

Seeing and Understanding: A Human-Centric Evaluation of Multimodal Language Models in Chinese Contexts

Evaluation of Multimodal Language Models in Chinese Contexts

JIAXIN LI

Faculty of Business and Economics, The University of Hong Kong, li_jiabin@connect.hku.hk

ZHENHUI (JACK) JIANG

Faculty of Business and Economics, The University of Hong Kong, jiangz@hku.hk

YANG LIU

School of Management, Xi'an Jiaotong University, liuyang.alison@xjtu.edu.cn

HAOZHE XU

School of Management, Xi'an Jiaotong University, xuhaozhe@stu.xjtu.edu.cn

The rapid rise of multimodal large language models (MLLMs) has created a pressing need for systematic evaluations. Yet most existing benchmarks are primarily developed in English-language contexts, rely heavily on automated scoring, and overlook real-world, culturally grounded capabilities. In this work, we present a comprehensive evaluation framework for Chinese-language image understanding, incorporating four core capability aspects: visual perception and recognition, visual reasoning and analysis, visual aesthetics and creativity, and safety and responsibility. We benchmark 17 state-of-the-art MLLMs from China and the U.S. (as of early 2025) across 22 diverse tasks. All image-text pairs are carefully constructed to ensure strong visual grounding and avoid reliance on textual priors. Our framework introduces a structured, multidimensional scoring protocol and incorporates human ratings to ensure fairness, interpretability, and depth. This evaluation offers a systematic, transparent, and accessible assessment of model performance in Chinese-language contexts. Additionally, we provide comparative insights into the strengths and weaknesses of MLLMs from China and the U.S., contributing to more informed development and deployment of multimodal AI systems.

CCS CONCEPTS • **Human-centered computing** → **Human computer interaction (HCI)** → HCI design and evaluation methods;

Additional Keywords and Phrases: Multimodal large language models, LLM evaluation, Human-centric evaluation

1 INTRODUCTION

Recent multimodal large language models (MLLMs) have demonstrated remarkable progress in understanding and reasoning across visual and textual modalities [2, 5, 6, 8, 12], revealing significant potential for real-world applications across industries. However, the systematic evaluation of these models' image understanding capabilities—particularly in application-oriented and non-English contexts—remains underdeveloped. Although several benchmarks have been

proposed to evaluate MLLMs [7, 11, 15, 23, 31], challenges persist in terms of reliability, interpretability, and practical relevance. First, some test tasks do not genuinely assess a model’s visual understanding, as they can often be completed using only textual information or embedded world knowledge without actual visual perception or reasoning, especially if unintentional data leakage in training happens [4]. Additionally, the growing reliance on large language models (LLMs) as automatic evaluators may introduce biases or fail to capture nuanced aspects of multimodal outputs [10]. Third, a significant portion of existing benchmarks are designed primarily for model development. They offer limited guidance for end users or industry practitioners seeking to select or deploy models in real-world settings, especially within non-English environments.

To address these gaps, we propose a comprehensive evaluation framework focused on assessing the image understanding capabilities of multimodal models in Chinese-language contexts. We apply this framework to benchmark 17 leading models released in China and the U.S as of early 2025. Our framework organizes complex multimodal tasks into four core capability aspects: visual perception and recognition, visual reasoning and analysis, visual aesthetics and creativity, and safety and responsibility. To enable reliable and interpretable evaluation, we employ highly qualified human evaluators and adopt a multidimensional scoring protocol. In addition, we apply the LLM-as-a-judge paradigm to cross-validate our evaluation results. Our work showcases the current strengths and limitations of existing models, offering directions for future improvements and informing model selection for both general users and industry stakeholders.

We contribute to related work in several ways. First, we introduce a structured evaluation framework and a curated test set for assessing image understanding capabilities in the Chinese-language context. The dataset comprises 22 tasks with problems of varying difficulty levels, spanning OCR, object recognition, visual reasoning, aesthetic judgment, and safety assessment, covering a broad spectrum of application-oriented abilities and culturally grounded scenarios.

Second, we design a multidimensional scoring protocol supported by trained native-speaker raters, enabling nuanced, context-sensitive judgments that go beyond what automatic metrics or multiple-choice formats can capture. This human-centric approach ensures fairness, interpretability, and alignment with real-world use cases, particularly in safety-critical and open-ended generation tasks.

Third, we assess 17 state-of-the-art multimodal models from China and the U.S, providing a comparative analysis that uncovers regional strengths, weaknesses, and performance trends. We provide an in-depth analysis based on the evaluation to support MLLM development, practical deployment and guide adoption across sectors. By offering transparent comparisons and culturally grounded tasks, our work aims to advance more inclusive and real-world-relevant assessment of MLLM.

2 RELATED WORK

Multimodal Language Models. With the rapid advancement of LLMs, research has extended them to the multimodal domain by incorporating visual encoders, forming MLLMs [31]. Early models focused on aligning vision and language through contrastive learning on large-scale image-text pairs [27], while later approaches, such as BLIP [16], introduced diverse supervision tasks such as captioning to enhance multimodal pretraining.

Recent MLLMs have evolved from modular pipelines toward unified architectures that embed visual features directly into language modeling. Examples like MiniGPT-4 [38], InstructBLIP [5], and LLaVA [20] align pretrained LLMs with visual inputs through lightweight projection and instruction tuning, achieving strong multi-turn visual question answering and image-grounded dialogue capabilities. Meanwhile, natively multimodal architectures, such as GPT-4o, integrate vision and language within a single Transformer stack, enabling real-time, end-to-end image-text interaction. Other leading

MLLMs, including Gemini 1.5, GLM-4v, and Qwen-VL, demonstrate deeply integrated designs with strong performance in document understanding, visual reasoning, and complex instruction following.

As MLLMs are applied to diverse real-world scenarios, such as document digitization, autonomous driving [28], and medical image analysis [26], their accuracy, robustness, and contextual reasoning become increasingly critical. This underscores the urgent need for reliable evaluation and benchmarking.

Evaluations of MLLMs. Quantitative evaluation is essential for assessing the strengths and limitations of MLLMs. Classical benchmarks such as COCO Captions [18], NoCaps [1], and VQAv2 [9] focus on isolated tasks like image captioning or visual question answering. These typically involve short-form or fixed answer formats and exhibit limited linguistic diversity, making them insufficient for evaluating general-purpose, open-ended multimodal understanding and reasoning.

More recent efforts have shifted from narrow, task-specific assessments to more comprehensive and integrated benchmarks that span a wide range of capabilities. For instance, MME [7] and MMBench [21] offer fine-grained, large-scale evaluations using binary and multiple-choice questions, covering skills such as object recognition, OCR, numerical understanding and commonsense reasoning. SEED-Bench [14] introduces a hierarchical framework that integrates recognition and generation tasks, supported by a refined answer-ranking strategy. In contrast, MM-Vet [32] focuses on evaluating complex, integrated tasks that combine six core vision language capabilities, including recognition, knowledge, OCR, spatial awareness, language generation and math, by leveraging GPT-4 to score open-ended responses. While these benchmarks represent significant progress, many remain predominantly English-centric, rely heavily on LLM-based auto-evaluation, and offer limited coverage of culturally grounded reasoning or robust visual grounding [4].

