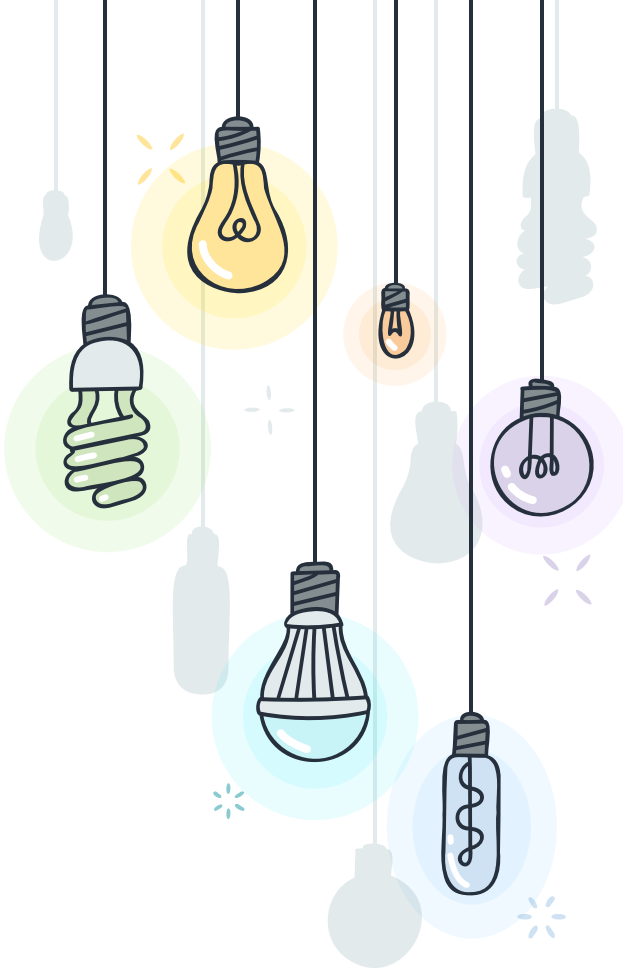
Part-III

Vision-Language Capability Evaluation



Kaipeng Zhang

Researcher

Shanghai AI Lab / Shanghai Innovation Institute

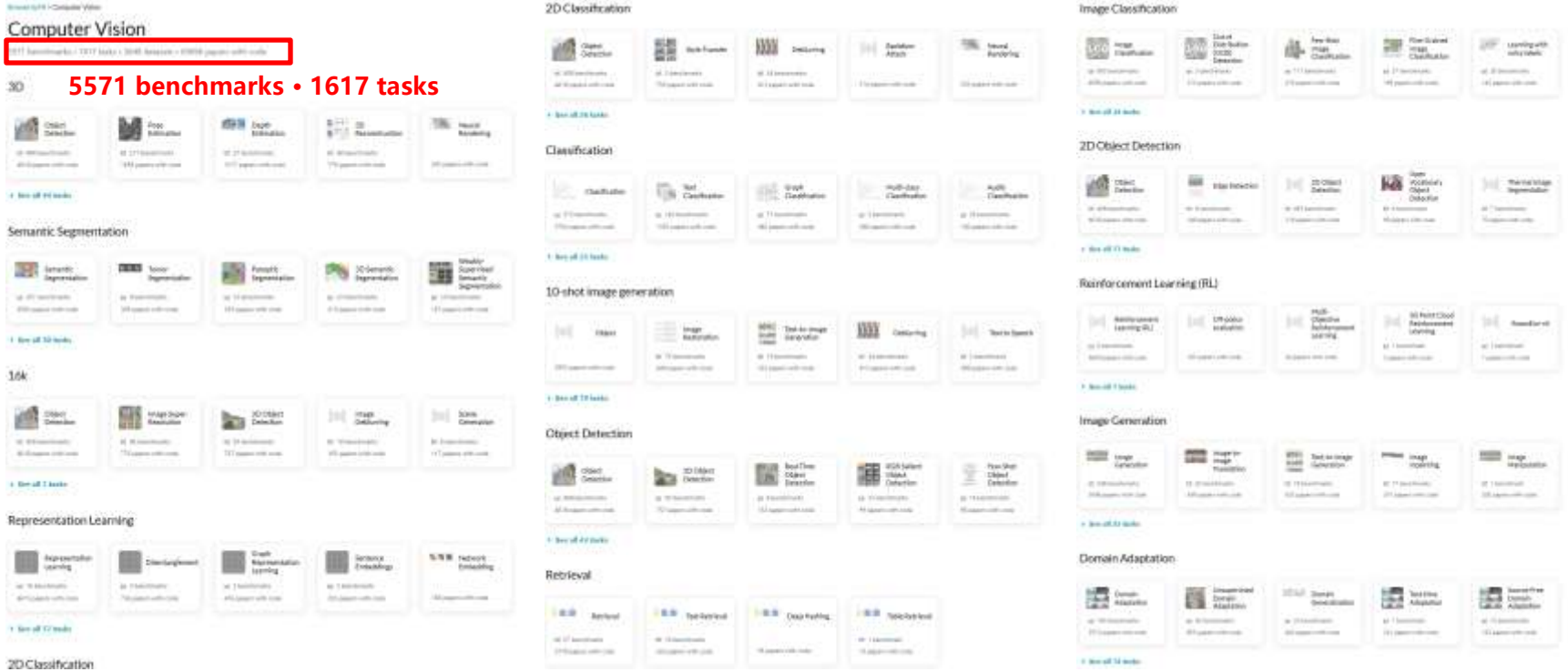
https://kpzhang93.github.io/

Table of Content

Table of Content

General Evaluation

Vision-centric evaluation

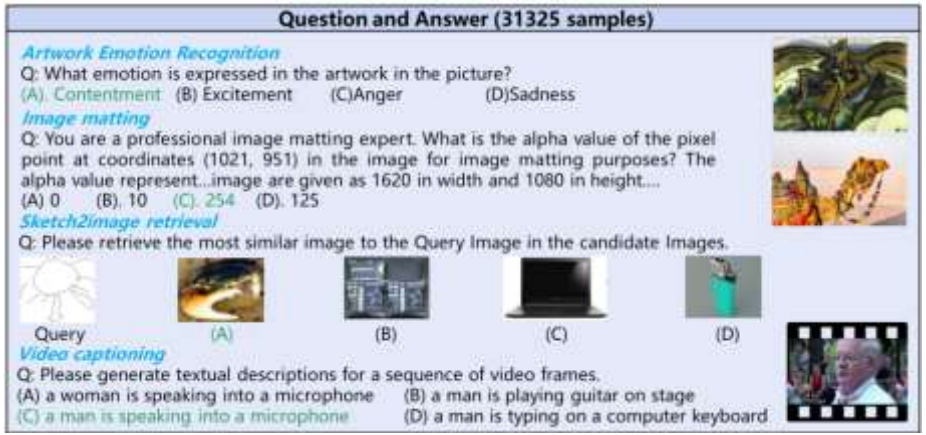


Datasets in Paper With Code

General Evaluation

Vision-centric evaluation

Convert to VQA format



Examples in MMT-Bench - multi-choice

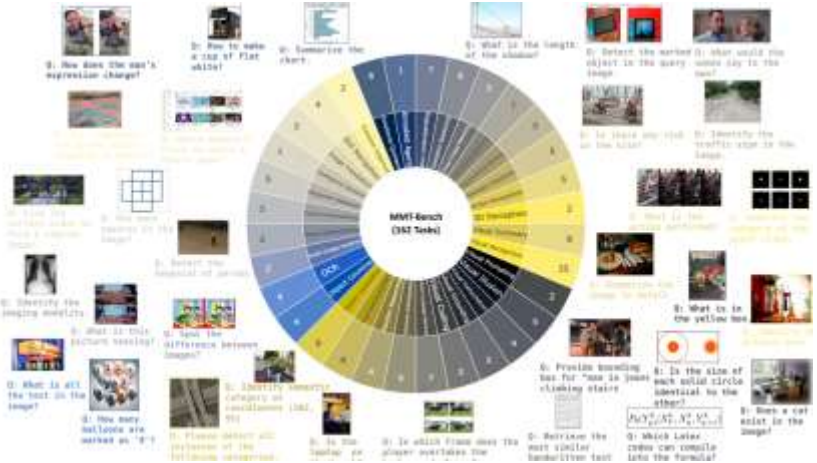
Examples in MME – Yes/No

MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI
MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models

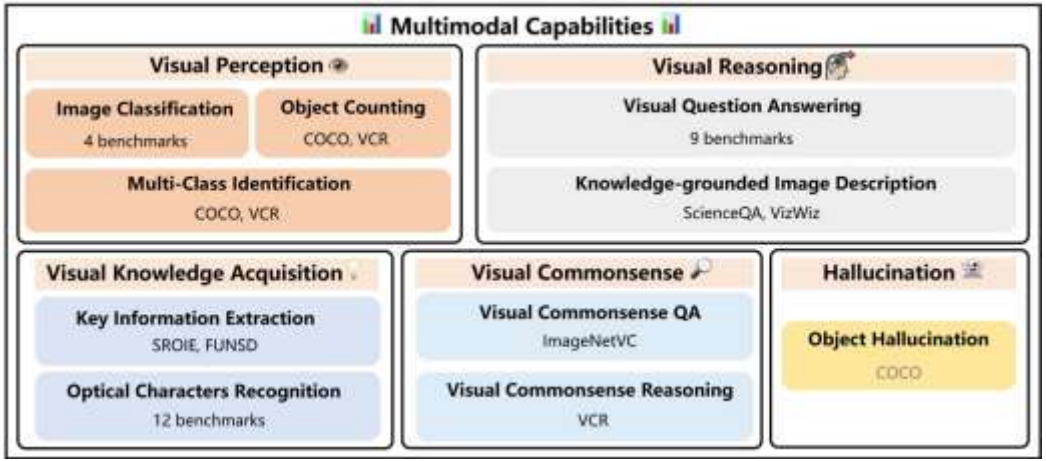
General Evaluation

Vision-centric evaluation

Convert to VQA format



MMT-Bench



LVLM-eHub

MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI
LVLM-EHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models

