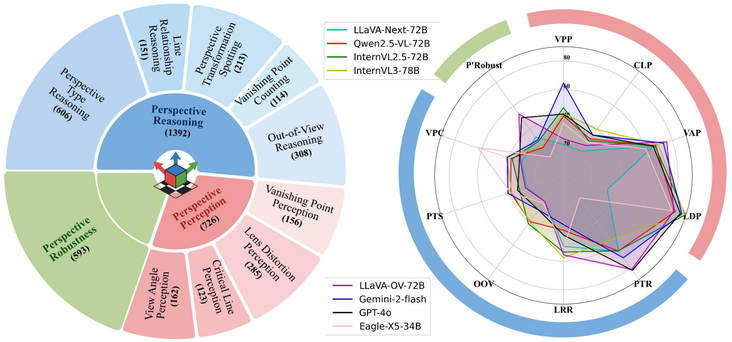
MMPerspective: Do MLLMs Understand Perspective? A Comprehensive Benchmark for Perspective Perception, Reasoning, and Robustness

Yolo Yunlong Tang[1]*, Pinxin Liu[1]*, Zhangyun Tan[1]*, Mingqian Feng[1], Rui Mao[1], Chao Huang[1], Jing Bi[1], Yunzhong Xiao[2], Susan Liang[1], Hang Hua[1], Ali Vosoughi[1], Luchuan Song[1], Zeliang Zhang[1], Chenliang Xu[1]

[1] University of Rochester, [2] Carnegie Mellon University, * Equal Contribution

Key Contribution

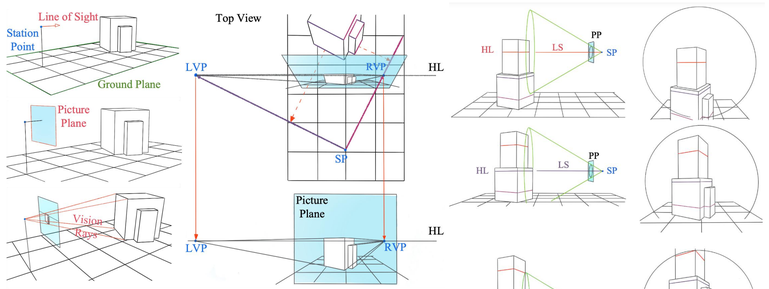
- We introduce MMPerspective, the first dedicated benchmark for evaluating perspective understanding in MLLMs, spanning 10 tasks across three dimensions, consisting of 2,711 instances and 5,083 QA pairs.
- We conduct a comprehensive evaluation of 43 representative MLLMs and reveal key limitations in perspective perception, reasoning, and robustness.
- We offer new insights into current model bottlenecks and provide guidance toward building geometry-aware, spatially grounded multimodal systems.



Left: MMPerspective benchmark consists of 2,711 instances and 5,083 QA pairs, hierarchically organized into 3 core categories (Perspective Perception, Reasoning, and Robustness).
Right: The accuracy of 8 representative MLLMs on 10 tasks of MMPerspective across the 3 categories.

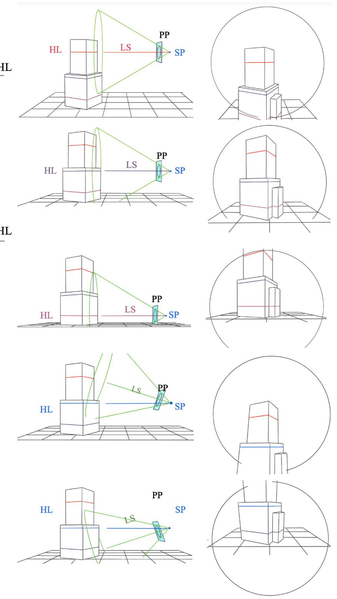
Preliminary: Perspective Foundation

Perspective is nothing more than a rational demonstration applied to the consideration of how objects in front of the eye transmit their image to it. — Leonardo da Vinci [1]