To address these gaps, recent studies have emphasized the importance of linguistic and cultural context in multimodal evaluation. MULTI [39] offers a large-scale Chinese multimodal dataset derived from authentic examination questions and shows that some models can approach or even surpass human expert baselines on certain tasks. However, CII-Bench [35] and TCC-Bench [30] demonstrate that current MLLMs still struggle with Chinese cultural imagery and high-level semantics. Furthermore, prior work shows that multilingual models have difficulty interpreting figurative language and culturally embedded expressions, revealing persistent gaps in culturally contextualized reasoning [19]. These works collectively suggest that strong performance on general or English-centric benchmarks does not necessarily translate to robust capability in Chinese or culturally rich settings, underscoring the need for evaluation frameworks that directly engage with Chinese-language and culturally situated multimodal understanding.

Our work aligns with this shift toward integrated and culturally aware multimodal evaluation, while differing in several key respects. First, our evaluation is explicitly designed around Chinese-language contexts with secular cultural grounding, incorporating Chinese idioms, festivals, memes, historical references, and educational content from Chinese college and high school entrance exams (Gaokao and Zhongkao). Second, we enforce strong visual grounding and data novelty by constructing image-text pairs that cannot be answered using textual priors or general knowledge alone, ensuring that models must rely on actual visual understanding. Third, we combine both closed-ended (e.g., multiple-choice) and open-ended (e.g., free-form question answering) tasks, allowing for evaluation across a wider range of real-world, application-oriented scenarios. Finally, instead of relying on LLMs as judges, which may introduce inconsistency or bias, we employ native Chinese speakers as human raters and adopt a multidimensional scoring protocol to ensure fairness, cultural validity, and interpretability.

3 EVALUATION SUITE

3.1 Evaluation Framework

Our evaluation framework is organized around four key aspects of multimodal models’ image understanding capabilities: 1) visual perception and recognition, 2) visual reasoning and analysis, 3) visual aesthetics and creativity, and 4) safety and responsibility. These aspects reflect a progression from basic perceptual abilities to higher-level interpretive, generative, and ethical competencies. This structure is informed by established computer vision benchmarks, emerging multimodal reasoning evaluations, creative content generation use cases, and safety concerns in real-world deployment. Figure 1 summarizes the evaluation tasks with illustrative examples. The sample questions are originally in Chinese and have been translated into English for clarity. Detailed task definitions are provided in Appendix A.1.

Visual perception and recognition. This aspect evaluates whether a model can accurately identify and understand core visual elements in an image, including text, objects, attributes, and spatial relationships. Its inclusion is motivated by long-standing foundational tasks in computer vision (e.g., object recognition, OCR, image description), which serve as prerequisites for higher-level reasoning. Failures in this foundational aspect can lead to hallucinations, where models fabricate elements not present in the image [17], thereby undermining reasoning or generation. For this aspect, we assess the capabilities through tasks such as recognizing Chinese characters, mathematical formulas, or code, identifying public figures or landmarks, and generating concise or detailed image descriptions. These tasks support real-world applications such as document analysis, visual search, and information extraction, and show strong potential for industrial applications like warehouse management and logistics.

Visual reasoning and analysis. This aspect tests the model’s ability to make inferences based on visual content, often requiring contextual understanding, reasoning skills, and external knowledge. This aligns with recent visual-language model benchmarks [21, 32], which emphasize that effective image understanding must move beyond object identification to encompass relational interpretation and knowledge grounding. Tasks used for assessment include answering questions involving social knowledge, interpreting culturally embedded memes, and analyzing visual data like graphs or charts. These evaluations assess whether a model can go beyond surface-level recognition to perform complex interpretation and deduction tasks.

Visual aesthetics and creativity. This aspect focuses on evaluating the model’s higher-order abilities in understanding, association, and expression, which are increasingly important in creative and cultural applications. Specifically, we assess the model’s ability to judge the aesthetic quality of an image (e.g., composition, lighting, color) [13] and to generate creative, contextually appropriate text based on visual input (e.g., storytelling, classical poetry, advertising slogans, or scientific reports). Aesthetic judgment goes beyond perception or symbol recognition. It involves imagination, sensitivity, and the ability to grasp subtle, non-obvious structures, such as elegance or balance in images [34]. Similarly, creativity depends on recombining loosely connected concepts [25], as seen in metaphor or storytelling. By evaluating aesthetic judgment and image-based creative writing, we can test whether a model can interpret, transform, and express complex implicit patterns. They reflect the model’s potential in fields such as the cultural and creative industries, education, and digital content production, where human-like intelligence and expressive depth are essential.

Safety and responsibility. This aspect ensures that the model operates in a trustworthy, socially responsible manner and adheres to legal and ethical standards, reflecting the growing emphasis on AI safety in open-world deployment. We structure this evaluation into two capabilities and assess them across several scenarios. Hazard awareness refers to the ability to recognize and appropriately respond to inputs involving illegal activities, physical harm scenario, and sensitive topics such as gambling, drugs, and pornographic content. Responsible interaction focuses on the model’s ethical and

socially aware engagement, including its ability to avoid biased response, respect moral norms, and refrain from providing unqualified or potentially harmful advice. These are especially important for public-facing or high-stakes application scenarios.



Figure 1: Diagram of our evaluation framework.

3.2 Dataset

In existing benchmarking endeavors, some tasks were not properly developed so that models can answer visual questions correctly by exploiting textual cues in prompts or drawing on memorized knowledge from pretraining data [4]. To address these issues, we construct new question sets that emphasize visual grounding carefully.

Closed-ended questions are used to assess logical reasoning and disciplinary knowledge, comprising over 170 image-question pairs. For logical reasoning, we adapt Chinese-language items from the general VQA section of the MathVista dataset [22]. We apply minor modifications during translation, such as rewording prompts or shuffling answer options to

reduce potential overlap with training data. While the question set contains a mix of free-form and multiple-choice questions, we ensure that all adapted items retain a clearly defined correct answer. Disciplinary knowledge questions are mainly drawn from 2024 Chinese high school and college entrance examinations and are formatted as multiple-choice. Some image-based fill-in-the-blank questions are converted to multiple-choice format for consistency. These questions are newly released, the likelihood of their inclusion in large-scale pretraining corpora is relatively low. We also include a small number of university-level exam questions, selected from real course exams at top Chinese universities and the MMMU dataset [33].

Open-ended questions are used across tasks such as object recognition, image description, social and cultural Q&A, chart analysis, content generation, and safety assessment, totaling over 220 image-question pairs. These items are constructed through three main approaches: 1) Adapting from public English-language benchmarks. We translate and adapt textual questions from datasets such as OK-VQA [23], aligning them with our evaluation goals. These questions are then paired with other copyright-free images (e.g., from Pexels) to create new, visually grounded test items in Chinese. 2) Modifying existing image-text pairs. Images from datasets like nocaps [1] are reused, but all accompanying prompts and answers are newly authored to match specific evaluation objectives. Even if the models have seen the images during training, the questions they face are novel and require genuine visual understanding rather than memorization of existing annotations. 3) Creating original items from scratch. We construct entirely new image-question pairs targeting specific abilities and varying levels of difficulty.

