Can MLLMs Understand the Deep Implication Behind Chinese Images?

Chenhao Zhang*, Xi Feng*, Yuelin Bai*, Xinrun Du*
Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu
Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang
Min Yang, Wenhao Huang, Chenghua Lin, Ge Zhang†, Shiwen Ni†

Huazhong University of Science and Technology,
Shenzhen Institute of Advanced Technology, CAS
01.ai, M-A-P, Hokudai, UCSB, University of Manchester, etc.

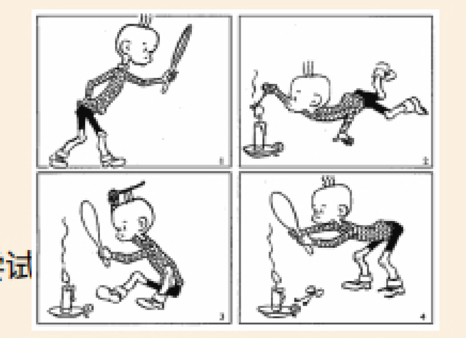
Introduction

As the capabilities of Multimodal Large Language Models (MLLMs) improve, the need for higher-order evaluation of them is increasing. However, there is a lack of work evaluating MLLM for **higher-order perception** and understanding of **Chinese visual content**. To address this, we introduce the **CII-Bench**, which aims to assess MLLMs' such capabilities for Chinese images. To ensure the authenticity of the Chinese context, images in CII-Bench are sourced from the Chinese Internet and manually reviewed, with corresponding answers also manually crafted. Additionally, CII-Bench incorporates images that represent Chinese traditional culture, such as famous Chinese traditional paintings, which can deeply reflect the model's understanding of Chinese traditional culture. Through experiments on multiple MLLMs using CII-Bench, significant findings emerged. There is a large gap between MLLMs and humans in performance. The highest MLLM accuracy is 64.4%, while the human average is 78.2% and the peak is 81.0%. MLLMs perform poorly on traditional culture images, indicating limitations in understanding high-level semantics and lacking a deep knowledge base of Chinese traditional culture. Moreover, most models have higher accuracy when image emotion hints are added to the prompts. We believe CII-Bench will help MLLMs better understand Chinese semantics and specific images, and move forward the development of expert artificial general intelligence (AGI).
Our project is publicly available at ***https://cii-bench.github.io***.

Overview

We introduce the **C**hinese **I**mage **I**mplication Understanding **Bench**mark **CII-Bench**, a new benchmark measuring the **higher-order perceptual**, **reasoning** and **comprehension** abilities of MLLMs when presented with **complex Chinese implication images**. These images, including abstract artworks, comics, memes, posters and paintings, possess visual implications that require an understanding of visual details and reasoning ability. CII-Bench reveals whether current MLLMs, leveraging their inherent comprehension abilities, can accurately decode the metaphors embedded within the complex and abstract information presented in these images.

生活 Life



Question: 这张图片有什么隐喻?
Option:
(A)有的人不充分考虑自身的情况就盲目的去追求某些事物。
(B)羽毛熄灭蜡烛是一种天马行空的想法，讽刺了不切实际的胡乱尝试
(C)每个人都有追求时尚的权利。
(D)羽毛熄灭蜡烛是一种天马行空的想法，虽然失败了，但这种创新值得赞扬。
(E)图中人物多次用羽毛熄灭蜡烛，赞扬了坚持不懈、百折不挠的精神。
(F)用羽毛熄灭蜡烛这种不合理的行为，讽刺了有的人做事不考虑周到，盲目的尝试。

Image Type: 多格漫画(Multi-panel Comic)
Rhetoric: 隐喻
Emotion: 消极
Difficulty Level: 简单

艺术 Art

Question:远处的小岛暗示了什么?
Option:
(A)远处的小岛被描绘为精神寄托的象征，代表了人们在现实世界中寻找精神慰藉和寄托的地方。
(B)远处的小岛与远处的棕榈树共同营造出一种与自然和谐共处的氛围，暗示着人与自然之间的和谐关系。
(C)远处的小岛象征着希望和目标，虽然距离遥远，但依旧可以到达。
(D)远处的小岛上有着特定的文化景观或历史遗迹，象征着特定的文化背景或历史时期，提醒人们关注和尊重历史与文化的重要性。
(E)远处的小岛作为远方的地标，象征着未知的领域或新的探索方向，鼓励人们勇敢地去探索未知。
(F)远处的小岛象征着个人内心深处的平静之地，是人们在面对外界压力和挑战时寻求内心平静和恢复的地方。

Image Type: 插画(Illustration)
Rhetoric: 象征
Emotion: 积极
Difficulty Level: 简单

社会 Society

Question: 这张图片有什么隐喻?
Option:
(A)坚持不懈是一种重要的美德。
(B)父母的行为习惯决定了孩子的未来。
(C)教育的失败是因为家长没有起到足够的监督作用。
(D)钢琴的学习应该从小做起并坚持下来，这才能够走向成功。
(E)有的家长把教育失败的原因归咎于孩子，却忽略了自身的原因。
(F)如果父母不以身作则成为榜样，那么将来孩子的教育一定失败。