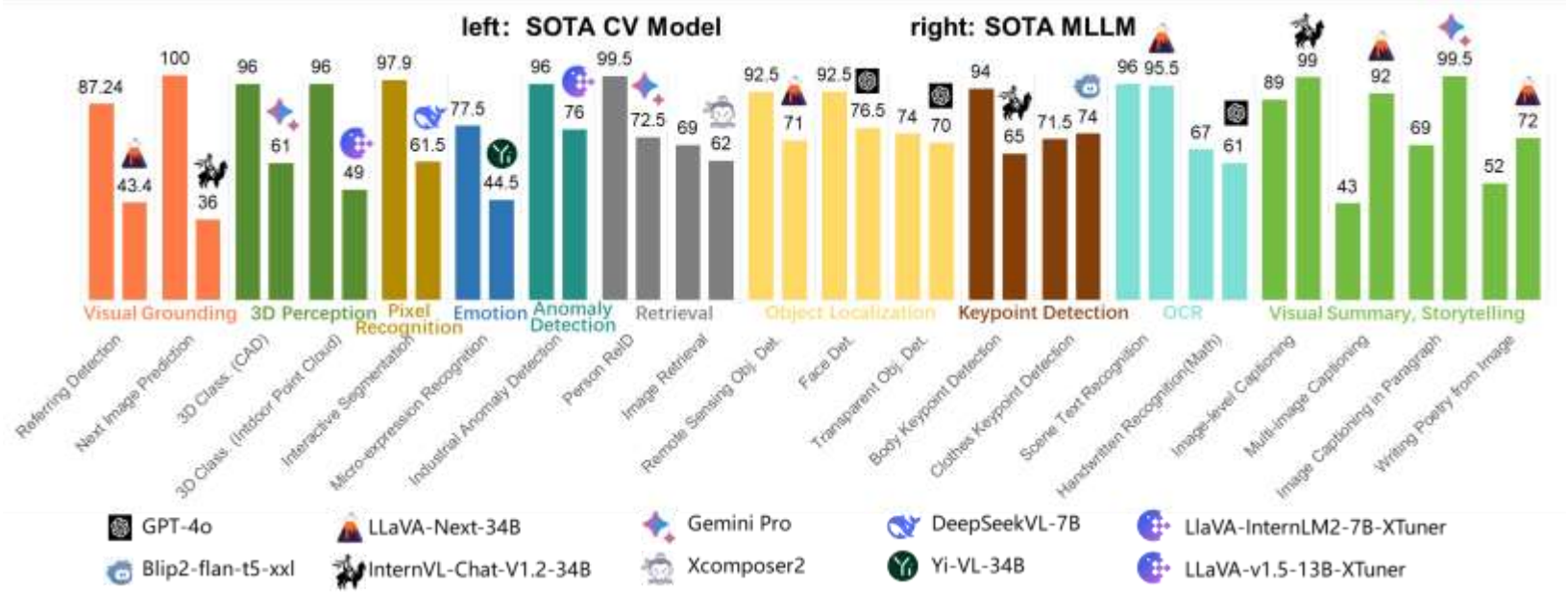
General Evaluation

Vision-centric evaluation

Rank	Method	Param (B)	Language Model	Vision Model	MMT-Bench_VAL
1	InternVL2-Llama3-76B	76	Llama-3-70B-Instruct	InternViT-6B	67.6
2	GPT-4o (0513, detail-high)				67.3
3	InternVL2-40B	40	Nous-Hermes-2-Yi-34B	InternViT-6B	66.1
4	GPT-4o (0513, detail-low)				65.6
5	Yi-Vision				62.9
6	Ovis1.5-Gemma2-9B	11.4	Gemma2-9B-It	SigLIP-400M	62.8
7	Claude3.5-Sonnet				61.8
8	GPT-4o-mini (0718, detail-high)				61.2
9	OmChat-v2.0-13B	13	Qwen2-7B	InternViT-6B	61
10	MiniCPM-V-2.6	8	Qwen2-7B	SigLIP-400M	60.8
11	Ovis1.5-Llama3-8B	8	Llama-3-8B-Instruct	SigLIP-400M	60.7
12	InternVL2-26B	26	InternLM2-20B	InternViT-6B	60.4

General Evaluation

Vision-centric evaluation



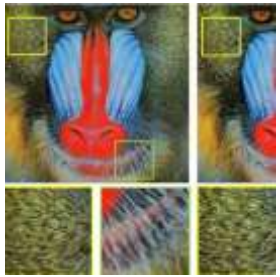
MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI

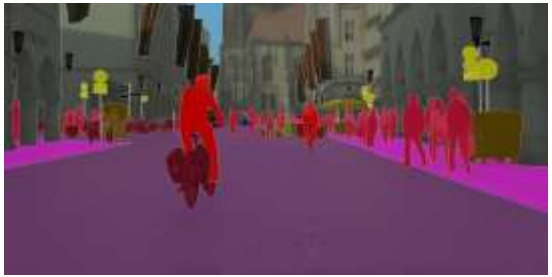
General Evaluation

Vision-centric evaluation

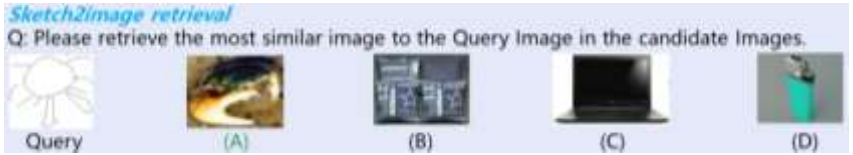
Convert to VQA format

Most MLLMs output text only





Some conversions are difficult



Difficulty control is challenging

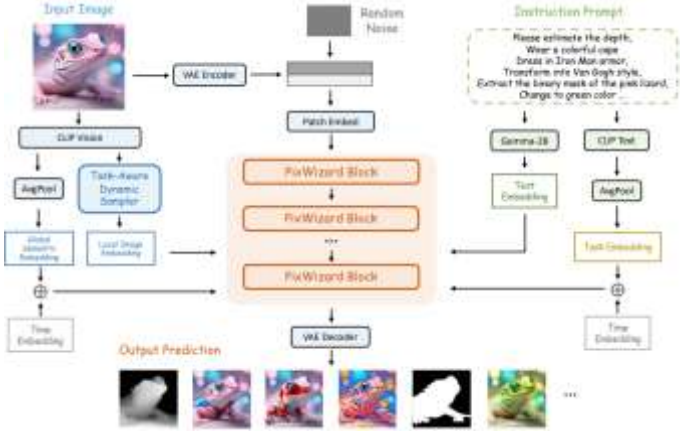
General Evaluation

Vision-centric evaluation

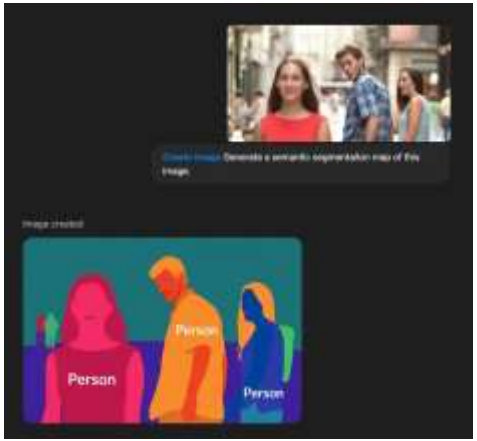
Use the original task format

More and more models can output images

Vitron



PixWizard



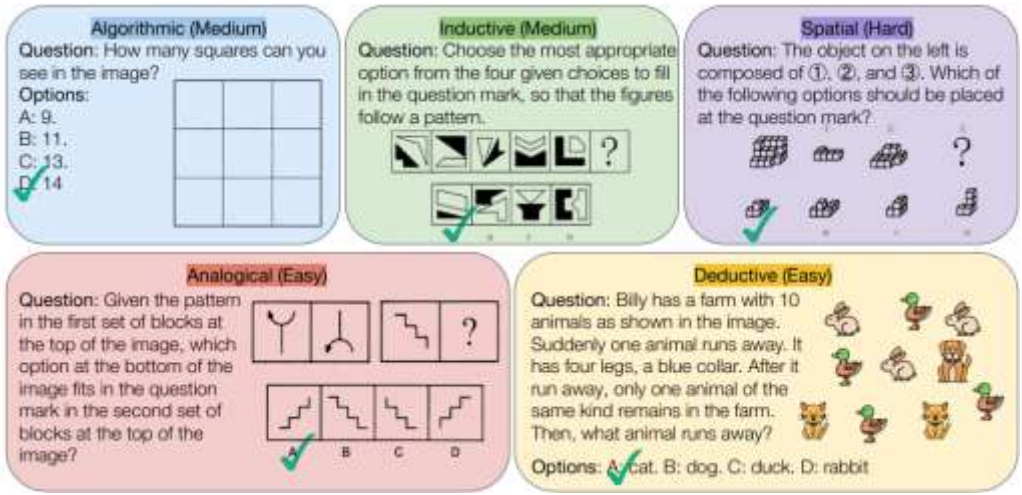
GPT-4o-Image

VITRON:A Unified Pixel-level Vision LLM forUnderstanding, Generating, Segmenting, Editing
PixWizard: Versatile Image-to-Image Visual Assistant with Open-Language Instructions

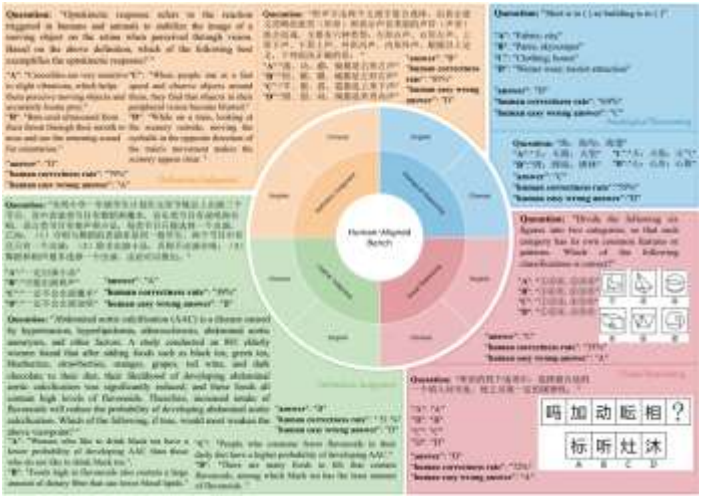
General Evaluation

Vision-centric evaluation

Visual Puzzle



VisualPuzzles



Human-Aligned Bench

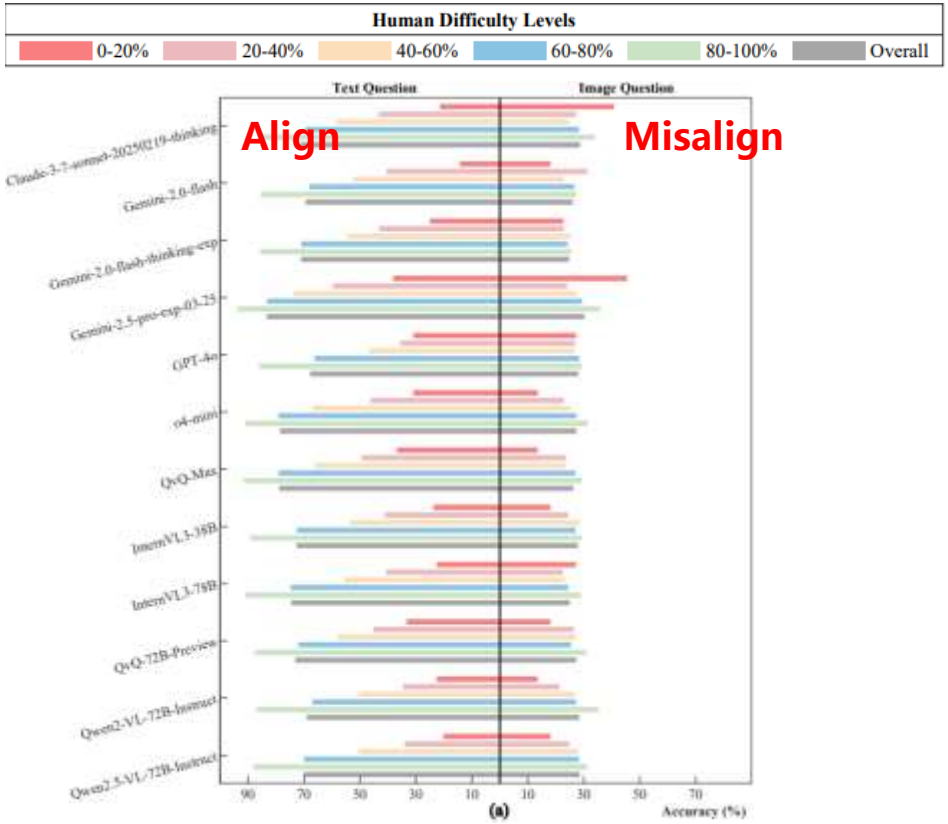
VisualPuzzles: Decoupling Multimodal Reasoning Evaluation from Domain Knowledge
Human-Aligned Bench: Fine-Grained Assessment of Reasoning Ability in MLLMs vs. Humans

