Image Understanding in Chinese Contexts: A Human-Centric Approach to Assess MLLMs from the US and China

Anonymous ACL submission

Abstract

The rapid rise of multimodal large language models (MLLMs) has created a pressing need for systematic evaluations of their performance. Most existing benchmarks are designed for English-language settings and rely heavily on automated scoring, leaving a significant gap in evaluating complex multimodal tasks in Chinese and culturally grounded scenarios. To address this, we introduce a comprehensive evaluation framework and a curated dataset for Chinese-language image understanding. Our framework encompasses four core capability aspects: visual perception and recognition, visual reasoning and analysis, visual aesthetics and creativity, and safety and responsibility. All image-text pairs are carefully constructed to ensure strong visual grounding. We benchmark 17 state-of-the-art MLLMs from the U.S. and China across 22 diverse tasks using a human-centric evaluation approach, supported by a multidimensional scoring protocol. Our findings show that GPT-4o and Claude lead across the four capability aspects, while models like Qwen-VL and Step-1V demonstrate particular strengths in visual perception tasks, especially in culturally specific scenarios. Additionally, we provide comparative insights into the strengths and limitations of U.S.- and China-developed models, offering guidance for more informed development and deployment of multimodal AI systems.

1 Introduction

Recent multimodal large language models (MLLMs) have demonstrated remarkable progress in understanding and reasoning across visual and textual modalities (OpenAI et al., 2024a; Alayrac et al., 2022; Huang et al., 2023; Li et al., 2023b; Driess et al., 2023; Dai et al., 2023; Gong et al., 2023; Liang et al., 2024), 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 (He et al., 2024; Li et al., 2023a; Fu et al., 2023; Liu et al., 2023b; Xu et al., 2023), 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 (Chen et al., 2024). Additionally, the growing reliance on large language models (LLMs) as automatic evaluators may introduce biases or fail to capture nuanced aspects of multimodal outputs (Gu et al., 2025). 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 issues, we propose a comprehensive evaluation framework focused on MLLMs' image understanding capabilities in Chinese-language contexts. We apply this framework to benchmark 17 leading models released in China and the United States 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 human expert raters and adopt a multidimensional scoring protocol. 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.

- We introduce a structured evaluation framework and a curated question set for assessing image understanding capabilities in Chinese-language contexts. 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.

- We design a multidimensional scoring protocol that incorporates expert human evaluations, 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.

- We assess 17 state-of-the-art MLLMs from China and the U.S, providing a comparative analysis that uncovers model strengths, weaknesses, and regional performance trends. By offering transparent evaluation and culturally grounded tasks, our work aims to advance more inclusive, meaningful, and real-world-relevant assessment of MLLMs.

2 Related Work

Multimodal Language Models. The success of large language models (LLMs) has spurred their adaptation to multimodal tasks through integration with visual encoders, leading to multimodal large language models (MLLMs). (Yin et al., 2024). Early approaches such as CLIP (Radford et al., 2021) focused on aligning vision and language through contrastive learning on large-scale image-text pairs , while subsequent models such as BLIP (Li et al., 2022) introduced diverse supervision tasks such as captioning to improve multimodal pretraining. However, both required separate vision-language pipelines, incurring error accumulation.

Recent MLLMs have evolved from modular pipelines toward unified architectures that embed visual features directly into language modeling. Leading adapter-based approaches like MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023), and LLaVA (Liu et al., 2023a) align pre-trained LLMs with visual inputs through lightweight projection and instruction tuning, enabling advanced capabilities in multi-turn visual question answering and image-based dialogue. Meanwhile, natively multimodal architectures, such as GPT-4o (OpenAI et al., 2024b) and Gemini 1.5 (Gemini et al., 2024), process vision and language in integrated architectures, supporting real-time, end-to-end image-text interaction.

As MLLMs are applied to various real-world scenarios, such as document digitization, autonomous driving (Wei et al., 2024), and medical image analysis (Moor et al., 2023), their accuracy, robustness, and contextual reasoning become increasingly critical. This underscores the urgent need for reliable evaluation and benchmarking (Huang and Zhang, 2024).

Evaluations of MLLMs. Quantitative evaluation is essential to assess the strengths and limitations of MLLM. Classical benchmarks such as COCO Captions (Lin et al., 2015), NoCaps (Agrawal et al., 2019), and VQAv2 (Goyal et al., 2017) focus on isolated tasks like image captioning or visual question answering. These typically involve fixed answer formats and narrow linguistic distributions, making them insufficient for evaluating general-purpose, open-ended multimodal understanding and reasoning.

