OIG-BENCH: A MULTI-AGENT ANNOTATED BENCHMARK FOR MULTIMODAL ONE-IMAGE GUIDES UNDERSTANDING

Jiancong Xie[1,2], **Wenjin Wang**[2], **Zhuomeng Zhang**[2], **Zihan Liu**[2], **Qi Liu**[2], **Ke Feng**[2], **Zixun Sun**[2], **Yuedong Yang**[1,3,*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Interactive Entertainment Group, Tencent Inc, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing (MOE), China
[*]Corresponding author, yangyd25@mail.sysu.edu.cn.

ABSTRACT

Recent advances in Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities. However, evaluating their capacity for human-like understanding in **One-Image Guides** remains insufficiently explored. One-Image Guides are a visual format combining text, imagery, and symbols to present reorganized and structured information for easier comprehension, which are specifically designed for human viewing and inherently embody the characteristics of human perception and understanding. Here, we present **OIG-Bench**, a comprehensive benchmark focused on One-Image Guide understanding across diverse domains. To reduce the cost of manual annotation, we developed a semi-automated annotation pipeline in which multiple intelligent agents collaborate to generate preliminary image descriptions, assisting humans in constructing image–text pairs. With OIG-Bench, we have conducted comprehensive evaluation of 29 state-of-the-art MLLMs, including both proprietary and open-source models. The results show that Qwen2.5-VL-72B performs the best among the evaluated models, with an overall accuracy of 77%. Nevertheless, all models exhibit notable weaknesses in semantic understanding and logical reasoning, indicating that current MLLMs still struggle to accurately interpret complex visual-text relationships. In addition, we also demonstrate that the proposed multi-agent annotation system outperforms all MLLMs in image captioning, highlighting its potential as both a high-quality image description generator and a valuable tool for future dataset construction. Datasets are available at https://github.com/XiejcSYSU/OIG-Bench.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have recently achieved groundbreaking progress Li et al. (2023); Liu et al. (2023); Zhu et al. (2023); Wang et al. (2024), demonstrating the ability to process and integrate information from multiple modalities, including text, images, and audio. These models have demonstrated impressive performance in tasks such as image captioning Bucciarelli et al. (2024) and visual question answering (VQA) Liu et al. (2024b). With the continuous improvement of model capabilities, how to scientifically and comprehensively evaluate the multimodal understanding and reasoning ability of MLLM has become a key issue to promote further.

One of the ultimate objectives of MLLMs is to demonstrate human-like perceptual and cognitive abilities. Nevertheless, evaluating the human-likeness of MLLMs remains a significant challenge. To assess such capabilities, existing MLLM evaluation benchmarks often focus on complex scenarios, particularly those involving visually complex and information-rich images, such as infographics. Existing benchmarks for infographic understanding have made valuable contributions Mathew et al. (2022); Masry et al. (2022); Xia et al. (2024); Li et al. (2024) by targeting structured visual data, including charts and tables. However, these datasets typically feature regular layouts and fixed el-

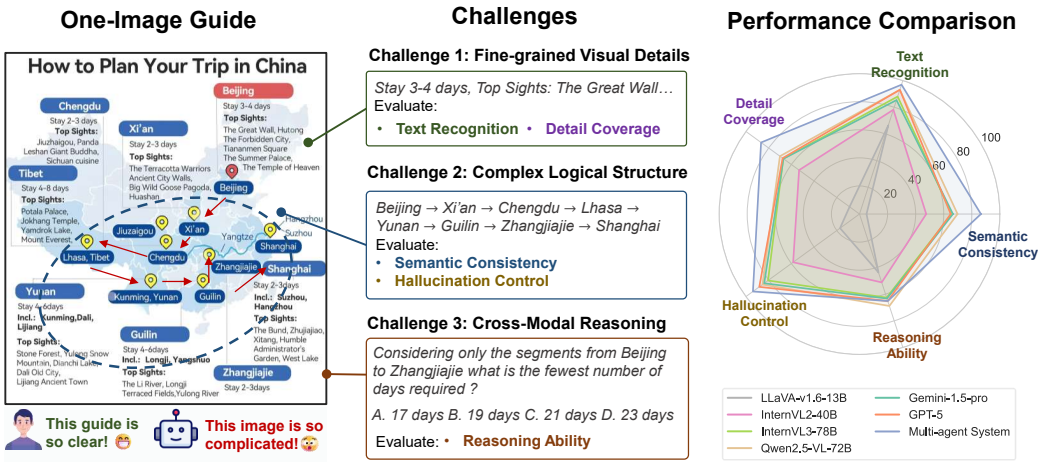


Figure 1: Left: An example of a One-Image Guide. Middle: Three major challenges for MLLMs in understanding One-Image Guides. Right: Results of six representative MLLMs and the proposed multi-agent annotation system. The left-skewed distribution indicates that MLLMs still perform poorly in logical reasoning and semantic understanding. In contrast, the multi-agent annotation system achieves the best overall performance.

ement positions, making them insufficient to fully assess a model's comprehensive reasoning and semantic understanding ability at a human-like level.

To bridge this gap, we focus on a visual information format that inherently embodies the characteristics of human-like perception and understanding: **One-Image Guide**. A One-Image Guide is a single image that presents rich task-relevant information in a compact, visually engaging format. (see Fig. 1 Left). It leverages human cognitive advantages such as parallel visual processing and pattern recognition, enabling rapid comprehension of the overall structure and efficient extraction of key information. While this format is highly accessible to humans, it poses significant challenges for MLLMs due to its fine-grained visual details, complex logical structure, and cross-modal reasoning (see Fig. 1 Middle). To achieve a complete and accurate understanding of one-image guides, MLLMs must possess fine-grained visual perception (e.g., text and details recognition) and spatial structure parsing capabilities (e.g., interpreting layouts, arrows, and segmented regions). These abilities must be effectively integrated to form coherent and contextually grounded semantic interpretations. As the One-Image Guide is optimized for human cognitive processing, the ability to comprehend it with human-like ease can serve as evidence of human-like visual understanding.

Here, we introduce OIG-Bench, a comprehensive benchmark for evaluating MLLMs on the understanding of one-image guides. OIG-Bench is a bilingual dataset containing 808 images spanning multiple domains. To construct this benchmark efficiently, we innovatively propose a semi-automated annotation pipeline based on a multi-agent collaboration framework. In this framework, multiple intelligent agents collaborate to generate detailed descriptions of one-image guides, which are then refined through human verification. This approach substantially reduces manual annotation effort while ensuring high-quality labels. Meanwhile, we also demonstrate that such multi-agent annotation system serves as an effective approach for high-quality image description generation (see Fig. 1 Right). Based on these annotated images, we further conduct extensive evaluations, encompassing 29 prominent MLLMs, including GPT-5, Qwen2.5-VL, and InternVL3. We design two evaluation tasks, including image description generation and VQA, to assess five key capabilities. Our evaluation shows that Qwen2.5-VL-72B achieves the best overall performance, but all models exhibit notable weaknesses in semantic consistency and reasoning. Through this evaluation, we not only benchmarked the current MLLMs but also provided key insights to drive the model to achieve more human-like perception and cognitive abilities in complex, information-intensive multimodal scenarios.

