The Perceptual Observatory

Characterizing Robustness and Grounding in MLLMs

Tejas Anvekar*, Fenil Bardoliya*, Pavan Turaga, Chitta Baral, Vivek Gupta

Project Code

Arizona State University

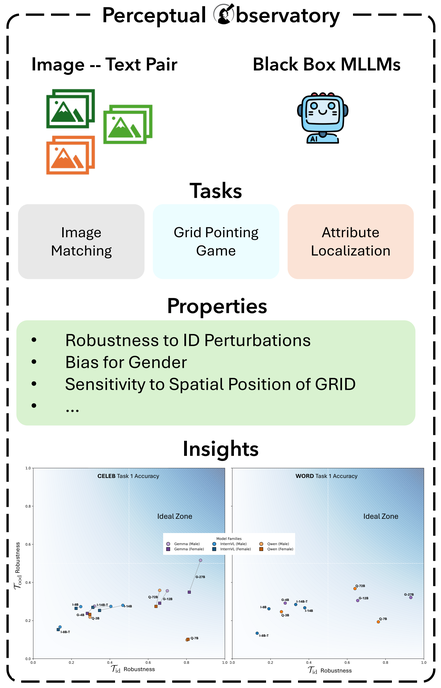
WACV 2026 TUCSON, AZ 3/6 - 3/10

Introduction

TL;DR;

Problem: Current MLLM benchmarks mostly measure end-task accuracy, so we can't tell whether gains come from **true visual grounding** or from **text/world-knowledge shortcuts**, especially when model families scale the language side while reusing similar vision encoders.

Solution: The **Perceptual Observatory** evaluates "how MLLMs see" using controlled tasks (image matching, grid pointing, attribute localization) on **faces and text-in-vision**, under **pixel corruptions** and **diffusion-based stylized illusions**, with metrics for robustness, spatial invariance, attribution fidelity, and bias.

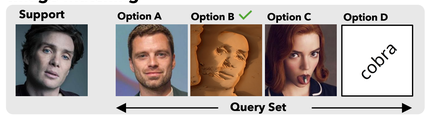


Task Modeling

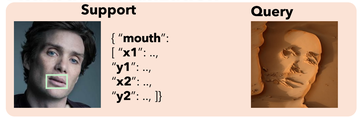
To **Assess, MLLMs** beyond **end-task** performance we model the tasks for**:**

- **Increasing cognitive load:** match → point → localize/transfer.
- **Spatial:** grid position invariance + attribute grounding.
- **Shift: ID** corruptions vs **OOD** stylized illusions (same layout, new style).

