II-Bench: An Image Implication Understanding Benchmark for Multimodal Large Language Models

Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, Hongquan Lin, Jiaming Li, Yuansheng Ni, Haihong Wu, Yaswanth Narsupalli, Zhigang Zheng, Chengming Li, Xiping Hu, Ruifeng Xu, Xiaojun Chen, Min Yang, Jiaheng Liu, Ruibo Liu, Wenhao Huang, Ge Zhang, Shiwen Ni.

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
01.ai

NeurIPS 2024 | Presenter: Ziqiang Liu | November, 2024

https://huggingface.co/datasets/m-a-p/II-Bench

Contents

中国科学院大学
University of Chinese Academy of Sciences

中国科学院大学
University of Chinese Academy of Sciences

Can MLLMs Understand the **Deep Implication** Behind Images?

Motivation

Image Implication Capabilties

中国科学院大学
University of Chinese Academy of Sciences

- Numerous challenging and comprehensive benchmarks have been proposed to more accurately assess the capabilities of MLLMs.
- There is a dearth of exploration of the **higher-order perceptual capabilities** of MLLMs.
- We propose the **I**mage **I**mplication understanding **Bench**mark, **II-Bench**, which aims to evaluate the model's higher-order perception of images.
- We believe that II-Bench will inspire the community to develop the next generation of MLLMs, advancing the journey towards expert artificial general intelligence (AGI).

How to construct II-Bench?

II-Bench

Data Curation Process

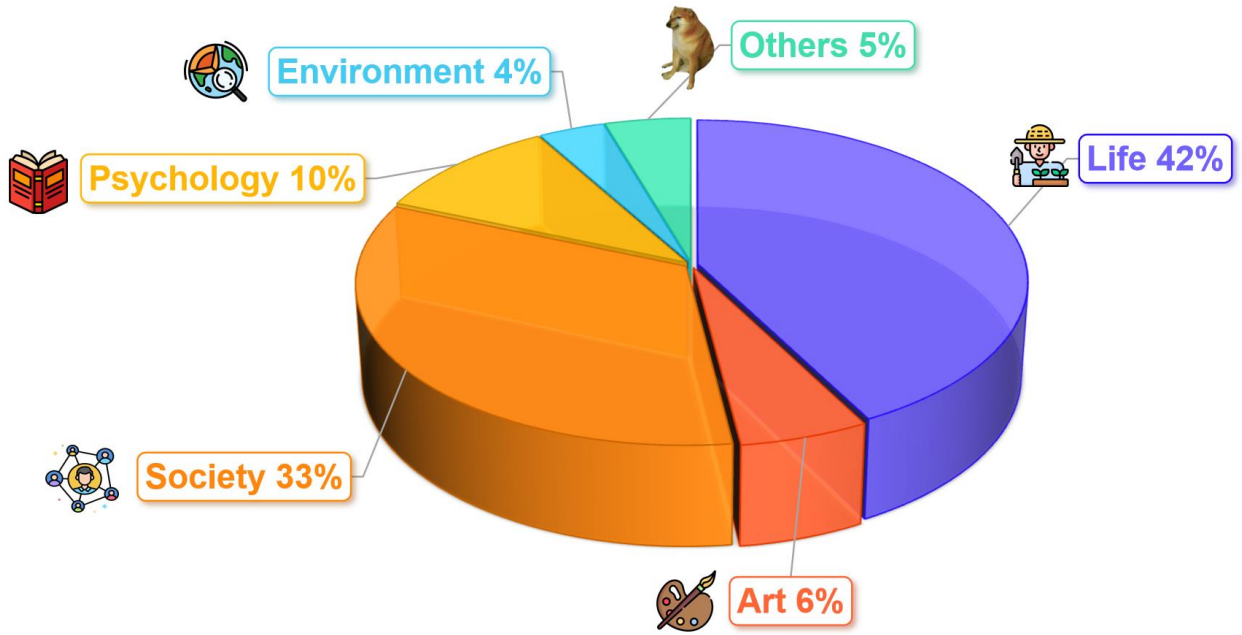
University of Chinese Academy of Sciences

- **Data Collection**: We collect 20,150 raw images from various renowned illustration websites, ensuring a sufficiently extensive raw dataset.
- **Data Filtration**: image deduplication -> text-to-image ratio control -> visual inspection
- **Data Annotation**: The annotators mark the images with their difficulty, image type, domain, and corresponding rhetoric first. An explanation of contained visual implications is then drafted for each image, Finally, the annotators devise 1-3 fine-grained questions per image, each with only one correct answer and five distractor options related to the implication nuances.
- **Data Quality Assurance**: Each question and option undergoes multiple rounds of meticulous manual annotation to ensure the distractors are sufficiently challenging and not easily distinguishable from the correct option and ensure consistency across different annotators.