2 RELATED WORKS

2.1 MULTIMODAL LARGE LANGUAGE MODELS

Multimodal Large Language Models (MLLMs) have rapidly advanced at the intersection of computer vision and natural language processing. Early systems used separate encoders and shallow fusion, exemplified by CLIPRadford et al. (2021). With the rise of large language models (LLMs) such as GPT-3Brown et al. (2020), visual encoders were integrated with generative language backbones to create general-purpose MLLMs. A typical architecture combines a pretrained vision encoder (e.g., CLIP-ViT, Swin TransformerLiu et al. (2021)), a modality alignment module (e.g., linear projection, Q-FormerLi et al. (2023)), and an LLM (e.g., LLaMATouvron et al. (2023), OPTZhang et al. (2022), GPT Brown et al. (2020)) for reasoning and generation.

Several influential MLLMs have been proposed in recent years. BLIP-2 Li et al. (2023) introduced a lightweight Q-Former to bridge frozen vision encoders and frozen LLMs, achieving strong performance with minimal training cost. MiniGPT-4 Zhu et al. (2023) demonstrated that aligning CLIP visual features with Vicuna's Team (2023) language space enables rich image-grounded conversations. LLaVA Liu et al. (2023) further improved instruction-following capabilities by fine-tuning on large-scale image–text instruction datasets. Commercial systems such as GPT-4o and Gemini have extended these capabilities to more modalities and complex reasoning tasks. Despite the rapid progress of MLLMs, systematic evaluation of their performance on text-rich, information-dense images remains insufficient, leaving a gap in understanding their true capabilities in such challenging scenarios.

2.2 MLLMS BENCHMARKS

The rapid growth of Multimodal Large Language Models (MLLMs) has led to numerous benchmarks for evaluating vision–language capabilities. Early evaluation datasets, such as VQAv2 Yang et al. (2022), COCO Caption Chen et al. (2015), and Flickr30k Entities Plummer et al. (2015), primarily targeted specific tasks like visual question answering, image captioning, and referring expression comprehension. With the emergence of instruction-following MLLMs, more comprehensive evaluation suites have been proposed to assess perception, reasoning, and knowledge-based abilities in a unified framework. For example, MMBench Liu et al. (2024b) evaluates models through carefully designed multiple-choice questions covering diverse multimodal skills, while MME Fu et al. (2024) provides large-scale, fine-grained assessments across perception, cognition, and reasoning dimensions.

While these benchmarks cover a broad spectrum of multimodal tasks, they may not fully capture the challenges posed by text-rich, information-dense images, which require accurate OCR, spatial layout understanding, and fine-grained multimodal reasoning. To address this, several specialized benchmarks have been developed. TextVQA Singh et al. (2019) focuses on natural scene images containing embedded text, evaluating models' abilities to read and reason over textual content. DocVQA Mathew et al. (2021) targets document images such as forms and invoices, emphasizing structured layout parsing. ChartQA Masry et al. (2022) assesses reasoning over charts and plots, requiring numerical understanding and visual–textual integration. Seed-Bench-2-PLUS further expands the evaluation scope to include text-rich images, charts, and diagrams. In the infographic domain, InfographicVQA Mathew et al. (2022) contains infographic-style images that combine visual elements, icons, and explanatory text, challenging models to understand across heterogeneous visual and textual components. In addition, reasoning-oriented benchmarks such as M^3CoT Chen et al. (2024a) and CoMT Cheng et al. (2025) are designed to evaluate MLLMs' ability to perform multi-hop reasoning, also posing a challenge to existing MLLMs

3 OIG-BENCH

The entire data construction process is illustrated in Figure 2. We adopt a semi-automated approach to collect, filter, and annotate the dataset. In this section, we detail the semi-automated data processing pipeline, as well as the evaluation tasks and data analysis of our benchmark.

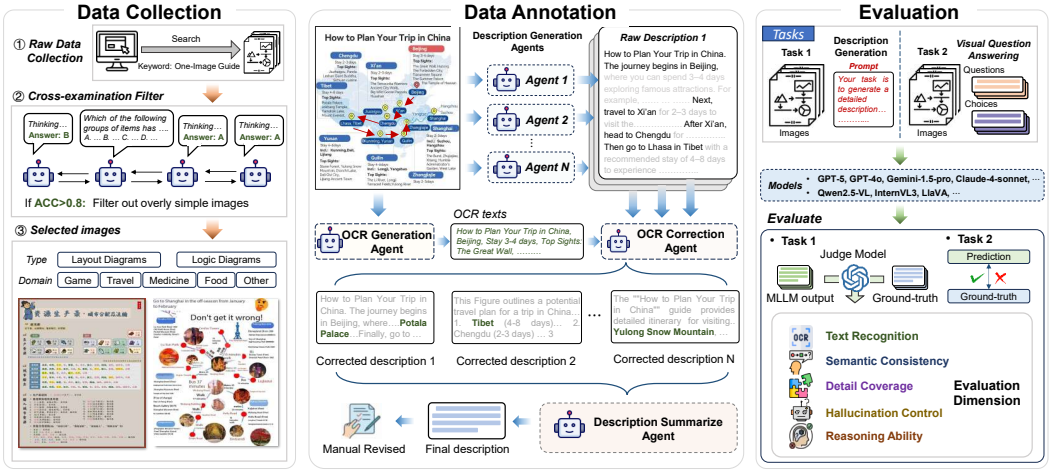


Figure 2: The One-Image Guide Data Processing Pipeline. We collect One-Image Guide images from the Internet and filter them using a cross-examination process with multiple MLLMs, where one model poses questions and others answer. Images with accuracy above 0.8 are considered too simple and removed. Next, we annotate the images using a multi-agent system, followed by manual verification. In the evaluation stage, we assess MLLMs on two tasks, namely description generation and VQA, covering five key capabilities.

3.1 Semi-Automated Data Processing Pipeline

3.1.1 Data Curation Process

Data Collection: We collected One-Image Guides from multiple online platforms, including Red Note, Baidu, and Google, anonymizing all real names to ensure privacy. We categorize the collected images into two types: **layout diagrams** that present ordered or ranked information through tables, columns, or lists (e.g., game character tier rankings, recommended travel destination lists), and **logic diagrams** that illustrate processes, causal relationships, or conceptual connections (e.g., game character relationship maps, travel route maps). Examples can be found in Figure 6 in Appendix A.1.

Data Filtration: To ensure that the dataset remained both high-quality and sufficiently challenging, we employed a filtering process assisted by MLLMs. We adopted a mutual cross-examination strategy in which a subset of MLLMs generated questions for each image while the remaining models attempted to answer them. If the average accuracy exceeded 80%, the image was deemed too simple and removed. A final manual review was conducted to eliminate any remaining low-quality samples, resulting in a curated set of 808 images for subsequent annotation and evaluation. More details can be found in Appendix A.2.

3.1.2 Multi-agent-based Description Generation

Manually creating accurate and information-rich textual descriptions for each image is not only time-consuming and labor-intensive, but also prone to inconsistencies in style and detail across annotators. To address these challenges, we design a multi-agent collaboration framework that integrates the complementary capabilities of multiple specialized agents.