General Evaluation

Vision-centric evaluation

Visual Puzzle

MLLMs' visual intelligence misalign with human intelligence

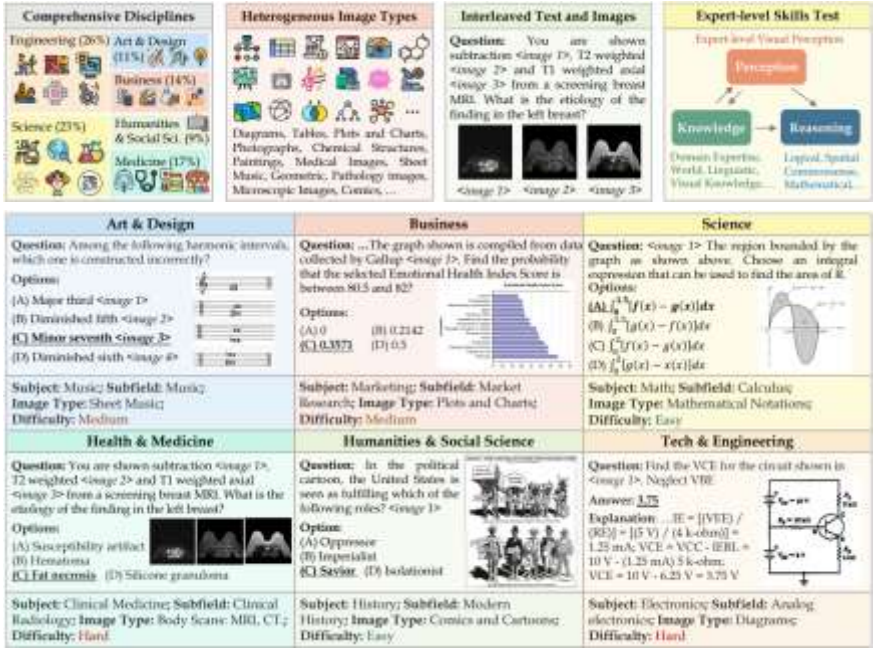


Human-Aligned Bench: Fine-Grained Assessment of Reasoning Ability in MLLMs vs. Humans

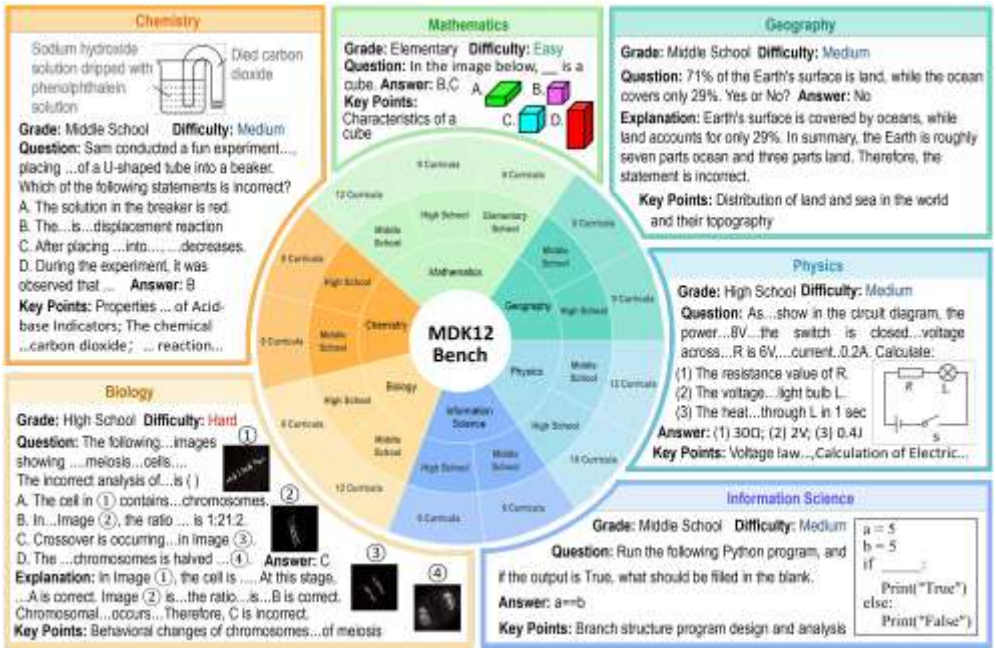
General Evaluation

Vision-Language

Multi-discipline



MMMU (college-level)



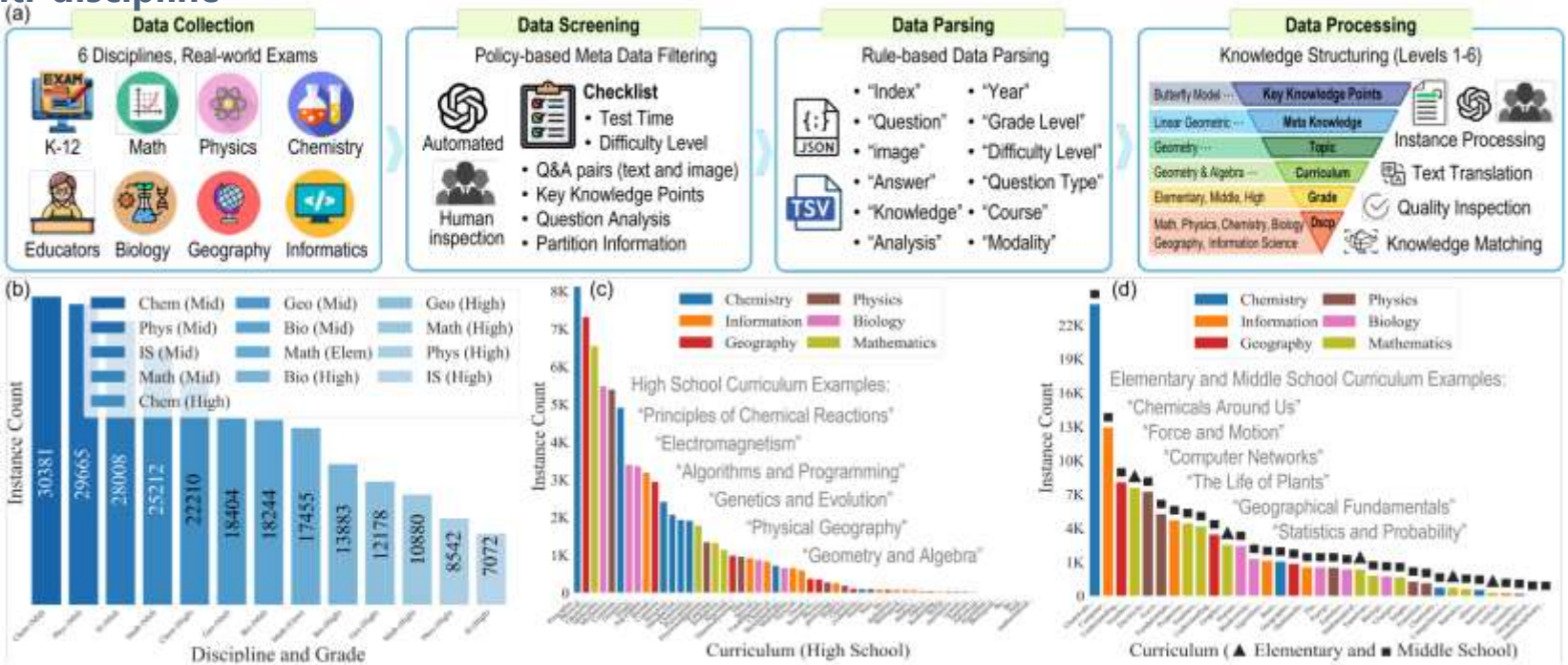
MDK12 (K12-level)

MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI
MDK12-Bench: A Multi-Discipline Benchmark for Evaluating Reasoning in Multimodal Large Language Models

General Evaluation

Vision-Language

Multi-discipline

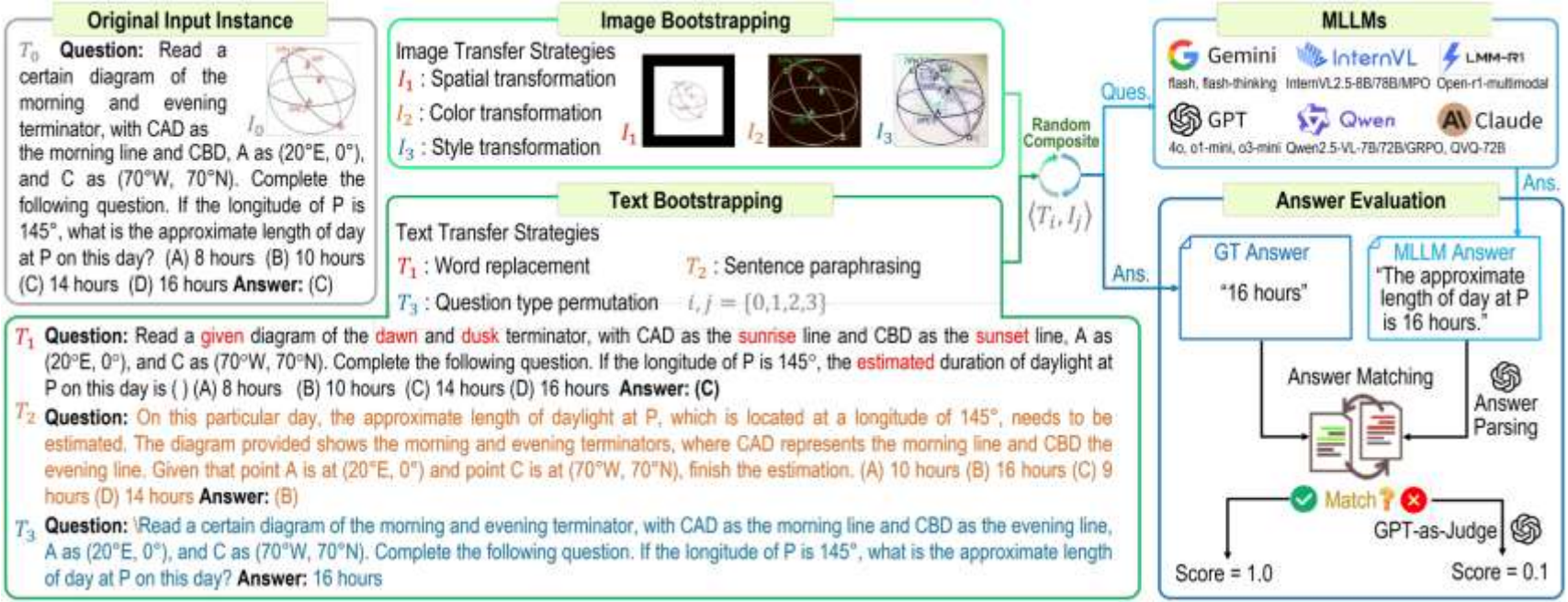


MDK12 (K12-level)

