☆ Part-III

Vision-Language Capability Evaluation

Kaipeng Zhang

Researcher *Shanghai Al Lab / Shanghai Innovation Institute*

https://kpzhang93.github.io/



* Table of Content

+ General Evaluation

- × Vision-centric
- × Vision-language
- × Interleaved Image-text Generation

+ Specialized Evaluation

- × Long-context and multi-image
- × Visual-spatial intelligence
- × Multi-turn conversation
- × GUI
- × Thinking with images

+ Evaluation Issues

- × Prompt sensitivity
- × Dynamic evaluation

* Table of Content

+ General Evaluation

- × Vision-centric
- × Vision-language
- × Interleaved Image-text Generation
- + Specialized Evaluation
 - × Long-context and multi-image
 - × Visual-spatial intelligence
 - × Multi-turn conversation
 - × GUI
 - × Thinking with images

+ Evaluation Issues

- × Prompt sensitivity
- × Dynamic evaluation

Vision-centric evaluation

Computer \	/ision				2D Classification
30 55	71 bench	marks •	1617 tas	ks	Caper Solution A differentiering A filling of the
		Det Lose Bandan 17 secondos	111 B 111 B 111 Basenate 1 Maartaat 1 Maartaat	The next leading	· Invation
s must remain	ation				in franksie Rosenikas
Arasti Aprender A di antanti Di ana atrad	New York Specialize 1 control on	Propint Imposibilities Int cases with come		For the second s	 ker af 10 kerke 10-shot image gene
1 five of 10 leads					
		Notices In Processor Transmitter		Initiation Generation Information Theorem	Object Detection
Representation Le	aming				Contactor Contactor A fragmentation
in Therese	Designed in restore	Cost References Internet	R Tanana	5.7 E recet Emissing	 beral de teste Rebrieval
1 Secul Crash					and American Constraint Constraint (constraint)

Classification					image Claisifica
1 325.	A location	Million States	(iii) Anal	Transform	1
an of the local					a manufacture
wification					2D Object Deter
C Gallate (Connell)	(j); 25. 	Elli Bankan Alianakan Alianakan	Multi-Say Gardhadan o Linnanni Miranni Miran	n dia managana di kana	Martin State
er of it tests					Table of solider of the
shot image ger	veration				Reinforcement I
-	Lange of the second	Maria Inge Salar Salar Inge Salar Salar Sa	Concerning	(1) because	Antipation and a second
an of the laster					 Annoid Chester
ject Detection					Image Generatio
<u> </u>	2016art Reaction a 1 Theorem	A resolution	A succession	The Sector Secto	
or of the second second					Florencies Education
trieval					Contraint results
and Annual	The Arrive	I an coprusing	Telebaland		A Constant
Plant of the	100,000,000,000	A search into	1.000		a pine of 14 meter

Image Classification

a transm	A constant		Rectification Contraction of Procession of procession	all antipath all formation
a man of 24 mater				
2D Object Detection	an			
	The boston	10 month	Contraction Contra	(ii) <u>service</u>
· for eff () take				
Reinforcement Les	writing (RL)			
ini inima ana	Int Display	International Statement	In the second se	(1), Andres a linear
A Real Property lies				
Image Generation				
Constant Constant Constant Constant		III Advantage	insp Sopering	The second
+ 3++ al. (1 miles				
Domain Adaptatio	0			
	insected Drain Malitim Comments	1910 Janat Booldada A Disease B Sametra		

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

Seneral 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

Vision-centric evaluation

Rank	Method	a Param (B)	Language Model	 Vision Model	MMT-Bench_VAL	- ^î
1	InternVL2-Llams3-768	76	Llama-3-708-Instruct	InternViT-68	67.6	
2	GPT-4c (0513, dotail-high)				67.3	
3	InternVL2-40B	40	Nous-Hermes-2-Y1-348	InternViT+68	66.1	
4	GPT-4o (0513, detail-low)				65.6	
5	Y1-V1sion				62.9	
6	Ovis1.5-Genna2-98	11.4	Genna2-98-It	SigLIP-400M	62.8	
7	Claude3.5-Sonnet				61.8	
8	GPT-4o-mini (0718, detail-high)				61.2	
9	OmChat-v2.0-138	13	Qwen2-7B	InternViT-68	61	
10	MiniCPM-V-2.6	8	Qwen2-78	SigLIP-400M	60.8	
11	Ovis1.5-Llama3-88	8	Llama-3-88-Instruct	SigLIP-400M	60.7	
12	InternVL2-268	26	InternLM2-208	InternViT-6B	60.4	

Vision-centric evaluation



General Evaluation

Vision-centric evaluation

Convert to VQA format



O latta cal?

