TaskGalaxy: Scaling Multi-modal Instruction Fine-tuning with Tens of Thousands Vision Task Types

KuaiShou MMU YuanQi Team

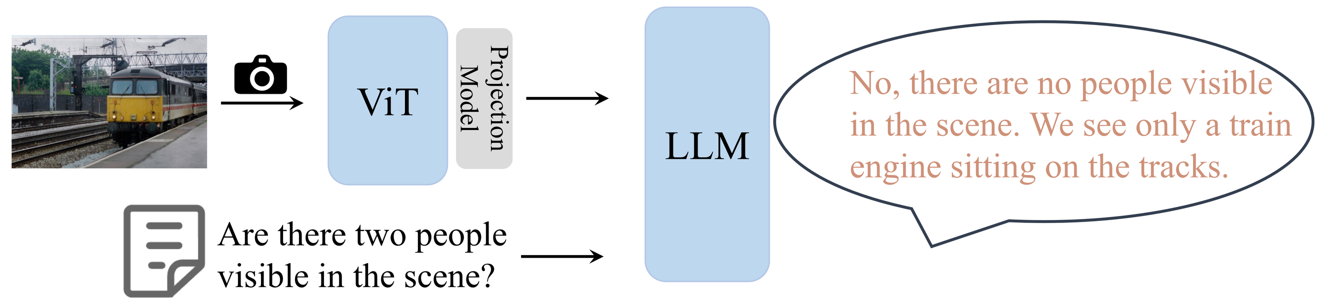
Jiankang Chen*, Tianke Zhang, Changyi Liu, Haojie Ding, **Bin Wen**†, Fan Yang, Tingting Gao, Di Zhang

ICLR International Conference On Learning Representations
2025 in Singapore

Introduction

■ Multi-modal Instruction Fine-tuning

➢ **Task**: Enable LLMs models to understand visual inputs and perform instruction-following outputs--dialogue capabilities based on visual data



➢ **Format**:

Predicted = MLLM(Input; θ), where Input = {<images>,<text>}

➢ **Core Challenge**: **Poor task diversity** limits the generalization ability of MLLMs, leading to biased outputs.

Motivation

■ Limitations of Existing Multi-modal Fine-tuning Datasets

➢ ***Insufficient Task-specific Data***: The poor diversity of VQA tasks leads to MLLMs being exposed to out-of-distribution(OOD) situations. 😭

Dataset	Instances	# Tasks	# Source
LLaVA Liu et al. (2024b)	150K	3	Synthetic
LAMM Yin et al. (2024)	196K	8	Synthetic
VL-Qwen Bai et al. (2023)	350K	Unknown	Synthetic
M3IT Li et al. (2023d)	2.4M	40	Private
mPlug-Owl Ye et al. (2023)	150K	3	Synthetic
Shikra Chen et al. (2023a)	156K	4	Synthetic
SVIT Zhao et al. (2023)	4.2M	4	Synthetic
MultiInstruct Xu et al. (2023)	510K	62	Synthetic
VISION-FLAN Xu et al. (2024)	1.6M	196	Public
TaskGalaxy (Ours)	431K	19227	Public+Synthetic

➢ ***Expensive, Time-consuming, Labor-intensive Annotations***: Involvement of a large number of specialized personnel, lack of automation. 😭

■ Rich Knowledge and Strong Ability of MLLMs

➢ **GPT-4o:** Powerful text understanding, visual-text understanding, reasoning abilities

➢ **CLIP:** Robust text-to-image alignment capabilities

➢ **GLM-4v, InternVL-Chat, InternVL2:** Strong visual understanding and question answering ability, skilled at solving multimodal tasks

