MLLM-CompBench: A Comparative Reasoning Benchmark for Multimodal LLMs

Acknowledgment
This research is supported by grants from the NSF (OAC- 2118240, OAC2112606).

Jihyung Kil*, Zheda Mai*, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, Wei-Lun Chao
The Ohio State University

Highlights

- **MLLM-CompBench**, a comprehensive benchmark to evaluate **comparative reasoning** ability in MLLMs.
- MLLM-CompBench comprises **8 relativities**, **14 datasets** with diverse domains, **40k** human annotated samples.
- MLLMs have difficulty in **existence**, **temporal**, **spatial** and **quantity** comparison.

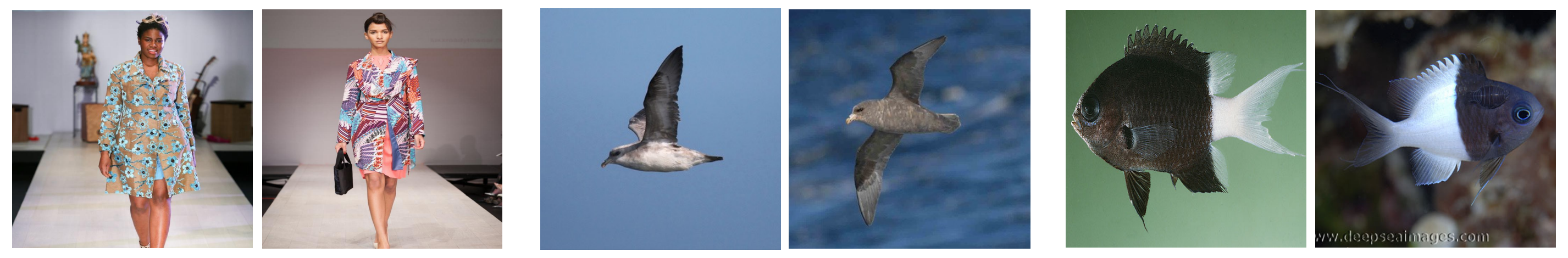
The ability to compare is important for AI models.

"Please buy the freshest apple for me"

Can MLLMs compare?

- Although MLLMs have achieved great performance in many visual tasks
- **Much less** attention has been paid to tasks involving **relativity** and **comparison** between **multiple** images for MLLMs.