II-Bench

Statistics

University of Chinese Academy of Sciences



- **1222** images
- **1434** questions
- **6** domains
- **6** categories
- **3** sentiments
- **3** difficulties
- **9** rhetorics

Experiments

Settings

University of Chinese Academy of Sciences

Model	Size	ViT	Projection Module	LLM
CogVLM2-Llama3-Chat [58]	19.5B	EVA2-CLIP-E	MLP	Llama-3-8B + Visual Expert
MiniCPM-Llama3-2.5 [23]	8.5B	SigLip-400M	Perceiver Resampler	Llama3-8B
InternVL-Chat-1.5 [8]	25.5B	InternViT-6B	MLP	InternLM2-20B
InternLM-XComposer2-VL [13]	7B	OpenAI ViT-Large	PLoRA	InternLM-2
DeepSeek-VL-Chat-7B [40]	7.3B	SAM-B + SigLIP-L	MLP	DeepSeek-LLM-7B
InstructBLIP-T5 [11]	4.0B/12.3B	ViT-g/14	MLP	FLAN T5 XL/XXL
BLIP-2 FLAN-T5 [33]	4.1B/12.1B	ViT-g/14	MLP	FLAN T5 XL/XXL
mPLUGw-OWL2 [62]	8.2B	ViT-L/14	Visual Abstractor	Llama-2-7B
Qwen-VL-Chat [3]	9.6B	ViT-bigG	VL Adapter	Qwen-7B
Yi-VL-34B-Chat [63]	7.1B/35.4B	CLIP ViT-H/14	MLP	Yi-34B-Chat
LLaVA-1.6-34B [35]	34.8B	ViT-L/14	MLP	Nous-Hermes-2-Yi-34B
Mantis-8B-siglip-llama3 [26]	8.5B	SigLIP	MLP	Llama-3-8B
Idefics2-8B [28]	8.4B	SigLIP	MLP	Mistral-7B

- Zero(Few)-Shot Prompting: 0, 1, 2, 3 shot(s)
- Chain of Thought Prompting
- Domain: give the image’s domain (e.g. life, environment) in the prompt
- Emotion: give the image’s emotion(e.g. positive, negative) in the prompt
- Rhetoric: give the rhetorical devices (e.g. metaphor, personification)