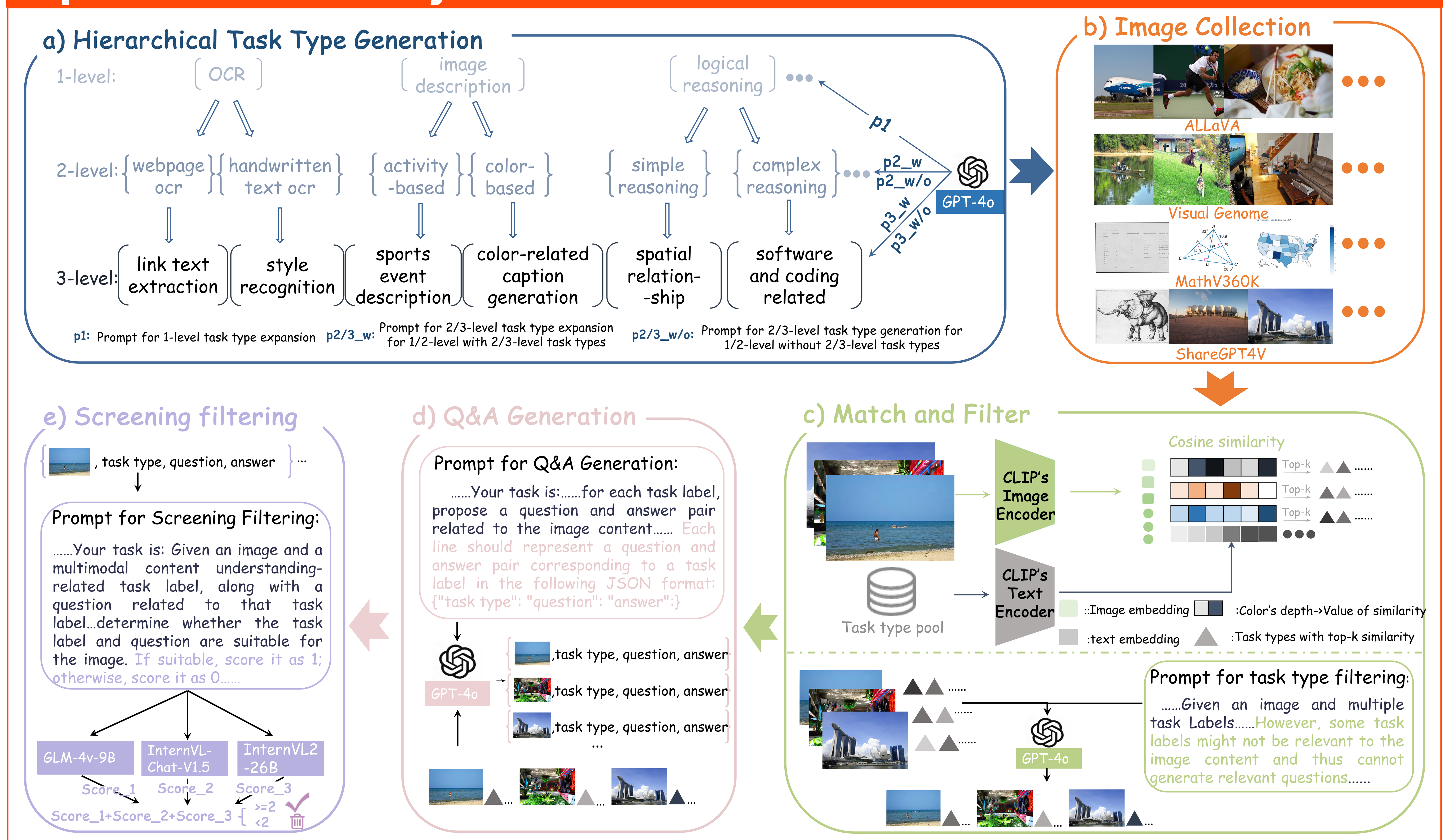
Contribution

➢ A novel multi-modal instruction fine-tuning dataset, ***TaskGalaxy***, which contains **tens of thousands** of vision task types and approximately 413k samples, addressing the limitation of task diversity in existing datasets

➢ An almost **fully automated** pipeline for creating a comprehensive fine-tuning dataset of diverse task types was designed, which can be **flexibly expanded** by incorporating high-quality images, task types, and question-answer samples

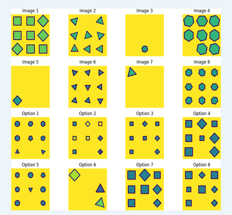
➢ Incorporating TaskGalaxy into the fine-tuning of LLaVA-v1.5 and InternVL-Chat-v1.0 resulted in **improvements across all 16 benchmarks** compared to fine-tuning with the original data, proving expanding the diversity of visual task types and high-quality question-answer pairs associated with these tasks significantly enhances the generalization capabilities of multimodal models