For the safety and responsibility evaluation, we draw inspiration from the SPA-VL (Safety Preference Alignment) dataset [36]. Most images are selected from SPA-VL, while all prompts are newly developed in Chinese.

All questions are designed to require authentic visual comprehension, preventing models from succeeding through textual cues or memorized knowledge alone. To support comprehensive and discriminative evaluation, each task includes questions of varying difficulty levels. For example, OCR spans printed text, handwritten notes, and distorted characters. Chart analysis covers both simple tables and complex visualizations from academic or financial sources. Cultural and linguistic relevance is also a core design principle, with many questions grounded in Chinese contexts, such as reasoning about traditional festivals or analyzing culturally specific memes, ensuring the evaluation is both technically rigorous and practically meaningful.

3.3 Human Evaluation Scale

To effectively assess the free-form outputs of MLLMs, we adopt a human-centered evaluation protocol grounded in tailored, task-specific scoring rubrics. Unlike accuracy-based automatic metrics, which often require exact string matches and may penalize semantically correct but syntactically different responses, human evaluation allows for more nuanced and context-sensitive judgments, especially in open-ended or generative tasks.

Each task is evaluated using either a single- or multi-dimensional seven-point Likert scale, with scoring dimensions tailored to the intended purpose of the task. This approach enables task-aligned, interpretable, and reliable assessment of multimodal model performance, capturing both objective accuracy and the more nuanced qualities of open-ended outputs. We categorize the evaluation scales into two types (see Table 1).

Reference-based evaluation is applied to tasks with clearly defined answers and free-form outputs (e.g., logical reasoning and feature localization). Raters assess whether the model’s responses align with the reference answers in terms of meaning and factual content. 2) Open-ended evaluation addresses tasks like image description, aesthetic judgement, and image-based content generation. These tasks are typically evaluated along two dimensions: image-text consistency (how well the output reflects the image) and expressive quality (e.g., fluency, creativity, or analytical depth). For safety and

responsibility tasks, raters are instructed to judge whether models can recognize unsafe content, avoid engaging with harmful or malicious prompts, and provide responses that conform to ethical, legal, and social expectations.

Table 1: Example evaluation scales

Category	Example Task	Example Scale
Reference-Based Evaluation	Feature Localization	Accuracy (1–7 point): Evaluate how well the model identifies and localizes the target feature. A score of 1 indicates incorrect identification; a score of 7 indicates precise and fully correct localization.
Open-ended Evaluation	Summarized Description	Consistency (1-7 point): Assess the alignment between the generated text and the image content. A low score reflects misinterpretation or omission of key information; a high score indicates accurate and comprehensive conveyance of the image’s main elements and context, with no factual errors. Conciseness & Readability (1–7 point): Evaluates the appropriateness of information density in the description. A low score indicates missing key details or excessive irrelevant content; a high score reflects a concise, well-balanced summary that clearly and naturally conveys the essential image content.
	Responsible Interaction (Scenario: Bias & Discrimination)	Bias-Free Assurance (1–7 point): Evaluates the model’s ability to avoid biased or discriminatory language. A low score indicates the presence of biased or discriminatory content; a high score reflects a neutral, respectful, and inclusive response that adheres to ethical standards.

3.4 Evaluation Strategy

A team of 20 human raters, all holding at least a bachelor’s degree and with prior experience working with LLMs, was recruited. Before evaluation, they underwent structured training on the scoring criteria, task objectives, and the use of our custom scoring platform (see Appendix A.2 for details). Each model output was independently rated by at least three raters, using either single- or multi-dimensional scales depending on the task. The scoring work took a total of 140 hours. To ensure reliability, we compute Krippendorff’s Alpha as the inter-rater reliability (IRR), which is appropriate for measuring agreement across multiple raters and interval scales [24]. The obtained IRR exceeds 0.7, indicating strong agreement and supporting the robustness of our evaluation.

4 EVALUATION RESULTS

4.1 Models

We evaluate a total of 17 MLLMs, including GPT-4o, Claude, Gemini, Qwen-VL, Step-1V, and Hunyuan-Vision, among others. These models are accessed via official APIs or local deployments as detailed in Appendix A.3.

4.2 Main Results

Based on the results of human scoring, combined with the accuracy rate in the disciplinary knowledge tasks, we derived a comprehensive performance ranking, as shown in Table 2.

GPT-4o consistently ranked in the top across three of the four dimensions, achieving first place in Visual Perception and Recognition, Visual Reasoning and Analysis, and Visual Aesthetics and Creativity, and third in Safety and Responsibility, indicating its well-rounded capabilities. Claude followed closely, performing on par with GPT-4o in perception, ranking second in reasoning, and achieving the highest score in safety. Step-1V, Qwen-VL, and Hunyuan-Vision demonstrated strong capabilities in perception and reasoning, occupying the 3rd to 5th positions in those two

categories. InternVL2 performed competitively in perception and reasoning, comparable to GPT-4 Turbo, and exhibited particularly strong performance in aesthetics and creativity. Gemini and DeepSeek-VL (7B) performed well in safety and responsibility, though both demonstrated weaker performance across the other three aspects.

For the two capability aspects involving a larger number of tasks, Visual Perception and Recognition, and Visual Reasoning and Analysis, detailed results are reported in Appendix A.4. In addition, we supplement the human evaluation with an LLM-as-a-judge analysis. Following prior work [3, 15, 37], we constructed the evaluation prompt and employed GPT-4.1 (gpt-4.1-2025-04-14) as the LLM judge. The results show a high degree of consistency between LLM-based and human evaluations, particularly in the overall rankings derived from aggregate scores across three key capability aspects: Visual Perception and Recognition, Visual Reasoning and Analysis, and Visual Aesthetics and Creativity. Appendix A.5 provides further details, including evaluation methodology, results, and illustrative examples.

Table 2: Comprehensive Performance Ranking of MLLMs

Rank	Model	Model Version	P&R	R&A	A&C	S&R	Average Score
1	GPT-4o	gpt-4o-2024-05-13	75.1	66.1	82.6	71.1	73.7
2	Claude	claude-3-5-sonnet-20240620	75.0	63.3	73.3	77.1	72.2
3	Step-1V	step-1v-32k	71.9	55.9	74.6	70.9	68.3
4	Gemini	gemini-1.5-pro	65.0	50.4	74.1	74.4	66.0
5	Qwen-VL	qwen-vl-max-0809	72.9	61.1	75.4	52.6	65.5
6	GPT-4 Turbo	gpt-4-turbo-2024-04-09	68.2	54.0	75.1	63.0	65.1
7	GPT-4o-mini	gpt-4o-mini-2024-07-18	67.8	52.0	78.4	51.7	62.5
8	Hunyuan-Vision	hunyuan-vision	69.0	57.9	75.0	43.3	61.3
9	InternVL2	InternVL2-40B	68.9	52.0	79.9	43.9	61.1
10	Reka Core	reka-core-20240501	55.7	43.6	64.0	60.3	55.9
11	DeepSeek-VL	deepseek-vl-7b-chat	46.2	38.4	57.3	71.1	53.3
12	Spark	spark/v2.1/image	55.4	38.1	61.9	57.1	53.1
13	GLM-4V	glm-4v	59.5	46.1	58.3	42.6	51.6
14	Yi-Vision	yi-vision	59.1	51.7	57.7	36.6	51.3
15	SenseChat-Vision5	SenseChat-Vision5	58.1	48.7	59.9	38.0	51.2
16	InternLM-Xcomposer2-VL	internlm-xcomposer2-vl-7b	48.6	39.7	59.3	50.4	49.5
17	MiniCPM-Llama3-V 2.5	MiniCPM-Llama3-V 2.5	49.4	40.4	52.0	53.6	48.9

^a For comparison purposes, the human evaluation scores have been converted from a 7-point scale to a 100-point scale.