General Evaluation

Vision-Language

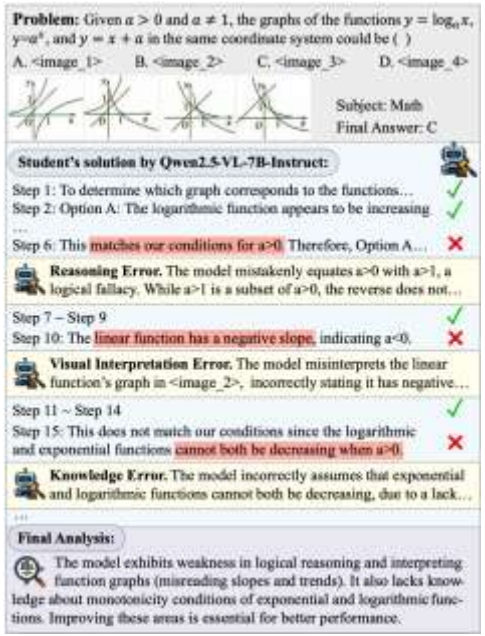
Multi-discipline



MDK12 (K12-level)

Vision-Language

Multi-discipline - Process-level evaluation



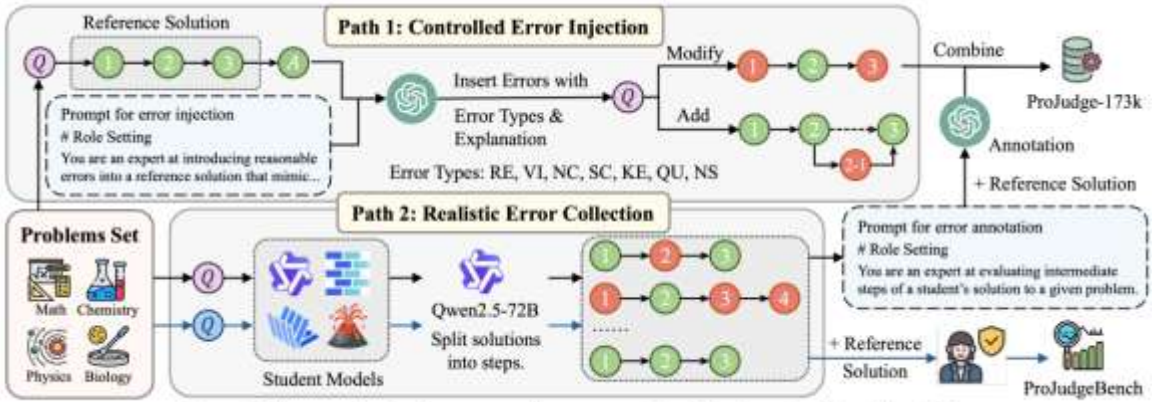
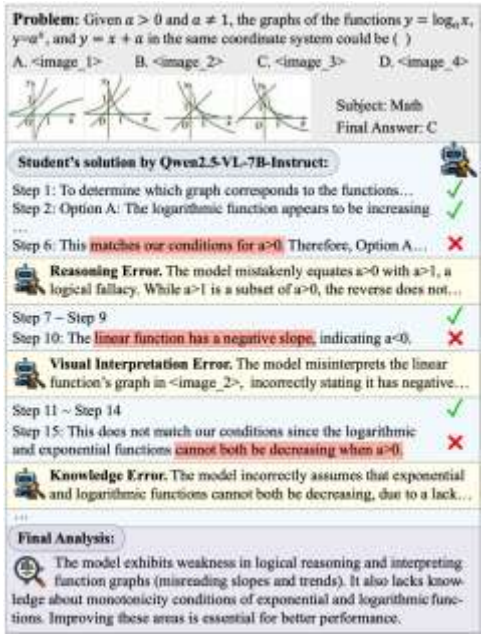


Figure 2. An overview for data construcion process of ProJudgeBench and ProJudge-173k.

Vision-Language

Multi-discipline - Process-level evaluation



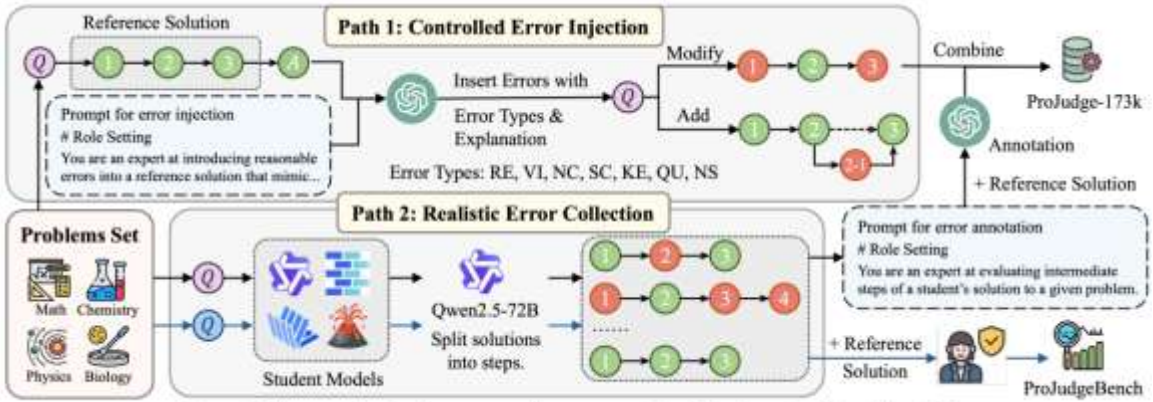


Figure 2. An overview for data construcion process of ProJudgeBench and ProJudge-173k.

General Evaluation

Vision-Language

Multi-discipline - Process-level evaluation

Model Name	Step Corr.	Error Types							
		Overall	RE.	VI.	NC.	SC.	KE.	QU.	NS.
Open-source MLLMs									
InternVL2.5-8B	25.58	6.77	8.19	0.	8.18	1.75	1.71	1.07	0.
InternVL2.5-26B	66.72	13.51	16.44	0.64	14.80	2.63	2.85	2.15	**4.16**
InternVL2.5-38B	**78.85**	17.12	17.95	**3.31**	27.08	8.33	**17.54**	**25.80**	2.08
MiniCPM-V-2_6	23.26	0.13	0.03	0.	1.68	0.	0.	0.	0.
Qwen2.5-VL-3B	11.47	0.84	0.89	0.	2.11	0.43	0.57	0.	0.
Qwen2-VL-7B	34.65	0.69	0.55	0.	1.55	0.43	2.71	0.	0.
Qwen2-VL-72B	77.93	**33.87**	**40.11**	1.49	**39.91**	**30.70**	10.27	4.83	2.08
Proprietary MLLMs									
Gemini-2.0-flash-exp	72.51	35.54	40.70	26.60	30.18	5.70	12.41	8.06	**8.33**
Gemini-2.0-thinking-exp	72.61	35.27	40.06	**27.67**	32.44	4.82	11.98	9.13	**8.33**
GPT-4o	**85.10**	**44.89**	**52.94**	9.61	**40.90**	**27.19**	**16.11**	**30.10**	2.08
Fine-tuned MLLMs on ProJudge-173k									
InternVL2.5-8B†	**84.50** +58.92	**45.39** +38.62	**53.97** +45.78	30.98 +30.98	22.14 +13.96	8.33 +6.58	**15.12** +13.95	3.22 +2.15	**10.41** +10.41
Qwen2.5-VL-3B†	81.29 +70.45	39.09 +38.25	46.36 +45.47	24.67 +24.67	24.54 +22.43	5.70 +5.27	12.69 +12.12	1.61 +1.61	4.16 +4.16
Qwen2-VL-7B†	83.72 +49.07	44.57 +43.88	52.29 +51.47	**31.62** +31.62	**27.36** +25.81	**11.40** +10.97	14.69 +11.98	**4.30** +1.30	6.25 +6.25

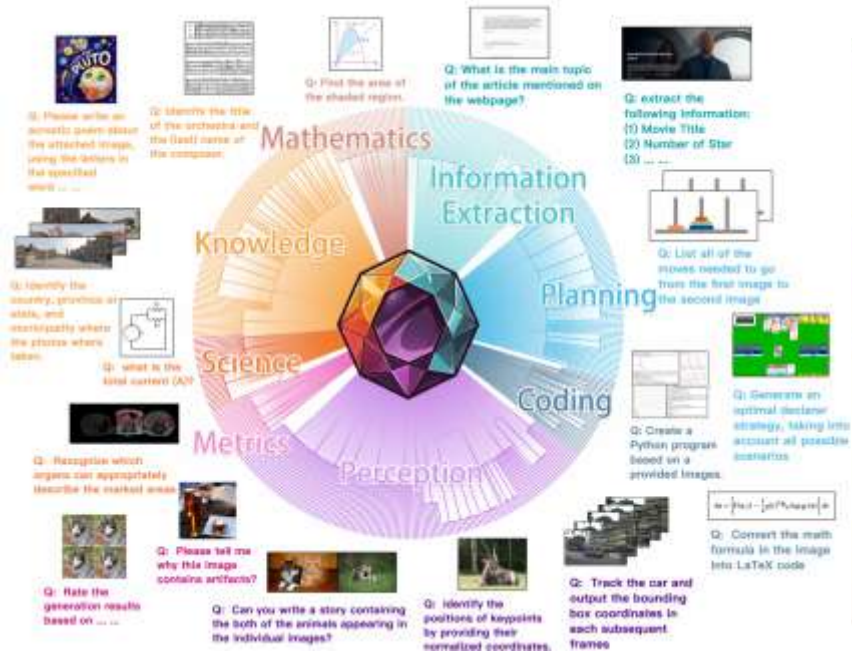
Step judge accuracy

ProJudge: A Multi-Modal Multi-Discipline Benchmark and Instruction-Tuning Dataset for MLLM-based Process Judges

