92.4**	**85.6**	**20.0**	18.8	**21.1**	**26.3**	**19.2**	**15.3**	**13.7**	22.9	9.0	6.5	10.5	12.0
MiniCPM-V	66.8	81.5	81.0	72.4	22.4	24.6	32.8	35.8	19.2	**19.2**	29.8	22.7	10.1	9.2	12.4	14.9
GPT-4o	100.0	94.5	89.2	78.3	27.4	28.1	27.5	35.0	20.9	24.0	43.1	38.8	14.1	13.3	14.2	16.7
Random (30 trials)	50.0		50.9		46.3		58.7		28.3		26.6		42.5		44.2	
Always "Yes"	50.0		47.2		61.2		68.7		0.0		0.0		0.0		100.0	

Table 2: A comprehensive evaluation of VLMs in egocentric relative FoR with reflected transformation, using an explicit camera perspective (`cam`) prompt, is conducted. The metrics considered include object hallucination (F1-score), accuracy (Acc), region parsing error (ε), prediction noise (η), standard deviation (σ), and consistency (c).

experiments on English conventions, where both *relative* and *intrinsic* FoRs are available, but there is a conventional preference for a *relative* FoR combined with a *reflected* coordinate transformation in the relative FoR [see 35, Table 5.4]. We further extend our setup to multilingual settings by evaluating models in 109 languages across 170 regions worldwide. To cover a variety of VLMs with different capabilities and training approaches, we evaluate the following models: InstructBLIP (7B/13B) [16], LLaVA v1.5 (7B/13B) [39], InternLM-XComposer2 (7B) [19], MiniCPM-Llama3-V v2.5 (8B) [29, 73], GLaMM (7B) [56], mBLIP-BLOOMZ-7B [24], GPT-4o [49].

Firstly, we find that **most VLMs prefer reflected coordinate transformation convention and egocentric relative frame of reference, as detailed in Section A.6.**

3.1 VLMs Fail to Adopt Alternative Frames of Reference Flexibly

We now address the research question: **can VLMs adopt different FoRs when perspectives are explicitly specified to disambiguate spatial expressions?** We again use `COMFORT-CAR`; however, instead of using the no-perspective prompt (`nop`), we require VLMs to follow one FoR by explicitly specifying the perspective (`cam`/`rel`/`add`) in the textual prompts (Table 4). Table 1 shows the results in accuracy and $\varepsilon^{\cos}$ and the performance compared to when no perspective is specified, and Table 9 in the appendix gives the complete evaluation. We find that all models, including the strong ones like GPT-4o and InternLM-XComposer2, show close-to-chance performance (50% accuracy) when being prompted to use the intrinsic or addressee-centered relative FoRs. Compared to the same probing setup without a perspective specified (`nop`), we find generally marginal improvements in region parsing error (ε), whereas the accuracy decreases. Overall, the results indicate that while VLMs can comprehend scenes using egocentric relative FoR, they struggle to adapt flexibly to alternative FoRs.

3.2 Spatial Representations in VLMs Are Not Robust and Consistent

In this section, we further ask: **are spatial representations in VLMs robust and consistent?** The considered metrics include accuracy (Acc), region parsing error (ε), prediction noise (η), standard

Language		English	Tamil	Hausa
Intrinsic		50.9	52.0	54.0
Ego-Rel	Ref.	**35.8**	**40.4**	**41.0**
	Rot.	57.3	55.2	56.1
	Tran.	53.7	51.1	53.0
Add-Rel	Ref.	58.8	52.2	52.8
	Rot.	51.3	52.9	55.3
	Tran.	56.1	56.1	56.1
GPT-4o Prefer		Ego-Ref.	Ego-Ref.	Ego-Ref.
Human Prefer		Ego-Ref.	Ego-Rot.	Ego-Trans.

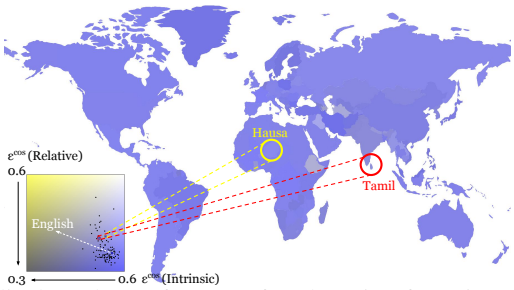


Figure 2: A visualization of the world map that displays the preference of each region for using the intrinsic FoR over the relative FoR. The plot is based on the top three spoken languages in each region, as reported by The World Factbook [8], and averages the cosine parsing error ($\varepsilon^{\cos}$, $\downarrow$), weighted by the speaking population. We present a quantitative comparison of English, Tamil, and Hausa, with the best-performing FoR marked in bold and the convention preferred by human speakers underlined.

deviation (σ), and consistency (c) as defined in Section 2. One commonly considered possibility that VLMs underperform is that they suffer from *object hallucination*, where they misperceive objects in the scenes [36, 13]. Following the object probing setups, we prompt the VLMs to inquire about the presence of an existing object and a non-existing object in the scene, and compute the F1-score (Table 2). We find that the BLIP models suffer from severe object hallucinations, which contribute to their underperformance in the previous evaluation. Many VLMs, despite showing decent performance metrics in terms of spatial understanding and reasoning accuracy, demonstrate a lack of robustness and consistency. For example, the spatial opposite consistency (c^{opp}) of GPT-4o is not significantly better than 30 random trials. In contrast, VLMs that have undergone supervised fine-tuning on spatial relation tasks have a more robust and consistent spatial representation. For instance, InternLM-XComposer2 and MiniCPM-V (on the `COMFORT-BALL`task, with no object hallucinations) show improved performance. On the other hand, although GLaMM is mechanistically grounded to objects and exhibits minimal object hallucination, its spatial understanding capability is poor. This suggests that improving visual entity grounding helps in recognizing individual objects but does not automatically translate to better spatial understanding between multiple objects.

3.3 A Cross-lingual and Cross-cultural Evaluation of Frame of Reference

All previous experiments are centered around English; however, individuals from diverse multilingual and cultural backgrounds may adopt different preferences and conventions to select their FoR in resolving ambiguities [45, 50, 5, 48]. Our next research question naturally arises: **Does multilingual VLMs faithfully follow the preferences and conventions (associated with different languages) to select the FoR?** To extend the study of preferred FoR from English to a multilingual setting, we evaluate 109 languages worldwide to investigate whether each language shows a preferred FoR. We translate the English prompts into the target languages using the Google Cloud Translate API. Given that the open-source language models either lack strong multilingual capabilities or underperform in previous evaluations, we study this problem on the GPT-4o model [49]. We follow the setup similar to Section A.6.2, but only evaluate the images corresponding to the four canonical directions using the `nop` prompt. For each language, we compute $\varepsilon^{\cos}$ for each FoR and coordinate transformation. Figure 2 presents a visualization of the world map, displaying the preference of each region for using the (object-centered) intrinsic FoR over the relative FoR, where the latter corresponds to a low $\varepsilon^{\cos}$ value. Table 10 summarizes the results for all languages tested.

Nearly all tested languages demonstrate a preference towards the relative FoR, except several underrepresented languages, such as Konkani, Kurdish, and Amharic, which exhibit near-random performance without a significant preference. In Figure 2, we present a classic comparison between English, Tamil, and Hausa similar to that of Levinson [35], with the best-performing FoR marked in bold, and the preferred convention by humans underlined. Although human speakers of these languages have different preferred coordinate transformation conventions, the English convention of reflected projection is observed for both Tamil and Hausa. Although, for example, Hausa permits an English-like interpretation of front-back relations, this interpretation is generally less favored and may confuse Hausa speakers [28]. This raises concerns that English may dominate the FoR preference conventions of other languages in multilingual VLMs.

Acknowledgments

This work was supported in part by NSF IIS-1949634, NSERC RGPIN-2024-04395, and the Microsoft Accelerate Foundation Models Research (AFMR) grant program. Ziqiao Ma is supported in part by the Weinberg Cognitive Science Fellowship. The authors would like to thank Yinpei Dai, Run Peng, Jung-Chun Liu, and Xuejun Zhang for proofreading and feedback.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Andrea Bender, Sarah Teige-Mocigemba, Annelie Rothe-Wulf, Miriam Seel, and Sieghard Beller. Being *In Front* is good—but where is *In Front* ? preferences for spatial referencing affect evaluation. *Cognitive Science*, 44(6):e12840, 2020.

[5] Jürgen Bohnemeyer, Katharine Donelson, Randi Tucker, Elena Benedicto, Alejandra Capistrán Garza, Alyson Eggleston, Néstor Hernández Green, María de Jesús Selene Hernández Gómez, Samuel Herrera Castro, Carolyn O'Meara, et al. The cultural transmission of spatial cognition: Evidence from a large-scale study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

[6] Stephen Butterworth et al. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541, 1930.

[7] Laura A Carlson-Radvansky and Gordon D Logan. The influence of reference frame selection on spatial template construction. *Journal of memory and language*, 37(3):411–437, 1997.

[8] Central Intelligence Agency. *The world factbook 2009*. Central Intelligence Agency, 2009. `https://www.cia.gov/the-world-factbook/field/languages/` [Accessed: (Jun 1st 2024)].

[9] Joyce Y Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. Language to action: towards interactive task learning with physical agents. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2–9, 2018.

[10] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[11] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[12] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023.

[13] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F Fouhey, and Joyce Chai. Multi-object hallucination in vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[14] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *arXiv preprint arXiv:2406.01584*, 2024.

[15] Noam Chomsky. *Aspects of the Theory of Syntax*. MIT press, 1965.

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[17] Eve Danziger. Deixis, gesture, and cognition in spatial frame of reference typology. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 34 (1):167–185, 2010.

[18] Vittoria Dentella, Fritz Günther, and Evelina Leivada. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120, 2023.

[19] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[20] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488. PMLR, 2023.

[21] Carola Eschenbach. Contextual, functional, and geometric components in the semantics of projective terms. In *Functional Features in Language and Space: Insights from Perception, Categorization, and Development*. Oxford University Press, 12 2004.

[22] Andrew U Frank. Formal models for cognition—taxonomy of spatial location description and frames of reference. *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge*, pages 293–312, 1998.

[23] Nancy Franklin, Linda A Henkel, and Thomas Zangas. Parsing surrounding space into regions. *Memory & Cognition*, 23:397–407, 1995.

[24] Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. mblip: Efficient bootstrapping of multilingual vision-llms. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 7–25, 2024.

[25] Dedre Gentner, Asli Özyürek, Özge Gürcanli, and Susan Goldin-Meadow. Spatial language facilitates spatial cognition: Evidence from children who lack language input. *Cognition*, 127 (3):318–330, 2013.

[26] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[27] William G Hayward and Michael J Tarr. Spatial language and spatial representation. *Cognition*, 55(1):39–84, 1995.

[28] Clifford Hill. Up/down, front/back, left/right. a contrastive study of hausa and english. *Here and there: Cross-linguistic studies on deixis and demonstration*, 1342, 1982.

[29] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[30] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. MMTom-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, 2024.

[31] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, 2023.

[32] Adam Kendon. Spacing and orientation in co-present interaction. *Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March 23-27, 2009, Revised Selected Papers*, pages 1–15, 2010.

[33] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[34] Stephen C Levinson. Frames of reference and molyneux's question: Crosslinguistic evidence. *Language and Space*, pages 109–170, 1996.

[35] Stephen C Levinson. *Space in language and cognition: Explorations in cognitive diversity*, volume 5. Cambridge University Press, 2003.

[36] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[37] Changsong Liu, Jacob Walker, and Joyce Y Chai. Ambiguities in spatial language understanding in situated human robot dialogue. In *2010 AAAI Fall Symposium Series*, 2010.

[38] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36, 2023.

[40] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[41] Gordon D Logan and Daniel D Sadler. A computational analysis of the apprehension of spatial relations. *Language and Space*, pages 493–530, 1996.

[42] Ziqiao Ma, Jiayi Pan, and Joyce Chai. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 524–544, 2023.

[43] Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. Towards a holistic landscape of situated theory of mind in large language models. *Findings of Empirical Methods in Natural Language Processing*, 2023.

[44] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023.

[45] Asifa Majid, Melissa Bowerman, Sotaro Kita, Daniel BM Haun, and Stephen C Levinson. Can language restructure cognition? the case for space. *Trends in cognitive sciences*, 8(3):108–114, 2004.