Pipeline —— *TaskGalaxy*



Task Types

➢100+ task type seeds→ GPT-4o → 19,227 task types

Images	Task types related to image content
	1.image description~culture-based~architecture and landmarks 2.content creation~virtual tour narration~city landmark tours 3.visual question answering~location-based visual question answering~identify specific landmark in image
	1.visual data mining~visual commonsense reasoning~spatial relationship reasoning 2.visual data mining~visual reasoning~logical reasoning
	1.text generation based on image~news article generation~contextual background generation 2.summarization~news article summarization 3.ocr~identification~newspaper ocr
	1.image description~activity-based~sports event description 2.analysis~athletic skills techniques~skill classification 3.image content understanding~content recognition~sportsmanship recognition

Cases

Task Type: image description~location-based~describe cityscapes

Question: Can you provide a detailed description of the cityscape in the image?

Answer: The image shows a panoramic view of a cityscape with a prominent clock tower in the foreground. The city is characterized by densely packed buildings with red-tiled roofs and a mix of modern and historical architecture. In the background, a range of mountains is visible, indicating the city is surrounded by natural scenery.

Task Type: suggestions~sports safety

Question: Based on the image, what general sports safety suggestions can be recommended?

Answer: General sports safety recommendations include providing all participants with proper, well-fitted protective gear such as helmets and gloves, ensuring all children are visibly marked for identification, maintaining a clear and supervised skiing area, offering quick access to first aid, and educating children on skiing safety and etiquette before allowing them on the slopes.

Task Type: storytelling based on image scenes~image description~historical context

Question: Create a short story based on the image scene, incorporating historical context.

Answer: In the bustling colonial city of Saigon, the main thoroughfare was alive with the clatter of rickshaws bearing their human cargo. The grand colonial edifices stood as silent witnesses to the daily ebb and flow of life. Amid the sea of conical hats, Linh, an ambitious young man, dreamed of one day owning his own rickshaw fleet. The imposing government building on the left side of the street symbolized power and authority, a place where Linh hoped to one day step inside not as a servant but as a respected businessman. Life in the colonial city was a delicate dance between aspiration and reality, shaped by the intricate tapestry of cultural influences and the enduring human spirit.

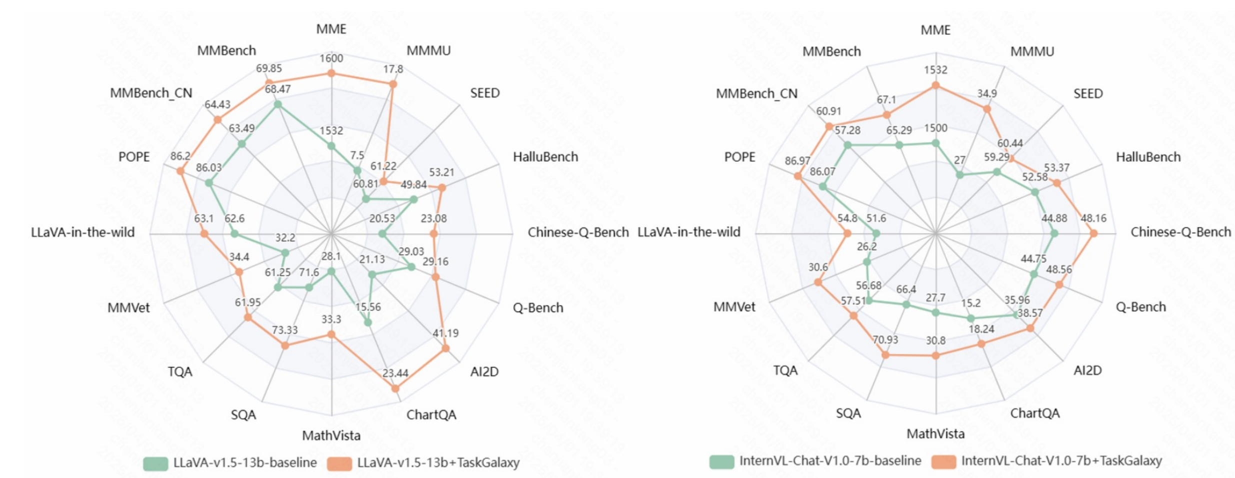
Task Type: ocr~logo recognition~logo-text association

Question: What text is associated with the logo?

