上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

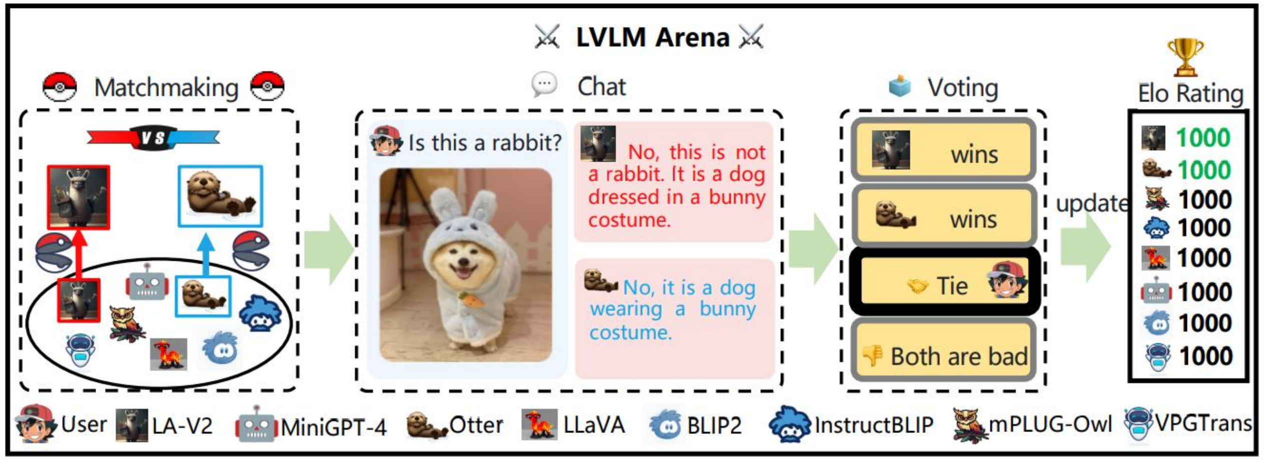
MMBench: Is Your Multi-Modal Model An **All-Around** Player?

Yuan Liu*, Haodong Duan*‡, Yuanhan Zhang*, Bo Li*, Songyang Zhang*, Wangbo Zhao,
Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu†, Kai Chen†, Dahua Lin†

* Equal Contribution ‡ Project Lead † Corresponding Author

Presenter: Haodong Duan
Oct, 2024

The study starts in May 2023. Back then, the major evaluation strategies are:



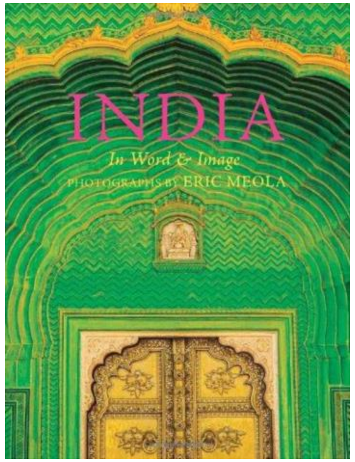
👉 **Subjective Evaluation** has following drawbacks:

1. Introduce **Human Biases**
2. Consume lots of **Resources**
3. Hard to **Reproduce**

👇 **Objective Evaluation (VQA) :**

1. Obtain metrics w. **rule-based** matching, suffer from **false-negative** samples
2. Lack a **holistic** benchmark

👇 **An example from OCRVQA:**



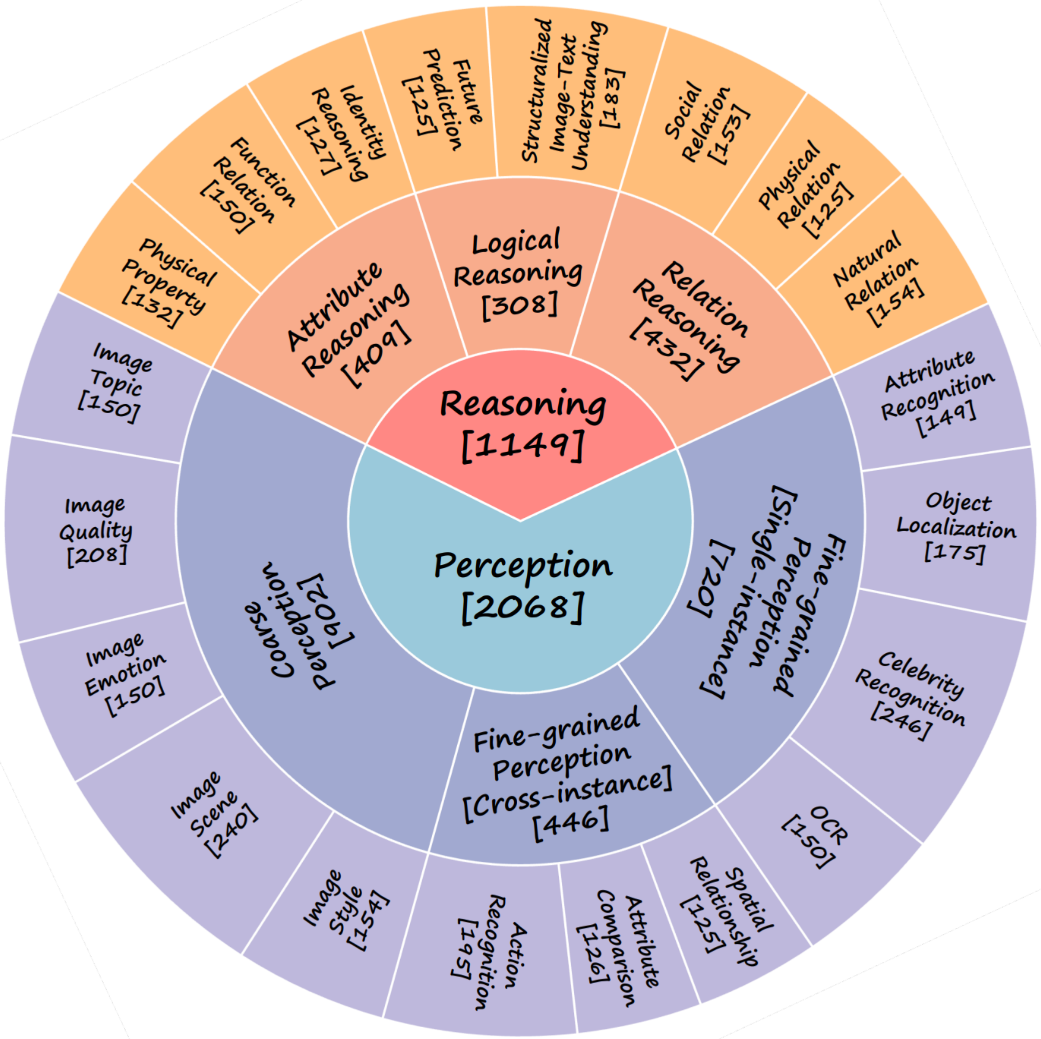
Q: What is the genre of this book?
GT: Arts & Photography
Pred (GPT-4v): The book titled "India: In Word & Image" is likely a photography or travel book that ... . The genre could be classified as travel photography, cultural exploration, or a photographic essay.

Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

The project aims at designing a new multi-modal benchmark featuring the following characteristics:

1. The benchmark needs to deliver objective & quantitative evaluation results, that is easily reproducible.
2. The benchmark needs to be comprehensive enough to cover as much multi-modal capabilities as possible.
3. The benchmark should conduct rigorous yet reasonable evaluation and mitigate the negative impact of false-negative samples

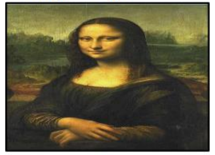
We first design a taxonomy of multi-modal capabilities:



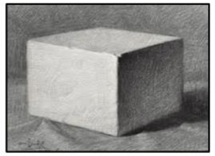
1. The taxonomy features 3 capability levels and 20 fine-grained capabilities.
2. The two most fundamental L-1 capabilities are perception & reasoning.

MMBench adopts the multi-choice format:

Image Style

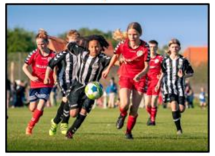


Q: Which category does this image belong to?
A. Oil Paiting
B. Sketch
C. Digital art
D. Photo
GT: A

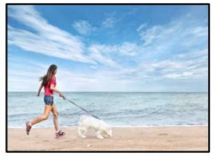


Q: Which category does this image belong to?
A. Oil Paiting
B. Sketch
C. Digital art
D. Photo
GT: B

Image Topic

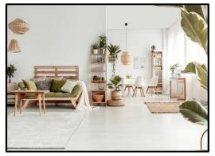


Q: Which of the following captions best describes this image?
A. A group of people playing soccer in a field
B. A woman walking her dog on a beach
C. A man riding a bicycle on a mountain trail
D. A child playing with a ball in a park
GT: A

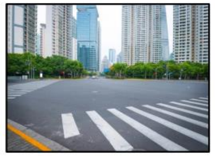


Q: Which of the following captions best describes this image?
A. A group of people playing soccer in a field
B. A woman walking her dog on a beach
C. A man riding a bicycle on a mountain trail
D. A child playing with a ball in a park
GT: B

Image scene

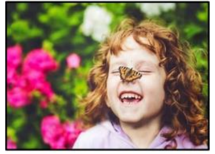


Q: What type of environment is depicted in the picture?
A. Home
B. shopping mall
C. Street
D. forest
GT: A

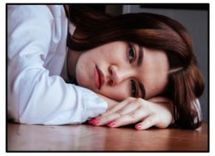


Q: What type of environment is depicted in the picture?
A. Home
B. shopping mall
C. Street
D. forest
GT: C

Image Mood

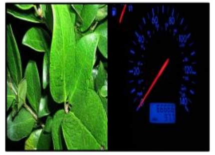


Q: Which mood does this image convey?
A. Cozy
B. Anxious
C. Happy
D. Angry
GT: C

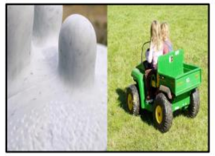


Q: Which mood does this image convey?
A. Sad
B. Anxious
C. Happy
D. Angry
GT: A

Image Quality



Q: Which image is more brightful?
A. The first image
B. The second image
GT: A

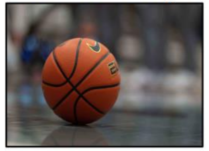


Q: which image is more colorful
A. The first image
B. The second image
GT: B

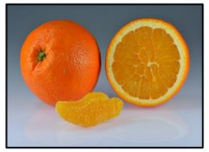
👈 Coarse Perception

Fine-grained Perception (Instance) 👇

Attribute Recognition

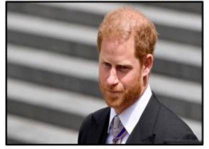


Q: What is the shape of this object?
A. Circle
B. Triangle
C. Square
D. Rectangle
GT: A