[46] Reinhard Moratz and Thora Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial cognition and computation*, 6(1):63–107, 2006.

[47] Edward Munnich, Barbara Landau, and Barbara Anne Dosher. Spatial language and spatial representation: A cross-linguistic comparison. *Cognition*, 81(3):171–208, 2001.

[48] Awino Ogelo and Emanuel Bylund. Spatial frames of reference in dholuo. *Language Sciences*, 104:101614, 2024.

[49] OpenAI. Hello gpt-4o, May 2024. URL `https://openai.com/index/hello-gpt-4o/`.

[50] Carolyn O'Meara and Gabriela Pérez Báez. Spatial frames of reference in mesoamerican languages. *Language Sciences*, 33(6):837–852, 2011.

[51] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024.

[52] Shannon M Pruden, Susan C Levine, and Janellen Huttenlocher. Children's spatial thinking: Does talk about the spatial world matter? *Developmental science*, 14(6):1417–1430, 2011.

[53] Jennie E Pyers, Anna Shusterman, Ann Senghas, Elizabeth S Spelke, and Karen Emmorey. Evidence from an emerging sign language reveals that language supports spatial cognition. *Proceedings of the National Academy of Sciences*, 107(27):12116–12120, 2010.

[54] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the Second Workshop on Open-Vocabulary 3D Scene Understanding*, 2024.

[55] Navid Rajabi and Jana Kosecka. Towards grounded visual spatial reasoning in multi-modal vision language models. In *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR): Harnessing Momentum for Science*, 2024.

[56] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[57] Terry Regier and Laura A Carlson. Grounding spatial language in perception: an empirical and computational investigation. *Journal of experimental psychology: General*, 130(2):273, 2001.

[58] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[59] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024.

[60] Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. Benchmarks as microscopes: A call for model metrology. In *Proceedings of the First Conference on Language Modeling*, 2024.

[61] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023.

[62] Haoyue Freda Shi. *Learning Language Structures through Grounding*. PhD thesis, Toyota Technological Institute at Chicago, 2024.

[63] Anna Shusterman and Peggy Li. Frames of reference in spatial language acquisition. *Cognitive psychology*, 88:115–161, 2016.

[64] Thora Tenbrink. Identifying objects on the basis of spatial contrast: An empirical study. In *International Conference on Spatial Cognition*, pages 124–146. Springer, 2004.

[65] Luca Tommasi and Bruno Laeng. Psychology of spatial cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(6):565–580, 2012.

[66] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212, 2021.

[67] Marina Vasilyeva and Stella F Lourenco. Development of spatial cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3):349–362, 2012.

[68] Alex Warstadt and Samuel R Bowman. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press, 2022.

[69] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[70] Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F Fouhey, and Joyce Chai. 3d-grand: A million-scale dataset for 3d-llms with better grounding and less hallucination. *arXiv preprint arXiv:2406.05132*, 2024.

[71] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2023.

[72] Shoubin Yu, Jaehong Yoon, and Mohit Bansal. Crema: Multimodal compositional video reasoning via efficient modular adaptation and fusion. *arXiv preprint arXiv:2402.05889*, 2024.

[73] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.

[74] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction in robotics. In *8th Annual Conference on Robot Learning*, 2024.

[75] 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.

[76] Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5718–5728, 2023.

[77] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

A Appendix

A.1 Extended Background and Related Work

The recent success of large language models has sparked breakthroughs in multi-modalities, leading to the development of many vision-language models [VLMs; 12, 49, 58, *inter alia*]. With some benchmarks developed to evaluate the downstream performance of these models [40, 75], there has been growing excitement around evaluations and analyses inspired by human cognitive capabilities such as referential grounding [42], compositional reasoning [44], visual illusions [76, 26], and theory of mind [30]. One direction among them that captures significant attention is spatial language understanding and reasoning, leading to several benchmarks [31, 38] and enhanced models [10, 14].

Indeed, spatial cognition is a crucial part of human cognitive capability, developed since infancy and continuing through the elementary school years [65, 67]. Language is closely intertwined with spatial cognition, with each contributing to the acquisition of the other [27, 57, 53, 52, 25]. While spatial language and non-linguistic spatial representations in memory are closely correlated and share foundational properties, they are, to some extent, divergent—spatial conventions are not consistently preserved across different languages or tasks, and humans demonstrate flexibility in using multiple coordinate systems for both non-linguistic reasoning and linguistic expressions [47, 63]. Thus, spatial language is inherently ambiguous, and as we quote:

> Languages just do turn out to use fundamentally different semantic parameters in their categorization of spatial relations—different coordinate systems, different principles for constructing such coordinate systems, yielding different categorizations of 'same' and 'different' across spatial scenes.
>
> *Stephen C. Levinson (2003)*

A.1.1 Spatial Language and Spatial Representation

Some projective terms, such as the English words *front*, *back*, *right*, and *left*, convey meanings of spatial relations [21]. These terms articulate the spatial relation between two entities within a designated *frame of reference* (FoR), often involving one entity as the reference object (*relatum*/*ground*) and another target object (*referent*/*figure*) that is positioned relative to the relatum along a specific axis/direction [34, 22]. In situated communication, speech act participants (e.g., an *addressee*) may also be considered [17]. To determine acceptable uses of various spatial relations, existing theories suggest that people fit *spatial templates*, which are centered on the relatum and aligned with the FoR [41], to parse out *regions of acceptability* of certain directions [23, 7].

Ambiguities in frame of reference. The choice of perspectives may lead to different FoRs, where Levinson [35] has identified three main types of FoR: *absolute*, *intrinsic*, and *relative*. The absolute FoR uses cardinal directions, such as *north* and *south*, as fixed bearings. The intrinsic FoR aligns the origin with the relatum, describing the referent's position relative to the relatum's inherent orientation. The relative FoR positions a *viewer* (egocentric or addressee) as the origin, focusing on the observer's intrinsic perspective. Liu et al. [37] have highlighted the ambiguities in situated communication among three variations of intrinsic and relative FoRs (Figure 3): the *egocentric relative*, the *addressee-centered relative*, and the *object-centered intrinsic* FoRs.[3] When not specified, these FoRs are not easily distinguishable based solely on their linguistic expressions [64]. To resolve the ambiguity, individuals from diverse linguistic and cultural backgrounds adopt different preferences and conventions in choosing FoRs [45, 50, 5, 4, 48].

Ambiguities in relative FoRs. The variations of relative FoRs form another source of ambiguity. After putting the origin of the coordination system on the viewer, multiple strategies specifying how to transform the axes can be considered (Figure 1b). Different languages use different transformation conventions to resolve the ambiguity of the front-back and left-right of a relatum [35, 63], including: (1) *translated* projection (e.g., Hausa) where the coordinate frame of the speaker is directly applied, (2) *rotated* projection (e.g., Tamil), where the coordinate frame of the speaker is transformed with a 180-degree rotation, and (3) *reflected* projection (e.g., English), where only the front-back axis is reversed.

[3] We exclude the absolute FoR from our study as it introduces little ambiguity [37].

Origin	Frame of Reference	Example (English)
Camera (Preferred)	Egocentric Relative FoR	(From the camera's viewpoint,) the ball is **behind** the car.
Addressee	Addressee-Centered Relative FoR	(From the woman's viewpoint,) the ball is to the **left** of the car.
Reference	Object-Centered Intrinsic FoR	(From the car's viewpoint,) the ball is to the **right** of the car.

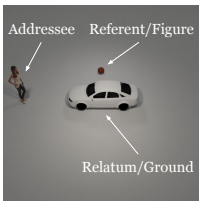
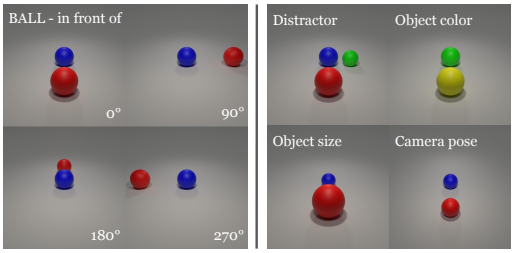
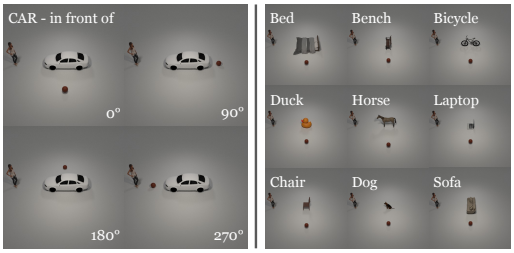


Figure 3: An illustrative example of how a frame of reference (FoR) specifies the reference system when describing the spatial relation between a target object (i.e., the ball) and a reference object (i.e., the car). When the FoR is not explicitly specified, English prefers an egocentric relative FoR, i.e., "the ball is behind the car." We study FoRs that lead to ambiguity [37].



(a) Sample images from COMFORT-BALL. The 4 images on the left are selected every 90° interval along the rotational path out of 36 images. The 4 images on the right illustrate variations with a distractor, different object colors, sizes, or camera poses.



(b) Sample images from COMFORT-CAR. The 4 images on the left are selected every 90° interval along the rotational path out of 36 images. The 9 images on the right are sample images of each variation with different relatum objects.

Figure 4: Examples from the COMFORT-BALL and COMFORT-CAR datasets.

A.1.2 Spatial Understanding in Vision-Language Models

Large language models (LLMs) have exhibited strong adaptability that extends beyond text, encompassing 2D and 3D vision [66, 2, 70], their affordances in the physical embodiment [20, 54, 74], and various other modalities [72]. Especially, a variety of vision-language models (VLM) have been developed by visual instruction tuning on paired text-image data [16, 39, 19]. With supervised fine-tuning using entity-phrase mappings in text-image pairs, grounded VLMs have been developed for fine-grained vision-language understanding at both the region [11, 3, 71, 51] and pixel level [33, 69, 56, 77].

Spatial understanding is known to be challenging even for state-of-the-art VLMs and is receiving increasing attention [1]. Besides using spatial language understanding modules [55], recent works such as SpatialVLM [10] and SpatialRGPT [14] improve spatial reasoning in VLMs by leveraging 3D VQA or scene graph data for supervised fine-tuning. Several benchmarks have been developed to evaluate spatial reasoning in VLMs. The Visual-Spatial Reasoning (VSR) [38] dataset contains 66 types of spatial relations in real text-image pairs. SpatialRGPT-Bench [14] builds on 3D annotations for indoor, outdoor, and simulated environments, focusing on 3D spatial cognition. What'sUp [31] curates images that vary only in the spatial relations of objects while keeping the object identity fixed, allowing for controlled evaluation. Still, these benchmarks consist of text-image pairs where objects may or may not indicate certain spatial relations. They overlook ambiguities related to the frame of reference, lack spatial continuity, and do not propose metrics to evaluate the robustness and consistency of spatial reasoning.

A.2 Dataset Details

A.2.1 Dataset Configurations

The entire data generation pipeline produces 720 English test cases in COMFORT-BALL, and 57.6k English test cases in COMFORT-CAR. For COMFORT-BALL: 1 object combination $\times$ 5 variants $\times$ 4 relations $\times$ 36 angles = 720 test cases. For COMFORT-CAR: 20 object combinations $\times$ 5 variants $\times$ 4 relations $\times$ 36 angles $\times$ 4 prompts = 57,600 test cases. The table below lists all possible variants and configurations for the dataset, and we describe our dataset configuration in detail as follows.

Test Case Setup	Possible Variants
Scene $\mathcal{S}$	COMFORT-BALL: **Relatum**: *red ball*; **Referent**: *blue ball*; 36 samples uniformly collected along a rotational path.
	COMFORT-CAR: **Relatum**: *basketball*; **Referent**: *horse*, *car*, *bench*, *laptop*, *rubber duck*, *chair*, *dog*, *sofa*, *bed*, *bicycle*; **Addressee**: woman; 36 samples uniformly collected along a rotational path.
Spatial Relation $\mathcal{R}$	*to the left of*, *to the right of*, *in front of*, *behind*
Frame of Reference $\mathcal{F}$	*egocentric relative, addressee-centered relative*, *object-centered intrinsic*
Language $\mathcal{L}$	See Table 10.

Table 3: A test case in COMFORTis defined as a 4-tuple in $\mathcal{S} \times \mathcal{R} \times (\mathcal{F} \cup \{\emptyset\}) \times \mathcal{L}$. This table enumerates all possible variants and configurations of the dataset.

