

CAN MLLMs UNDERSTAND THE DEEP IMPLICATION BEHIND CHINESE IMAGES?

Chenhao Zhang^{1,2*} Xi Feng^{2,3*} Yuelin Bai^{2*} Xinrun Du^{4,5*}
 Jinchang Hou^{2,3} Kaixin Deng⁶ Guangzeng Han⁷ Qinrui Li⁸ Bingli Wang⁹ Jiaheng Liu⁴
 Xingwei Qu¹⁰ Yifei Zhang¹¹ Qixuan Zhao^{2,3} Yiming Liang¹² Ziqiang Liu² Feiteng Fang^{2,3}

Min Yang² Wenhao Huang⁵ Chenghua Lin¹¹ Ge Zhang^{4,5†} Shiwen Ni^{2†}

¹Huazhong University of Science and Technology ²Shenzhen Institute of Advanced Technology, CAS

³University of Science and Technology of China ⁴M-A-P ⁵01.ai ⁶CDUT ⁷University of Memphis

⁸University of California, Santa Barbara ⁹SICAU ¹⁰University of Manchester ¹²SWU ¹²UCAS

ABSTRACT

As the capabilities of Multimodal Large Language Models (MLLMs) continue to improve, the need for higher-order capability evaluation of MLLMs is increasing. However, there is a lack of work evaluating MLLM for higher-order perception and understanding of Chinese visual content. To fill the gap, we introduce the **Chinese Image Implication understanding Benchmark, CII-Bench**, which aims to assess the higher-order perception and understanding capabilities of MLLMs for Chinese images. CII-Bench stands out in several ways compared to existing benchmarks. Firstly, 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 extensive experiments on CII-Bench across multiple MLLMs, we have made significant findings. Initially, a substantial gap is observed between the performance of MLLMs and humans on CII-Bench. The highest accuracy of MLLMs attains 64.4%, whereas human accuracy averages 78.2%, peaking at an impressive 81.0%. Subsequently, MLLMs perform worse on Chinese traditional culture images, suggesting limitations in their ability to understand high-level semantics and lack a deep knowledge base of Chinese traditional culture. Finally, it is observed that most models exhibit enhanced accuracy when image emotion hints are incorporated into the prompts. We believe that CII-Bench will enable MLLMs to gain a better understanding of Chinese semantics and Chinese-specific images, advancing the journey towards expert artificial general intelligence (AGI). Our project is publicly available at <https://cii-bench.github.io/>.



Figure 1: Comparision of Chinese and English image implications. Chinese images often embody richer scenes and deeper implications with Chinese traditional culture compared with the straightforward and explicit symbolism in English images.

*Equal Contribution. ✉ ch_zhang@hust.edu.cn; fengxi@ustc.edu

†Corresponding authors. ✉ gezhang@umich.edu; sw.ni@siat.ac.cn

1 INTRODUCTION

With the rapid advancement of artificial intelligence, Multimodal Large Language Models (MLLMs) (Liu et al., 2023b; Li et al., 2023c; Ye et al., 2023; Tong et al., 2024) have demonstrated exceptional performance across various domains, including natural language processing (Chowdhary & Chowdhary, 2020; Luo et al., 2024; Zhang et al., 2024a) and computer vision (Lu et al., 2022; Li et al., 2023b;a; Xu et al., 2023; Fu et al., 2023; Cai et al., 2023; Zhang et al., 2023; Chen et al., 2024b; Jin et al., 2024). These models are not only capable of processing and generating text but also excel at integrating and interpreting information across multiple modalities, such as images, sound, and video. However, despite the significant progress made in tasks like image recognition and generation, a crucial research question remains: Can these models truly understand and interpret images that have deep implications? (Liu et al., 2024b) construct an English image implication understanding dataset, II-Bench, and the experiments on MLLMs and human subjects reveal a substantial gap in the models' higher-order perception abilities, particularly in nuanced emotional understanding and profound meaning extraction, when compared to humans. Unfortunately, the rapid advancement of MLLMs has led to significant performance improvements. For instance, Claude-3.5-Sonnet has achieved an impressive accuracy of 80.9% on II-Bench, approaching the average human accuracy of 90.3%. This progress underscores the need for more challenging benchmarks that incorporate richer scenes and deeper implications to continue pushing the boundary of image implication understanding task.

In contrast to English images, Chinese images often embody richer scenes (Xu, 2023) and deeper implications as Figure 1 shows. For instance, Chinese traditional landscape paintings not only depict natural scenery but also convey profound philosophical concepts, such as the harmony between humans and nature, through artistic techniques like the interplay of void and solid, the use of negative space, and the brushwork. As the famous Chinese poet Su Shi noted, “Poetry and painting share the same essence, embodying both craftsmanship and purity”. The depth of Chinese images lies not only in their aesthetic appeal but also in the underlying spirit and philosophy they express. Similarly, New Year paintings, as a significant carrier of Chinese traditional culture, typically use symbolism and implication to convey wishes for good fortune, prosperity, and peace. Unlike the directness often found in English imagery, Chinese images emphasize the creation of atmosphere and subtle expression, requiring viewers to possess certain cultural knowledge to accurately grasp their meanings. This cultural disparity leads to significant differences in the modes of expression and meaning conveyed between Chinese and English images, highlighting the need to consider cultural context when evaluating the capability of MLLMs to understand the deep implications of images.

To address this gap, we develop CII-Bench, a benchmark designed to comprehensively test the higher-order perception, reasoning, and understanding abilities of models within a Chinese context. This benchmark allows us to gain a clearer understanding of these models' interpretive capacities, offering new insights into their application in cross-cultural environments, and thus advancing the research and development of MLLMs.

As illustrated in Figure 2, CII-Bench comprises 698 images and 800 multiple-choice questions spanning six domains: Life, Art, Society, Politics, Environment, and Chinese Traditional Culture. Moreover, to ensure diversity, CII-Bench includes six types of images: Illustration, Meme, Poster, Single-panel Comic, Multi-panel Comic, and Painting. By employing images of various types and from different domains, the benchmark provides a more robust evaluation of models' comprehension and reasoning abilities.

We conduct extensive experiments to evaluate CII-Bench on MLLMs that support Chinese and deeply evaluate the model's grasp of Chinese traditional culture. Our key contributions are as follows:

- We introduce CII-Bench, the first benchmark designed to assess the understanding of implications in Chinese images, which poses a significant challenge to current MLLMs.

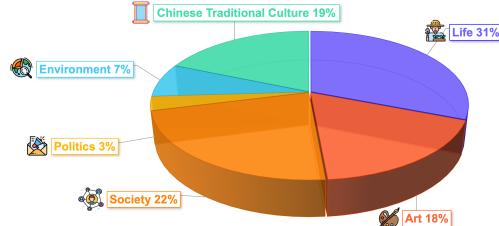


Figure 2: Composition of CII-Bench.

- We design a comprehensive evaluation metric based on GPT-4o to evaluate Chinese traditional culture. This metric aligns more closely with human annotations and is better suited for evaluating Chinese traditional painting.
- Our experimental findings are as follows: (1) 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%. (2) 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%. (3) 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. (4) 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.

2 RELATED WORK

2.1 MULTIMODAL LARGE LANGUAGE MODELS

With the rapid development of large language models (LLMs) (Aakanksha et al., 2022; Won et al., 2022; Chiang et al., 2023; Touvron et al., 2023; OpenAI, 2023a;b; Team, 2024; Cai et al., 2024), Multimodal Large Language Models (MLLMs) have made significant improvements. Many works incorporate additional module inputs on LLMs, effectively bridging the gap between visual and language. BLIP-2 (Li et al., 2023c) encodes images using ViT (Dosovitskiy et al., 2020) and employs a Q-Former to map visual features into the language space. LLaVA (Liu et al., 2023b;a; 2024a; Li et al., 2024a) utilizes an MLP as the connector between the visual encoder and the LLM backbone. Similarly, mPLUG-Owl2 (Ye et al., 2023) employs a modality-adaptive module to facilitate the collaboration between visual and language modalities by mapping them into a unified representation space. Subsequent works (Wang et al., 2023; Lu et al., 2024; Chen et al., 2024c; Young et al., 2024; Laurençon et al., 2024; GLM et al., 2024; Yao et al., 2024; Anthropic, 2024; Wang et al., 2024) further enhance MLLMs by designing novel modules for more sufficient modality alignment.

2.2 MLLM BENCHMARKS

The rapid advancement of MLLMs has emphasized the critical need for comprehensive evaluation frameworks within the research community. Initial benchmarks primarily focused on specific tasks, such as visual question answering (VQA) (Antol et al., 2015; Goyal et al., 2017; Kafle & Kanan, 2017; Singh et al., 2019; Hudson & Manning, 2019) and image captioning (Lin et al., 2014; Agrawal et al., 2019; Plummer et al., 2015). While these benchmarks have yielded significant insights, they fall short in providing a holistic assessment of MLLMs across the broader spectrum of multimodal perception and reasoning capabilities. To address this limitation, recent studies have developed more comprehensive evaluation approaches (Xu et al., 2023; Fu et al., 2023; Lu et al., 2022; Cai et al., 2023; Zhang et al., 2023; He et al., 2024; Chen et al., 2024b). For instance, MMBench (Liu et al., 2023c) and SEED (Li et al., 2023b;a) assess models' capabilities through common-sense questions, employing multiple-choice formats to evaluate various dimensions of ability. To assess specialized expertise, MMMU (Yue et al., 2023) and CMMMU (Zhang et al., 2024b) utilize content derived from exams and textbooks, enhancing the evaluation of domain-specific knowledge. Furthermore, Cambrian-1 (Tong et al., 2024) introduces a novel vision-centric benchmark (CV-Bench) to repurpose standard vision tasks for multimodal evaluation.

2.3 IMAGE IMPLICATION UNDERSTANDING

Image implication understanding represents a more complex and challenging task than conventional image understanding. This advanced cognitive process necessitates multi-hop reasoning ability and sophisticated theory of mind (ToM), capabilities that are intrinsic to human cognition (Desai et al., 2022; Hessel et al., 2023; Yang et al., 2024; Zhong et al., 2024; Strachan et al., 2024; Street et al., 2024; Horvitz et al., 2024). II-Bench (Liu et al., 2024b) is the first benchmark specifically designed to evaluate MLLMs in both image understanding and reasoning through English image implication.