Description Generation Agent: Given an input image, the process begins with several state-of-the-art multimodal large language models (MLLMs) acting as description generation agents. Each of these agents independently analyzes the visual content and produces an initial description, capturing salient objects, attributes, and relationships from its own perspective. Specifically, we employ GPT-4o, Claude-4-Sonnet, Gemini-1.5-Pro, and Qwen2.5-VL-78B as the description generation agents. This diversity in initial outputs helps mitigate the bias or omission that may occur when relying on a single model.

OCR Generation Agent: In parallel, the image is processed by an OCR generation agent, which is responsible for detecting and extracting any textual content embedded within the image, such as labels, annotations, or captions. This agent employs an OCR-specific model, PaddleOCR, rather than a large language model, as specialized OCR systems are generally more accurate and efficient for text recognition tasks.

OCR Correction Agent: Once both the initial descriptions and OCR results are obtained, an OCR correction agent refines each initial description by cross-referencing it with the extracted text. This process corrects transcription errors, fills in missing textual details, and ensures that all in-image text is faithfully represented in the description. We implement this agent using GPT-4.1, leveraging its strong language understanding and reasoning capabilities to integrate OCR outputs with visual descriptions effectively.

Description Summarize Agent: Finally, a description summarize agent aggregates all the corrected descriptions into a single, coherent, and comprehensive final description. This agent consolidates overlapping information, resolves inconsistencies across different sources, and ensures that the final output is logically structured and semantically complete. We implement this agent using GPT-4.1. By integrating the strengths of visual understanding, text recognition, and summarization, the multi-agent pipeline produces descriptions that are both accurate and information-rich, providing a robust foundation for subsequent annotation and evaluation tasks.

This multi-agent-based description generation system can serve as an effective tool for producing high-quality image descriptions by leveraging the complementary strengths of multiple specialized agents. Finally, all automatically generated descriptions are manually reviewed and corrected by human annotators, and the refined results are used as the final ground-truth annotations for the images. The prompts for each agent are provided in the Appendix A.6.

3.2 Evaluation Tasks

To comprehensively assess the model's capabilities, we employ two evaluation tasks: Description Generation and Visual Question Answering (VQA).

(1) Description Generation. Given an image, the model is required to produce a coherent and semantically accurate textual description. The evaluation focuses on four core capabilities:

- **Text Recognition:** Measures the correctness of transcribed textual content from the visual input, reflecting the model's OCR performance.
- **Semantic Consistency:** Evaluates the alignment between the generated description and the ground truth in terms of factual correctness, logical coherence, and temporal or numerical accuracy.
- **Detail Coverage:** Assesses whether the generated description captures all relevant and salient details present in the image, including objects, attributes, and contextual information.
- **Hallucination Control:** Identifies instances where the model introduces content not supported by the visual input, indicating over-generation or fabrication.

For each aspect, we employ GPT-4.1 to compare the model-generated description with the human-verified ground truth and assign a score from 0 to 1 on a six-level rating scale (0.0, 0.2, 0.4, 0.6, 0.8, 1.0). To validate the reliability of GPT-4.1 scoring, we additionally conducted human evaluations on several representative models. As shown in Table 5 and Figure 8 in Appendix A.4, the GPT-4.1 scores exhibit strong consistency with human scores, indicating the robustness and accuracy of the automated evaluation.

(2) VQA: In the VQA task, the model is required to answer natural language questions based on the given visual input. This task is specifically designed to evaluate the model's **Reasoning Ability**. To construct the question set, we leverage the ground-truth descriptions in our dataset and employ GPT-4.1 to automatically generate 3-5 questions for each description. The questions are intentionally constructed to require multi-step reasoning, such as combining multiple pieces of information, performing temporal or numerical inference, or integrating contextual cues. Model performance is evaluated in terms of accuracy.

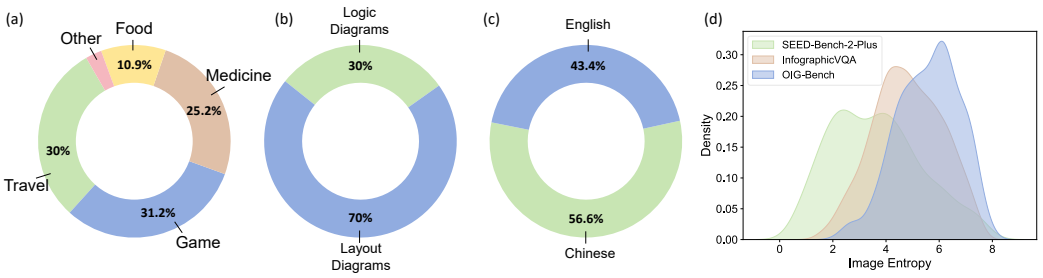


Figure 3: Statistics of OIG-Bench. (a) Domain distribution: Game (48.8%), Travel (28.7%), Food (9.1%), Medicine (8.9%), and Other. (b) Diagram type distribution: Layout Diagrams (70.5%) and Logic Diagrams (29.5%). (c) Kernel density estimation of image entropy, comparing OIG-Bench with SEED-Bench-2-Plus and InfographicVQA.

3.3 Dataset Analysis

Our proposed OIG-Bench contains a total of 808 one-image guide images accompanied by 2,800 questions. The dataset spans multiple domains, including game, travel, medicine, and food, with the game and travel domains being the most represented, accounting for 31.2% and 30% of all images (Fig. 3(a)), respectively. Each question is provided with multiple-choice options. The average question length is approximately 67.5 words, while the average option length is around 74.9 words. In our benchmark, the OIG images can be broadly classified into two categories, with logic diagrams accounting for 30% and layout diagrams for 70% of the dataset (Fig. 3(b)). Furthermore, the dataset is bilingual, with Chinese images accounting for 56.6% and English images for 43.4% (Fig. 3(c)).

To quantitatively assess the visual complexity of our dataset, we compute the image entropy for each image. Image entropy measures the amount of information contained in an image and is defined as:

$$H = -\sum_{i=0}^{L-1} p_i log_2 p_i, \tag{1}$$

where L denotes the number of possible pixel intensity levels, and p_i represents the probability of intensity level i in the image. Higher entropy values indicate greater variability and complexity in pixel distribution, which typically correspond to richer visual content.

We compare the image entropy of OIG-Bench with SEED-Bench-2-Plus and InfographicVQA. As shown in Figure 3(d), OIG-Bench has a higher average entropy (5.8) than InfographicVQA (4.8) and SEED-Bench-2-Plus (3.2). This substantial difference suggests that OIG-Bench images are more visually complex and information-dense, thereby posing greater challenges for multimodal large language models.