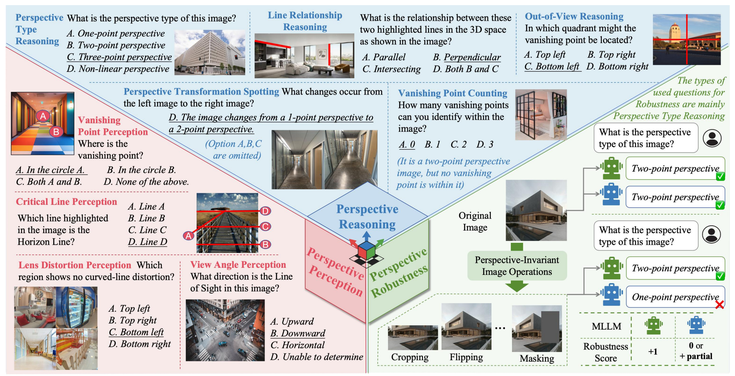


Perspective illustration with terminology. The figure is adapted from Robertson and Bertling (2013) [2].

Perspective describes how a 3D scene is projected onto a 2D image based on the viewer's position. The Station Point (SP) denotes the viewer or camera location, the Picture Plane (PP) is where the projection occurs, the Horizon Line (HL) corresponds to the viewer's eye level, and Vanishing Points (VPs) are where sets of parallel lines in 3D appear to converge, forming 1-, 2-, or 3-point perspective.



The relationship between Station Point (SP), Picture Plane (PP), Line of Sight (LS), and Horizon Line (HL) in perspective drawing.



MMPerspective benchmark overview. We introduce 10 tasks spanning 3 complementary dimensions of perspective understanding: Perspective Perception, Reasoning, and Robustness.

