OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning

Ling Fu[1] **Zhebin Kuang**[1] **Jiajun Song**[1] **Mingxin Huang**[2] **Biao Yang**[1]
Yuzhe Li[1] **Linghao Zhu**[1] **Qidi Luo**[1] **Xinyu Wang**[3] **Hao Lu**[1] **Zhang Li**[1]
Guozhi Tang[4] **Bin Shan**[4] **Chunhui Lin**[4] **Qi Liu**[4] **Binghong Wu**[4]
Hao Feng[4] **Hao Liu**[4] **Can Huang**[4] **Jingqun Tang**[4] **Wei Chen**[1]
Lianwen Jin[2] **Yuliang Liu**[1*] **Xiang Bai**[1*]

[1]Huazhong University of Science and Technology [2]South China University of Technology
[3]University of Adelaide [4]ByteDance

Abstract

Scoring the Optical Character Recognition (OCR) capabilities of Large Multimodal Models (LMMs) has witnessed growing interest. Existing benchmarks have highlighted the impressive performance of LMMs in text recognition; however, their abilities in certain challenging tasks, such as text localization, handwritten content extraction, and logical reasoning, remain underexplored. To bridge this gap, we introduce **OCRBench v2**, a large-scale bilingual text-centric benchmark with currently the most comprehensive set of tasks ($4\times$ more tasks than the previous multi-scene benchmark OCRBench), the widest coverage of scenarios (31 diverse scenarios), and thorough evaluation metrics, with $10,000$ human-verified question-answering pairs and a high proportion of difficult samples. Moreover, we construct a private test set with $1,500$ manually annotated images. The consistent evaluation trends observed across both public and private test sets validate the OCRBench v2's reliability. After carefully benchmarking state-of-the-art LMMs, we find that most LMMs score below 50 (100 in total) and suffer from five-type limitations, including less frequently encountered text recognition, fine-grained perception, layout perception, complex element parsing, and logical reasoning. The benchmark and evaluation scripts are available at https://github.com/Yuliang-Liu/MultimodalOCR.

1 Introduction

The emergence of Large Language Models (LLMs) [1, 2, 3] has greatly improved the understanding and generation of structured text. However, in reality, much of the textual content is unstructured; it appears within images, videos, and other non-textual media in varied positions, orientations, and shapes. The need for processing such unstructured content leads to the study of Large Multimodal Models (LMMs) [4, 5, 6] that extend the text-only LLMs to additional modalities. By pretraining on multimodal data, LMMs acquire the zero-shot ability to interpret across diverse media, such as recognizing and understanding complex visual scene text [7]. Such capability represents a significant advancement over standard Optical Character Recognition (OCR), because LMMs not only spot text but also interpret its semantic relevance to a scene.

Compared with classic OCR that typically relies on task-specific models to spot text, the increasing capability of LMMs to process multimodal inputs has opened new potential to redefine the area of

*Corresponding authors

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

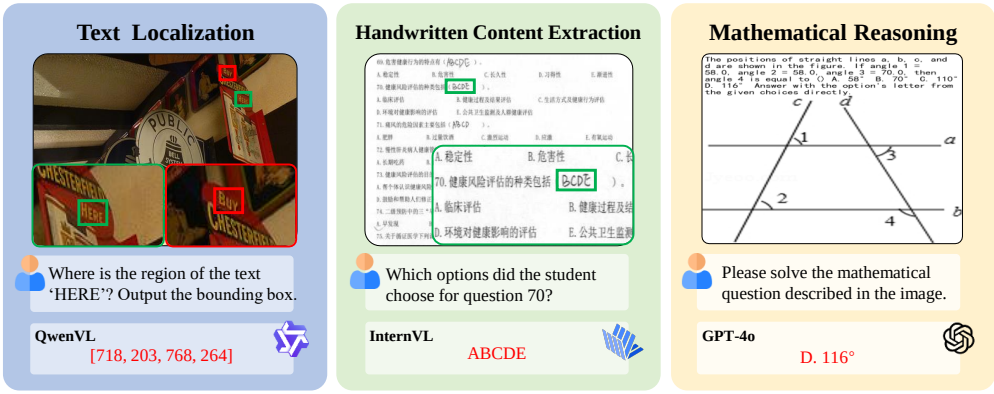


Figure 1: **Large multimodal models struggle with text-intensive tasks accurately**. They are prone to errors in tasks like text localization, handwritten content extraction, and mathematical reasoning, revealing limitations in tackling complex textual information within images.

OCR. OCR has therefore become an important aspect of recent LMM evaluations. Some text-focused tasks have been included in standard benchmarks to assess the proficiency of LMMs in recognizing and interpreting textual content [8, 9]. Typically, text-based Visual Question Answering (VQA) datasets [10, 11, 12] are repurposed to evaluate OCR by framing generic VQA into questions that require accurate reading of embedded text. However, many of these datasets are initially created for classic OCR models, which are of limited diversity, depth, and suitability for evaluating LMMs. A common drawback is that, many questions lack sufficient complexity to assess the reasoning abilities of LMMs on scene text, and some can even be answered without visual input [13, 12].