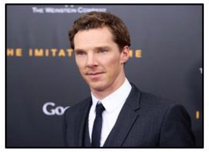


Q: what is the color of this object?
A. Purple
B. Pink
C. Gray
D. Orange
GT: D

Celebrity Recognition

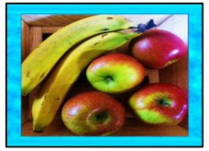


Q: Who is this person
A. David Beckham
B. Prince Harry
C. Daniel Craig
D. Tom Hardy
GT: B

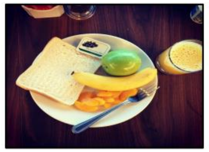


Q: Who is this person
A. Benedict Cumberbatch
B. Idris Elba
C. Ed Sheeran
D. Harry Styles
GT: A

Object Localization

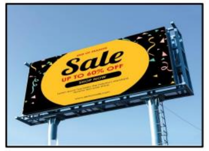


Q: How many apples are there in the image? And how many bananas are there?
A. 4 apples and 2 bananas
B. 3 apples and 3 banana
C. 2 apples and 4 bananas
D. 4 apples and 1 bananas
GT: A

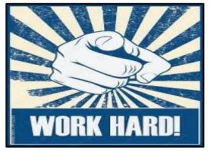


Q: Which corner is the juice?
A. Up
B. Down
C. Left
D. Right
GT: D

OCR

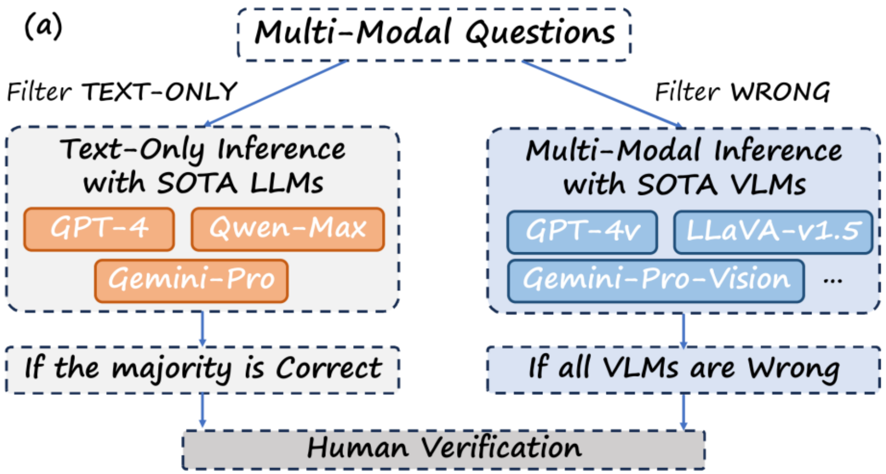


Q: What does this outdoor billboard mean?
A. Smoking is prohibited here.
B. Something is on sale.
C. No photography allowed
D. Take care of your speed.
GT: B



Q: What does this picture want to express?
A. We are expected to care for green plants.
B. We are expected to care for the earth.
C. We are expected to stay positive.
D. We are expected to work hard.
GT: D

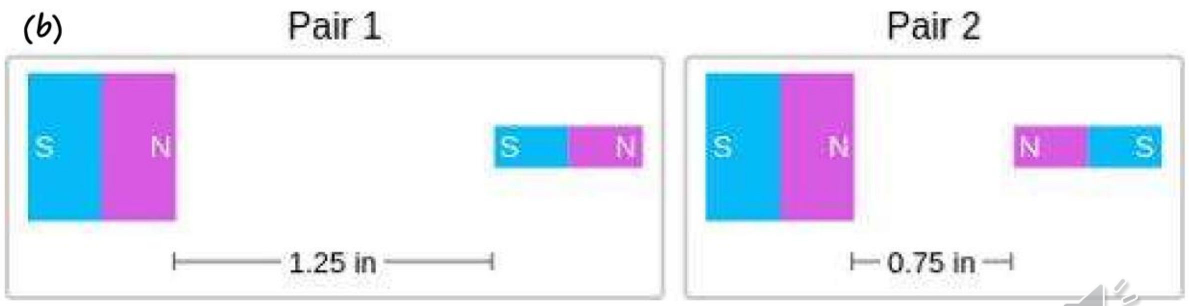
Quality Control is Crucial



Translation 👉 :
We translate all questions to Chinese with LLM, and then perform manual screening and correction.

👉 Semi-Automated Screening:

1. Questions that can be correctly answered by LLMs are removed.
2. If all SOTA VLMs failed to solve a question (w. Circular), the question will be tagged and manually checked.



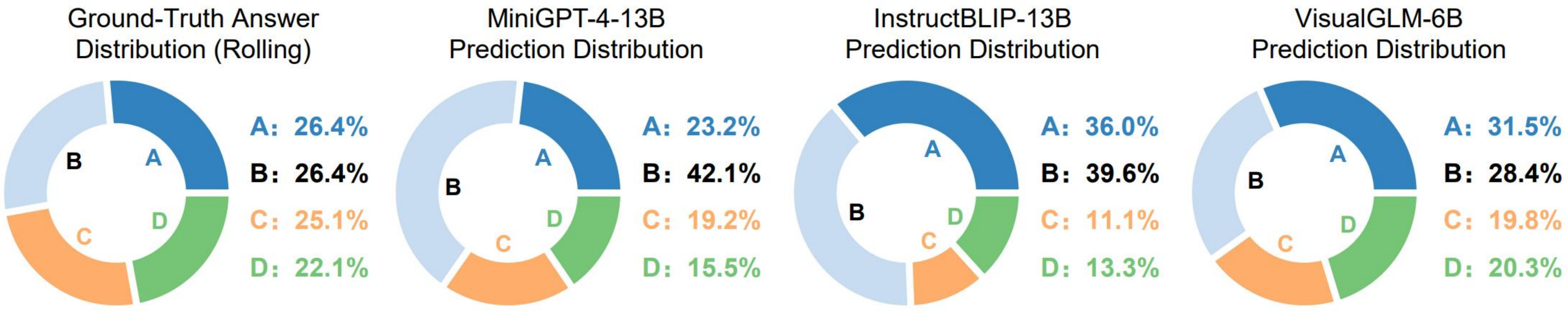
English Version (Original)

QUESTION. Think about the magnetic force between the magnets in each pair. Which of the following statements is true?
A. The magnitude of the magnetic force is smaller in Pair 2.
B. The magnitude of the magnetic force is smaller in Pair 1.
C. The magnitude of the magnetic force is the same in both pairs.

