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

Ziqiang Liu[1,2]* **Feiteng Fang**[1,3]* **Xi Feng**[1,3]* **Xinrun Du**[4,14]* **Chenhao Zhang**[1,6]*
Zekun Wang[12,14] **Yuelin Bai**[1,2] **Qixuan Zhao**[1,3] **Liyang Fan**[1] **Chengguang Gan**[7]
Hongquan Lin[1,3] **Jiaming Li**[1,2] **Yuansheng Ni**[9] **Haihong Wu**[1,3] **Yaswanth Narsupalli**[5]
Zhigang Zheng[1] **Chengming Li**[10] **Xiping Hu**[10] **Ruifeng Xu**[11] **Xiaojun Chen**[8] **Min Yang**[1]
Jiaheng Liu[12] **Ruibo Liu**[13] **Wenhao Huang**[14] **Ge Zhang**[4,14,15]† **Shiwen Ni**[1]†
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]University of Science and Technology of China [4]M-A-P [5]IIT Kharagpur
[6]Huazhong University of Science and Technology [7]Yokohama National University
[8]Shenzhen University [9]Zhejiang University [10]Shenzhen MSU-BIT University
[11]Harbin Institute of Technology (Shenzhen) [12]Beihang University
[13]Dartmouth College [14]01.ai [15]University of Waterloo

Abstract

The rapid advancements in the development of multimodal large language models (MLLMs) have consistently led to new breakthroughs on various benchmarks. In response, numerous challenging and comprehensive benchmarks have been proposed to more accurately assess the capabilities of MLLMs. However, there is a dearth of exploration of the higher-order perceptual capabilities of MLLMs. To fill this gap, we propose the **I**mage **I**mplication understanding **Bench**mark, **II-Bench**, which aims to evaluate the model's higher-order perception of images. Through extensive experiments on II-Bench across multiple MLLMs, we have made significant findings. Initially, a substantial gap is observed between the performance of MLLMs and humans on II-Bench. The pinnacle accuracy of MLLMs attains 74.8%, whereas human accuracy averages 90%, peaking at an impressive 98%. Subsequently, MLLMs perform worse on abstract and complex images, suggesting limitations in their ability to understand high-level semantics and capture image details. Finally, it is observed that most models exhibit enhanced accuracy when image sentiment polarity hints are incorporated into the prompts. This observation underscores a notable deficiency in their inherent understanding of image sentiment. We believe that II-Bench will inspire the community to develop the next generation of MLLMs, advancing the journey towards expert artificial general intelligence (AGI). II-Bench is publicly available at `https://huggingface.co/datasets/m-a-p/II-Bench`.

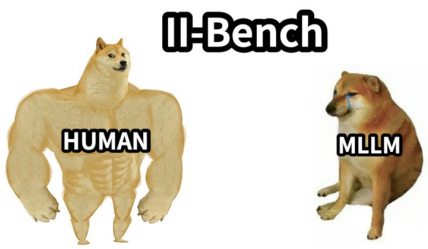

Figure 1: Implication: a significant gap exists between humans and MLLMs on II-Bench.

*Equal Contribution.
†Corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

1 Introduction

In recent years, the development of Multimodal Large Language Models (MLLMs) [2, 12, 33, 52, 66] has significantly advanced our ability to understand and generate content across various modalities, including text, images, and audio. Leveraging sophisticated architectures and vast amounts of data, MLLMs have demonstrated remarkable performance in image captioning [18, 24, 45], visual question answering[44, 51], video understanding and generation[46, 70], etc.

Nevertheless, comprehensively evaluating the performance of these models remains a challenge. While benchmarks exist for multimodality, such as ScienceQA [43], MMMU [67], there is a dearth of exploration of the higher-order perceptual capabilities of MLLMs, which refer to nuanced emotional understanding and profound meaning extraction.

Philosopher Suzanne Langer once noted, "Art is the creation of forms symbolic of human feeling." This profoundly summarizes how images often embody human emotions and serve as a conduit for personal views and cultural narratives. Therefore, understanding the meaning of images requires not only meticulous observation but also an exploration of the human emotions and cultural contexts they reflect. In real life, many artworks, comics, and posters are imbued with rich meanings, and artists convey their insights to the audience through these works. These abstract and complex images pose a significant challenge for MLLMs, as the models must possess advanced higher-order perceptual capabilities to accurately understand the human emotions conveyed in the pictures and infer the deeper meanings the creators intend to express. The evaluation of high-order perception is crucial for artificial general intelligence (AGI) because it encompasses the ability to understand complex, non-literal aspects of visual information, much like humans do, which includes recognizing and processing implications, emotional cues, synesthesia or other cognitive features that go beyond mere object recognition or factual knowledge extraction [6, 55]. This implies that a machine's understanding of the world should also incorporate these implications to achieve a more human-like comprehension, which fosters a deeper, more nuanced understanding, serving as an important means of exploration from perceptual intelligence to cognitive intelligence.

However, an effective benchmark for higher-order perceptual capability measurement is notably absent in the current landscape. To fill this gap, we introduce II-Bench, a comprehensive benchmark designed to assess MLLMs' higher-order perceptual, reasoning and comprehension abilities. This holistic evaluation enables us to gain a deeper insight into the models' true capabilities, thereby fostering advancements in multimodal AI research.

As illustrated in Figure 2, II-Bench comprises 1,222 images, spanning six domains: life, art, society, psychology, environment and others. Furthermore, II-Bench encompasses diverse categories of images, including illustrations, memes, posters, comics, logos and paintings. By utilizing images from multiple domains and categories, the model's comprehension and reasoning abilities can be evaluated more objectively and comprehensively.

Figure 2: Composition of II-Bench.

We conduct extensive experiments to evaluate II-Bench on 20 MLLMs. Our main contributions and findings are as follows:

- We introduce II-Bench, the first Image Implication Understanding Benchmark, which is very challenging for current MLLMs.
- A significant difference exists in performance between humans and MLLMs: the highest accuracy achieved by the model is 74.8%, whereas the average accuracy for humans is 90%, with the highest reaching 98%.
- Closed-source models often outperform open-source ones, while the performance gap between the leading closed-source model and the leading open-source model is minimal, only about 1%.
- Models perform worse in domains containing abstract and complex information, such as Art and Psychology, compared to Environment, Life, Society and other domains.
- Incorporating additional emotional polarity information of images into prompts generally enhances model scores, indicating that models lack sufficient emotional understanding of images, leading to misinterpretation of implicit meanings.

Our aim with II-Bench is to evaluate MLLMs' higher-order perception of images. We believe that II-Bench will inspire the community to create the next generation of MLLMs, propelling us further on the path toward sophisticated arti cial general intelligence (AGI).

2 Related Work

2.1 Multimodal Large Language Models

Given that advanced large language models (LLMs) exhibit sophisticated reasoning abilities, strong generality, and extensive world knowledge [48, 49], current multimodal LLMs (MLLMs) typically involve integrating additional modules to align non-textual modality features with the language space. For example, BLIP-2 [35] encodes images using ViT [16] and employs a Q-Former to map visual features into the language space. Similarly, LLaVA [38, 7, 9] utilizes an MLP as the connector between the visual encoder and the LLM backbone. These architectural designs not only incorporate visual representations into the LLMs but also preserve the advanced capabilities inherent to LLMs. Recent studies have demonstrated that current MLLMs are capable of understanding human minds, reasoning with scienti c gures, etc. [4, 49], due to the success of unlocking the abilities of LLM backbones in multimodal settings. Nonetheless, despite the strong implication understanding abilities of LLMs [59], there is limited research on the implication understanding of images by current MLLMs, and our work addresses this gap for the rst time.