生活 Life	艺术 Art	社会 Society
<p>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)每个人都想追求时尚的权利。 (D)羽毛熄灭蜡烛是一种天马行空的想法, 虽然失败了, 但这种创新值得称赞。 (E)图中人物多次用羽毛熄灭蜡烛, 表扬了坚持不懈、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。</p> <p>Image Type: 多格漫画(Multi-panel Comic) Rhetoric: 隐喻 Emotion: 消极 Difficulty Level: 简单</p>	<p>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)遥远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)远处的小岛意味着希望和目标, 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, 是人们面对外界压力时能够暂时寻求内心平静和惬意的地方。</p> <p>Image Type: 插画(Illustration) Rhetoric: 象征 Emotion: 积极 Difficulty Level: 简单</p>	<p>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, 却忽略了自身的原凶。 (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。</p> <p>Image Type: 单格漫画(Single-panel Comic) Rhetoric: 对比 Emotion: 消极 Difficulty Level: 简单</p>
<p>中华传统文化 Chinese Traditional Culture</p> <p>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)酒身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已难求的悲伤。 (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。</p> <p>Image Type: 绘画(Painting) Rhetoric: 隐喻 Emotion: 积极 Difficulty Level: 困难</p>	<p>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。</p> <p>Image Type: 海报(Poster) Rhetoric: 象征 Emotion: 消极 Difficulty Level: 中等</p>	<p>政治 Politics</p> <p>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, 人们也可能怀抱对和平的渴望。</p> <p>Image Type: 插画(Illustration) Rhetoric: 隐喻, 对比 Emotion: 消极 Difficulty Level: 困难</p>

Figure 3: CII-Bench examples sampled from each domain. The English version in Appendix B.

3 THE CII-BENCH

3.1 OVERVIEW OF CII-BENCH

We present the **Chinese Image Implication Understanding Benchmark** (CII-Bench), a novel benchmark designed to assess the perceptual, reasoning, and comprehension abilities of MLLMs in the context of Chinese imagery. This benchmark includes a diverse range of visual content such as traditional Chinese traditional artworks, comics, posters, and Chinese Internet memes, all rich in visual information and cultural significance. The main goal of CII-Bench is to evaluate if current MLLMs can leverage their understanding and knowledge of Chinese culture to accurately interpret the deeper implications and abstract information within these images.

CII-Bench comprises 698 images across various categories, with detailed classification and domain statistics provided in Appendix A. These images are manually collected and annotated by 30 undergraduate students from different disciplines and institutions, sourced from several well-known image websites. Each image is paired with 1 to 3 multiple-choice questions, each offering six options with only one correct answer. One fixed question asks, “What is the implication in this image?” Additional questions for the same image probe different levels of understanding, such as overarching interpretation and nuanced details. The benchmark includes 800 multiple-choice questions, with 765 for the test set and 35 for developing and validating few-shot tasks. Figure 3 provides representative examples from CII-Bench.

3.2 DATA CURATION PROCESS

3.2.1 DATA COLLECTION

We collect 17,695 raw images from various renowned illustration websites, ensuring a sufficiently extensive raw dataset. Our collectors are well instructed to adhere to copyright and license regulations, avoiding data from sites prohibiting copy and redistribution. For detailed information on the specific websites from which we collect images, please refer to Appendix C.

3.2.2 DATA FILTRATION

After collecting the raw images, we meticulously design a three-stage data filtering process: In the first stage, we focus on image deduplication. We utilize image similarity algorithms for pixel-level comparison to eliminate duplicates and preserve dataset uniqueness; In the second stage, we regulate text prevalence in images. Optical Character Recognition (OCR) technology identifies textual areas and disqualifies images exceeding set text-area ratios, maintaining a visual-centric dataset; In the third stage, images undergo rigorous visual inspection, discarding those without metaphorical depth based on strict criteria. This process refines the dataset, rejecting over 95% of initial images and securing under 1,000 high-quality ones.

3.2.3 DATA ANNOTATION

The annotation process for the benchmark was meticulously designed through several steps to ensure rigor and precision as following. The detailed annotation protocol can be found in Appendix C.

Preparation and Consistency Check: Before formal annotation, annotators first acquaint themselves with standard templates and guidelines. A pre-annotation round on a shared image batch ensures uniform standard understanding, with discrepancies resolved through discussion.

Multiple Rounds of Annotation and Cross-Validation: To reduce bias, each image receives annotations from two different annotators. Cross-validation follows, with a third-party review for significant discrepancies, guaranteeing accuracy.

Refinement of Annotation Content: Annotators annotate each image’s difficulty, type, emotional label, domain, and rhetorical devices based on specific criteria, ensuring consistency and comparability. They also craft 1 to 3 refined questions per image, each with one correct answer among five distractor options, including the default question, “What is the implication in this image?”

Context Analysis: During the annotation process, annotators assess the image’s cultural and background significance, especially for implications and rhetorical devices, consulting relevant materials for accuracy.

Post-Annotation Review: Upon completion, annotations undergo a thorough quality review for any oversight, errors, or inconsistencies. Based on the evaluation results, feedback is provided to the annotators, with re-annotations as necessary to maintain data quality.

3.3 DATASET STATISTICS

CII-Bench comprises 698 images, each accompanied by 1 to 3 multiple-choice questions, totaling 800 questions. We randomly select 35 of these questions to construct a few-shot development set and validation set. On average, each question is approximately 11 characters long, while each option has an average length of 28 characters. Additionally, each image is supplemented with a manually written description by the annotators, which provides a detailed explanation of the image’s content, nuances, and the human interpretation of its deep implication.

CII-Bench covers images across six distinct domains: Life, Art, Society, Politics, Environment, and Chinese Traditional Culture. The types of images are diverse, including Illustration, Meme, Poster, Single-panel Comic, Multi-panel Comic, and Painting. Based on human understanding, these images are categorized into three levels of difficulty: Easy, Medium, and Hard. Moreover, the images are classified according to the emotional information they convey: Positive, Neutral, or Negative. Each image is also manually annotated with the rhetorical devices employed, including Metaphor, Exaggeration, Symbolism, Visual Dislocation, Antithesis, Analogy, Personification, and Contrast. Detailed statistical information is provided in Appendix A.

4 EXPERIMENT

We conduct systematic experiments on both open-source and closed-source MLLMs using CII-Bench. For each model, we employ eight different configurations: None (zero-shot), 1-shot, 2-shot, 3-shot, CoT, Domain, Emotion, and Rhetoric. “None” represents the use of a standard prompt without any additional information. “Emotion” indicates the inclusion of information related to the

emotional polarity of the image (e.g., positive, negative) in the prompt, “Domain” involves adding information about the image’s domain (e.g., life, art), and “Rhetoric” refers to including details about the rhetorical devices used in the image (e.g., metaphor, contrast) in the prompt. Additionally, to verify the necessity of images in problem-solving, we select a portion of LLMs to complete tasks without image input. For consistency across all MLLMs and LLMs, we use identical prompts and experiment setup, with specific details available in Appendix D.

4.1 BASELINES

MLLMs. To comprehensively evaluate CII-Bench, we carefully select a diverse range of MLLMs, encompassing both open-source and closed-source models, with the aim of covering a wide spectrum of model characteristics and scales. These models span parameter sizes from 7B to 100B, ensuring that models of varying complexity and capability are thoroughly assessed. In selecting the models, we focus on the following key aspects: 1) model diversity, 2) Open-Source vs. Closed-Source models, and 3) model parameter scaling law.

LLMs. To verify the critical role of images in answering questions, we specifically design an experiment in which some LLMs participate in the task without any image input. The purpose of this experiment is to assess whether these models can accurately understand the questions and make correct choices in the absence of image information, thereby further demonstrating the importance of images in the comprehension and problem-solving process. We select DeepSeek-67B, LLaMA-3-8B, and Qwen2-7b as the LLMs used in this experiment.

Evaluation. We use accuracy as the primary evaluation metric, multi-choice format questions and answer extraction method, which are widely used in previous benchmarks such as Helleswag (Zellers et al., 2019), MMMU (Yue et al., 2023), CMMMU (Zhang et al., 2024b), MMLU (Li et al., 2024b) and so on. Since CII-Bench is entirely composed of multiple-choice questions, the evaluation process only requires extracting the selected option from the model’s response, which simplifies the complexity of rule design. It is important to note that when models use chain-of-thought (CoT) prompts, the responses may include intermediate steps. Therefore, the evaluation rules must be sufficiently robust, or the model’s output must follow a fixed format. If the selected option cannot be extracted from the model’s response, the model is considered to have answered the question incorrectly. For the detailed statistics of the model output, please see Appendix F. For reference, we also select three Chinese PhD students to evaluate human performance on CII-Bench.

4.2 MAIN RESULTS

In this section, we conduct a comprehensive comparison of the performance of various MLLMs, LLMs, and humans on CII-Bench. Detailed results across different domains and emotional dimensions are presented in Table 1, while different image types, difficulty levels, and rhetoric can be found in Appendix E. The main experimental results and findings are summarized as follows:

4.2.1 NATURAL CHALLENGES OF CII-BENCH

This benchmark presents a significant challenge for current models. Notably, despite GPT-4o being an advanced model, its accuracy is only 54.1%, indicating substantial room for improvement. This reflects the rigorous and demanding nature of the benchmark. Further analysis reveals that most models perform worst in the domain of Chinese traditional culture, highlighting a significant deficiency in their understanding of Chinese cultural nuances. It is also noteworthy that human performance in this domain is not ideal, as questions related to Chinese traditional culture often require deep cultural knowledge. The lack of this knowledge base poses difficulties for both models and humans when dealing with Chinese cultural content. In addition, text-only models like DeepSeek-67B-Chat only get 27.1% accuracy, which shows that most of the questions in CII-Bench require image information to be answered correctly, proving that CII-Bench is highly dependent on visual content (Chen et al., 2024a).

4.2.2 GAP BETWEEN HUMANS AND MLLMs

The results indicate a significant gap between human performance and multimodal large models (MLLMs) on CII-Bench. Human participants achieved an average accuracy of 78.2%, with the

Model	Overall (800)	Life (216)	Art (123)	Society (157)	Politics (21)	Env. (51)	CTC (130)	Positive (220)	Negative (247)	Neutral (231)
<i>Open-source Models</i>										
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	<u>55.1</u>	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	<u>70.8</u>	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	<u>54.8</u>	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	<u>57.9</u>	<u>55.8</u>	<u>55.1</u>	<u>61.9</u>	62.5	<u>70.7</u>	<u>52.6</u>	54.4	<u>58.0</u>	<u>60.8</u>
Qwen2-VL-72B	64.4	61.7	61.2	68.0	79.2	75.9	59.9	62.7	63.8	66.4
<i>Closed-source Models</i>										
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	<u>61.9</u>	52.6	62.5	46.6	<u>53.3</u>	52.7	56.5	53.0
Qwen-VL-MAX	56.9	53.3	59.2	58.8	62.5	<u>67.2</u>	52.6	53.9	58.3	58.0
Gemini-1.5 Pro	60.1	60.0	63.3	<u>62.4</u>	70.8	62.1	51.1	<u>54.8</u>	65.6	59.4
GLM-4V	60.9	<u>55.0</u>	59.9	66.5	<u>66.7</u>	79.3	55.5	<u>58.5</u>	64.5	59.4
<i>Text-Only Models</i>										
Llama-3-8B-Instruct	21.7	22.2	26.9	18.6	<u>25.0</u>	27.8	<u>20.4</u>	21.2	<u>24.4</u>	19.5
DeepSeek-67B-Chat	<u>27.1</u>	<u>26.6</u>	<u>32.7</u>	30.9	20.0	<u>35.2</u>	18.2	<u>25.7</u>	22.2	<u>33.2</u>
Qwen2-7B-Instruct	32.5	33.2	34.6	30.9	35.0	40.7	28.5	33.6	30.4	33.6
<i>Humans</i>										
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

Table 1: Overall results of different MLLMs, LLMs and humans on different domains and emotions. The best-performing model in each category is **in-bold**, and the second best is underlined.

highest accuracy reaching 81.0%. In contrast, the best-performing closed-source model, GLM-4V, achieved an accuracy of 60.9%, while the top open-source model, Qwen2-VL-72B, scored 64.4%. These findings highlight the substantial disparity between human abilities and even the most advanced models in understanding image implications. The highest accuracy achieved by the models is considerably lower than the average human score, indicating that multimodal large models still face significant challenges in this domain.