4 Experiments

4.1 Experiment Setups

We evaluate a total of 29 models, including 5 leading proprietary (closed-source) models and 24 SOTA open-source models. The proprietary models include GPT5, GPT-4o, Claude-4-sonnet, Claude-3-7-sonnet, and Gemini-1.5-pro. The open-source models include InternVL3.5 Wang et al. (2025), InternVL3 Zhu et al. (2025), InterVL2.5 Chen et al. (2024b), InterVL2 Chen et al. (2024c), Qwen2.5-VL Bai et al. (2025), Qwen2-VL Wang et al. (2024), Yi-VL Young et al. (2024), LLaVA-v1.6 Liu et al. (2024a), and LLaVA-v1.5 Liu et al. (2024a), covering various model scales from 6B to 78B. To ensure fairness and comparability, all models are prompted using the same instruction template and set with a temperature of 0. The maximum completion token length is fixed at 2048. For subsequent few-shot experiments, we reserve two images from each domain, with the remaining images used exclusively for evaluation. The initial descriptions from the multi-agent annotation sys-

Table 1: Performance comparison of proprietary and open-source MLLMs on OIG-Bench. The overall score (%) is calculated as the mean of the scores (%) across five dimensions. The best and second-best results among evaluated MLLMs are highlighted in bold and underlined, respectively.

Model	Scale	Overall	Description Generation				VQA
			Semantic Consistency	Text Recognition	Detail Coverage	Hallucination Control	Reasoning Ability
Open-Source MLLMs							
LLaVA-1.5	7B	30.63	6.10	71.69	8.49	26.19	38.67
	13B	29.68	4.11	76.93	7.45	13.95	43.98
LLaVA-1.6	7B	22.31	2.79	53.11	4.78	11.44	39.41
	13B	28.21	5.00	66.40	8.74	17.12	43.78
Yi-VL	6B	34.89	20.42	75.35	10.54	18.49	47.64
	34B	36.82	24.46	76.79	9.68	19.55	49.61
InternVL2	26B	56.21	46.71	75.50	52.75	58.78	50.31
	40B	57.29	47.52	78.11	53.20	58.42	51.19
InternVL2.5	8B	60.42	53.65	84.60	51.03	61.15	53.67
	26B	63.03	54.96	85.30	57.70	63.06	56.14
	38B	65.21	55.14	87.61	57.60	66.98	58.74
	78B	68.44	59.11	87.90	57.53	73.85	63.79
InternVL3	9B	63.44	54.37	83.29	58.69	68.74	54.12
	14B	66.23	57.70	84.40	61.04	73.59	56.41
	38B	68.82	58.60	86.76	66.53	74.55	57.64
	78B	72.38	<u>65.69</u>	<u>87.99</u>	67.13	80.54	62.57
InternVL3.5	8B	64.22	54.37	83.29	58.69	68.74	55.99
	14B	69.68	60.00	83.79	67.07	74.90	62.65
	38B	70.13	60.95	84.29	66.30	75.62	63.49
Qwen2-VL	7B	68.31	56.89	91.04	57.97	80.32	55.31
	72B	73.39	64.11	90.97	64.15	83.57	62.14
Qwen2.5-VL	7B	71.10	59.73	91.08	60.68	82.39	61.64
	32B	<u>76.18</u>	<u>69.84</u>	92.25	**70.74**	80.32	<u>67.77</u>
	72B	**77.56**	**70.48**	<u>93.02</u>	<u>70.07</u>	85.42	**68.83**
Proprietary MLLMs							
Gemini-1.5-pro	–	73.61	66.54	85.31	66.98	83.93	63.30
Claude-3-7-sonnet	–	65.54	58.69	75.35	63.79	68.67	63.22
Claude-4-sonnet	–	67.36	61.99	76.20	65.93	75.52	59.14
GPT-4o	–	71.93	63.77	77.61	65.38	<u>87.95</u>	64.93
GPT-5	–	75.41	64.96	**93.05**	67.70	**88.24**	65.13
Multi-agent System	–	86.24	86.49	96.85	86.67	93.83	67.34

tem are included in the evaluation, but as it cannot perform VQA directly, we combine its generated description with the question and use GPT-4.1 to produce the answer for fair comparison.

4.2 Overall Performance of Open- and Closed-source Models

Table 1 summarizes the results for the description generation and VQA tasks. Overall, closed-source models generally perform better than open-source models, particularly in hallucination control. In contrast, open-source models exhibit a larger variance in performance across different tasks and model scales. Nevertheless, several open-source models are able to match or even surpass closed-source models. In particular, Qwen2.5-VL-72B achieves the highest overall score among all evaluated models, outperforming the best closed-source model, GPT-5, by 2.6%. Qwen2.5-VL-72B also excels in key metrics such as semantic consistency and reasoning ability, highlighting its advanced capability in interpreting complex visual content.

Across all models, we observe that current MLLMs demonstrate strong performance in text recognition. However, their performance in semantic consistency, which is crucial for ensuring that generated descriptions accurately reflect the visual content, remains considerably weaker. As model

Table 2: Inference time of single-model and multi-agent settings on OIG-Bench.

Model	Agents	Inference time per image (s)	Overall score
Qwen2.5-VL-72B	–	26.34	77.56
GPT-4o	–	5.13	71.93
Multi-agent	GPT-4o, Claude-4, Gemini-1.5-pro, Qwen2.5-VL-72B	66.39	**86.24**
Multi-agent	GPT-4o, Claude-4, Gemini-1.5-pro	38.34	<u>82.49</u>

scale increases, for example, in the Qwen family from 7B to 72B, we see consistent improvements across five capabilities, with the largest gains in semantic consistency and hallucination control. Nevertheless, several metrics, particularly text recognition and detail coverage, tend to plateau as model size grows, likely due to the fixed capacity of the vision encoder, which limits the extraction of fine-grained visual information. For instance, InternVL2.5-26B, InternVL2.5-38B, and InternVL2.5-78B all employ a 6B-parameter ViT, yet their performance in text recognition and detail coverage shows little difference.

4.3 Performance of the Multi-Agent Annotation System

The proposed multi-agent annotation system yields the strongest description-generation results in our benchmark. It achieves 86.4% in semantic consistency, 96.8% in text recognition, 86.6% in detail coverage, and 93.8% in hallucination control, exceeding any single model baseline on each metric.

The gains can be attributed to the complementary strengths of multiple specialized agents, which can reduce hallucinations while improving coverage of fine-grained details in text-rich scenes. Notably, the multi-agent system involves invoking several large-scale model agents (GPT-4o, Claude-4, Gemini-1.5-pro, and Qwen2.5-VL-72B), inevitably introducing higher latency and computational costs. In particular, Qwen2.5-VL-72B that we deploy locally delivers the best performance but also incurs the longest inference time, averaging 26 seconds per image as shown in Table 2. In contrast, GPT-4o, accessed via its official API, requires only 5 seconds. The complete multi-agent annotation system takes 66 seconds per image, which is considerably longer than that of a single-model setup. To mitigate this, we removed Qwen2.5-VL-72B from the description generation agent, reducing the inference time while maintaining an overall score that still surpasses any single-model baseline. This highlights an inherent trade-off between performance and efficiency in multi-agent designs.

4.4 Performance across Different Image Domains and Types

Figure 4 presents the performance of the evaluated models across different image domains and types. The results indicate that model scores vary substantially across domains: models generally achieve higher scores in the medical and food domains, while consistently performing worse in the gaming and travel domains. This disparity may be attributed to the open-world, natural, and icon-rich scenes that are typical of gaming and travel images, which pose greater challenges for accurate visual understanding and description generation. When comparing image types, we observe that layout diagrams are generally easier for the models than logic diagrams. This suggests that tasks involving spatial arrangement and structural recognition are more manageable for current MLLMs, whereas those requiring complex reasoning remain more challenging.

