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

A Technical Appendices and Supplementary Material

This supplementary material contains the following content:

- **Sec. A.1**: Comparison experiments between LMMs and some text-centric expert models.
- **Sec. A.2**: Data collection.
- **Sec. A.3**: Task definitions.
- **Sec. A.4**: Additional statistics of *OCRBench v2*.
- **Sec. A.5**: Evaluation metrics.
- **Sec. A.6**: Experimental setting for the evaluation process.
- **Sec. A.7**: Compute resources for the evaluation process.
- **Sec. A.8**: Evaluation results for LMMs on *OCRBench v2*.
- **Sec. A.9**: Potential factors affecting OCR capabilities
- **Sec. A.10**: Visualization samples for task examples.
- **Sec. A.11**: Visualization samples for failure cases.
- **Sec. A.12**: Discussion of broader impacts.
- **Sec. A.13**: Discussion of limitations.

A.1 Comparison with LMMs and Text-centric Expert Models

Comparison with text recognizers. We compare LMMs with several representative scene text recognizers, including CRNN [1], ABINet [2], ASTER [3], MASTER [4], and SVTR [5], on the text recognition task. The weights of these models are loaded from mmocr[1]. The results are shown in Tab. 1, where we selected 5 representative LMMs, including Qwen2.5VL-7B [6], InternVL3-14B [7], GPT4o [8], Gemini1.5-Pro [9], and Step-1V [10]. The results demonstrate that LMMs exhibit remarkable text recognition capabilities, validating our motivation to evaluate LMMs on more challenging OCR-related tasks.

Comparison with text spotters. We also compare LMMs with ABCNet series [11, 12] and TESTR [13] on the text spotting task. The ABCNet series utilize the official weights[2], and TESTR is also initialized with its publicly released checkpoint[3]. These models were fine-tuned with Total-Text [14]. The results are shown in Tab. 2. Although LMMs demonstrate promising capabilities in text recognition, there remains notable potential for improvement in the text spotting task.

Comparison with GOT. We notice a recent work, GOT [15], that can parse the textual elements within images. We conduct comparison experiments between GOT and some representative LMMs, and the results are shown in Tab. 3. We observe that LMMs show advantages in general text recognition, while GOT demonstrates better performance in the document parsing task.

[1] `https://github.com/open-mmlab/mmocr`
[2] `https://github.com/aim-uofa/AdelaiDet`
[3] `https://github.com/mlpc-ucsd/TESTR`

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Method	Accuracy
CRNN [1]	38.1
ABINet [2]	62.4
ASTER [3]	50.0
MASTER [4]	54.1
SVTR [5]	57.8
Qwen2.5VL-7B [6]	73.0
InternVL3-14B [7]	71.1
GPT4o [8]	74.1
Gemini1.5-Pro [9]	64.1
Step-1V [10]	75.4

Table 1: Comparison between LMMs and text recognizers.

Method	F1 score
ABCNet [11]	32.2
ABCNetV2 [12]	44.2
TESTR [13]	51.8
Qwen2.5VL-7B [6]	1.2
InternVL3-14B [7]	11.2
Gemini1.5-Pro [9]	13.5
GPT4o [8]	0
Step-1V [10]	7.2

Table 2: Comparison between LMMs and text spotters.

A.2 Data Collection

Text Recognition. The data for text recognition task are sampled from ICDAR2013 [16], SVT [17], IIIT5K [18], ICDAR2015 [19], SCUT-CTW1500 [20], COCO-Text [21], CUTE80 [22], TotalText, SVTP [23], WordArt [24], NonSemanticText [25], IAM [26], ORAND-CAR-2014 [27], HOST [28], and WOST [28]. Meanwhile, CAPTCHA (Completely Automated Public Turing Test to Tell Humans Apart) images are sourced from a CAPTCHA dataset[4] and a number CAPTCHA dataset[5]. Additionally, dot matrix images in the text recognition task are manually collected from the web page.

Fine-grained Text Recognition. In the fine-grained text recognition task, images are sampled from the test sets of Fox [29], Totaltext, COCO-Text, CTW1500 [30], and ICDAR2015. We use the original annotations for Fox, while the other datasets are manually re-annotated.