4.2.3 MODEL PERFORMANCE ACROSS DIFFERENT DOMAINS AND EMOTIONS

In terms of domain performance, our results in Table 1 indicate that the models generally perform better in the Environment and Politics domains, achieving higher accuracy. Conversely, the accuracy is lower in the Life and Society domains, proving that everyday metaphors are generally more difficult in the Chinese context. The lowest score for the Chinese Traditional Culture and Art domains, which shows that while the models generalize well in common domains, they struggle with the more abstract and logically demanding information found in Chinese Traditional Culture and Art.

From an emotional perspective, the models tend to exhibit higher accuracy when the image implications convey negative emotions, while accuracy is the lowest for images with positive emotions. This discrepancy highlights that the models' preferences do not align with those of humans, as humans are significantly more sensitive to positive implications. The performance of the model is opposite to the conclusion shown in II-Bench (Liu et al., 2024b), reflecting the obvious difference in emotional expression in the Chinese and English contexts.

4.2.4 ANALYSIS ON DIFFERENT PROMPT SKILLS

Analysis of Chain-of-Thought (CoT). In Table 2, we evaluate the impact of Chain-of-Thought (CoT) prompting on model performance. The results indicate that CoT does not significantly improve the accuracy of the models. In some cases, particularly with smaller open-source models, the accuracy even declined when CoT was used. For example, MiniCPM-v2.6 scores 45.0% without CoT, but this drops to 38.9% with CoT; similarly, LLaVA-1.6-72B scores decrease from 48.0% to 45.3%.

Model	None	CoT	Domain	Emotion	Rhetoric
<i>Open-source Models</i>					
Qwen-VL-Chat	34.3	34.0	32.1	35.0	33.4
idefics2-8b	36.3	33.3	37.5	38.6	37.4
MiniCPM-Llama3-2.5	40.4	35.8	41.1	39.0	34.8
CogVLM2-Llama3-Chinese-Chat	43.4	42.6	43.5	44.0	43.4
MiniCPM-v2.6	45.0	38.9	44.4	45.4	45.4
LLaVA-1.6-34B	46.0	44.5	46.4	47.1	45.4
LLaVA-1.6-72B	48.0	45.3	47.3	48.6	45.4
Qwen2-VL-7B	49.6	50.0	51.0	50.8	49.3
GLM-4V-9b	50.3	49.1	49.9	51.1	49.5
InternVL2-Llama3-76B	52.9	52.6	54.1	52.8	53.5
InternVL2-8B	53.1	47.9	53.5	56.3	53.8
InternVL2-40B	<u>57.9</u>	<u>57.6</u>	<u>57.1</u>	<u>60.0</u>	<u>57.9</u>
Qwen2-VL-72B	64.4	62.1	66.0	64.3	63.0
<i>Closed-source Models</i>					
GPT-4o	54.1	54.9	55.4	54.9	51.9
Claude-3.5-Sonnet	54.1	51.6	56.4	53.5	54.9
Qwen-VL-MAX	56.9	54.0	<u>59.1</u>	<u>59.9</u>	54.8
Gemini-1.5 Pro	<u>60.1</u>	<u>54.1</u>	59.0	57.9	<u>55.6</u>
GLM-4V	60.9	48.8	60.4	60.6	58.8

Table 2: Overall results of different prompts on CII-Bench. The label (*Emotion*, *Domain*, *Rhetoric*) means providing corresponding information for the images in the prompt. The best-performing model in each category is **in-bold**, and the second best is underlined.

Upon analyzing the models’ responses, we find that those models showing a decrease in accuracy with CoT often suffer from overinterpretation, where questions that were initially answered correctly are misinterpreted after CoT is applied. Additionally, for questions that were originally answered incorrectly, CoT does not lead to significant improvements and sometimes even causes confusion, such as selecting multiple options. However, for most models, the probability of failing to extract an answer option from the response decreases after using CoT, which explains why some models show improved accuracy with CoT.

Analysis of Different Types and Domains. To evaluate the impact of different label information on model accuracy, we conduct an ablation study by providing relevant label information (such as emotion, domain, and rhetoric) in the prompts. The results in Table 2 show that emotion labels significantly improve model accuracy, followed by domain and rhetoric labels, both of which exhibit similar effectiveness.

This result aligns with human intuition. The answer options typically include negative, positive, and neutral choices. When the model receives emotional information, it can eliminate some irrelevant options, naturally leading to higher accuracy. In contrast, domain and rhetoric information generally do not effectively help the model eliminate options, resulting in more limited improvements. Additionally, from a model training perspective, models tend to have a more mature understanding of emotions, while specific nouns in rhetoric and domain labels are often custom-defined. During pre-training, the model may not have encountered a large number of descriptions for such specific nouns, making these labels less helpful in improving accuracy.

Analysis of Few-shot Examples. The results in Table 3 indicate that few-shot examples do not improve the models’ accuracy. Specifically, performance declines as the number of examples increases. This decline can be attributed to the models’ inferior capabilities in handling multiple images compared to single images, leading to a decrease in accuracy with a higher number of shots. Furthermore, as the number of shots increases, the input length also extends, and the models’ ability to process long texts is inadequate, resulting in suboptimal performance with long contexts.

4.3 EVALUATION OF CHINESE TRADITIONAL CULTURE

The Chinese traditional culture category is a distinctive feature of the CII-Bench dataset, where MLLMs consistently score the lowest. Therefore, we need a deeper evaluation of this field to analyze

Model	None	1-shot	2-shot	3-shot
Qwen2-VL-7B	49.6	44.1	39.3	37.5
GPT-4o	54.1	51.8	49.5	49.1
Claude-3.5-Sonnet	54.1	55.4	55.3	55.4
InternVL2-40B	57.9	53.0	47.1	41.9
Gemini-1.5 Pro	60.1	57.4	55.8	55.4

Table 3: Few-shot results of different models on the CII-Bench.

the extent to which MLLM understands Chinese traditional culture. We chose to deeply analyze MLLM’s understanding of Chinese traditional culture by evaluating Chinese traditional paintings.

4.3.1 EVALUATION METRIC

Chinese traditional painting, a cornerstone of Chinese traditional culture, encompasses a rich tapestry of styles and techniques developed over millennia. These paintings are typically categorized based on their subject matter (e.g., landscape paintings, flower-and-bird paintings, figure paintings, and New Year paintings) or their stylistic and skill (e.g., court paintings, meticulous brush paintings, freehand brush paintings, and color-and-ink paintings). Each category embodies unique characteristics that reflect China’s artistic evolution and philosophical underpinnings.

To comprehensively assess MLLMs’ understanding of Chinese traditional paintings, we develop a multifaceted evaluation metric. This metric is designed to probe both the surface-level information readily apparent in the artwork and the deeper culture and history that informs its creation and interpretation. Our evaluation metric encompasses five key perspectives: ***Surface-level Information***, ***Aesthetic Characteristics***, ***Brush and Ink Skills***, ***Culture and History***, and ***Deep Implications***. For each perspective, we give its detailed description in Figure 4.

Evaluation Metric	Evaluation Standard
1. Surface-level Information:	[1 point]: The description of the picture is incomplete and does not include any background information. It only mentions the most obvious elements in the picture, lacking recognition and understanding of traditional cultural elements.
2. Aesthetic Characteristics:	[2 points]: The description of the image is relatively complete, but the background information is limited. Some traditional cultural elements can be identified, but their meaning is not explained in depth. There is a lack of insight into the aesthetic characteristics or skills of the image.
3. Brush and Ink Skills:	[3 points]: The description of the image is comprehensive and includes some background information. The traditional cultural elements can be accurately identified and briefly explained. There is a preliminary understanding of the aesthetic characteristics or skills of the picture, but there is a lack of in-depth analysis. Some historical or cultural background is mentioned, but it is not fully expanded.
4. Culture and History:	[4 points]: The description of the image is detailed and contains rich background information. It is able to explain the significance of traditional cultural elements in depth. It provides a relatively in-depth analysis of the aesthetic characteristics and skills of the image. It provides relevant historical and cultural background and attempts to explore the implication of the image, but the implication is misunderstood.
5. Deep Implications:	[5 points]: The description of the image is comprehensive and accurate, and the background information is in-depth and comprehensive. It can not only accurately interpret traditional cultural elements, but also reveal their deep meaning. It has unique insights into the aesthetic characteristics and skills of the image. It deeply expounds on the relevant historical and cultural background, and can see through the philosophical thoughts, humanistic spirit or social values contained in the image, showing a deep understanding of Chinese traditional culture.

Figure 4: Evaluation metric and evaluation standard of Chinese traditional painting.

4.3.2 LLM-BASED CHINESE TRADITIONAL PAINTING AUTOMATIC EVALUATION

To evaluate Chinese traditional painting comprehension in MLLMs, we develop an LLM-based evaluation standard based on evaluation metrics, as illustrated in Figure 4. Our experiment utilize the CTC domain data from CII-Bench, comprising 130 Chinese traditional paintings. We employ human-written descriptions and implication interpretations as ground truth. We choose GPT-4o to

generate descriptions for these images, which are subsequently scored using GPT-4o and our evaluation standard. Please see the evaluation prompt in Appendix D. To validate the model’s scoring efficacy, we enlist three PhD students well-versed in Chinese metaphorical imagery to independently score the 130 paintings.

The model-human scoring consistency reached 98%, affirming the method’s validity for assessing Chinese traditional painting comprehension. Table 4 presents the detailed model scores. Analysis of these results, in conjunction with our evaluation standard, reveals insights across three dimensions: overall performance, difficulty levels, and emotions. The overall score of 2.71 indicates that while MLLM is able to observe the surface-level information of paintings, it has a large gap with humans in deeply interpreting the complex cultural elements contained in Chinese traditional art. In terms of difficulty evaluation, the model is consistent with human cognition, while in terms of emotion, the model has a higher negative score, indicating that the model can identify negative implications in paintings, such as using the past to satirize the present, and not appreciating talents.