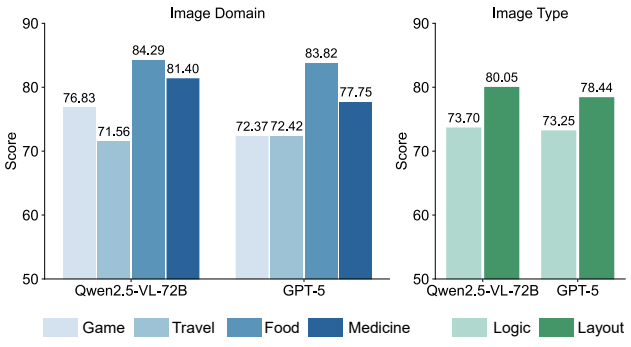


Figure 4: Performance comparison of two representative MLLMs across different image domains and types.

Table 3: Overall results of different prompts on OIG-Bench.

Model	Prompt	Overall	Description Generation				VQA
			Semantic Consistency	Text Recognition Accuracy	Detail Coverage	Hallucination Detection	Reasoning Ability
GPT-4o	Base	71.93	63.77	77.61	65.38	87.95	64.93
	CoT	71.26 (-0.67)	63.59 (-0.18)	76.14 (-1.47)	65.17 (-0.21)	84.31 (-3.64)	67.11 (+2.18)
	1-shot	71.28 (-0.65)	63.12 (-0.65)	77.49 (-0.12)	64.54 (-0.84)	86.14 (-1.81)	65.13 (+0.20)
	2-shot	71.40 (-0.53)	63.45 (-0.32)	77.31 (-0.30)	64.37 (-1.01)	86.37 (-1.58)	65.49 (+0.56)
Qwen2.5-VL-72B	Base	77.56	70.48	93.02	70.07	85.42	68.83
	CoT	77.15 (-0.41)	69.17 (-1.31)	93.12 (+0.10)	69.73 (-0.34)	84.02 (-1.40)	69.71 (+0.88)
	1-shot	77.00 (-0.56)	69.10 (-1.38)	93.06 (+0.04)	69.01 (-1.06)	84.82 (-0.60)	68.97 (+0.14)
	2-shot	76.57 (-0.99)	68.45 (-2.03)	92.25 (-0.77)	68.39 (-1.68)	84.62 (-0.80)	69.12 (+0.29)

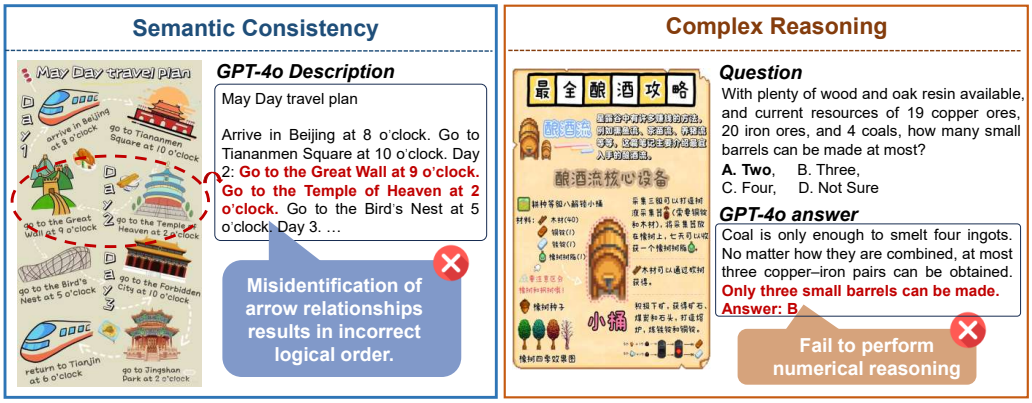


Figure 5: Case analysis of model errors.

4.5 Analysis on Different Prompt Skills

Table 3 shows the impact of Base, CoT, and few-shot prompts on image description generation and VQA performance. We observe that CoT and few-shot prompts yield small gains in VQA reasoning, yet they reduce description quality. The largest declines occur in hallucination control and detail coverage. One possible explanation is that CoT prompts tend to elicit more exploratory outputs, which can be beneficial for arithmetic or aggregation questions in VQA but may harm the quality of free-form descriptions by introducing additional hallucinated content.

4.6 Error Analysis

This subsection provides a case study for analyzing two major weaknesses of current models: semantic consistency and complex reasoning. One of the most common errors arises from incorrect identification of arrow relationships, which leads to faulty logical sequencing. This issue is particularly pronounced when the two entities connected by an arrow are spatially close in the image. As illustrated in Figure 5, GPT-4 reverses the visiting order between the Great Wall and the Temple of Heaven. Another frequent error involves the inability to handle complex reasoning tasks. The model demonstrates a pronounced weakness in multi-step inference, particularly in scenarios involving arithmetic reasoning. Overall, these observations not only highlight the current limitations of the models but also inform potential improvements.

5 Discussion

In this work, we present OIG-Bench, the first benchmark specifically designed for evaluating One-Image Guide understanding in MLLMs. We propose a semi-automated benchmark construction method and systematically evaluate 29 MLLMs. Experimental results show that current MLLMs still exhibit limitations in semantic understanding and logical reasoning. We hope OIG-Bench to provide a challenging, realistic platform that drives MLLMs toward human-level understanding of complex visual information.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Davide Bucciarelli, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Personalizing multimodal large language models for image captioning: an experimental analysis. In *European Conference on Computer Vision*, pp. 351–368. Springer, 2024.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M ^{3} cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*, 2024a.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024c.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23678–23686, 2025.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Vicuna Team. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. *Vicuna: An open-source chatbot impressing gpt-4 with*, 90, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, et al. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*, 2024.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 3081–3089, 2022.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Appendix

A.1 Details of OIG-Bench

OIG-Bench is a bilingual evaluation dataset of One-Image Guides, containing both English and Chinese samples. It primarily covers four domains: travel, gaming, food, and medical. Figure 6 presents representative examples from OIG-Bench,

Figure 6: OIG-Bench examples sampled from each domain.

A.2 Details of Data Filtering

We first collected 2,891 images from the internet using "one-image guide" as the primary search keyword. In the first stage, we performed a coarse manual screening to remove images that did not conform to the definition of a One-Image Guide, such as those lacking a clear integration of text, imagery, and symbols. This step resulted in 1,982 remaining images.

In the second stage, we applied an automated cross-validation system involving four advanced MLLMs: GPT-4o, Gemini-1.5-Pro, Claude-4, and Qwen2.5-VL-72B. The process was conducted in a round-robin manner. In each round, one model generated a multiple-choice question based on the content of a One-Image Guide, and the remaining models answered it. This procedure continued until each model had served as the question generator once. Images with an overall answering accuracy greater than 80% across all rounds were considered overly simple and thus excluded from the benchmark. This filtering step reduced the dataset to 1,023 images.