2.2 MLLM Benchmarks

The evolution of MLLMs has underscored the importance of comprehensive evaluations within the research community. Initial benchmarks primarily targeted singular tasks, such as the visual question answering (VQA) tasks [21, 19, 29, 54, 27] and the image captioning tasks [36, 1, 50]. While notable achievements have been recorded on these benchmarks, they fall short of thoroughly evaluating MLLMs across the broader spectrum of multimodal perception and reasoning. To bridge this gap, recent studies have aimed at evaluating models from various perspectives [40, 32, 31, 62, 17, 43, 5, 71, 20]. For example, MMBench [40] and SEED [32, 31] explore models' capabilities through common-sense questions, featuring multiple-choice questions across various dimensions of ability. To assess specialized expertise, MMMU [67] and CMMMU [69] leverage content from exams and textbooks to enhance domain-speci c knowledge evaluation.

However, MMStar [8] pointed out that the model can answer some benchmarks' questions without images, and there is a risk of data leakage during training. We nd that these benchmarks mostly test knowledge or just simple image understanding and don't assess logic and reasoning skills. Image implication understanding represents a more challenging task compared to image understanding, necessitating multi-hop reasoning ability and theory of mind (ToM) [14, 23, 63, 72, 56, 57]—the sophisticated capability intrinsic to human cognition. II-Bench is a benchmark designed to evaluate MLLMs' prowess in both image understanding and reasoning through image implication.

3 The II-Bench

3.1 Overview of II-Bench

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

II-Bench contains a total of 1,222 various images. The speci c image types and domain statistics can be seen in Figure 5 of the Appendix A. These images are manually collected and annotated by 50 undergraduate students from various disciplines and institutions, with sources from multiple renowned 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,

Figure 3: II-Bench examples sampled from each domain. The pictures include life, art, society, psychology, environment and other domains. Understanding these images and completing the corresponding questions require a certain level of comprehension.

symbolism, and detailed understanding of the images. The benchmark includes a total of 1,434 multiple-choice questions, with 1,399 questions used to construct the test set and 35 questions used to construct the development and validation set for few-shot tasks. Figure 3 shows representative examples of II-Bench.

3.2 Data Curation Process

Data Collection. We collect 20,150 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. We have also taken special measures to prevent data leakage. For detailed information on the specific websites from which we collect images and our data leakage prevention measures, please refer to Appendix B.

Data Filtration. After collecting the raw images, we carefully design a three-stage data filtration procedure. In Stage 1, dedicated to image deduplication, we utilize image similarity algorithms to perform pixel-based comparisons which allows the identification and elimination of copies and close variants, rendering the dataset unique. Stage 2, focused on text-to-image ratio control, we use Optical Character Recognition (OCR) to locate text portions in the images. We then calculate the area occupied by text relative to the total image area. Images are removed if the text-to-image ratio breaches the threshold, ensuring that the dataset remains visually dominant. In Stage 3, an exhaustive visual inspection is conducted by humans. Our specific screening protocol is mandated to identify and discard images lacking implications. This strategic exclusion ensures that irrelevant and poor-quality images are weeded out, enhancing the meaningfulness and quality of data retained for further processing. After these filtration stages, we have eliminated over 90% of the original images, leaving us with fewer than 2,000 images.

Data Annotation. We forward the annotation sources to the crowdsourcing annotators and perform three steps of data annotation using our carefully devised annotation protocol. The annotators mark the images with their dif culty, image type, domain, and corresponding rhetoric rst. An explanation of contained visual implications is then drafted for each image. Finally, the annotators devise 1-3 ne-grained questions per image, each with only one correct answer and ve distractor options related to the implication nuances. The detailed annotation protocol is in Appendix B.

Data Quality Assurance. To ensure the quality, dif culty of questions and distractor options and consistency among annotators, we have implemented a rigorous multi-step process. Speci cally, each question and option undergoes multiple rounds of meticulous manual annotation to ensure the distractors are suf ciently challenging and not easily distinguishable from the correct option and ensure consistency across different annotators. If any question is found to be insuf ciently challenging or answers from different annotators are not consistent, it is revised accordingly. This iterative process continues until at least 4 out of 5 independent reviewers agree that the question is accurate, reasonable, dif cult and reach a consensus.

3.3 Dataset Statistics

II-Bench comprises 1,222 images, each accompanied by 1 to 3 multiple-choice questions, totaling 1,434 questions. We randomly select 35 of these questions to construct a few-shot development set and validation set. The average question length is approximately 17 words, and the average option length is 14 words. Each image also includes a description manually annotated by annotators, explaining the human interpretation of the image's implication.

II-Bench encompasses images from six distinct domains: Life, Art, Society, Psychology, Environment and Others. It features a diverse array of image types, including Illustrations, Memes, Posters, Multi-panel Comics, Single-panel Comics, Logos and Paintings. The images are classi ed based on human understanding into three levels of dif culty: Easy, Middle and Hard. Additionally, they are categorized by the emotional tone they convey: Positive, Neutral or Negative. Furthermore, each image is manually annotated with rhetorical devices such as Metaphor, Exaggeration, Symbolism, Contrast, Visual Dislocation, Antithesis, Analogy, Personi cation and Others. The detailed statistical information can be found in Table 1.

Statistics	
Total Questions	1,434
Total Images	1,222
Dev : Validation : Test	15 : 20 : 1,187
Easy : Medium : Hard	708 : 385 : 129
Average Question Length	16.91
Average Option Length	14.05
Average Explanation Length	170.47
Metaphor	955
Exaggerate	191
Symbolism	236
Visual Dislocation	71
Antithesis	27
Analogy	38
Personi cation	108
Contrast	226
Other	47

Statistics	
Life	516 (42.23%)
Art	70 (5.73%)
Society	408 (33.39%)
Psychology	127 (10.39%)
Environment	44 (3.60%)
Other	57 (4.66%)
Positive	169 (13.83%)
Neutral	702 (57.45%)
Negative	351 (28.72%)
Illustration	374 (28.70%)
Meme	269 (20.64%)
Poster	111 (8.52%)
Multi-panel Comic	311 (23.87%)
Single-panel Comic	90 (6.91%)
Logo	59 (4.53%)
Painting	89 (6.83%)

Table 1: Statistics of II-Bench.

4 Experiment

We conduct experiments on II-Bench using both open-source and closed-source MLLMs. For each model, we employ eight different settings: 1-shot, 2-shot, 3-shot, zero-shot (None), CoT,

Domain, Emotion and Rhetoric. "Emotion" denotes prompts where the model is informed about the emotional polarity of the images(e.g., positive, negative), "Domain" involves adding information about the image's domain (e.g., life, environment) to the prompt, and "Rhetoric" signi es prompt with information about the rhetorical devices used in the image (e.g., metaphor, personi cation), while "None" indicates the use of standard prompts without any additional information. Uniform prompts are applied across all MLLMs, with detailed speci cations available in the Appendix C. All experiments are conducted on NVIDIA A800 GPUs.

4.1 Baselines

MLLMs. Table 2 provides an overview of the studied MLLMs, highlighting differences in their architectures and parameters. Notably, InternLM-XComposer2 [15] attempts to modify the projection module in LLaVA architecture to better align multiple modalities. Meanwhile, CogVLM [60] integrates a visual expert into the large language model, enabling a deep fusion of vision and language features without compromising performance on NLP tasks.