Model	Overall	Easy	Middle	Difficult	Positive	Negative	Neutral
GPT-4o	2.71	3.0	3.2	2.35	2.63	3.0	2.82

Table 4: Overall result of Chinese traditional painting.

4.4 ERROR ANALYSIS

To conduct a comprehensive error analysis of GPT-4o’s performance (under CoT setting) on CII-Bench, we randomly select a total of 100 erroneous samples from various domains, distributed according to their proportions in the dataset. These samples are subjected to in-depth analysis by expert annotators. As illustrated in Figure 5, GPT-4o’s errors can be categorized into the following types: Information Neglect, Misunderstanding of Visual Information, Over-Inference, Superficial Reasoning, and Lack of Cultural Background Knowledge. For detailed analysis of cases, please see the Appendix G.

Information Neglect (36%):

Complex images contain both visual and textual elements. Sole reliance on visual information makes accurate interpretation challenging due to diversity in meaning. Incorporating textual information clarifies the author’s emotional intent, aiding accurate interpretation. Unfortunately, GPT-4o frequently overlooks key visual (13%) and textual (23%) information. When directly asked about these elements, we find that GPT-4o can often answer correctly, indicating two main issues: 1) Insufficient image recognition abilities, and 2) Significant shortcomings in multimodal fusion, leading to underutilization of acquired information.

Over-Inference (25%):

During answer construction, distractors are included at surface and deep levels. GPT-4o often selects more exaggerated, deep-level incorrect options, ignoring narrower but correct ones, especially in Chinese memes. This suggests that GPT-4o has a preference for selecting abstract options.

Lack of Cultural Background Knowledge (16%):

CII-Bench requires a model’s deep understanding of Chinese traditional culture. Lacking knowledge of traditional symbols, historical figures, and classical allusions, GPT-4o struggles with interpreting deeper implications within images. Despite reasonable Chinese language handling, the model’s cultural deficiency affects its reasoning and performance.

Superficial Reasoning (12%):

Understanding extended meanings within images is crucial. However, GPT-4o often only focus on surface-level elements, neglecting the deep implications and deeper cultural connotations behind

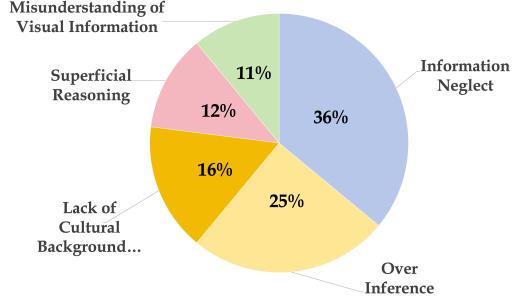


Figure 5: GPT-4o error responses distribution.

them. This superficial reasoning hinders the model from fully grasping profound messages that the artist or designer intends to convey.

Misunderstanding of Visual Information (11%):

Accurate identification of visual information is vital. We find that GPT-4o sometimes misidentifies visual elements within images, particularly when dealing with abstract images. The abstract nature of such images often stems from the inclusion of exaggerated imaginative elements, sometimes even defying physical laws. Therefore, correctly identifying these abstract elements requires the model to have a deep understanding of the essence of objects, a capability that current models clearly do not yet possess.

5 DISCUSSION

5.1 INTERPRETABILITY ANALYSIS OF CHINESE IMAGE IMPLICATIONS

The essence of Chinese image implications is deeply rooted in deep cultural heritage and complex contextual associations, which enables them to convey profound messages through nuanced expressions. For example, in traditional Chinese art forms such as landscape and New Year paintings, the imagery transcends mere depiction of nature or daily occurrences. Instead, it embodies emotions, philosophical insights, and societal norms through metaphorical and highly symbolic expressions. These symbols, like the pine tree, plum blossom, and crane, are not superficial meaning but are steeped in centuries of cultural tradition, representing resilience, purity, and longevity.

However, deciphering these complex messages can be challenging, particularly for those unfamiliar with the cultural and historical narratives tied to these symbols. This contrasts with English image implications, which often convey messages through more straightforward and explicit symbolism. As a result, the interpretability of Chinese image implications depends to some extent on reconstructing and resonating with the cultural context, which is what makes them unique: their meaning is not only visual but also culturally resonant, bridging across time and space.

Moreover, the interpretability of Chinese image implications has new changed in the modern era. Globalization and the surge of internet culture have intertwined foreign elements with traditional Chinese culture, birthing new symbols and implications. This intersection introduces additional layers of meaning, complicating the interpretation of traditional symbols.

5.2 WHY CHOOSE CHINESE TRADITIONAL PAINTINGS TO EVALUATE CHINESE TRADITIONAL CULTURE?

The imagery associated with Chinese traditional culture often embodies complex implications, encompassing customs, historical anecdotes, and legendary tales, making direct evaluation particularly challenging. Chinese traditional painting, intrinsically intertwined with Chinese traditional culture, offers a viable proxy for this assessment. The unique value of Chinese traditional painting lies in its embodiment of Chinese cultural connotations, aesthetic implications, and distinctive artistic expression. The core philosophical concepts of Confucianism, Taoism, and Buddhism, along with their humanistic essence, have consistently permeated the entire trajectory of Chinese painting history. Consequently, we have chosen to evaluate MLLMs' comprehension of Chinese traditional culture through an in-depth analysis of their understanding of Chinese traditional paintings.

6 CONCLUSION

The development of CII-Bench marks a significant step forward in evaluating the capabilities of multimodal large models (MLLMs) and brings us closer to achieving expert artificial general intelligence (AGI). This benchmark promotes a deeper exploration of the higher-order theory of mind in MLLMs. Experimental results indicate that current MLLMs still exhibit a significant gap compared to humans in understanding the implications of images within a Chinese context. We found that most MLLMs lack a deep knowledge base of Chinese traditional culture, leading to a superficial understanding of this cultural content. Finally, the experiments showed that incorporating image emotion hints into prompts often improves model performance, suggesting that models still struggle

with emotional understanding, which in turn leads to misinterpretation of implications. We believe that CII-Bench will inspire the academic community to further develop the next generation of multimodal foundational models that move toward expert AGI.

LIMITATIONS

We acknowledge several limitations in our study. Although CII-Bench is comprehensive, subjective elements can result in varying interpretations, impacting result consistency. In addition, in order to ensure high quality and practicability, our benchmark is not particularly large. The evaluation metrics may not fully capture the advanced understanding and reasoning capabilities of AI systems. These limitations underscore the necessity for continuous refinement and expansion of our benchmarks. Future work will focus on developing and incorporating more stringent and objective test sets to enhance the reliability and validity of our benchmark.

ETHICS STATEMENT

In developing CII-Bench, we strictly adhere to ethical guidelines and legal regulations, ensuring fairness, transparency, inclusivity and respect for all stakeholders. We stress the importance of safeguarding privacy and intellectual property rights, underscoring our commitment to responsible and lawful data management. We have taken steps to anonymize any personal data to protect privacy and have made every effort to minimize harmful or biased content. However, we recognize that biases can inadvertently arise and some information may be potentially offensive. We are committed to continuous monitoring and improvement to mitigate such biases. Furthermore, we encourage users of our dataset to employ it responsibly and to consider the ethical implications of their work, particularly in applications that may impact individuals or communities.

REFERENCES

- Sharan Aakanksha, Maarten Jacob, Adam Gaurav, Chung Paul, Sebastian Charles, Kensen Parker, Joshua Sasha, et al. Palm: Scaling language modeling with pathways. 2022.
- Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. 2019.
- Anthropic. Claude 3.5 sonnet model card addendum. 2024.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. 2015.
- Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Xing Luo, Chenyu Yi, and Alex Kot. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. *arXiv preprint arXiv:2312.02896*, 2023.
- Zheng Cai, Maosong Cao, Haojong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, et al. Internlm2 technical report, 2024.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024a.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, et al. Are we on the right way for evaluating large vision-language models?, 2024b.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024c.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023.

- KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pp. 603–649, 2020.
- Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. *AAAI*, 2022.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. 2017.
- Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning, 2024.
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest. 2023.
- Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. Getting serious about humor: Crafting humor datasets with unfunny large language models. 2024.
- Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. 2019.
- Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal theory of mind question answering. 2024.
- Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. 2017.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, et al. Llava-onevision: Easy visual task transfer, 2024a.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023a.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023b.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, et al. Cmmlu: Measuring massive multitask language understanding in chinese, 2024b.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023c.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. 2014.

- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.
- Haotian Liu, Chunyuan Li, Qingsong Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next/>, 2024a.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023c.
- Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, et al. Ii-bench: An image implication understanding benchmark for multimodal large language models, 2024b.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, et al. Deepseek-vl: Towards real-world vision-language understanding, 2024.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. 2022.
- Zihan Luo, Xiran Song, Hong Huang, Jianxun Lian, Chenhao Zhang, Jinqi Jiang, and Xing Xie. Graphinstruct: Empowering large language models with graph understanding and reasoning capability, 2024.
- OpenAI. Chatgpt. <https://chat.openai.com/>, 2023a.
- OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023b.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. 2015.
- Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. 2019.
- James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 2024.
- Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Robin IM Dunbar, et al. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*, 2024.
- Gemini Team. Gemini: A family of highly capable multimodal models, 2024.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, , et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.
- Chung Won, Hou Le, Longpre Shayne, Zoph Barret, Tay Yi, Fedus William, Li Yunxuan, et al. Scaling instruction-finetuned language models, 2022.

- Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. LvLM-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.
- Qingshu Xu. Comparing covid-19 metaphors in chinese and english social media with critical metaphor analysis. *Frontiers in Psychology*, 2023.
- Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. Can large multimodal models uncover deep semantics behind images?, 2024.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.
- Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019.
- Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. 2024a.
- Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, et al. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark, 2024b.
- Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *arXiv preprint arXiv:2306.05179*, 2023.
- Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation, 2024.

A STATISTICS OF CII-BENCH

Statistics		Statistics	
Total Questions	800	Life	216 (30.95%)
Total Images	698	Art	123 (17.62%)
Dev : Validation : Test	15 : 20 : 765	Society	157 (22.49%)
Easy : Medium : Hard	305 : 282 : 111	Environment	51 (7.31%)
Average Question Length	10.54	Politics	21 (3.01%)
Average Option Length	28.31	Chinese Traditional Culture	130 (18.62%)
Average Explanation Length	121.06		
Metaphor	562	Positive	220 (31.52%)
Exaggerate	121	Neutral	247 (35.39%)
Symbolism	236	Negative	231 (33.09%)
Visual Dislocation	42		
Antithesis	13	Illustration	178 (25.50%)
Analogy	19	Meme	145 (20.77%)
Personification	73	Poster	87 (12.46%)
Contrast	87	Multi-panel Comic	34 (4.87%)
		Single-panel Comic	143 (20.49%)
		Painting	119 (17.05%)

Table 5: General statistics of CII-Bench.

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%)

Figure 6: CII-Bench specific image type and domain statistics.

