The Perceptual Observatory

Characterizing Robustness and Grounding in MLLMs

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

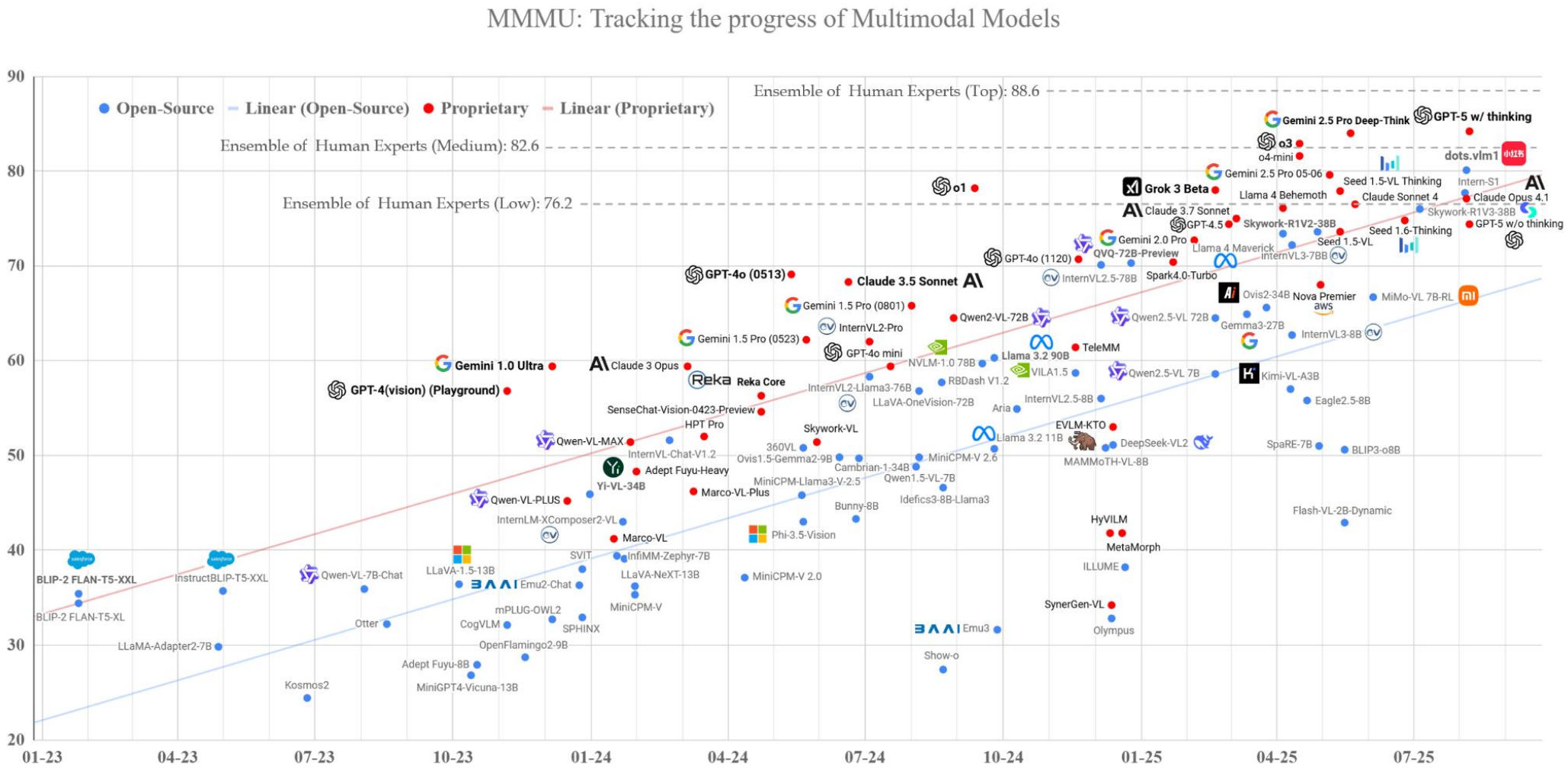
Arizona State University

Arizona State University

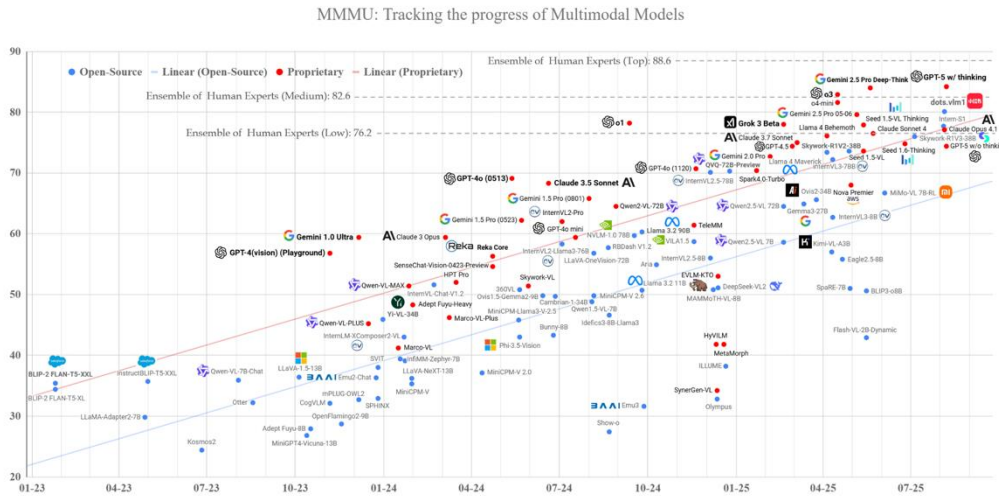
* Equal Contribution

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

Introduction

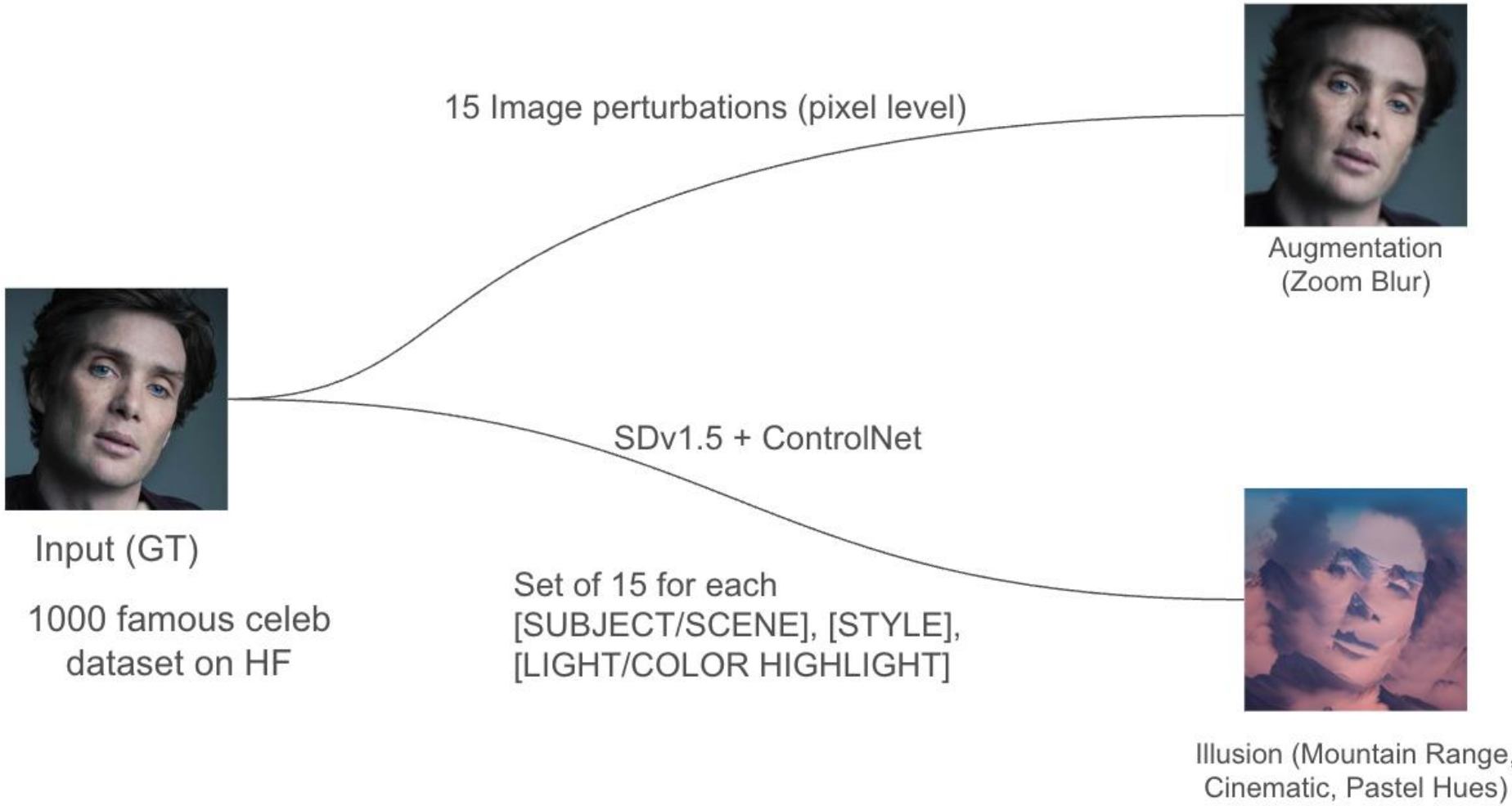


Introduction (NLP Folks do LM Scaling)

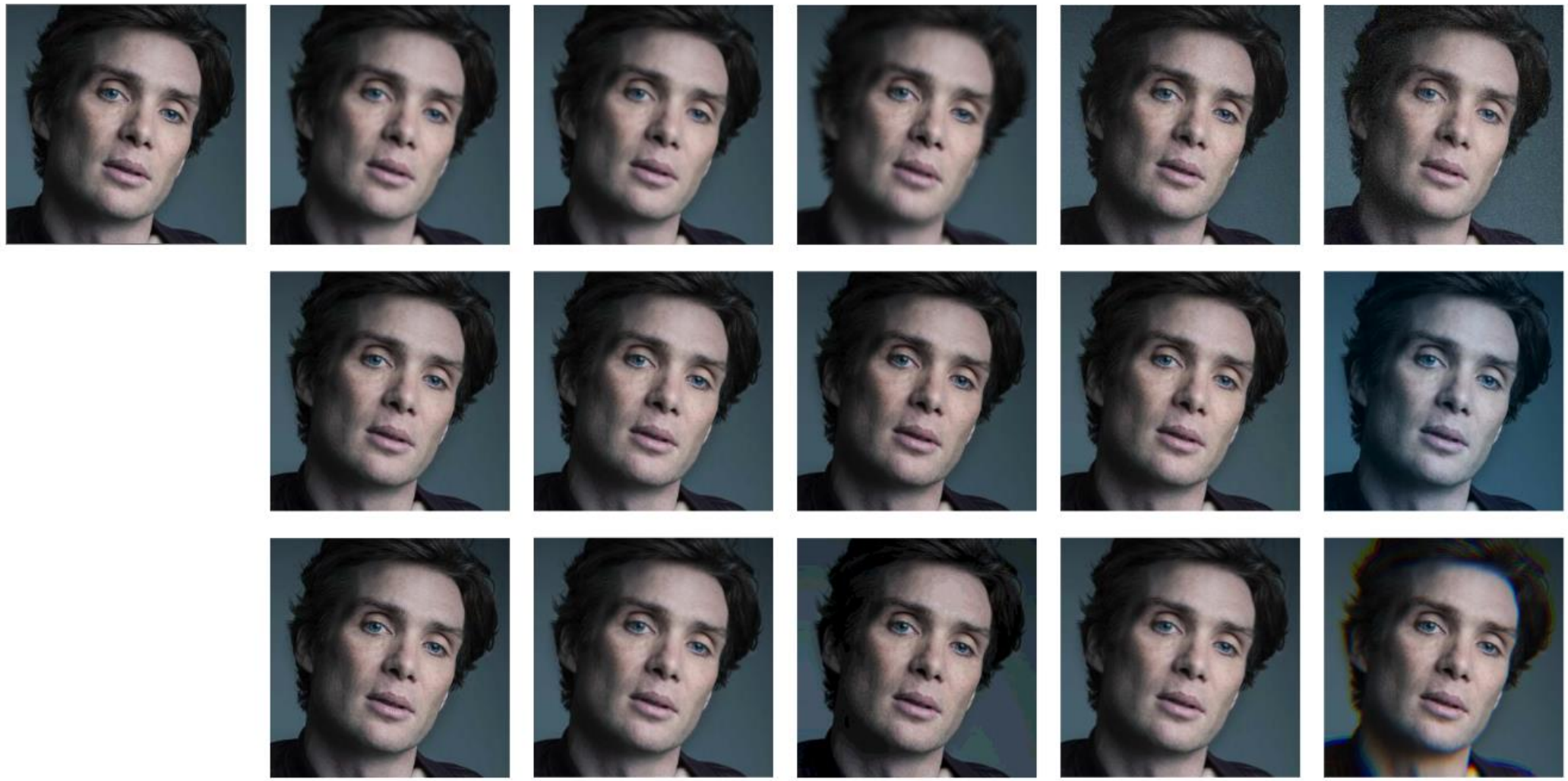


Configuration	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-72B
Vision Transformer (ViT)			
Hidden Size	1280	1280	1280
# Layers	32	32	32
# Num Heads	16	16	16
Intermediate Size	3456	3456	3456
Patch Size	14	14	14
Window Size	112	112	112
Full Attention Block Indexes	{7, 15, 23, 31}	{7, 15, 23, 31}	{7, 15, 23, 31}
Vision-Language Merger			
In Channel	1280	1280	1280
Out Channel	2048	3584	8192
Large Language Model (LLM)			
Hidden Size	2048	3,584	8192
# Layers	36	28	80
# KV Heads	2	4	8
Head Size	128	128	128
Intermediate Size	4864	18944	29568
Embedding Tying	✓	✗	✗
Vocabulary Size	151646	151646	151646
# Trained Tokens	4.1T	4.1T	4.1T