Full-page OCR. The data sources for full-page OCR task include Fox, HierText [31], CTW [32], RCTW-17 [33], ReCTS [34], LSVT2019 [35], M6Doc [36], and CDLA[6].

Text Grounding. The images for the text grounding task are sampled from testset of Totaltext, COCO-Text, CTW1500, and ICDAR2015. QA pairs and bounding boxes annotations are based on their official OCR annotations.

VQA with Position. The images used for VQA with position task are sampled from the test sets of TextVQA [37] and RICO [38], with QA pairs and bounding box annotations derived from their original datasets.

Text Spotting. The data sources for the text spotting task include Totaltext, COCO-Text, CTW1500, and ICDAR2015.

Key Information Extraction. The data sources for key information extraction task include FUNSD [39], SROIE [40], POIE [41], M6Doc, XFUND [42], ICDAR2023-SVRD [43], and a private dataset of photographed receipts.

Key Information Mapping. The data sources for the key information mapping task include FUNSD and POIE.

[4] https://aistudio.baidu.com/datasetdetail/159309

[5] https://www.heywhale.com/mw/dataset/5e5e56b6b8dfce002d7ee42c/file

[6] https://github.com/buptlihang/CDLA

Method	Rec	FG-Rec	Full-Rec	Doc-Parse
GOT [15]	64.1	52.9	73.3	53.9
Qwen2.5VL-7B [6]	73.0	36.4	84.2	39.1
InternVL3-14B [7]	71.1	36.4	83.0	36.9
GPT4o [8]	74.1	13.8	54.1	35.9
Gemini1.5-Pro [9]	64.1	22.9	83.9	40.5
Step-1V [10]	76.8	24.8	74.8	36.0

Table 3: Comparison between LMMs and GOT [15].

Handwritten Content Extraction. This task's data is our private data, which contains real exam paper data with student information removed and manually annotated QA pairs.

Table Parsing. The images for table parsing task are selected from MMTab [44], WTW [45], TabRecSet [46] and flush table recognition competition[7].

Chart Parsing. The data sources for the chart parsing task come from OneChart [47] and MMC [48].

Document Parsing. The data sources for document parsing task come from DoTA [49], DocVQA [50], M6Doc, and CDLA.

Formula Recognition. The data sources for the formula Recognition task includes HME100K [51], IM2LATEX-100K [52], M2E [53], MathWriting [54], MLHME-38K[8], CASIA-CSDB [55], and some private data.

Math QA. The data sources for the math QA task includes MathMatics [56], MathVerse [57], MathVision [52], and MathVista [58].

Text Counting. The data for the text counting task are collected from IIIT5K, SVT, ICDAR2013, HierText, and TotalText.

Cognition VQA. The data sources for the cognition VQA task include EST-VQA [59], OCRVQA [60], ST-VQA [61], TEXTVQA, DIR300 [62], ChartQA [63], DVQA [64], PlotQA [65], InfoVQA [66], WTW, PubTabNet [67], WTQ [68], CORD [69], LLaVAR [70], WebSRC [71], DocVQA, M6Doc, XFUND, Publaynet [72], RVL-CDIP [73], ScreenQA [74], SlideVQA [75], a movie poster collection dataset[9], a website screenshot collection dataset[10], and a private receipt photograph dataset.

Diagram QA. The data sources for the diagram QA task include AI2D [76] and TextBookQA [77].

Document Classification. The images for the document classification task are collected from RVL-CDIP.

Reasoning VQA. The reasoning VQA task shares some common data sources with the cognition VQA task. Additionally, portions of the reasoning VQA dataset are drawn from MMSI [78] and CMMMU [79].

Science QA. The images and annotations of the science QA task are collected from ScienceQA [80] and MMMU-Pro [81]

APP Agent. The data source of the APP agent task is RICO.

ASCII Art Classification. The data sources for the ASCII art classification task is ASCIIEval [82].