Attribute

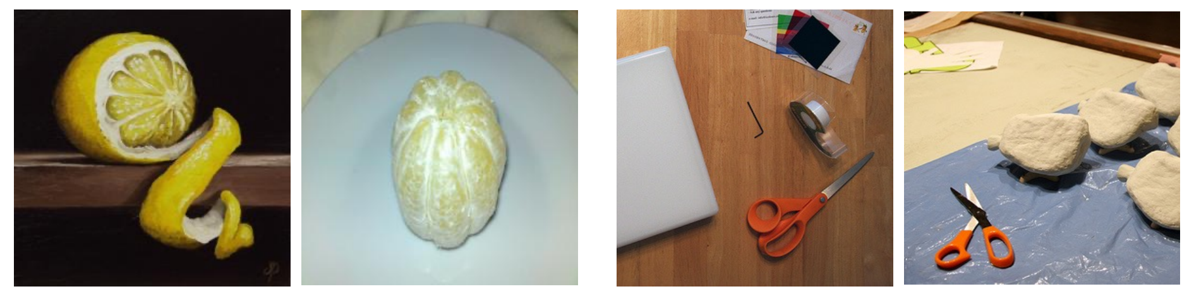


Q: Which coat is more floral?
👥: Left ⊛: Right

Q: Which bird has more grey on its breast?
👥: Right ⊛: Left

Q: Which fish has more evenly split colors?
👥: Right ⊛: Left

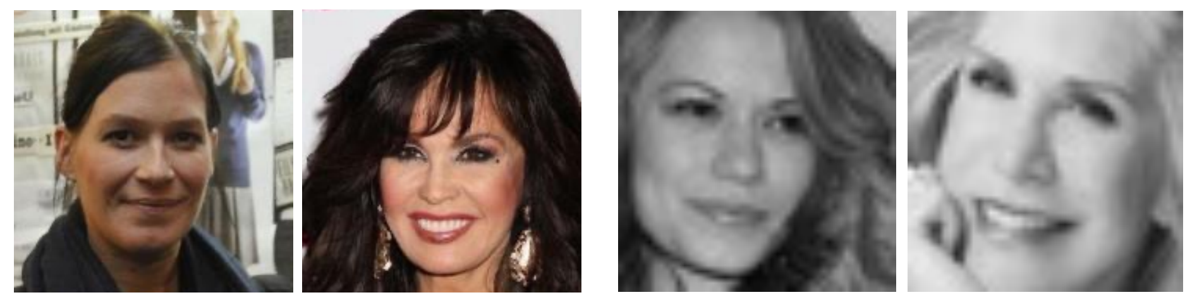
State



Q: Which lemon is more peeled?
👥: Right ⊛: Left

Q: Which scissor is more opened?
👥: Right ⊛: Left

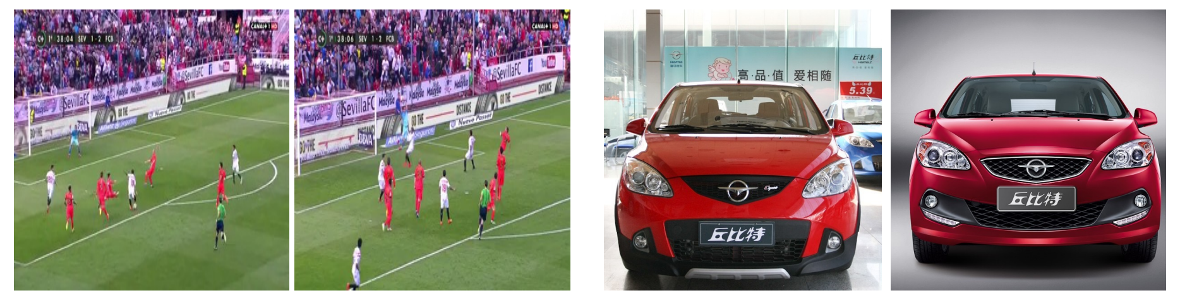
Emotion



Q: Which person smiles more?
👥: Right ⊛: Left

Q: Which person feels happier?
👥: Right ⊛: Left

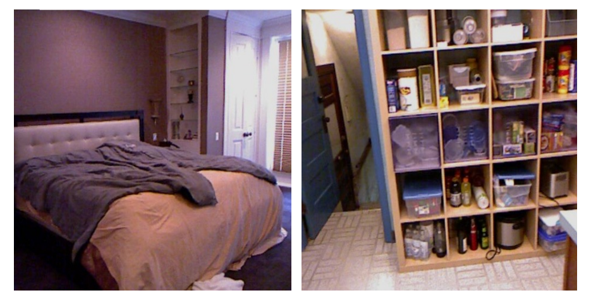
Temporal



Q: Which frame happened first?
👥: Left ⊛: Right

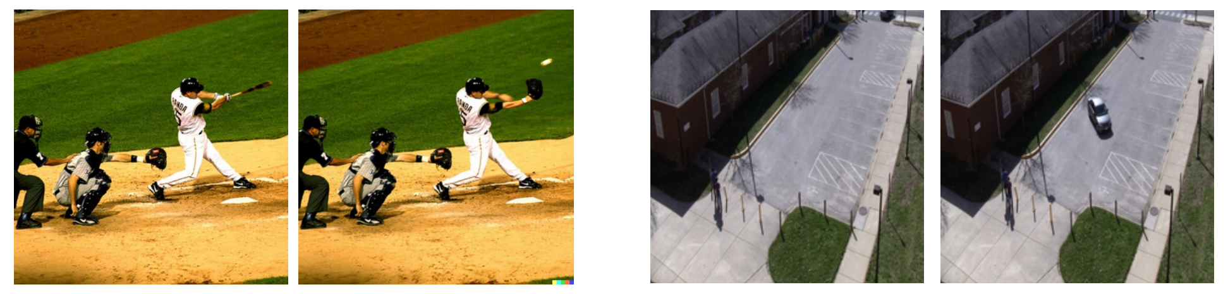
Q: Which car is newer by release year?
👥: Right ⊛: Left

Spatial



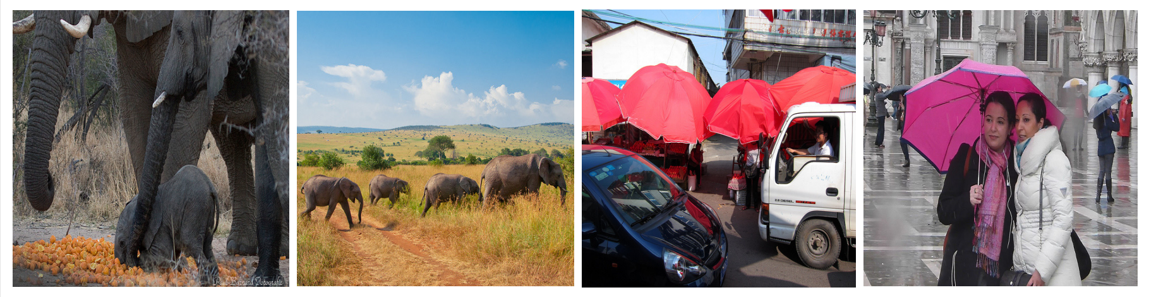
Q: Which shelves is closer to the camera?
👥: Right ⊛: Left