B CII-BENCH EXAMPLES OF ENGLISH VERSION

Life	Art	Society
<p>Question: What metaphor does this image contain?</p> <p>Option:</p> <p>(A) Some people pursue certain things considering their own circumstances.</p> <p>(B) Everyone has the right to pursue a feather is an imaginative idea, satirizing unrealistic and reckless attempts.</p> <p>(C) Everyone has the right to pursue fashion.</p> <p>(D) Extinguishing a candle with a feather is an imaginative idea; although it failed, this kind of innovation is commendable.</p> <p>(E) The repeated attempts by the person in the picture to extinguish the candle with a feather praise the spirit of perseverance and resilience.</p> <p>(F) The unreasonable act of trying to extinguish a candle with a feather satirizes those who act without careful consideration and blindly try things.</p> <p>Image Type: Multi-panel Comic Rhetoric: Metaphor Emotion: Negative Difficulty Level: Easy</p>	<p>Question: What does the distant island suggest?</p> <p>Option:</p> <p>(A) The distant island is depicted as a symbol of spiritual sustenance, representing a place where people seek solace and refuge in the real world.</p> <p>(B) The distant tree, together with the distant river, creates an atmosphere of harmony with nature, suggesting a harmonious relationship between humans and the natural world.</p> <p>(C) The distant island is a symbol of a goal, which, although far away, is still attainable.</p> <p>(D) The distant island contains specific cultural landscapes or historical sites.</p> <p>(E) The distant island, as a landmark in the distance, reminds people of the importance of paying attention to and respecting history and culture.</p> <p>(F) The distant island symbolizes challenges, a place of inner peace within an individual, a place where people seek calm and recovery when facing external pressure and challenges.</p> <p>Image Type: Illustration Rhetoric: Symbolism Emotion: Positive Difficulty Level: Easy</p>	<p>Question: What metaphor does this image contain?</p> <p>Option:</p> <p>(A) Perseverance is an important virtue.</p> <p>(B) Parents' behavior and habits determine the future of their children.</p> <p>(C) The failure of education is due to parents not providing enough supervision.</p> <p>(D) Piano learning should start at an early age and be persisted with, as this is the path to success.</p> <p>(E) Some parents blame the failure of education on their children, while neglecting their own responsibility.</p> <p>(F) If parents do not lead by example and become role models, their children's education will inevitably fail.</p> <p>Image Type: Single-panel Comic Rhetoric: Contrast Emotion: Negative Difficulty Level: Easy</p>
<p>Chinese Traditional Culture</p> <p>Question: What metaphor does this image contain?</p> <p>Option:</p> <p>(A) The desolate winter scene hints at the character's longing for the arrival of spring and the revival of all things.</p> <p>(B) Viewing the snowy landscape among the snowflakes, the character's serene and detached attitude towards worldly matters.</p> <p>(C) Enjoying the snow alone implies the character's inner loneliness and the sorrow of finding it hard to meet a kindred spirit.</p> <p>(D) The act of looking up suggests that the character is in deep thought.</p> <p>(E) Being alone hints at the character's longing for family and homeland.</p> <p>(F) The bleak winter scene hints at the character's inner sadness.</p> <p>Image Type: Painting Rhetoric: Metaphor Emotion: Positive Difficulty Level: Hard</p>	<p>Environment</p> <p>Question: What metaphor does this image contain?</p> <p>Option:</p> <p>(A) Symbols that creatures in the natural world are severely impacted by human activities, even facing the threat of extinction.</p> <p>(B) This image represents the rapid development of industrial technology, implying that future life will be more convenient and prosperous.</p> <p>(C) The image depicts the application of environmental technologies, depicting a harmonious coexistence between industry and nature.</p> <p>(D) Suggests that people have the ability to mitigate damage to the natural environment through changes in behavior, the adoption of new technologies, and the implementation of environmental policies, making sustainable development and ecological balance possible.</p> <p>(E) The image symbolizes humanity's conquest of nature by altering the Earth's surface through technology.</p> <p>(F) Expresses deep concern about environmental protection and care of the natural environment while pursuing industrial development.</p> <p>Image Type: Poster Rhetoric: Symbolism Emotion: Negative Difficulty Level: Medium</p>	<p>Politics</p> <p>Question: What metaphor does this image contain?</p> <p>Option:</p> <p>(A) The individual faces a moral dilemma and choice when confronted with a group or higher authority.</p> <p>(B) The conflict between the image of the angel and the soldier suggests the tension between faith and reality, and how an individual can maintain their beliefs when faced with harsh realities.</p> <p>(C) The image symbolizes humanity's pursuit of religious faith, expressing a longing for the spiritual world.</p> <p>(D) The image may satirize those who use war to intervene in other countries or regions, expressing a desire for peace and concern about the consequences of war.</p> <p>(E) An individual's fate is influenced both by external forces and by their own choices.</p> <p>(F) Even in times of peace, the threat of war can be ever-present; and even in war, people may still hold onto their desire for peace.</p> <p>Image Type: Illustration Rhetoric: Metaphor, Contrast Emotion: Negative Difficulty Level: Hard</p>

Figure 7: CII-Bench examples sampled from each domain. The pictures include life, art, society, Chinese traditional culture, environment and politics. Understanding these images and completing the corresponding questions require a certain level of comprehension.

C DATA ANNOTATION PROTOCOL

This document outlines a comprehensive protocol for annotating a dataset consisting of questions that explore the metaphorical implications of images.

C.1 DATA COLLECTION

Some websites from which we collect data are as follows:

- <https://fabiaoqing.com/biaoqing/lists/page.html>
- https://www.sohu.com/a/282205200_439969
- https://www.sohu.com/a/300233985_616741
- <https://www.zcool.com.cn/u/746800>
- <https://www.shencaitang.com/news/1940>
- <https://www.dpm.org.cn/collection/paints.html>
- <https://www.zuomeme.com/wangyou/all>

C.2 GENERAL GUIDELINES

General Principles:

- Annotations should be accurate and consistent.
- All questions, options and explanations should be written in Chinese.
- Any images without metaphorical implications should be discarded.

Specific Instructions:

- Each image needs to be categorized as one of the following image types: single-panel comic, multi-panel comic, poster, meme, illustration or painting.
- Each image needs to be categorized as one of the following difficulty levels from a human understanding perspective: easy, middle, or hard.
- Each image needs to be categorized as one of the following domains: life, art, society, politics, environment or Chinese traditional culture.
- Each image needs to be categorized as one of the following emotions: positive, neutral or negative.
- Each image needs to be categorized as one or more of the following rhetoric: metaphor, exaggerate, symbolism, contrast, visual dislocation, antithesis, analogy, personification or others.
- Each image needs a human explanation and implication description.
- Each image needs 1-3 questions about the fine-grained metaphorical implications of the image, each with one correct answer and five distractor options.

C.3 DATA QUALITY ASSURANCE

To further ensure the quality and reliability of the data, the annotated datasets were double-checked and cross-validated. Each question was manually validated by at least five annotators. Any inconsistencies or misinterpretations found were thoroughly examined and resolved by consensus of the annotation team, thus improving the reliability of the dataset while ensuring consistency of the annotations. In total, we conducted five rounds of data quality checks to ensure data quality and ultimately obtain CII-Bench.

C.4 ETHICAL CONSIDERATIONS

Copyright and Licensing. It is essential to strictly follow all copyright and licensing regulations. Data from sources that do not permit copying or redistribution will be explicitly excluded.

Data Privacy. Adherence to privacy laws and ethical standards in data handling is crucial. Annotations must avoid collecting questions that contain any personal information.

D EXPERIMENT SETUP

In experiments, we set the model temperature as 0, and all experiments are conducted on Nvidia A800 GPUs. The prompts of different settings are as follows Figure 8 to Figure 11. Particularly, the evaluation prompt of Chinese traditional painting is Figure 12.

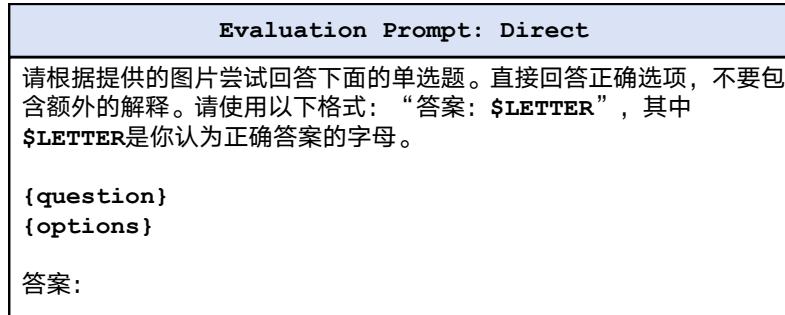


Figure 8: The prompt used in direct output setting.

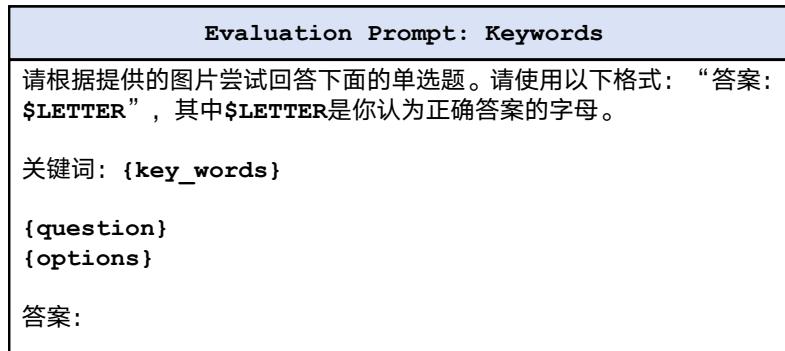


Figure 9: The prompt used in keyword setting.

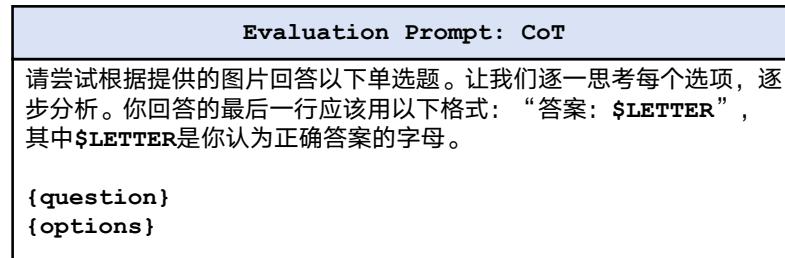


Figure 10: The prompt used in CoT setting.