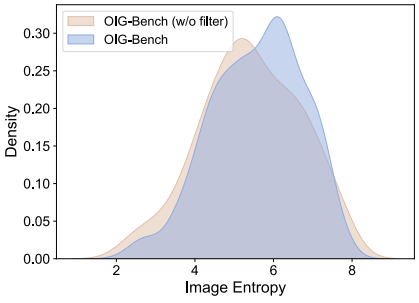


Figure 7: Kernel density estimation of image entropy of OIG-Bench before and after cross-validation filtering.

Finally, we conducted a fine-grained manual inspection to remove images with excessively low resolution or overly simple content. After this final filtering stage, 808 high-quality One-Image Guide images were retained for the benchmark. To validate the effective-

ness of the proposed cross-validation system, we analyzed and compared the image entropy distributions before and after this filtering stage. As shown in Figure 7, the post-filtering entropy distribution exhibits a clear rightward shift relative to the pre-filtering distribution, indicating that the filtering process successfully removed overly simple images while retaining those with richer informational content.

A.3 Details of Evaluated MLLMs

Table 4 provides an overview of the open-source MLLMs, highlighting differences in their architectures and parameters.

Table 4: The architecture and size of different models.

Model	Scale	Vision Part	Language Part
LLaVA-1.5	7B	CLIP ViT-L/14@336px	Vicuna-7B
	13B	CLIP ViT-L/14@336px	Vicuna-13B
LLaVA-1.6	7B	CLIP ViT-L/14@336px	Vicuna-7B
	13B	CLIP ViT-L/14@336px	Vicuna-13B
Yi-VL	6B	CLIP VIT-H/14	Yi-6B-Chat
	34B	CLIP VIT-H/14	Yi-34B-Chat
Qwen2-VL	7B	DFN(ViT-625M)	Qwen2-7B
	72B	DFN(ViT-625M)	Qwen2-78B
Qwen2.5-VL	7B	-	Qwen2.5-7B
	32B	-	Qwen2.5-32B
	72B	-	Qwen2.5-72B
InternVL2	26B	InternViT-6B-448px-V1-5	Internlm2-chat-20b
	40B	InternViT-6B-448px-V1-5	Nous-Hermes-2-Yi-34B
InternVL2.5	8B	InternViT-300M-448px-V2_5	Internlm2_5-7b-chat
	26B	InternViT-6B-448px-V2_5	Internlm2_5-20b-chat
	38B	InternViT-6B-448px-V2_5	Qwen2.5-32B-Instruct
	78B	InternViT-6B-448px-V2_5	Qwen2.5-72B-Instruct
InternVL3	9B	InternViT-300M-448px-V2_5	Internlm3-8b-instruct
	14B	InternViT-300M-448px-V2_5	Qwen2.5-14B
	38B	InternViT-6B-448px-V2_5	Qwen2.5-32B
	78B	InternViT-6B-448px-V2_5	Qwen2.5-72B
InternVL3.5	8B	InternViT-300M-448px-V2_5	Qwen3-8B
	14B	InternViT-300M-448px-V2_5	Qwen3-14B
	38B	InternViT-6B-448px-V2_5	Qwen3-32B

A.4 Analysis of Automated Evaluation

In this study, we employed GPT-4.1 to automatically evaluate the accuracy of descriptions generated by MLLMs across five assessment dimensions, using human-annotated descriptions as references. To verify the reliability of the automatic scoring, we additionally conducted human evaluations on two representative models: GPT-5 and Qwen2.5-VL-72B. Each image was independently scored by three human annotators, and the average score was used for comparison. As shown in Table 5, the human scores and the GPT-4.1 automatic scores exhibit strong alignment. Figure 8 further illustrates the comparison of overall scores for each image, where the Pearson correlation coefficients between human and GPT-4.1 scores are 0.93 for Qwen2.5-VL-72B and 0.94 for GPT-5, confirming that GPT-4.1 can serve as a reliable proxy for evaluation in this benchmark.

Table 5: Comparison between GPT-4.1 automatic scoring and human scoring.

Model	Score	Semantic Consistency	Text Recognition Accuracy	Detail Coverage	Hallucination Detection
Qwen2.5-VL-72B	GPT-4.1	70.48	93.02	70.07	85.42
Qwen2.5-VL-72B	Huamn	69.37	95.21	71.69	86.97
GPT-5	GPT-4.1	64.96	93.05	67.70	88.24
GPT-5	Human	62.15	94.68	69.13	88.37

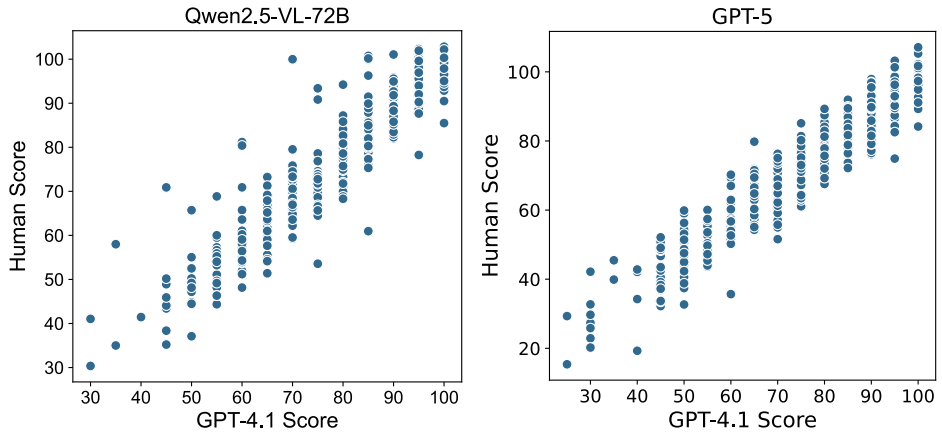


Figure 8: Correlation between GPT-4.1 and human overall scores.

A.5 Performance Across Different Image Pixels

Figure 9 (left) presents the pixel distribution of images in OIG-Bench. We divided all images into three categories: low-resolution images with a total pixel count below 1.5M, medium-resolution images between 1.5M and 2.5M, and high-resolution images above 2.5M. Among these, medium-resolution images are the most common, accounting for 60.2% of the dataset. Figure 9 (right) shows the performance of Qwen2.5-VL-72B across the three resolution categories. We observe that as resolution increases, the model's scores on semantic consistency and detail coverage decrease noticeably, suggesting that higher-resolution images may contain more complex visual information and finer details, which increase the difficulty of accurate description generation. In contrast, text recognition and hallucination control do not exhibit a performance drop with increasing resolution, indicating that these aspects are less sensitive to image resolution within the tested range.