Data Generation



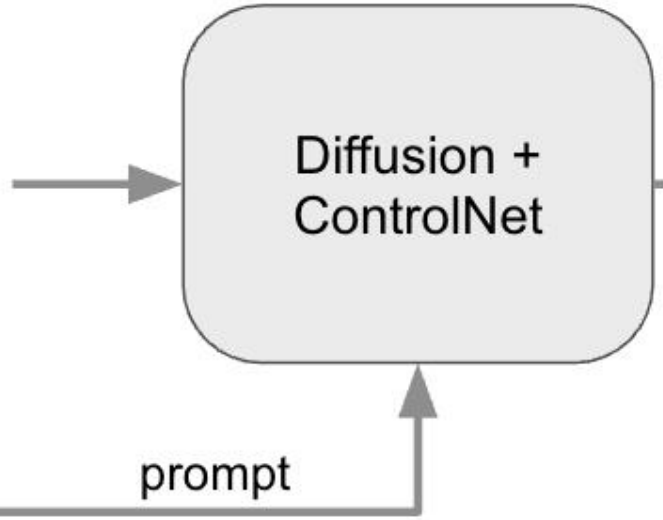
Data Generation (ID)



Data Generation (OOD)

N L P

"A photo realistic group of people on street"



Data Generation (OOD)

Motivation

Motivation

Motivation

Motivation

WACV

Task-1: Image Matching

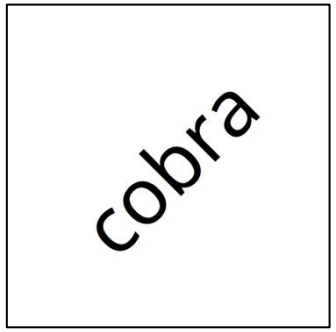
Support

Option A

Option B

Option C

Option D



Which images match from options match the query?