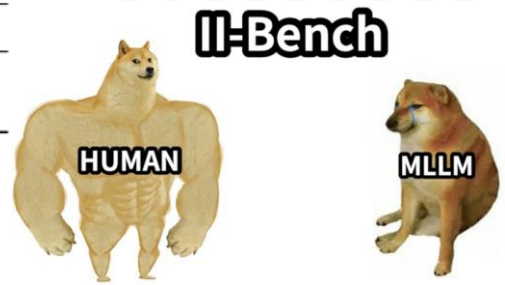
Experiments

Main Results

University of Chinese Academy of Sciences

	Overall (1,399)	Life (585)	Art (85)	Society (461)	Psy. (152)	Env. (51)	Others (65)	Positive (196)	Neutral (789)	Negative (414)
Open-source Models										
InstructBLIP-T5-XL	47.3	45.6	48.2	48.8	44.7	52.9	50.8	46.9	48.3	45.4
BLIP-2 FLAN-T5-XL	52.8	53.0	58.8	52.5	42.8	64.7	58.5	56.1	52.9	51.0
mPLUGw-OWL2	53.2	54.0	56.5	50.5	52.0	60.8	56.9	55.6	52.6	53.1
Qwen-VL-Chat	53.4	53.2	49.4	52.1	50.0	60.8	72.3	56.1	52.6	53.6
InstructBLIP-T5-XXL	56.7	56.2	58.8	58.6	45.4	64.7	64.6	63.3	56.1	54.6
Mantis-8B-siglip-Llama3	57.5	56.8	61.2	57.5	53.9	64.7	61.5	59.2	58.0	55.6
BLIP-2 FLAN-T5-XXL	57.8	57.1	63.5	57.0	53.3	66.7	66.2	67.9	57.2	54.3
DeepSeek-VL-Chat-7B	60.3	59.0	58.8	58.4	61.8	68.6	76.9	65.8	60.1	58.0
Yi-VL-6B-Chat	61.3	60.9	63.5	60.7	56.6	66.7	72.3	61.7	61.7	60.1
InternLM-XComposer2-VL	62.1	61.7	62.4	62.3	58.6	70.6	66.2	65.8	63.0	58.7
InternVL-Chat-1.5	66.3	63.6	65.9	68.5	65.8	64.7	76.9	73.5	65.4	64.5
Idefics2-8B	67.7	67.2	**74.1**	67.7	62.5	74.5	70.8	68.9	67.0	68.4
Yi-VL-34B-Chat	67.9	67.5	70.6	67.7	63.8	70.6	76.9	74.0	68.2	64.5
MiniCPM-Llama3-2.5	69.4	68.4	71.8	69.4	64.5	**80.4**	78.5	75.0	69.3	66.9
CogVLM2-Llama3-Chat	70.3	68.9	68.2	70.9	67.8	72.5	**86.2**	69.9	71.1	69.1
LLaVA-1.6-34B	**73.8**	**73.8**	71.8	**73.3**	**71.1**	78.4	81.5	**79.1**	**72.9**	**72.9**
Closed-source Models										
GPT-4V	65.9	65.0	69.4	65.3	59.9	76.5	80.0	69.4	66.0	64.0
GPT-4o	72.6	72.5	72.9	73.3	68.4	76.5	75.4	78.6	71.2	72.5
Gemini-1.5 Pro	73.9	73.7	**74.1**	74.4	63.2	**80.4**	83.1	**80.1**	70.8	**75.4**
Qwen-VL-MAX	**74.8**	**74.7**	71.8	**74.6**	**73.0**	76.5	**84.6**	**80.1**	**74.5**	72.9
Humans										
Human_avg	90.3	90.0	88.2	91.4	86.6	96.1	92.3	84.7	89.1	92.2
Human_best	**98.2**	**97.9**	**98.8**	**98.3**	**97.4**	**100.0**	**100.0**	**98.0**	**98.0**	**98.8**

There is still **a huge gap** between humans and MLLMs.



Experiments

Main Results

University of Chinese Academy of Sciences

	Overall (1,399)	Life (585)	Art (85)	Society (461)	Psy. (152)	Env. (51)	Others (65)	Positive (196)	Neutral (789)	Negative (414)
Open-source Models										
InstructBLIP-T5-XL	47.3	45.6	48.2	48.8	44.7	52.9	50.8	46.9	48.3	45.4
BLIP-2 FLAN-T5-XL	52.8	53.0	58.8	52.5	42.8	64.7	58.5	56.1	52.9	51.0
mPLUGw-OWL2	53.2	54.0	56.5	50.5	52.0	60.8	56.9	55.6	52.6	53.1
Qwen-VL-Chat	53.4	53.2	49.4	52.1	50.0	60.8	72.3	56.1	52.6	53.6
InstructBLIP-T5-XXL	56.7	56.2	58.8	58.6	45.4	64.7	64.6	63.3	56.1	54.6
Mantis-8B-siglip-Llama3	57.5	56.8	61.2	57.5	53.9	64.7	61.5	59.2	58.0	55.6
BLIP-2 FLAN-T5-XXL	57.8	57.1	63.5	57.0	53.3	66.7	66.2	67.9	57.2	54.3
DeepSeek-VL-Chat-7B	60.3	59.0	58.8	58.4	61.8	68.6	76.9	65.8	60.1	58.0
Yi-VL-6B-Chat	61.3	60.9	63.5	60.7	56.6	66.7	72.3	61.7	61.7	60.1
InternLM-XComposer2-VL	62.1	61.7	62.4	62.3	58.6	70.6	66.2	65.8	63.0	58.7
InternVL-Chat-1.5	66.3	63.6	65.9	68.5	65.8	64.7	76.9	73.5	65.4	64.5
Idefics2-8B	67.7	67.2	**74.1**	67.7	62.5	74.5	70.8	68.9	67.0	68.4
Yi-VL-34B-Chat	67.9	67.5	70.6	67.7	63.8	70.6	76.9	74.0	68.2	64.5
MiniCPM-Llama3-2.5	69.4	68.4	71.8	69.4	64.5	**80.4**	78.5	75.0	69.3	66.9
CogVLM2-Llama3-Chat	70.3	68.9	68.2	70.9	67.8	72.5	**86.2**	69.9	71.1	69.1
LLaVA-1.6-34B	**73.8**	**73.8**	71.8	**73.3**	**71.1**	78.4	81.5	**79.1**	**72.9**	**72.9**
Closed-source Models										
GPT-4V	65.9	65.0	69.4	65.3	59.9	76.5	80.0	69.4	66.0	64.0
GPT-4o	72.6	72.5	72.9	73.3	68.4	76.5	75.4	78.6	71.2	72.5
Gemini-1.5 Pro	73.9	73.7	**74.1**	74.4	63.2	**80.4**	83.1	**80.1**	70.8	**75.4**
Qwen-VL-MAX	**74.8**	**74.7**	71.8	**74.6**	**73.0**	76.5	**84.6**	**80.1**	**74.5**	72.9
Humans										
Human_avg	90.3	90.0	88.2	91.4	86.6	96.1	92.3	84.7	89.1	92.2
Human_best	**98.2**	**97.9**	**98.8**	**98.3**	**97.4**	**100.0**	**100.0**	**98.0**	**98.0**	**98.8**