More recently, several customized benchmarks [14, 15, 16, 17, 18] have explored the OCR capabilities of LMMs. For example, OCRBench [14] consolidates 5 core text-oriented tasks to evaluate LMM performance across traditional OCR functions. Other datasets, such as ComTQA [19] and ChartX [20], focus on structured text interpretation like table and chart understanding. While such effort represents a leap over standard OCR benchmarks, they remain limited in both data diversity and quantity (see Tab. 1), often leading to rapid performance saturation. For example, recent LMMs such as Qwen2.5-VL [21] have achieved 96.4% accuracy on the DocVQA dataset [22], nearly matching human performance at 98.1%, and 88.8% on OCRBench [14]. This raises an important question for the community: *Do models perform well enough on text-oriented visual understanding tasks in the LMM era, or do existing benchmarks fail to capture the broader challenges in diverse environments?*

To answer the question above, we conducted preliminary tests with several state-of-the-art LMMs, including Qwen2.5-VL-7B [21], InternVL3-14B [23], and GPT-4o [24]. These tests assessed performance on text-oriented tasks, such as text localization, handwritten content extraction, and document-based logical reasoning. As illustrated in Fig. 1, each model can fail on one of the text-intensive tasks. These failures reveal a gap in detailed visual perception across different models, which constrains their effectiveness in tasks requiring accurate text localization, recognition, and contextual understanding within images. Recent benchmarks, such as OmniDocBench [25], CC-OCR [26], and MMLONGBENCH-DOC [27], have broadened evaluation to cover more comprehensive scenarios, including fine-grained document parsing and multi-page document understanding. Their analyses reveal the limited capabilities of LMMs for practical OCR applications and highlight the growing need for benchmarks that allow for more robust and varied evaluation of LMMs.

To bridge this gap, we propose *OCRBench v2*, a comprehensive benchmark designed to assess LMMs across diverse text-oriented visual understanding tasks. As shown in Fig. 3, *OCRBench v2* assesses eight core text-reading abilities, including *text recognition*, *text referring*, *text spotting*, *relation extraction*, *element parsing*, *mathematical calculation*, *visual text understanding*, and *knowledge reasoning*, organized into a total of 23 concrete tasks. This benchmark provides $10,000$ high-quality, human-validated instruction-response pairs and also six types of evaluation metrics, which offers a rigorous framework for evaluating LMM performance in complex, practical OCR scenarios. For better evaluation quality, we further collect and label $1,500$ additional text-images from scratch, reserved as the private test set. This private data serves as an independently curated test set to validate model generalization. In summary, the contributions of this work are three-fold:

Table 1: Comparison between the proposed benchmark and existing text-centric datasets.

Benchmark	#Scenario	#Task	#Image	#Instruction
OCRbench [14]	~ 14	5	0.9k	1k
Seed-bench-2-plus [15]	~ 8	1	0.6k	2.3k
CONTEXTUAL [16]	~ 11	1	0.5k	0.5k
Fox [17]	2	9	0.7k	2.2k
MMTab-eval [28]	1	9	23k	49k
ComTQA [19]	1	4	1.6k	9k
ChartX [20]	1	7	6k	6k
MMC [29]	1	9	1.7k	2.9k
OmniDocBench [25]	9	5	1k	1k
MMLONGBENCH-DOC [27]	7	2	6.4k	1.1k
OCRBench v2 (Ours)	31	23	9.5k	10k

- *OCRBench v2*: an improved benchmark designed to assess eight core OCR competencies and covers 23 tasks across 31 diverse scenarios, which provides a thorough evaluation framework encapsulating fundamental and advanced text-centric challenges.
- We systematically evaluate state-of-the-art LMMs, ranging from commercial APIs to open-source models, which establishes broad baselines for OCR performance and enables a comparative understanding of model capabilities across varied text-oriented visual understanding tasks.
- We provide a detailed analysis to identify factors affecting the OCR capabilities of LMMs. The analysis examines performance across various dimensions such as model generalization to diverse text types, model robustness, and the ability to tackle complex visual-textual relations.

2 Related Work

OCR-Enhanced LMMs. Inspired by LLMs, visual encoders are integrated into them to create LMMs capable of processing both images and text. Early LMMs exhibit strong zero-shot OCR capabilities, motivating the exploration of text-centric LMMs. For instance, some work [30, 31] use text-centric instruction-tuning to enhance OCR-related abilities. But they are restricted to low-res inputs, limiting the ability to recognize dense and small text. To address this, several studies [32, 33, 34] shift attention to increasing the input resolution. As the resolution of inputs increases, so does computational cost. To tackle this issue, TextMonkey [7] introduces a Token Resampler to compress redundant visual feature tokens, mPLUG-DocOwl2 [35] presents a DocCompressor module for compressing high-res images, and DocKylin [36] adopts adaptive pixel slimming and dynamic token slimming modules to reduce redundant regions. To enhance layout perception, DocLayLLM [37] integrates layout information into LMMs inputs, LayTokenLLM [38] shares position IDs between text and layout tokens, DocMark [39] utilizes adaptive generation of markup languages to build structured document representations, while Marten [40] introduces an additional mask generator during pre-training. Despite strong results on existing benchmarks, challenges remain unsolved in certain key areas such as text localization, entity extraction, and logical reasoning.