Task-1: Image Matching

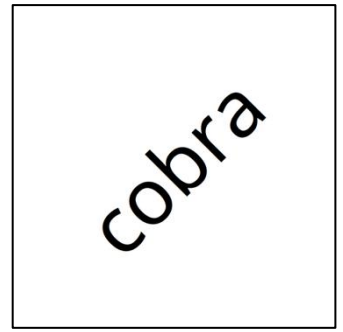
Support

Option A

Option B

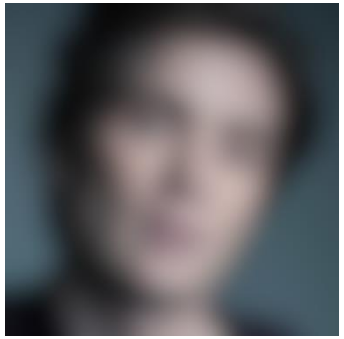
Option C

Option D



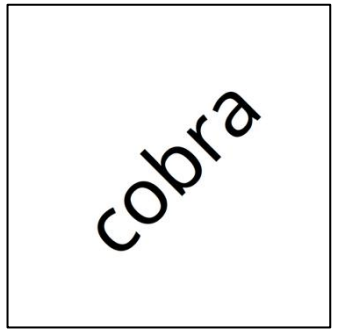
Support

Option A



Option B

Option C



Option D

We Assess All tasks for Different types of Perturbation

Task-1: Image Matching

Support

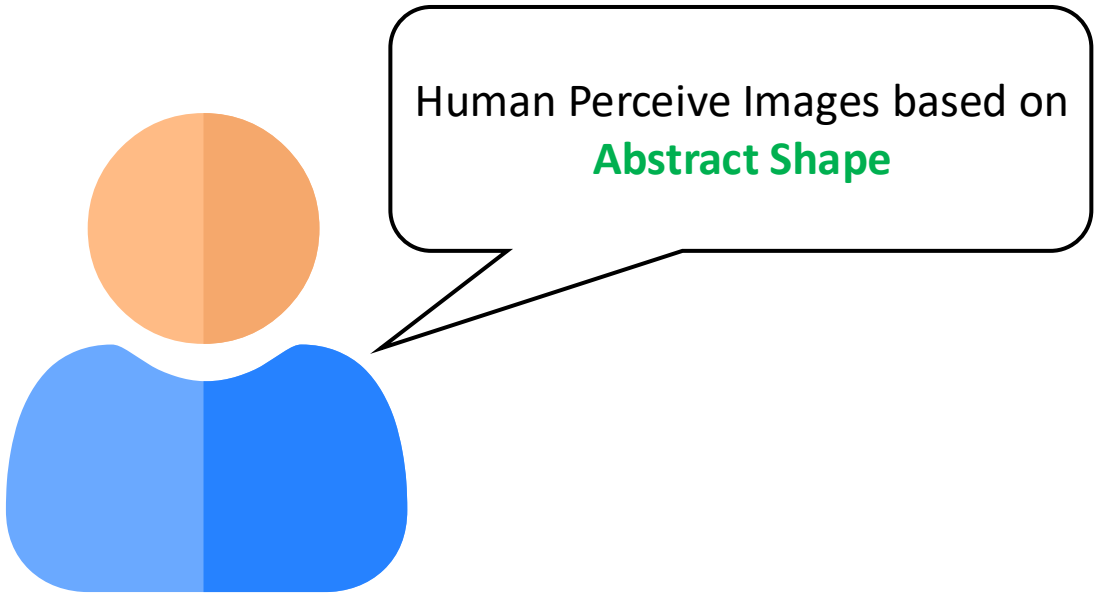
Option A

Option B

Option C

Option D

cobra



Task-1: Image Matching

Support

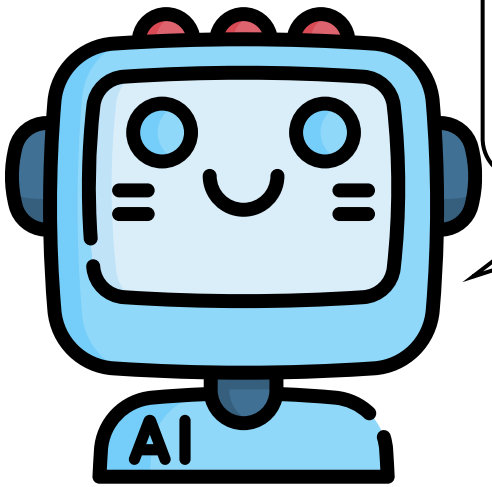
Option A

Option B

Option C

Option D

cobra



MLLMs Perceive Images based on **Semantics**

Task-2: Grid Pointing Game

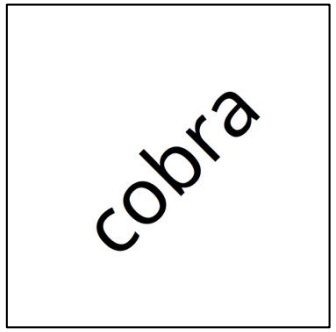
Support

Option A

Option B

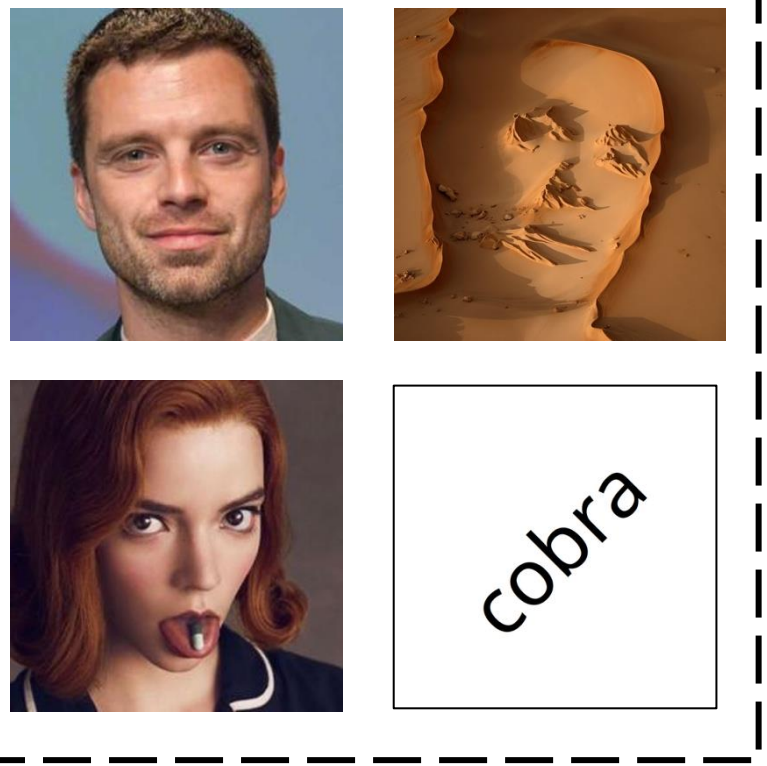
Option C

Option D



Task-2: Grid Pointing Game

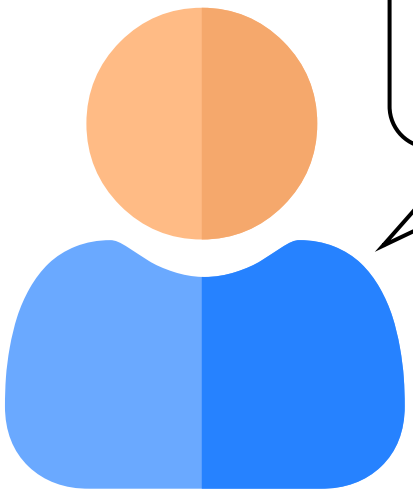