General Evaluation

Vision-Language

Real-world application



MEGA-Bench

Output types

MEGA-Bench: Scaling Multimodal Evaluation to over 500 Real-World Tasks

General Evaluation

Interleaved Image-text Generation

General Evaluation

Interleaved Image-text Generation

Method	Human Evaluation				GPT Evaluation				IntJudge Evaluation			
	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)
Human	83.28%	86.03%	68.17%	78.55%	82.49%	82.69%	82.03%	82.43%	87.46%	91.49%	75.49%	84.23%
GPT-4o+DALL-E3	78.42%	81.39%	65.21%	75.15%	85.70%	85.99%	85.58%	85.82%	85.02%	86.92%	72.22%	80.68%
Gemini1.5+Flux	65.57%	65.82%	49.31%	61.85%	71.75%	71.76%	71.12%	71.56%	68.30%	69.73%	54.47%	65.41%
SEED-X	51.98%	49.49%	34.70%	49.65%	54.82%	55.12%	54.11%	55.03%	49.86%	49.58%	33.57%	49.72%
Anole	51.90%	52.17%	36.46%	51.52%	53.36%	53.13%	52.58%	53.10%	53.42%	52.04%	33.92%	51.33%
SEED-LLaMA	44.30%	42.12%	29.11%	44.56%	40.96%	40.87%	40.46%	40.96%	50.13%	47.71%	31.57%	48.48%
Emu2	40.89%	37.07%	23.42%	41.84%	41.72%	41.63%	40.58%	41.85%	36.28%	33.79%	21.87%	39.51%
Show-o	36.28%	34.02%	21.63%	39.84%	30.77%	30.22%	29.61%	30.62%	31.49%	21.08%	12.48%	32.87%
NExT-GPT	33.67%	26.93%	17.09%	35.36%	22.61%	22.39%	22.11%	22.74%	30.96%	21.70%	13.36%	32.58%
MiniGPT-5	30.69%	26.72%	17.11%	35.09%	28.64%	28.37%	28.02%	28.64%	24.47%	15.46%	9.91%	27.85%
GILL	25.80%	19.57%	12.71%	30.23%	30.55%	30.24%	29.65%	30.62%	24.87%	19.72%	12.82%	30.32%

Table 2. Comparison of model win rates evaluated by human, GPT-4o, and our IntJudge under FDT and different tie metrics. FDT: Force Dividing Tie metric. w/o Tie: Non-tie case. w/ Tie (0) and w/ Tie Tie (.5): Count a tie as 0 and 0.5 wins for a model in a battle, respectively.

Evaluator	FDT				w/ Tie				w/o Tie			
	Average	Seen	Unseen	HM	Average	Seen	Unseen	HM	Average	Seen	Unseen	HM
Random	49.83%	49.86%	49.79%	49.83%	32.60%	32.03%	33.18%	32.60%	50.00%	48.36%	51.89%	50.06%
GPT-4o	71.08%	73.33%	68.77%	70.98%	51.93%	54.95%	48.82%	51.70%	74.58%	77.54%	71.43%	74.36%
InternLMX2.5-7B	56.81%	55.73%	57.92%	56.81%	40.26%	40.19%	40.33%	40.26%	61.05%	61.21%	60.97%	61.09%
Qwen2-VL-7B	61.61%	61.59%	61.63%	61.61%	32.81%	31.16%	34.50%	32.75%	80.77%	81.15%	80.23%	80.69%
IntJudge-7B (Ours)	**82.42%**	**84.05%**	**80.75%**	**82.37%**	**66.45%**	**69.02%**	**63.80%**	**66.31%**	**91.11%**	**92.38%**	**89.55%**	**90.94%**

Table 3. Agreement rate between different MLLM-based judges and human judgments in different metrics. HM: Harmonic Mean.

General Evaluation

Interleaved Image-text Generation

Method	Human Evaluation				GPT Evaluation				IntJudge Evaluati		
	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)	FDT	w/o Tie	w/ Tie (0
Human	83.28%	86.03%	68.17%	78.55%	82.49%	82.69%	82.03%	82.43%	87.46%	91.49%	75.49%
GPT-4o+DALL-E3	78.42%	81.39%	65.21%	75.15%	85.70%	85.99%	85.58%	85.82%	85.02%	86.92%	72.22%
Gemini1.5+Flux	65.57%	65.82%	49.31%	61.85%	71.75%	71.76%	71.12%	71.56%	68.30%	69.73%	54.47%
SEED-X	51.98%	49.49%	34.70%	49.65%	54.82%	55.12%	54.11%	55.03%	49.86%	49.58%	33.57%
Anole	51.90%	52.17%	36.46%	51.52%	53.36%	53.13%	52.58%	53.10%	53.42%	52.04%	33.92%
SEED-LLaMA	44.30%	42.12%	29.11%	44.56%	40.96%	40.87%	40.46%	40.96%	50.13%	47.71%	31.57%
Emu2	40.89%	37.07%	23.42%	41.84%	41.72%	41.63%	40.58%	41.85%	36.28%	33.79%	21.87%
Show-o	36.28%	34.02%	21.63%	39.84%	30.77%	30.22%	29.61%	30.62%	31.49%	21.08%	12.48%
NExT-GPT	33.67%	26.93%	17.09%	35.36%	22.61%	22.39%	22.11%	22.74%	30.96%	21.70%	13.36%
MiniGPT-5	30.69%	26.72%	17.11%	35.09%	28.64%	28.37%	28.02%	28.64%	24.47%	15.46%	9.91%
GILL	25.80%	19.57%	12.71%	30.23%	30.55%	30.24%	29.65%	30.62%	24.87%	19.72%	12.82%

Table 2. Comparison of model win rates evaluated by human, GPT-4o, and our IntJudge under FDT and different tie metrics. F
Dividing Tie metric. w/o Tie: Non-tie case. w/ Tie (0) and w/ Tie Tie (.5): Count a tie as 0 and 0.5 wins for a model in a battle, re

Evaluator	FDT				w/ Tie				w/o Tie		
	Average	Seen	Unseen	HM	Average	Seen	Unseen	HM	Average	Seen	Unseen
Random	49.83%	49.86%	49.79%	49.83%	32.60%	32.03%	33.18%	32.60%	50.00%	48.36%	51.89%
GPT-4o	71.08%	73.33%	68.77%	70.98%	51.93%	54.95%	48.82%	51.70%	74.58%	77.54%	71.43%
InternLMX2.5-7B	56.81%	55.73%	57.92%	56.81%	40.26%	40.19%	40.33%	40.26%	61.05%	61.21%	60.97%
Qwen2-VL-7B	61.61%	61.59%	61.63%	61.61%	32.81%	31.16%	34.50%	32.75%	80.77%	81.15%	80.23%
IntJudge-7B (Ours)	**82.42%**	**84.05%**	**80.75%**	**82.37%**	**66.45%**	**69.02%**	**63.80%**	**66.31%**	**91.11%**	**92.38%**	**89.55%**

Table 3. Agreement rate between different MLLM-based judges and human judgments in different metrics. HM: Harmonic M

Friday, June 13

7:00 - 18:00 **Registration / Badge Pickup** (Summit Lobby)
7:00 - 18:00 **Registration / Badge Pickup** (ExHall Concourse)
7:00 - 17:00 **Press Room** (203 B)
7:00 - 17:00 **Mother's Room** (Level 1 near Room 101 and on Level 3 near Exhibit Hall D)
7:00 - 17:00 **Prayer or Quiet Room** (203 A)
7:00 - 9:00 **Breakfast** (ExHall C)
8:00 - 8:30 **Poster Setup** (ExHall D)
8:30 - 9:00 **Welcome & Awards** (Karl Dean Ballroom)

9:00 - 10:15 **Oral Session 1A: Image and Video Synthesis** (Karl Dean Ballroom)

Award candidate paper

1 Motion Prompting: Controlling Video Generation with Motion Trajectories, *Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, Deqing Sun*

2 Go-with-the-Flow: Motion-Controllable Video Diffusion Models Using Real-Time Warped Noise, *Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, Ning Yu*

3 LookingGlass: Generative Anamorphoses via Laplacian Pyramid Warping, *Pascal Chang, Sergio Sancho, Jingwei Tang, Markus Gross, Vinicius Azevedo*

4 Alias-Free Latent Diffusion Models: Improving Fractional Shift Equivariance of Diffusion Latent Space, *Yifan Zhou, Zeqi Xiao, Shuai Yang, Xingang Pan*

5 RandAR: Decoder-only Autoregressive Visual Generation in Random Orders, *Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T. Freeman, Yu-Xiong Wang*

9:00 - 10:15 **Oral Session 1B: Interpretability and Evaluation** (ExHall A2)

1 OpenING: A Comprehensive Benchmark for Judging Open-ended Interleaved Image-Text Generation, *Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, Lirui Zhao, Shuo Liu, Tianhua Li, Yuxuan Xie, Xiaojun Chang, Yu Qiao, Wenqi Shao, Kaipeng Zhang*

2 [illegible] Transformer Attributions, *Farnoush Mehri, Mahdieh Soleymani Baghshah, Mohammad Taher Pilehvar*

3 Do We Always Need the Simplicity Bias? Looking for Optimal Inductive Biases in the Wild, *Damien Teney, Liangze Jiang, Florin Gogianu, Ehsan Abbasnejad*

4 Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models, *Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, Aniruddha Kembhavi*

5 Rethinking Vision-Language Model in Face Forensics: Multi-Modal Interpretable Forged Face Detector, *Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, Xiaoming Liu*

General Evaluation

Unified multimodal understanding and generation

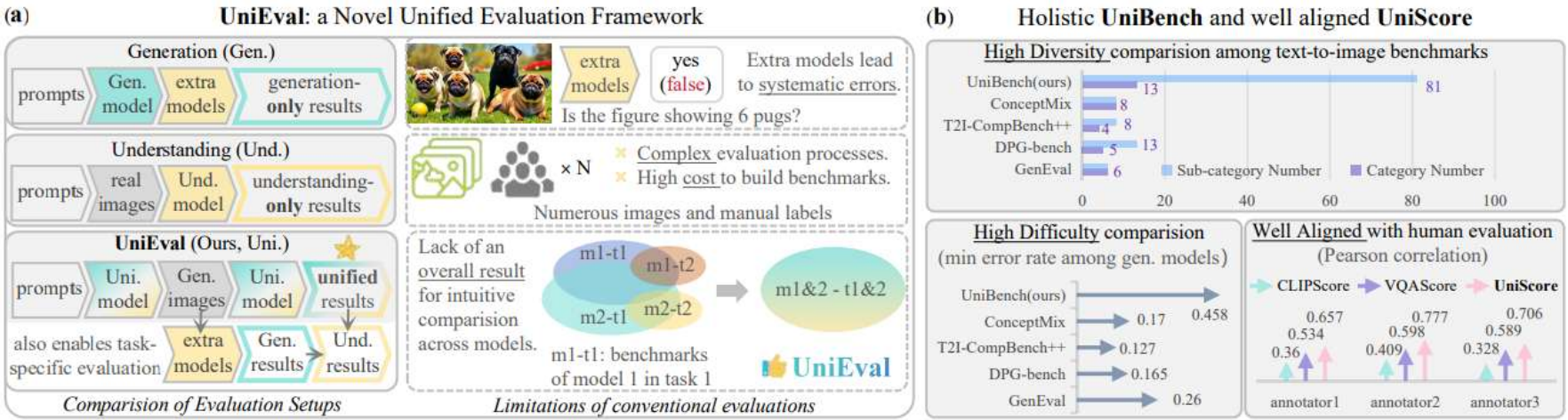
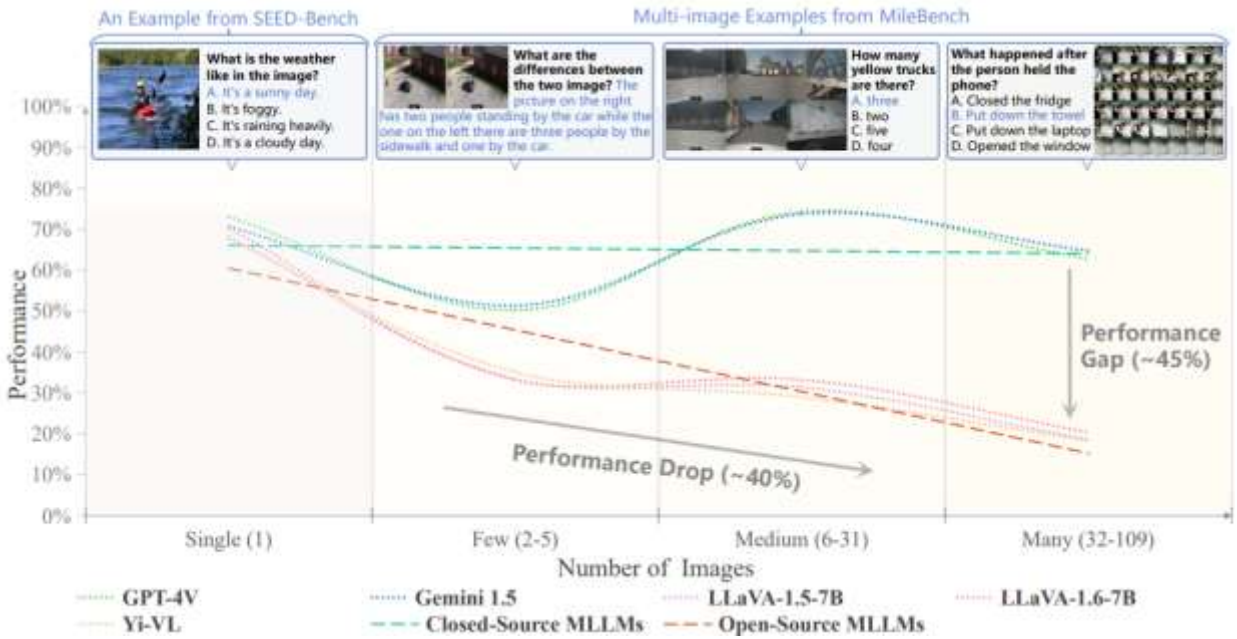