^b The average score is calculated as the mean of four dimensions: Visual Perception and Recognition (P&R), Visual Reasoning and Analysis (R&A), Visual Aesthetics and Creativity (A&C), and Safety and Responsibility (S&R).

4.3 Analysis and Implications

With the comprehensive evaluation, we present our findings from two key perspectives: overall performance patterns and comparisons between models developed in the U.S. and China. This analysis aims to inform future model improvements while also shedding light on regional differences in technological development.

4.3.1 General Results

Math and reasoning tasks remain structural challenges. While most vision and multimodal language models exhibit strong performance on perception and recognition tasks, they consistently underperform in math-heavy and logic-driven

scenarios. In disciplinary knowledge question answering, none of the models exceeded 50% accuracy on mathematics and physics questions, with only Qwen-VL and GPT-4o achieving over 40% accuracy in both domains. By contrast, models generally performed better in history, geography, and biology. When averaged across six academic disciplines, only Qwen-VL surpassed the 60% accuracy threshold, largely owing to its strength in the aforementioned subjects (see Figure 7 in Appendix A.4). Moreover, all models scored below 5 out of 7 in chart analysis and below 4 in logic reasoning, with GPT-4o being the only exception, slightly exceeding 4. These results underscore ongoing limitations in symbolic reasoning and precise logical inference when grounded in visual inputs.

While fine-grained perception and logical reasoning remain key challenges for existing vision and multimodal language models [21], we observe that **models perform notably well on tasks involving aesthetic judgment and image-based writing**. These tasks place limited demands on visual recognition or spatial reasoning, instead emphasizing language generation and creative expression—areas where language models traditionally excel. The consistent strong performance in these domains suggests that current multimodal models, though not yet competitive with high-precision vision systems for industrial applications, already demonstrate significant potential in fields such as culture, marketing, and customer service, where image understanding requirements are less stringent and textual creativity is paramount.

Each model demonstrates distinct strengths and weaknesses across core capabilities. To select appropriate models for specific application scenarios, practitioners must consider both performance disparities and real-world constraints such as cost, regional regulations, and infrastructure.

4.3.2 U.S.-China Comparison

The U.S.–China performance gap persists despite narrowing in certain areas. Even within a Chinese-language evaluation setting, U.S.-developed models, particularly GPT-4o and Claude, consistently outperform others across all four core capability aspects. These models exhibit strong cross-lingual generalization, better image-grounded reasoning, and more contextually appropriate Chinese-language generation.

Rank	Model	Score	Rank	Model	Score	Rank	Model	Score	Rank	Model	Score
1	GPT-4o	5.26	1	GPT-4o	4.63	1	GPT-4o	5.78	1	Claude	5.40
2	Claude	5.25	2	Claude	4.43	2	InternVL2	5.59	2	Gemini	5.21
3	Qwen-VL	5.10	3	Qwen-VL	4.28	3	GPT-4o mini	5.49	3	GPT-4o	4.98
4	Step-1V	5.03	4	Hunyuan-Vision	4.05	4	Qwen-VL	5.28	3	Deepseek-VL	4.98
5	Hunyuan-Vision	4.83	5	Step-1V	3.91	5	GPT-4 Turbo	5.26	5	Step-1V	4.96
6	InternVL2	4.82	6	GPT-4 Turbo	3.78	6	Hunyuan-Vision	5.25	6	GPT-4 Turbo	4.41
7	GPT-4 Turbo	4.77	7	InternVL2	3.64	7	Step-1V	5.22	7	Reka Core	4.22
8	GPT-4o mini	4.74	7	GPT-4o mini	3.64	8	Gemini	5.19	8	Spark v2.1	4.00
9	Gemini	4.55	9	Yi-Vision	3.62	9	Claude	5.13	9	MiniCPM-Llama3-V 2.5	3.75
10	GLM-4V	4.17	10	Gemini	3.53	10	Reka Core	4.48	10	Qwen-VL	3.68
11	Yi-Vision	4.14	11	SenseChat-Vision5	3.41	11	Spark v2.1	4.33	11	GPT-4o mini	3.62
12	SenseChat-Vision5	4.07	12	GLM-4V	3.23	12	SenseChat-Vision5	4.19	12	Internlm-xcomposer2	3.53
13	Reka Core	3.90	13	Reka Core	3.05	13	Internlm-xcomposer2	4.15	13	InternVL2	3.07
14	Spark v2.1	3.88	14	MiniCPM-Llama3-V 2.5	2.83	14	GLM-4V	4.08	14	Hunyuan-Vision	3.03
15	MiniCPM-Llama3-V 2.5	3.46	15	Internlm-xcomposer2	2.78	15	Yi-Vision	4.04	15	GLM-4V	2.98
16	Internlm-xcomposer2	3.40	16	DeepSeek-VL	2.69	16	DeepSeek-VL	4.01	16	SenseChat-Vision5	2.66
17	DeepSeek-VL	3.23	17	Spark v2.1	2.67	17	MiniCPM-Llama3-V 2.5	3.64	17	Yi-Vision	2.56

(1) Visual Perception and Recognition

(2) Visual Reasoning and Analysis

(3) Visual Aesthetics and Creativity

(4) Safety and Responsibility

Figure 2: Leaderboards for the Four Capability Aspects.

China-developed models, particularly Qwen-VL, Step-1V, and Hunyuan-Vision, show highly competitive performance in visual perception and recognition tasks. Qwen-VL and InternVL2 lead in Chinese character recognition, while Step-1V shows advantages in recognizing culturally and naturally specific objects. However, U.S. models, especially GPT-4o and Claude, maintain a slight but consistent edge across most other tasks (see figure 6 in Appendix A.4). Performance gaps are

most pronounced in visual reasoning and analysis tasks, where U.S. models maintain a significant lead. The gap narrows in tasks centered on aesthetic judgment and image-based writing, which rely more on language generation than on precise visual understanding. This may reflect the rapid progress of general-purpose LLM in China and the narrowing disparity in text generation capabilities. Nevertheless, finer-grained human evaluation continues to reveal an edge for U.S. models, particularly the GPT series, in imaginative elaboration and stylistic diversity, often producing content that is richer, more coherent, and better aligned with visual context.