A.2.2 List of Evaluated Languages

We started with 132 candidate languages supported by Google Translate API.[4] We removed 23 languages from our multilingual evaluation due to their failure to adhere to instructions for generating "yes" and "no" predictions, or because they did not pass the back-translation test for quality control: Aymara, Bambara, Croatian, Dhivehi, Dogri, Ewe, Guarani, Hmong, Kyrgyz, Luganda, Malayalam, Meiteilon (Manipuri), Mizo, Odia (Oriya), Punjabi, Quechua, Samoan, Tatar, Telugu, Tigrinya, Uyghur, Xhosa, Yoruba.

A.2.3 Task Formulation

Following the setups in object hallucination evaluation [36, 13], we formulate the task as a spatial relation inference problem. In this task, a VLM $\mathcal{M}$ is presented with an RGB image x_{img} and a textual question x_{query}. The image shows the egocentric perception of a scene $s \in \mathcal{S}$, where $\mathcal{S}$ is the set of possible scenes in which the referent moves along a rotational trajectory with a constant radius from the relatum. In contrast to fixing the referents on the standard canonical axes, this setup better mirrors the spatial continuity in common real-world scenarios.

A.2.4 Scene Setup

In COMFORT, there are configurations determined by whether the relatum has an intrinsic semantic front:

- **COMFORT-BALL**: When the relatum is non-fronted (e.g., Figure 1b), we focus on the ambiguity of FoR conventions associated with different languages. The split involves an observer's egocentric perception of a referent (e.g., a red ball) and a non-fronted relatum (e.g., a blue ball). We further randomize the dataset with object-level (colors, sizes, and shapes) and scene-level variations (camera positions and distractors) to consider more diverse yet reasonable settings (Figure 4a).
- **COMFORT-CAR**: When the relatum is fronted (e.g., Figure 1a), multiple FoRs can be explicitly adopted to interpret the scene. A COMFORT-CARimage, therefore, involves the egocentric perception of a referent, a fronted relatum, and an additional human addressee. One can interpret the spatial relations using either the Camera, Addressee, or Relatum (C/A/R) as the origin to resolve the reference frame ambiguity. COMFORT-CARhas a set of 10 realistic objects in a typical household or outdoor scene, including *horse*, *car*, *bench*, *laptop*, *rubber duck*, *chair*, *dog*, *sofa*, *bed*, and *bicycle*, all of which have a clear semantic front. We use a basketball as the referent and vary the relatum. In addition to these objects, we include a human addressee in the scene. To disentangle different FoRs as much as possible, we let the addressee face right, and let the relatum face either left or right in the rendered images from the rendering camera's perspective (Figure 4b).

A.2.5 Language Query Setup

Given a pair of referent [A] and a relatum [B], together with a spatial relation of interest, the query is posed as "Is [A] [relation] [B]?" Depending on whether or not and which FoR is specified, the query is appended after four different perspective prompts (Table 4): no perspective (nop), camera perspective (cam), addressee perspective (add), and relatum perspective (rel). We only query from the camera egocentric perspective (cam) for COMFORT-BALL, focusing on the ambiguity introduced by variations of the relative FoR. For COMFORT-CAR, we use all four possible language prompts to study how ambiguity in the reference system is resolved. Overall, for English, the above data generation

[4] https://cloud.google.com/translate

Origin	Prompt Template
`nop`	Is `[A]` `[relation]` `[B]`?
`cam`	From the camera's viewpoint, is `[A]` `[relation]` `[B]`?
`add`	From the `[addressee]`'s viewpoint, is `[A]` `[relation]` `[B]`?
`rel`	From the `[relatum]`'s viewpoint, is `[A]` `[relation]` `[B]`?

Table 4: The origins of each coordinate system and the corresponding prompt templates for querying the FoR given a referent-relation-relatum triple.

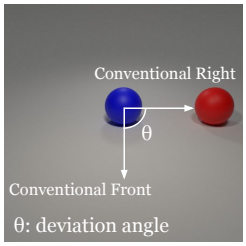


Figure 5: A red ball with a deviation angle $\theta = 90°$ relative to the conventional front (English) of the blue ball.

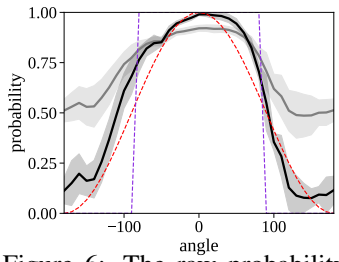


Figure 6: The raw probability $p(\theta)$ in gray, normalized probability $\widehat{p}(\theta)$ in black, and two reference probability $\lambda^{\text{hemi}}(\theta)$ and $\lambda^{\cos}(\theta)$ in purple and red.

pipeline leads to 720 test cases in `COMFORT-BALL`, and 57.6k test cases in `COMFORT-CAR`. The same method for dataset synthesis can be generalized to any other language; however, for computational efficiency, we only include the scenes corresponding to the four most prototypical directions (i.e., left, right, front, and back) in our multilingual analysis.

A.3 Metrics

Quantitatively assessing the spatial understanding and reasoning capabilities of models is challenging for two reasons. First, the physical world is continuous, and spatial relations may extend beyond the precise canonical front-back and left-right axes. As noted by Carlson-Radvansky and Logan [7], there exists *regions of acceptability* where, for instance, an object slightly to the front-left might still be considered being "in front." Second, as Dentella et al. [18] pointed out, language models are biased towards affirmative responses. However, the intermediate representations may be sensitive to variations in input and, to some extent, align with human perceptions of spatial cues. Based on these concerns and findings, we introduce multiple metrics to evaluate the models' competence to enable more nuanced analyses in addition to accuracy that measures performance.

Unless further clarified, we adopt a right-handed coordinate system with the thumb pointing upwards when describing angles. We define the *deviation angle* $\theta \in (-180°, 180°]$ as the angular displacement from the canonical direction r to the vector connecting the relatum and target. For example, in Figure 5, the deviation angle of canonical right from canonical front is $\theta = 90°$. Following Carlson-Radvansky and Logan [7], we define the acceptable region for a spatial relation r as the 180-degree hemisphere centered at the corresponding canonical direction. For a VLM $\mathcal{M}$ and a test case indexed by i, we let $P_i(\textit{response}; \mathcal{M})$ denote the probability of *response* $\in \{\text{Yes}, \text{No}\}$ assigned by $\mathcal{M}$, and abbreviate it as $P_i(\textit{response})$ if there is no confusion.

Accuracy. Given a spatial relation r in the textual prompt, we assess whether the assigned response probabilities correspond to whether the referent lies within the acceptable region defined by the relatum and r. Formally, we define the local probability of the model responding 'Yes' by $p_i = P_i(\text{Yes}) / [P_i(\text{Yes}) + P_i(\text{No})]$. We consider the inference correct if (1) the scene falls into the acceptability region and $p_i > 0.5$ or (2) the scene falls out of the acceptability region and $p_i \leq 0.5$.

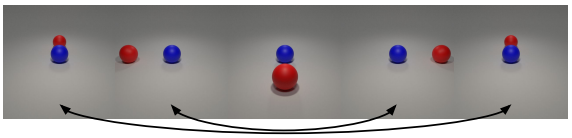
Region Parsing Error. To mitigate the known bias towards affirmative answers, where the expectation of p_i is generally higher than 0.5, we normalize it across all image-prompt pairs, resulting in the normalized probability

$$\widehat{p}_i := \frac{p_i - \min_j p_j}{\max_j p_j - \min_j p_j}.$$

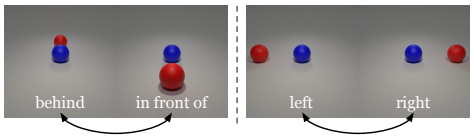
We adopt the root mean square error (RMSE) between the normalized acceptance probability $\widehat{p}$ and reference probability threshold λ^{ref} that represents the actual regions of acceptability,

$$\varepsilon^{\text{ref}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\widehat{p}_i - \lambda^{\text{ref}})^2},$$

where λ^{ref} denotes the reference of the assigned probability, analogically to ground-truth labels in machine learning terms.



(a) An illustration of the spatial symmetry with respect to the (conventional) front. As the red ball rotates around the blue ball, spatial symmetry consistency ensures that each symmetric pair, with different deviation angles θ and $-\theta$, has the same probability of being identified as the front.



(b) Antonyms for spatial opposite consistency, e.g., When evaluating if the red ball is to the left of the blue ball, spatial opposite consistency ensures the probability of accepting a sample as left equals the probability of identifying it as not right.

Figure 7: Illustrations for the consistency metrics defined in COMFORT.

We introduce two analytically and geometrically motivated proposals defining λ^{ref}, λ^{hemi} and $\lambda^{\cos}$, based on hemispheres and cosine of angles, respectively. First, the hemisphere-based reference λ^{hemi} is defined as

$$\lambda^{\text{hemi}}(\theta) := \begin{cases} 1 & \text{if } \theta \in (-90°, 90°) \\ 0 & \text{if } \theta \in (-180°, -90°] \cup [90°, 180°]. \end{cases}$$

Here, $\theta = 0°$ corresponds to the most prototypical spatial relation, and $\theta = 180°$ corresponds to the opposite. Intuitively, $\lambda^{\text{hemi}} = 1$ denotes the test case falls into the acceptable region defined by the textual prompt, and otherwise not.

The second reference is derived from the cosine of the deviation angle. Matching the range of the cosine function to that of probability, i.e., $[0, 1]$, we define the cosine-based reference $\lambda^{\cos}(\theta)$ by

$$\lambda^{\cos}(\theta) := [\cos(\theta) + 1]/2.$$

Figure 6 shows an example of the vanilla probability curve $p(\theta)$ from LLaVA-v1.5-7B [39], normalized probability curve $\widehat{p}(\theta)$, and two reference curves $\lambda^{\text{hemi}}(\theta)$ and $\lambda^{\cos}(\theta)$. We report both $\varepsilon^{\text{hemi}}(\theta)$ and $\varepsilon^{\cos}(\theta)$ in experiments. We also note that in human spatial cognition, the regions of acceptability are neither mutually exclusive 90° quadrants nor overlapping 180° hemispheres, as they vary across individuals and depend on the situational context [23]. We discuss the limitations of this design in Section B.4.

A.4 Robustness Metrics

Standard deviation. In COMFORT, some images depict variations of the same scene, sharing identical spatial relations between the referent and the object but differing in terms of object colors, sizes, or distractors. When the spatial relation and the query text remain unchanged, an ideal model should have consistent predictions for all variations. To measure the robustness of the model prediction, we report the average standard deviation of the predicted probability $\widehat{p}_i$ across all deviation angles $\sigma := \text{avg}_\theta \sigma(\theta)$.

Prediction noise. Since our data is generated through interpolation, ideally, if a model well understands spatial relations, the probability curve with respect to the deviation angle should be low-frequency (i.e., smooth) rather than high-frequency (i.e., noisy). Therefore, we measure the noise by the RMSE, denoted by η, between the predicted probability and a Butterworth Low Pass Filter [LPF; 6]:

$$\eta := \sqrt{\frac{1}{n}\sum_{i=1}^{n} [\widehat{p}_i - \text{LPF}(\widehat{p}_i)]^2}.$$

A smaller value of η indicates that the probability curve is smoother, which is more desirable.