Table of Content

Specialized Evaluation

Multi-images

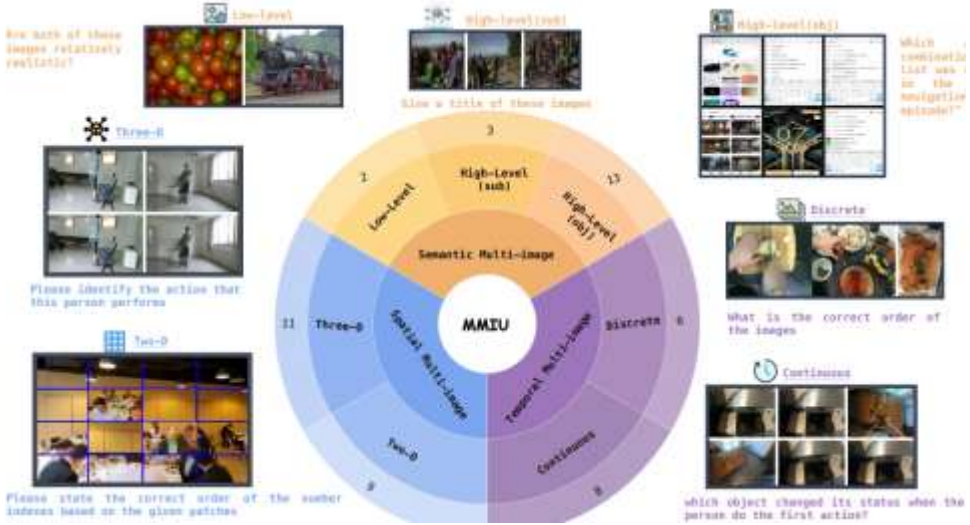


Task definition

MILEBENCH: Benchmarking MLLMs in Long Context

Specialized Evaluation

Multi-images



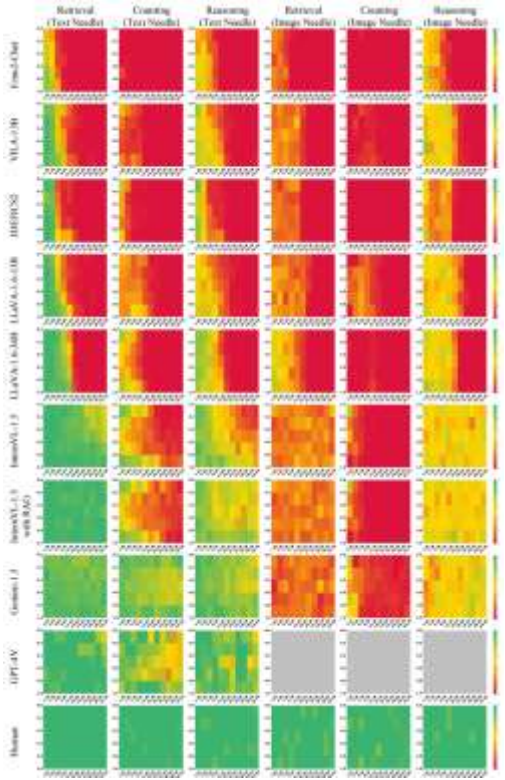
Task definition

Examples

Specialized Evaluation

Long context

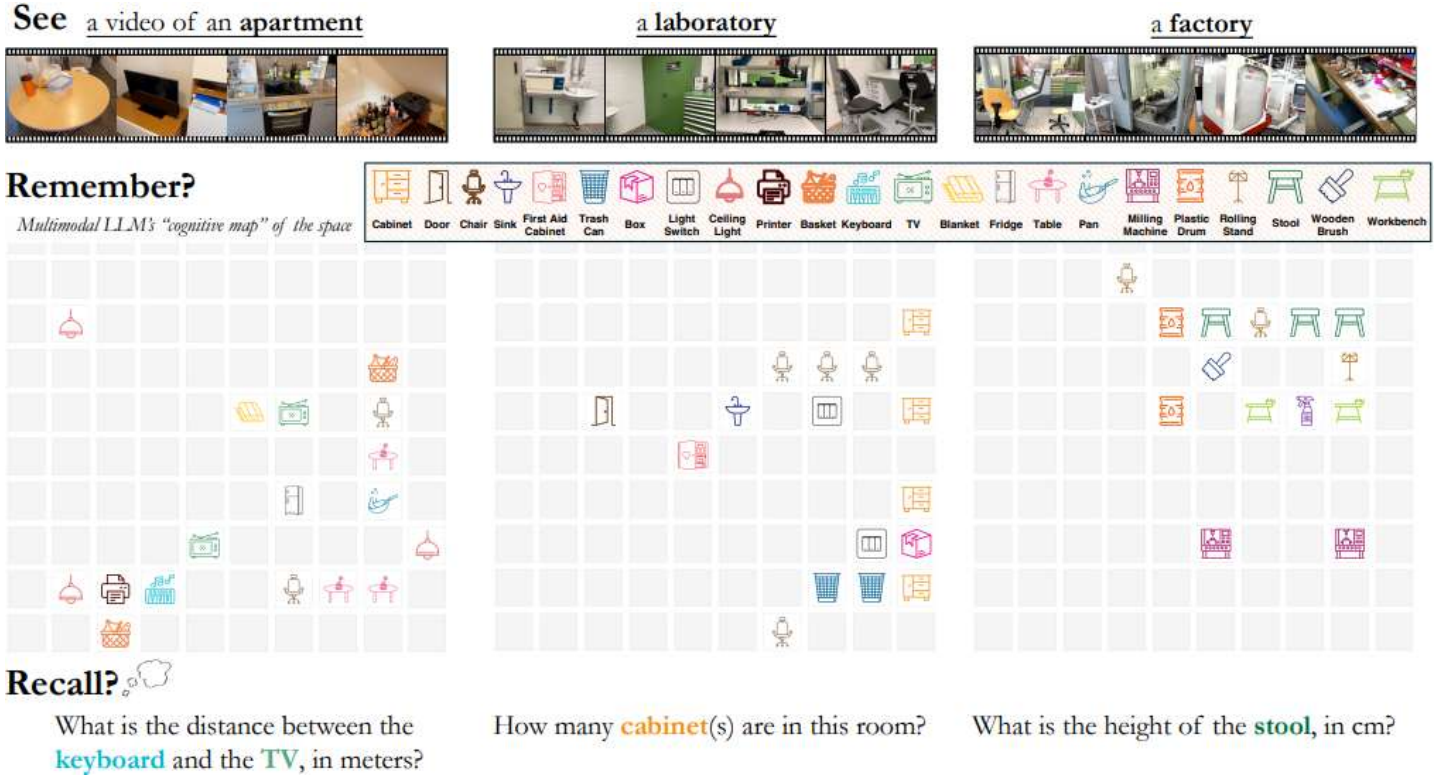
(a) Retrieval-Text-Needle (b) Counting-Text-Needle (c) Reasoning-Text-Needle (d) Retrieval-Image-Needle (e) Counting-Image-Needle (f) Reasoning-Image-Needle