Text Translation. The datasets collected for text translation task includes memes[11], MSRA-TD500 [83], MTWI2018 [84], M6Doc, ICDAR2023-SVRD, EST-VQA, RCTW17 [85],

[7] https://github.com/10jqka-aicubes/table-recognition

[8] https://ai.100tal.com/icdar

[9] https://www.kaggle.com/datasets/neha1703/movie-genre-from-its-poster

[10] https://huggingface.co/datasets/Zexanima/website_screenshots_image_dataset/tree/main

[11] https://www.kaggle.com/datasets/dvishal485/meme-challenge?resource=download

Scene	Number	Scene	Number	Scene	Number
Schematic diagram	1238	Scientific paper	799	Word	728
Table(filled)	705	Chart	620	Receipts	609
Questions	581	Mathematical formula	475	Product labels	434
Phone screenshot	431	Indoor scenes	395	Industry research reports	343
Poster	264	Street scene	224	ASCII Art	199
Shop sign	189	Financial reports	153	Chemical formula	149
Textbook	148	Magazine	146	Email	111
Web screenshot	99	Details page	95	Verification code	87
Resumes	67	Illustration	61	Newspaper	52
Road signs	43	Menus	31	Notify	30
Questionnaire	29				

Table 4: The number of images included in each scene category in public data.

DAST1500 [86], XFUND, ArT2019 [87], ChartQA, CDLA, ICDAR2015, SlideVQA, Fintabnet [88], ScienceQA, InfoVQA, COMICS-Dialogue[12], and ExpressExpense SRD[13].

A.3 Task Definitions

In this section, we introduce the definition of each task, and the visualizations for each task can be found in Sec. A.10.

Text Recognition. Text recognition refers to the fundamental OCR ability on text image patches, which asks LMMs to read the text content. To comprehensively evaluate LMMs' text recognition ability across diverse scenarios, our collection incorporates various text types, including regular text, irregular text, artistic text, handwriting text, digit string text, non-semantic text, occluded text, doc matrix text, and CAPTCHA text.

Fine-grained Text Recognition. This task requires LLMs to read and comprehend textual content within the given region. It evaluates LLMs' fine-grained perception capabilities in understanding text in natural scenes and documents.

Full-page OCR. Full-page OCR [29] task requires LMMs to extract and recognize all text content from the given images. Converting text into digital format facilitates subsequent processing and analysis of text images.

Text grounding. In this task, users would provide a text string and require LMMs to locate its specific location, evaluating LMMs' fine-grained perception capabilities.

VQA with Position. For VQA with position task, LMMs need to not only respond to the question but also provide the exact position coordinates that directly correspond to the answer. We ask LMMs to output both information in JSON format for convenient evaluation, and the coordinates are required to be normalized with image sizes and scaled to the range of $[0, 1000]$.

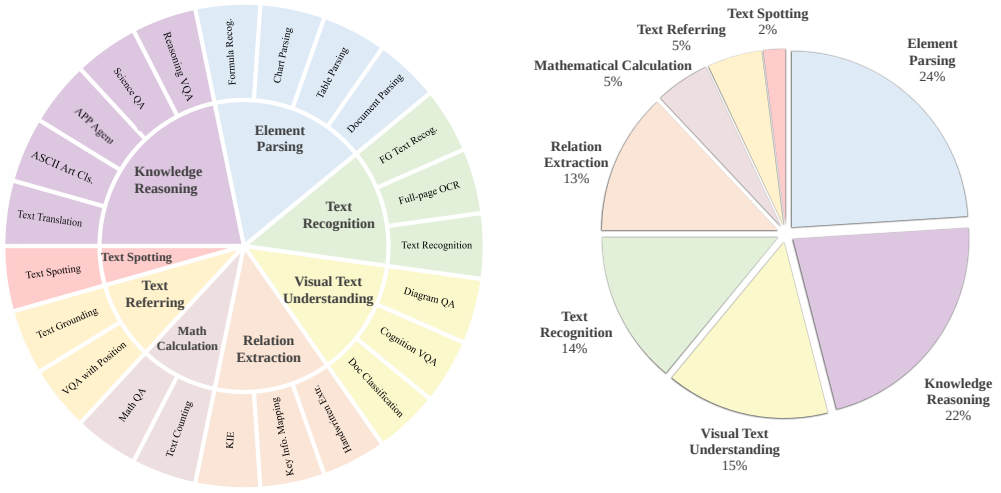
Text Spotting. Text spotting task needs LMMs to output the localization and content of all appeared text simultaneously. Due to the interference of background elements and the large number of text instances, this task demands high fine-grained perception capabilities from the model. Besides, the coordinates are required to be normalized with image sizes and scaled to the range of $[0, 1000]$.