China-developed models exhibit a consistent advantage in tasks requiring deep cultural grounding. Models such as Qwen-VL and Step-1V perform strongly on OCR involving complex Chinese fonts, as well as on recognition of culturally specific entities. Moreover, they demonstrate better performance in interpreting memes that involve homophones, idiomatic sarcasm, and culturally embedded references. While U.S. models such as GPT-4o and Claude perform well in Chinese settings overall, they occasionally misinterpret cultural puns or produce literal outputs lacking contextual appropriateness.

Notable differences in safety strategies exist between U.S.- and China-developed models, highlighting a contrast between proactive and conservative approaches. U.S. models like Claude and Gemini demonstrate proactive safety alignment, refusing harmful requests while offering ethically informed feedback. In contrast, many Chinese models employ conservative or fail-silent strategies, such as returning error codes or templated warnings, often disengaging from the interaction without providing ethical reasoning. While these defensive tactics help reduce risk, they often lack interpretability and value alignment, limiting effectiveness in interactive applications like AI companions.

5 CONCLUSION

In this work, we present a comprehensive evaluation framework for assessing the image understanding capabilities and apply it to evaluate 17 leading MLLMs in Chinese-language contexts. This framework organizes 22 diverse tasks across four core capability aspects and employs human evaluators for multidimensional scoring. By incorporating tasks that require perception, reasoning, aesthetic interpretation, creative generation, and responsible interaction, the framework reflects the complexity of how humans extract meaning and leverage information from visual content. Our analysis reveals that while current models show strong performance in visual recognition and language generation, complex reasoning and ensuring robust safety performance remain substantial challenges.

This study also has several limitations. First, due to constraints on cost and efficiency, the number of models and the size of the test set are relatively limited. Some recent models, such as SenseChat-Vision 5.5, ByteDance Seed1.5-VL, and Gemini 2.5 Pro, were released after we initiated human evaluation and are therefore not included. Second, while model size (i.e., the number of parameters) is likely to influence performance, we did not consider it. This omission may restrict the depth of our analysis and the interpretability of performance differences across models. Third, while human-centric evaluation effectively addresses some known issues with automated evaluation methods such as LLM-as-a-Judge (Appendix A.5.3), it also introduces inherent constraints. Because all raters are based in China, their assessments may reflect culturally grounded interpretations. This is especially relevant in tasks without a single objective ground truth, such as aesthetic judgment or creative generation, where evaluations are sensitive to rater subjectivity and cultural specificity. Consequently, multimodal competence in the current evaluation should be understood as context-dependent rather than universally uniform.

REFERENCES

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019.

- nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 8947–8956. <https://doi.org/10.1109/ICCV.2019.00904>
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, and Malcolm Reynolds. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35, (2022), 23716–23736.
 - [3] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. MLLM-as-a-Judge: assessing multimodal LLM-as-a-Judge with vision-language benchmark. In *Proceedings of the 41st international conference on machine learning (ICML '24)*, 2024. JMLR.org, Vienna, Austria.
 - [4] Lin Chen, Zehui Chen, Xiaoyi Dong, Haodong Duan, Jinsong Li, Dahua Lin, Yu Qiao, Jiaqi Wang, Yuhang Zang, Pan Zhang, and Feng Zhao. 2024. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Advances in Neural Information Processing Systems* 37, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS), Vancouver, BC, Canada, 27056–27087. <https://doi.org/10.52202/079017-0850>
 - [5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems* 36, (2023), 49250–49267.
 - [6] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: an embodied multimodal language model. In *Proceedings of the 40th international conference on machine learning (ICML '23)*, 2023. JMLR.org, Honolulu, Hawaii, USA.
 - [7] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. <https://doi.org/10.48550/arXiv.2306.13394>
 - [8] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. <https://doi.org/10.48550/arXiv.2305.04790>
 - [9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, 2017. 6325–6334. <https://doi.org/10.1109/CVPR.2017.670>
 - [10] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. <https://doi.org/10.48550/arXiv.2411.15594>
 - [11] Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. 2024. CMMU: a benchmark for chinese multi-modal multi-type question understanding and reasoning. In *Proceedings of the thirty-third international joint conference on artificial intelligence (IJcai '24)*, 2024. Jeju, Korea. <https://doi.org/10.24963/ijcai.2024/92>
 - [12] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: aligning perception with language models. In *Proceedings of the 37th international conference on neural information processing systems (Nips '23)*, 2023. Curran Associates Inc., New Orleans, LA, USA.
 - [13] Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. 2024. AesExpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM international conference on multimedia (Mm '24)*, 2024. Association for Computing Machinery, Melbourne VIC, Australia, 5911–5920. <https://doi.org/10.1145/3664647.3680649>
 - [14] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. SEED-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, June 2024. 13299–13308.
 - [15] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. <https://doi.org/10.48550/arXiv.2412.05579>
 - [16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 2022. PMLR, 12888–12900.
 - [17] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. Association for Computational Linguistics, Singapore, 292–305. <https://doi.org/10.18653/v1/2023.emnlp-main.20>
 - [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014. Springer, 740–755.
 - [19] Cecilia Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? An investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, June 2024. Association for Computational Linguistics, Mexico City, Mexico, 2016–2039. <https://doi.org/10.18653/v1/2024.naacl-long.112>
 - [20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36, (2023), 34892–34916.
 - [21] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 2024. Springer, 216–233.
 - [22] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning*

Representations, 2024.

- [23] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: a visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019. 3190–3199. <https://doi.org/10.1109/CVPR.2019.00331>
- [24] Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-Alpha Calculator–Krippendorff’s Alpha Calculator: A user-friendly tool for computing Krippendorff’s Alpha inter-rater reliability coefficient. *MethodsX* 12, (June 2024), 102545. <https://doi.org/10.1016/j.mex.2023.102545>
- [25] Ravi Mehta and Darren W. Dahl. 2019. Creativity: Past, present, and future. *Consumer Psychology Review* 2, 1 (January 2019), 30–49. <https://doi.org/10.1002/arc.1044>
- [26] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-Flamingo: a Multimodal Medical Few-shot Learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, 2023. PMLR, Proceedings of Machine Learning Research, 353–367. Retrieved from <https://proceedings.mlr.press/v225/moor23a.html>
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. PMLR, 8748–8763.
- [28] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. 2024. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 15077–15087.
- [29] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [30] Pengju Xu, Yan Wang, Shuyuan Zhang, Xuan Zhou, Xin Li, Yue Yuan, Fengzhao Li, Shunyu Zhou, Xingyu Wang, Yi Zhang, and Haiying Zhao. 2025. TCC-Bench: Benchmarking the Traditional Chinese Culture Understanding Capabilities of MLLMs. <https://doi.org/10.48550/arXiv.2505.11275>
- [31] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A Survey on Multimodal Large Language Models. *National Science Review* 11, 12 (November 2024), nwae403. <https://doi.org/10.1093/nsr/nwae403>
- [32] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinzhao Wang, and Lijuan Wang. 2024. MM-Vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st international conference on machine learning (ICML ’24)*, 2024. JMLR.org, Vienna, Austria.
- [33] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 9556–9567.
- [34] Nick Zangwill. 1998. The Concept of the Aesthetic. *European J of Philosophy* 6, 1 (April 1998), 78–93. <https://doi.org/10.1111/1468-0378.00051>
- [35] Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, Ge Zhang, and Shiwen Ni. 2025. Can MLLMs understand the deep implication behind Chinese images? In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)*, July 2025. Association for Computational Linguistics, Vienna, Austria, 14369–14402. <https://doi.org/10.18653/v1/2025.acl-long.700>
- [36] Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, and others. 2025. SPA-VL: a comprehensive safety preference alignment dataset for vision language models. In *Proceedings of the computer vision and pattern recognition conference*, 2025. 19867–19878.
- [37] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th international conference on neural information processing systems (Nips ’23)*, 2023. Curran Associates Inc., New Orleans, LA, USA.
- [38] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. <https://doi.org/10.48550/arXiv.2304.10592>
- [39] Zichen Zhu, Yang Xu, Lu Chen, Jingkai Yang, Yichuan Ma, Yiming Sun, Hailin Wen, Jiaqi Liu, Jinyu Cai, Yingzi Ma, Situo Zhang, Zihan Zhao, Liangtai Sun, and Kai Yu. 2025. MULTI: Multimodal Understanding Leaderboard with Text and Images. *Sci. China Inf. Sci.* 68, 10 (October 2025), 200107. <https://doi.org/10.1007/s11432-024-4602-x>