Chinese Version (Translated)

QUESTION. 考虑每对磁铁之间的磁力。以下哪个陈述是正确的?
A. 第二对磁铁之间的磁力大小较小。
B. 第一对磁铁之间的磁力大小较小。
C. 两对磁铁之间的磁力大小相同。

CircularEval is adopted to provide rigorous evaluation results



VLMs may have different preferences over choices, which introduces significant biases

Circular Evaluation

The original VL problem:
Q: How many apples are there in the image?
A. 4; B. 3; C. 2; D. 1 GT: A

4 Passes in Circular Evaluation (choices with circular shift):
1. Q: How many apples are there in the image? Choices: A. 4; B. 3; C. 2; D. 1. VLM prediction: A. GT: A ✓
2. Q: How many apples are there in the image? Choices: A. 3; B. 2; C. 1; D. 4. VLM prediction: D. GT: D ✓
3. Q: How many apples are there in the image? Choices: A. 2; B. 1; C. 4; D. 3. VLM prediction: B. GT: C ✗
4. Q: How many apples are there in the image? Choices: A. 1; B. 4; C. 3; D. 2. VLM prediction: B. GT: B ✓

VLM failed at pass 3. Thus wrong.

Under CircularEval, a VLM correctly solve a MCQ only if it succeeds in all circular passes

CircularEval vs. VanillaEval

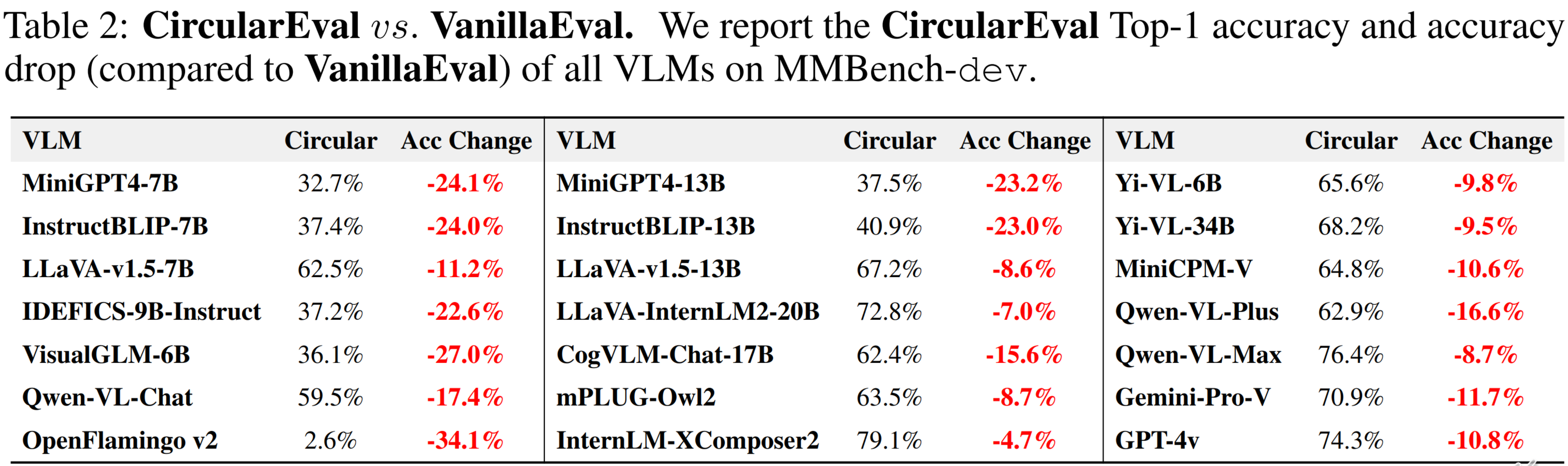


Table 2: **CircularEval** *vs.* **VanillaEval.** We report the **CircularEval** Top-1 accuracy and accuracy drop (compared to **VanillaEval**) of all VLMs on MMBench-`dev`.

VLM	Circular	Acc Change	VLM	Circular	Acc Change	VLM	Circular	Acc Change
MiniGPT4-7B	32.7%	**-24.1%**	**MiniGPT4-13B**	37.5%	**-23.2%**	**Yi-VL-6B**	65.6%	**-9.8%**
InstructBLIP-7B	37.4%	**-24.0%**	**InstructBLIP-13B**	40.9%	**-23.0%**	**Yi-VL-34B**	68.2%	**-9.5%**
LLaVA-v1.5-7B	62.5%	**-11.2%**	**LLaVA-v1.5-13B**	67.2%	**-8.6%**	**MiniCPM-V**	64.8%	**-10.6%**
IDEFICS-9B-Instruct	37.2%	**-22.6%**	**LLaVA-InternLM2-20B**	72.8%	**-7.0%**	**Qwen-VL-Plus**	62.9%	**-16.6%**
VisualGLM-6B	36.1%	**-27.0%**	**CogVLM-Chat-17B**	62.4%	**-15.6%**	**Qwen-VL-Max**	76.4%	**-8.7%**
Qwen-VL-Chat	59.5%	**-17.4%**	**mPLUG-Owl2**	63.5%	**-8.7%**	**Gemini-Pro-V**	70.9%	**-11.7%**
OpenFlamingo v2	2.6%	**-34.1%**	**InternLM-XComposer2**	79.1%	**-4.7%**	**GPT-4v**	74.3%	**-10.8%**