Existence



Q: What is the most obvious difference between two images?
👥: Baseball bat ⊛: None
👥: Car ⊛: People

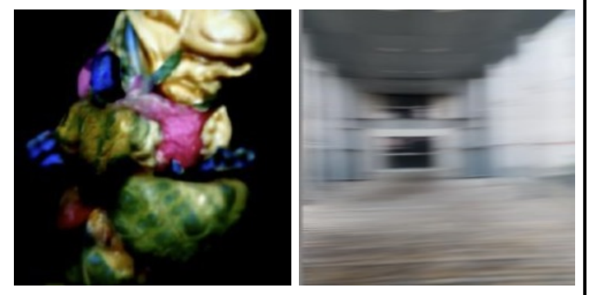
Quantity



Q: Which image has more elephants?
👥: Right ⊛: Left

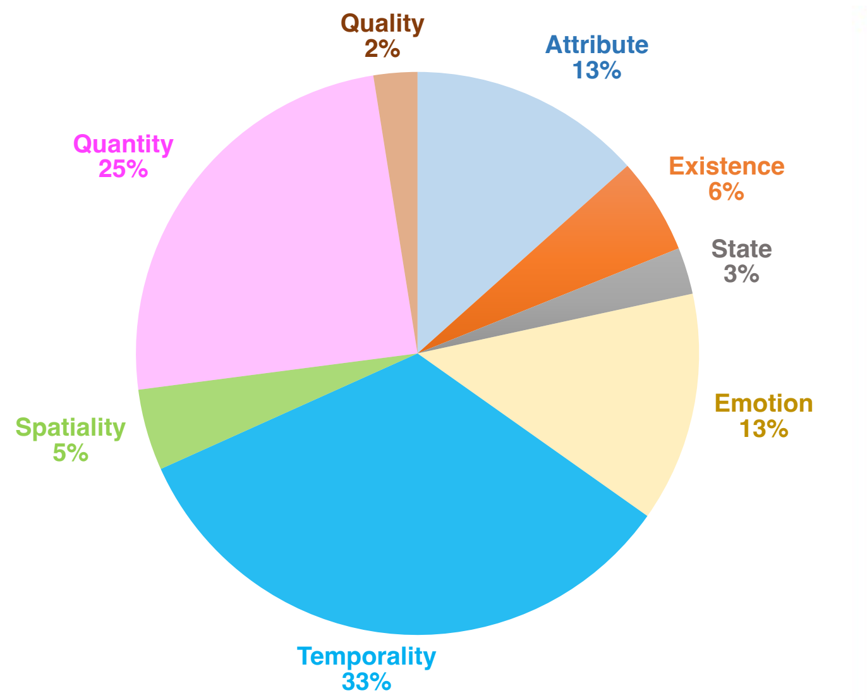
Q: Which image has more umbrellas?
👥: Right ⊛: Left

Quality



Q: Which image is more affected by motion blur?
👥: Right ⊛: Left

- A pair of **visually** or **semantically** relevant images
- A question about their **relativity**