Benchmarks for Text-Centric LMMs. Previous efforts have focused on creating scenario-specific benchmarks to assess LMMs. For example, DocVQA [22], ChartQA [41], Infographics VQA [42], and TextVQA [10] evaluate models on document understanding, chart reasoning, infographic interpretation, and scene text comprehension, respectively. To broaden evaluation scope, OCRBench [14] introduces a holistic evaluation framework covering five text-oriented tasks, while CONTEXTUAL [16] and SEED-Bench-2-Plus [15] introduce context-sensitive and diverse real-world images. Other benchmarks target specific challenges such as dense text understanding [43], complex structure parsing [26], and fine-grained document analysis [25]. To provide a more thorough assessment, some benchmarks design multiple tasks within a specific scenario. TableVQA-Bench [18], MMTab [28], and ComTQA [19] explore table-based tasks, while ChartY [44], ChartX [20], and MMC [29] focus on chart information extraction and reasoning. OmniDocBench [25] focuses on document parsing tasks and provides a comprehensive evaluation framework. Recently, DUDE [45], MM-NIAH [46], MP-DocVQA [47], MMLONGBENCH-DOC [27], and LongDocURL [48] explore the long document understanding capability of LMMs. In this work, we establish *OCRBench v2*, a systematic benchmark to reveal the limitations of LMMs in diverse single-image, text-related scenarios.

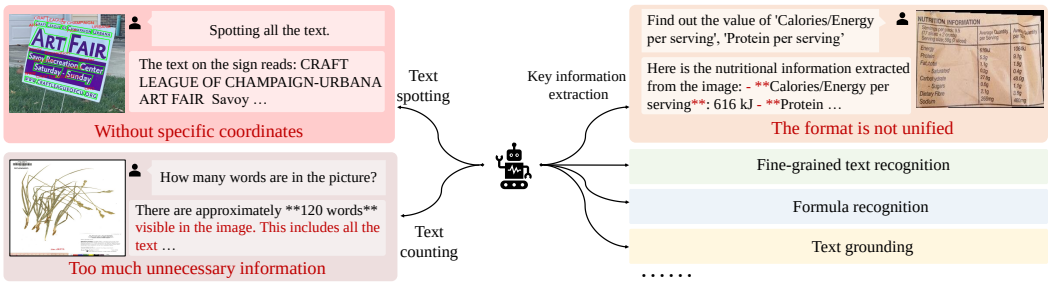


Figure 2: As evaluation for LMMs expands to diverse text-oriented tasks, existing datasets often require task-specific handling, making unified and scalable evaluation difficult.

3 Why Do We Need OCRBench v2?

Limitations of Existing Benchmarks. Recent evaluations of LMMs' OCR capabilities have made significant progress, yet most existing benchmarks exhibit limitations. Datasets like DocVQA, ChartQA, and TextVQA are often narrow in scope, focusing predominantly on text recognition within specific domains such as forms, tables, or documents. While useful for isolated capabilities, they fall short in task diversity, instruction complexity, and structured output formats that better reflect the multimodal nature of LMMs. In particular, many of these benchmarks were originally tailored for traditional OCR systems that prior to the emergence of LMMs. Furthermore, as illustrated in Fig. 2, complex task-specific processes are needed for LMMs when extended to more text-oriented tasks, which limits the evaluation of their broader capabilities. In this spirit, OCRBench v2 aims to evaluate OCR systems in terms of what ultimately matters: can a model recognize, understand, and reason over visual text to produce correct and meaningful answers?

The Necessity of Unified Multi-task Evaluation. With the emergence of LMMs, current models now excel at end-to-end performance across diverse tasks. Therefore, modern OCR goes beyond basic character recognition. Real-world documents often involve complex layouts and semantic structures that demand contextual understanding and reasoning. To assess these multi-task models, unified benchmarks like LongDocURL [48], OmniDocBench [25], CCOCR [26], OCRBench [14], CONTEXTUAL [16], SEED-Bench-2-Plus [15], have been proposed and successfully demonstrated the value of evaluating text-oriented models across diverse tasks. These benchmarks show the importance of unified evaluation frameworks in guiding model development. However, as model capabilities expand, existing benchmarks with limited task coverage result in fragmented and sometimes misleading insights. To address this, a unified benchmark is essential to: 1) *Understand generalization*: Can a model perform consistently across varied text-centric tasks? 2) *Diagnose failure models*: Does a model that excels in recognition also succeed in reasoning, localization, and parsing? 3) *Guide model development*: Unified evaluation provides clearer signals for architecture and training improvements.

As shown in Fig. 3, *OCRBench v2* tackles this by combining 23 tasks under 8 core capabilities within one framework. This holistic design enables systematic comparison of models and highlights trade-offs (e.g., performance on reasoning vs. recognition) that isolated benchmarks cannot reveal.

How OCRBench v2 Addresses the Gaps. *OCRBench v2* is a comprehensive, and high-difficulty benchmark specifically built to evaluate LMMs in realistic OCR settings, with key advantages: 1) *Breadth of coverage*: With 31 scenarios, we ensure diverse contextual challenges; 2) *Task variety*: The benchmark spans 8 OCR-related capabilities, many of which are poorly handled by current LMMs; 3) *Instruction complexity*: Human-authored prompts and structured outputs (e.g., Markdown, JSON, LaTeX) raise the bar beyond simple answer extraction; 4) *Private evaluation test set*: To prevent overfitting and training contamination, we additionally provide a private test set.