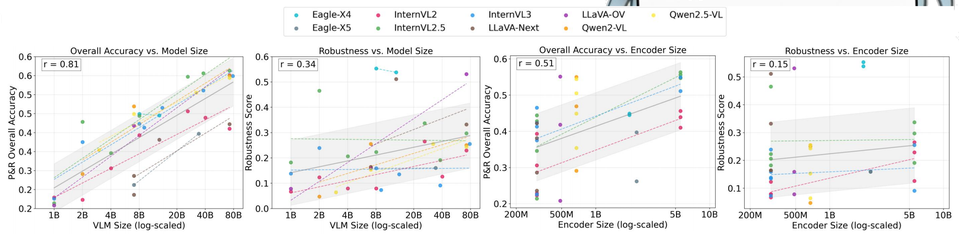
Experiments

Model	Perspective Perception				Perspective Reasoning					P'Percep & P'Reason			Robustness	
	VPP	CLP	VAP	LDP	PTR	LRR	OVR	PTS	VPC	P Acc	R Acc	Overall	Graded	Binary
MLLMs: < 7B														
InternVL2.5-2B	**47.4**	22.8	13.0	**65.3**	**62.2**	31.8	16.6	30.0	**50.0**	37.1	**38.1**	**37.7**	**59.1**	**46.5**
Qwen2.5-VL-3B	27.6	22.8	56.8	55.1	32.3	32.5	15.9	**39.4**	44.7	40.6	33.0	36.3	22.2	6.4
InternVL2.5-4B	32.1	26.0	**59.3**	64.2	28.2	30.5	10.7	37.1	36.8	**45.4**	28.7	36.1	25.0	20.6
InternVL3-2B	22.4	**28.5**	50.0	44.6	43.1	31.1	**34.4**	25.4	43.0	36.4	35.4	35.8	39.0	23.9
InternVL2-4B	26.9	12.2	54.3	60.4	18.0	**40.4**	18.8	24.4	45.6	38.4	29.4	33.4	14.5	7.9
Qwen2-VL-2B	12.2	19.5	49.4	35.8	23.3	24.5	28.9	32.9	47.4	29.2	31.4	30.4	18.0	4.7
InternVL3-1B	19.9	13.0	53.7	20.7	16.3	8.6	23.7	21.6	47.4	26.8	23.5	25.0	16.1	13.8
InternVL2-1B	20.5	20.3	15.4	24.2	24.1	11.3	24.0	22.1	44.7	20.1	25.2	23.0	18.2	6.7
LLaVA-OV-1B	13.5	14.6	35.8	24.2	15.2	19.2	19.5	22.1	40.4	22.0	23.3	22.7	13.0	7.8
InternVL2-2B	26.9	26.0	3.1	36.8	18.8	12.6	23.1	21.1	34.2	23.2	22.0	22.5	19.3	12.3
InternVL2.5-1B	14.7	23.6	0.6	33.0	20.1	11.3	13.3	34.7	45.6	18.0	25.0	21.9	19.0	18.2
MLLMs: 7B - 9B														
InternVL2.5-8B	38.5	17.9	53.1	75.4	40.8	48.3	34.7	24.9	67.5	46.2	43.3	**44.6**	38.7	22.3
Qwen2.5-VL-7B	35.3	29.3	**70.4**	73.7	42.4	44.4	32.1	28.6	44.7	52.1	38.5	44.5	33.2	15.3
Qwen2-VL-7B	34.6	25.2	63.0	64.2	57.1	49.0	27.3	31.0	46.5	46.7	42.2	44.2	46.9	25.5
InternVL3-9B	37.2	**33.3**	63.0	77.5	30.7	**53.0**	27.9	23.9	43.9	52.8	35.9	43.4	19.2	7.3
InternVL3-8B	**42.3**	27.6	67.9	**81.8**	38.1	46.4	20.8	23.9	32.5	**54.9**	32.3	42.4	29.1	15.9
LLaVA-OV-7B	34.0	**33.3**	51.2	57.9	44.9	**53.0**	19.8	**35.2**	49.1	44.1	40.4	42.0	36.1	15.9
Eagle-X4-8B	39.1	17.1	46.9	47.7	**65.3**	37.1	18.2	32.9	**68.4**	37.7	**44.4**	41.4	**60.7**	**55.3**
InternVL2-8B	33.3	19.5	59.3	73.3	27.1	36.4	**42.5**	22.1	48.2	46.4	35.3	40.2	19.9	7.9
LLaVA-Next-m-7B	35.9	21.1	35.2	50.5	17.7	37.7	15.6	27.2	46.5	35.7	28.9	31.9	17.9	16.4
Eagle-X5-7B	25.0	26.0	24.7	34.7	22.1	46.4	15.6	20.7	42.1	27.6	29.4	28.6	18.4	15.9
LLaVA-Next-v-7B	16.7	20.3	40.7	39.6	16.3	44.4	19.8	16.4	7.0	29.3	20.8	24.6	16.7	16.4
MLLMs: 10B - 30B														