Key Information Extraction. The key information extraction task is to extract the necessary information from densely arranged text. In this task, we provide some desired entities as keys and demand LMMs to output the corresponding values to form the output JSON string.

Key Information Mapping. In this task, we provide a set of entity keys and their corresponding values in the prompt. The LMMs are then asked to match and pair these keys with their respective values into groups.

[12] https://huggingface.co/datasets/lmms-lab/M4-Instruct-Data

[13] https://expressexpense.com/blog/free-receipt-images-ocr-machine-learning-dataset/



(a) Overview of the eight and associated tasks in OCRbench v2

(b) The ratios of each type of tasks in OCRBench v2

Figure 1: **Overview of the eight testable text-reading capabilities and associated tasks in OCRBench v2**. Each color represents a distinct capability type.

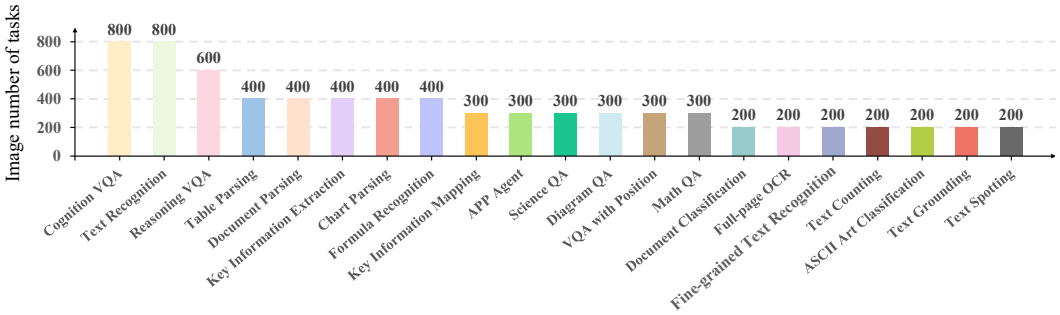


Figure 2: The quantity distribution of English tasks of public data.

Handwritten Content Extraction. To investigate the information extraction capabilities of LMMs in educational scenarios, we collect some Chinese examination papers, containing both printed question text and handwritten student responses. There are four types of questions in these examination papers, including single-choice, multiple-choice, true or false, and brief response questions. The prompts require LMMs to extract the handwritten content for specific questions.

Table Parsing. Table parsing task requires LMMs to parse the given table into structured text, including Markdown and HTML format.

Chart Parsing. Apart from tables, charts can also be converted to structured information. In this task, LLMs are required to transform visual charts into JSON format.

Document Parsing. In the document parsing task, both text and the complex elements, including charts, tables, and formulas, are required to be parsed.

Formula Recognition. This task asks LMMs to recognize the given formula in the LaTeX format. The collection includes mathematical and chemical formulas.

Math QA. Math QA task evaluates the LMMs' mathematical calculation ability. In particular, we render the mathematical problem description and related figures into images and ask LMMs to answer the questions within the images.

Text Counting. Text counting task is built to evaluate the quantity property perceiving ability of LMMs, including the character frequency in words and the word counting in the given image.