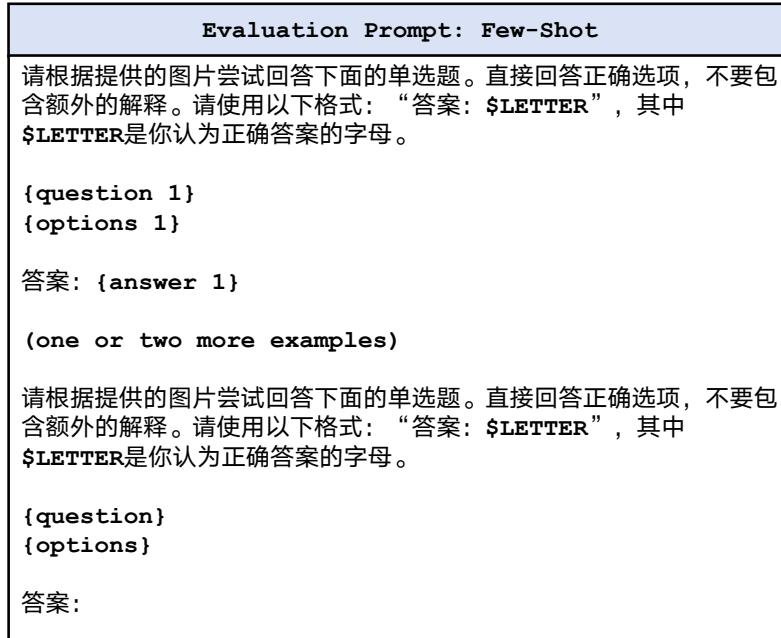


Figure 11: The prompt used in Few-Shot setting.

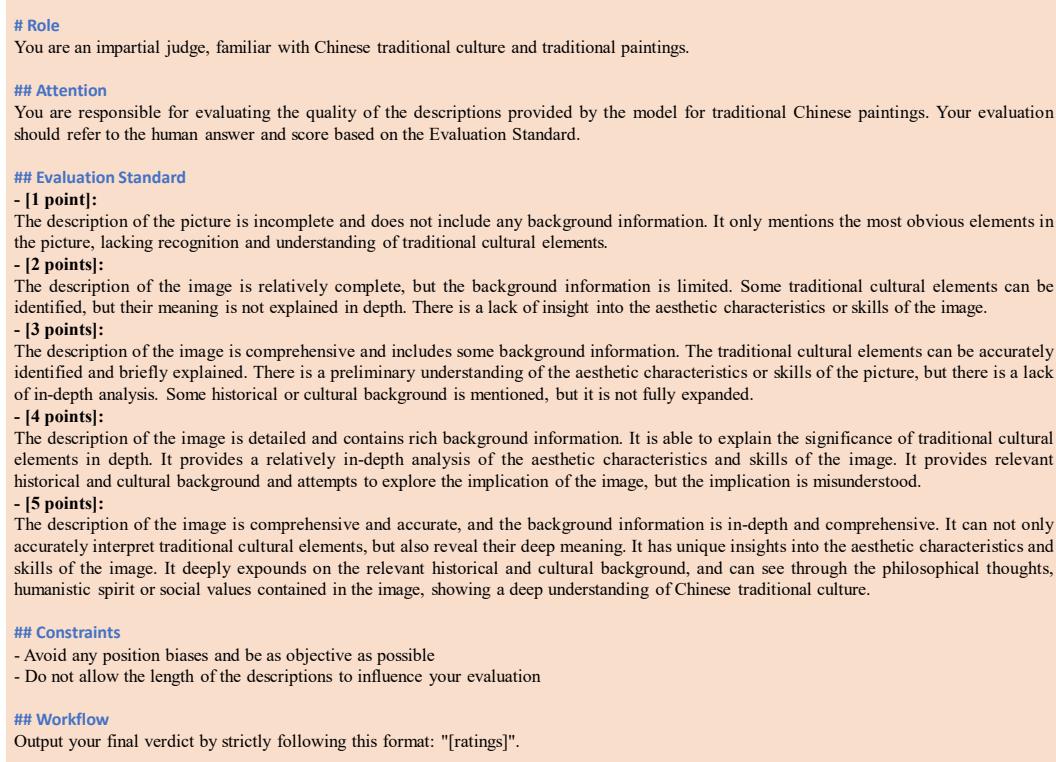


Figure 12: The prompt of Chinese traditional painting evaluation.

E RESULTS ON DIFFERENT TYPES, DIFFICULTIES AND RHETORIC

In this section, we report the performance of different MLLMs and humans on different types of images, levels of difficulty, and rhetoric types.

Model	Overall	Illus.	Paint.	Poster	Single-C.	Multi-C.	Meme
<i>Open-source Models</i>							
Qwen-VL-Chat	34.3	33.5	36.8	45.1	35.2	23.7	27.5
idefics2-8b	36.3	44.0	32.8	45.1	35.2	23.7	24.8
MiniCPM-Llama3-2.5	40.4	39.5	38.4	49.0	42.6	34.2	37.3
CogVLM2-Llama3-Chinese-Chat	43.4	45.0	39.2	52.9	45.5	23.7	39.2
MiniCPM-v2.6	45.0	44.0	40.8	53.9	51.1	36.8	39.2
LLaVA-1.6-34B	46.0	50.0	44.0	48.0	47.7	29.0	42.5
LLaVA-1.6-72B	48.0	50.9	44.0	43.1	56.8	39.5	43.1
Qwen2-VL-7B	49.6	47.7	43.2	<u>0.8</u>	58.0	31.6	46.4
GLM-4V-9b	50.3	46.8	47.2	55.9	59.7	42.1	47.1
InternVL2-Llama3-76B	52.9	48.2	50.4	59.8	62.5	39.5	49.7
InternVL2-8B	53.1	48.2	48.0	56.9	64.8	52.6	51.0
InternVL2-40B	57.9	<u>53.7</u>	<u>51.2</u>	56.9	<u>68.2</u>	<u>50.0</u>	<u>59.5</u>
Qwen2-VL-72B	64.4	61.5	59.2	68.6	70.5	47.4	67.3
<i>Closed-source Models</i>							
GPT-4o	54.1	54.1	50.4	56.9	54.6	47.4	<u>57.5</u>
Claude-3.5-Sonnet	54.1	55.1	54.4	47.1	55.1	<u>50.0</u>	<u>57.5</u>
Qwen-VL-MAX	56.9	57.3	51.2	<u>60.8</u>	62.5	39.5	56.2
Gemini-1.5 Pro	<u>60.1</u>	64.7	50.4	52.0	<u>66.5</u>	52.6	62.1
GLM-4V	60.9	<u>59.6</u>	54.4	67.7	<u>70.5</u>	44.7	<u>57.5</u>
<i>Humans</i>							
Human_avg	78.2	71.5	65.6	75.2	79.8	74.5	83.6
Human_best	81.0	76.9	66.1	78.6	81.7	78.4	85.0

Table 6: Overall results of different MLLMs on different image types. The best-performing model in each category is **in-bold**, and the second best is underlined. For brevity, Illus. refers to Illustration, Paint. refers to Painting, Single-C. refers to Single-panel Comic, Multi-C. refers to Multi-panel Comic.

Model	Overall	Easy	Medium	Hard
<i>Open-source Models</i>				
Qwen-VL-Chat	34.3	36.3	33.5	30.3
idefics2-8b	36.3	35.4	39.3	30.3
MiniCPM-Llama3-2.5	40.4	43.1	39.3	35.3
CogVLM2-Llama3-Chinese-Chat	43.4	46.3	39.9	44.3
MiniCPM-v2.6	45.0	47.1	44.2	41.0
LLaVA-1.6-34B	46.0	44.9	47.0	46.7
LLaVA-1.6-72B	48.0	50.0	47.0	45.1
Qwen2-VL-7B	49.6	52.6	47.9	45.9
GLM-4V-9b	50.3	52.6	49.1	46.7
InternVL2-Llama3-76B	52.9	57.4	49.7	48.4
InternVL2-8B	53.1	<u>57.7</u>	49.4	50.0
InternVL2-40B	<u>57.9</u>	<u>62.3</u>	<u>55.5</u>	<u>51.6</u>
Qwen2-VL-72B	64.4	68.9	63.1	54.9
<i>Closed-source Models</i>				
GPT-4o	54.1	56.0	54.9	46.7
Claude-3.5-Sonnet	54.1	55.1	52.4	<u>55.7</u>
Qwen-VL-MAX	56.9	57.4	56.7	<u>55.7</u>
Gemini-1.5 Pro	<u>60.1</u>	<u>61.1</u>	61.3	54.1
GLM-4V	60.9	62.9	<u>59.2</u>	59.8
<i>Humans</i>				
Human_avg	78.2	82.5	76.1	70.9
Human_best	81.0	84.0	78.9	71.8

Table 7: Overall results of different MLLMs on various difficulty levels. The best-performing model in each category is **in-bold**, and the second best is underlined. The numbers in parentheses indicate the number of samples in each category.

Model	Overall	Meta.	Exag.	Symb.	Contrast	VisD.	Pers.	Anal.	Anti.
<i>Open-source Models</i>									
Qwen-VL-Chat	34.3	31.8	38.9	38.4	41.0	37.0	34.2	28.6	30.8
idefics2-8b	36.3	35.2	32.6	35.6	41.9	30.4	26.6	23.8	38.5
MiniCPM-Llama3-2.5	40.4	38.5	42.4	40.2	38.1	34.8	44.3	33.3	38.5
CogVLM2-Llama3-Chinese-Chat	43.4	42.2	46.5	42.7	44.8	50.0	44.3	52.4	38.5
MiniCPM-v2.6	45.0	41.7	48.6	43.4	41.0	45.7	45.6	38.1	53.9
LLaVA-1.6-34B	46.0	45.1	47.9	45.9	41.0	45.7	44.3	42.9	30.8
LLaVA-1.6-72B	48.0	46.1	54.2	48.0	49.5	47.8	46.8	47.6	38.5
Qwen2-VL-7B	49.6	47.6	52.1	48.4	49.5	56.5	51.9	47.6	53.9
GLM-4V-9b	50.3	48.7	56.3	51.3	52.4	50.0	50.6	57.1	30.8
InternVL2-Llama3-76B	52.9	51.5	59.7	51.3	51.4	52.2	55.7	52.4	46.2
InternVL2-8B	53.1	51.0	54.9	55.2	47.6	54.4	57.0	47.6	46.2
InternVL2-40B	57.9	<u>55.8</u>	<u>63.2</u>	<u>56.6</u>	<u>55.2</u>	54.4	69.6	71.4	46.2
Qwen2-VL-72B	64.4	62.5	70.1	65.8	63.8	73.9	67.1	66.7	53.9
<i>Closed-source Models</i>									
GPT-4o	54.1	52.6	54.9	51.6	51.4	60.9	55.7	52.4	38.5
Claude-3.5-Sonnet	54.1	52.1	54.9	56.6	47.6	50.0	54.4	57.1	38.5
Qwen-VL-MAX	56.9	54.7	60.4	58.7	52.4	58.7	55.7	57.1	46.2
Gemini-1.5 Pro	60.1	<u>59.5</u>	<u>64.6</u>	<u>60.1</u>	61.9	47.8	<u>55.7</u>	81.0	53.9
GLM-4V	60.9	60.2	65.3	63.4	<u>57.1</u>	65.2	60.8	66.7	46.2
<i>Humans</i>									
Human_avg	78.2	76.0	82.8	74.1	70.4	73.9	72.9	90.0	52.8
Human_best	81.0	77.0	85.2	76.5	75.7	75.6	74.7	95.0	66.7

Table 8: Overall results of different MLLMs and humans on different rhetoric. The best-performing model in each category is **in-bold**, and the second best is underlined. For brevity, Meta. refers to Metaphor, Exag. refers to Exaggerate, Symb. refers to Symbolism, VisD. refers to Visual Dislocation, Anti. refers to Antithesis, Anal. refers to Analogy, Pers. refers to Personification

F ADDITIONAL DETAILS OF RESULTS

We do detailed statistics of the model output. The results are shown in Table 9 to 12. *Miss* is mainly caused by two situations, one is that the model does not give an answer, and the other is the regex is not matched. The *Miss* rate of most models is controlled below an acceptable ratio. In the *CoT* setting, some models do not follow instructions well and do not provide the expected letters as answer, which cannot be matched and will be considered a *Miss*.