Model	Size	ViT	Projection Module	LLM
CogVLM2-Llama3-Chat [60]	19.5B	EVA2-CLIP-E	MLP	Llama-3-8B + Visual Expert
MiniCPM-Llama3-2.5 [25]	8.5B	SigLip-400M	Perceiver Resampler	Llama3-8B
InternVL-Chat-1.5 [10]	25.5B	InternViT-6B	MLP	InternLM2-20B
InternLM-XComposer2-VL [15]	7B	OpenAI ViT-Large	PLoRA	InternLM-2
DeepSeek-VL-Chat-7B [42]	7.3B	SAM-B + SigLIP-L	MLP	DeepSeek-LLM-7B
InstructBLIP-T5 [13]	4.0B/12.3B	ViT-g/14	MLP	FLAN T5 XL/XXL
BLIP-2 FLAN-T5 [35]	4.1B/12.1B	ViT-g/14	MLP	FLAN T5 XL/XXL
mPLUGw-OWL2 [64]	8.2B	ViT-L/14	Visual Abstractor	Llama-2-7B
Qwen-VL-Chat [3]	9.6B	ViT-bigG	VL Adapter	Qwen-7B
Yi-VL-34B-Chat [65]	7.1B/35.4B	CLIP ViT-H/14	MLP	Yi-34B-Chat
LLaVA-1.6-34B [37]	34.8B	ViT-L/14	MLP	Nous-Hermes-2-Yi-34B
Mantis-8B-siglip-llama3 [28]	8.5B	SigLIP	MLP	Llama-3-8B
Ide cs2-8B [30]	8.4B	SigLIP	MLP	Mistral-7B

Table 2: The architecture and size of different models.

Evaluation. Accuracy is used as our main evaluation metric. Given that II-Bench comprises entirely multiple-choice questions, the evaluation merely involves extracting the selected options from the model's responses, thereby simplifying the rule design complexity. Notably, when the model employs chain-of-thought (CoT) prompting, the responses generate intermediate steps. This necessitates that the designed rules possess suf cient robustness or that the model outputs answers in a xed format. If the options cannot be extracted from the model's response, it is deemed that the model has answered the current question incorrectly. For the robust answer extraction method and the detailed statistics of the model output, please see Appendix E. For reference, we also assessed human performance on II-Bench.

4.2 Main Results

In this section, we present a comprehensive comparison of different MLLMs and humans on II-Bench. The detailed results of different domains and emotions are in Table 3. The detailed results of different image types, levels of dif culty, and rhetoric are in Appendix D. The main experimental results and ndings are summarized below:

4.2.1 Gap between Humans and MLLMs

The results indicate a signi cant disparity between humans and MLLMs on II-Bench. Human participants achieve an average accuracy of 90.3%, with the highest accuracy reaching 98.2%. In comparison, the best closed-source model, Qwen-VL-MAX, achieves an accuracy of 74.8%, while the best open-source model, LLaVA-1.6-34B, scores 73.8%. These results highlight the substantial gap between human capabilities and current state-of-the-art models in understanding image implications. The highest accuracy of the models is substantially lower than the average human score, underscoring the challenges that MLLMs face in this domain.

	Overall (1,399)	Life (585)	Art (85)	Society (461)	Psy. (152)	Env. (51)	Others (65)	Positive (196)	Neutral (789)	Negative (414)
Open-source Models										
InstructBLIP-T5-XL	47.3	45.6	48.2	48.8	44.7	52.9	50.8	46.9	48.3	45.4
BLIP-2 FLAN-T5-XL	52.8	53.0	58.8	52.5	42.8	64.7	58.5	56.1	52.9	51.0
mPLUGw-OWL2	53.2	54.0	56.5	50.5	52.0	60.8	56.9	55.6	52.6	53.1
Qwen-VL-Chat	53.4	53.2	49.4	52.1	50.0	60.8	72.3	56.1	52.6	53.6
InstructBLIP-T5-XXL	56.7	56.2	58.8	58.6	45.4	64.7	64.6	63.3	56.1	54.6
Mantis-8B-siglip-Llama3	57.5	56.8	61.2	57.5	53.9	64.7	61.5	59.2	58.0	55.6
BLIP-2 FLAN-T5-XXL	57.8	57.1	63.5	57.0	53.3	66.7	66.2	67.9	57.2	54.3
DeepSeek-VL-Chat-7B	60.3	59.0	58.8	58.4	61.8	68.6	76.9	65.8	60.1	58.0
Yi-VL-6B-Chat	61.3	60.9	63.5	60.7	56.6	66.7	72.3	61.7	61.7	60.1
InternLM-XComposer2-VL	62.1	61.7	62.4	62.3	58.6	70.6	66.2	65.8	63.0	58.7
InternVL-Chat-1.5	66.3	63.6	65.9	68.5	65.8	64.7	76.9	73.5	65.4	64.5
Ide cs2-8B	67.7	67.2	74.1	67.7	62.5	74.5	70.8	68.9	67.0	68.4
Yi-VL-34B-Chat	67.9	67.5	70.6	67.7	63.8	70.6	76.9	74.0	68.2	64.5
MiniCPM-Llama3-2.5	69.4	68.4	71.8	69.4	64.5	80.4	78.5	75.0	69.3	66.9
CogVLM2-Llama3-Chat	70.3	68.9	68.2	70.9	67.8	72.5	86.2	69.9	71.1	69.1
LLaVA-1.6-34B	73.8	73.8	71.8	73.3	71.1	78.4	81.5	79.1	72.9	72.9
Closed-source Models										
GPT-4V	65.9	65.0	69.4	65.3	59.9	76.5	80.0	69.4	66.0	64.0
GPT-4o	72.6	72.5	72.9	73.3	68.4	76.5	75.4	78.6	71.2	72.5
Gemini-1.5 Pro	73.9	73.7	74.1	74.4	63.2	80.4	83.1	80.1	70.8	75.4
Qwen-VL-MAX	74.8	74.7	71.8	74.6	73.0	76.5	84.6	80.1	74.5	72.9
Humans										
Human_avg	90.3	90.0	88.2	91.4	86.6	96.1	92.3	84.7	89.1	92.2
Human_best	98.2	97.9	98.8	98.3	97.4	100.0	100.0	98.0	98.0	98.8

Table 3: Overall results of different MLLMs and humans on different domains and emotions. The best-performing model in each category is in bold, and the second best is underlined.

4.2.2 Disparity between Open-source and Closed-source Models

The results on II-Bench reveal that closed-source models generally perform better, with open-source models exhibiting a larger variance. However, some open-source models show excellent performance. The highest scores for open-source and closed-source models are LLaVA-1.6-34B (73.8%) and Qwen-VL-MAX (74.8%), respectively. Top open-source models like CogVLM2-Llama3-Chat-19B, MiniCPM-Llama3-2.5, Yi-VL-34B-Chat, Ide cs2-8B, and InternVL-Chat-1.5 outperform the closed-source model GPT-4V's 65.9% accuracy but fall short of GPT-4o's 72.6%.

According to our analysis, the image implication understanding not only tests the model's image understanding ability but also tests the model's multi-hop reasoning ability. From the image understanding perspective, top open-source MLLMs perform closely to GPT-4V on various OCR-related benchmarks [41, 47, 53] and general multimodal benchmarks [39, 67, 40, 32, 31]. In terms of logical reasoning, multi-hop reasoning ability is crucial, and LLMs used in MLLMs like Llama3-Chat-8B, InternLM2-Chat-20B, and Yi-34B-Chat exhibit strong performance in reasoning and mathematics benchmarks [5, 68, 22, 39, 11]. Conversely, InstructBLIP-T5-XL, with weaker multi-hop reasoning ability from its language model Flan-T5-XL, shows the lowest accuracy at 47.3%.

4.2.3 Model Performance across Different Domains and Emotions

In terms of domain performance, our results in Table 3 indicate that the models generally perform better in the Environment, Other, Life and Society domains, achieving higher accuracy. Conversely, the accuracy is lower in the Art and Psychology domains, which suggests that while the models generalize well in common domains, they struggle with the more abstract and logically demanding information found in Art and Psychology.