Ultimately, *OCRBench v2* fills a critical gap by offering a unified and challenging benchmark that reflects the practical needs of OCR in the LMM era. It not only measures what current models can do, but more importantly, reveals what they still cannot.

Design Rationale: Focusing on Single-Image Text Tasks. While designing *OCRBench v2*, we focus on challenges in single-image, text-related scenarios, and do not extend our study to multi-image tasks. This design choice is grounded in two considerations: 1) Single-image understanding is the

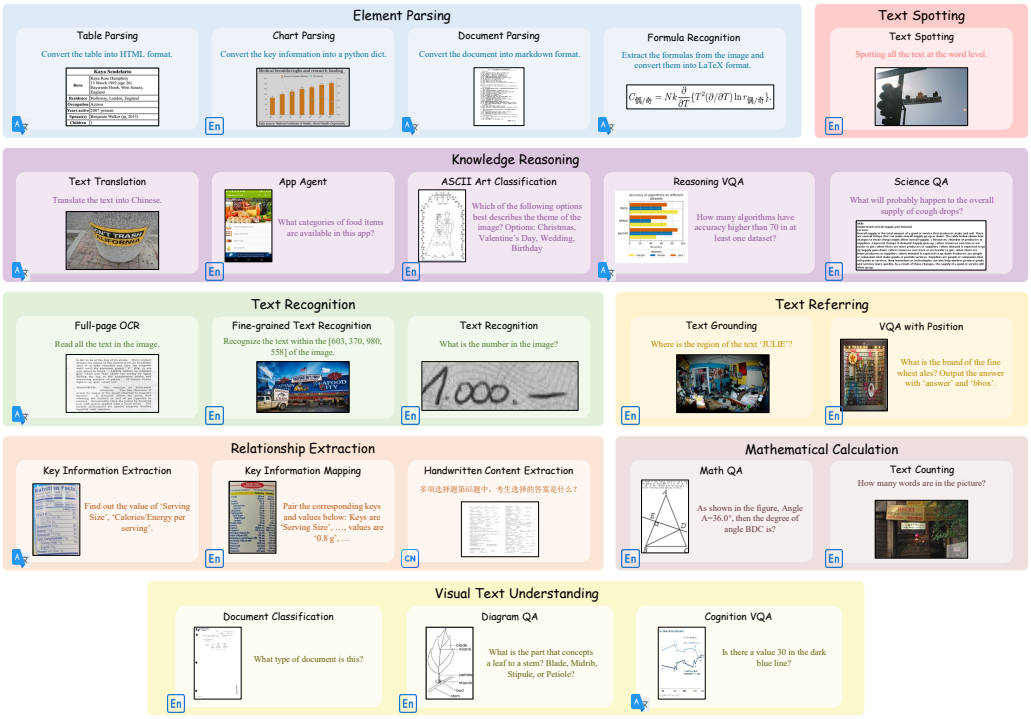


Figure 3: **Sample visualizations for each task.** OCRBench v2 comprises 23 sub-tasks grouped under 8 core OCR capabilities. Tasks marked with A contain both English and Chinese instructions, while other tasks are either English-only En or Chinese-only CN (Zoomed in for better clarity).

foundation for more complex multimodal tasks. Many existing models still perform unsatisfactorily in various single-image scenarios, which motivates our work; 2) Given long-context inputs, multi-page tasks have more emphasis on long-sequence modeling, requiring specific benchmarks to assess this capability individually. For example, MMLONGBENCH-DOC focuses on evaluating the ability of LMMs to locate and understand content across pages in long documents.

Private Dataset for Reliable Evaluation. To further enhance the assessment quality, we also construct a private test set. This data comprises $1,500$ manually collected text-rich images with human-annotated labels, covering 23 tasks aligned with the distribution of the public data. Among the private data, 735 images were manually captured, and 765 images were sourced from unlabeled data with diverse scenarios. The data sources include printed books, e-books, scanned documents, and web content. During data collection and annotation, we meticulously curated samples to align with practical text-oriented applications. Given that benchmarks may be contaminated in massive internet-scraped pre-training data of LMMs, this data will not be released. Instead, we maintain a regularly updated leaderboard to reflect the performance of advanced LMMs. Moreover, consistent performance trends and model rankings observed on both the public and private test sets (see Section 5.2) indicate the benchmark's well-founded design and its effectiveness in identifying model capabilities.

4 Benchmark Construction

In this section, we describe the task description, annotation curation, statistics, and evaluation criteria.

4.1 Task Description

To provide a comprehensive evaluation framework for text-reading tasks, we categorize OCR capabilities into eight core areas, each encompassing specific sub-tasks that address various aspects of

text comprehension and interpretation. Fig. 3 exhibits samples for each task, with visual inputs and corresponding instructions. Detailed descriptions of these core capabilities are as follows.

Text Recognition. This fundamental capability focuses on perceiving textual content. The related tasks include (fine-grained) text recognition and full-page OCR.

Text Referring. Determining the location of texts accurately is necessary for real-world OCR applications. This ability is evaluated with text grounding and VQA with position tasks.

Text Spotting. Text spotting is a widely studied OCR task that requires models to output both the location and content of text. We consider it a distinct capability due to this unique output format.

Relation Extraction. Given that texts are often densely arranged in images, the ability to extract and map visual components is essential. This capability is assessed through key information extraction, key information mapping, and handwritten content extraction.