Yes, that's a call Specifically, it looks like a very young kitter - appeare to be an orange or ginger tably kitten with beautiful blue eyes. The kitten looks quite small and fluffy, probably just a few weeks old based on its size and factal farmers.

*

C (D Q Reby -

Piepty to Claude	
	and contract of the second

Most MLLMs output text only



Some conversions are difficult



Difficulty control is challenging

Seneral Evaluation

Vision-centric evaluation

Use the original task format

More and more models can output images





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

Vision-centric evaluation

Visual Puzzle



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



Vision-Language Multi-discipline



MMMU (college-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

MDK12 (K12-level)

Vision-Language

Multi-discipline



MDK12-Bench: A Multi-Discipline Benchmark for Evaluating Reasoning in Multimodal Large Language Models

Vision-Language

Multi-discipline



MDK12-Bench: A Multi-Discipline Benchmark for Evaluating Reasoning in Multimodal Large Language Models

Vision-Language Multi-discipline - Process-level evaluation





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

ProJudge

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

Vision-Language Multi-discipline - Process-level evaluation







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

ProJudge

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

Vision-Language Multi-discipline - Process-level evaluation

	Step				Error	Types			
Model Name	Corr.	Overall	RE.	VI.	NC.	SC.	KE.	QU.	NS.
			Open-	source MLL	Ms				
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	о.	0.
Qwen2-VL-72B	77.93	33.87	40.11	1.49	39.91	30.70	10.27	4.83	2.08
			Propi	rietary MLL	Ms				
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
		Fi	ne-tuned MI	LMs on Pro	Judge-173k				
InternVL2.5-8B [†]	84.50	45.39+38.02	53.97	30.98	22.14+11.ml	8.33+0.58	15.12	3.22+215	10.41
Qwen2.5-VL-3B [†]	81.29+70.05	39.09+88.05	46.36-05-07	24.67-31.67	24.54+22.43	5.70+5.27	12.69-12.12	1.61-1.61	4.16+1.16
Qwen2-VL-7B [†]	83.72 - 0 07	44.57+01.00	52.29+51.07	31.62	27.36 25.81	11.40+10.97	14.69+11.98	4.30-1.30	6.25

Step judge accuracy

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

Vision-Language

Real-world application



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

Interleaved Image-text Generation



Interleaved Image-text Generation

		Human I	evaluation	1		GPT Ev	aluation	(IntJudge Evaluation				
Method	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)	FDT	w/o Tie	w/ Tic (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-40+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-40, 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		F	DT		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-40	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.

OpenING: A Comprehensive Benchmark for Judging Open-ended Interleaved Image-Text Generation

Interleaved Image-text Generation

		Human I	Evaluation	1		GPT Ev	aluation	IntJudge Evaluation				
Method	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 (6	
Human	83.28%	86.03%	68.17%	78.55%	82.49%	82.69%	82.03%	82.43%	87.46%	91.49%	75.49%	
GPT-40+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-40, 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		F	т			w/	Tie	w/o Tie			
	Average	Seen	Unseen	HM	Average	Seen	Unseen	HM	Average	Seen	Unsee
Random	49.83%	49.86%	49.79%	49.83%	32.60%	32.03%	33.18%	32.60%	50.00%	48.36%	51.899
GPT-40	71.08%	73.33%	68.77%	70.98%	51.93%	54.95%	48.82%	51.70%	74.58%	77.54%	71.439
InternLMX2.5-7B	56.81%	55.73%	57.92%	56.81%	40.26%	40.19%	40.33%	40.26%	61.05%	61.21%	60.979
Qwen2-VL-7B	61.61%	61.59%	61.63%	61.61%	32.81%	31.16%	34.50%	32.75%	80.77%	81.15%	80.239
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 N

OpenING: A Comprehensive Benchmark for Judging Open-ended Interleaved Image-Text Generation