From an emotional perspective, the models tend to exhibit higher accuracy when the image implications convey positive emotions, while accuracy is the lowest for images with negative emotions. This discrepancy highlights that the models' preferences do not align with those of humans, as humans are signi cantly more sensitive to negative implications. Additionally, the results suggest that the models are overly biased towards positive responses, potentially re ecting a positive emotion bias in the training data.

Models	None	CoT	Domain	Emotion	Rhetoric
Open-source Models					
InstructBLIP-T5-XL	47.3	30.0	47.8	49.8	47.6
BLIP-2 FLAN-T5-XL	52.8	42.0	51.4	51.8	51.5
mPLUGw-OWL2	53.2	54.2	54.5	55.0	55.7
Qwen-VL-Chat	53.4	51.6	54.9	57.0	54.0
InstructBLIP-T5-XXL	56.7	50.8	56.7	58.7	56.0
Mantis-8B-siglip-Llama3	57.5	56.7	57.1	57.0	58.0
BLIP-2 FLAN-T5-XXL	57.8	42.5	57.5	58.4	57.3
DeepSeek-VL-Chat-7B	60.3	59.2	60.4	63.3	59.8
Yi-VL-6B-Chat	61.3	60.8	60.8	62.8	60.4
InternLM-XComposer2-VL	62.1	60.7	60.9	61.5	61.6
InternVL-Chat-1.5	66.3	63.3	66.6	67.4	65.6
Ide cs2-8B	67.7	67.7	67.0	68.6	66.6
Yi-VL-34B-Chat	67.9	67.6	67.7	70.1	67.6
MiniCPM-Llama3-2.5	69.4	67.4	70.3	70.8	69.3
CogVLM2-Llama3-Chat-19B	70.3	69.3	69.1	71.7	69.3
LLaVA-1.6-34B	73.8	60.0	73.1	75.3	73.3
Closed-source Models					
GPT-4V	65.9	68.4	66.0	68.3	69.3
GPT-4o	72.6	75.7	72.6	74.2	71.3
Gemini-1.5 Pro	73.9	68.2	73.1	70.5	71.3
Qwen-VL-MAX	74.8	74.1	74.1	75.5	73.6

Table 4: Overall results of different prompts on II-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.

4.2.4 Analysis on different prompt skills

We present a comprehensive analysis of prompt skills, with detailed results in Table 4.

Analysis of Chain-of-Thought (CoT). The Chain-of-Thought (CoT) prompting skill is evaluated to determine its impact on model performance in Table 4. The results indicate that CoT has no significant effect on improving accuracy. In some cases, particularly with smaller open-source models, the accuracy even decline when CoT is used. For example, CogVLM2-Llama3-Chat-19B scores 70.3% without CoT and drops to 69.3% with CoT, InternVL-Chat-1.5 scores 66.3% and 63.3% as the same.

One primary reason for the observed decline in performance with CoT prompting is that many models fail to adhere to the required format. Specifically, CoT prompts require models to output both an analysis and a final answer in a fixed format. However, many models only provide the analysis without concluding with the final answer. We manually check the outputs and find that models either fail to explicitly generate the answer option after the analysis (instead of generating the content of the answer) or select multiple options, which reflect the decline in instruction following ability, leading to the failure of regex matching. An obvious example is BLIP-2 FLAN-T5-XXL, where using the CoT prompt results 15.8% increase in responses that fail to match our regex compared to the direct answer prompt. This issue is highlighted in Appendix E, where we present statistics on model outputs. The lack of a final answer in CoT responses leads to extraction failures, which negatively impacts performance.

Another contributing factor is that CoT prompting does not universally enhance performance across all types of tasks. In evaluations such as MMLU [61], C-Eval [26], where the primary focus is not on logical reasoning or mathematical problem-solving, CoT prompting often does not lead to better results. The same phenomenon is also observed in our experiments, which contrast with tasks that inherently benefit from step-by-step reasoning where CoT can be more effective. These findings align with other benchmarks [69, 34, 21].

Analysis of Different Types and Domains. To evaluate the impact of different label information on model accuracy, we conduct an ablation study by providing corresponding label information (Emotion, Domain, Rhetoric) for the images in the prompt. The results in Table 4 indicate that

Emotion labels signi cantly enhance model accuracy, followed closely by Domain and Rhetoric labels, which exhibit similar effectiveness.

This outcome is consistent with the human perspective of image implication comprehension. Emotion labels likely provide more intuitive and salient cues that align closely with human interpretative processes, thereby facilitating better model performance. In contrast, Domain and Rhetoric labels, while still bene cial, are not as immediately intuitive or universally applicable, thus resulting in slightly lower effectiveness in improving model accuracy. At the same time, from the perspective of model training, the model has a normal understanding of emotion, unlike the speci c nouns we de ne ourselves in the Rhetoric and Domain labels. The model does not see many descriptions of such speci c nouns during pre-training, which does not help improve accuracy.

Analysis of Few-shot Examples. The results in Table 5 demonstrate that few-shot examples do not enhance the accuracy of the models. Speci cally, the performance tends to drop as more examples are provided. This can be attributed to the models' inferior multi-image capabilities compared to their single-image capabilities, leading to a decline in accuracy with an increasing number of shots. Additionally, as the number of shots increases, the input length becomes longer, and the model's long text ability is insuf cient, resulting in poor long context performance. An example is Qwen-VL-Max, where inputs exceeding 6,000 tokens cause errors. Moreover, chat models generally exhibit good instruction following ability, reducing the necessity for few-shot examples.

Model	0-shot	1-shot	2-shot	3-shot
Qwen-VL-Chat	53.4	43.3	47.9	41.1
Mantis-8B-siglip-Llama3	57.5	55.3	54.2	54.9
GPT-4V	65.9	65.5	67.7	67.1
Ide cs2-8B	67.7	64.1	62.4	59.5
Gemini-1.5 Pro	73.9	73.2	73.8	74.1
Qwen-VL-Max	74.8	74.5	69.6	53.6*

Table 5: Few-shot results of different models on the II-Bench. * means exceeds the context length.

4.3 Error Analysis

In order to perform a comprehensive error analysis of GPT-4V's performance on II-Bench, we randomly select 100 erroneous samples from each domain, in proportion to their representation in the dataset. These samples are meticulously analyzed by expert annotators. As illustrated in Figure 4, GPT-4V's errors can be categorized into the following types: Metaphorical Misunderstanding, Detail Misunderstanding, Detail Ignorance, Surface-Level Interpretation, Reasoning Error, Reject to Answer and Answer Extraction Error. This error analysis is crucial for gaining deeper insights into the capabilities of MLLMs and identifying the current limitations in image comprehension tasks. Understanding these shortcomings can guide researchers in developing and training more robust and performant models in the future. A selection of 77 notable cases, along with detailed analyses, is included in Appendix G, providing further insights into the nature of these errors.

Reminder: although we ltered and sifted as much as possible, some of the negative cases in the appendix are offensive to certain groups of people.

Metaphorical Misunderstanding (36%): Metaphorical Misunderstanding is a common error that GPT-4V makes when generating responses based on image comprehension. This indicates that the model has misunderstood the implications or symbolic meanings within the images. There are two main reasons for this. First, the model might grasp certain aspects of the image's meaning, but its overall understanding of the image's theme is incorrect, as exempli ed by Fig.G28. Second, some implications and hidden meanings require speci c knowledge to be understood, and the model's internal knowledge might not cover

Figure 4: GPT-4V error response distribution.

these areas, leading to an incorrect interpretation of the image's deeper meaning.

Detail Misunderstanding (21%): Detail Misunderstanding is another common mistake made by GPT-4V. Understanding details is very important for models, as inaccuracies in understanding details can sometimes affect how the model interprets the meaning of images. For instance, in Fig.G11, GPT-4V has an error in understanding the details, resulting in an incorrect response.