Image Type: 单格漫画(Single-panel Comic)
Rhetoric: 对比
Emotion: 消极
Difficulty Level: 简单

中华传统文化 Chinese Traditional Culture

Question: 这张图片有什么隐喻?
Option:
(A)萧瑟的冬景暗示了人物对于春天到来、万物复苏的渴望。
(B)孤身赏雪景暗示了图片中人物淡然、超脱世俗的心境。
(C)独自一人欣赏雪景暗示了人物内心的孤独和知己难求的悲伤。
(D)抬头的动作暗示了人物的思考。
(E)独身一人暗示了人物对于亲人和家乡的怀念。
(F)萧瑟的冬景暗示了人物内心的悲伤。

Image Type: 绘画(Painting)
Rhetoric: 隐喻
Emotion: 积极
Difficulty Level: 困难

环境 Environment

Question: 这张图片有什么隐喻?
Option:
(A)象征着自然界的生物受到人类活动的严重影响，甚至面临灭绝的威胁。
(B)这张图片表现了工业技术的飞速发展，暗示着未来生活将更加便利和富裕。
(C)这张图片旨在宣传新型环保技术的应用，表现工业与自然和谐共处的美好愿景。
(D)暗示了人们有能力通过改变行为模式、采用新技术、实施环保政策等方式，来减轻对自然环境的破坏，实现可持续发展和生态平衡的可能。
(E)表达了人类对自然界的彻底征服，通过技术改变地表环境。
(F)表达了对环境污染和生态破坏的深刻忧虑，它提醒观者在追求工业发展的同时，不应忽视对自然环境的保护和珍惜。

Image Type: 海报(Poster)
Rhetoric: 象征
Emotion: 消极
Difficulty Level: 中等

政治 Politics

Question: 这张图片有什么隐喻?
Option:
(A)个体在面对群体或更高权威时，所面临的道德困境和选择。
(B)天使和士兵形象之间的冲突暗示了信仰与现实之间的张力，以及个体在面对残酷现实时，如何坚持自己的信仰。
(C)图片象征了人类对宗教信仰的追求，表达了对精神世界的渴望。
(D)图片可能讽刺了那些以战争干预其他国家或地区的行为，表达了对和平的渴望与对战争后果的担忧。
(E)个人的命运既受到外力的影响，也取决于个人的选择。
(F)即使在和平时期，战争的威胁也可能随时存在；而即使在战争中，人们也可能怀抱着对和平的渴望。

Image Type: 插画(Illustration)
Rhetoric: 隐喻、对比
Emotion: 消极
Difficulty Level: 困难

Statistics

CII-Bench contains a total of **698** various Chinese images. These images are manually collected and annotated by 30 undergraduate students from various disciplines and institutions, with sources from multiple renowned Chinese illustration websites. Each image is manually designed with one to three multiple-choice questions, each with six options and only one correct answer. The questions cover the metaphors, symbolism, and detailed understanding of the images. The benchmark includes a total of **800** multiple-choice questions, with 765 questions used to construct the test set and 35 questions used to construct the development and validation set for few-shot tasks.

Statistics	
Total Questions	800
Total Images	698
Dev : Validation : Test	15 : 20 : 765
Easy : Medium : Hard	305 : 282 : 111
Average Question Length	10.54
Average Option Length	28.31
Average Explanation Length	121.06
Metaphor	562
Exaggerate	121
Symbolism	236
Visual Dislocation	42
Antithesis	13
Analogy	19
Personification	73
Contrast	87

Statistics	
Life	216 (30.95%)
Art	123 (17.62%)
Society	157 (22.49%)
Environment	51 (7.31%)
Politics	21 (3.01%)
Chinese Traditional Culture	130 (18.62%)
Positive	220 (31.52%)
Neutral	247 (35.39%)
Negative	231 (33.09%)
Illustration	178 (25.50%)
Meme	145 (20.77%)
Poster	87 (12.46%)
Multi-panel Comic	34 (4.87%)
Single-panel Comic	143 (20.49%)
Painting	119 (17.05%)

Illustration	Meme	Poster
Life (19, 10.68%)	Life (138, 95.17%)	Life (8, 9.20%)
Art (79, 44.38%)	Art (0, 0%)	Art (33, 37.93%)
Society (58, 32.58%)	Society (4, 2.76%)	Society (3, 3.45%)
CTC (0, 0%)	CTC (0, 0%)	CTC (7, 8.05%)
Environment (12, 6.74%)	Environment (3, 2.07%)	Environment (36,41.37%)
Politics (10, 5.62%)	Politics (0, 0%)	Politics (0, 0%)