LLM choice-extractor to reduce false-negative samples

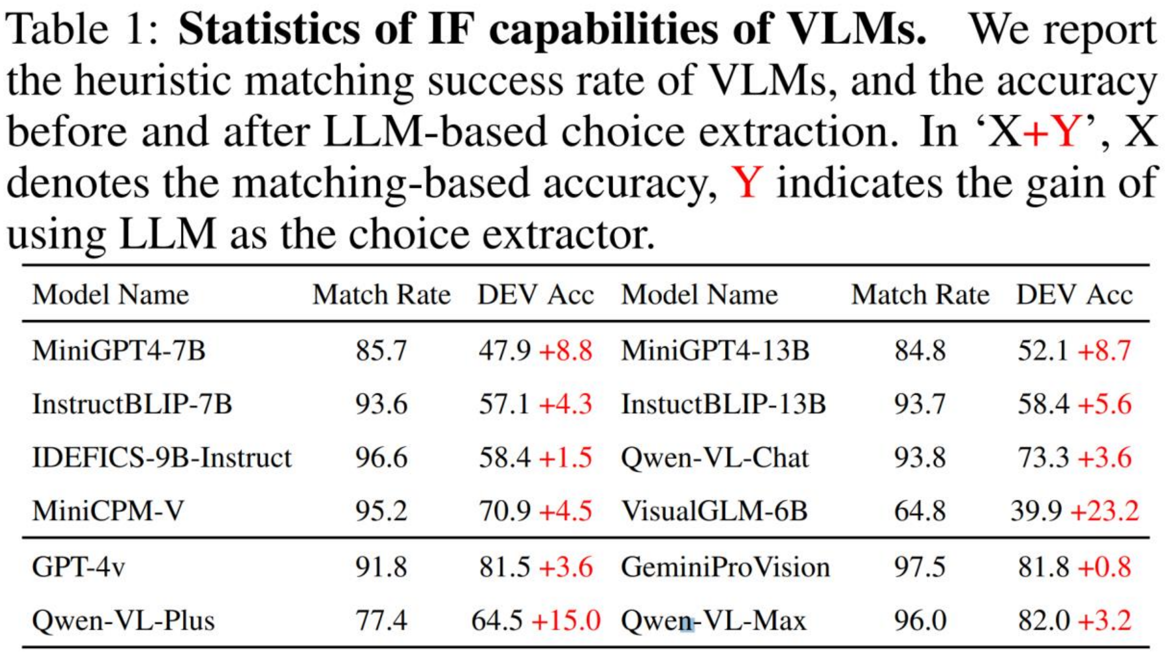
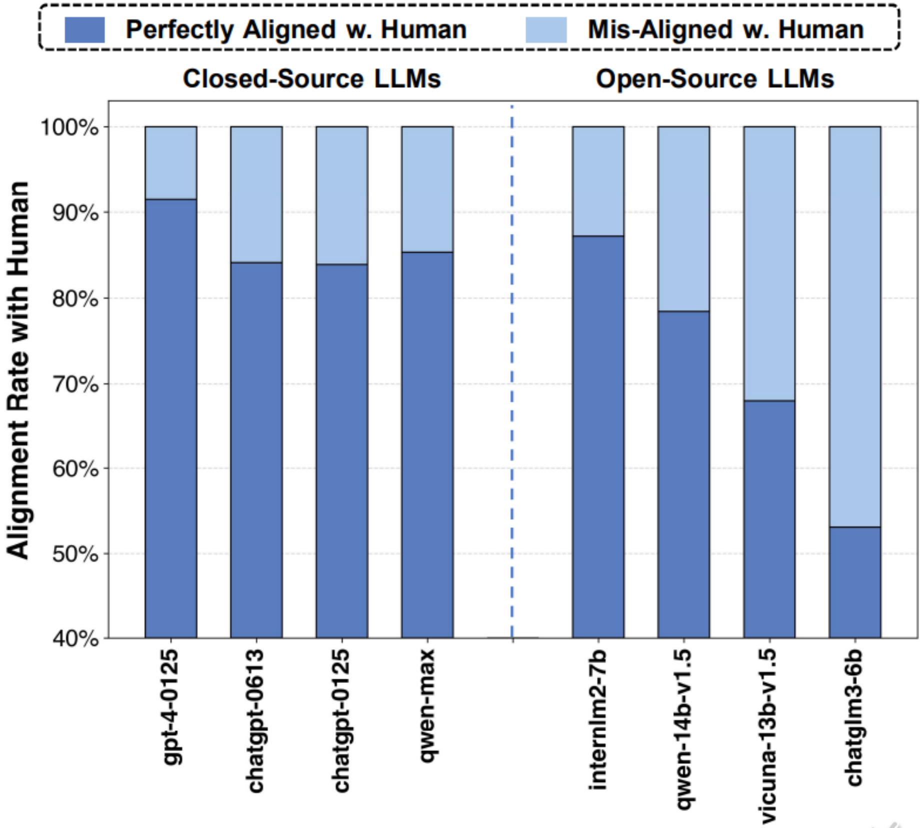


Table 1: **Statistics of IF capabilities of VLMs.** We report the heuristic matching success rate of VLMs, and the accuracy before and after LLM-based choice extraction. In ‘X+Y’, X denotes the matching-based accuracy, Y indicates the gain of using LLM as the choice extractor.