Other Errors: The remaining errors are detail ignorance (11%), surface-level interpretation (15%), reasoning error (12%), reject to answer (4%), and answer extraction error (1%). The description of these errors can be found in Appendix F.

5 Conclusion

The development of II-Bench for assessing the capabilities of MLLMs represents a signi cant milestone in the journey towards achieving Expert AGI, marking a step into higher-order theory of mind in the exploration of the capabilities of MLLMs. The experimental results show that the current state-of-the-art MLLMs are good at understanding the surface content of image, but the gap between the understanding of image implication and humans is still huge. We found that including information about the emotional polarity of the image in the prompts usually improves the model score, suggesting that the model lacks suf cient emotional understanding of the image, which leads to misinterpretation of the implied meaning. Moreover, we found that humans would implicitly understand neutral and negative emotions much better than models. The vast majority of MLLMs perceive positive emotions better than neutral and negative emotions, and we think that the distribution of training data for MLLMs is more skewed toward positive emotions. We believe II-Bench will stimulate the community to build next generation multimodal foundation models towards expert AGI.

Limitations

We acknowledge several limitations in our study. While II-Bench is comprehensive, the inclusion of subjective elements can lead to varying interpretations, potentially affecting result consistency. Additionally, our benchmark focuses on speci c domains, covering only a portion of human knowledge. The evaluation metrics might not entirely re ect the sophisticated understanding and reasoning abilities of advanced AI systems. These limitations highlight the need for ongoing re nement and expansion of our benchmarks. In future work, we aim to develop and incorporate more stringent and objective test sets to enhance reliability and validity of our benchmark.

Ethics Statement

In developing II-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 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.

Acknowledgement

This work was supported by National Key Research and Development Program of China (2022YFF0902100), China Postdoctoral Science Foundation (2024M753398), Postdoctoral Fellowship Program of CPSF (GZC20232873), Guangdong Basic and Applied Basic Research Foundation (2023A1515110718 and 2024A1515012003), and Shenzhen Basic Research Foundation (JCYJ20210324115614039).

References

[1] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson. Nocaps: Novel object captioning at scale. 2019.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. 2015.

[3] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 2023.

[4] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, et al. Sparks of arti cial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv: 2303.12712, 2023.

[5] R. Cai, Z. Song, D. Guan, Z. Chen, X. Luo, C. Yi, and A. Kot. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. arXiv preprint arXiv:2312.02896, 2023.

[6] C. Cazeaux. Metaphor and Continental Philosophy: From Kant to Derrida. Routledge, London, 2007.

[7] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin. Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793, 2023.

[8] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, et al. Are we on the right way for evaluating large vision-language models?, 2024.

[9] L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, B. Lin, Z. Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. arXiv preprint arXiv:2406.04325, 2024.

[10] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. 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, 2024.

[11] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, et al. Training veri ers to solve math word problems, 2021.

[12] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, et al. A survey on multimodal large language models for autonomous driving. 2024.

[13] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. NIPS, 2024.

[14] P. Desai, T. Chakraborty, and M. S. Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. AAAI, 2022.

[15] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, X. Wei, S. Zhang, H. Duan, M. Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. arXiv preprint arXiv:2401.16420, 2024.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2020.

[17] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023.

[18] T. Ghandi, H. Pourreza, and H. Mahyar. Deep learning approaches on image captioning: A review. ACM Computing Surveys, 2023.

[19] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. 2017.

[20] Z. He, X. Wu, P. Zhou, R. Xuan, G. Liu, X. Yang, Q. Zhu, and H. Huang. Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning, 2024.

[21] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. ICLR, 2021.

[22] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. NeurIPS, 2021.

[23] J. Hessel, A. Marasovic, J. D. Hwang, L. Lee, J. Da, R. Zellers, R. Mankoff, and Y. Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. 2023.

[24] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 2019.

[25] J. Hu, Y. Yao, C. Wang, S. Wang, Y. Pan, Q. Chen, T. Yu, H. Wu, Y. Zhao, H. Zhang, X. Han, Y. Lin, J. Xue, D. Li, Z. Liu, and M. Sun. Large multilingual models pivot zero-shot multimodal learning across languages. arXiv preprint arXiv:2308.12038, 2023.

[26] Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, J. Lei, Y. Fu, M. Sun, and J. He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models, 2023.

[27] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. 2019.

[28] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen. Mantis: Interleaved multi-image instruction tuning. arXiv preprint arXiv:2405.01483, 2024.

[29] K. Ka e and C. Kanan. An analysis of visual question answering algorithms. 2017.

[30] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh. What matters when building vision-language models? arXiv preprint arXiv:2405.02246, 2024.

[31] B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan. Seed-bench-2: Benchmarking multimodal large language models. arXiv preprint arXiv:2311.17092, 2023.

[32] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125, 2023.

[33] C. Li. Large multimodal models: Notes on cvpr 2023 tutorials. arXiv preprint arXiv:2306.14895, 2023.

[34] H. Li, Y. Zhang, F. Koto, Y. Yang, H. Zhao, et al. Cmmlu: Measuring massive multitask language understanding in chinese, 2024.

[35] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. 2014.

[37] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744, 2023.

[38] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. arXiv preprint arXiv:2304.08485, 2023.

[39] H. Liu, Z. Zheng, Y. Qiao, H. Duan, Z. Fei, F. Zhou, W. Zhang, et al. Mathbench: Evaluating the theory and application pro ciency of llms with a hierarchical mathematics benchmark, 2024.

[40] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281, 2023.

[41] Y. Liu, Z. Li, B. Yang, C. Li, X. Yin, et al. On the hidden mystery of ocr in large multimodal models, 2024.

[42] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, Y. Sun, et al. Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525, 2024.

[43] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. 2022.

[44] S. Lu, M. Liu, L. Yin, Z. Yin, X. Liu, and W. Zheng. The multi-modal fusion in visual question answering: a review of attention mechanisms. PeerJ Computer Science, 2023.

[45] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, and T. Mei. Semantic-conditional diffusion networks for image captioning. 2023.

[46] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424, 2023.

[47] M. Mathew, D. Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021.

[48] OpenAI. Chatgpt. https://chat.openai.com/, 2023.

[49] OpenAI. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[50] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. 2015.

[51] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. 2024.

[52] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque. Integrating multimodal information in large pretrained transformers. 2020.

[53] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, et al. Towards vqa models that can read, 2021.

[54] A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach. Towards vqa models that can read. 2019.

[55] G. Stellardi. Heidegger and Derrida on Philosophy and Metaphor: Imperfect Thought for a Humanity Books, Amherst, N.Y., 2000.

[56] J. W. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, et al. Testing theory of mind in large language models and humans. Nature Human Behaviour, 2024.

[57] W. Street, J. O. Siy, G. Keeling, A. Baranes, B. Barnett, M. McKibben, T. Kanyere, A. Lentz, R. I. Dunbar, et al. Llms achieve adult human performance on higher-order theory of mind tasks. arXiv preprint arXiv:2405.18870, 2024.

[58] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, et al. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022.

[59] L. Wachowiak and D. Gromann. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. 2023.

[60] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, et al. Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079, 2023.

[61] C. Won, H. Le, L. Shayne, Z. Barret, T. Yi, F. William, L. Yunxuan, et al. Scaling instruction- netuned language models, 2022.

[62] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. arXiv preprint arXiv:2306.09265, 2023.

[63] Y. Yang, Z. Li, Q. Dong, H. Xia, and Z. Sui. Can large multimodal models uncover deep semantics behind images?, 2024.

[64] Q. Ye, H. Xu, J. Ye, M. Yan, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv preprint arXiv:2311.04257, 2023.