Element Parsing. LMMs face the need of parsing complex elements for downstream applications. This ability is evaluated via table parsing, chart parsing, document parsing, and formula recognition.

Mathematical Calculation. Math calculation is essential for LMMs to address numerical reasoning tasks. Hence, text counting is introduced to assess the textual perception ability. Besides, we enhance the math QA data by rendering textual questions into images, accompanied by geometric figures.

Visual Text Understanding. To tackle sophisticated tasks involving human interaction, LMMs need to comprehend the semantic information of texts, a capability we term visual text understanding. This ability is evaluated by document classification and diagram QA. Additionally, we include basic VQA instructions where answers are located directly within the image, which refers to cognition VQA.

Knowledge Reasoning. Some tasks require complex inference and world knowledge, including science QA, APP agent interactions, ASCII art classification, text translation, and reasoning VQA (where answers are not directly visible in images).

4.2 Annotation Curation

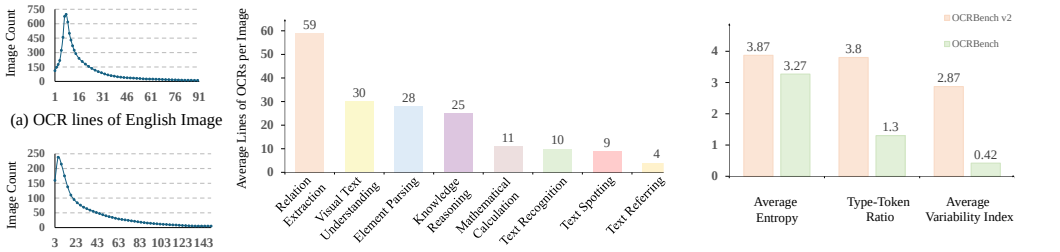
Dataset Collection. To ensure data diversity, we manually harvest and screen 81 text-rich academic datasets. To ensure diverse scenario coverage, we also supplement them with additional private data. In all, our dataset comprises 31 typical scenarios (see Tab. 11 for the full list).

Annotation Protocol. Before starting the annotation, we conducted thorough discussions to establish clear guidelines. For example, in questions involving numbers such as dates, amounts, or frequencies, answers were required to include all common formats—Arabic numerals, English abbreviations, and full English expressions. For coordinate-related questions, all coordinate values in the answers were normalized to a 0–1000 scale based on the image size to ensure consistency across varying image resolutions. In cases where multiple correct answers were possible, all valid answers were included. For the "read all text" task, we required that the answer follow a natural reading order from left to right and from top to bottom. Based on these guidelines, 15 professional annotators carried out the annotation work. Each annotator strictly adhered to the instructions and created QA pairs along with the relevant coordinate information, depending on the task requirements.

Manual Verification. To ensure data quality, we perform a manual cross-validation process to ensure accuracy and quality. Specifically, each annotated example was first completed by one annotator, then reviewed by a second annotator to verify the correctness. If disagreements or ambiguities arose, the case was escalated to a third annotator for judgment. In instances where consensus could not be reached among all three annotators, the corresponding instruction was excluded from the dataset. Finally approximately 1% annotations are corrected.

4.3 Statistics of OCRBench v2

Here we present the OCR-related statistics and the measurement of prompt quality. As shown in Fig. 4 (a) and (b), we count the distribution of line-level OCR results of $7,400$ English and $2,600$ Chinese images. And Fig. 4 (c) exhibits the average number of line-level OCR results per category. These statistics demonstrate that the text information is sufficiently rich in *OCRBench v2*. In addition, Fig. 4 (d) compares the Average Entropy, Type-Token Ratio, and Average Variability Index of the



(a) OCR lines of English Image

(b) OCR lines of Chinese Image

(c) Statistics by Average OCR Lines of Capabilities

(d) Comparison between OCRBench v2 and OCRBench

Figure 4: **OCR-related statistics and prompt quality assessment of OCRBench v2**.

Table 2: **Evaluation of existing LMMs on English tasks of OCRBench v2's public data**. "Recognition", "Referring", "Spotting", "Extraction", "Parsing", "Calculation", "Understanding", and "Reasoning" refer to text recognition, text referring, text spotting, relation extraction, element parsing, mathematical calculation, visual text understanding, and knowledge reasoning, respectively. Higher values indicate better performance. Best performance is in boldface, and the second best is underlined. The notations apply to all subsequent figures.