Recent MLLM evaluation efforts have shifted from narrow, task-specific assessments to more comprehensive and integrated benchmarks that span a wide range of capabilities. For instance, MME (Fu et al., 2023) and MMBench (Liu et al., 2023b) 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-2 (Li et al., 2023a) introduces a hierarchical framework that integrates recognition and generation tasks, supported by a refined answer-ranking strategy. In contrast, MM-Vet (Yu et al., 2024) 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 evaluate open-ended responses. While recent benchmarks represent significant progress toward evaluating MLLMs, most remain constrained by English-centric design, limited task diversity, over-reliance on LLM-based auto-evaluation, and insufficient control over data contamination or visual grounding (Chen et al., 2024).

Figure 1: Diagram of our evaluation framework. Sample questions are translated from Chinese.

Our work aligns with this trend toward integrated multimodal evaluation and builds upon the capabilities emphasized in prior benchmarks such as MM-Vet (Yu et al., 2024) and MME (Fu et al., 2023). Yet, our work distinguishes itself in several important ways. First, our benchmark is designed specifically for the Chinese-language context, addressing the linguistic and cultural limitations of existing English-centric benchmarks. Second, we enforce strong visual grounding and data novelty by carefully designing new image-text pairs that cannot be answered using textual priors or general knowledge alone. This ensures 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 incorporate expert human raters and adopt a multidimensional scoring protocol to ensure fairness 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 represent a progression from basic to advanced skills and include both technical and ethical considerations. This structure mirrors the layered competencies required for real-world, commercial applications and socially aligned deployment of MLLMs. Figure 1 summarizes the evaluation tasks and shows illustrative

examples, and Appendix A provides detailed task definitions.

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. Failures in this foundational aspect can lead to hallucinations, where models fabricate elements not present in the image (Li et al., 2023c), thereby undermining reasoning or generation. For this aspect, we assessed 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 reasoning skills and external knowledge. Tasks used for assessment include answering questions involving social knowledge, interpreting culturally embedded memes, analyzing visual data like graphs or charts, and solving other complex reasoning problems. 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 through tasks that require more than factual knowledge. Specifically, we assess the model's ability to judge the aesthetic quality of an image (e.g., composition, lighting, color) (Huang et al., 2024) 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 (Zangwill, 1998). Similarly, creativity depends on recombining loosely connected concepts (Mehta and Dahl, 2019), as seen in metaphor or storytelling. By evaluating aesthetic judgment and image-based creative writing, we can test whether a model can perceive, interpret, and transform 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 serves as essential constraints on model behavior in real-world applications. It ensures that the model operates in a trustworthy, socially responsible manner and adheres to legal and ethical standards. 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.

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 (Chen et al., 2024). 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 (Lu et al., 2023). 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 exam questions are newly released, the likelihood of their inclusion in 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 (Yue et al., 2024).

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): Evaluate 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 content.
	Responsible Interaction (Scenario: Bias & Discrimination)	Bias-Free Assurance (1–7 point): Evaluate 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 aligns with ethical and social standards.

Table 1: Example evaluation scales.

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 (Marino et al., 2019), 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 (Agrawal et al., 2019) are reused, but all accompanying questions 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 can only be answered accurately through genuine visual understanding. 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 (Zhang et al., 2024). Most images are selected from SPA-VL, while all prompts are newly developed in Chinese.

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-centric 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. These scales allow for task-aligned, interpretable, and reliable assessment of multimodal model performance, capturing both objective accuracy and the more subtle qualities of open-ended model outputs. We categorize the evaluation scales into two types (see Table 1): 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.

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 B 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 calculate inter-rater reliability (IRR) for each task. Across all tasks, the IRR value exceeds 0.7, indicating strong agreement among raters and validating the robustness of the scoring process.

4 Evaluation Results

4.1 Models

We evaluate a total of 17 MLLMs, including GPT-4o (OpenAI et al., 2024b), Claude, Gemini (Gemini et al., 2024), Qwen-VL (Bai et al., 2023), Step-1V, Hunyuan-Vision, and Deepseek-VL (Lu et al., 2024) among others. These models are accessed via official APIs or local deployments as detailed in Appendix C.

4.2 Main Results

Based on the results of human scoring, combined with the accuracy rate in the disciplinary knowledge tasks, we derive a comprehensive performance ranking, as shown in Table 2. More detailed evaluation results can be found in Appendix D.