[65] A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, H. Li, J. Zhu, J. Chen, J. Chang, et al. Yi: Open foundation models by 01. ai. arXiv preprint arXiv:2403.04652, 2024.

[66] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490, 2023.

[67] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502, 2023.

[68] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence?, 2019.

[69] G. Zhang, X. Du, B. Chen, Y. Liang, T. Luo, T. Zheng, K. Zhu, Y. Cheng, et al. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark, 2024.

[70] H. Zhang, X. Li, and L. Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023.

[71] W. Zhang, S. M. Aljunied, C. Gao, Y. K. Chia, and L. Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. arXiv preprint arXiv:2306.05179, 2023.

[72] S. Zhong, Z. Huang, S. Gao, W. Wen, L. Lin, M. Zitnik, and P. Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation, 2024.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately re ect the paper's contributions and scope? [Yes] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Limitations.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] See Ethics Statement.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Abstract.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Abstract.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identi able information or offensive content? [Yes] See Ethics Statement
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix B
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Image type and domain statistics

Figure 5: II-Bench speci c image type and domain statistics.

B Data Annotation Protocol

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

B.1 Data Collection

Some websites from which we collect data are as follows:

- https://www.davidebonazzi.com
- https://www.boredpanda.com
- https://themindsjournal.com
- https://naldzgraphics.net/satirical-illustrations-agim-sulaj
- https://www.pinterest.co.uk
- https://www.asafhanuka.com/the-realist

B.2 Data Leakage Prevention

We have made efforts to mitigate data leakage risk:

- Recent Data Collection We have made an effort to collect data from a recent time, which aims to reduce the likelihood that these speci c images are included in the training sets of existing models.
- Brand new, Manually Constructed Questions The questions we posed are entirely new and manually constructed, which are not present in any model's training set. Even if the images themselves are part of the training data, it is unlikely to correctly answer our high-level implication questions.

- Multimodal Data Leakage Assessment. We conduct evaluations where models are tested without providing the corresponding images. The experimental results are shown in the table 6, which indicate that models perform badly without the images, suggesting that the models do not possess pre-existing knowledge of the speci c questions and images, thereby indicating that there is no risk of data leakage.

These measures collectively help ensure that our evaluations are fair and that the risk of data leakage is minimized.

Model	Mode	Accuarcy
deepseek-vl-7b-chat	w. image	60.3
deepseek-vl-7b-chat	w/o. image	36.8 (-23.5)
Qwen-VL-Chat	w. image	53.4
Qwen-VL-Chat	w/o. image	38.0 (-15.4)
ide cs2-8b	w. image	67.7
ide cs2-8b	w/o. image	47.8 (-19.9)

Table 6: The architecture and size of different models.

B.3 General Guidelines

General Principles:

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

Speci c Instructions:

- Each image needs to be categorized as one of the following image types: single-panel comic, multi-panel comic, poster, logo, meme, illustration or painting.
- Each image needs to be categorized as one of the following dif culty 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, psychology, environment or others.
- 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, personi cation or others.
- Each image needs a human explanation.
- Each image needs 1-3 questions about the ne-grained metaphorical implications of the image, each with one correct answer and ve distractor options.

B.4 Additional 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 four 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 four rounds of data quality checks to ensure data quality and ultimately obtain II-Bench.

B.5 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. Annotators must avoid collecting questions that contain any personal information.

C Prompts

In experiments, the prompts of different settings are as follows:

C.1 None

```
instruction: "Instruction: Please try to answer the single-answer multiple choice
    question below based on the picture provided."
prompt_format:
  - |
    Question: {}
    (A) {}
    (B) {}
    (C) {}
    (D) {}
    (E) {}
    (F) {}
    Answer:
```

C.2 Few-shot

```
instruction:
  - |
    Instruction: Please try to answer the single-answer multiple choice question
        below based on the example(with answer) and the corresponding picture.
  - |
    Instruction: Please try to answer the single-answer multiple choice question
        below based on the examples(with answers) and the corresponding pictures.
prompt_format:
  - |
    Question: In the comic image, what deeper societal commentary might Barry ' s
        costume choice at the party represent?
    Picture: <Picture {}>
    (A) The backlash faced when challenging traditional roles.
    (B) The struggle to fit in while also standing out in social circles.
    (C) The challenge of maintaining personal identity in group dynamics.
    (D) The discomfort caused by confronting controversial or taboo topics in
        social settings.
    (E) The effects of poor decision-making on interpersonal relationships.
    (F) The significance of color coordination in party costumes to enhance the
        festive atmosphere.
    Answer: (D)
  - |
    Question: What hidden message can be inferred about the dynamics of fame and
        the collective cultural memory from the text and images of Brendan Fraser
        within the meme?
    Picture: <Picture {}>
    (A) The meme suggests that the public and media often overlook certain
        celebrities in favor of others due to shifting trends and narratives in
        popular culture.
    (B) The imagery suggests that personal struggles of celebrities are often
        overlooked by the public and media.
    (C) It points to a discrepancy between the talent and contributions of
        celebrities and their recognition in the media.
    (D) The focus on Brendan Fraser is meant to highlight how male fashion trends
        drastically changed from the 90s to the present.
    (E) Brendan Fraser is depicted as the quintessential 90s figure, indicating
        that he defined the entire decade ' s style and sensibilities.
    (F) The meme indicates that celebrities who maintain a consistent public
        image are more likely to remain in the spotlight.
```

```
    Answer: (A)
  - |
    Question: What is the metaphorical significance of the glowing eye in this
    ↵   image?
    Picture: <Picture {}>
    (A) It represents the ever-present nature of surveillance in society.
    (B) It symbolizes enlightenment and the pursuit of knowledge.
    (C) It signifies wisdom and the foresight of a leader.
    (D) It depicts the uninterrupted attention and care from protectors.
    (E) It represents the vigilance and unending watchfulness of authority.
    (F) It conveys the omnipresent gaze of societal norms and expectations.
    Answer: (E)
  - |
    Question: {}
    Picture: <Picture {}>
    (A) {}
    (B) {}
    (C) {}
    (D) {}
    (E) {}
    (F) {}
    Answer:
```

C.3 Keywords

The keywords here include one of the following: emotion, domain, rhetoric.

```
instruction: "Instruction: Please try to answer the single-answer multiple choice
↵   question below based on the picture and the key words."
prompt_format:
  - |
    Key words: {}
    Question: {}
    (A) {}
    (B) {}
    (C) {}
    (D) {}
    (E) {}
    (F) {}
    Answer:
```

C.4 CoT

```
instruction: "Instruction: Please try to answer the single-answer multiple choice
↵   question below based on the picture provided. Let ' s think through each
↵   option. Let ' s think step by step."
prompt_format:
  - |
    Question: {}
    (A) {}
    (B) {}
    (C) {}
    (D) {}
    (E) {}
    (F) {}
    Explanation:
    Answer:
```

D Results on Different Types, Dif culties and Rhetoric

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

D.1 Image Types and Dif culty

We present a comprehensive comparison of different MLLMs and humans on image types and different levels of dif culty, with detailed results in Table 7.