Support



Task-2: Grid Pointing Game

Support

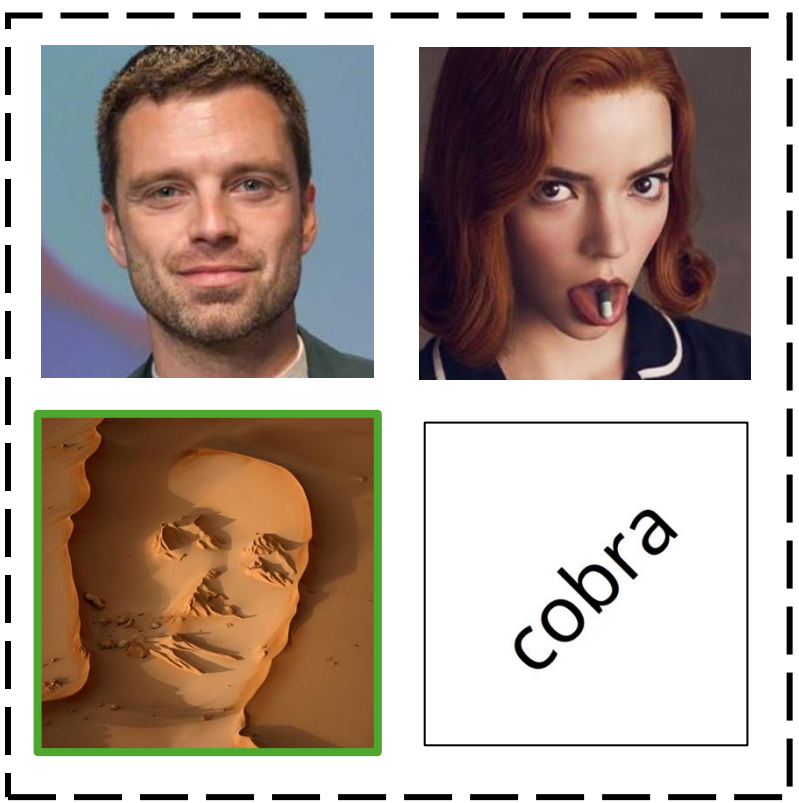
cobra

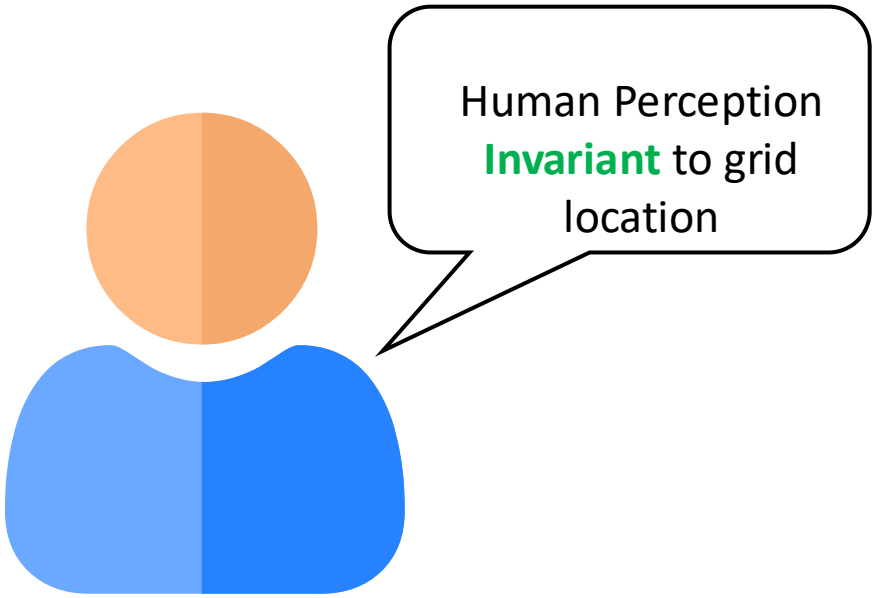


Human Perception **Invariant** to grid location

Task-2: Grid Pointing Game

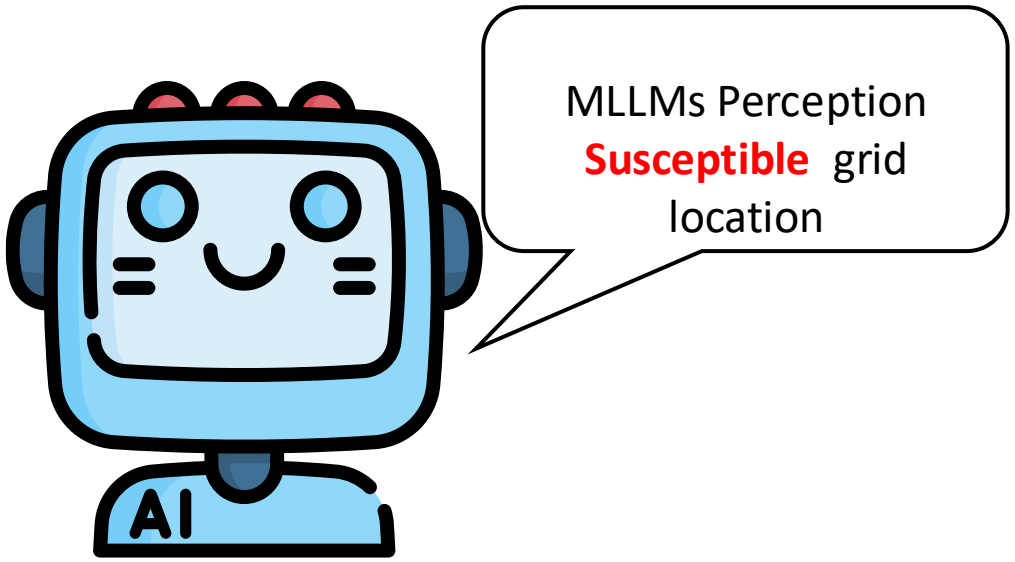
Support





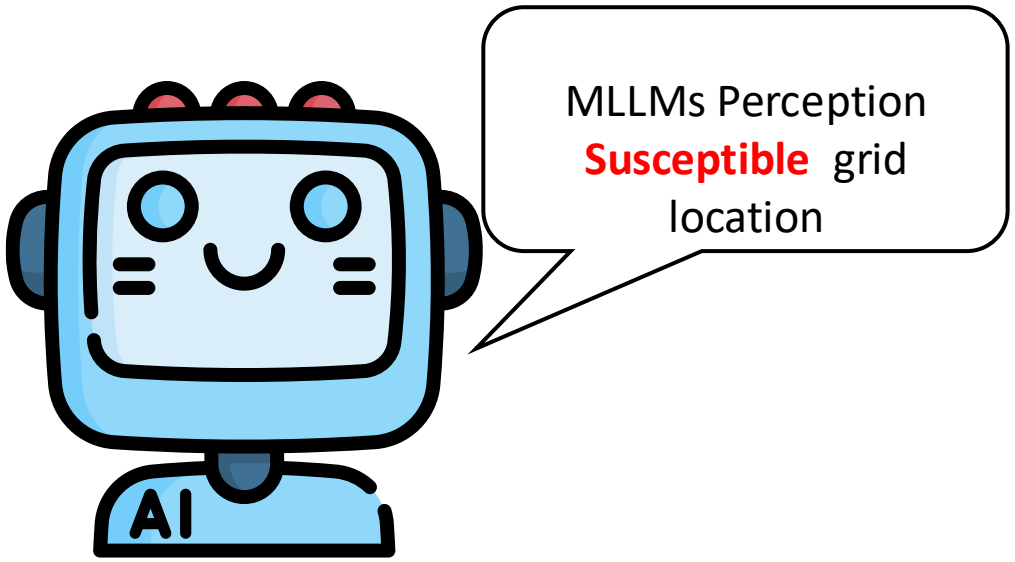
Task-2: Grid Pointing Game

Support



Task-2: Grid Pointing Game

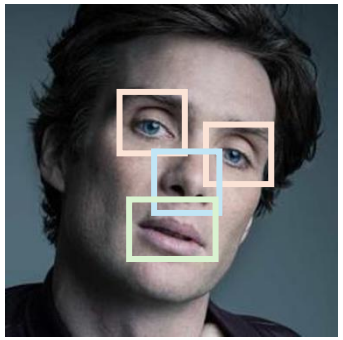
Support



Task-3: Attribute Localization

Full Guidance

Support

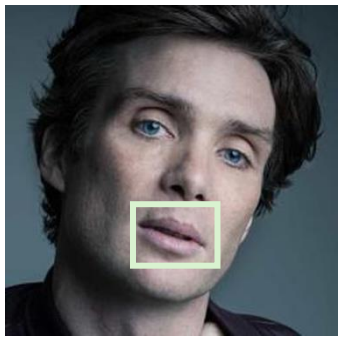


{ “**mouth**”:
[“**x1**”: ..,
“**y1**”: ..,
“**x2**”: ..,
“**y2**”: ..,
]}

Can Models Localize with Full Guidance?

Partial Guidance

Support



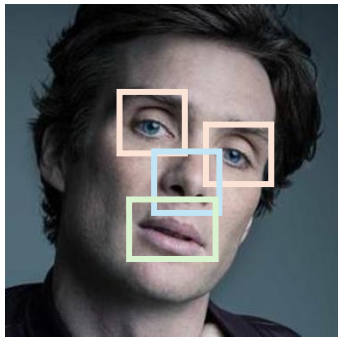
{ “**mouth**”:
[“**x1**”: ..,
“**y1**”: ..,
“**x2**”: ..,
“**y2**”: ..,
]}

Can Models Localize with Partial Guidance?