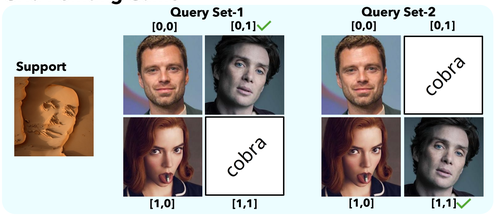
Image Matching



Attribute Localization



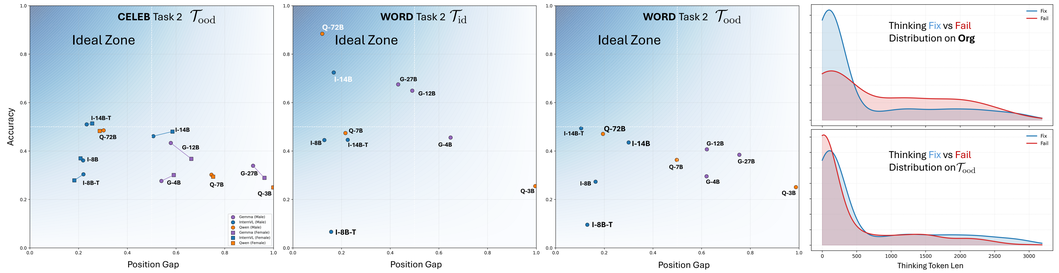
Grid Pointing Game



Benchmark Results

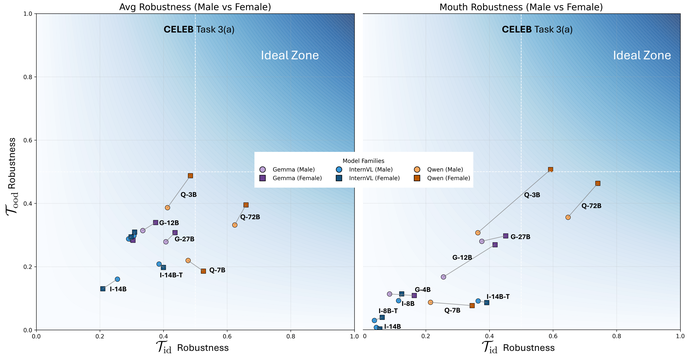
#	CELEB Task1 Org	CELEB Task1 $\mathcal{T}_{id}$	CELEB Task1 $\mathcal{T}_{ood}$	CELEB Task2 Org	CELEB Task2 $\mathcal{T}_{id}$	CELEB Task2 $\mathcal{T}_{ood}$	CELEB Task3 Org	CELEB Task3 $\mathcal{T}_{id}$	CELEB Task3 $\mathcal{T}_{ood}$	WORD Task1 Org	WORD Task1 $\mathcal{T}_{id}$	WORD Task1 $\mathcal{T}_{ood}$	WORD Task2 Org	WORD Task2 $\mathcal{T}_{id}$	WORD Task2 $\mathcal{T}_{ood}$	WORD Task3 Org	WORD Task3 $\mathcal{T}_{id}$	WORD Task3 $\mathcal{T}_{ood}$
Qwen 2.5 - VL																		
3B	33.66	29.57	22.57	25.00	24.95	24.98	90.57	90.23	**85.33**	21.00	25.66	24.66	25.75	25.53	25.08	97.54	97.48	**95.35**
7B	**78.21**	**80.52**	10.00	64.75	65.66	29.81	99.65	99.17	16.52	**75.00**	**76.26**	19.33	43.75	47.43	36.33	**99.99**	**99.95**	54.39
Δ	44.55	50.95	-12.57	39.75	40.71	04.83	09.08	08.94	-68.81	54.00	50.60	-05.33	18.00	21.90	11.25	02.45	02.47	-40.96
72B	51.48	65.10	**31.61**	**98.25**	**98.66**	**48.40**	**99.99**	**99.93**	39.65	56.00	64.06	**36.80**	**87.75**	**88.40**	**47.06**	98.64	97.97	50.55
Δ	17.82	35.53	09.04	73.25	73.71	23.42	09.42	09.70	-45.68	35.00	38.40	12.14	62.00	62.87	21.98	01.10	00.49	-44.80
Gemma 3																		
4B	24.75	30.09	25.67	45.00	47.10	28.85	**99.83**	**99.72**	**61.92**	23.00	27.66	29.20	46.25	45.56	29.56	99.99	99.99	**99.69**
12B	55.44	67.98	32.34	**84.75**	**84.44**	**40.03**	99.54	98.47	51.86	66.00	65.46	30.53	64.50	64.93	**40.68**	99.99	99.99	99.28
Δ	30.69	37.89	06.67	39.75	37.34	11.18	-00.29	-01.25	-10.06	43.00	37.80	01.33	18.25	19.37	11.12	00.00	00.00	-00.41
27B	**78.21**	**84.35**	**43.16**	71.50	68.63	31.40	99.58	97.38	25.86	**96.00**	**93.20**	**32.13**	**69.00**	**67.50**	38.46	**99.99**	**99.99**	80.27
Δ	53.46	54.26	17.49	26.50	21.53	02.55	-00.25	-02.34	-36.06	73.00	65.54	02.93	22.75	21.94	08.90	00.00	00.00	-19.42
InternVL 3.5																		
8B	18.81	23.56	26.86	81.75	79.71	36.56	**99.99**	**99.99**	**66.84**	13.00	19.11	26.20	47.00	44.53	27.31	**99.99**	**99.99**	**99.64**
– <T>	03.96	13.53	15.97	29.25	31.73	26.66	96.99	95.50	17.79	08.00	12.99	13.40	06.25	06.33	09.56	72.99	66.76	47.82
Δ	-14.85	-10.03	-10.89	-52.50	-47.98	-09.90	-03.00	-04.49	-49.05	-05.00	-06.12	-12.80	-40.75	-38.20	-17.75	-27.00	-33.23	-51.82
14B	27.72	**40.66**	26.66	**98.50**	**97.91**	47.08	49.41	53.39	52.34	**34.00**	**37.86**	26.73	**71.25**	**72.40**	43.55	98.61	98.12	95.42
– <T>	**27.72**	35.77	**27.19**	89.25	92.53	**51.18**	85.04	81.49	11.17	27.00	33.13	**28.46**	44.00	44.59	**49.41**	99.22	97.30	92.89
Δ	00.00	-04.89	00.53	-09.25	-05.38	04.10	35.63	28.10	-41.17	-07.00	-04.73	01.73	-27.25	-27.81	05.86	00.61	-00.82	-02.53
Human																		
	100	100	89.11	100	100	87.55	95.66	93.22	81.88	100	100	83.86	100	100	79.98	99.99	99.99	99.99

Insights and Findings



- **OOD stylized "illusions" break perception:** Under **OOD**, models drift out of the *Ideal Zone* accuracy drops while **position gap** rises, showing sensitivity to style shifts even when layout is preserved.
- **WORD ≈ saturated, CELEB is brittle: WORD** tasks stay near-ceiling in **ID** and degrade less than **CELEB**, where faces show sharp robustness and grounding failures under OOD.
- **Scaling ≠ robustness:** Larger LMs often improve clean / ID accuracy, but **robustness and spatial invariance are non-monotonic** bigger models don't reliably land in the Ideal Zone under OOD.
- **"Thinking" doesn't reliably help seeing:** <think> sometimes improves clean cases, but frequently **fails under OOD**; fixes concentrate at short chains, while long chains often still fail suggesting narration > perception.

These plots summarize **Task 3(a) semi-guided attribute localization on CELEB**: even with a hint box, most models remain far from the *Ideal Zone*, and performance drops sharply from **ID augmentations** to **OOD stylized illusions** ; the effect can worsen when you zoom into a specific attribute like **mouth**, where several models show larger errors and clearer **male-female gaps** (gender bias).



Conclusion: Current MLLMs do not reliably perceive grounded attribute structure under distribution shift, so better vision-language alignment is needed beyond scaling the language side.

Connect With
anvekar-tejas
anvekartejas@gmail.com

Connect With
fenil-bardoliya
fenildbardoliya@gmail.com