InternVL2.5-26B	41.7	**35.0**	55.6	**81.8**	65.5	**46.4**	43.5	**34.3**	46.5	**53.5**	**47.2**	**50.0**	52.9	33.7
InternVL3-14B	39.1	26.0	**73.5**	73.3	36.5	34.4	**54.5**	28.2	54.4	53.0	41.6	46.7	27.3	13.5
InternVL2-26B	28.2	**35.0**	61.1	74.0	50.7	41.7	28.9	28.6	43.0	49.6	38.6	43.5	44.1	26.5
Eagle-X4-13B	**42.3**	26.8	41.4	44.6	65.8	20.5	28.2	31.0	**57.9**	38.8	40.7	39.8	**60.7**	**53.8**
LLaVA-Next-13B	7.7	17.1	54.3	34.7	**66.7**	24.5	13.0	26.8	43.9	28.5	35.0	32.1	59.7	51.1
MLLMs: 30B - 70B														
InternVL2.5-38B	**46.8**	**36.6**	67.9	89.5	58.4	51.7	38.3	**44.1**	44.7	60.2	**47.5**	**53.1**	41.6	19.1
InternVL3-38B	45.5	35.0	**71.0**	**90.9**	37.3	43.0	**56.8**	37.6	43.0	**60.6**	43.5	51.1	23.9	9.1
Qwen2.5-VL-32B	35.9	22.8	68.5	73.7	**62.0**	37.7	33.8	35.2	45.6	50.2	42.9	46.1	**48.8**	**25.5**
Eagle-X5-34B	36.5	28.5	60.5	79.6	19.5	51.0	24.0	39.0	**63.2**	51.3	39.3	44.6	18.7	16.0
InternVL2-40B	26.3	22.0	66.0	76.1	43.2	**55.0**	27.3	25.8	47.4	47.6	39.7	43.2	29.5	12.6
MLLMs: > 70B														
InternVL3-78B	43.6	**39.8**	69.8	89.1	55.9	**57.6**	40.3	38.0	**42.1**	**60.6**	46.8	**52.9**	43.6	25.5
InternVL2.5-72B	**47.4**	30.1	67.3	**89.5**	65.2	53.6	**41.9**	32.4	37.7	58.6	46.2	51.7	56.3	29.7
Qwen2.5-VL-72B	41.7	31.7	67.9	82.1	65.3	38.4	39.9	**39.0**	38.6	55.8	44.3	49.4	49.9	24.3
Qwen2-VL-72B	34.6	18.7	70.4	82.5	68.8	52.3	38.6	35.2	**42.1**	51.5	**47.4**	49.2	51.3	25.0
LLaVA-OV-72B	25.6	26.0	**75.9**	81.1	**81.4**	55.6	22.4	28.2	31.6	52.2	43.8	47.5	**71.8**	**53.1**
LLaVA-Next-72B	21.8	21.1	66.0	32.3	65.7	49.7	22.4	27.2	30.7	35.3	39.1	37.4	55.6	33.2
InternVL2-72B	26.9	18.7	57.4	56.8	56.1	47.0	24.7	24.4	7.9	40.0	32.0	35.6	43.9	22.9
MLLMs: Proprietary														
Gemini-2-flash (CoT)	**69.2**	**49.6**	72.8	87.4	78.7	32.5	**40.9**	39.9	43.9	**69.8**	**47.2**	**57.2**	50.5	24.8
GPT-4o (CoT)	45.5	46.3	70.4	**88.8**	81.4	**47.0**	34.4	37.6	34.2	62.7	46.9	54.0	69.4	**49.9**
Gemini-2-flash	64.7	35.0	**73.5**	87.0	71.3	34.4	29.9	**40.8**	41.2	65.0	43.5	53.1	56.8	30.7
GPT-4o	42.9	35.0	66.0	86.0	**82.0**	41.7	29.9	33.8	32.5	57.5	44.0	50.0	**71.9**	**49.9**
Gemini-1.5-flash (CoT)	30.1	28.5	66.7	79.3	51.0	39.7	20.1	31.5	35.1	51.1	35.5	42.4	37.8	11.6
GPT-4o-mini	35.3	24.4	43.2	71.6	43.1	29.8	14.6	31.0	**45.6**	43.6	32.8	37.6	28.7	10.8
Gemini-1.5-flash	26.9	25.2	59.3	70.5	26.4	27.8	18.2	26.8	22.8	45.5	24.4	33.8	20.6	10.6