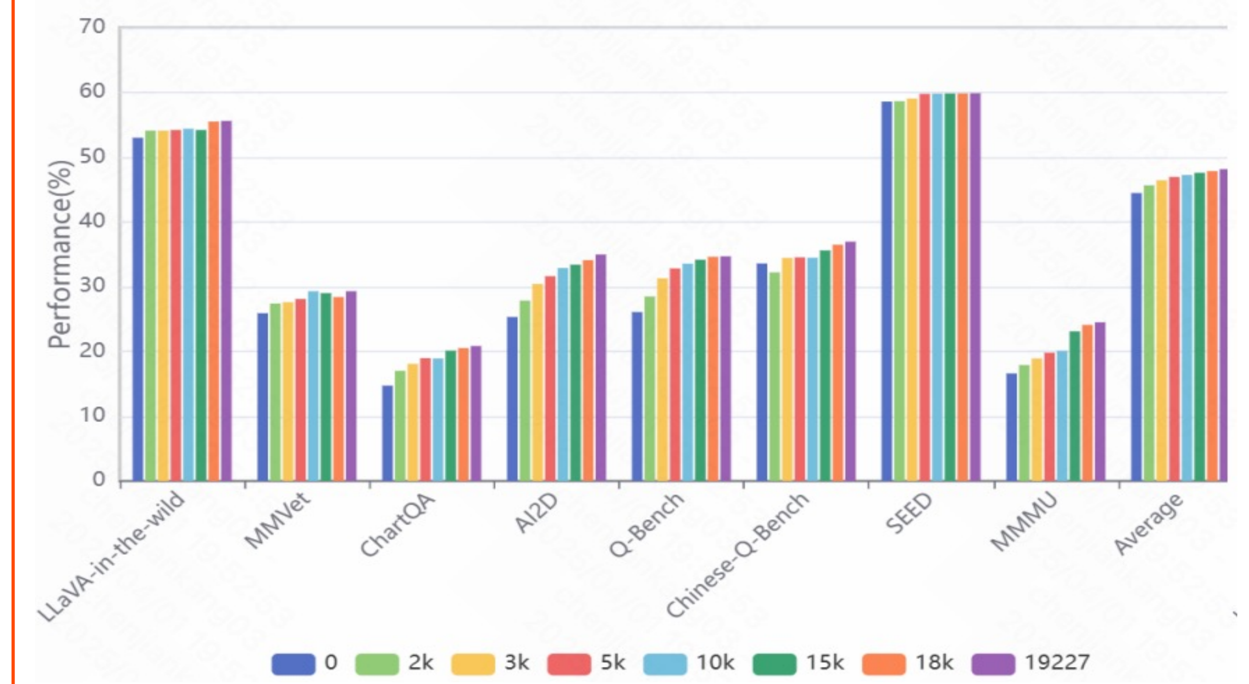
Answer: The text associated with the logo is 'ВС РФ' and 'ВЗВОД ПРОТИВОДИВЕРСИОННОЙБОРЬБЫ'.