A.5 Consistency Metrics

Spatial symmetric consistency. A critical aspect of consistent spatial reasoning is geometric symmetry. As our tested target object rotates around the relatum in a circular path that is spatially symmetric, we expect the probabilities of an ideal VLM to consistently reflect geometric symmetry (Figure 7a). For a pair of test cases, indexed by i and j, that have the same configurations but opposite deviation angles, i.e., $\theta_i + \theta_j = 0°$, we define the symmetry consistency:

$$c^{\text{sym}} := \sqrt{\frac{2}{n-2}\sum_{i,j} (\widehat{p}_i - \widehat{p}_j)^2}.$$

Model	Back $\varepsilon^{\cos}$ (↓)		Front $\varepsilon^{\cos}$ (↓)		Left $\varepsilon^{\cos}$ (↓)		Right $\varepsilon^{\cos}$ (↓)		Aggregated			Preferred
	Same	**Rev.**	Same	**Rev.**	**Same**	Rev.	**Same**	Rev.	Tran.	Rot.	**Ref.**	
InstructBLIP-7B	45.6	**39.0**	**31.6**	52.0	**37.2**	48.0	47.5	**37.8**	**40.5**	44.2	43.9	-
InstructBLIP-13B	**40.9**	45.5	46.0	**37.4**	**43.4**	44.9	45.6	**41.6**	44.0	**42.3**	43.0	-
mBLIP	**51.2**	53.7	51.2	**47.9**	**52.4**	53.5	54.6	**46.8**	52.3	**50.5**	52.1	-
GLaMM	58.3	**33.3**	43.9	**42.9**	**38.3**	51.8	**17.3**	63.7	39.5	47.9	**33.0**	**Ref.**
LLaVA-1.5-7B	54.0	**32.9**	59.1	**24.8**	**11.9**	70.0	**13.0**	68.5	34.5	49.0	**20.7**	**Ref.**
LLaVA-1.5-13B	61.8	**19.2**	56.0	**27.7**	**31.7**	61.8	**24.3**	64.3	43.4	43.2	**25.7**	**Ref.**
XComposer2	73.2	**17.9**	74.5	**20.7**	**20.1**	80.9	**21.3**	81.1	47.3	50.1	**20.0**	**Ref.**
MiniCPM-V	70.9	**21.9**	64.3	**26.9**	**19.7**	74.1	**21.1**	73.3	44.0	49.1	**22.4**	**Ref.**
GPT-4o	75.7	**28.2**	73.6	**32.0**	**24.3**	80.8	**25.1**	80.8	49.7	55.5	**27.4**	**Ref.**

Table 5: Preferred coordinate transformation mapping from the egocentric viewer (camera) to the relatum in the relative FoR. The cosine region parsing errors $\varepsilon^{\cos}$ are computed against both the Same and Reversed directions relative to the egocentric viewer's coordinate system. For example, native English speakers typically prefer a Reflected transformation, which maintains the lateral (left/right) axis but reverses the sagittal (front/back) axis relative to the viewer (Figure 1). We determine the preferred transformation based on the aggregated performance, with "-" for no significant preference.

Spatial opposite consistency. Similarly, we expect the probabilities of an ideal VLM to consistently reflect geometric opposition (Figure 7b). For example, the probability that a sample is accepted by the spatial relation "to the left" should be identical to the probability that it is rejected by "to the right." For a pair of opposite spatial relation $r, \text{opp}(r) \in \mathcal{R}$ with the same configurations including the deviation angles θ_i, the opposition consistency is given as:

$$c^{\text{opp}} := \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\widehat{p}_i^{\,r} + \widehat{p}_i^{\,\text{opp}(r)} - 1)^2}.$$

A.6 Additional Results, Figures and Tables

A.6.1 Most VLMs Prefer Reflected Coordinate Transformation Convention

In this section, we address the research question: **do VLMs have a preferred coordinate transformation convention, and if so, what is it?** Experiments are conducted on `COMFORT-BALL`using the camera perspective prompt (`cam`) that explicitly specifies an egocentric relative FoR (Table 5). Table 7 in the appendix shows the complete evaluation including $\varepsilon^{\text{hemi}}$ and $\varepsilon^{\cos}$.

We observe that almost all VLMs demonstrate a clear preference over the *reflected* transformation similar to English, except the BLIP series. Still, some models are also affected by the ambiguity of multiple transformation conventions. With the textual prompting specifying a relation, at $\theta = 0$, GPT-4o and LLaVA-1.5-13B show a sharp drop of performance and a significant variance for `behind` and `right`, respectively (Figure 8), indicating that some models are sensitive to other transformations.

A.6.2 Most VLMs Prefer Egocentric Relative Frame of Reference

We now attempt to answer the research question: **do VLMs have a preferred frame of reference, and if so, what is it?** We conduct our study on `COMFORT-CAR`using the no perspective prompt (`nop`) that deliberately leaves the FoR ambiguous. When calculating the performance with respect to relative FoRs (either egocentric or addressee-centered), we assume a reflected coordinate transformation convention. Table 6 shows the results of preferred FoR in English measured by the region parsing error $\varepsilon^{\cos}$, and Table 8 in the appendix shows the complete evaluation including both $\varepsilon^{\text{hemi}}$ and $\varepsilon^{\cos}$. Almost all VLMs demonstrate a significant preference for the *egocentric relative* FoR similar to English, again, except for the BLIP series. Additionally, the models' performances are inconsistent across different spatial relations—models generally perform better in the lateral directions (left and right) than the sagittal ones (front and behind), even in competitive industry models like GPT-4o. For instance, GLaMM does not show a very strong preference when resolving ambiguities along the sagittal axes, but it demonstrates a significant preference when resolving lateral ambiguity.

Model	Back $\varepsilon^{\cos}$ (↓)			Front $\varepsilon^{\cos}$ (↓)			Left $\varepsilon^{\cos}$ (↓)			Right $\varepsilon^{\cos}$ (↓)			Aggregated			Prefer
	Ego.	Int.	Add.	**Ego.**	Int.	Add.	**Ego.**	Int.	Add.	**Ego.**	Int.	Add.	**Ego.**	Int.	Add.	
InstructBLIP-7B	41.0	**38.6**	**38.6**	**40.9**	46.9	46.9	45.6	**32.5**	51.9	39.6	51.2	**31.8**	**41.8**	42.3	42.3	-
InstructBLIP-13B	**32.9**	34.4	34.4	52.5	**48.5**	**48.5**	47.8	56.2	**27.8**	40.6	**27.6**	56.6	43.5	**41.7**	41.8	-
mBLIP-BLOOMZ	**52.2**	53.2	53.2	45.3	**44.6**	**44.6**	47.8	**47.6**	48.1	45.4	48.4	**42.4**	47.7	48.4	**47.1**	-
GLaMM	**28.0**	49.1	49.1	**30.0**	40.2	40.2	**14.0**	56.8	41.5	**13.7**	53.0	46.6	**21.4**	49.8	44.4	**Ego.**
LLaVA-1.5-7B	**20.9**	43.0	43.0	34.5	**32.6**	**32.6**	**13.4**	53.5	47.4	**14.3**	53.6	49.3	**20.8**	45.7	43.1	**Ego.**
LLaVA-1.5-13B	**31.9**	38.8	38.8	**24.8**	57.1	57.1	**11.7**	51.1	51.1	**27.5**	57.4	48.7	**24.0**	51.1	48.9	**Ego.**
XComposer2	**12.7**	49.3	49.3	**15.2**	48.3	48.3	**18.8**	61.2	53.7	**16.5**	58.4	54.5	**15.8**	54.3	51.4	**Ego.**
MiniCPM-V	**34.2**	40.7	40.7	**35.5**	53.4	53.4	**18.0**	53.9	58.4	**19.0**	58.1	52.7	**26.7**	51.5	51.3	**Ego.**
GPT-4o	38.3	**36.7**	**36.7**	**43.1**	50.2	50.2	**34.7**	59.3	56.5	**24.3**	57.3	61.7	**35.1**	50.9	51.3	**Ego.**

Table 6: Preferred frame of reference in VLMs. Models' Cosine Region Parsing Errors $\varepsilon^{\cos}$ are computed against the Intrinsic FoR (relatum origin), Egocentric relative FoR (camera origin), and Addressee-centric relative FoR (addressee origin). English typically prefers an egocentric relative FoR. We determine the preferred FoR based on the aggregated performance, with "–" indicating no significant preference.

	Back						Front					
	Same			Reversed			Same			Reversed		
	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$
InstructBLIP-7B	47.2	58.4	45.6	47.2	53.8	39.0	47.2	47.5	31.6	47.2	64.6	52.0
InstructBLIP-13B	47.2	55.9	40.9	47.2	56.6	45.5	47.2	60.0	46.0	47.2	53.0	37.4
mBLIP	56.1	60.2	51.2	47.2	64.8	53.7	51.1	61.4	51.2	47.8	58.0	47.9
GLaMM	47.2	71.1	58.3	47.2	46.3	33.3	47.2	55.4	43.9	47.2	55.9	42.9
LLaVA-1.5-7B	47.2	66.7	54.0	47.2	47.0	32.9	47.2	71.0	59.1	47.2	36.4	24.8
LLaVA-1.5-13B	47.2	73.8	61.8	47.2	36.3	19.2	42.8	67.3	56.0	51.7	39.1	27.7
XComposer2	13.3	84.5	73.2	90.0	26.3	17.9	15.0	85.8	74.5	85.0	31.6	20.7
MiniCPM-V	13.9	84.1	70.9	80.6	35.6	21.9	26.1	77.0	64.3	75.0	35.3	26.9
GPT-4o	16.7	87.3	75.7	87.8	30.3	28.2	25.6	82.4	73.6	80.0	40.2	32.0

	Left						Right					
	Same			Reversed			Same			Reversed		
	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$
InstructBLIP-7B	47.2	51.5	37.2	47.2	61.6	48.0	47.2	61.4	47.5	47.2	52.0	37.8
InstructBLIP-13B	47.2	54.2	43.4	47.2	57.0	44.9	47.2	58.1	45.6	47.2	52.5	41.6
mBLIP	47.2	59.8	52.4	47.2	64.2	53.5	47.8	65.7	54.6	47.8	56.4	46.8
GLaMM	47.2	48.9	38.3	47.2	65.5	51.8	79.4	29.8	17.3	15.0	76.2	63.7
LLaVA-1.5-7B	47.2	25.3	11.9	47.2	83.4	70.0	47.2	26.0	13.0	47.2	80.9	68.5
LLaVA-1.5-13B	62.8	39.1	31.7	31.7	76.8	61.8	91.1	35.8	24.3	8.9	79.3	64.3
XComposer2	97.8	11.3	20.1	3.3	95.6	80.9	96.7	15.2	21.3	3.3	95.8	81.1
MiniCPM-V	86.1	27.7	19.7	9.4	88.1	74.1	82.2	32.7	21.1	12.2	87.0	73.3
GPT-4o	94.4	20.4	24.3	11.1	92.6	80.8	94.4	19.0	25.1	11.1	92.8	80.8

	Aggregated									
	Translated			Rotated			Reflected			Preferred Transform
	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$	Acc%	$\varepsilon^{\text{hemi}}_{\times 10^2}$	$\varepsilon^{\cos}_{\times 10^2}$	
InstructBLIP-7B	47.2	54.7	40.5	47.2	58.0	44.2	47.2	57.8	43.9	Not Significant
InstructBLIP-13B	47.2	57.1	44.0	47.2	54.8	42.3	47.2	55.5	43.0	Not Significant
mBLIP	50.6	61.8	52.3	47.5	60.9	50.5	47.5	62.1	52.1	Not Significant
GLaMM	55.3	51.3	39.5	39.2	61.0	47.9	55.3	45.2	33.0	Reflected
LLaVA-1.5-7B	47.2	47.3	34.5	47.2	61.9	49.0	47.2	33.7	20.7	Reflected
LLaVA-1.5-13B	61.0	54.0	43.4	34.9	57.9	43.2	63.2	37.6	25.7	Reflected
XComposer2	55.7	49.2	47.3	45.4	62.3	50.1	92.4	21.1	20.0	Reflected
MiniCPM-V	52.1	55.4	44.0	44.3	61.5	49.1	81.0	32.8	22.4	Reflected
GPT-4o	57.8	52.3	49.7	47.5	64.0	55.5	89.2	27.5	27.4	Reflected

Table 7: The full results for testing the preferred coordinate transformation mapping from the viewer to the relatum in the relative frame of reference.

B Discussion and Conclusions

B.1 Do vision-language models represent space and how?