	Overall (1,399)	Illu. (436)	Meme (292)	Poster (133)	MPC (359)	SPC (104)	Logo (71)	Paint. (101)	Easy (786)	Mid. (465)	Hard (148)
Open-source Models											
InstructBLIP-T5-XL	47.3	40.8	53.8	51.9	47.4	45.2	57.8	44.6	50.1	44.7	39.9
BLIP-2 FLAN-T5-XL	52.8	44.5	59.6	60.2	54.3	54.8	69.0	47.5	56.1	49.3	46.0
mPLUGw-OWL2	53.2	43.1	63.0	59.4	56.3	55.8	63.4	43.6	56.0	50.5	46.6
Qwen-VL-Chat	53.4	42.7	64.0	61.7	56.3	43.3	57.8	55.5	56.7	51.8	40.5
InstructBLIP-T5-XXL	56.7	47.9	67.1	63.2	58.5	51.9	60.6	54.5	58.8	55.9	48.0
Mantis-8B-siglip-llama3	57.5	47.7	66.1	65.4	59.6	58.6	69.0	55.5	58.9	56.6	52.7
BLIP-2 FLAN-T5-XXL	57.8	47.7	66.1	65.4	59.6	58.6	69.0	55.5	58.9	56.6	52.7
DeepSeek-VL-Chat-7B	60.3	47.7	70.2	72.2	65.7	59.6	67.6	51.5	64.3	57.0	49.3
Yi-VL-6B-Chat	61.3	53.2	68.5	63.9	62.4	63.5	74.6	59.4	64.1	59.4	52.0
InternLM-XComposer2-VL	62.1	53.0	68.8	65.4	66.6	60.6	74.7	60.4	65.3	60.4	50.7
InternVL-Chat-1.5	66.3	54.6	78.1	71.4	71.6	66.4	71.8	59.4	69.7	64.3	54.1
Ide cs2-8B	67.7	58.5	77.4	76.7	68.8	59.6	81.7	66.3	68.8	69.5	56.1
Yi-VL-34B-Chat	67.9	56.7	81.9	70.7	71.6	60.6	77.5	58.4	71.1	66.7	54.7
MiniCPM-Llama3-2.5	69.4	61.9	80.5	79.0	69.1	65.4	77.5	63.4	70.2	69.7	64.2
CogVLM2-Llama3-Chat	70.3	60.8	82.9	75.9	73.5	66.4	74.7	60.4	74.2	66.9	60.8
LLaVA-1.6-34B	73.8	62.8	84.6	80.5	80.5	67.3	80.3	67.3	77.5	71.4	61.5
Closed-source Models											
GPT-4V	65.9	55.1	79.8	73.7	69.1	64.4	67.6	58.4	69.6	61.9	58.8
GPT-4o	72.6	64.7	81.2	78.2	76.9	72.1	80.3	66.3	76.6	67.5	67.6
Gemini-1.5 Pro	73.9	66.7	82.2	79.7	74.6	70.2	81.7	74.3	75.1	74.2	66.9
Qwen-VL-MAX	74.8	65.1	84.3	85.0	78.0	71.2	73.2	75.3	77.4	73.3	66.2
Humans											
Human_avg	90.3	90.3	89.6	88.4	90.8	92.3	92.3	93.6	90.7	90.1	88.5
Human_best	98.2	98.4	99.3	99.3	96.7	97.1	100.	99.0	98.1	98.3	98.7

Table 7: Overall results of different MLLMs and humans on different image types and different levels of dif culty. The best-performing model in each category is in bold, and the second best is underlined. For brevity, Illu. refers to Illustration, MPC refers to Multi-panel Comic, SPC refers to Single-panel Comic, Paint. refers to Painting and Mid. refers to Middle.

D.2 Rhetoric

We present a comprehensive comparison of different MLLMs and humans on on different rhetoric, with detailed results in Table 8.

	Overall (1,399)	Meta. (1106)	Exag. (227)	Symb. (271)	VisD. (88)	Anti. (35)	Anal. (42)	Pers. (128)	Contrast (274)	Others (55)
Open-source Models										
InstructBLIP-T5-XL	47.3	47.6	44.9	49.8	45.5	57.1	42.9	50.8	50.7	41.8
BLIP-2FLAN-T5-XL	52.8	53.6	48.9	52.8	46.6	54.3	45.2	54.7	58.4	49.1
mPLUGw-OWL2	53.2	53.4	51.5	49.8	44.3	45.7	47.6	55.5	50.7	56.4
Qwen-VL-Chat	53.4	52.9	52.9	50.2	45.5	45.7	59.5	57.8	55.1	47.3
InstructBLIP-T5-XXL	56.7	57.8	57.3	53.5	51.1	51.4	42.9	63.3	60.2	50.9
Mantis-8B-siglip-llama3	57.5	56.6	56.8	53.1	58.0	48.6	64.3	60.9	60.2	63.6
BLIP-2FLAN-T5-XXL	57.8	58.4	55.1	56.5	56.8	54.3	52.4	64.1	59.9	52.7
DeepSeek-VL-Chat-7B	60.3	59.8	56.8	54.6	53.4	65.7	54.8	61.7	66.1	60.0
Yi-VL-6B-Chat	61.3	61.1	59.0	59.0	58.0	54.3	64.3	61.7	63.1	54.5
InternLM-XComposer2-VL	62.1	61.1	57.3	62.4	56.8	54.3	66.7	71.1	63.9	67.3
InternVL-Chat-1.5	66.3	65.7	64.8	64.2	60.2	57.1	64.3	76.6	68.2	65.5
Ide cs2-8B	67.7	67.7	67.8	63.5	68.2	77.1	66.7	66.4	70.8	70.9
Yi-VL-34B-Chat	67.9	67.7	64.8	60.5	69.3	65.7	71.4	73.4	70.8	65.5
MiniCPM-Llama3-2.5	69.4	69.6	68.7	66.1	63.6	68.6	69.0	65.6	72.3	65.5
CogVLM2-Llama3-Chat	70.3	70.8	72.2	64.2	62.5	71.4	78.6	70.3	72.6	60.0
LLaVA-1.6-34B	73.8	73.1	73.1	68.6	68.2	80.0	71.4	77.3	75.5	74.5
Closed-source Models										
GPT-4V	65.9	65.2	60.8	61.6	67.0	80.0	69.0	72.7	68.2	70.9
GPT-4o	72.6	71.3	69.2	70.5	63.6	71.4	78.6	78.1	72.6	74.5
Gemini-1.5 Pro	73.9	74.0	75.8	68.3	70.5	68.6	78.6	75.0	74.5	69.1
Qwen-VL-MAX	74.8	73.9	74.0	67.5	68.2	74.3	71.4	78.9	79.2	81.8
Humans										
Human_avg	90.3	90.1	89.9	91.3	88.6	88.6	86.9	94.1	90.0	88.2
Human_best	98.2	98.1	98.2	98.9	100.0	94.3	97.6	97.7	96.7	100.0

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 and Pers. refers to Personi cation.

E Additional Details of Results

E.1 Answer Extraction

To mitigate the issue of models understanding the question but failing to output the correct option, we have implemented a robust answer extraction method and we use the same method for both CoT and non-CoT responses to ensure consistency. Specically, our extraction method involves the following steps:

- Pattern Matching: We first search for the presence of option labels within the model's output using specic patterns.
- Frequency Analysis: If multiple option labels are found, we count their occurrences and select the most frequently occurring label.

E.2 Statistics of Output

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 2%, which is 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 as Miss. For convenience of presentation, some model names are abbreviated. The specic meanings of these abbreviations are consistent with the full model names used elsewhere in the paper.