Small Disparity between Open-source and Closed-source Models

Experiments

Main Results

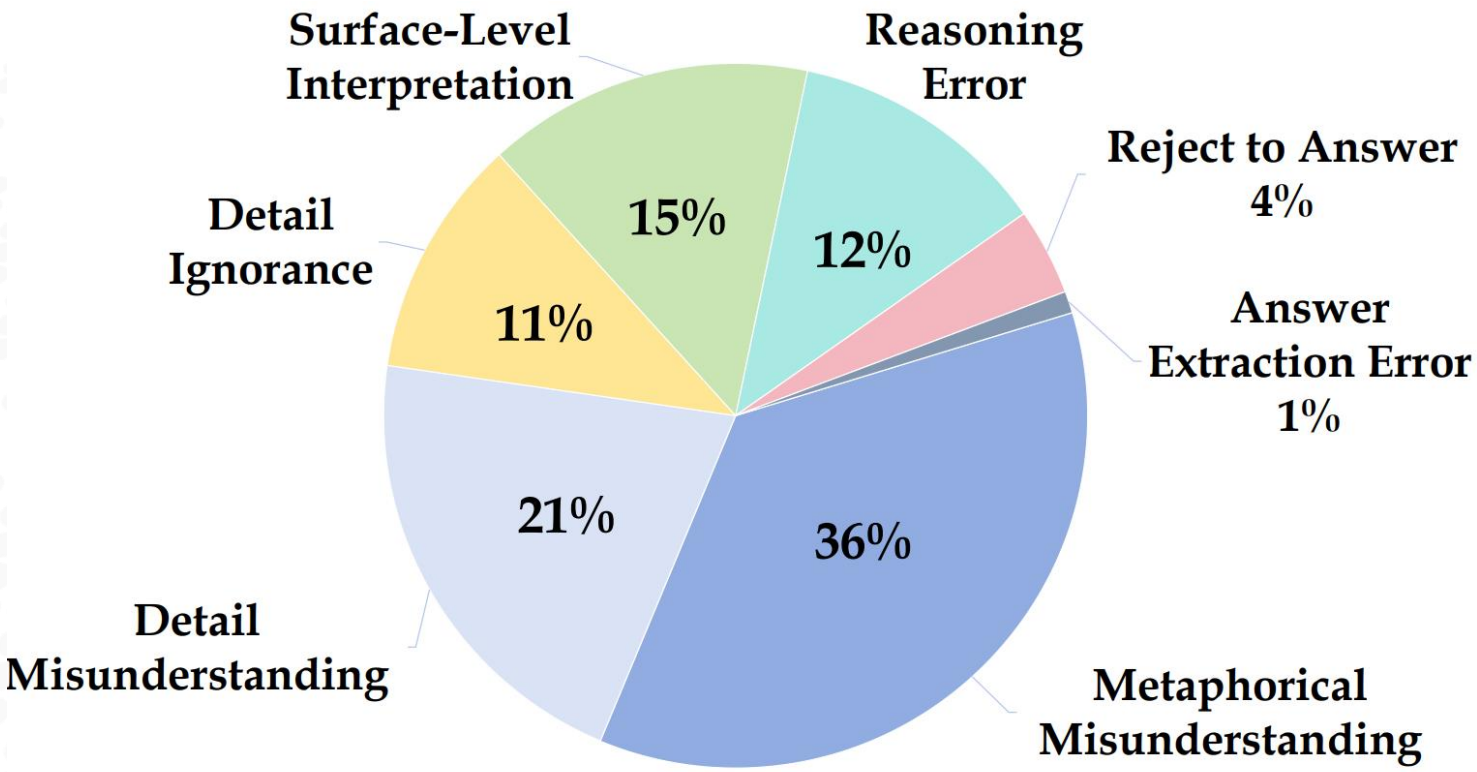
University of Chinese Academy of Sciences