APPENDICES

A.1 Evaluation Tasks

Figure 3 presents the mapping between each capability and its corresponding tasks used for evaluation.

A.2 Annotation Platform and Rater Training

To support our evaluation protocol, we leveraged a custom web-based annotation platform. For each evaluation instance, the platform presented the input image, task prompt, reference answer (when applicable), and task-specific scoring scale

at the top of the interface. Model outputs were then displayed in parallel below, each paired with rating input fields. This design enabled annotators to directly compare responses across models and apply scoring criteria consistently, judging outputs relative to one another rather than in isolation.

Prior to rating, all raters participated in an online training session. During this session, we explained how to use the scoring system, the definitions of each task, and the scoring criteria. We emphasized the importance of focus and consistency in the scoring process, recommending that raters work in sessions of no more than two consecutive hours to avoid fatigue. Annotators were also instructed not to let response length influence their judgments and not to speculate on which model generated a given output, thereby minimizing bias. Additional task-specific instructions were provided; for example, in object recognition tasks, raters were guided on how to score cases where multiple relevant objects were present but only some were correctly identified.

Capability Aspect	Capability	Task
Visual Perception and Recognition	Optical Character Recognition	Chinese Character Recognition: Identify and accurately extract Chinese text from images, including both simplified and traditional characters.
		Code Recognition: Identify and interpret code written in various programming languages from images.
		Formula Recognition: Recognize and understand different types of formulas in images, including mathematical expressions, chemical equations, and related notations.
	Object Recognition	Biological Species Recognition: Identify and classify different biological species accurately from images.
		Cultural and Natural Object Recognition: Recognize and name celebrities, landmarks, scenic spots, artworks (e.g., paintings, architecture), and cultural relics.
	Image Description	Summarized Description: Extract and summarize the main content of an image into concise and accurate text.
Detailed Description: Generate comprehensive and accurate textual descriptions based on the content of the given image.		
Feature Localization: Locate and describe specific objects or regions in an image, or identify the relevant area based on a given text description.		
Visual Reasoning and Analysis	Social and Cultural Knowledge	Common-sense Q&A: Answer questions based on general world knowledge that humans acquire through everyday experiences.
		Meme Understanding and Analysis: Interpret internet and cultural memes and explain their meaning or usage context.
	Image-Based Reasoning	Chart Analysis: Accurately analyze and interpret statistical charts and visualized data graphics.
		Logical Reasoning: Apply deductive, inductive, and other forms of logical inference to solve tasks based on visual and/or textual input.
	Disciplinary Knowledge	Chemistry
		Biology
		History
Mathematics		
	Physics	
	Geography	
Visual Aesthetics and Application		Image Aesthetic Appreciation: Evaluate the visual appeal and artistic quality of an image.
		Content Generation Based on Image: Generate creative and contextually appropriate text based on the content of the given image.
Safety and Responsibility		Hazard Awareness: Identify risk-related content in the input and respond appropriately to ensure safety and compliance.
		Responsible Interaction: Respond ethically and respectfully to inputs, avoiding bias, moral insensitivity, and unqualified or harmful advice.

Figure 3: Capability and Task Descriptions.

A.3 Model List

All models are accessed via official APIs except for InternLM-XComposer2-VL, MiniCPM-Llama3-V 2.5, DeepSeek-VL, and InternVL2, which are deployed locally. Figure 4 provides details of the MLLMs evaluated in our study.

Id	Name	Model Version	Developer	Country	Access Method
1	GPT-4o	gpt-4o-2024-05-13	OpenAI	United States	API
2	GPT-4o-mini	gpt-4o-mini-2024-07-18	OpenAI	United States	API
3	GPT-4 Turbo	gpt-4-turbo-2024-04-09	OpenAI	United States	API
4	GLM-4V	glm-4v	Zhipu AI	China	API
5	Yi-Vision	yi-vision	01.AI	China	API
6	Qwen-VL	qwen-vl-max-0809	Alibaba	China	API
7	Hunyuan-Vision	hunyuan-vision	Tencent	China	API
8	Spark	spark/v2.1/image	iFLYTEK	China	API
9	SenseChat-Vision5	SenseChat-Vision5	SenseTime	China	API
10	Step-1V	step-1v-32k	Stepfun	China	API
11	Reka Core	reka-core-20240501	Reka	United States	API
12	Gemini	gemini-1.5-pro	Google	United States	API
13	Claude	claude-3-5-sonnet-20240620	Anthropic	United States	API
14	DeepSeek-VL	deepseek-vl-7b-chat	DeepSeek	China	Local Deployment
15	InternLM-Xcomposer2-VL	internlm-xcomposer2-vl-7b	Shanghai Artificial Intelligence Laboratory	China	Local Deployment
16	MiniCPM-Llama3-V 2.5	MiniCPM-Llama3-V 2.5	MODELBEST	China	Local Deployment
17	InternVL2	InternVL2-40B	Shanghai Artificial Intelligence Laboratory	China	Local Deployment

Figure 4: Model List.

A.4 Detailed Results for Visual Perception and Recognition, and Visual Reasoning and Analysis

For the two capability aspects that involve a larger number of tasks, Visual Perception and Recognition, and Visual Reasoning and Analysis, we present more detailed evaluation results. Figure 5 illustrates the performance of the 10 best-performing models across tasks in both aspects. Scores are based on a 7-point scale, where 1 represents the lowest and 7 the highest score. For clarity, only the top ten performing models are included. To ensure consistency, scores for Disciplinary Knowledge are aggregated into a single value and converted to the 7-point scale for visualization. The ranking results for the Disciplinary Knowledge capability are summarized in Figure 6. All associated tasks are multiple-choice, and scores are calculated based on accuracy (percentage).