It is insufficient to answer this question by simply querying the model with text-image pairs and comparing the output with a fixed ground truth. We must, at least, query the models with awareness of the ambiguity in FoRs, which is essential in determining how the scenes in the physical world

	Back									Front								
	Egocentric			Intrinsic			Addressee			Egocentric			Intrinsic			Addressee		
	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$
InstructBLIP-7B	47.2	51.4	41.0	47.2	53.0	38.6	47.2	53.0	38.6	47.2	54.2	40.9	47.2	60.7	46.9	47.2	60.7	46.9
InstructBLIP-13B	47.2	43.5	32.9	47.2	48.9	34.4	47.2	48.9	34.4	47.2	66.5	52.5	47.2	61.1	48.5	47.2	61.1	48.5
mBLIP-BLOOMZ	52.8	62.1	52.2	52.8	63.9	53.2	52.8	63.9	53.2	52.8	56.4	45.3	52.8	55.5	44.6	52.8	55.5	44.6
GLaMM	47.2	45.6	31.9	47.2	51.0	38.8	47.2	51.0	38.8	47.2	37.9	24.8	47.2	69.6	57.1	47.2	69.6	57.1
LLaVA-1.5-7B	49.2	41.6	28.0	47.5	60.3	49.1	47.5	60.3	49.1	48.6	43.2	30.0	48.6	52.9	40.2	48.6	52.9	40.2
LLaVA-1.5-13B	50.8	36.8	20.9	48.6	54.7	43.0	48.6	54.7	43.0	47.2	46.5	34.5	47.2	47.3	32.6	47.2	47.3	32.6
XComposer2	91.4	25.0	12.7	53.6	59.9	49.3	53.6	59.9	49.3	87.8	26.6	15.2	55.0	59.3	48.3	55.0	59.3	48.3
MiniCPM-V	66.4	46.5	34.2	60.8	51.3	40.7	60.8	51.3	40.7	57.5	45.0	35.5	50.8	64.6	53.4	50.8	64.6	53.4
GPT-4o	64.2	49.1	38.3	66.4	45.4	36.7	66.4	45.4	36.7	58.1	54.8	43.1	53.6	61.0	50.2	53.6	61.0	50.2

	Left									Right								
	Egocentric			Intrinsic			Addressee			Egocentric			Intrinsic			Addressee		
	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$
InstructBLIP-7B	47.2	59.0	45.6	47.2	45.3	32.5	47.2	62.0	51.9	47.2	53.1	39.6	47.2	61.7	51.2	47.2	45.3	31.8
InstructBLIP-13B	47.2	59.7	47.8	47.2	70.2	56.2	47.2	39.6	27.8	47.2	53.6	40.6	47.2	39.5	27.6	47.2	70.8	56.6
mBLIP-BLOOMZ	52.8	58.2	47.8	52.8	59.7	47.6	52.8	58.4	48.1	52.8	57.7	45.4	52.8	60.6	48.4	52.8	53.8	42.4
GLaMM	75.8	22.3	11.7	46.4	62.0	51.1	52.5	62.3	51.1	60.8	41.8	27.5	44.7	68.5	57.4	53.1	58.7	48.7
LLaVA-1.5-7B	76.7	25.6	14.0	33.9	68.2	56.8	64.4	52.7	41.5	56.4	28.5	13.7	44.2	64.6	53.0	52.5	57.3	46.6
LLaVA-1.5-13B	81.7	23.7	13.4	42.2	65.0	53.5	57.2	58.5	47.4	86.7	26.8	14.3	47.8	64.0	53.6	52.2	59.9	49.3
XComposer2	95.0	18.8	18.8	45.6	70.5	61.2	54.4	64.0	53.7	96.1	17.1	16.5	47.8	68.1	58.4	52.2	64.6	54.5
MiniCPM-V	93.3	20.4	18.0	52.2	64.3	53.9	47.8	68.0	58.4	91.7	22.6	19.0	46.1	68.3	58.1	53.9	62.5	52.7
GPT-4o	78.6	42.1	34.7	48.1	69.4	59.3	51.9	65.8	56.5	93.9	21.8	24.3	52.8	67.0	57.3	47.2	71.0	61.7

	Aggregated									
	Egocentric			Intrinsic			Addressee			Preferred FoR
	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	
InstructBLIP-7B	47.2	54.4	41.8	47.2	55.2	42.3	47.2	55.2	42.3	Not Significant
InstructBLIP-13B	47.2	55.8	43.5	47.2	54.9	41.7	47.2	55.1	41.8	Not Significant
mBLIP-BLOOMZ	52.8	58.6	47.7	52.8	59.9	48.4	52.8	57.9	47.1	Not Significant
GLaMM	57.8	36.9	24.0	46.4	62.8	51.1	50.0	60.4	48.9	Egocentric Relative
LLaVA-1.5-7B	57.7	34.7	21.4	43.5	61.5	49.8	53.3	55.8	44.4	Egocentric Relative
LLaVA-1.5-13B	66.6	33.5	20.8	46.5	57.7	45.7	51.3	55.1	43.1	Egocentric Relative
XComposer2	92.6	21.9	15.8	50.5	64.4	54.3	53.8	61.9	51.4	Egocentric Relative
MiniCPM-V	77.2	33.7	26.7	52.5	62.1	51.5	53.3	61.6	51.3	Egocentric Relative
GPT-4o	73.7	42.0	35.1	55.2	60.7	50.9	54.8	60.8	51.3	Egocentric Relative

Table 8: The full results for testing the preferred frame of reference in VLMs.

	Back									Front								
	Egocentric			Intrinsic			Addressee			Egocentric			Intrinsic			Addressee		
	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$
InstructBLIP-7B	47.2	56.4	45.1	47.2	54.3	41.2	47.2	56.0	42.8	47.2	56.2	42.0	47.2	56.4	43.6	47.2	56.3	43.3
InstructBLIP-13B	47.2	49.2	38.1	47.2	54.0	40.4	47.2	53.8	40.9	47.2	63.0	49.8	47.2	58.6	46.2	47.2	59.7	47.6
mBLIP-BLOOMZ	52.9	65.4	55.3	52.4	64.7	54.8	50.8	66.3	57.1	52.6	66.2	56.3	52.1	63.8	52.8	53.1	67.1	58.9
GLaMM	47.2	46.2	32.7	47.2	54.8	42.4	47.2	62.6	49.9	47.2	40.4	25.3	47.2	55.1	41.6	47.2	51.0	38.3
LLaVA-1.5-7B	49.0	41.6	27.6	47.4	56.3	45.7	46.2	66.7	55.0	47.5	39.4	25.2	47.4	52.9	39.8	47.2	41.1	27.5
LLaVA-1.5-13B	47.2	38.3	22.4	47.2	53.2	41.1	47.2	52.1	39.9	47.2	48.8	36.8	47.2	54.7	41.2	47.2	41.9	26.8
XComposer2	65.4	40.6	26.0	52.2	57.9	47.0	54.0	58.5	47.5	86.9	27.0	17.1	52.1	58.9	47.8	53.1	57.8	46.6
MiniCPM-V	55.7	48.6	36.0	46.9	57.7	45.9	53.8	47.4	36.3	54.7	45.9	35.0	52.2	58.4	45.9	52.2	58.5	46.9
GPT-4o	69.0	41.3	28.7	59.7	50.0	37.7	56.4	48.7	36.0	58.6	52.5	40.3	52.1	57.4	45.1	48.3	60.0	46.9

	Left									Right								
	Egocentric			Intrinsic			Addressee			Egocentric			Intrinsic			Addressee		
	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$	$\mathrm{Acc}_{\%}$	$\varepsilon^{\mathrm{hemi}}_{\times 10^2}$	$\varepsilon^{\mathrm{cos}}_{\times 10^2}$
InstructBLIP-7B	47.2	56.3	43.3	47.2	56.0	43.0	47.2	57.9	47.1	47.2	56.8	43.5	47.2	52.9	41.5	47.2	54.4	41.0
InstructBLIP-13B	47.2	58.0	46.2	47.2	61.7	48.7	47.2	46.5	33.8	47.2	53.5	41.1	47.2	49.8	37.6	47.2	62.6	49.4
mBLIP-BLOOMZ	51.4	65.6	55.4	46.4	67.0	56.4	47.2	64.6	54.8	50.7	65.3	54.4	48.2	63.5	52.8	47.2	62.9	52.3
GLaMM	47.2	29.6	16.9	47.2	57.7	45.8	47.2	53.7	41.5	47.2	34.3	18.3	47.2	58.7	47.1	47.2	53.2	41.6
LLaVA-1.5-7B	64.9	23.7	12.1	50.4	60.2	48.9	49.9	56.3	45.3	59.3	25.2	8.7	47.9	59.8	48.5	49.6	56.7	45.7
LLaVA-1.5-13B	47.2	29.2	18.3	47.2	59.5	47.0	47.2	53.5	41.2	64.7	32.1	17.9	47.4	61.6	50.7	48.5	58.6	47.8
XComposer2	95.6	18.4	16.4	49.7	64.8	54.5	54.0	62.4	51.3	94.4	19.2	15.7	49.9	64.9	54.8	51.8	64.3	53.8
MiniCPM-V	89.4	24.2	13.3	50.4	60.9	50.0	52.1	60.6	49.9	89.9	24.3	14.2	50.1	60.6	49.5	53.3	58.0	47.3
GPT-4o	91.7	24.0	22.8	52.5	60.1	48.6	46.7	59.9	47.6	93.9	22.1	20.5	49.2	59.4	47.0	45.1	61.0	49.1

Table 9: The full results for benchmarking perspective-taking performance in VLMs.

are mapped to spatial expressions [35]. Our experiments confirm that many VLMs are equipped with reasonable spatial representations through vision-language training alone; in particular, most VLMs clearly prefer the egocentric relative FoR with reflected projection, aligning with English conventions. However, our results also show these representations lack robustness and consistency in a continuous space. Similar experimental setups can yield widely varying performance across