Model Name	Match Rate	DEV Acc	Model Name	Match Rate	DEV Acc
MiniGPT4-7B	85.7	47.9 +8.8	MiniGPT4-13B	84.8	52.1 +8.7
InstructBLIP-7B	93.6	57.1 +4.3	InstuctBLIP-13B	93.7	58.4 +5.6
IDEFICS-9B-Instruct	96.6	58.4 +1.5	Qwen-VL-Chat	93.8	73.3 +3.6
MiniCPM-V	95.2	70.9 +4.5	VisualGLM-6B	64.8	39.9 +23.2
GPT-4v	91.8	81.5 +3.6	GeminiProVision	97.5	81.8 +0.8
Qwen-VL-Plus	77.4	64.5 +15.0	Qwen-VL-Max	96.0	82.0 +3.2

General VLMs (including GPT-4v, Gemini, etc.) do not perform IF optimization for MCQ problems.

Using LLM to extract choice labels can help to reveal the real performance of those VLMs.



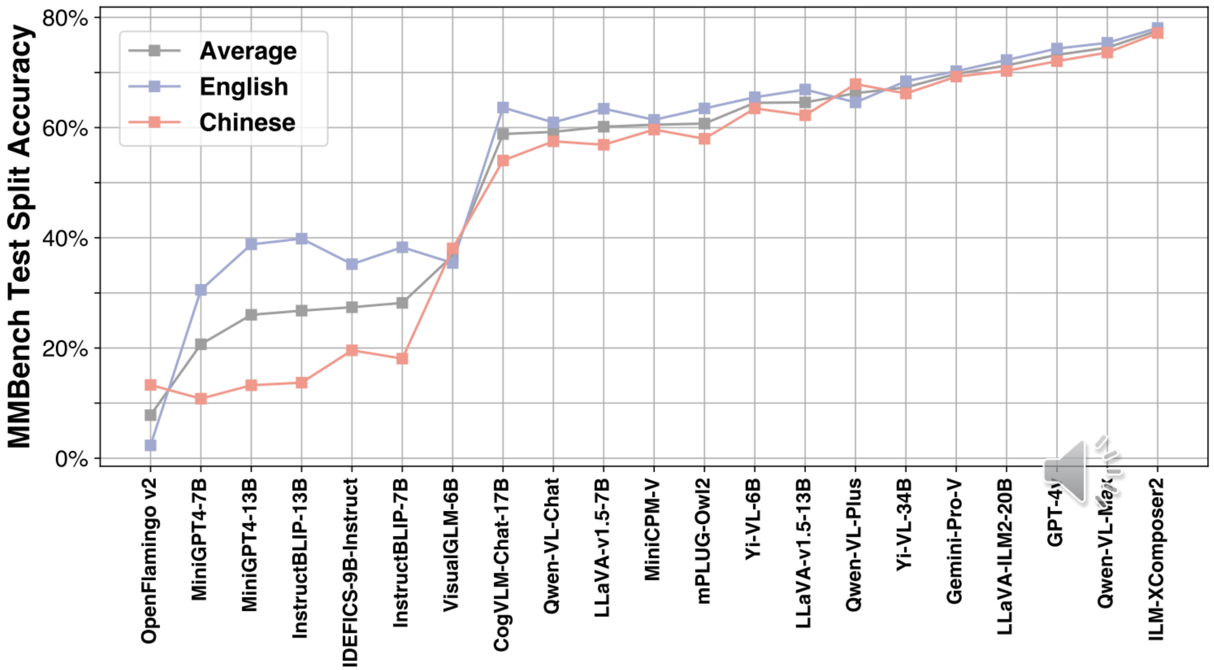
We quantitatively measured the alignment rates between different LLMs and Human on the choice extraction task.

Main Results (Mar. 2024)

Model	Overall	CP	FP-S	FP-C	AR	LR	RR
Large Language Models							
GPT-4-Turbo (0125) [37]	2.9%	0.6%	1.2%	4.1%	3.7%	4.9%	7.4%
OpenSource VLMs							
OpenFlamingo v2 [4]	2.3%	1.1%	3.5%	1.5%	5.3%	0.0%	2.7%
MiniGPT4-7B [56]	30.5%	37.0%	31.8%	17.2%	49.8%	9.2%	25.6%
IDEFICS-9B-Instruct [26]	35.2%	48.3%	31.3%	29.6%	47.8%	11.4%	25.2%
VisualGLM-6B [14]	35.4%	40.2%	38.5%	26.2%	47.8%	19.6%	29.5%
InstructBLIP-7B [12]	38.3%	46.7%	39.0%	31.8%	55.5%	8.7%	31.0%
MiniGPT4-13B [56]	38.8%	44.6%	42.9%	23.2%	64.9%	8.2%	32.9%
InstructBLIP-13B [12]	39.8%	47.2%	42.9%	21.0%	60.4%	12.5%	38.8%
Qwen-VL-Chat* [6]	60.9%	68.5%	67.7%	50.2%	78.0%	37.0%	45.7%
MiniCPM-V [39]	61.4%	65.6%	69.4%	51.3%	70.6%	35.3%	59.7%
LLaVA-v1.5-7B [32]	63.4%	70.0%	68.0%	57.7%	77.6%	33.2%	56.2%
mPLUG-Owl2 [50]	63.5%	68.1%	69.1%	55.8%	78.4%	37.0%	57.0%
CogVLM-Chat-17B [47]	63.6%	72.8%	66.6%	55.4%	71.4%	33.7%	62.0%
Yi-VL-6B* [2]	65.5%	72.8%	72.9%	56.2%	75.5%	41.3%	55.4%
LLaVA-v1.5-13B [32]	66.9%	73.1%	72.4%	60.3%	75.5%	35.9%	65.5%
Yi-VL-34B* [2]	68.4%	72.0%	78.0%	54.7%	81.2%	38.6%	68.2%
LLaVA-InternLM2-20B [11]	72.3%	78.3%	76.6%	68.2%	78.4%	46.2%	69.4%
InternLM-XComposer2* [13]	78.1%	80.4%	83.5%	73.0%	83.7%	63.6%	74.4%
Proprietary VLMs							
Qwen-VL-Plus [6]	64.6%	66.5%	79.1%	50.2%	73.9%	42.9%	57.8%
Gemini-Pro-V [44]	70.2%	70.0%	78.9%	65.9%	82.9%	46.2%	65.9%
GPT-4v [37]	74.3%	77.6%	73.8%	71.5%	85.3%	63.6%	68.6%
Qwen-VL-Max [6]	75.4%	74.8%	87.2%	67.0%	85.3%	54.9%	70.5%