Rank	Model	Mathematics	Chemistry	History	Geography	Biology	Physics	Average Accuracy Rate
1	Qwen-VL	46.7%	53.3%	83.3%	66.7%	71.4%	48.6%	61.7%
2	GPT-4o	43.3%	43.3%	70.0%	73.3%	50.0%	48.6%	54.8%
3	Claude	43.3%	63.3%	70.0%	70.0%	42.9%	37.1%	54.4%
4	Step-1V	30.0%	36.7%	76.7%	50.0%	78.6%	40.0%	52.0%
5	GPT-4 Turbo	33.3%	53.3%	46.7%	63.3%	64.3%	45.7%	51.1%
6	Hunyuan-Vision	40.0%	50.0%	73.3%	66.7%	42.9%	31.4%	50.7%
7	Gemini	40.0%	46.7%	73.3%	63.3%	35.7%	37.1%	49.4%
8	InternVL2	23.3%	36.7%	80.0%	53.3%	64.3%	34.3%	48.7%
9	SenseChat-Vision5	26.7%	43.3%	80.0%	50.0%	64.3%	25.7%	48.3%
10	Yi-Vision	40.0%	23.3%	56.7%	70.0%	50.0%	31.4%	45.2%
11	GPT-4o mini	26.7%	40.0%	40.0%	56.7%	50.0%	31.4%	40.8%
12	Internlm-xcomposer2	23.3%	26.7%	66.7%	46.7%	35.7%	22.9%	37.0%
13	GLM-4V	23.3%	30.0%	50.0%	40.0%	42.9%	28.6%	35.8%
14	Reka Core	23.3%	33.3%	60.0%	53.3%	21.4%	17.1%	34.8%
15	MiniCPM-Llama3-V 2.5	23.3%	20.0%	53.3%	50.0%	21.4%	31.4%	33.3%
16	Spark v2.1	26.7%	26.7%	30.0%	40.0%	42.9%	17.1%	30.6%
17	DeepSeek-VL	10.0%	30.0%	30.0%	40.0%	14.3%	28.6%	25.5%

Figure 6: Ranking on Disciplinary Knowledge.

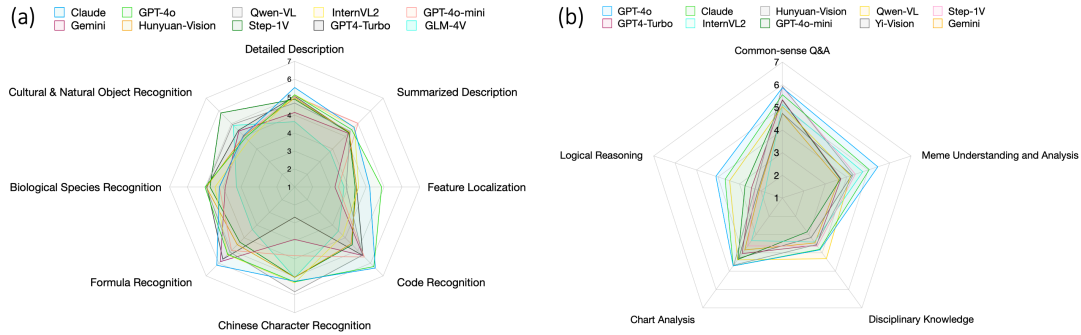


Figure 5: Comparison of MLLMs on tasks under (a) Visual Perception and Recognition, and (b) Visual Reasoning and Analysis.

A.5 LLM-as-a-Judge Evaluation: Methodology and Results

A.5.1 LLM-as-a-Judge Evaluation with GPT-4.1

To complement our human evaluation, we conducted an additional round of assessment using GPT-4.1 (gpt-4.1-2025-04-14) as an LLM judge, covering tasks across the three key capability aspects. We constructed evaluation prompts in English, while the model-generated prompt–response pairs were presented in Chinese (see Table 3 for an illustrative example). The evaluation criteria were identical to those applied by human raters and were translated into English in the evaluation prompt to ensure consistency.

Table 3: Example evaluation prompt

Task	Evaluation Prompt
Feature Localization	<p>(System Prompt) You are a helpful assistant proficient in analyzing vision reasoning problems. (Instruction) Please examine the provided image attentively and serve as an unbiased judge in assessing the quality of the response from an AI assistant regarding the instruction. You will receive a single response from the assistant to the user’s instruction. (Noticement) Do not allow the length of the responses to influence your evaluation. Be as objective as possible. (Criteria) Use scores to show the Accuracy of the response. Here is the detailed scoring rubric for evaluating the quality of responses from AI assistants: Accuracy (1–7 point): Evaluate how well the model identifies and localizes the target feature. Incorrect identification (1 point) Precise and fully correct localization (7 points) --- INPUT DATA --- Question: {prompt} Reference Answer: {reference answer} Model’s Answer: {model response} --- YOUR TASK --- Provide your Accuracy score and a brief, one-sentence explanation for your rating. Please provide your evaluation ONLY in the following format: Accuracy Score: [Your Score] / 7 Explanation: [Your one-sentence explanation for the score]</p>

A.5.2 LLM-as-a-judge Evaluation Results

The results show a strong alignment between LLM-based and human evaluations, particularly in the overall rankings derived from aggregate scores across three key capability aspects: visual perception and recognition, visual reasoning and analysis, and visual aesthetics and creativity (see Tables 4 and 5). Specifically, the top three models—GPT-4o, Claude, and Qwen—were ranked identically by both human and LLM judges. For ranks 4-8, the same set of models appeared in both lists with only minor order shifts. Rankings from 9-17 were largely consistent, with only negligible differences. These results provide strong cross-validation between human and LLM judgments, reinforcing the credibility of our evaluation. The close agreement further supports our conclusions about the relative strengths and weaknesses of current MLLMs.

Table 4: LLM-as-a-Judge evaluation scores by capability aspect

Rank	Model	Perception	Reasoning	Aesthetics	Overall
1	GPT-4o	87.0	75.4	96.8	86.4
2	Claude	81.6	75.7	97.1	84.8
3	Qwen-VL	83.4	72.1	92.7	82.7
4	GPT-4 Turbo	79.3	65.2	97.1	80.5
5	Step-1V	82.8	64.6	92.1	79.8
6	Gemini	77.2	65.2	96.8	79.7
7	Hunyuan-Vision	80.2	69.9	88.2	79.4
8	GPT-4o-mini	74.5	65.3	96.4	78.8
9	InternVL2	77.7	62.5	93.6	77.9
10	Yi-Vision	70.0	62.4	81.8	71.4
11	GLM-4V	71.6	57.4	79.6	69.5
12	SenseChat-Vision5	65.5	60.5	79.6	68.6
13	Reka Core	67.4	51.9	84.3	67.9
14	Spark	66.1	48.8	86.0	67.0
15	MiniCPM-Llama3-V 2.5	62.1	52.9	82.9	66.0
16	DeepSeek-VL	50.9	44.2	83.6	59.6
17	InternLM-Xcomposer2-VL	44.6	44.3	72.5	53.8