Method	Recognition	Referring	Spotting	Extraction	Parsing	Calculation	Understanding	Reasoning	Average
Open-source LMMs									
LLaVA-Next-8B [49]	41.3	18.8	0	49.5	21.2	17.3	55.2	48.9	31.5
LLaVA-OV-7B [50]	46.0	20.8	0.1	58.3	25.3	23.3	64.4	53.0	36.4
Monkey [51]	35.2	0	0	16.6	16.3	14.4	59.8	42.3	23.1
TextMonkey [7]	39.1	0.7	0	19.0	12.2	19.0	61.1	40.2	23.9
Molmo-7B [52]	52.4	21.3	0.1	45.5	7.6	28.5	65.3	55.0	34.5
Cambrian-1-8B [53]	45.3	21.5	0	53.6	19.2	19.5	63.5	55.5	34.7
Pixtral-12B [54]	48.9	21.6	0	66.3	35.5	29.8	66.9	53.7	40.3
Qwen2.5-VL-7B [21]	<u>68.8</u>	25.7	1.2	<u>80.2</u>	30.4	38.2	73.2	56.2	46.7
InternVL3-14B [23]	67.3	<u>36.9</u>	<u>11.2</u>	**89.0**	38.4	38.4	**79.2**	60.5	**52.6**
Deepseek-VL2-Small [55]	62.7	28.0	0.1	77.5	32.7	14.3	<u>77.1</u>	53.9	43.3
MiniCPM-o-2.6 [56]	66.9	29.5	0.5	70.8	33.4	31.9	69.9	57.9	45.1
GLM-4V-9B [57]	61.8	22.6	0	71.7	31.6	22.6	72.1	58.4	42.6
Ovis2-8B [58]	**73.2**	24.6	0.7	62.4	**44.8**	40.6	72.7	**62.6**	47.7
Closed-source LMMs									
GPT-4o [1]	61.2	26.7	0	77.5	36.3	<u>43.4</u>	71.1	55.5	46.5
GPT-4o-mini [59]	57.9	23.3	0.6	70.8	31.5	38.8	65.9	55.1	43.0
Gemini-Pro [60]	61.2	**39.5**	**13.5**	79.3	<u>39.2</u>	**47.7**	75.5	59.3	<u>51.9</u>
Claude3.5-sonnet [61]	62.2	28.4	1.3	56.6	37.8	40.8	73.5	<u>60.9</u>	45.2
Step-1V [62]	67.8	31.3	7.2	73.6	37.2	27.8	69.8	58.6	46.7

questions between *OCRBench v2* and OCRBench. *OCRBench v2* presents higher values across all three metrics, indicating more diverse, less redundant, and structurally varied questions. This suggests it provides a more comprehensive and challenging benchmark for LMMs.

4.4 Evaluation Criteria

We adopt six types of evaluation metrics tailored to specific task categories. In the following, we present an overview of the evaluation metrics and their applicability to specific tasks.

Parsing Type. To evaluate the element parsing ability of LMMs, we assess their performance in transforming input images into structured formats, including HTML, Markdown, and JSON. TEDS [63] is employed to measure the structural similarity between outputs and the desired format.

Localization Type. For text referring, the IoU score is applied to quantify the distance between the predicted regions and the ground truth.

Extraction Type. To evaluate relation extraction, we employ the F1 score to assess key information extraction and mapping. Since this evaluation requires structural extraction of information from the output of LMMs, the format is provided in the given prompt.

Long Reading Type. To assess performance on long text reading tasks, BLEU [64], METEOR [65], F1 score, and edit distance are used to assess the similarity between predicted text and ground truth.

Table 3: **Evaluation of existing LMMs on Chinese tasks of OCRBench v2's public data**. "LLM Size" indicates the number of parameters of the language model employed in each method.

Method	LLM Size	Recognition	Extraction	Parsing	Understanding	Reasoning	Average
Open-source LMMs							
LLaVA-Next-8B [49]	8B	5.7	2.9	12.2	7.5	17.2	9.1
LLaVA-OV-7B [50]	8B	14.8	15.7	13.7	16.0	28.7	17.8
Monkey [51]	8B	4.6	11.2	8.4	21.5	20.0	13.1
TextMonkey [7]	8B	23.5	14.8	8.4	19.9	12.2	15.8
Molmo-7B [52]	8B	7.1	15.0	9.2	9.0	23.7	12.8
Cambrian-1-8B [53]	8B	5.3	14.9	12.6	8.5	8.1	9.9
Pixtral-12B [54]	12B	13.4	10.9	21.0	7.0	20.7	14.6
Qwen2.5-VL-7B [21]	8B	**75.3**	61.4	**41.8**	59.3	40.4	55.6
InternVL3-14B [23]	14B	66.2	**64.8**	33.5	**63.4**	**50.6**	**55.7**
Deepseek-VL2-Small [55]	16B	60.9	50.6	28.3	53.0	20.5	42.7
MiniCPM-o-2.6 [56]	7B	53.0	49.4	27.1	43.5	32.7	41.1
GLM-4V-9B [57]	9B	24.4	60.6	20.4	52.8	25.2	36.6
Ovis2-8B [58]	7B	72.2	50.8	37.7	47.9	37.4	49.2
Closed-source LMMs							
GPT-4o [1]	-	21.6	53.0	29.8	38.5	18.2	32.2
GPT-4o-mini [59]	-	13.1	38.9	27.2	28.8	16.9	25.0
Gemini-Pro [60]	-	52.5	47.3	30.9	51.5	33.4	43.1
Claude3.5-sonnet [61]	-	21.0	56.2	35.2	55.0	30.5	39.6
Step-1V [62]	-	56.7	41.1	37.6	38.3	39.2	42.6

Counting Type. In text counting, LMMs are required to count the number of text instances. Thus, we use the L1 distance to measure the absolute difference between predicted and ground truth counts. The final score is then normalized to the range of $[0, 1]$ based on the ground truth.

Basic VQA Type. For questions where the original data provides options, we use exact string matching to compute accuracy. In other cases, we follow the approach of OCRBench to check whether the ground truth is contained in the prediction for short answers (fewer than 5 words) and employ ANLS to measure prediction quality for longer answers (5 words or more).