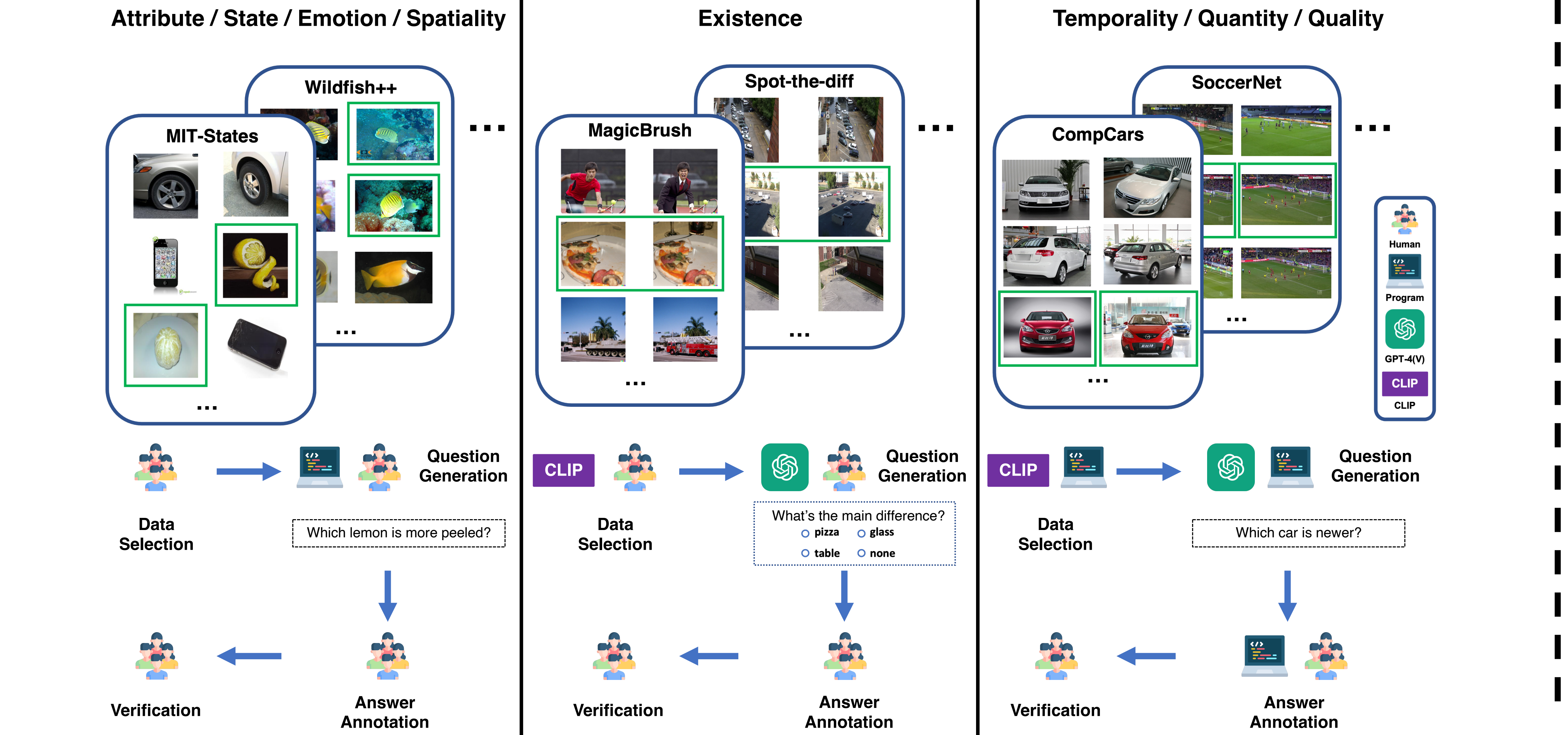
- **8** relativities
- **14** datasets with **diverse** domains
- **~40k human** annotated samples

Relativity	Dataset	Domain	# our samples
Attribute	MIT-States [24] Fashionpedia [26] VAW [46] CUB-200-2011 [59] Wildfish++ [69]	Open Fashion Open Bird Fish	0.2K 2.4K 0.9K 0.9K 0.9K
Existence	MagicBrush [65] Spot-the-diff [25]	Open Outdoor Scene	0.9K 1.2K
State	MIT-States [24] VAW [46]	Open Open	0.6K 0.5K
Emotion	CelebA [34] FER-2013 [20]	Face Face	1.5K 3.8K
Temporality	SoccerNet [19] CompCars [63]	Sport Car	8.3K 5K
Spatiality	NYU-Depth V2 [54]	Indoor Scene	1.9K
Quantity	VQAv2 [21]	Open	9.8K
Quality	Q-Bench2 [66]	Open	1K
Total	-	-	39.8K