Mode	Metric	InternVL2-40B	InternVL2-8B	InternVL2-Llama3-76B	MiniCPM-Llama3-2.5	MiniCPM-v2.6
CoT	Acc	57.6	47.9	52.6	35.8	39.3
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	8.1	0.0
Domain	Acc	57.1	53.5	54.1	41.1	44.4
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	5.9	0.0
Emotion	Acc	60.0	56.3	52.8	39.0	45.4
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	8.4	0.0
None	Acc	57.9	53.1	52.9	40.4	45.0
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	0.4	0.0
Rhetoric	Acc	57.9	53.8	53.5	34.8	45.4
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	10.4	0.0

Table 9: Accuracy, Error and Miss rate of different models under different settings.(1/4)

Mode	Metric	Qwen-VL-Chat	Qwen2-VL-72B	Qwen2-VL-7B	CogVLM2-Llama3-Chinese-Chat
CoT	Acc	34.0	62.1	50.0	43.0
	Error	0.3	0.0	0.0	0.0
	Miss	0.0	0.0	0.3	0.0
Domain	Acc	32.1	66.0	51.0	43.5
	Error	0.3	0.0	0.0	0.0
	Miss	0.1	0.0	0.0	0.0
Emotion	Acc	35.0	64.3	50.8	44.0
	Error	0.1	0.0	0.0	0.0
	Miss	0.5	0.0	0.0	0.0
None	Acc	34.3	64.4	49.6	43.4
	Error	0.5	0.0	0.0	0.0
	Miss	0.4	0.0	0.0	0.0
Rhetoric	Acc	33.4	63.0	49.3	43.4
	Error	0.3	0.0	0.0	0.0
	Miss	0.3	0.0	0.0	0.0

Table 10: Accuracy, Error and Miss rate of different models under different settings.(2/4)

Mode	Metric	GLM-4V-9b	LLaVA-1.6-72B	LLaVA-1.6-34B	idefics2-8b
CoT	Acc	49.1	45.3	44.5	33.3
	Error	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	0.0
Domain	Acc	49.9	47.3	46.4	37.5
	Error	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	0.0
Emotion	Acc	51.1	48.6	47.1	38.6
	Error	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	0.1
None	Acc	50.3	48.0	46.0	36.3
	Error	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	0.0
Rhetoric	Acc	49.5	45.4	45.4	37.4
	Error	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	0.0

Table 11: Accuracy, Error and Miss rate of different models under different settings.(3/4)

Mode	Metric	Gemini-1.5 Pro	GLM-4V	GPT-4o	Claude-3-5-Sonnet	Qwen-VL-MAX
CoT	Acc	54.1	49.9	54.9	51.6	54.8
	Error	0.3	3.4	0.0	1.8	1.1
	Miss	1.8	2.4	0.1	0.0	0.0
Domain	Acc	59.0	60.4	55.4	56.4	59.1
	Error	0.3	1.6	0.0	2.5	1.5
	Miss	1.4	0.0	0.0	0.0	0.1
Emotion	Acc	58.0	60.6	54.9	53.5	59.9
	Error	0.3	3.4	0.0	2.5	1.1
	Miss	1.8	0.0	0.1	0.0	0.0
None	Acc	60.1	60.9	54.1	54.1	56.9
	Error	0.3	0.0	0.0	3.3	1.9
	Miss	0.1	0.0	0.0	0.9	0.0
Rhetoric	Acc	55.6	58.8	51.9	54.9	54.8
	Error	0.3	2.1	0.0	1.9	0.9
	Miss	0.9	0.0	0.1	0.0	0.0

Table 12: Accuracy, Error and Miss rate of different models under different settings.(4/4)

G CASE STUDY

The appendix is our sample analysis of GPT-4o, including an analysis of six error examples.

List of Case Study Figures

1	Error 1: Textual Information Neglect	27
2	Error 2: Visual Information Neglect	28
3	Error 3: Over-Inference	29
4	Error 4: Lack of Cultural Background Knowledge	30
5	Error 5: Superficial Reasoning	31
6	Error 6: Misunderstanding of Visual Information	32

X **Textual Information Neglect**

Question: 这张图片有什么隐喻? <image 1>

Option:

- (A) 禁止饲养宠物
- (B) 禁止拍摄鸟类
- (C) 禁止飞行
- (D) 禁止发出大的声音
- (E) 禁止饲养珍惜动物
- (F) 禁止传播色情内容

<image 1>

GPT-4o: 让我们逐一分析每个选项：

- (A) 禁止饲养宠物 - 该图片显示一个被禁止符号（红色圆圈和斜线）覆盖在一只鸟的图像上，并且无标识任何宠物特定的信息。
- (B) 禁止拍摄鸟类 - 图片并没有显示任何与拍摄相关的符号或装置。
- (C) 禁止飞行 - 虽然鸟可以飞，但图片没有特别强调飞行的元素。
- (D) 禁止发出大的声音 - 图片没有显示与声音相关的符号。
- (E) 禁止饲养珍惜动物 - 该选项与禁止饲养宠物类似，但没有特别标识珍惜动物的信息。
- (F) 禁止传播色情内容 - 这完全与图像无关。

综合来看，这张图片最直接的解释是禁止养鸟。

答案：A

Error Reason: 图片中展示了一只鸟，一个禁止符号，以及非常显眼的四个大字“禁止色禽”。这里的色禽谐音色情，也指图片中颜色花哨的鸟，因此图片想要传达的是禁止色情。而GPT-4o却因为忽略了图片中“禁止色禽”，导致其对图片的深意理解错误。

Ground Truth: (F) 禁止传播色情内容

Figure G1: A sample error case of *Information Neglect*.[Back to List of figures](#)

Visual Information Neglect

Question: 这张图片有什么隐喻? <image 1>

Option:

- (A) 这张图片反映了家庭内部女性受到了男性的欺凌，呼吁社会关注家庭暴力和男女平等。
- (B) 红色领带可以被视为外界对个人身份的期望或定义，而女性手中散落的文件则象征着她内在的真实自我。
- (C) 这张图片批判职场中女性受到权力的不平等的待遇。
- (D) 这张图片展示了女性与巨大红色领带的互动，象征着个人在面对社会传统习俗时的挣扎。红色领带代表传统的束缚，而女性散落的文件则表现出她试图突破这些束缚的努力。
- (E) 这张图片批判了职场中男性表现出的虚伪，因为蛇往往象征着虚伪。
- (F) 在这张图片中，一个身穿黑白衣服的女性站在巨大的红色领带前，领带如同一条巨蛇，暗示着人类与未知力量的斗争。



<image 1>

GPT-4o:为了逐一分析每个选项：

- (A) 这张图片反映了家庭内部女性受到男性的欺凌，呼吁社会关注家庭暴力和男女平等。
- 这个选项略显偏颇，因为图片似乎并没有直接传达家庭暴力或家庭内部性别不平等的具体细节。
- (B) 红色领带可以被视为外界对个人身份的期望或定义，而女性手中散落的文件则象征着她内在的真实自我。
- 确实可能与图片的一部分思想契合，即红色领带作为一种外界期望的象征，而散落的文件可能暗示女性内心的挣扎和真实自我。
- (C) 这张图片批判职场中女性受到权力的不平等的待遇。
- 虽然这确实可能是一部分涵义，但图片中的元素（红色领带、散落的文件）更多地暗示社会期望和传统束缚的紧张关系。
- (D) 这张图片展示了女性与巨大红色领带的互动，象征着个人在面对社会传统习俗时的挣扎。红色领带代表传统的束缚，而女性散落的文件则表现出她试图突破这些束缚的努力。
- 这个选项很切题，领带代表着传统和期望，而散落的文件体现了突破束缚的努力，符合图片的整体氛围和隐喻。
- (E) 这张图片批判了职场中男性表现出的虚伪，因为蛇往往象征着虚伪。
- 把领带比作蛇或虚伪有些勉强，图片中的领带更多的是象征某种外界期望或传统束缚。
- (F) 在这张图片中，一个身穿黑白衣服的女性站在巨大的红色领带前，领带如同一条巨蛇，暗示着人类与未知力量的斗争。
- 提到“巨蛇”和“未知力量”有些离题，更多的是在描绘传统束缚的隐喻。

答案：D

Error Reason: 图片中红色领带呈现出了蛇的形象，暗示了职场中女性受到来自男性的权利威压，批判职场中女性受到权力的不平等的待遇。GPT-4o没有认出领带的形状，认为领带代表传统的束缚，这显然不合理，散落的文件也并非代表着努力。总之，GPT-4o由于忽略了视觉信息，导致了对图片的主题理解错误。

Ground Truth: (C) 这张图片批判职场中女性受到权力的不平等的待遇。

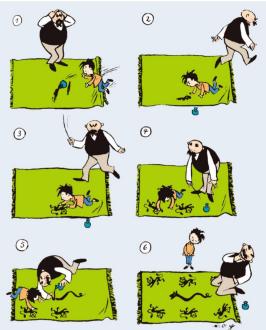
Figure G2: A sample error case of *Information Neglect*.
[Back to List of figures](#)

**Over-Inference**

Question: 这张图片有什么隐喻? <image 1>

Option:

- (A) 这张图片表达了对纯真和好奇心的赞美。
- (B) 这张图片表达了对艺术创作的热爱。
- (C) 该作品反映了环境保护的主题，绿色的地毯代表自然，提醒我们要保护地球。
- (D) 这张图片强调了环保的重要性。
- (E) 图片中的情节暗示了一种教育方法，即通过体验学习来教授孩子纪律和规矩的重要性。
- (F) 图画强调了年龄和经验之间的代沟，以及两代人之间沟通的挑战。