Needle in a multimodal haystack

Specialized Evaluation

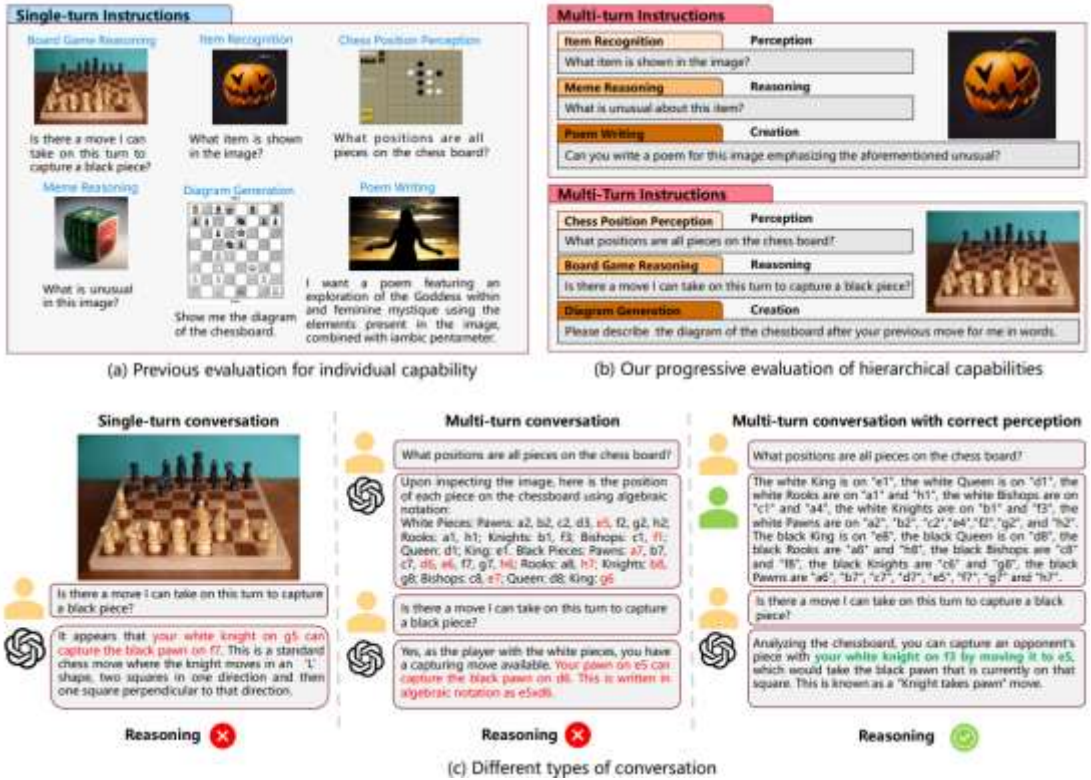
Visual-spatial intelligence



Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces

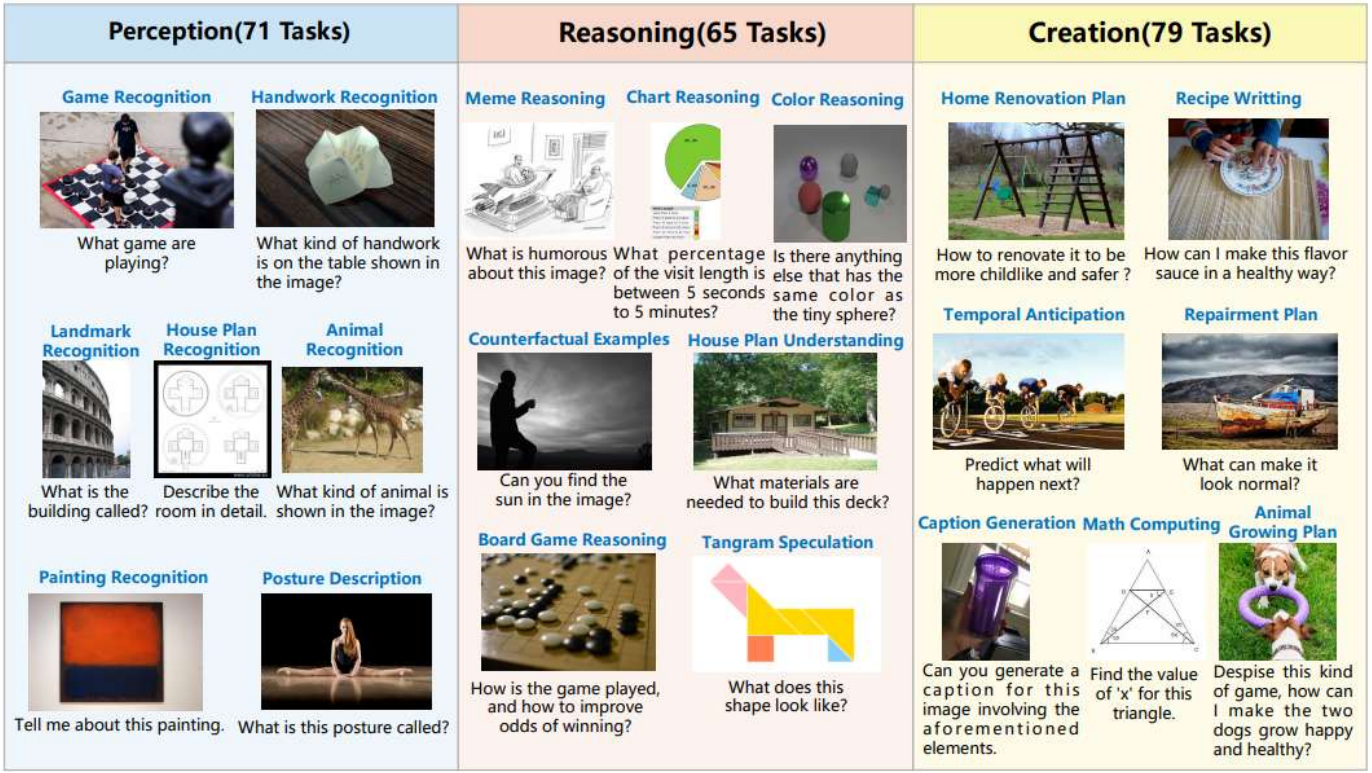
Specialized Evaluation

Multi-turn conversation



Specialized Evaluation

Multi-turn conversation

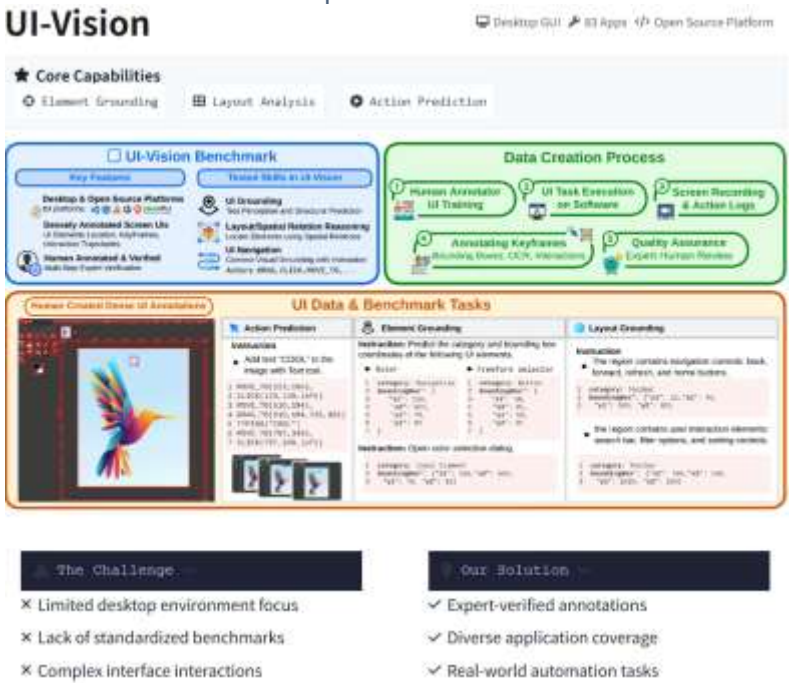


ConvBench: A Multi-Turn Conversation Evaluation Benchmark with Hierarchical Capability for Large Vision-Language Models

Specialized Evaluation

GUI

Perception and Interaction



Cross-app Interaction

UI-Vision: A Desktop-centric GUI Benchmark for Visual Perception and Interaction
GUI Odyssey: A Comprehensive Dataset for Cross-App GUI Navigation on Mobile Devices.

Specialized Evaluation

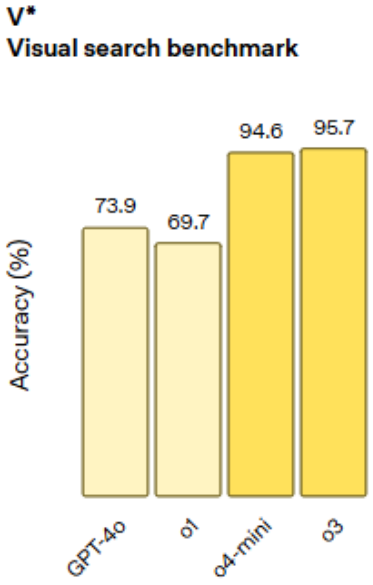
Thinking with images

which stop is this, and what is the frequency of the bus at this stop? search the internet if needed!

Analyzed image

Specialized Evaluation

Thinking with images



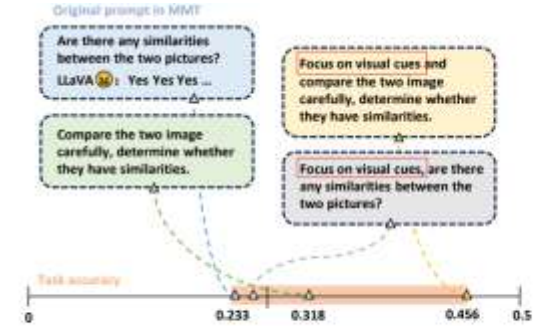
V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs

Table of Content

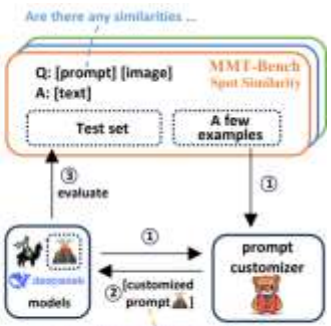
Evaluation Issues

Prompt sensitivity

Model	# IT-Data		BLINK			MMB			SeedB			TMA			MMMU			Overall
			S	ID	OOD	S	ID	OOD	S	ID	OOD	S	ID	OOD	S	ID	OOD	
7B / 8B Models																		
LLaVA-1.5-7B	665K	Avg.	43.67	37.26	39.72	70.00	68.55	69.20	60.67	57.35	56.16	37.00	42.94	42.60	36.67	37.19	36.16	48.94
		Max-Min	8.00	15.00	15.00	18.00	16.00	9.00	5.00	18.00	16.00	14.00	26.00	18.00	4.00	14.00	13.00	13.93
LLaVA-Next-7B	760k	Avg.	45.33	38.92	37.64	62.67	60.43	58.08	70.00	63.29	62.16	50.67	44.06	44.60	33.67	31.51	29.24	48.95
		Max-Min	7.00	16.00	12.00	10.00	20.00	9.00	2.00	18.00	10.00	16.00	17.00	11.00	2.00	18.00	8.00	11.73
Qwen-VL-7B	50M	Avg.	36.00	34.44	34.04	50.67	47.51	47.16	30.67	29.66	28.80	31.67	29.76	30.76	25.67	28.06	28.40	34.18
		Max-Min	4.00	9.00	8.00	3.00	11.00	11.00	10.00	17.00	12.00	9.00	19.00	14.00	2.00	17.09	11.00	10.47
Qwen-VL-Chat-7B	50M	Avg.	31.67	40.09	40.28	62.67	74.02	75.16	36.00	58.77	58.32	39.33	51.39	51.48	39.00	36.49	36.36	50.08
		Max-Min	4.00	21.00	20.00	3.00	17.00	14.00	2.00	20.00	13.00	8.00	17.00	12.00	10.00	16.00	10.00	12.47
IDEFICS2-8B	1.8M	Avg.	39.33	48.97	46.36	71.00	70.75	70.28	43.33	53.36	54.04	36.00	47.40	46.20	29.33	27.48	28.36	47.28
		Max-Min	4.00	17.00	10.00	6.00	11.00	9.00	7.00	16.00	17.00	8.00	20.00	17.00	3.00	14.00	11.00	11.33
LLaVA-1.5-7B w / 5K templates	665K	Avg.	46.33	43.29	45.44	68.67	71.66	73.20	64.33	65.33	66.16	52.00	51.79	52.64	39.33	37.48	37.32	54.18
		Max-Min	5.00	13.00	2.59	10.00	12.00	8.00	3.00	11.00	6.00	4.00	22.00	10.00	9.00	11.00	6.00	8.84
13B Models																		
LLaVA-1.5-13B	665K	Avg.	40.00	38.75	41.20	72.33	73.42	73.24	67.00	68.87	66.92	54.00	52.36	52.24	37.33	39.00	37.20	54.13
		Max-Min	7.00	16.00	14.00	3.00	12.00	6.00	5.00	9.00	10.00	8.00	16.00	15.00	6.00	16.00	10.00	10.20
LLaVA-Next-13B	760k	Avg.	39.67	40.72	38.16	64.67	65.47	63.40	68.33	68.78	66.88	54.67	51.33	47.68	31.00	33.23	33.80	51.06
		Max-Min	1.00	15.00	13.00	9.00	19.00	15.00	1.00	12.00	11.00	5.00	21.00	14.00	2.00	21.00	10.00	11.27
LLaVA-1.5-13B w / 100 templates	665K	Avg.	37.67	41.22	42.68	70.00	73.08	73.68	69.33	69.37	69.08	51.33	50.49	50.68	39.67	43.23	44.40	55.21
		Max-Min	14.00	15.00	8.00	12.00	10.00	10.00	3.00	7.00	5.00	1.00	12.00	5.00	7.00	15.00	15.00	9.27



(a) Example of prompt sensitivity in multi-modal benchmark.