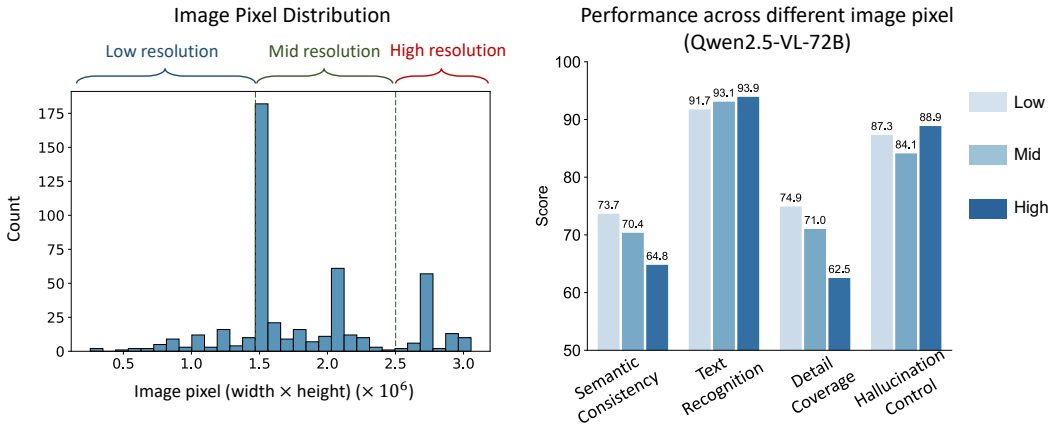


Figure 9: Left: The image pixel distribution of OIG-Bench. Right: Performance of Qwen2.5-VL-72B across different image resolution.

A.6 PROMPT TEMPLATES

Prompt to generate description of One-Image Guide (Description Generation Agent).

User: <Image>. Your task is to generate a detailed description for a 'One-Image Guide'. Clearly convey the content of the guide by following these guidelines:

1. Carefully read and understand all the textual information in the guide image, ensuring you grasp every step and key point.

2. Present each step in a detailed and well-organized manner, following the logical sequence of the guide.

3. Use clear and easy-to-understand language. Only describe the content shown in the image; do not include any information that is not mentioned in the image, and do not provide additional introductions.

4. Ensure the description is complete and contains all necessary information.

Prompt for OCR correction of generated descriptions (OCR Correction Agent).

User: Your task is to revise the large language model's description of a 'One-Image Guide' based on the OCR model's extracted text.

Below is the large language model's output description:

<LLM Description>

{*description*}

</LLM Description>

Below is the OCR model's output:

<OCR Output>

{*OCR_text*}

</OCR Output>

During the calibration process, compare the LLM's description with the OCR text, and use the context from the OCR text to identify any parts of the LLM's description, such as locations, people, or objects that are inconsistent with the OCR text. Modify only the inconsistent parts

in the LLM's description. Be careful not to change the logical structure of the LLM's output description.

First, briefly explain your analysis and reasoning process, then output the revised content. The output format should be as follows:

<Thought>

Provide a detailed explanation of your reasoning process, showing how you compared and integrated the description content to arrive at the correct description.

</Thought>

<Corrected Result>

Write the corrected content here.

</Calibration Result>

Prompt to summarize multiple corrected descriptions (Description Summarization Agent).

User: Your task is to analyze and compare multiple descriptions from large language models for a 'One-Image Guide' image, and summarize the correct description. Below are the descriptions provided by multiple large language models, each starting with 'Description i:'.

<Description Collection>

{*description*}

</Description Collection>

Some of these descriptions are correct, while others contain errors. Please follow the steps below:

1. Compare and synthesize the multiple descriptions, identifying similarities and differences in their content.

2. Content that is consistent across most descriptions can be considered correct and should be retained.

3. If a model's output clearly deviates from the others, it can be considered incorrect.

4. First, briefly explain your analysis and reasoning process, pointing out the correct and incorrect content in each description, and then output a complete, correct description.

The output format should be as follows:

<Thought>

Provide a detailed explanation of your reasoning process, showing how you compared the descriptions to arrive at the correct one.

</Thought>

<Description>

Write the complete, correct description here.

</Description>

Prompt to generate visual question and answer pairs.

User: Your task is to create 3-5 high-difficulty reasoning-based single-choice questions based on the provided <Reference Text>.

Requirements:

1. The questions must be based on the content of the <Reference Text>, but cannot directly copy the original text. They should require reasoning, summarization, or multi-step logical analysis to arrive at the answer.

2. Each question must have 4 options (A, B, C, D), with only one correct option, and you must provide the correct answer.

3. Do not directly give the answer in the question, and do not include obvious hints in the options.

4. The question, options, and answer should be separated by — on the same line in the following format:

Question content— A. xxx, B. xxx, C. xxx, D. xxx — Correct answer

Each question should occupy one line.

Here is the Reference Text:

<Reference Text>: {*description*}

Prompt to answer the question with a given image.

User: <Image>. You are a visual question answering system. You are given an image, a question, and several options. Carefully observe the content of the image and reason out the option that best matches the facts.

Note:

1. Base your answer only on the information in the image and the question; do not introduce external knowledge.

2. There is only one correct option.

3. Only output the answer (A, B, C, or D); do not output anything else, including your reasoning process.

<Question>

{*question*}

</Question>

<Options>

{*choice*}

</Options>

Prompt to evaluate the generated description.

User: You are an "One-Image Guide" image description analysis expert. You will be given two inputs:

Ground Truth: The correct "one-image guide" description manually extracted from the image.

Model Prediction: The description generated by the model for the same image.

Your task:

1. Carefully compare the Ground Truth and the Model Prediction.

2. Identify the errors in the Model Prediction and classify them into the following four categories (note: each error can only be assigned to the single most appropriate category, and must not be counted in multiple categories):

- Text Recognition Accuracy: Errors in recognizing text from the image, such as OCR mistakes, spelling errors, etc. Any spelling difference counts as an error, regardless of whether it affects overall understanding.

- Detail Coverage: Missing important details present in the Ground Truth, such as missing locations, activities, attributes, or contextual information. If details are only simplified, replaced with synonyms, or merged in the description, but the core information remains, it is not considered missing. Only when a detail is completely absent or replaced with irrelevant information should it be considered missing.

- Hallucination Control: Content generated that does not exist in the Ground Truth, such as invented locations, events, details, etc. Whether it is a variation, extension, or complete fabrication, it belongs to this category.

- Semantic Consistency: Inconsistencies with the Ground Truth in logic, sequence, location, time, relationships, etc. Minor descriptive differences (synonyms, rhetorical changes) should not affect scoring. Only when the difference causes factual errors, logical conflicts, sequence reversal, or errors in location/time/quantity should points be deducted.

Scoring Rules: (0–1 points):

1.0 point: Completely correct, no errors at all.

0.8 point: Only very few minor errors, not affecting overall understanding.

0.6 point: A certain number of moderate errors, affecting partial understanding.

0.4 point: Many errors, seriously affecting understanding.

0.2 point: Very many errors, almost impossible to understand correctly.

0.0 point: Completely wrong or irrelevant to the Ground Truth.

Output format requirement (must strictly follow the JSON structure below) :

{ "Text Recognition Accuracy": { "Error Description": "Detailed description of errors in text recognition.", "Score": 0-1 }, "Detail Coverage": { "Error Description": "Detailed description of missing details.", "Score": 0-1 }, "Hallucination Control": { "Error Description": "Detailed description of fabricated content.", "Score": 0-1 }, "Semantic Consistency": { "Error Description": "Detailed description of semantic consistency errors.", "Score": 0-1 } }

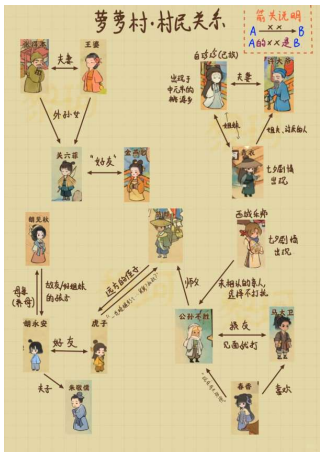
Considering that OIG-Bench is a bilingual dataset, for images containing Chinese text or originating from Chinese sources, all associated prompts were translated into Chinese, and the evaluated models were instructed to generate responses in Chinese to ensure linguistic consistency between the input and output.