Mode	Metric	BLIP2-XL	BLIP2-XXL	CogVLM2	DeepSeek	InsBLIP-XL
CoT	Acc	42.0	42.5	69.3	59.2	30.0
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	11.7	15.8	0.0	0.2	11.7
Domain	Acc	51.4	57.5	69.1	60.4	47.8
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	0.0	0.0
Emotion	Acc	51.8	58.4	71.7	63.3	49.8
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.1	0.1	0.0	0.0	0.0
None	Acc	52.8	57.8	70.3	60.3	47.3
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	0.1	0.0
Rhetoric	Acc	51.5	57.3	69.3	59.8	47.6
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.0	0.0	0.0	0.0	0.0

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

Mode	Metric	InsBLIP-XXL	XComposer2	InternVL	LLaVA-1.6	MiniCPM-2.5
CoT	Acc	50.8	60.7	63.3	60.0	67.4
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	2.2	2.3	0.1	12.4	0.0
Domain	Acc	56.7	60.9	66.6	73.1	70.3
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.4	0.0	0.0	0.0	0.0
Emotion	Acc	58.7	61.5	67.4	75.3	70.8
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.4	0.1	0.0	0.0	0.0
None	Acc	56.7	62.1	66.3	73.8	69.4
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.4	0.0	0.0	0.1	0.0
Rhetoric	Acc	56.0	61.6	65.6	73.3	69.3
	Error	0.0	0.0	0.0	0.0	0.0
	Miss	0.4	0.0	0.0	0.0	0.0

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

Mode	Metric	mPLUGw-OWL2	GPT-4o	Yi-VL-34B	Yi-VL-6B
CoT	Acc	54.2	75.7	67.6	60.8
	Error	0.0	0.1	0.0	0.0
	Miss	0.2	10.7	0.0	0.0
Domain	Acc	54.5	72.6	67.7	60.8
	Error	0.0	0.0	0.0	0.0
	Miss	0.0	5.2	0.0	0.1
Emotion	Acc	55.0	74.2	70.1	62.8
	Error	0.0	0.1	0.0	0.0
	Miss	0.0	0.3	0.0	0.1
None	Acc	53.2	72.6	67.9	61.3
	Error	0.0	0.0	0.0	0.0
	Miss	0.0	0.2	0.0	0.0
Rhetoric	Acc	54.6	71.3	67.6	60.4
	Error	0.0	0.1	0.0	0.0
	Miss	0.0	0.1	0.0	0.0

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

Mode	Metric	GPT-4V	Qwen-Chat	Qwen-MAX	Gemini1.5	Mantis	Ide cs2
CoT	Acc	68.4	51.6	74.1	68.2	56.7	67.7
	Error	0.4	0.0	0.4	0.0	0.0	0.0
	Miss	0.4	10.7	0.1	0.3	0.0	0.1
Domain	Acc	66.0	54.9	74.1	73.1	57.1	67.0
	Error	0.1	0.0	0.6	0.0	0.0	0.0
	Miss	2.6	5.2	0.0	1.3	0.0	0.1
Emotion	Acc	68.3	57.0	75.5	70.5	57.0	68.6
	Error	0.6	0.0	0.6	0.0	0.0	0.0
	Miss	2.0	5.1	0.1	2.1	0.0	0.1
None	Acc	65.9	53.4	74.8	73.9	57.5	67.7
	Error	0.9	0.0	0.4	0.0	0.0	0.0
	Miss	4.2	7.0	0.1	1.3	0.0	0.0
Rhetoric	Acc	69.3	54.0	73.6	71.3	58.0	66.6
	Error	0.1	0.0	0.6	0.0	0.0	0.0
	Miss	1.9	6.4	0.4	1.1	0.0	0.0
1-shot	Acc	65.5	43.3	74.5	73.2	55.3	64.1
	Error	0.3	0.0	0.4	0.1	0.0	0.0
	Miss	1.6	1.9	0.6	0.7	0.0	0.0
2-shot	Acc	67.7	47.9	69.6	73.8	54.2	62.4
	Error	0.2	0.0	0.4	0.1	0.0	0.0
	Miss	0.5	5.7	1.0	0.4	0.0	0.0
3-shot	Acc	67.1	41.1	53.6	74.1	54.9	59.5
	Error	0.6	0.0	16.7	0.0	0.0	0.0
	Miss	0.4	5.4	0.1	0.2	0.0	0.0

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

F Other Errors

Detail Ignorance (11%): Detail Ignorance refers to GPT-4V overlooking certain crucial details or elements in images. This oversight can sometimes lead to 'Metaphorical Misunderstanding.' Many images convey implications and meanings through their details. By fully utilizing the content of an image and not ignoring any details, one can better understand and uncover the hidden meanings within the image.

Surface-Level Interpretation (15%): This error occurs when GPT-4V focuses only on the super- cial information in images while ignoring their deeper meanings. It may only offer a super cial interpretation of the images, failing to grasp their complexity or multifaceted meanings. For example, in Fig.G12, GPT-4V noticed only the surface-level information of 'Internet' and overlooked the deeper element of 'the emotional value of the meme,' thus providing an incorrect response.

Reasoning Error (12%): Reasoning errors may arise even when GPT-4V accurately understands the content of an image, such as in the illustration of Fig.G18. In such instances, errors occur during complex problem-solving that demands advanced logical and mathematical reasoning. This type of error often stems from the model's limited capabilities in handling intricate logic and mathematics, highlighting areas where GPT-4V requires further re nement to improve its reasoning accuracy.

Reject to Answer (4%): Reject to Answer is a common error encountered in GPT-4V. This phenomenon typically manifests for two primary reasons. First, the model may determine that the information provided in the question and accompanying images is insuf cient to deduce the underlying meanings or implications, thereby rendering it incapable of ascertaining the correct answer, such as Fig.G38. The second reason involves the model's evaluation of the content as potentially harmful or inappropriate. In such cases, the model opts to withhold a response as a preventive measure against disseminating sensitive or damaging information. This safety mechanism is crucial in maintaining ethical standards and preventing the propagation of harmful content. However, this can also lead to frustrations when users expect a response that the model is programmed to avoid for ethical reasons.

Answer Extraction Error (1%): Answer Extraction Error refers to the phenomenon of extracting incorrect answers using a regularization formula from GPT-4V's output. This issue may arise due to GPT-4V's weak instruction-following capabilities on certain questions, where it fails to generate answers in the correct format according to the rules.

G Case Study

The appendix is our sample analysis of GPT-4V, including an analysis of 49 error examples and 28 correct examples.

List of Case Study Figures

Figure G1: A sample correct case of the domain.
Back to List of Figures

Figure G2: A sample correct case of the domain.
Back to List of Figures

Figure G3: A sample correct case in the Life domain.
Back to List of Figures

Figure G4: A sample correct case of the domain.
Back to List of Figures

Figure G5: A sample correct case of the domain.
Back to List of gures

Figure G6: A sample correct case of the Lock domain.
Back to List of Figures

Figure G7: A sample error case of the domain.
Back to List of Figures

Figure G8: A sample error case of the domain.
Back to List of Figures

Figure G9: A sample error case of the domain.
Back to List of gures

Figure G10: A sample error case in the Life domain.
Back to List of Figures

Figure G11: A sample error case in the Life domain.
Back to List of Figures

Figure G12: A sample error case in the Lunar domain.
Back to List of Figures

Figure G13: A sample error case in the Life domain.
Back to List of Figures

Figure G14: A sample error case in the Life domain.
Back to List of Figures

Figure G15: A sample error case in the Life domain.
Back to List of Figures

Figure G16: A sample error case in the Life domain.
Back to List of Figures

Figure G17: A sample error case in the Life domain.
Back to List of Figures

Figure G18: A sample error case in the Life domain.
Back to List of Figures

Figure G19: A sample error case in the Life domain.
Back to List of Figures

Figure G20: A sample error case in the Life domain.
Back to List of Figures

Figure G21: A sample correct case of Society domain.
Back to List of Figures

Figure G22: A sample correct case of Society domain.
Back to List of gures

Figure G23: A sample correct case of Society domain.
Back to List of gures

Figure G24: A sample correct case of Society domain.
Back to List of gures

Figure G25: A sample correct case of Society domain.

Back to List of Figures

Figure G26: A sample correct case of Society domain.

Back to List of Figures

Figure G27: A sample correct case of Society domain.
Back to List of gures

Figure G28: A sample error case of Society domain.
Back to List of Figures

Figure G29: A sample error case of Society domain.
Back to List of Figures