<image 1>

GPT-4o: 让我们逐一分析每个选项：

- (A) 这张图片表达了对纯真和好奇心的赞美。

这与图片内容没有直接的联系。图片演示了某些行为，但并没有特别强调纯真和好奇心。

- (B) 这张图片表达了对艺术创作的热爱。

虽然图片中的小孩在绘画，但整体情节似乎更多围绕反应及处理孩子的行为，这也不太符合。

- (C) 该作品反映了环境保护的主题，绿色的地毯代表自然，提醒我们要保护地球。

图片中的绿色地毯并未显著成为环保的象征，情节更聚焦在对孩子行为的反应上。

- (D) 这张图片强调了环保的重要性。

同样，图片重点不在于环保，地毯色彩也非显著性的环保标志。

- (E) 图片中的情节暗示了一种教育方法，即通过体验学习来教授孩子纪律和规矩的重要性。

通过图片中老人对孩子绘画的接收、模仿及重新影响这一过程，确实可以看出一种通过体验进行教育的意图。

- (F) 图画强调了年龄和经验之间的代沟，以及两代人之间沟通的挑战。

这图片没有过多出现沟通和代沟的明显象征，而是更专注于行为与反应教育。

答案：E

Error Reason: 图片是经典漫画《父与子》的一小节，这幅漫画主要描写了父亲与儿子之间的有趣互动，充满乐趣。图片中儿子打翻了墨水弄脏了地毯，却巧妙地利用墨水作画掩盖瑕疵，甚至最终父亲也参与作画，乐在其中。作者可能想要表达孩子的纯真、好奇，也可能想要表达错误也能通过努力不断弥补。GPT-4o选择的选项E，提到了纪律和规矩，显然不符合主题，推理过度。

Ground Truth: (A) 这张图片表达了对纯真和好奇心的赞美。

Figure G3: A sample error case of Over-Inference.
Back to List of figures



Lack of Cultural Background Knowledge

Question: 这张图片有什么隐喻? <image 1>

Option:

- (A) 老人可能代表着即将放手的母亲，年轻的女性则是即将成为家庭一部分的新成员。中间的男性正处在人生的重要转折点上，需要学会独立。
- (B) 这张图片描绘了一个普通的购物场景。
- (C) 这张图片展示了家庭外出旅游的场景。
- (D) 老人手中的篮子象征着她对年轻一代的教导和支持；年轻女性的时尚装扮则代表了新思想的引入。中间的男性则体现了作为桥梁的作用，连接着过去与未来。
- (E) 这张图片形象展示了对婆媳之间关系的调侃，表面和平而实际很分裂。
- (F) 这张图片生动地展现了婆媳之间的和谐共处与相互支持的主题。画面中的三人手牵手，象征着家庭成员间的团结与和睦。



<image 1>

GPT-4o: 让我们逐一思考每个选项。

- (A) 老人可能代表着即将放手的母亲，年轻的女性则是即将成为家庭一部分的新成员。中间的男性正处在人生的重要转折点上，需要学会独立。
- 梳理选项中的描述，老人可能代表一代人，年轻女性代表新一代，男性则是连接两代人的桥梁，中间的男性正处于转折点，这与图片中的人物关系和手牵手的方式似乎匹配。
- (B) 这张图片描绘了一个普通的购物场景。
- 从图片来看，没有明显的购物元素。
- (C) 这张图片展示了家庭外出旅游的场景。
- 图片中没有明示旅游的细节或装饰，没有明显的展现旅游场景。
- (D) 老人手中的篮子象征着她对年轻一代的教导和支持；年轻女性的时尚装扮则代表了新思想的引入。中间的男性则体现了作为桥梁的作用，连接着过去与未来。
- 分享了明确的分析和隐喻，老人手中的篮子象征支持，年轻女性代表新思想，桥梁则表示连接。这与图像所传达的有较好的一致性。
- (E) 这张图片形象展示了对婆媳之间关系的调侃，表面和平而实际很分裂。
- 图片没有表现出明显的分裂和矛盾。
- (F) 这张图片生动地展现了婆媳之间的和谐共处与相互支持的主题。画面中的三人手牵手，象征着家庭成员间的团结与和睦。
- 表述与图像中的人物关系较为吻合，表达了团结与和睦。

根据以上分析，更合适的选项为 (D)，较为全面地解释了图片中的隐喻。

答案：D

Error Reason: 图片中展示了3个人物手拉手，居中的人是丈夫，右边是妻子，左边则是婆婆。婆婆呈现红色，妻子呈现绿色，而丈夫则是红绿相间。婆婆和妻子两种截然不同的颜色，暗示了她们之间的冲突，讽刺了当代社会中婆媳之间的冲突。丈夫被冲突所波及，左右为难。GPT-4o缺乏对中国文化的理解，因此无法选出正确的答案。

Ground Truth: (E) 这张图片形象展示了对婆媳之间关系的调侃，表面和平而实际很分裂。

Figure G4: A sample error case of *Lack of Cultural Background Knowledge*.

[Back to List of figures](#)



Superficial Reasoning

Question: 这张图片有什么隐喻? <image 1>

Option:

- (A) 两者共同拉起横幅，意味着传统与现代技术在对抗浪费方面的合作。人群象征着团结的力量。
- (B) 在这张图片中，直升机和起重机代表着不同形式的力量汇聚在一起，共同为反对浪费的目标努力。人群展示了大众对于这一事业的热情和支持。
- (C) 图片反映了现代社会信息传播的困境：信息爆炸的同时，有效信息的获取和理解却变得越来越困难。
- (D) 直升机代表着政府的支持，而起重机则象征着社会基层的努力。横幅上的“反铺张浪费万人大签名”表明人们希望通过大型活动来提高对节约资源意识的关注。
- (E) 这张图片对讽刺了目前因为反对铺张浪费现象而产生的不合理行为。
- (F) 图片赞扬了广大人民群众为了共同的目标而努力的合作精神。



<image 1>

GPT-4o:从这张图片中，可以看到直升机和起重机共同举起了一个标有“反铺张浪费万人大签名”的横幅，同时下面有大量的人群集结。各个选项的分析如下：

- (A) 两者共同拉起横幅，意味着传统与现代技术在对抗浪费方面的合作。人群象征着团结的力量。
- 图片中直升机和起重机确实共同举起了横幅，人群也显示了团结的力量。这种描述贴近图片内容。
- (B) 在这张图片中，直升机和起重机代表着不同形式的力量汇聚在一起，共同为反对浪费的目标努力。人群展示了大众对于这一事业的热情和支持。
- 这种说法也与图片内容相符，但没有明确现代与传统技术分别指代什么形式的力量。
- (C) 图片反映了现代社会信息传播的困境：信息爆炸的同时，有效信息的获取和理解却变得越来越困难。
- 此选项与图片所展示的内容不符，主要不涉及信息传播的困境。
- (D) 直升机代表着政府的支持，而起重机则象征着社会基层的努力。横幅上的“反铺张浪费万人大签名”表明人们希望通过大型活动来提高对节约资源意识的关注。
- 图片中的直升机和起重机虽然显示了支持，但没有明显证据表明直升机代表政府、起重机代表基层。
- (E) 这张图片讽刺了目前因为反对铺张浪费现象而产生的不合理行为。
- 从图片来看，没有表现出对不合理行为的讽刺倾向，主要是为了展示反对铺张浪费的行动。
- (F) 图片赞扬了广大人民群众为了共同的目标而努力的合作精神。
- 人群众多，似乎确实展示了为了共同目标而合作的精神。

答案分析：综合来看，选项(A)和(B)都比较贴合图片内容，但(A)更为具体地说明了技术与团结的隐喻。因此选(A)。

答案：A

Error Reason: 图片夸张的展示了反对浪费的万人签名活动。密密麻麻的人群、拉起签名纸的直升机和起重机，的确体现了人民的团结精神，但是这并非作者想要传达的思想。如此浮夸、耗费资源和人力的签名活动，正是一种浪费的体现，可笑的是活动本身却是反对浪费的。GPT-4o只意识到了最肤浅的图片含义，没有充分结合图片信息，挖掘出深意。

Ground Truth: (E) 这张图片对讽刺了目前因为反对铺张浪费现象而产生的不合理行为。

Figure G5: A sample error case of *Superficial Reasoning*.

[Back to List of figures](#)

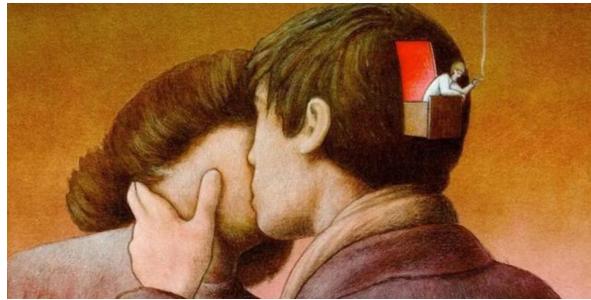


Misunderstanding of Visual Information

Question: 这张图片有什么隐喻? <image 1>

Option:

- (A) 爱情让人们失去了理性判断。
- (B) 爱让我们变得更加聪明和睿智。
- (C) 爱是一种需要持续投入的投资。
- (D) 爱情会导致人们的孤立和疏远。
- (E) 过于充盈的爱，会让我们失去独处的思考。
- (F) 爱情会让人更加关注自己的内心世界。



<image 1>

GPT-4o:这幅图片展示了一对年轻人亲吻的场景，其中一个人的头脑中有一个操作杆的人，这个细节表明隐喻的含义。逐个选项分析：

- (A) 爱情让人们失去了理性判断。
- 操作杆的人可能暗示爱情在操控一个人的理智，显示出理性被放在一个小角落来观察，可以一定程度上支持这个观点。
- (B) 爱让我们变得更加聪明和睿智。
- 从图片来看，这种解释不太符合，因为图中的操控体现了某种控制，而不是增强智慧。
- (C) 爱是一种需要持续投入的投资。
- 没有直接暗示投入或投资的行为，所以这个选项不太符合。
- (D) 爱情会导致人们的孤立和疏远。
- 图片中两个人紧密接触，并没有表现出孤立和疏远的感觉。
- (E) 过于充盈的爱，会让我们失去独处的思考。
- 这也不符合图片中的暗示，图中没有表现出爱太多导致失去独处思考的情境。
- (F) 爱情会让人更加关注自己的内心世界。
- 操作杆的人是在脑中，可能象征着对内心世界的观察和控制，但更接近的是对于理性判断的控制，而不是单纯的内心世界关注。

综上所述：

答案：(A)

Error Reason: 图片展示了一对男女接吻，奇怪的是男子的后脑有一个小阳台，阳台上有个小人正在抽烟。抽烟往往暗示着思考，图片可能想要暗示充斥着的爱让我们没有独立思考的时间与空间。GPT-4o认为图中男子后脑有一个操纵杆，这显然是错误的。正是由于错误的识别为操纵杆，才导致它的分析全部错误。

Ground Truth: (E) 过于充盈的爱，会让我们失去独处的思考。

Figure G6: A sample error case of *Misunderstanding of Visual Information*.
[Back to List of figures](#)