Multi-panel Comic	Single-panel Comic	Painting
Life (25, 73.53%)	Life (27, 18.88%)	Life (0, 0%)
Art (0, 0%)	Art (11, 7.69%)	Art (0, 0%)
Society (8, 23.53%)	Society (88, 61.54%)	Society (0, 0%)
CTC (0, 0%)	CTC (4, 2.80%)	CTC (119, 100%)
Environment (0, 0%)	Environment (2, 1.40%)	Environment (0, 0%)
Politics (1, 2.94%)	Politics (11, 7.69%)	Politics (0, 0%)

Main Results

Model	Overall (800)	Life (216)	Art (123)	Society (157)	Politics (21)	Env. (51)	CTC (130)	Positive (220)	Negative (247)	Neutral (231)
Open-source Models										
Qwen-VL-Chat	34.3	27.9	34.7	32.5	45.8	55.2	36.5	34.0	35.1	33.6
idefics2-8b	36.3	25.0	46.3	38.1	41.7	56.9	32.9	32.8	39.1	36.4
MiniCPM-Llama3-2.5	40.4	36.3	45.6	37.1	50.0	51.7	40.2	43.2	37.0	41.3
CogVLM2-Llama3-Chinese-Chat	43.4	37.1	48.3	42.3	54.2	63.8	40.2	40.3	45.7	43.8
MiniCPM-v2.6	45.0	37.5	47.6	49.5	58.3	55.2	42.3	45.6	44.6	44.9
LLaVA-1.6-34B	46.0	40.8	55.1	42.8	45.8	62.1	43.1	44.4	48.2	45.2
LLaVA-1.6-72B	48.0	43.8	48.3	49.5	70.8	60.3	43.8	41.5	52.5	49.2
Qwen2-VL-7B	49.6	42.5	51.7	54.1	62.5	65.5	44.5	50.2	47.5	51.2
GLM-4V-9b	50.3	46.7	48.3	53.6	54.2	62.1	48.2	51.9	52.9	46.3
InternVL2-Llama3-76B	52.9	50.8	53.7	51.0	58.3	67.2	51.1	54.8	51.8	52.3
InternVL2-8B	53.1	49.2	53.1	55.7	62.5	63.8	50.4	50.6	53.3	55.1
InternVL2-40B	57.9	55.8	55.1	61.9	62.5	70.7	52.6	54.4	58.0	60.8
Qwen2-VL-72B	**64.4**	**61.7**	**61.2**	**68.0**	**79.2**	**75.9**	**59.9**	**62.7**	**63.8**	**66.4**
Closed-source Models										
GPT-4o	54.1	54.1	55.8	52.1	50.0	63.8	51.8	51.9	56.2	54.1
Claude-3.5-Sonnet	54.1	52.1	61.9	52.6	62.5	46.6	53.3	52.7	56.5	53.0
Qwen-VL-MAX	56.9	53.3	59.2	58.8	62.5	67.2	52.6	53.9	58.3	58.0
Gemini-1.5 Pro	60.1	**60.0**	**63.3**	62.4	**70.8**	62.1	51.1	54.8	**65.6**	**59.4**
GLM-4V	**60.9**	55.0	59.9	**66.5**	66.7	**79.3**	**55.5**	**58.5**	64.5	**59.4**
Text-Only Models										
Llama-3-8B-Instruct	21.7	22.2	26.9	18.6	25.0	27.8	20.4	21.2	24.4	19.5
DeepSeek-67B-Chat	27.1	26.6	32.7	**30.9**	20.0	35.2	18.2	25.7	22.2	33.2
Qwen2-7B-Instruct	**32.5**	**33.2**	**34.6**	**30.9**	**35.0**	**40.7**	**28.5**	**33.6**	**30.4**	**33.6**
Humans										
Human_avg	78.2	81.0	67.7	82.7	87.7	84.0	65.9	77.9	75.2	81.6
Human_best	**81.0**	**83.2**	**73.6**	**87.2**	**89.5**	**86.0**	**66.7**	**78.2**	**78.8**	**83.3**

- Gap between Humans and MLLMs
There is a notable performance gap between MLLMs and humans. Models demonstrate the highest accuracy of **64.4%**, while human accuracy average at **78.2%** and best at **81.0%**.

- Disparity between Open-source and Closed-source Models
Closed-source models generally **outperform** open-source models, but the best-performing open-source model surpasses the top closed source model, with a difference of more than **3%**.

- Model Performance across Different Domains and Emotions
Models perform significantly worse in **Chinese traditional culture** compared to other domains, indicating that current models still lack sufficient understanding of Chinese culture. Further analysis shows that GPT-4o can only observe the **surface-level information**, it's difficult to deeply interpret the complex cultural elements contained in Chinese traditional painting.

- Analysis on different prompt skills
Incorporating image **emotion hints** into prompts generally improves model scores, indicating that models struggle with **emotional understanding**, leading to misinterpretation of the implicit meanings in the images.

Other Information

■ **Paper & Code**
PDF: https://arxiv.org/abs/2410.1385
Code: https://github.com/MING-ZCH/CII-Bench

■ **Authors**
Chenhao Zhang: ch_zhang@hust.edu.cn
Shiwen Ni: sw.ni@siat.ac.cn

ACL 2025 VIENNA