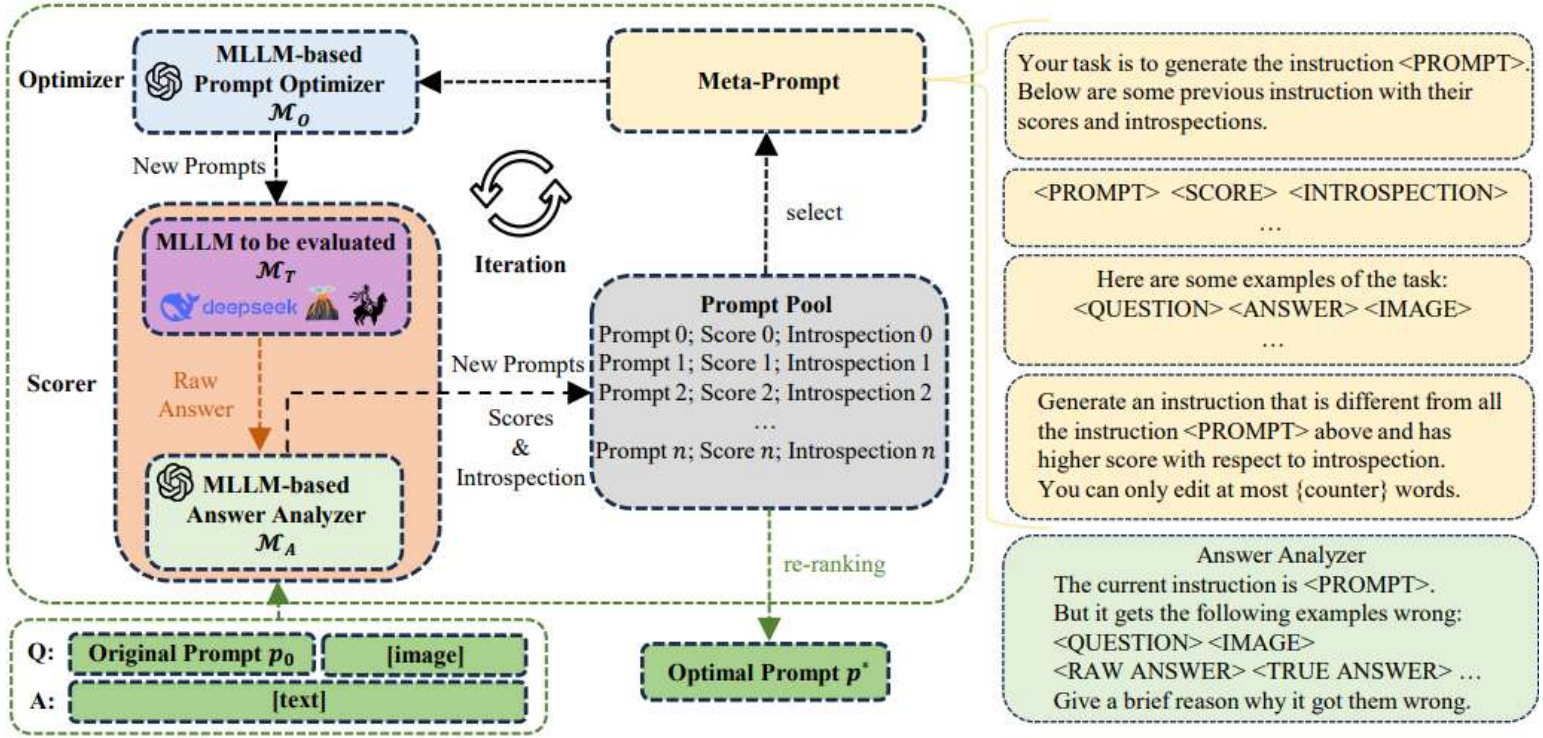


(b) Framework of TP-Eval.

Investigating the Scaling Effect of Instruction Templates for Training Multimodal Language Model
TP-Eval: Tap Multimodal LLMs' Potential in Evaluation by Customizing Prompts

Evaluation Issues

Prompt sensitivity



TP-Eval: Tap Multimodal LLMs' Potential in Evaluation by Customizing Prompts

Evaluation Issues

Prompt sensitivity

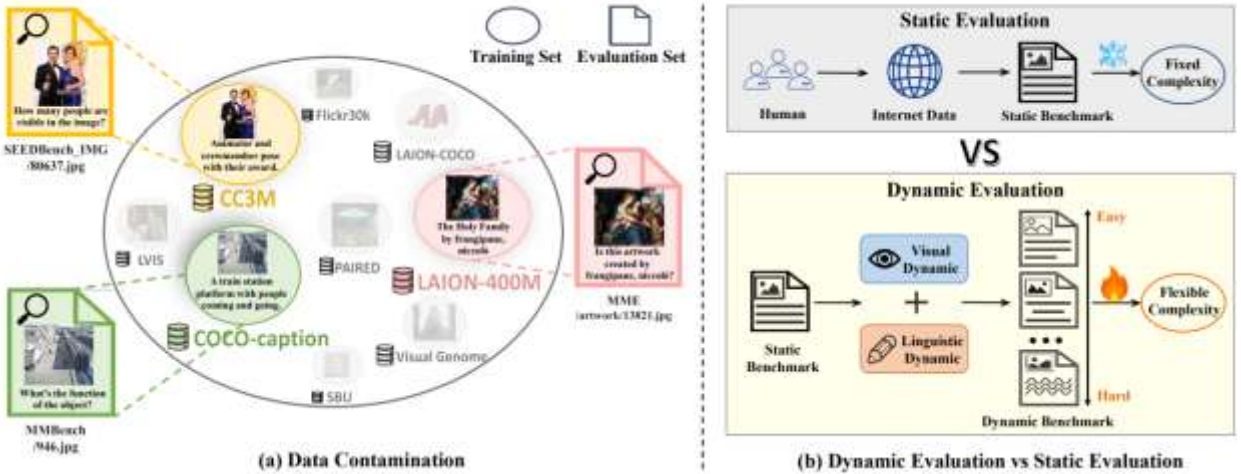
Model	Original Score	TP-Eval Score
LLaVA-1.5-7B	50.4	54.4
DeepSeek-VL-7B	55.2	57.3
Mini-InternVL-Chat-4B-V1-5	54.6	56.9

Task name	Original prompt	Zero-shot	Few-shot
helmet anomaly detection	0.65	0.86	0.92
artwork emotion recognition	0.3	0.33	0.41
spot similarity	0.23	0.42	0.52

TP-Eval: Tap Multimodal LLMs' Potential in Evaluation by Customizing Prompts

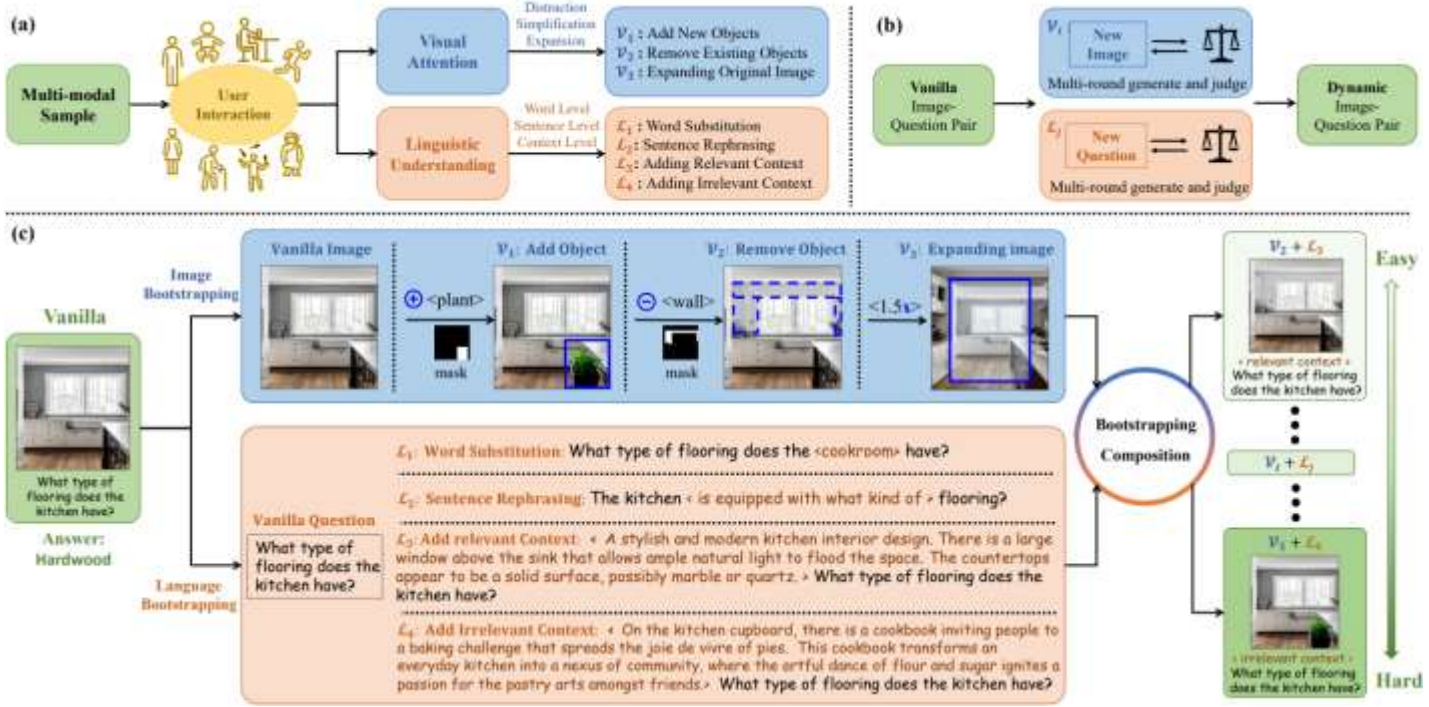
Evaluation Issues

Dynamic evaluation



Evaluation Issues

Dynamic evaluation



Dynamic Multimodal Evaluation with Flexible Complexity by Vision-Language Bootstrapping

Conclusion

- Unified understanding and generation evaluation and process evaluation are promising
- More and more specialized capability evaluation are coming along with model capability improving

(e.g., visual tool-use in o3)

- Data leakage and prompt sensitivity during evaluation

Thanks!

Any questions?