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

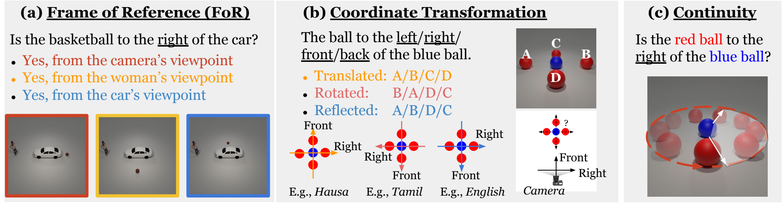
Zheyuan Zhang*[1], Fengyuan Hu*[1], Jayjun Lee*[1]
Freda Shi[2,3], Parisa Kordjamshidi[4], Joyce Chai[1], Ziqiao Ma[1]

ICLR International Conference On Learning Representations

Motivation

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 consistent across languages or tasks;
- Humans demonstrate flexibility in using multiple coordinate systems for both non-linguistic reasoning and linguistic expressions;
- Do vision-language models represent space, and how? Alignment with humans' resolution of spatial ambiguities is largely under-explored.



Background

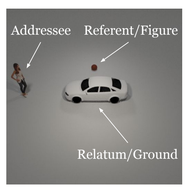
People may use different *frames of reference* (FoR; [1-3], *inter alia*) to resolve ambiguity about the underlying coordinate system.

- The intrinsic FoR aligns the origin with the relatum, and describes the referent's position relative to the relatum's inherent orientation;
- The relative FoR positions a *viewer* (egocentric or addressee) as the origin and focuses on the observer's intrinsic perspective.

Ambiguities in relative frames of reference:

- *Translated*, where the coordinate frame of the speaker is directly applied;
- *Rotated*, where the coordinate frame of the speaker is transformed with a 180-degree rotation;
- *Reflected*, where only the front-back axis is reversed.

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.



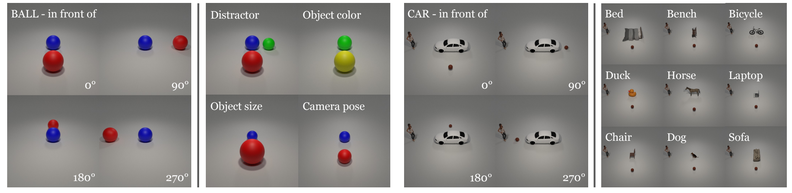
References

[1] Space in language and cognition: Explorations in cognitive diversity. *Stephen C. Levinson*. Cambridge University Press, 2003.
[2] Ambiguities in spatial language understanding in situated human robot dialogue. *Changsong Liu, Jacob Walker, Joyce Y Chai*. AAAI Fall Symposium, 2010.
[3] Frames of reference in spatial language acquisition. *Anna Shusterman, Peggy Li*. Cognitive psychology, 2016.
[4] The cultural transmission of spatial cognition: Evidence from a large-scale study. *Jürgen Bohnemeyer, et al.* CogSci, 2014.

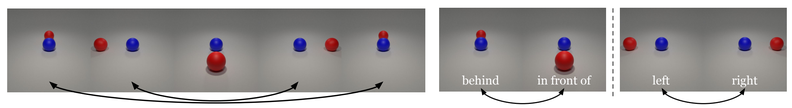
Takeaways (TL;DR)

- Do vision-language models represent space, and how? **VLMs aligns with English conventions.**
- Are VLMs robust and consistent? **No!**
- Can VLMs take alternative perspectives? **No!**
- Can multilingual VLMs adhere to cultural conventions? **No, English tends to dominate other languages.**

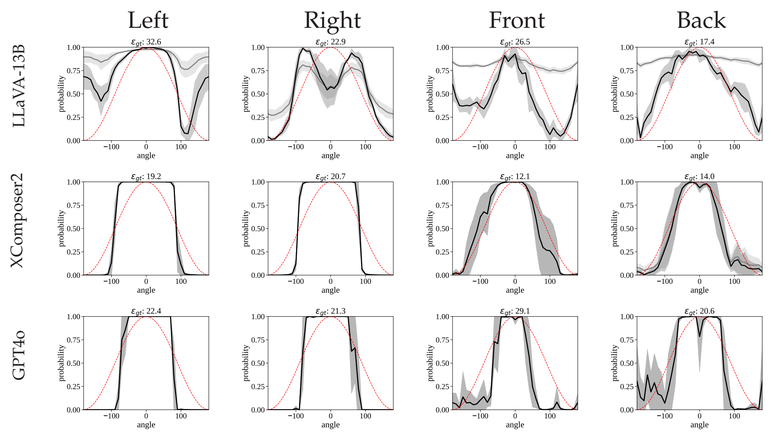
Dataset and Metrics



- **Accuracy.** 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.** We normalize p_i across all image-prompt pairs, and compute the RMSE against the reference probability threshold (defined by hemispheres and cosine of angles) that represents the actual regions of acceptability.
- **Robustness.** Standard deviation and prediction noise.
- **Consistency.** Spatial symmetry and opposition.



Visualization of Region Parsing Error



Prefer Transformation Convention

Most VLMs prefer *reflected* transformation similar to English.

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.	
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.**
GLaMM	58.3	**33.3**	43.9	**42.9**	**38.3**	51.8	**17.3**	63.7	39.5	47.9	**33.0**	**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.**

Prefer Frame of Reference

Most VLMs prefer the *egocentric relative* FoR similar to English.

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.	
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.**
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.**
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.**

Failure in Perspective Taking

While VLMs can comprehend scenes using egocentric relative FoR, they struggle to adapt flexibly to alternative FoRs.

Model	Egocentric		Intrinsic		Addressee		Aggregated	
	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$ (↓)	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$ (↓)	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$ (↓)	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$ (↓)
LLaVA-1.5-13B	$51.6_{(-15.0)}$	$23.9_{(+3.1)}$	$47.3_{(+0.8)}$	$45.0_{(-0.7)}$	$47.5_{(-3.8)}$	$\mathbf{38.9}_{(-4.2)}$	$48.8_{(-6.0)}$	$35.9_{(-0.6)}$
GLaMM	$47.2_{(-10.6)}$	$23.3_{(-0.7)}$	$47.2_{(+0.8)}$	$\mathbf{44.2}_{(-6.9)}$	$47.2_{(-2.8)}$	$42.8_{(-6.1)}$	$47.2_{(-4.2)}$	$36.8_{(-4.6)}$
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	$78.3_{(+4.6)}$	$28.1_{(-7.0)}$	$\mathbf{53.4}_{(-1.9)}$	$44.6_{(-6.3)}$	$49.1_{(-5.7)}$	$44.9_{(-6.4)}$	$60.3_{(-1.0)}$	$39.2_{(-6.6)}$

Lack of Robustness and Consistency

Despite decent performance in accuracy, VLMs demonstrate a lack of robustness and consistency.

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
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
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
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	

Cross-lingual Evaluation

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.