Cognition VQA. In *OCRBench v2*, we split text-centric VQA instructions into cognition VQA and Reasoning VQA based on whether the answers can be directly found in the images. Cognition VQA

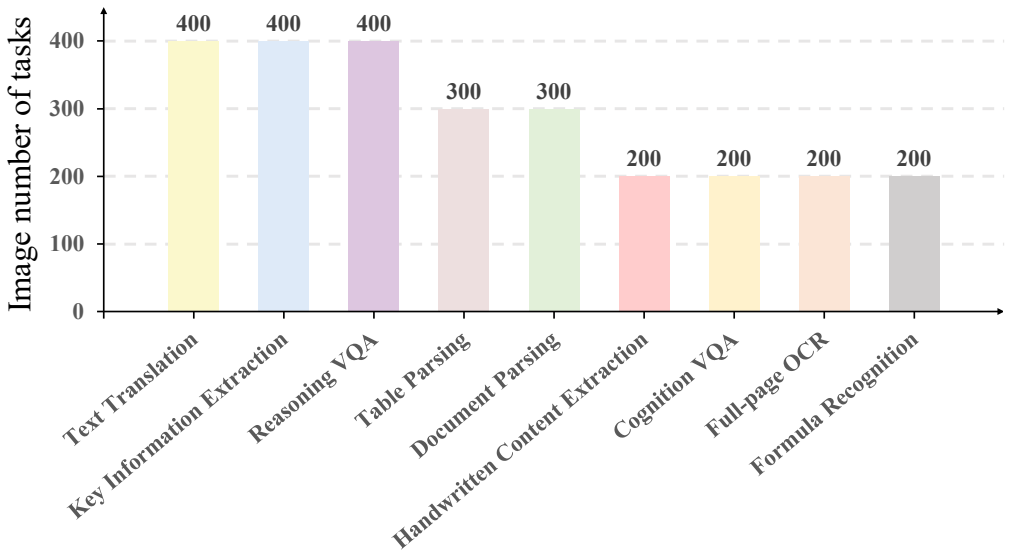


Figure 3: The quantity distribution of Chinese tasks of public data.

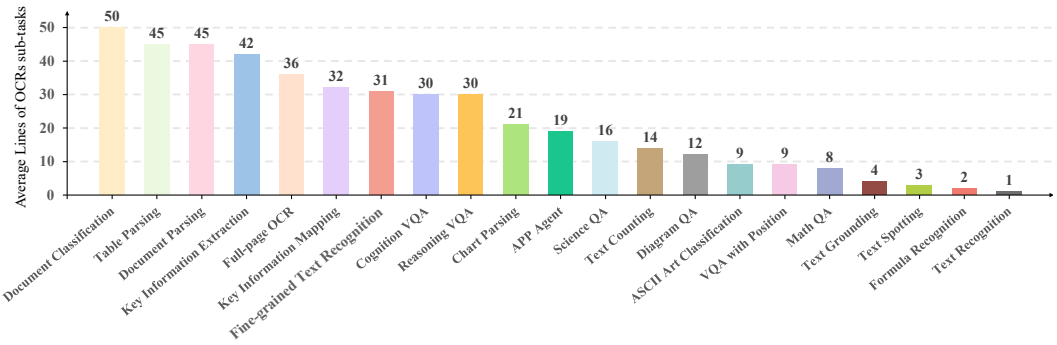


Figure 4: The OCR lines distribution of English tasks of public data.

task refers to the instructions where answers are explicitly present in the given image. This task evaluates the fundamental text-centric question-answering ability based on visual content.

Diagram QA. In the diagram QA task, LMMs need to respond to the question about the given diagrams, reflecting LMMs' ability to understand the relationship between the visual elements.

Document Classification. Document classification task asks LMMs to classify the category of the given document image. The included categories are letters, forms, emails, handwritten documents, advertisements, scientific reports, scientific publications, specifications, file folders, news articles, budgets, invoices, presentations, questionnaires, resumes, and memos.

Reasoning VQA. In reasoning VQA tasks, the answers often do not directly appear in the image. This forces LMMs to perform logical reasoning to respond to questions based on visual information.

Science QA. In the Science QA task, LMMs are required to respond to the scientific problem. We use PaddleOCR[14] to extract text from the collected images and filter out those with fewer than four OCR results. Additionally, when extra subject-related knowledge is provided by the source, we incorporate it by rendering it into the images.

APP Agent. For the APP agent task, LMMs need to understand the relationship between textual content, icons, and world knowledge to respond to the question from the user, simulating the real-world application scene.