👈Evaluation Results on MMBench-Test

Performance Comparison: EN vs. CN 👇



Our evaluation service has processed 20,000+ submissions (As of Sep. 2024)

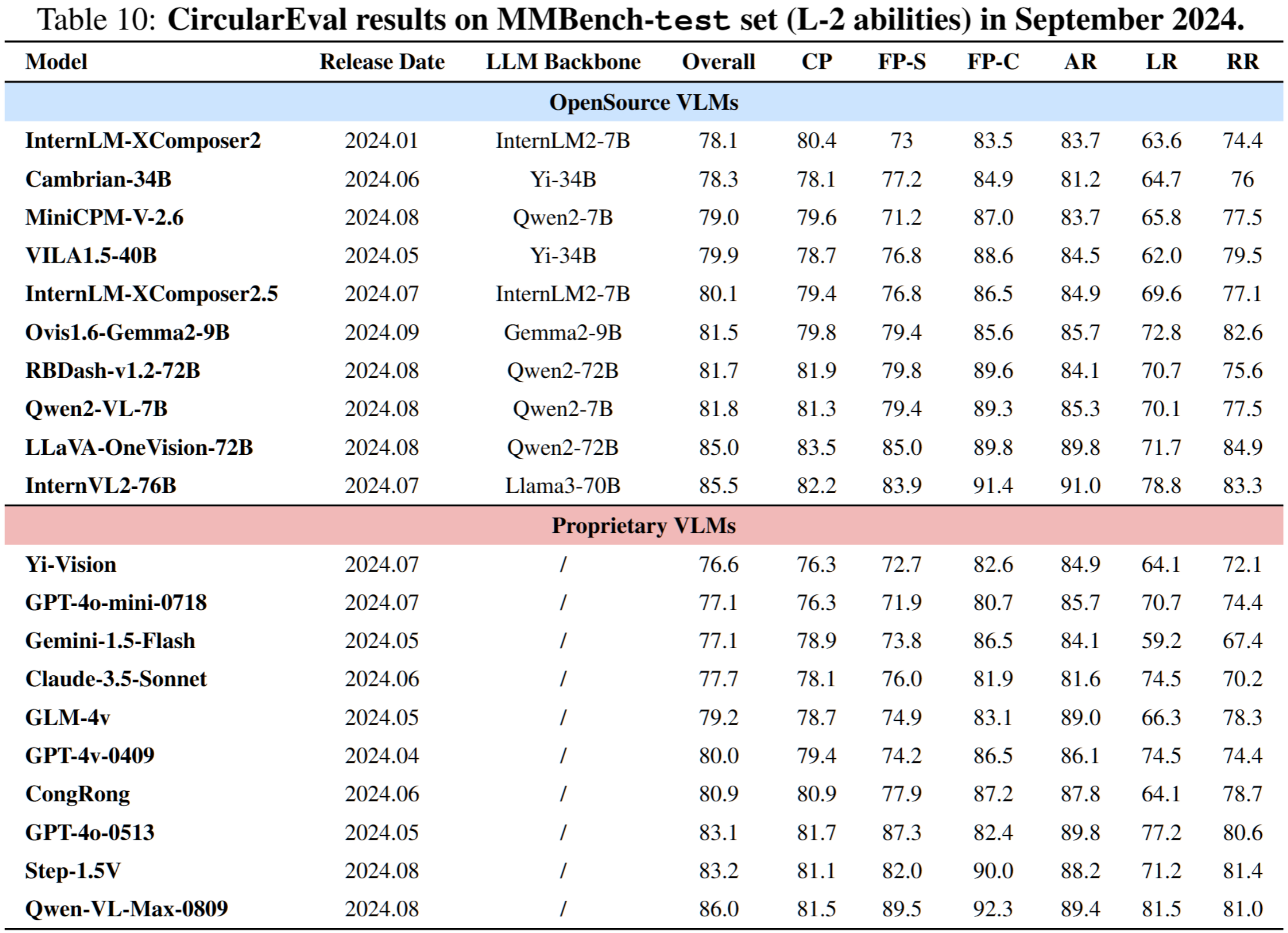


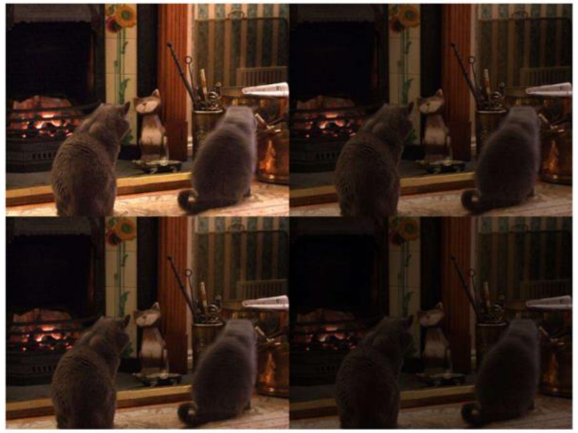
Table 10: **CircularEval results on MMBench-test set (L-2 abilities) in September 2024.**