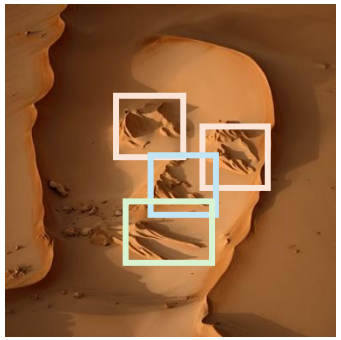
Task-3: Attribute Localization

Full Guidance

Support

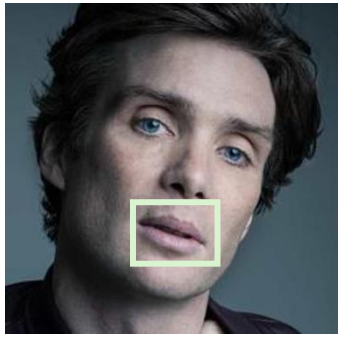


{ “**mouth**”:
[“**x1**”: ..,
“**y1**”: ..,
“**x2**”: ..,
“**y2**”: ..,
]}

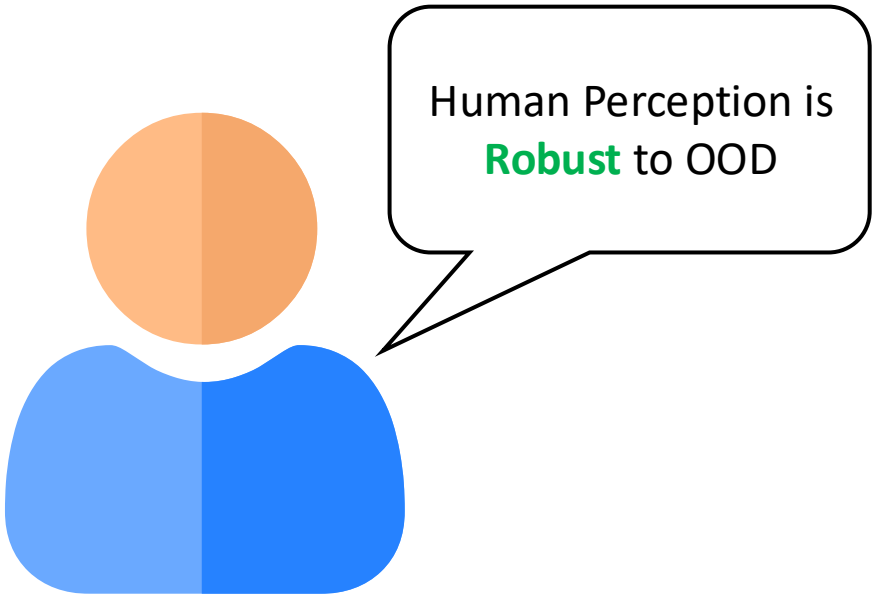


Partial Guidance

Support



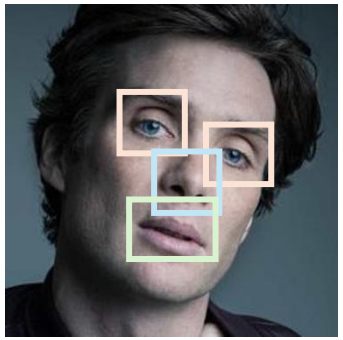
{ “**mouth**”:
[“**x1**”: ..,
“**y1**”: ..,
“**x2**”: ..,
“**y2**”: ..,
]}



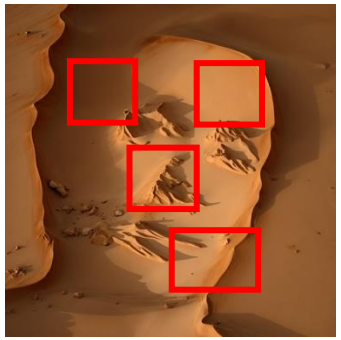
Task-3: Attribute Localization

Full Guidance

Support

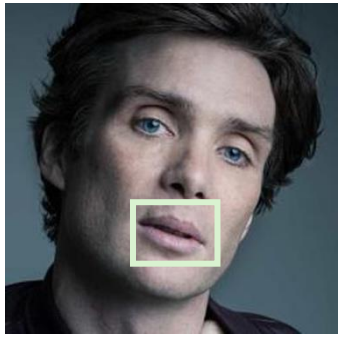


{ “**mouth**”:
[“**x1**”: ..,
“**y1**”: ..,
“**x2**”: ..,
“**y2**”: ..,
]}

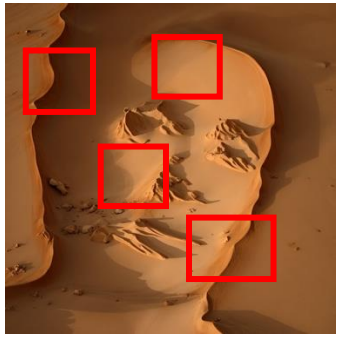


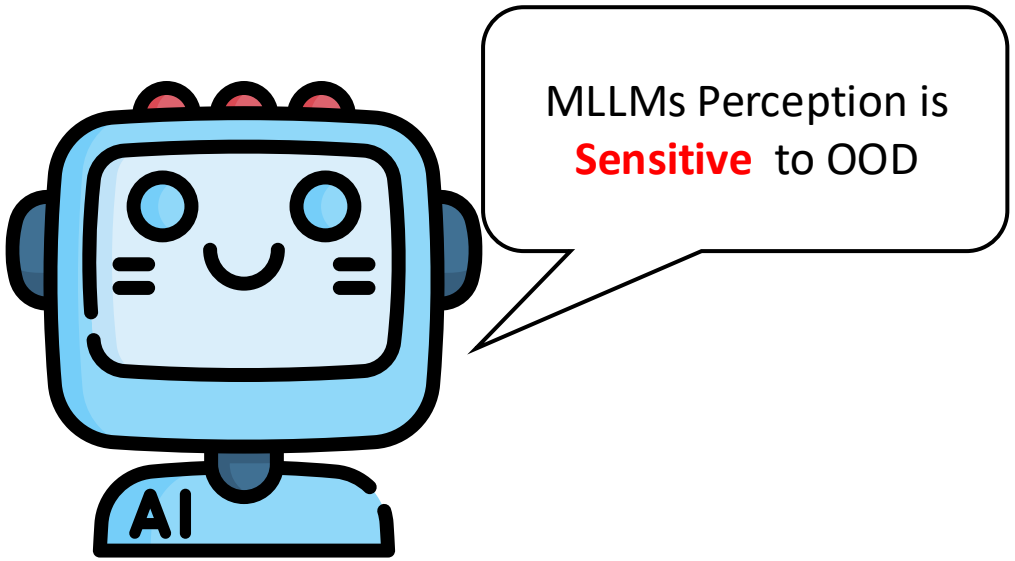
Partial Guidance

Support



{ “**mouth**”:
[“**x1**”: ..,
“**y1**”: ..,
“**x2**”: ..,
“**y2**”: ..,
]}





Our Findings – Main Results

Table 1. Table summarizes robustness of MLLMs for ID vs OOD across both dataset across all task, here Task3(b) is dubbed as Task 3. Δ refers to difference between current vs smallest among family ex: Qwen7B – Qwen3B, and also difference between thinking (<T>) vs non-thinking.

	CELEB									WORD								
	Task1			Task2			Task3			Task1			Task2			Task3		
#	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$	Org	$\mathcal{T}_{id}$	$\mathcal{T}_{ood}$
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

Stylized OOD perturbations **disproportionately harm mid-scale models (7–14B)**, and even large models (e.g., Qwen-72B, Gemma-27B) **don't reliably stay robust**.