GPT-4o consistently ranks at the top across three of the four evaluation aspects, securing first place in Visual Perception and Recognition, Visual Reasoning and Analysis, and Visual Aesthetics and Creativity, and placing third in Safety and Responsibility, indicating its well-rounded capabilities. Claude follows 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 show strong capabilities in perception and reasoning, occupying the 3rd to 5th positions in those two categories. InternVL2 performs competitively in perception and reasoning, comparable to GPT-4 Turbo, and exhibits particularly strong performance in aesthetics and creativity. Gemini and DeepSeek-VL (7B) perform well in safety and responsibility, though both show weaker performance across the other three aspects.

Rank	Model	P&R	R&A	A&C	S&R	Ave.
1	GPT-4o	75.1	66.1	82.6	71.1	73.7
2	Claude	75.0	63.3	73.3	77.1	72.2
3	Step-1V	71.9	55.9	74.6	70.9	68.3
4	Gemini	65.0	50.4	74.1	74.4	66.0
5	Qwen-VL	72.9	61.1	75.4	52.6	65.5
6	GPT-4 Turbo	68.2	54.0	75.1	63.0	65.1
7	GPT-4o-mini	67.8	52.0	78.4	51.7	62.5
8	Hunyuan-Vision	69.0	57.9	75.0	43.3	61.3
9	InternVL2	68.9	52.0	79.9	43.9	61.1
10	Reka Core	55.7	43.6	64.0	60.3	55.9
11	DeepSeek-VL	46.2	38.4	57.3	71.1	53.3
12	Spark	55.4	38.1	61.9	57.1	53.1
13	GLM-4V	59.5	46.1	58.3	42.6	51.6
14	Yi-Vision	59.1	51.7	57.7	36.6	51.3
15	SenseChat-Vision5	58.1	48.7	59.9	38.0	51.2
16	InternLM-Xcomposer2-VL	48.6	39.7	59.3	50.4	49.5
17	MiniCPM-Llama3-V 2.5	49.4	40.4	52.0	53.6	48.9

Table 2: Comprehensive Performance Ranking of MLLMs. For comparison purposes, the human evaluation scores have been converted from a 7-point scale to a 100-point scale. The Average Accuracy Rate (Avg.) is the mean across four capability aspects: 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 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 exceeds 50% accuracy on mathematics

Rank	Model	Score
1	GPT-4o	5.26
2	Claude	5.25
3	Qwen-VL	5.10
4	Step-1V	5.03
5	Hunyuan-Vision	4.83
6	InternVL2	4.82
7	GPT-4 Turbo	4.77
8	GPT-4o mini	4.74
9	Gemini	4.55
10	GLM-4V	4.17
11	Yi-Vision	4.14
12	SenseChat-Vision5	4.07
13	Reka Core	3.90
14	Spark v2.1	3.88
15	MiniCPM-Llama3-V 2.5	3.46
16	Internlm-xcomposer2	3.40
17	DeepSeek-VL	3.23

(1) Visual Perception and Recognition

Rank	Model	Score
1	GPT-4o	4.63
2	Claude	4.43
3	Qwen-VL	4.28
4	Hunyuan-Vision	4.05
5	Step-1V	3.91
6	GPT-4 Turbo	3.78
7	InternVL2	3.64
7	GPT-4o mini	3.64
9	Yi-Vision	3.62
10	Gemini	3.53
11	SenseChat-Vision5	3.41
12	GLM-4V	3.23
13	Reka Core	3.05
14	MiniCPM-Llama3-V 2.5	2.83
15	Internlm-xcomposer2	2.78
16	DeepSeek-VL	2.69
17	Spark v2.1	2.67

(2) Visual Reasoning and Analysis

Rank	Model	Score
1	GPT-4o	5.78
2	InternVL2	5.59
3	GPT-4o mini	5.49
4	Qwen-VL	5.28
5	GPT-4 Turbo	5.26
6	Hunyuan-Vision	5.25
7	Step-1V	5.22
8	Gemini	5.19
9	Claude	5.13
10	Reka Core	4.48
11	Spark v2.1	4.33
12	SenseChat-Vision5	4.19
13	Internlm-xcomposer2	4.15
14	GLM-4V	4.08
15	Yi-Vision	4.04
16	DeepSeek-VL	4.01
17	MiniCPM-Llama3-V 2.5	3.64

(3) Visual Aesthetics and Creativity

Rank	Model	Score
1	Claude	5.40
2	Gemini	5.21
3	GPT-4o	4.98
3	Deepseek-VL	4.98
5	Step-1V	4.96
6	GPT-4 Turbo	4.41
7	Reka Core	4.22
8	Spark v2.1	4.00
9	MiniCPM-Llama3-V 2.5	3.75
10	Qwen-VL	3.68
11	GPT-4o mini	3.62
12	Internlm-xcomposer2	3.53
13	InternVL2	3.07
14	Hunyuan-Vision	3.03
15	GLM-4V	2.98
16	SenseChat-Vision5	2.66
17	Yi-Vision	2.56