Code	Language	Intrinsic	Egocentric			Addressee			Code	Language	Intrinsic	Egocentric			Addressee		
			Ref.	Rot.	Tran.	Ref.	Rot.	Tran.				Ref.	Rot.	Tran.	Ref.	Rot.	Tran.
af	Afrikaans	50.9	33.7	57.8	49.2	56.6	55.5	57.5	ku	Kurdish	56.5	49.5	54.4	53.1	53.1	55.2	54.0
ak	Akan	51.8	39.6	52.2	48.8	50.4	50.6	53.8	la	Latin	52.2	43.9	49.8	55.5	55.1	50.8	56.6
am	Amharic	52.1	47.4	60.7	50.9	56.8	54.2	57.6	lb	Luxembourgish	54.7	35.6	57.6	50.3	58.6	53.0	59.9
ar	Arabic	55.7	35.8	59.0	51.0	56.6	55.8	59.8	ln	Lingala	52.6	45.7	50.3	59.4	54.6	51.3	58.0
as	Assamese	51.6	40.8	55.3	51.6	48.8	52.8	56.0	lo	Lao	55.8	40.1	55.0	53.7	54.7	55.7	55.5
az	Azerbaijani	49.8	41.9	56.2	51.6	50.7	52.4	55.5	lt	Lithuanian	54.4	35.5	58.3	51.6	57.4	56.5	59.1
be	Belarusian	54.4	39.7	61.7	46.5	51.1	51.9	58.9	lv	Latvian	55.8	35.5	57.7	53.8	57.9	58.7	58.8
bg	Bulgarian	56.9	32.8	56.0	51.4	55.7	55.6	58.9	mai	Maithili	52.7	39.8	55.1	49.9	51.2	50.6	56.5
bho	Bhojpuri	51.5	42.8	58.3	47.5	52.3	51.4	56.5	mg	Malagasy	47.1	37.1	52.2	48.1	53.1	50.9	53.5
bn	Bengali	55.8	34.5	57.1	50.2	53.8	56.1	57.4	mi	Maori	52.0	36.6	58.5	47.6	52.0	51.9	58.1
bs	Bosnian	55.1	35.2	58.5	49.6	54.3	53.1	59.4	mk	Macedonian	54.8	37.0	59.1	49.5	56.5	56.0	58.1
ca	Catalan	55.6	34.7	56.9	53.0	56.0	56.4	59.7	mn	Mongolian	54.1	36.7	56.8	47.4	54.7	53.6	54.7
ceb	Cebuano	52.9	40.0	52.7	54.9	55.3	49.5	60.3	mr	Marathi	52.9	34.5	55.0	48.3	51.7	52.4	55.4
ckb	Sorani	50.4	36.1	53.3	50.6	50.3	52.0	56.3	ms	Malay	54.2	33.1	55.8	50.5	55.9	55.4	57.9
co	Corsican	57.6	35.4	57.4	54.3	58.8	58.8	59.7	mt	Maltese	53.4	37.5	56.1	49.2	50.8	53.8	55.5
cs	Czech	56.4	35.7	58.2	52.4	57.0	58.5	58.0	my	Myanmar	54.9	39.3	58.7	51.8	54.2	56.1	58.1
cy	Welsh	55.1	36.7	59.5	48.7	54.9	54.8	58.5	nb	Norwegian	55.1	34.7	57.0	52.1	58.1	57.6	57.3
da	Danish	54.9	33.0	55.2	53.1	57.6	58.3	56.9	ne	Nepali	53.1	39.4	58.4	47.3	52.9	54.1	54.0
de	German	55.7	36.2	58.4	52.8	56.9	56.6	60.3	nl	Dutch	51.7	34.5	56.3	48.3	53.5	51.6	57.8
el	Greek	54.4	34.4	57.1	52.2	57.1	57.5	57.7	nso	Sepedi	53.8	42.6	51.0	57.1	46.3	54.2	53.7
en	English	50.9	35.8	57.3	53.7	58.8	51.3	58.8	ny	Nyanja	53.7	34.5	56.6	48.0	54.4	52.8	56.7
eo	Esperanto	58.0	34.3	56.4	54.6	58.2	58.2	60.2	om	Oromo	51.1	43.5	57.3	50.6	54.9	54.8	52.9
es	Spanish	56.9	36.2	58.1	53.3	57.0	58.5	59.0	pl	Polish	55.8	32.9	55.8	52.5	55.1	55.1	59.5
et	Estonian	53.7	35.1	56.0	51.7	55.0	54.8	58.5	ps	Pashto	53.0	34.6	57.4	48.9	53.7	54.6	57.4
eu	Basque	56.8	34.3	56.8	53.2	56.7	57.1	59.5	pt	Portuguese	56.3	35.9	58.2	51.9	59.1	59.3	57.6
fa	Persian	55.8	32.1	55.3	49.8	54.4	53.8	58.0	ro	Romanian	57.1	34.8	57.0	53.8	58.2	58.6	59.1
fi	Finnish	53.9	33.7	56.7	50.8	56.3	56.1	57.6	ru	Russian	56.2	36.9	58.8	53.0	56.8	56.3	60.8
fil	Filipino	50.9	31.1	54.1	49.2	54.3	54.0	55.7	rw	Kinyarwanda	53.2	35.2	56.7	48.9	54.4	54.1	57.2
fr	French	58.0	35.2	57.4	53.7	58.6	58.5	59.4	sa	Sanskrit	51.9	41.2	54.1	51.9	51.7	56.4	51.6
fy	Frisian	53.9	38.2	58.9	49.6	53.4	53.2	59.2	sd	Sindhi	51.3	40.3	56.5	49.1	54.8	49.8	57.4
ga	Irish	54.0	33.2	55.3	49.2	52.7	55.7	53.9	si	Sinhala	52.4	38.4	54.6	48.6	53.4	51.5	56.6
gd	Scots Gaelic	53.9	35.4	58.5	49.6	54.7	55.8	58.1	sk	Slovak	56.1	37.7	57.1	54.7	57.2	56.7	59.8
gl	Galician	56.6	37.1	59.0	53.4	57.9	57.9	60.0	sl	Slovenian	55.8	36.5	59.3	49.5	53.9	54.3	58.9
gom	Konkani	51.1	53.1	55.5	50.5	52.5	54.9	51.8	sn	Shona	56.0	34.7	56.0	52.2	54.8	55.6	58.5
gu	Gujarati	52.6	36.6	54.2	50.9	55.5	55.3	53.8	so	Somali	53.7	34.3	56.4	48.2	50.0	51.6	58.4
ha	Hausa	54.0	41.0	56.1	53.0	52.8	55.3	56.1	sq	Albanian	53.6	35.1	56.4	49.0	52.6	50.3	60.1
haw	Hawaiian	55.3	42.2	62.1	51.5	60.5	56.2	60.8	sr	Serbian	55.4	34.5	57.2	50.9	52.5	55.0	58.8
he	Hebrew	56.5	36.4	58.8	52.5	57.1	56.5	60.3	st	Sesotho	53.9	38.4	55.4	51.0	51.3	54.4	55.8
hi	Hindi	52.5	37.8	56.6	49.1	54.5	54.6	54.2	su	Sundanese	51.3	36.7	55.0	50.0	53.7	50.4	57.7
ht	Haitian Creole	56.1	36.0	58.3	53.6	58.4	58.2	59.6	sv	Swedish	54.0	33.5	56.7	51.7	55.8	56.3	58.2
hu	Hungarian	55.2	35.0	57.5	50.7	56.1	56.8	56.7	sw	Swahili	55.3	34.2	56.8	52.4	57.2	56.5	58.4
hy	Armenian	52.2	35.4	56.7	48.8	53.6	52.5	57.2	ta	Tamil	52.0	40.4	55.2	51.1	52.2	52.9	54.6
id	Indonesian	55.9	35.6	58.1	52.2	57.1	57.8	58.1	tg	Tajik	55.7	36.7	57.7	49.7	55.4	56.6	56.3
ig	Igbo	54.5	33.8	56.7	47.4	53.6	53.2	55.3	th	Thai	55.5	35.4	57.9	50.8	56.2	57.8	57.3
ilo	Ilocano	50.8	44.6	46.7	58.9	48.9	57.0	48.7	tk	Turkmen	52.3	45.3	59.0	51.5	52.6	51.2	59.1
is	Icelandic	55.9	34.2	57.0	52.2	56.5	58.0	57.3	tr	Turkish	55.3	33.6	56.3	50.8	56.2	57.0	56.3
it	Italian	56.8	35.6	57.6	53.6	57.9	58.2	59.6	ts	Tsonga	49.4	44.6	50.0	53.7	53.3	51.5	53.6
ja	Japanese	54.7	34.5	56.9	50.4	54.4	55.9	57.3	uk	Ukrainian	56.6	36.1	58.8	50.1	56.8	55.7	59.7
jv	Javanese	53.5	35.3	57.7	51.0	55.7	54.8	58.9	ur	Urdu	52.3	34.6	56.7	49.7	54.0	55.1	57.0
ka	Georgian	51.1	34.8	54.3	50.6	52.0	54.0	55.3	uz	Uzbek	52.6	34.5	56.4	48.1	51.7	53.2	56.6
kk	Kazakh	52.6	36.5	58.8	50.4	54.0	56.1	56.9	vi	Vietnamese	53.9	34.6	58.5	48.6	55.7	56.1	56.8
km	Khmer	55.6	37.6	60.2	50.5	56.9	56.6	59.3	yi	Yiddish	56.7	36.5	57.9	53.5	56.8	57.3	60.0
kn	Kannada	52.3	40.8	53.2	49.5	49.6	51.9	53.4	zh	Yoruba	54.6	35.5	58.3	51.4	56.9	57.2	57.8
ko	Korean	53.6	36.5	59.1	49.7	53.3	53.5	59.8	zu	Chinese	55.6	35.9	57.7	53.1	55.3	56.5	60.2
kri	Krio	58.3	36.2	57.1	51.2	56.1	53.2	60.2									

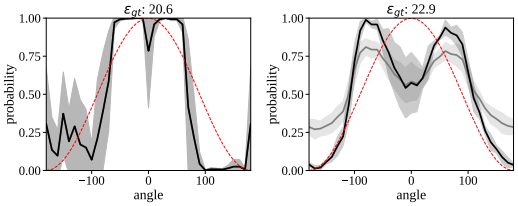
Table 10: The full results for the cross-lingual and cross-cultural evaluation of the preferred frame of reference in VLMs.

different spatial relations—for example, GPT-4 shows minimal preferences for the egocentric relative FoR along the sagittal axis but a significant preference along the lateral one (Table 6). As a result, VLMs demonstrate unsatisfactory consistency in their spatial performance (Table 2). Future work is necessary to improve the consistency and robustness of spatial representations in these models.

Our research also aligns with the growing trend of grounding linguistic analysis in rich modalities representing the real world [9, 62]. Language is not text in isolation; its meaning is significantly enriched when grounded in real-world contexts. For example, the ambiguity of spatial terms, as discussed in this paper, becomes meaningful only when combined with FoRs, and these FoRs are much more intuitively illustrated when visual cues are available. We advocate for future work that extends linguistic analysis to more grounded settings.

B.2 Perspective taking as a prerequisite of human-like spatial reasoning

Most languages support multiple FoRs.[5] The ability to understand and reason about space from a non-egocentric perspective is an important foundation of the Theory of Mind, a basic building block of our situated communication skill that allows us to infer others' mental states [43]. One of our key findings is that VLMs still struggle to adopt alternative FoRs flexibly, even when provided with explicit perspective-taking instructions (Section 3.1). We hypothesize that this phenomenon may come from a reporting bias in the image-text datasets available on the internet—it is natural to take the reflected relative FoR to view images presented on a screen, but this does not always apply in real-world applications. To address this issue, we suggest future work extend the current 2D VLMs to the 3D domain, by considering camera poses and multiview data [70] for training.



(a) `Behind` in GPT-4o. (b) `Right` in LLaVA-13B.

Figure 8: At $\theta = 0$, some models show sensitivity to multiple conventions.

B.3 Cross-cultural conventions in cross-lingual spatial understanding

The conventions for resolving spatial ambiguities are not uniform, as individuals from diverse linguistic and cultural backgrounds select their FoR differently. Cultural conventions can even be transmitted as individuals are exposed to other languages. Interestingly, Bohnemeyer et al. [5] found that among native speakers of Indigenous languages (with various preferences in FoRs), those more proficient in Spanish tend to use the reflected relative FoR (Spanish convention) more in their native language. This phenomenon has led to their Linguistic Transmission Hypothesis: "Using any language or linguistic variety – independently of its structures – may facilitate the acquisition of cultural practices of nonlinguistic cognition shared among the speakers of the language." Analogously, our experiment raises important concerns that English may dominate the FoR preference conventions of other languages in multilingual VLMs. This is not surprising, as current training recipes for multilingual multimodal language models heavily rely on machine-translated captions [12, 24]. However, this practice can be problematic. For instance, Hausa prefers an interpretation where the "front" aligns with the English concept of "back," [28], where this approach may lead to English conventions overshadowing those of other languages. At a high level, this issue is not limited to spatial reasoning—for example, Shi et al. [61] have demonstrated that English is always the best chain-of-thought language for math reasoning with multilingual LLMs, no matter what language is used for the problem description. To enable similar linguistic transmission in AI models, exposure to naturally generated multilingual data is crucial [59].

B.4 Limitations and Future Work

The acceptance regions. Using cosine and hemisphere as acceptance regions is analytical but might not capture some human cognitive biases. In reality, regional angles might not be uniformly distributed per relation, nor are they exactly 90 degrees. These angles vary across individuals and cultures [23].

Spatial relations. This work primarily focuses on the most basic types of spatial relations (front-back and left-right). However, many other relations exist, such as *away from* and *near* [41, 38]. Additionally, not all languages possess terms for "left," "right," "front," and "back." Some languages, like Guugu Yimithirr, use only absolute frames of reference instead [35].

Camera angle and occlusion. Currently, there is minimal occlusion, and the camera angle is high. Languages may differ in whether the speaker emphasizes these factors, such as the preference to use "behind" in cases of occlusion [35].

Pragmatic aspect of spatial cognition. Many conversational and pragmatic aspects of spatial cognition are simplified in this work, such as F-Formation [32] and human-robot interaction [37]. For example, in human-robot interaction settings, users prefer an addressee-centered frame of reference to facilitate the robot's comprehension of referents [46].

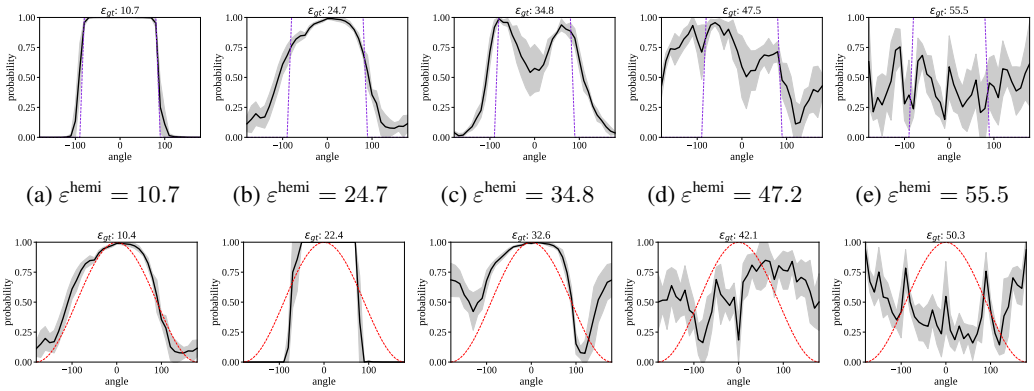
[5] Some languages, in very rare cases, have only one available spatial frame of reference. For example, Guugu Yimithirr exclusively uses the absolute FoR [35].