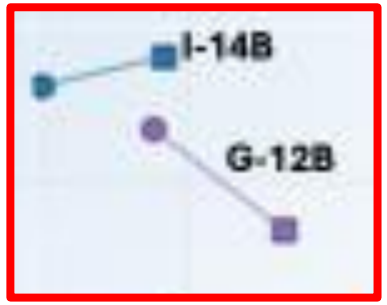
Robustness doesn't consistently improve with size (e.g., Gemma-12B can be **more brittle than Gemma-4B**).

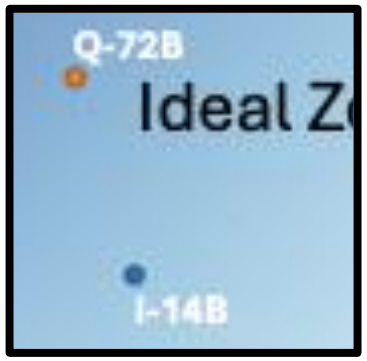
Improvements are largely driven by **LM-side capacity** (decoder/projector), since the **vision encoder is held fixed** suggesting "better narrative" rather than better seeing.

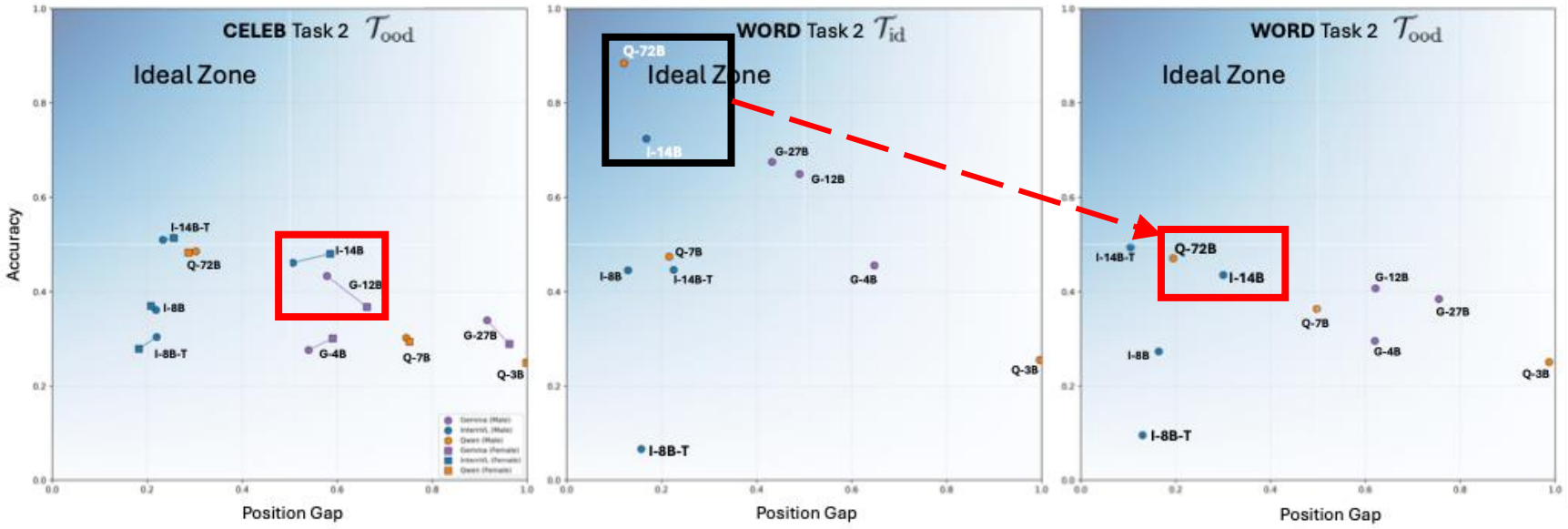
Decode-time reasoning can help clean/ID performance but **reduces transfer retention on faces**, and InternVL "thinking" variants can underperform non-thinking on these tasks.

Humans achieve **>95%** on identity/spatial tasks with only mild OOD degradation so the gap is model-driven, not dataset difficulty.

Our Findings – For Task 2 (Grid Pointing Game)