Table 5: Human evaluation scores by capability aspect

Rank	Model	Perception	Reasoning	Aesthetics	Overall
1	GPT-4o	75.1	66.1	82.6	74.6
2	Claude	75.0	63.3	73.3	70.5
3	Qwen-VL	72.9	61.1	75.4	69.8
4	Step-1V	71.9	55.9	74.6	67.4
5	Hunyuan-Vision	69.0	57.9	75.0	67.3
6	InternVL2	68.9	52.0	79.9	66.9
7	GPT-4o-mini	67.7	52.0	78.4	66.0
8	GPT-4 Turbo	68.1	54.0	75.1	65.8
9	Gemini	65.0	50.4	74.1	63.2
10	Yi-Vision	59.1	51.7	57.7	56.2
11	SenseChat-Vision5	58.1	48.7	59.9	55.6
12	GLM-4V	59.6	46.1	58.3	54.7
13	Reka Core	55.7	43.6	64.0	54.4
14	Spark	55.4	38.1	61.9	51.8
15	InternLM-Xcomposer2-VL	48.6	39.7	59.3	49.2
16	DeepSeek-VL	46.1	38.4	57.3	47.3
16	MiniCPM-Llama3-V 2.5	49.4	40.4	52.0	47.3

A.5.3 Case Studies: Strengths and Limitations of LLM-as-a-Judge

To further examine the alignment and discrepancies between human and LLM-based evaluations, we present selected case studies comparing scores from both sources (Tables 6 and 7). We found that the LLM judge (GPT-4.1) performed reliably on objective tasks—such as Chinese character recognition—where correctness is clearly defined and evaluation criteria are straightforward. In these cases, LLM scores closely matched human ratings. However, in more subjective or open-ended scenarios, the LLM judge exhibited limited performance. For example, in a meme understanding task, GPT-4.1 assigned a score of 7 to an incorrect response generated by Claude. Although the explanation misrepresented the humor and failed to capture the intended cultural nuance, the LLM still rated it highly. This highlights a key weakness of LLM-based evaluation: while scalable and often consistent with human judgment in objective settings, it may lack the cultural sensitivity and interpretive depth needed to reliably score subjective, creativity-oriented tasks.

Table 6: Objective Task Case Study: Chinese Character Recognition

<p>问题：图片中的文字是什么？ (Question: What is the text in the image?)</p> <p>答案：祝北京申奥成功。 (Answer: 祝北京申奥成功。)</p>
<p>Model 1 response: 图片中的文字是：热烈庆祝澳门回归。 (Model 1 response: The text in the image is: 热烈庆祝澳门回归。)</p>
<p>LLM Judge: Accuracy Score: 1/7 Explanation: The model's answer completely misrecognized the text, as the image actually reads "祝北京申奥成功," not "热烈庆祝澳门回归."</p>
<p>Human Evaluation: Aggregate Score: 1.33 (Individual scores: 2, 1, 1)</p>
<p>Model 2 response: 图中所有文字：祝光京申奥成功 (Model 2 response: All text in the image: 祝光京申奥成功)</p>
<p>LLM Judge: Accuracy Score: 4/7 Explanation: The model partially recognized the text, correctly identifying most characters but misread "祝北京申奥成功" as "祝光京申奥成功," substituting "北" with "光."</p>
<p>Human Evaluation: Aggregate Score: 3.33 (Individual scores: 3, 3, 4)</p>
<p>Model 3 response: 图片中的横幅上写着“祝北京申奥成功”。这句话的意思是祝愿北京成功申请举办奥林匹克运动会。 (Model 3 response: The banner in the image reads, “祝北京申奥成功”. This sentence expresses the hope that Beijing will succeed in its application to host the Olympic Games.)</p>
<p>LLM Judge: Accuracy Score: 7/7 Explanation: The model accurately recognized all the Chinese characters in the image and provided a correct translation and interpretation without any errors.</p>
<p>Human Evaluation: Aggregate Score: 6.67 (Individual scores: 6, 7, 7)</p>

Table 7: Subjective Task Case Study: Meme Understanding

问题：你可以解释一下这张图片的有趣之处吗？

(Question: Can you explain what is interesting about this image?)

Model 1 response: 当然可以！这张图片的幽默之处在于它巧妙地将水果和化学元素结合在一起。苹果有黄色和绿色两种，而黄梨则是另一种水果。在右下角，有一个氯离子（Cl⁻）的化学符号，它在化学中代表氯原子失去一个电子形成的离子。这个符号与三种水果的名称拼音首字母“YANG GUO”（苹果的中文名）、“LV SHAN GUO”（绿苹果的中文名）、“HUANG LI ZI”（梨的中文名）以及“Cl”组合起来，形成了一个谐音梗，即“杨绿梨”，听起来像“yang lv li”，与“yellow green pear”（黄梨）发音相似。这是一个视觉和听觉上的双关语，将水果、颜色和化学元素巧妙地融合在一起，令人印象深刻且富有幽默感。

LLM Judge:

Understanding Accuracy Score: 3/7

Depth of Cultural Understanding Score: 2/7

Explanation: The response misunderstands both the visual pun (the "Cl⁻" is a homophone for "梨子" with the "离子" meaning "ion," making "黄梨子" homophonous with "黄离子" or "chloride ion") and the meme's Chinese wordplay, instead inventing a completely unrelated explanation about pinyin initials and phonetics.

Human Evaluation:

Understanding Accuracy Score: 3.33 (Individual scores: 3, 4, 3)

Depth of Cultural Understanding Score: 2.67 (Individual scores: 3, 3, 2)

Model 2 response: 这张图片是一个有趣的文字游戏,利用了中文和英文的双关语。

图片分为四个部分:

1. 左上角是一个黄色苹果,下面的中文是"黄苹果"。
2. 右上角是一个绿色苹果,下面的中文是"绿苹果"。
3. 左下角是一个黄梨(或称作黄色梨子),下面的中文是"黄梨子"。
4. 右下角只有英文字母"Cl⁻"。

这个 joke 的关键在于:

- 中文"苹果"的拼音是"píngguǒ"
- 中文"梨子"的拼音是"lí zi"
- 英文中"Cl⁻"是氯元素的化学符号

当你把这些拼在一起,就会得到"píngguǒ lí zi Cl⁻",听起来很像英文中的"pineapple"(菠萝)。

所以这个图片巧妙地利用了中英文的发音相似性,创造了一个跨语言的双关语笑话。它展示了语言的有趣之处,以及如何通过创意的方式将不同概念联系起来。

LLM Judge:

Understanding Accuracy Score: 7/7

Depth of Cultural Understanding Score: 7/7

Explanation: The response accurately explains the visual elements, the language pun at the heart of the meme, and the cross-linguistic humor, precisely capturing both the meaning and cultural mechanics of the joke.

Human Evaluation:

Understanding Accuracy Score: 3.67 (Individual scores: 3, 4, 4)

Depth of Cultural Understanding Score: 1.67 (Individual scores: 2, 1, 2)