5 Results and Findings

Here we first benchmark state-of-the-art LMMs on *OCRBench v2*, presenting the quantitative analysis, then summarize key findings of current limitations for LMMs. All results are presented as percentages.

5.1 Baselines

The tested LMMs in this section includes LLaVA-Next-8B [49], LLaVA-OV-7B [50], Monkey [51], TextMonkey [7], Molmo-7B [52], Cambrian-1-8B [53], Pixtral-12B [54], Qwen2.5-VL-7B [21], InternVL3-14B [23], Deepseek-VL2-Tiny [55], MiniCPM-o-2.6 [56], GLM-4v-9B [57], Ovis2-8B [58], GPT4o [24], GPT4o-mini [59], Gemini-1.5-Pro [60], Claude3.5-sonnet [61], and Step-1V [62]. More LMM evaluation results can be found in Tabs. 12, 13, 14, and 15.

5.2 Main Results

Evaluation results on public data are shown in Tab. 2 and Tab. 3. While LMMs perform well on some basic capabilities such as text recognition and visual text understanding, most LMMs achieve low scores in other capabilities, such as text spotting and element parsing, mostly below 50. In particular, some LMMs show significant limitations in text spotting capabilities, failing to precisely locate and recognize the texts. Additionally, LMMs demonstrate inadequate abilities in element parsing and mathematical calculation, which are crucial for complicated tasks like document analysis and mathematical reasoning. Besides, after comparing the performance of LMMs on visual text understanding and knowledge reasoning capabilities, we find that they perform poorly in knowledge reasoning. This suggests the deficiency of LMMs in logical reasoning.

Evaluation results on private data are shown in Tab. 4 and Tab. 5. We observe similar evaluation trends to those in the public test set experiments. Overall, LMMs exhibit unsatisfactory performance in text referring, text spotting, element parsing, mathematical calculation, and knowledge reasoning capabilities. In addition, closed-source LMMs outperform their open-source counterparts, demon-

Table 4: **Evaluation of existing LMMs on English tasks of OCRBench v2's private data**.

Method	Recognition	Referring	Spotting	Extraction	Parsing	Calculation	Understanding	Reasoning	Average
Open-source LMMs									
LLaVA-Next-8B [49]	41.4	17.0	0	49.0	12.9	16.1	60.9	30.5	28.5
LLaVA-OV-7B [50]	45.4	18.5	0	60.0	15.5	32.0	59.0	39.3	33.7
Monkey [51]	31.5	0.1	0	34.4	26.3	17.7	61.4	22.4	24.2
TextMonkey [7]	39.8	1.6	0	27.6	24.8	10.2	62.3	21.2	23.4
Molmo-7B [52]	40.8	19.5	0	51.7	10.0	33.9	67.0	48.0	33.9
Cambrian-1-8B [53]	44.0	19.0	0	52.3	19.0	20.7	64.0	39.3	32.3
Pixtral-12B [54]	45.1	21.8	0	71.6	21.7	30.4	77.3	39.5	38.4
Qwen2.5-VL-7B [66]	51.5	24.5	3.1	64.8	13.1	53.3	78.6	45.5	41.8
InternVL3-14B [23]	55.8	24.5	2.1	89.3	21.0	59.5	72.0	50.0	46.8
Deepseek-VL2-Small [55]	56.6	23.7	0	86.4	18.9	30.6	72.2	39.5	41.0
MiniCPM-o-2.6 [56]	54.1	24.7	0.3	74.4	17.6	39.2	75.7	47.0	41.6
GLM-4v-9B [57]	52.7	20.6	0	79.4	15.9	21.5	74.7	32.0	37.1
Ovis2-8B [58]	54.2	20.9	0	83.6	24.2	54.7	74.1	57.3	46.1
Closed-source LMMs									
GPT-4o [1]	58.6	23.4	0	87.4	23.1	51.6	74.4	**62.3**	47.6
GPT-4o-mini [59]	55.3	21.8	0	85.4	20.6	45.2	75.5	49.0	44.1
Gemini1.5-Pro [60]	**59.1**	**41.2**	**6.6**	**89.5**	22.4	54.7	**78.8**	60.3	**51.6**
Claude3.5-sonnet [61]	52.9	24.9	2.5	86.9	23.8	**61.4**	74.4	53.0	47.5
Step-1V [62]	56.7	27.4	2.6	86.3	**33.3**	42.6	76.6	48.7	46.8

Table 5: **Evaluation of existing LMMs on Chinese tasks of OCRBench v2's private data**.