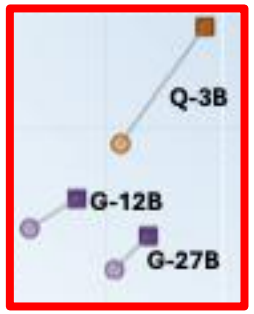


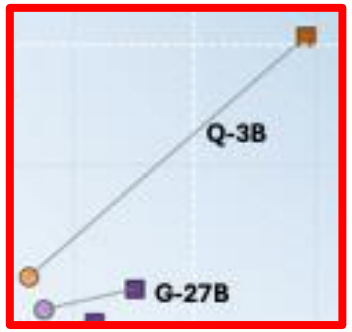


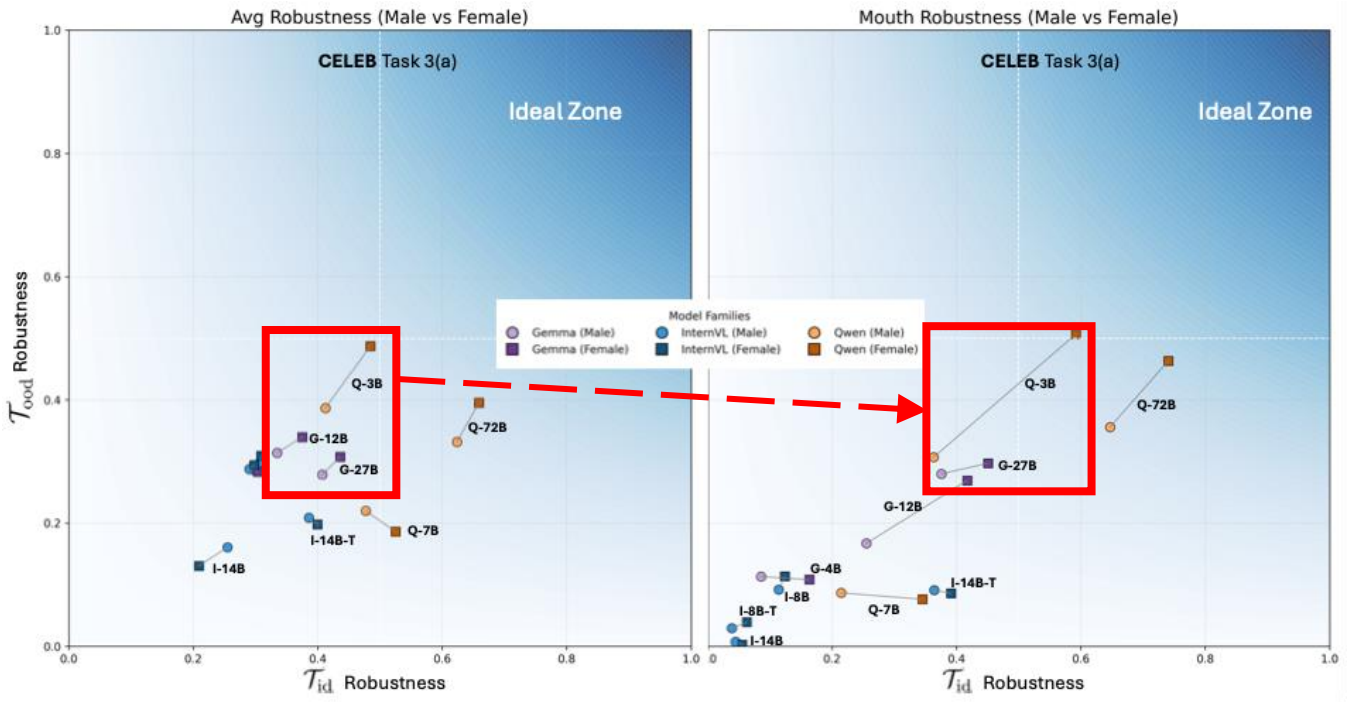


- Model have Gender Bias with Position Gap
- Bigger Models initially at Ideal Zone for ID perturbation Suddenly Drop to as OOD perturbations are tested

Our Findings – For Task 3 (Attribute Localization)

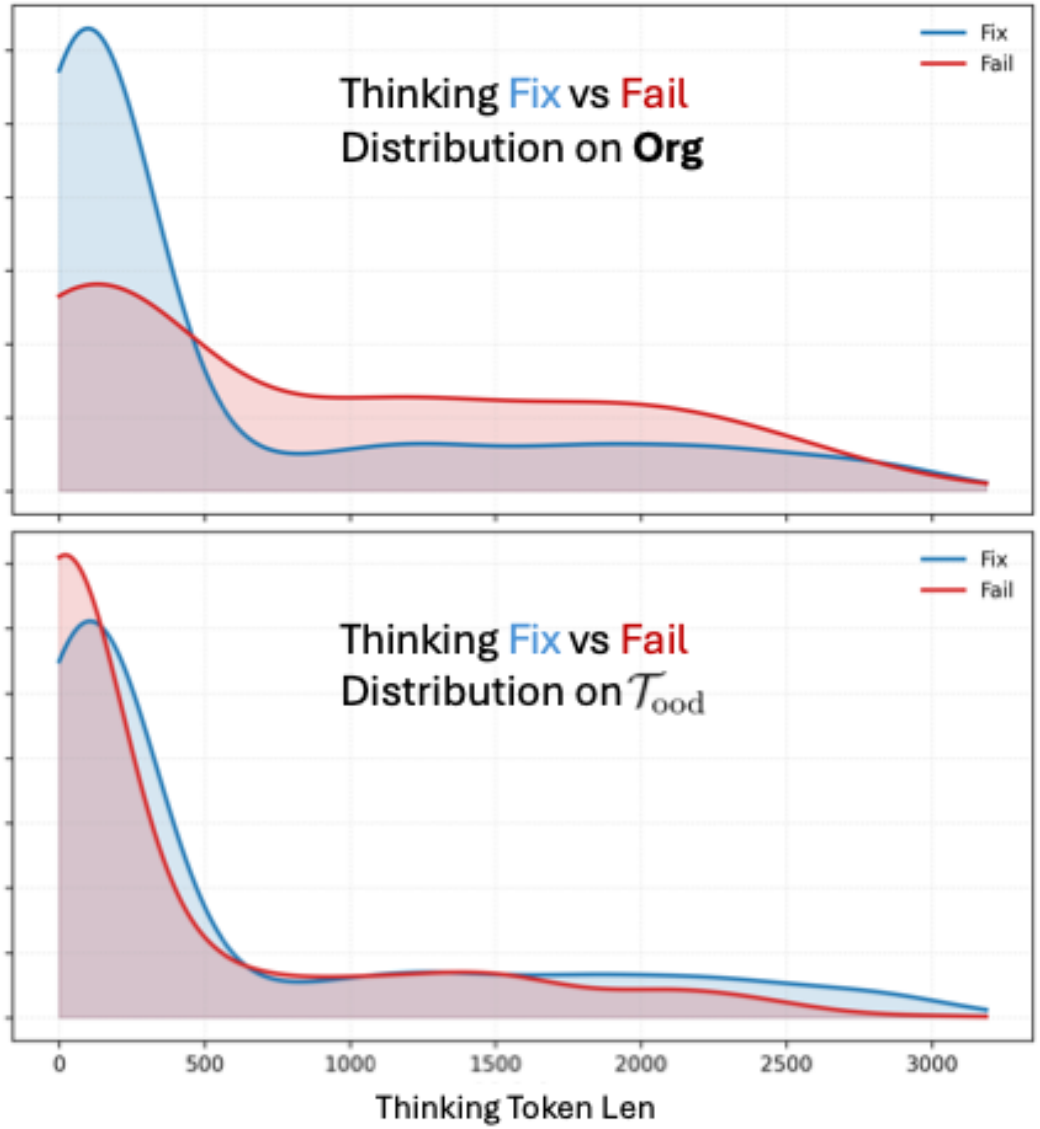






- Smaller Models are susceptible to OOD
- Avg Robustness to individual attribute performance reveal much higher gender bias.

Our Findings – For <Thinking>



- **On clean/Org:** "Fix" cases are concentrated at **short thinking lengths** (early peak), while "Fail" has a **heavier mid-length tail** → *overthinking often correlates with failure on easy/clean inputs*.
- **On OOD** both Fix/Fail peak at **very short** thinking lengths, but the **long-thinking tail is mostly Fix** → *extra compute helps when inputs are genuinely hard / shifted*.
- **Separation is clearer on Org than OOD:** on OOD the curves overlap more at short lengths, so **thinking is less diagnostic** unless the model commits to longer reasoning.

Closing Notes

Closing notes

- **Robustness is the bottleneck:** strongest failures come from **stylized OOD shifts**, not clean/ID noise.
- **Bigger isn't always safer:** robustness gains are **non-monotonic with scale** some mid/large models still break under shift.
- **Thinking is conditional:** decode-time reasoning helps **selectively** (hard/OOD), but can **waste** **compute or hurt** on easy/clean cases.
- **Key gap:** models often produce better *explanations* without better ***perception*** suggesting limits in visual grounding.

Future directions

- **Adaptive "thinking" policies:** trigger longer reasoning only when uncertainty/OOD is detected; cap tokens on confident clean cases.
- **Perception-first training:** add objectives that enforce **visual** **grounding** (counterfactuals, hard negatives, localization/region supervision).
- **OOD-aware evaluation + training:** build **style-shift benchmarks** and train with controlled stylization to improve transfer.
- **Model design:** strengthen the **vision–language interface** rather than only scaling the LM.
- **Reliability tools:** calibrate confidence, add **OOD detectors**, and use abstain/verify behaviors when perception is uncertain.

Try Perceptual Observatory Yourself

https://coral-lab-asu.github.io/PerceptualObservatory/

Code

Project

Thank You