Multilingual prompts. In this work, we used machine-translated text to construct the multilingual portion of the dataset. Although we verified data quality through back translation, incorporating human annotations in the future would be a valuable future step.

B.5 List of Evaluated Languages

We started with 132 candidate languages supported by Google Translate API.[6] We removed 23 languages from our multilingual evaluation due to their failure to adhere to instructions for generating "yes" and "no" predictions, or because they did not pass the back-translation test for quality control: Aymara, Bambara, Croatian, Dhivehi, Dogri, Ewe, Guarani, Hmong, Kyrgyz, Luganda, Malayalam, Meiteilon (Manipuri), Mizo, Odia (Oriya), Punjabi, Quechua, Samoan, Tatar, Telugu, Tigrinya, Uyghur, Xhosa, Yoruba.

B.6 Visualizations of Region Parsing Error



(a) $\varepsilon^{\text{hemi}} = 10.7$ (b) $\varepsilon^{\text{hemi}} = 24.7$ (c) $\varepsilon^{\text{hemi}} = 34.8$ (d) $\varepsilon^{\text{hemi}} = 47.2$ (e) $\varepsilon^{\text{hemi}} = 55.5$

(f) $\varepsilon^{\text{cos}} = 10.4$ (g) $\varepsilon^{\text{cos}} = 22.4$ (h) $\varepsilon^{\text{cos}} = 32.6$ (i) $\varepsilon^{\text{cos}} = 42.1$ (j) $\varepsilon^{\text{cos}} = 50.3$

Figure 9: ε visualization: (a-e) correspond to $\varepsilon^{\text{hemi}}$, and (f-j) correspond to ε^{cos}.

B.7 Computing Resources

English evaluations require one NVIDIA A40 GPU (48GB). Multilingual evaluations require OpenAI and Google Cloud APIs and it can run on CPU with stable internet connection.

[6] https://cloud.google.com/translate

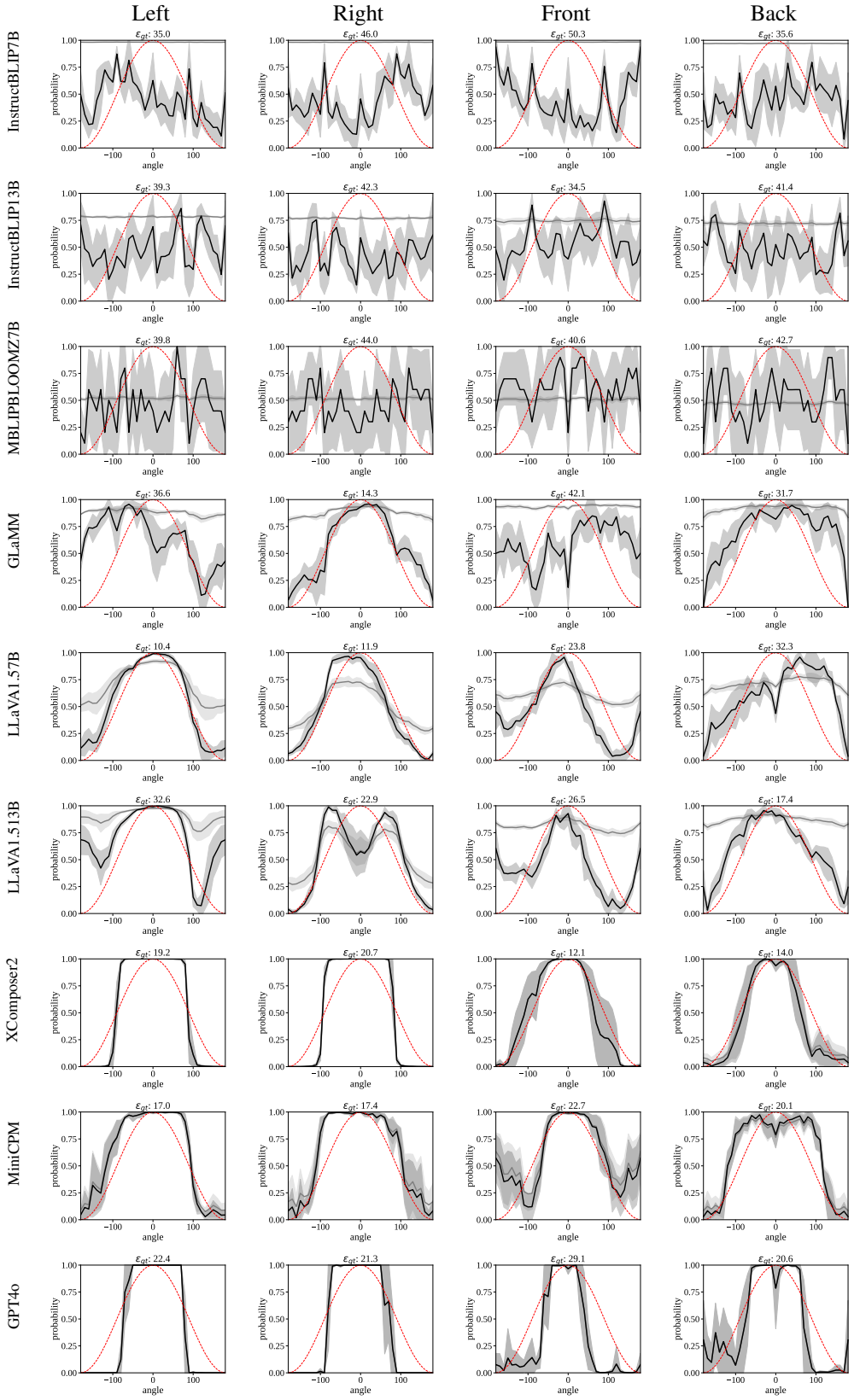


Figure 10: All prediction plots for each model on COMFORT-BALL using the camera perspective prompt (cam). The raw probability $p(\theta)$ in gray, normalized probability $\widehat{p}(\theta)$ in black, and the reference probability $p_{\cos}(\theta)$ of cam in red.

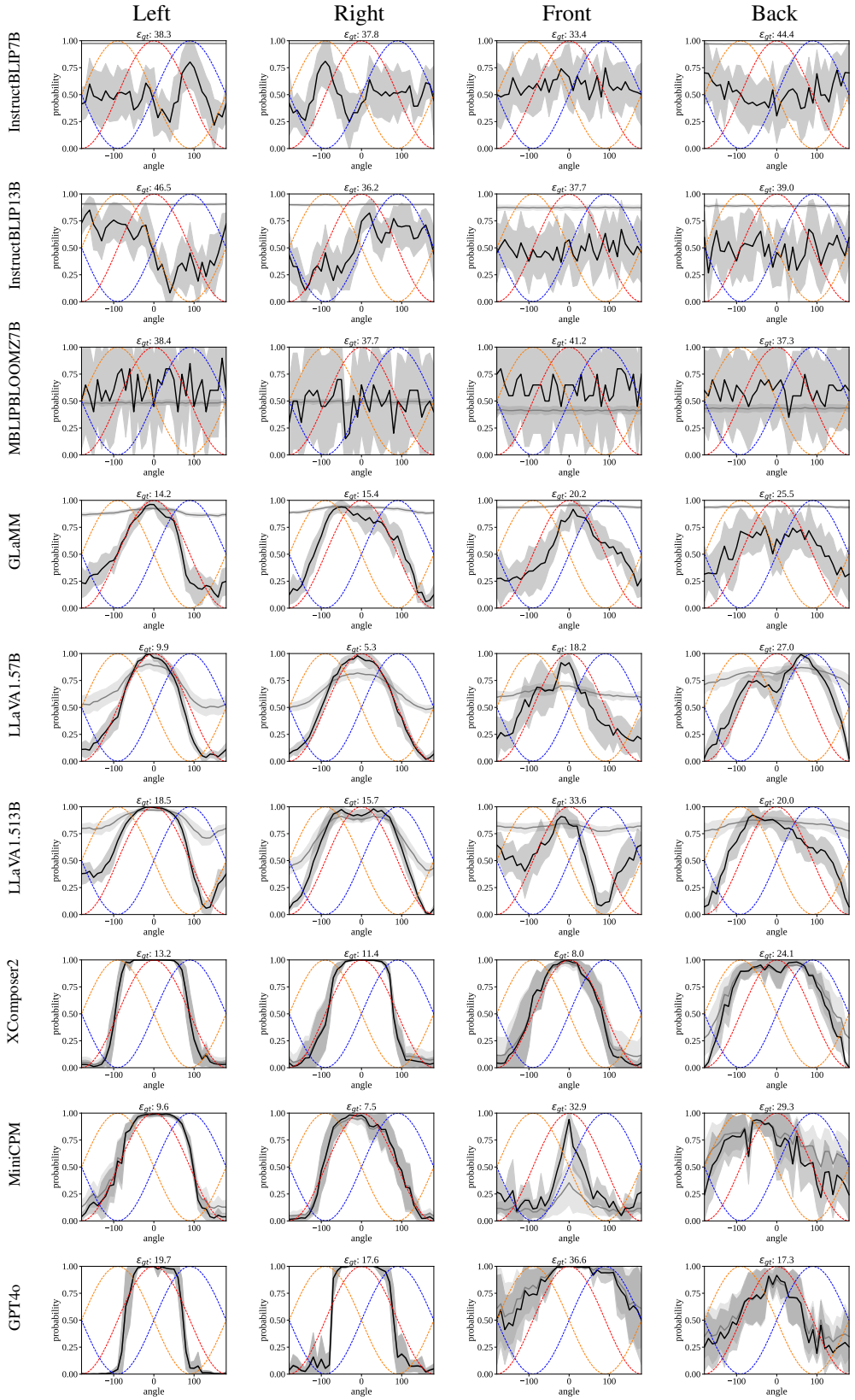


Figure 11: All prediction plots for each model on COMFORT-CAR using the camera perspective prompt (cam). The raw probability $p(\theta)$ in gray, normalized probability $\widehat{p}(\theta)$ in black, and the reference probabilities $p_{\cos}(\theta)$ of cam in red, add in orange, rel in blue. To avoid overlapping reference probabilities of add and rel, we use plots on COMFORT-CAR with relatum facing left for left and right relations and COMFORT-CAR with relatum facing right for front and behind relations.