	Overall (1,399)	Life (585)	Art (85)	Society (461)	Psy. (152)	Env. (51)	Others (65)	Positive (196)	Neutral (789)	Negative (414)
Open-source Models										
InstructBLIP-T5-XL	47.3	45.6	48.2	48.8	44.7	52.9	50.8	46.9	48.3	45.4
BLIP-2 FLAN-T5-XL	52.8	53.0	58.8	52.5	42.8	64.7	58.5	56.1	52.9	51.0
mPLUGw-OWL2	53.2	54.0	56.5	50.5	52.0	60.8	56.9	55.6	52.6	53.1
Qwen-VL-Chat	53.4	53.2	49.4	52.1	50.0	60.8	72.3	56.1	52.6	53.6
InstructBLIP-T5-XXL	56.7	56.2	58.8	58.6	45.4	64.7	64.6	63.3	56.1	54.6
Mantis-8B-siglip-Llama3	57.5	56.8	61.2	57.5	53.9	64.7	61.5	59.2	58.0	55.6
BLIP-2 FLAN-T5-XXL	57.8	57.1	63.5	57.0	53.3	66.7	66.2	67.9	57.2	54.3
DeepSeek-VL-Chat-7B	60.3	59.0	58.8	58.4	61.8	68.6	76.9	65.8	60.1	58.0
Yi-VL-6B-Chat	61.3	60.9	63.5	60.7	56.6	66.7	72.3	61.7	61.7	60.1
InternLM-XComposer2-VL	62.1	61.7	62.4	62.3	58.6	70.6	66.2	65.8	63.0	58.7
InternVL-Chat-1.5	66.3	63.6	65.9	68.5	65.8	64.7	76.9	73.5	65.4	64.5
Idefics2-8B	67.7	67.2	**74.1**	67.7	62.5	74.5	70.8	68.9	67.0	68.4
Yi-VL-34B-Chat	67.9	67.5	70.6	67.7	63.8	70.6	76.9	74.0	68.2	64.5
MiniCPM-Llama3-2.5	69.4	68.4	71.8	69.4	64.5	**80.4**	78.5	75.0	69.3	66.9
CogVLM2-Llama3-Chat	70.3	68.9	68.2	70.9	67.8	72.5	**86.2**	69.9	71.1	69.1
LLaVA-1.6-34B	**73.8**	**73.8**	71.8	**73.3**	**71.1**	78.4	81.5	**79.1**	**72.9**	**72.9**
Closed-source Models										
GPT-4V	65.9	65.0	69.4	65.3	59.9	76.5	80.0	69.4	66.0	64.0
GPT-4o	72.6	72.5	72.9	73.3	68.4	76.5	75.4	78.6	71.2	72.5
Gemini-1.5 Pro	73.9	73.7	**74.1**	74.4	63.2	**80.4**	83.1	**80.1**	70.8	**75.4**
Qwen-VL-MAX	**74.8**	**74.7**	71.8	**74.6**	**73.0**	76.5	**84.6**	**80.1**	**74.5**	72.9
Humans										
Human_avg	90.3	90.0	88.2	91.4	86.6	96.1	92.3	84.7	89.1	92.2
Human_best	**98.2**	**97.9**	**98.8**	**98.3**	**97.4**	**100.0**	**100.0**	**98.0**	**98.0**	**98.8**

Model Performance across **Different Domains and Emotions**

Analysis

Error Analysis



Metaphorical Misunderstanding is a most common error that GPT-4V makes when generating responses based on image comprehension.

Future

MLLM

AGI

中国科学院大学
University of Chinese Academy of Sciences

Thanks for your attention!

Contact:
zq.liu4@siat.ac.cn

Paper:
https://arxiv.org/pdf/2406.05862

Code:
https://github.com/II-Bench/II-Bench

Huggingface:
https://huggingface.co/datasets/m-a-p/II-Bench