Friday, June 13

7:0	0 - 18:00	Registration / Badge Pickup (Summit Lobby)
7:0	0-18:00	Registration / Badge Pickup (ExHall Concourse)
7:0	0-17:00	Press Room (203 B)
7:0	10 - 17:00	Mother's Room (Level 1 near Room 101 and on Level 3 near Eshibit Hall D)
7:0	0 - 17:00	Prover or Quiet Boom (203 A)
7:0	0-9:00	Breakfast (ExHall C)
	0 - 8:30	Poster Satury (ExHall D)
2.1	00.0	Walcome & Awards (Karl Desc Ballroom)
-	- 3.00	Helcome & Heleros (nan Dean bandon)
2.0	10 - 10:15	Oral Session 1A: Image and Video Synthesis (Karl Dean Ballroom)
1 2 3 4 4	22 - Awa Motion IP Insjectori Forrester Yuar Andrew C Go-with-I Using Re Weng Xii Yitong De Paul Deb Looking C Warping, Markus G Asias-Free Equivaria Shuai Yar Bandom	nd candidate paper rompting: Controlling Video Generation with Motion es. Daniel Geng, Charles Hermann, Junhwa Hur, Colis, Serena Zhang, Tabias Plaff, Talana Lopez-Guevara, a. Kichael Rubinstein, Chen Sur, Oliver Wang, Ywens, Deging Sun he-Flow: Motion-Eosthollable Video Diffusion Models ai-Time Warped Noise, Ayon Burgert, Yuancheng Xu, n, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, ing, Lingnio Li, Mohen Mousavi, Michael Ryco, werc, Ning Yu lass: Generative Anamorphoses via Laplacian Pyramid Pasical Chang, Sergio Suncho, Angwei Tang, ross, Wincias Azevedo Labart Diffusion Models: Improving Fractional Shift nee of Diffusion Latert Space, Yifan Zhou, Zegi Xiao, g, Xingang Pan Decoder-only Autoregressive Visual Generation in Orders, Zir Peng, Tenyuan Zhang, Fojun Luan, Yunee Mar,
	Heo Tan,	Kai Zhang, William T. Freeman, Yo-Xiong Wang
9:0	10 - 10:15	Oral Session 18: Interpretability and Evaluation (ExHall A2.)
	OpenING Interleave Jlajun Sol Hao Zhar Yuxuan X	: A Comprehensive Benchmark for Judging Open-ended Id Image-Text Generation, Penghei Zhou, Xiaopeng Peng, yg. Chuanhao Li, Zhaopen Xu, Ne Yang, Ziyao Guo, yg. Yuqi Lin, Yutei He, Linu Zhue, Shoo Liu, Tianhua Li, ie, Xiaojun Chang, Yu Qiao, Wengi Shao, Kalpeng Zhang
8	Transform	 Balancing Gradient How for University Better Vision her Attributions, Fandoun Mehri,
3	Mahdieh Do We Al Inductive Florin Go	Soleymani Bagtshah, Mohawmad Taher Pilehvar ways Need the Simplicity Bias? Looking for Optimal Biases in the Wild, Damien Teney, Liangze Nang, glanu, Ehsan Abbesnejad
4 22	Molmo ar Vision-Las Rohum Tri Niklas Mu Enin Brans Mark Yatz Favyon B Armavi Ci Aaron Sar Tanmay C Crystal N Oscar Ms Hunnanel	di PhiMa: Open Weigitti and Open Date for State-of-tee-At- rguage Models, Matt Dailis, Christopher Clark, Sangho Lee, Jaahi, Yuu Yang, Jie Sung Park, Molwimmadeazo Salehi, inninghot Kyle Lu, Luca Soldaini, Jesen Lu, Tahri Anderson, tono, Kiama Ehani, Huong Ngo, YenSung Chen, Ajay Patel, kar, Chris Galisaon-Barch, Andrew Head, Rose Hendrin, satami, El VanderBilt, Nathan Lambert, Yvonne Choa, Istada, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, and, Byran Bischell, Patel Walahi, Chris Newell, Piper Wolfen, and Sophie Latenschrift, Cartha Wolf, Cartasa Schoenick, end, Barya Kimina, Luca Wolfa, Nash A. Smith, end, Barya Kimina, Laca Wolfa, Nash A. Smith, end, Barya Kimina, Laca Wolfa, Andre A. Smith, et Hajahirzi, Ross Girabick, Al Farthard, Aniroddha Kembhaid
	Multi-Mo Xiufeng S	dal Interpretable Forged Face Detector, Xiao Guo, Iong, Yue Zhang, Xiaohong Liu, Xiaoming Liu

General Evaluation

Unified multimodal understanding and generation



Content :: Table of Content

+ General Evaluation

- × Vision-centric
- × Vision-Language
- × Interleaved Image-text Generation

+- Specialized Evaluation

- × Long-context and multi-image
- × Visual-spatial intelligence
- × Multi-turn conversation
- × GUI
- × Thinking with images