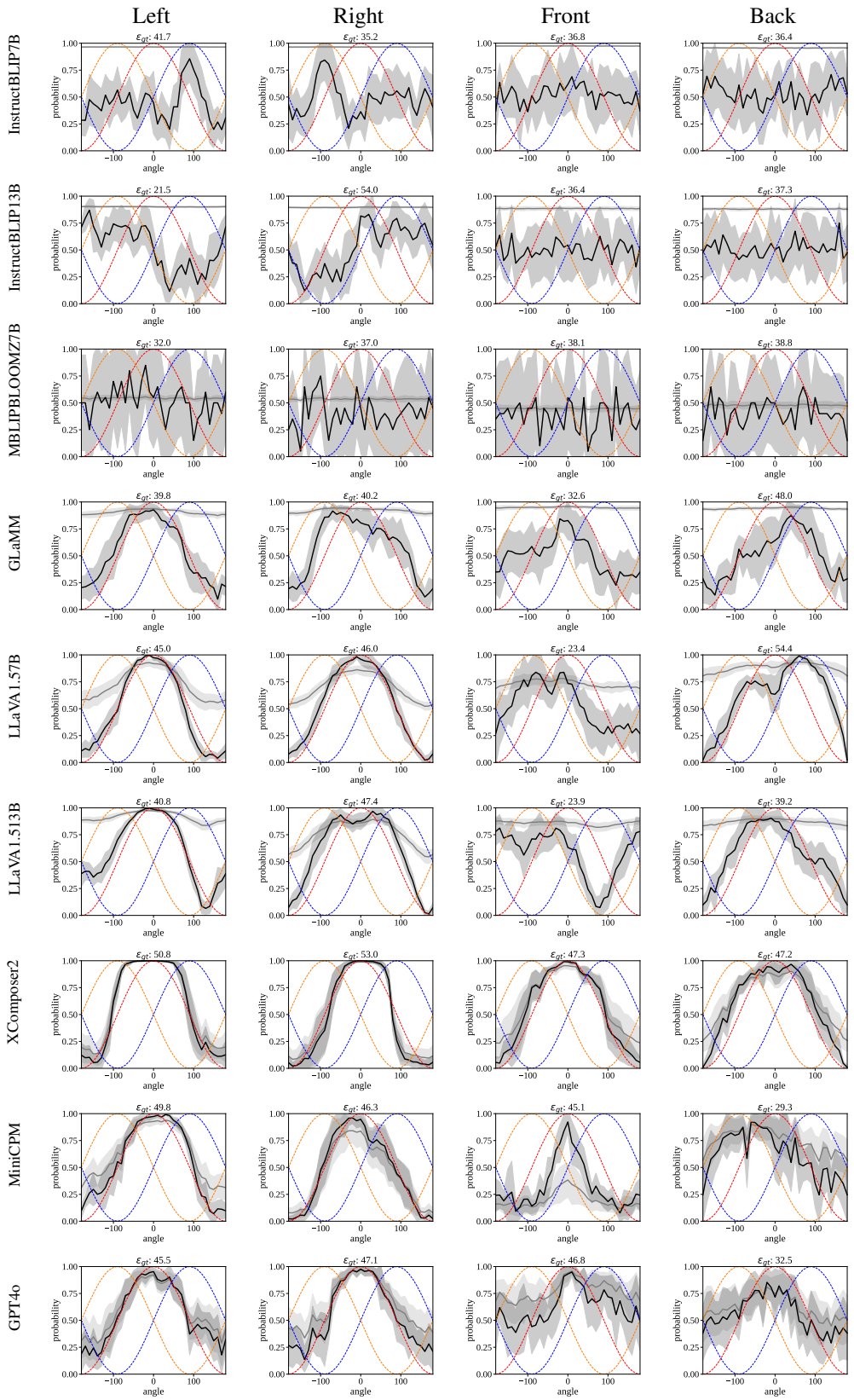


Figure 12: All prediction plots for each model on COMFORT-CAR using the addressee perspective prompt (add). The raw probability $p(\theta)$ in gray, normalized probability $\widehat{p}(\theta)$ in black, and the reference probabilities $p_{\cos}(\theta)$ of cam in red, add in orange, rel in blue. To avoid overlapping reference probabilities of add and rel, we use plots on COMFORT-CAR with relatum facing left for left and right relations and COMFORT-CAR with relatum facing right for front and behind relations.

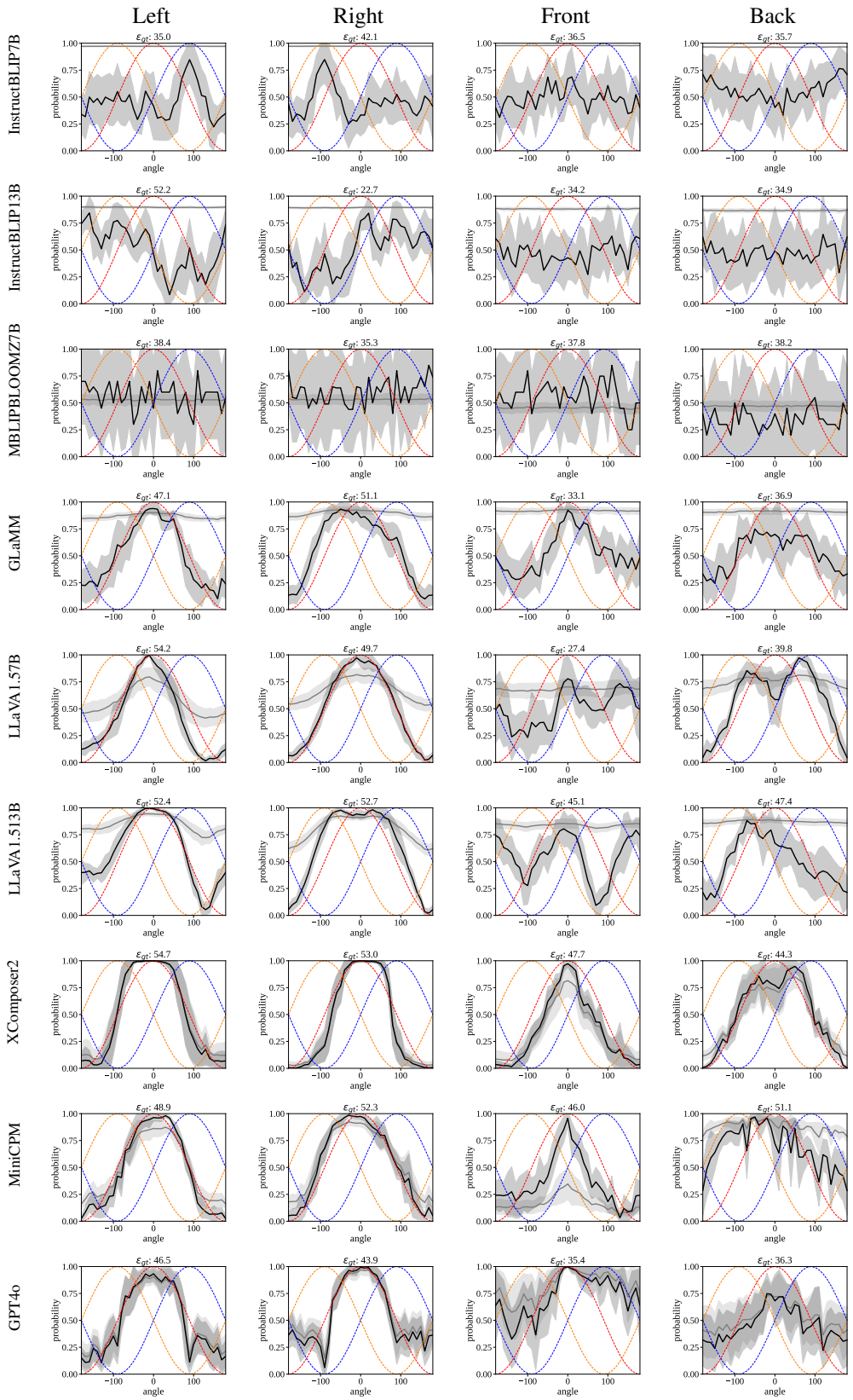


Figure 13: All prediction plots for each model on COMFORT-CAR using the relatum perspective prompt (rel). The raw probability $p(\theta)$ in gray, normalized probability $\widehat{p}(\theta)$ in black, and the reference probabilities $p_{\cos}(\theta)$ of cam in red, add in orange, rel in blue. To avoid overlapping reference probabilities of add and rel, we use plots on COMFORT-CAR with relatum facing left for left and right relations and COMFORT-CAR with relatum facing right for front and behind relations.

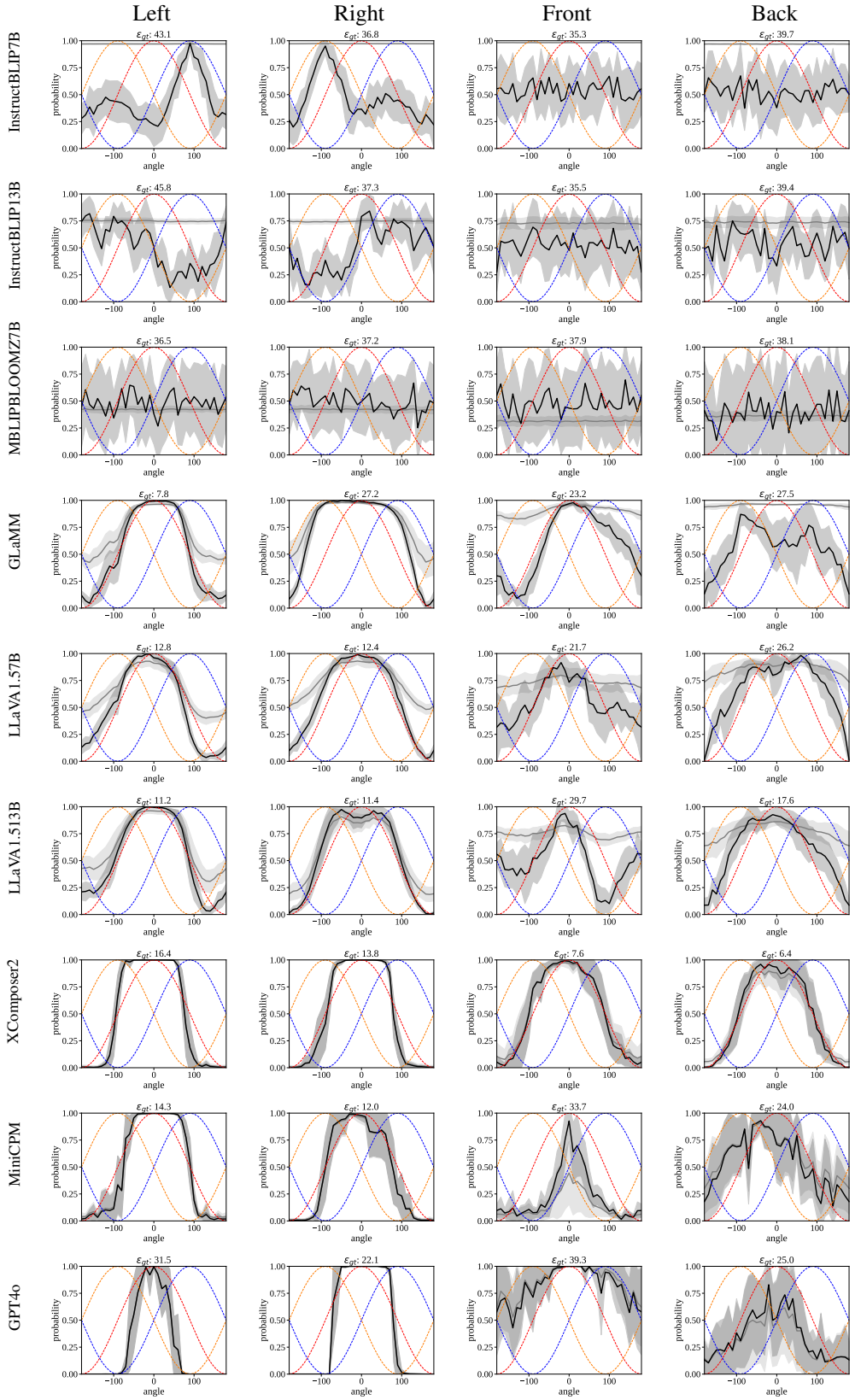


Figure 14: All prediction plots for each model on COMFORT-CAR without perspective prompt (nop). The raw probability $p(\theta)$ in gray, normalized probability $\widehat{p}(\theta)$ in black, and the reference probabilities $p_{\cos}(\theta)$ of cam in red, add in orange, rel in blue. To avoid overlapping reference probabilities of add and rel, we use plots on COMFORT-CAR with relatum facing left for left and right relations and COMFORT-CAR with relatum facing right for front and behind relations.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The claims made in the paper are supported by the experimental results.
   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: Due to the page limit, we include the limitations in the Appendix.
   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [NA]
   Justification: The paper does not include any theoretical result.
   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: Our paper has enough information to reproduce the results.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: The data and code will be released publicly upon acceptance.
   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: We have the necessary details for understanding the results.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment Statistical Significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [No]
   Justification: There's no training component in our paper and we evaluate logits from the language decoder in vision-language models (VLMs) without sampling so they are deterministic.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.
8. **Experiments Compute Resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: We include this in Appendix.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: The research was conducted in full compliance with the NeurIPS Code of Ethics.
   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [Yes]
    Justification: Yes, our paper has positive societal impacts as we value pluralistic alignment towards multilingualism.
    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
    Answer: [No]
    Justification: Our paper does not train new models.
    Guidelines:
    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**
    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
    Answer: [Yes]
    Justification: The existing assets used in this paper are properly credited and respected.
    Guidelines:
    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**
    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
    Answer: [Yes]
    Justification: The new dataset introduced in the paper is well documented.
    Guidelines:
    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
    Answer: [NA]
    Justification: The paper does not involve crowdsourcing experiments and research with human subjects.
    Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**
    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]
Justification: The paper does not involve crowdsourcing experiments and research with human subjects.
Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.