Method	LLM Size	Recognition	Extraction	Parsing	Understanding	Reasoning	Average
Open-source LMMs							
LLaVA-Next-8B [49]	8B	2.8	0.9	14.9	20.0	7.4	9.2
LLaVA-OV-7B [50]	8B	5.4	13.6	20.3	34.0	13.6	17.4
Monkey [51]	8B	1.5	28.4	29.1	40.0	8.3	21.5
TextMonkey [7]	8B	10.5	15.2	30.2	44.0	7.6	21.5
Molmo-7B [52]	8B	3.4	29.8	6.6	24.0	11.1	15.0
Cambrian-1-8B [53]	8B	2.4	19.8	26.7	36.0	7.6	18.5
Pixtral-12B [54]	12B	6.2	22.3	11.4	26.0	14.0	16.0
Qwen2.5-VL-7B [66]	8B	24.4	**78.9**	33.1	**82.0**	29.0	49.5
InternVL3-14B [23]	14B	62.1	59.5	33.2	80.0	29.2	52.8
DeepSeek-VL2-Small [55]	16B	51.6	56.3	27.8	79.6	25.3	48.1
MiniCPM-o-2.6 [56]	7B	54.0	62.4	24.1	68.0	29.8	47.7
GLM-4v-9B [57]	9B	60.6	65.2	32.4	**82.0**	18.2	51.7
Ovis2-8B [58]	7B	61.0	67.7	**43.6**	**82.0**	25.6	**56.0**
Closed-source LMMs							
GPT-4o [1]	-	41.7	52.1	29.0	76.0	29.4	45.7
GPT-4o-mini [59]	-	20.0	53.6	27.9	66.0	19.6	37.4
Gemini1.5-Pro [60]	-	**71.4**	63.8	30.5	**82.0**	29.9	55.5
Claude3.5-sonnet [61]	-	34.2	62.5	35.2	78.0	**32.2**	48.4
Step-1V [62]	-	65.2	64.9	33.1	78.0	25.5	53.4

strating stronger generalization capabilities. The consistent results across both public and private test sets confirm the soundness of *OCRBench v2*'s task design, data collection process, and evaluation metrics, and demonstrate its effectiveness in revealing the capability limitations of current LMMs.

5.3 Main Findings

We provide in-depth analyses for LMMs' common limitations, including rare text recognition, fine-grained spatial perception, layout perception, complex element analysis, and logical reasoning.

Finding 1. LMMs still face challenges with less frequently encountered texts, such as dot matrix texts and mathematical formulas. This performance gap highlights the continuing challenges LMMs face in real-world text recognition. For instance, occluded text, CAPTCHA, and dot-matrix text are considered low-frequency text, whereas other types belong to high-frequency text. Tab. 6 shows the performance of some LMMs on high-frequency and low-frequency texts. Notably, recognition accuracy varies significantly across these categories. For example, InternVL3-14B achieves 79.1% on high-frequency texts but drops to 46.7% on low-frequency ones.

Finding 2. Current LMMs still exhibit limited performance in tasks requiring precise spatial understanding, such as text referring and text spotting. For instance, when provided with coordinate information as input, many models are able to output the relevant content from captions or chapters.

Table 6: **LMMs' performance on high- and low-frequency words**.

Category	Pixtral-12B [54]	Cambrian-1-8B [53]	InternVL3-14B [23]	Qwen2.5-VL-7B [66]
High Frequency	58.3	59.8	79.1	84.5
Low Frequency	23.6	40.2	46.7	53.3

However, almost all models struggle to accurately retrieve the corresponding text from documents with dense text based on given coordinates. We investigate the content response accuracy and the IoU score for answer region localization in the VQA with position task. Tab. 7 suggests that although LMMs can roughly identify where the answer is located, they struggle to output the exact region.

Finding 3. While LMMs achieve good performance on basic text recognition, they struggle with complex layouts such as overlapping or rotated texts. For example, GPT-4o fails to detect the characters in overlapping handwritten text and misrecognizes numbers in 90° rotated images, revealing LMMs' limitations in handling texts with complex layouts. Rotating images in the DocVQA dataset led to a significant performance drop of 55.7% for InternVL3-14B (from 90.9% to 35.2%).

Table 7: **LMMs' performance on VQA with position task**.

Category	Pixtral-12B [54]	Cambrian-1-8B [53]	InternVL3-14B [23]	Qwen2.5-VL-7B [66]
Content Accuarcy	68.8	71.7	78.3	75.2
IoU Accuracy	1.7	0.0	12.9	9.6

Finding 4. LMMs still struggle to parse text into structured formats in downstream applications such as document digitalization. For instance, InternVL3-14B achieves 94.4% accuracy in unpaired entities matching, but its performance drops to 84.9% in key information extraction, where the model is required to identify the corresponding value given an entity. The performance further degrades in element parsing tasks that demand structured outputs.

Finding 5. Despite recent advances, LMMs still face challenges in complex mathematical and textual reasoning tasks. To assess their capabilities, we evaluated InternVL3-14B on the private test set covering reasoning VQA, ScienceQA, and APP agent tasks. Questions were categorized into five types: common sense reasoning, visual-text understanding, pattern recognition, calculation, and expert knowledge. Human ratings showed the model achieved accuracies of 72.9%, 83.0%, 69.2%, 56.5%, and 71.8%, respectively, indicating notable variation.

6 Conclusion

In this work, we introduce *OCRBench v2*, a comprehensive benchmark designed to evaluate the OCR capabilities of LMMs. Covering 23 tasks across 31 diverse scenarios, our benchmark systematically assesses eight core capabilities that are essential for text-oriented visual understanding tasks. It includes $10,000$ high-quality QA pairs and six rigorous evaluation metrics. In addition, we curate a private test set of $1,500$ manually labeled images to ensure robust generalization evaluation. Leveraging this benchmark, we conduct extensive experiments on representative LMMs. Through in-depth analysis of experimental results, we identify critical limitations of current models and uncover key factors that affect their OCR performance. We hope *OCRBench v2* could aid future research on enhancing LMMs' text understanding ability.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant Nos. 62225603 and 62206104).