Model	Release Date	LLM Backbone	Overall	CP	FP-S	FP-C	AR	LR	RR
			OpenSource VLMs						
InternLM-XComposer2	2024.01	InternLM2-7B	78.1	80.4	73	83.5	83.7	63.6	74.4
Cambrian-34B	2024.06	Yi-34B	78.3	78.1	77.2	84.9	81.2	64.7	76
MiniCPM-V-2.6	2024.08	Qwen2-7B	79.0	79.6	71.2	87.0	83.7	65.8	77.5
VILA1.5-40B	2024.05	Yi-34B	79.9	78.7	76.8	88.6	84.5	62.0	79.5
InternLM-XComposer2.5	2024.07	InternLM2-7B	80.1	79.4	76.8	86.5	84.9	69.6	77.1
Ovis1.6-Gemma2-9B	2024.09	Gemma2-9B	81.5	79.8	79.4	85.6	85.7	72.8	82.6
RBDash-v1.2-72B	2024.08	Qwen2-72B	81.7	81.9	79.8	89.6	84.1	70.7	75.6
Qwen2-VL-7B	2024.08	Qwen2-7B	81.8	81.3	79.4	89.3	85.3	70.1	77.5
LLaVA-OneVision-72B	2024.08	Qwen2-72B	85.0	83.5	85.0	89.8	89.8	71.7	84.9
InternVL2-76B	2024.07	Llama3-70B	85.5	82.2	83.9	91.4	91.0	78.8	83.3
			Proprietary VLMs						
Yi-Vision	2024.07	/	76.6	76.3	72.7	82.6	84.9	64.1	72.1
GPT-4o-mini-0718	2024.07	/	77.1	76.3	71.9	80.7	85.7	70.7	74.4
Gemini-1.5-Flash	2024.05	/	77.1	78.9	73.8	86.5	84.1	59.2	67.4
Claude-3.5-Sonnet	2024.06	/	77.7	78.1	76.0	81.9	81.6	74.5	70.2
GLM-4v	2024.05	/	79.2	78.7	74.9	83.1	89.0	66.3	78.3
GPT-4v-0409	2024.04	/	80.0	79.4	74.2	86.5	86.1	74.5	74.4
CongRong	2024.06	/	80.9	80.9	77.9	87.2	87.8	64.1	78.7
GPT-4o-0513	2024.05	/	83.1	81.7	87.3	82.4	89.8	77.2	80.6
Step-1.5V	2024.08	/	83.2	81.1	82.0	90.0	88.2	71.2	81.4
Qwen-VL-Max-0809	2024.08	/	86.0	81.5	89.5	92.3	89.4	81.5	81.0

The leaderboard provides evaluation results of ～200 different VLMs.

Full Leaderboard

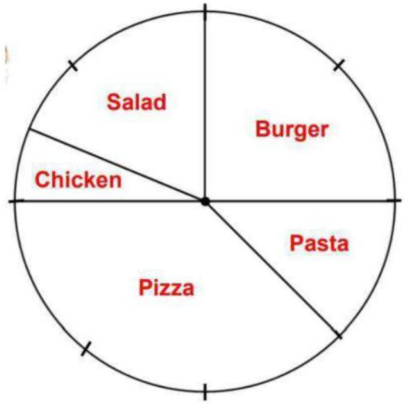
The most challenging fine-grained capabilities



Q. Which image is the second brightest?
A. upper-left
B. upper-right
C. lower-left
D. lower-right
Answer: C

A_{max}=61.3%

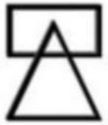
(a). Image Quality



Q. The graph shows the meals purchased in a restaurant in one day. What is the least popular meal?
A. Salad
B. Burger
C. Chicken
D. Pasta
Answer: C

A_{max}=61.5%

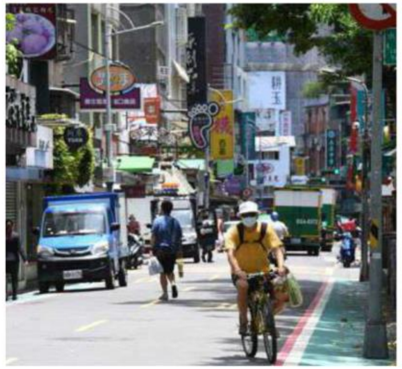
(b). Structralized Image-Text Understanding



Q. What is the positional relationship between the two shapes in the picture?
A. The two shapes are positioned apart or separated from each other.
B. The two shapes are tangentially positioned or externally tangent to each other.
C. The two shapes intersect with each other.
D. One shape is contained within the other or there is an inner shape enclosed by an outer shape.
Answer: C

A_{max}=68.0%

(c). Spatial Relationship



Q. From the perspective of the driver of the blue truck, in what position is the person riding a bike relative to the blue truck?
A. Left front
B. Right front
C. Right rear
D. Left rear
Answer: A

A_{max}=64.0%

(d). Physical Relation Reasoning

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Thanks for your attention!

Our Poster is at 182 this afternoon
Welcome to chat!

Also, if you want to learn about
the latest work of the team:

VLMEvalKit

Prism

MMBench-Video