Evaluation

❖ **Benefits of high task type coverage in TaskGalaxy for the SFT stage**



❖ **More Task Types, the Better !**



❖ **TaskGalaxy: Leading the Way in Instruction-Tuning Performance**

Model	Method	MME	MMB	MMB^{CN}	POPE	LLaVA^{W}	MMVet	TQA	SQA	MathVista
LLaVA-v1.5-7B	Baseline	1476	63.29	56.45	86.30	47.70	24.70	57.59	68.77	28.20
	ShareGPT-4V	1501	65.97	59.10	86.29	49.20	29.00	57.56	70.60	28.20
	LLaVA-OneVision	1251	59.79	52.84	83.90	51.20	**29.60**	52.99	**73.19**	28.20
	ALLaVA-4V	1474	60.13	55.39	84.21	38.00	27.00	53.77	70.05	29.20
	Cambrian-1	1494	61.08	54.46	85.46	52.00	25.70	55.17	71.03	29.10
	TaskGalaxy	**1520**	**66.62**	**59.43**	**86.40**	**52.30**	28.60	**58.08**	71.06	**29.30**
		ChartQA	AI2D	Q-Bench	Q-Bench^{CN}	HalluBench	SEED	MMMU	Average (w/o MME)	
	Baseline	14.40	25.29	24.89	31.26	47.95	58.62	19.70	43.62	
	ShareGPT-4V	17.84	27.08	26.22	32.51	48.79	59.26	15.60	44.48	
	LLaVA-OneVision	18.72	27.95	24.48	33.51	47.74	30.06	17.00	41.68	
	ALLaVA-4V	17.00	23.73	23.95	33.18	48.73	40.52	18.10	41.53	
	Cambrian-1	20.07	29.46	26.70	33.61	50.78	49.52	19.30	44.23	
	TaskGalaxy	19.90	**32.70**	**30.24**	**34.01**	**50.95**	**59.32**	**20.70**	**46.49**	
InternVL-Chat-v1.0-7B		MME	MMB	MMB^{CN}	POPE	LLaVA^{W}	MMVet	TQA	SQA	MathVista
	Baseline	1488	64.86	56.41	86.03	48.30	25.70	55.29	65.63	27.00
	ShareGPT-4V	1191	47.08	40.63	82.62	30.00	17.90	44.21	64.55	27.20
	LLaVA-OneVision	1350	61.23	54.74	67.94	32.50	19.40	37.74	66.29	25.00
	ALLaVA-4V	1425	62.76	52.78	84.50	21.50	23.50	48.04	66.29	29.40
	Cambrian-1	1481	60.22	53.01	84.17	43.70	26.80	52.61	67.71	**33.00**
	TaskGalaxy	**1512**	**65.03**	**57.91**	**86.23**	**52.30**	**30.10**	**56.15**	**68.88**	30.10
		ChartQA	AI2D	Q-Bench	Q-Bench^{CN}	HalluBench	SEED	MMMU	Average (w/o MME)	
	Baseline	14.12	35.92	42.89	43.73	51.94	59.06	26.90	47.17	
	ShareGPT-4V	14.52	35.59	46.69	36.38	52.36	47.24	30.30	42.48	
	LLaVA-OneVision	13.76	22.75	40.08	42.89	53.39	40.87	24.60	40.20	
	ALLaVA-4V	12.99	28.28	42.87	44.16	51.41	48.36	27.30	42.94	
	Cambrian-1	**16.00**	36.69	48.00	41.33	**54.63**	56.24	30.60	46.98	
	TaskGalaxy	15.16	**37.69**	**48.21**	**46.32**	53.00	**60.44**	**32.80**	**49.63**	

❖ **The Benefits of Chain-of-Thought**

Model	Method	MME	MMB	LLaVA^{W}	MathVista	ChartQA	Q-Bench	MMMU	Average(w/o MME)
LLaVA-v1.5-7B	Baseline	1506	64.69	53.0	26.7	14.72	26.08	16.6	44.46
	Baseline+max_5	1506	65.80	53.4	27.3	20.20	36.48	17.4	46.61
	Baseline+max_5 (CoT)	1523	66.72	64.7	27.9	20.96	43.27	19.3	47.92

❖ **TaskGalaxy remains strong on advanced architecture.**

Model	Method	MME	MMB	MMB^{CN}	POPE	LLaVA^{W}	MMVet	TQA	SQA	MathVista
InternVL-Chat-V2.0-8B	Baseline	1536	68.52	66.46	86.30	**63.20**	46.17	66.24	90.58	50.10
	TaskGalaxy	**1565**	**73.88**	**70.79**	**86.90**	62.85	**48.86**	**70.49**	**92.71**	**52.31**
		ChartQA	AI2D	Q-Bench	Q-Bench^{CN}	HalluBench	SEED	MMMU	Average (w/o MME)	
	Baseline	**76.64**	75.88	57.79	56.98	57.51	62.72	40.50	65.86	
	TaskGalaxy	76.56	**76.75**	**59.65**	**57.12**	**58.99**	**64.25**	**41.22**	**67.81**	