A.7 Additional Error Cases

Figures 10 and 11 present two additional error case analyses. Figure 10 illustrates that when the font style of the game task name is highly stylized, the model's OCR capability drops significantly. Moreover, when the number of arrows is large, the model's perception of relationships becomes weaker, leading to highly confused character relationship recognition. In Figure 11, the error shown in one scenario is that the arrows and textual descriptions are mismatched: the image describes traveling from the Oriental Pearl Tower to Lujiazui by metro, whereas the model incorrectly recognizes the route as arriving at the Oriental Pearl Tower by metro.

A.8 LLM Usage Statement

We used a large language model to help improve the clarity, grammar, and readability of the manuscript. The authors carefully reviewed and edited all LLM-assisted text to ensure accuracy.



GPT-4o Description

Zhang Huaidong and Wang Po are husband and wife, and they have a granddaughter named Guan Liufei. Guan Liufei and Jin Yange are good friends. Zizhenzhen (deceased) and Grandpa Xu were husband and wife, and they had a daughter named Xu Chenzhen. Xu Chuan and Ah Kuan are brothers and sisters. Ah Kuan and a teacher from the Western Regions who was born in Qixi Theater are good friends. He Lang, a teacher from the Western Regions and a member of the Qixi Drama Club, is a sister and brother. Helang and Gongsun Wusheng are martial brothers, while Gongsun Wusheng and Matthew Wei are bad friends, and they will fight each other when they meet. Matthew likes the fragrance of spring. Chunxiang is the disciple of Gongsun Wuying. Huzi and Lang Jianqiu are good friends, and Lang Jianqiu is Lang Yong'an's mother (foster mother). Lang Yong'an and Zhu Jingru are husband and wife.

Text recognition accuracy:{
Error description: "There were multiple spelling or recognition errors in the character names predicted by the model. For example, 'Zhang Huaidong' should be 'Zhang Deben', 'Zizhenzhen' should be 'Bai Zhenzhen', 'Matthew Wei' should be 'Ma Dawei', 'He Lang' should be 'Huolang', 'Xiyu Teacher' should be 'Xiyu Musician', and 'Zhu Jingru' should be 'Fuzi (Zhu Jingru)'. In addition, names such as' Xu Chenchen 'and' Akuan 'did not appear in Ground Truth, which may be misidentification or fiction. There is also 'Chunxiang is a disciple of Gongsun', which should be 'good sisters/companion'. These are all issues related to the accuracy of text recognition. ",
Rating: 2
},
Detail coverage:{
Error description: "The model misses many important details: Bai Zhenzhen's appearance in Taoyuan Township on Zhongyuan Festival, the relationship between Qingyi and Uncle Xu's brother-in-law/annoying person, Qingyi's appearance in the Qixi story, Lang Yong'an's friendship with Huzi and related descriptions (such as" I'm a little annoyed with him now! "," people outside the city... dress strangely "), the goods man is Huzi's distant nephew, Gongsun Zui and Chunxiang are good sisters/companions, arrows at the top right corner of the relationship map, etc. Some teacher-student relationships, and the time of the story are not covered.",
Rating: 2
},
Hallucination control:{
Error description: "In the model prediction, many characters and relationships are fictionalized: 'Xu Chuanchuang', 'Ah Kuan', 'Western teacher from Qixi Drama Club', 'He Lang from Qixi Drama Club', 'He Lang and Gongsun Kui are brothers',' Chun Xiang is Gongsun Kui's apprentice ',' Lang Yongan and Zhu Jingru are husband and wife ', etc. are not shown in Ground Truth, which belongs to fiction. ",
Rating: 1
},
Semantic consistency:{
Error description: "The model is seriously inconsistent in terms of character relationship, logical order, kinship, etc. For example, Zhang Huaidong and Wang Po are husband and wife (Zhang Deben), Xu Chuan and Ah Kuan are sister brother relationships, Zhu Jingru and Lang Yong'an are husband and wife (teachers and students), Chunxiang is Gongsun Qiong's apprentice (a good sisters/companion), He Lang and Gongsun Qiong are teacher brothers (no such relationship), Hu Zi and Lang Jianqiu are friends (Lang Yong'an and Hu Zi should be friends), etc. Multiple relationships are confused, resulting in a serious deviation from the original picture in the overall semantic understanding.",
Rating: 1
}

Figure 10: Error case analysis. Note that the original images and model outputs are in Chinese; for ease of presentation, they are translated into English here.

GPT-4o Description

Warning: Shanghai Might Steal Your Heart (and Camera Roll)
Day 1

1) Nanjing Road
- Entrance fee: Free
- Transportation: Subway Line 1/10
- Directions/Reason: Come out from Exit 1 of the subway station and walk to Nanjing Road to see the most prosperous side of Shanghai.
- Next move: Walk for 15 minutes.

2) Oriental Pearl Tower
- Getting there: 20-minute subway ride on Line 2
- Entrance fee: 199 yuan
- Transportation: Subway Line 2/10
- Reason: It's a must-visit landmark in Shanghai. You can overlook the whole city from the top of the tower.

3) Lujiazui
- Next move to here: Walk for about 4 minutes
- Entrance fee: Free
- Transportation: Subway Line 2
- Reason: Navigate to Jinmao Junyue Hotel. There's an electrical box at the entrance. Ask a friend to squat down and take an upward-angled photo for a great shot (using a bottle opener + syringe + egg beater pose).

4) Shanghai IFC
- Entrance fee: Free
- Transportation: Subway Line 2
- Reason: It's renowned for its high-end brand positioning and unique design with stone-like appearance.
- Next move: 20-minute subway ride on Line 2.

5) The Bund Wharf
- Entrance fee: 2 yuan for the Bund ferry
- Transportation: Subway Line 2
- Reason: Spend 2 yuan on the ferry, feel the evening breeze on the Huangpu River, and experience the prosperity of Shanghai. It's only 10 minutes but beautiful and worthwhile.

Text recognition accuracy:{
Error description: " All location names predicted by the model are consistent with Ground Truth, without spelling errors, character errors, or number recognition errors. ",
Rating: 5
},
Detail coverage:{
Error description: "" The core information is not missing. The overall detail coverage is good",
Rating: 5
},
Hallucination control :{
Error description: " No fictional attractions or content have appeared. All attractions and their details are consistent with Ground Truth, with no fabricated information ",
Rating: 5
},
Semantic consistency:{
Error description: " In terms of traffic sequence, the model describes the arrival mode of 'Oriental Pearl Tower' as' 20 minute subway ride on Line 2 ', while the original text is' Walk for 15 minutes from Nanjing Road', which creates a logical inconsistency between traffic mode and time ",
Rating: 3
}

Figure 11: Error case analysis.