(4) Safety and Responsibility

Figure 2: Leaderboards for the Four Capability Aspects. Scores are based on a 7-point scale, where 1 indicates the lowest and 7 the highest rating. For consistency, the score for Disciplinary Knowledge in the Visual Reasoning and Analysis aspect is converted from an accuracy percentage to a 7-point scale.

and physics questions, with only Qwen-VL and GPT-4o achieving over 40% accuracy in both domains. In contrast, models generally perform better in history, geography, and biology. When averaged across six academic disciplines, only Qwen-VL surpasses the 60% accuracy threshold, largely owing to its strength in the aforementioned subjects (see Appendix D). Moreover, all models score 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.

Models perform notably well on tasks involving aesthetic judgment and image-based writing. These tasks place limited demands on fine-grained 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.

Our findings emphasize the importance of **task-specific model selection in real-world applications.** As no single model consistently excels across all capability dimensions—perception, reasoning, aesthetics, and safety—practitioners should carefully align their model choice with the demands of the intended use case, while also account for real-world constraints such as cost, regional regulations, and infrastructure. For example, GPT-4o demonstrates strong overall performance, making it a reliable option for general-purpose deployment. Qwen-VL shows particular strengths in culturally nuanced visual perception and domain-specific knowledge. InternVL2 not only performs competitively in these areas but also offers notable advantages for on-premise deployment, making it especially suitable for use cases that require local operation without relying on cloud-based access. These insights underscore the need for a targeted evaluation and selection strategy that takes into account both technical performance and contextual fit, rather than relying solely on aggregate benchmark rankings.

4.3.2 U.S.-China Comparison

The U.S.–China performance gap persists despite narrowing in certain areas. Even within a Chinese-language testing context, U.S. models (e.g., GPT-4o, Claude) consistently lead across core capability aspects such as perception, reasoning, and creativity. These models exhibit strong cross-lingual generalization, better image-grounded reasoning, and contextual grounded generation.

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 3 in Appendix D). Performance gaps are most pronounced in visual reasoning and analytical 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 in Chinese-language contexts and apply it to evaluate 17 leading MLLMs. This human-centric assessment organizes 22 diverse tasks across four core capability aspects, uniquely employing expert human raters for multidimensional scoring. Our comparative analysis reveals that while models show strong performance in perception and language generation, complex reasoning and ensuring robust safety performance remain significant hurdles for current MLLMs.

Limitations

This study has several limitations. First, due to constraints on cost and efficiency, the number of models and test instructions included in the evaluation is relatively limited. In addition, several recent model versions—such as ByteDance Seed1.5-VL (Guo et al., 2025), SenseChat-Vision 6, and Gemini 2.5 Pro—were released after the initiation of our human scoring process and were therefore not incorporated into the benchmark. 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.

References

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)*, pages 8947–8956. ArXiv:1812.08658 [cs].

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a Visual Language Model for Few-Shot Learning.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint*. ArXiv:2308.12966 [cs].

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint*. ArXiv:2305.06500 [cs].

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, and 3 others. 2023. PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint*. ArXiv:2303.03378 [cs].

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. *arXiv preprint*. ArXiv:2306.13394 [cs].

Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*. ArXiv:2403.05530 [cs].

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. *arXiv preprint*. ArXiv:2305.04790 [cs].

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. *arXiv preprint*. ArXiv:1612.00837 [cs].

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. *arXiv preprint*. ArXiv:2411.15594 [cs].

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, and 178 others. 2025. Seed1.5-VL Technical Report. *arXiv preprint*. ArXiv:2505.07062 [cs].

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. *arXiv preprint*. ArXiv:2401.14011 [cs].

Jiaxing Huang and Jingyi Zhang. 2024. A Survey on Evaluation of Multimodal Large Language Models. *arXiv preprint*. ArXiv:2408.15769 [cs].

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. *arXiv preprint*. ArXiv:2302.14045 [cs].

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. 2024. AesBench: An Expert Benchmark for Multimodal Large Language Models on Image Aesthetics Perception. *arXiv preprint*. ArXiv:2401.08276 [cs].

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023a. SEED-Bench-2: Benchmarking Multimodal Large Language Models. *arXiv preprint*. ArXiv:2311.17092 [cs].

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint*. ArXiv:2301.12597 [cs].

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint*. ArXiv:2201.12086 [cs].

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A Survey of Multimodel Large Language Models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, Xi' an China. ACM.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. *arXiv preprint*. ArXiv:1405.0312 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual Instruction Tuning. *arXiv preprint*. ArXiv:2304.08485 [cs].