+ Evaluation Issues

- × Prompt sensitivity
- × Dynamic evaluation

Multi-images



MILEBENCH: Benchmarking MLLMs in Long Context

Specialized Evaluation

Multi-images



Task definition

Examples

Long context (Test Monday) (Thru Needle) (Test Needla) (Image Needle) (Image Needle) (Image Needle A.C. A.C. A.S. The Over regiment to the Department of A prop. Barranc BreakAD Tarts on the Device with VPF Belginers pick up the picers offset floreds. The pain Can we put talk about the fuminatar, because however, because the second second second to come to the Huge Hore 11 contentant lighterth Medda has been served bring machinest and property in equal proposition. Some description or relation in served My driver inches[at my very supprised, had by whether the second s now to had undynetical our rectinged arothing said Region A property new barry tant. You would have much may back to the collings. He know that I just on one of the many solf users that where up and The Brood All using how parties depending on the of starge wary as bosses, collegest entered workspace. The second set of a farmer, The new manife diamong tool fets trop view and Sample manufacture and and a feature of the second state I'm about as quark in the London Kyy, at a unitator with a waviden leg. Reprin H. and There are Directly Wandyeds that will work a financial pair water proof workshow of a second Sympose learch of their measurement parts are pro-Another responses use of 2001 to in predicted A, has bade ton: Yes, there was so come 8 10 aging offects. 1MI are include an aging weather The key have to visit reprincing the attributer well. based on self lengton. This is inservice dependences. He possile any over good in the completion. In The occurs through have whereast at loant 27 lives in the Balgian province of Liege, horse to Pepareter or Paren Assess by there I man of and other towns in the Meane Saxie. The Drift Centernetwork for Brunt AD lists you import and When I want't daughing from the uniting in seried work with generativy them Sailid Works or other praga . Region C. Justit to H. remains dry. And mis afpartness we bealled out in a catamacan and inage CAD programs. You use animale stude interestly, to flow problem, bottor that a boat large. Of Givne the senand incase, which has the forever right corner nationality, can you tell which of the following images in the missing part? Q: Which of the following images approve in a cortain image of the above dominant? Q: Please help the little pengain collect the Q. Please help me puller: the number of 🥨 to you'l though in the observe documents, and the sumber of banana, ... Assessed in the format like Q: What does the seastery anvel? baa.d. O: Which region stays dry thering the electro? A. Bernette A. Begins C. (e) Counting-Image-Needle (il) Retrieval-Image-Needle (f) Reasoning-Image-Needle (a) Retrieval-Text-Needle (b) Counting-Text-Needle (c) Reasoning-Text-Needle

Retrieval.

Contrary

Residence

Barrenal

Coming

No. of Concession, Name

Visual-spatial intelligence



What is the distance between the **keyboard** and the **TV**, in meters?

How many cabinet(s) are in this room?

What is the height of the stool, in cm?

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

Multi-turn conversation







(c) Different types of conversation

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

Multi-turn conversation



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

GUI

Perception and Interaction **UI-Vision** 🖵 Desktop Gull 🏓 83 Apps 4/2 Open Source Platform * Core Capabilities O ilament frounding I Lavout Analysia O Action Prediction UI-Vision Benchmark **Data Creation Process** Harton Arabetator UI Task Execution Scient Recording Desition & Open Source Platforms 8 Sh Desarating ill Training Aution Loga on Scholars explorations and that the Operation Densaty Annohated Science, Uni-Laysia Tapata Heatin Resources Entertainen 1 partellari, 6 da Propinsia. Annutating Keyharan Loopet Ethnisette using Tannal Balances Dustky Assurance U. Non-testion tini into Discour Children Making AGent Human Neules Names Sciences & Verified Annual state, children wanty on **UI Data & Benchmark Tasks** S. Denni Genantry R. Action Prediction Layest Dreaming testracilistic Predict the category and townilling hos-VANADAR Martin actions coordinates of the billingted Lit alertants. · AM INT COR. IN THE The regist contains margatise commit back forward refresh, and heres furthers. integer setti Tour cost. Aster · freehen inlation sense and total tensor 11 PORCETTA, 1984, 1984 1 Ann W. and anot. 701100.0041 10441, 705 0 etc. 1044, 1044, ad-THERE AND ! THEN." the legal contests and menalital demants elpel, 793 (97, 244) search tax. Hise systems, and setting contacts. ARRITED AND LOTT man actives: Open only presented status AND DESCRIPTION OF ANY ADDRESS OF A DESCRIPTION OF A DESC month specer : d'art' los, "ed" i son The Challenge Our Holution

- × Limited desktop environment focus
- × Lack of standardized benchmarks
- × Complex interface interactions

- Expert-verified annotations
 Diverse application coverage
- Real-world automation tasks

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.