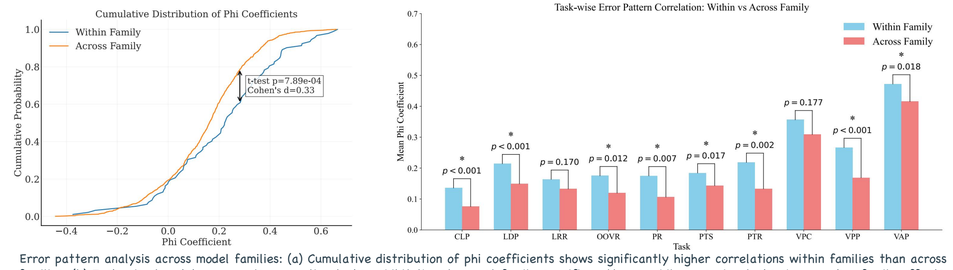
Performance of MLLMs on MMPerspective. Models are grouped by size and ranked by overall accuracy. Best scores in each group are bolded.

	Perspective Perception				Perspective Reasoning					P'Percep & P'Reason			Robustness
	VPP	CLP	VAP	LDP	PTR	LRR	OVR	PTS	VPC	P Acc	R Acc	Overall	P'Robust
GPT-4o	+2.56	+11.38	+4.32	+2.81	-0.66	+5.30	+4.55	+3.76	+1.75	+5.27	+2.94	+3.97	+0.00
Gemini-1.5-flash	+3.21	+3.25	+7.41	+8.77	+24.59	+11.92	+1.95	+4.69	+12.28	+5.66	+11.09	+8.67	+4.72
Gemini-2-flash	+4.49	+14.63	-0.62	+0.35	+7.43	-1.99	+11.04	-0.94	+2.63	+4.71	+3.63	+4.11	+15.18
Average Δ	+3.42	+9.76	+3.70	+3.98	+10.45	+5.08	+5.84	+2.50	+5.56	+5.21	+5.89	+5.59	+6.63

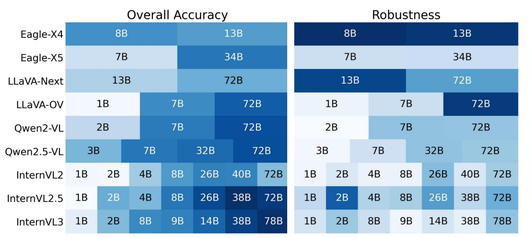
Chain of Thought (CoT) prompting improves MLLM performance on perspective tasks.



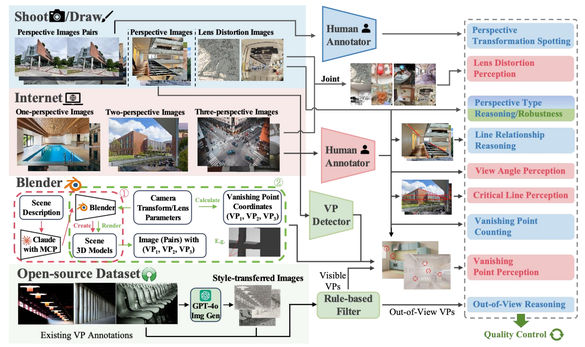
Correlation analysis between performance and size across MLLM families. Total model scaling strongly impacts perspective understanding, while vision encoder size has a limited influence on robustness.



Error pattern analysis across model families: (a) Cumulative distribution of phi coefficients shows significantly higher correlations within families than across families. (b) Task-wise breakdown reveals perception tasks exhibit the strongest family-specific patterns, while reasoning tasks show weaker family effects.



Heatmaps illustrating the relationship between model size and performance, measured by P&R Overall Accuracy and Robustness.



Data Curation Pipeline for MMPerspective.

Findings

(1) Current multimodal large models still struggle to truly understand perspective. (2) Models perform reasonably on simple tasks but drop sharply when tasks require deeper geometric reasoning or viewpoint inference. (3) CoT modestly improves model performance and robustness on perspective-related tasks by encouraging stepwise deduction. (4) while architectural and training choices strongly influence perspective perception biases, spatial reasoning challenges present consistent difficulties across all model families.

References & More Resources

[1] Leonardo Da Vinci. The notebooks of Leonardo da Vinci, volume 1. Courier Corporation, 2012.

[2] Scott Robertson and Thomas Bertling. How to Draw: drawing and sketching objects and environments from your imagination. Designstudio Press, 2013.

[3] Tang, Yunlong, Pinxin Liu, Zhangyun Tan, et al. "Mmperspective: Do mllms understand perspective? a comprehensive benchmark for perspective perception, reasoning, and robustness." In The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track. 2025.

[4] https://yunlong10.github.io/MMPerspective/

Project Page