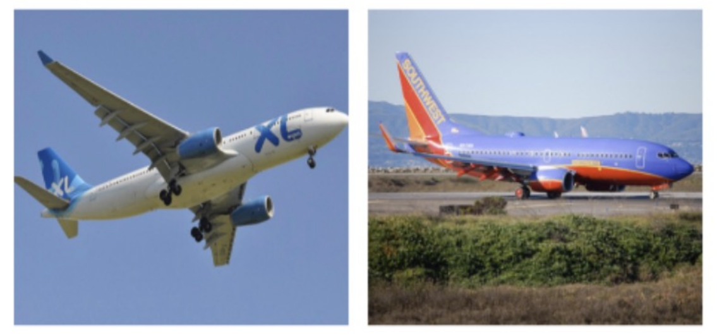
Model	Attribute					Exist.		State		Emot.		Temp.		Spat.	Quan.	Qual.	Avg
	ST	FA	VA	CU	WF	MB	SD	ST	VA	CE	FE	SN	CC	ND	VQ	QB	
GPT-4V	**91.8**	**89.0**	76.9	71.4	**72.1**	**58.3**	41.9	**92.2**	**87.8**	91.8	83.4	**71.4**	**73.7**	56.1	**63.8**	**73.0**	**74.7**
Gemini1.0-Pro	71.9	76.3	69.3	59.9	54.9	53.7	**53.0**	81.8	70.7	60.6	71.2	55.1	58.2	56.6	54.6	59.5	63.0
LLaVA-1.6	84.9	72.1	**77.7**	**72.6**	68.7	26.5	20.7	89.7	79.3	**96.2**	**83.5**	51.0	50.2	**67.2**	50.1	64.8	66.0
VILA-1.5	69.9	66.2	70.9	55.9	52.0	49.5	36.8	71.9	74.5	57.1	55.6	51.1	52.9	51.8	47.7	64.8	58.0
Chance level	50.0	50.0	50.0	50.0	50.0	8.6	9.7	50.0	50.0	50.0	50.0	50.0	50.0	50.0	33.3	37.4	43.1

MLLMs have difficulty in **existence**, **temporal**, **spatial** and **quantity** comparison.

Curation Pipeline

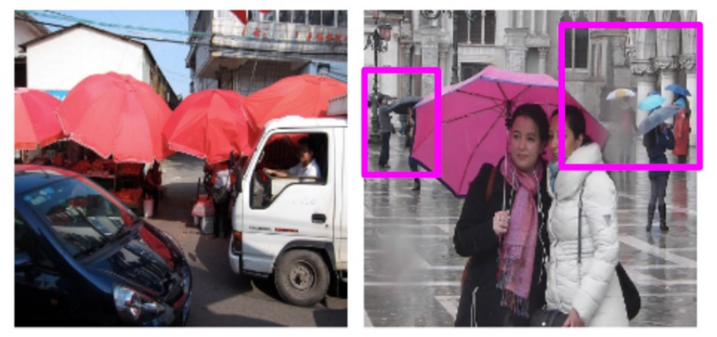


Error Analysis



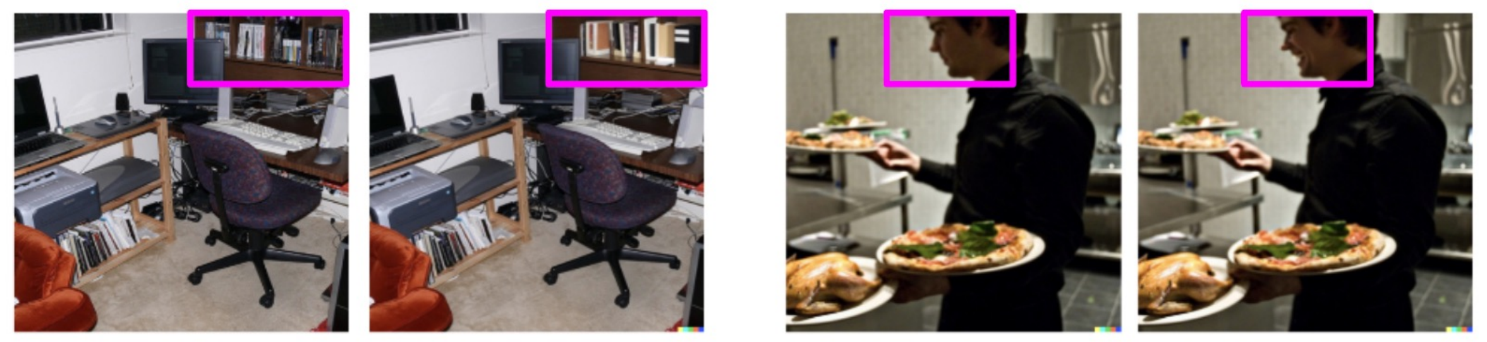
Q: Which plane is bluer?
👥: Right ⊛: Left

Differentiate colors between **objects** and **background**



Q: Which image has more umbrellas?
👥: Right ⊛: Left

Count **small** or **distant** objects



Q: What is the most obvious difference between two images?
👥: Books ⊛: None
👥: Waiter ⊛: None

Identify objects within **crowded** scenes

Recognize **out-of-focus** details