Cross-app Interaction



31

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



Route to the second



Analyzad maga

Thinking with images

V*





Carle of Content

+ General Evaluation

- × Vision-centric
- × Vision-Language
- × Interleaved Image-text Generation

+ Specialized Evaluation

- × Long-context and multi-image
- × Visual-spatial intelligence
- × Multi-turn conversation
- × GUI
- × Thinking with images

+ Evaluation Issues

- × Prompt sensitivity
- × Dynamic evaluation

Prompt sensitivity

Model	att.Date			BLINK	È		MMB			SeedB	Rbee		TMA		MMMU			Overall
and a second sec			. 11	ID.	000	5	ID	COD	- 5	iD	COD	- 5	ID	000	5	ID.	000	
						787	Hi Mad	élik										
11.00.15.70	1000	Nrg.	0.67	17.26	38,72	20.09	68,55	(9.2)	-60.67	.57.35	36.10	37.00	42.94	42.00	30.47	37.19	30,16	48.94
PERSONAL PARTY	10000	Man-Min	8.00	13.00	15,00	18.00	16.00	0.00	310	18.00	16:00	14.00	36.00	18.00	4,00	34.00	13.00	12.93
Dr. Alta March 70	200.00	Avg.	45.33	脉蛇	37.64	校校	60.43	58.08	70.00	43.29	07.10	50.67	64.05	44.60	13.67	角角	29.24	48,95
TT9 STORESCAR	1.0008	Max-Min	7.00	16.00	1200	10.00	20000	9.00	2.00	18.00	10.00	16.00	12.00	11.00	2.00	18.00	8.00	11.79
400	1000.0	Avg.	36.00	34.44	34.04	52.12	47.51	47.38	30.67	29.86	28.80	31.67	29.76	30.76	25.67	29.06	25,40	34.18
Contraction in the second	3.04	Man-Mill	4.00	9.00	8.00	3.00	11.00	13.00	10.00	17.00	12:00	9.00	19.00	14.00	1.00	17.09	11:00	10.47
0	iline a	Avg.	31.67	41.09	40.28	12.67	74,01	75.30	56.00	\$8,77	56.32	39.33	51.35	51.48	39.00	36.49	31.50	50.08
Control Contraction	2004	Max-Min	4.00	21.08	2000	3.00	12,00	14.00	2.00	31.00	13/00	8.00	17.00	12.00	10.00	38.00	10.00	12.47
Internet and	1.854	Aviz.	39.33	43,97	46.36	T1.00	70.75	78.28	43.30	51.36	54.04	36.00	47,40	46.39	29.39	27.48	28.36	47.28
11.42FR.52-88		Max-Mirr	4.00	17,00	10.00	6.00	11.00	9.00	7.00	36.00	17.00	8.00	2000	17.00	3.00	14.00	11.00	11.39
	12210	Avg.	-	61.29	45.44	18.17	TL00	73.20	68.33	63,33	66,00	12.00	10,79	52.64	19.35	22.48	37.32	74,10
trace-ra-towy or eathered	101.18.	Max-Min	E.00	13.00	4.55	30.00	12.00	8.00	3.00	11.00	6.00	4.00	22.00	10.00	9.00	11.00	4.00	8.84
						.13	B Mode	la										
A POWER DOWN	(harder	Neg.	40.00	加井	\$1.39	72.33	73.62	71,34	67.00	68.87	10.00	54.00	32.36	\$2,34	37,33	34.00	37.30	38.13
LLAVA-L2-LN	680E	Max-Min	7.00	16.00	14.00	3.00	12.00	6.00	5,00	9.00	10.00	6.00	16.00	15.00	6.00	36.00	10.00	10.20
11.00 20 3 4 4	12222	Airg.	39.67	83.72	35.10	64.67	65.47	63.40	08.33	88.78	66.58	14.67	31.33	47.68	11.00	31,23	33,40	31.00
LLAVA-Netz-Life	- 7008	Mau-Min	1.00	11.00	13.00	9,00	19,00	15.00	1.00	12.00	11.00	3.00	31.00	14.00	1.00	21.00	10.00	11.27
	Vice	Nvg.	37.67	41.22	42.68	20.00	11.88	7444	49.33	49.37	64.58	51.33	50.00	55.68	28.67	45.25	44.00	31.71
tracerts mess/ no imposs	PROFE.	Man-Min	14.00	1500	8.00	12.00	10.00	10.00	3.00	7.00	5.00	1.00	12.00	5.00	9.00	15.00	15.00	+27



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

Are there any similarities ...



(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



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

Dynamic evaluation



Dynamic evaluation



Dynamic Multimodal Evaluation with Flexible Complexity by Vision-Language Bootstrapping

- + 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?