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint*. ArXiv:2307.06281 [cs].

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv preprint*. ArXiv:2403.05525 [cs].

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. *arXiv preprint*. ArXiv:1906.00067 [cs].

Ravi Mehta and Darren W Dahl. 2019. Creativity: Past, present, and future. *Consumer Psychology Review*, 2(1):30–49.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023. Med-Flamingo: a Multimodal Medical Few-shot Learner. *arXiv preprint*. ArXiv:2307.15189 [cs].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024a. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024b. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint*. ArXiv:2103.00020 [cs].

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. *arXiv preprint*. ArXiv:2402.05746 [cs].

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. *arXiv preprint*. ArXiv:2306.09265 [cs].

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):nwae403. ArXiv:2306.13549 [cs].

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. *arXiv preprint*. ArXiv:2308.02490 [cs].

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 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*, pages 9556–9567.

Nick Zangwill. 1998. The Concept of the Aesthetic. *European Journal of Philosophy*, 6(1):78–93.

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, and 1 others. 2024. Spa-vl: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint*. ArXiv:2304.10592 [cs].

A Evaluation Tasks

We illustrate how each core capability is assessed by mapping it to the specific tasks used in our evaluation framework, as detailed in Figure 5.

B Design of the Scoring Platform

To support our evaluation protocol, we leveraged a custom web-based annotation platform. For each evaluation instance, the platform displays the input image, task prompt, reference answer (when applicable), and task-specific scoring guidelines at the top of the interface. Below this, the outputs from all evaluated models are shown in parallel, each accompanied by rating input fields. This layout enables raters to directly compare model outputs and apply the scoring criteria more consistently.

C 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 6 provides details of the MLLMs evaluated in our study.

D More Experimental Results

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. Model performance on tasks under these two aspects is illustrated in Figures 3 and 4, respectively. Scores are based on a 7-point scale, where 1 represents the lowest and 7 the highest rating. For clarity, only the top ten performing models are included.

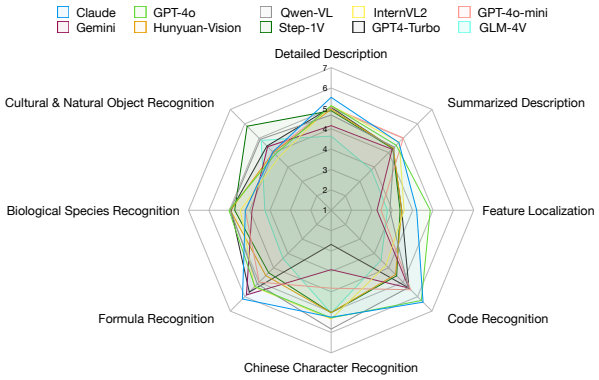

Figure 3: Comparison of 10 advanced MLLMs on 8 tasks under Visual Perception and Recognition. Each subtask is evaluated on a scale from 1 to 7, where 1 represents the lowest and 7 the highest rating.

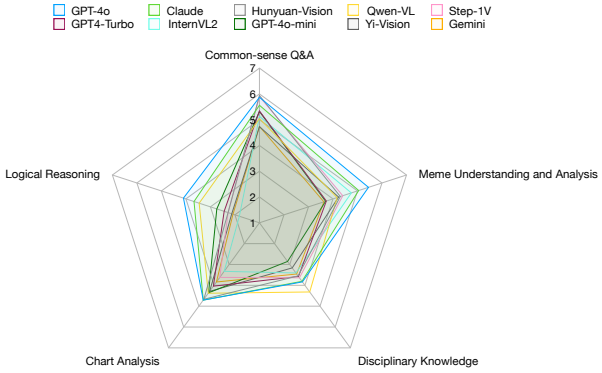

Figure 4: Comparison of 10 advanced MLLMs on tasks under Visual Reasoning and Analysis. The maximum possible score for each subtask is 7. The scores for Disciplinary Knowledge are aggregated into a single value and converted to the 7-point scale for visualization.

The ranking results for the Disciplinary Knowledge capability are summarized in Table 3. All associated tasks are multiple-choice, and scores are calculated based on accuracy (percentage).

Rank	Model	MA.	CH.	HI.	GE.	BI.	PH.	Avg.
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

Table 3: Rankings on Disciplinary Knowledge. All scores represent accuracy percentages. The Average Accuracy Rate (Avg.) is the mean across all listed subjects. Abbreviations: MA. (Mathematics), CH. (Chemistry), HI. (History), GE. (Geography), BI. (Biology), PH. (Physics).

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 5: Capability and Task Descriptions.

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 6: Model List.