[14] `https://github.com/PaddlePaddle/PaddleOCR`

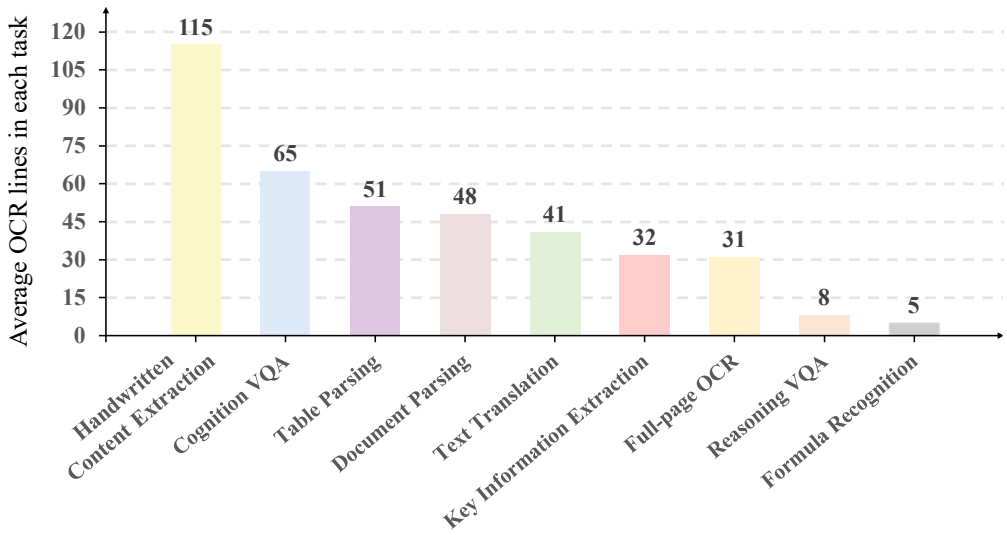


Figure 5: The OCR lines distribution of Chinese tasks of public data.

ASCII Art Classification. We incorporate a recent image classification task that uses images composed purely of ASCII characters [82]. This task is included in *OCRBench v2* to evaluate LMMs' ability to assess LMMs' pattern recognition and visual abstraction abilities.

Text Translation. In the text translation task, LMMs need to execute translation between Chinese and English texts, evaluating LMMs' semantic understanding abilities.

A.4 Additional Statistics of OCRBench v2

Scene Coverage. Our dataset can be divided into 31 classic scenes according to the scene of the image. The specific scenes and the corresponding number of pictures are shown in Tab. 4.

Statistics of each task. Fig. 1 shows an overview of each task in *OCRBench v2*.The distribution of 23 tasks in *OCRBench v2* is displayed in Fig. 2 and Fig. 3. Additionally, we calculate and present the average number of OCR text lines per task in Fig. 4 and Fig. 5. As illustrated in these figures, the task distribution is well-balanced, with each task containing adequate textual information for analysis.

A.5 Evaluation Metrics

Parsing Type. We use Tree-Edit-Distance-based Similarity (TEDS) [89] to evaluate parsing tasks, which require LMMs to transform the images to structured formats. Tree Edit Distance (TED) refers to the minimum number of edits to transform one tree into another. TEDS is based on TED to calculate the similarity of two trees. Assuming T_1 and T_2 are two different trees, $TED(T_1, T_2)$ refers to their TED, and the TEDS is defined as:

$$TEDS(T_1, T_2) = 1 - \frac{TED(T_1, T_2)}{\max(|T_1|, |T_2|)}, \tag{1}$$

where $|T_1|$ and $|T_2|$ is the number of nodes of trees, $TED(T_1, T_2)$ can be calculated by dynamic programming algorithm. If T_1 and T_2 are identical, then their TEDS equals 1. As the structural difference between two trees increases, their TED value becomes larger, resulting in the TEDS approaching 0.

Localization Type. In the text referring and spotting tasks, LMMs are required to provide regression bounding boxes of target objects. IoU score is adopted to measure the distance between the predicted regions and the ground truth.

$$IoU(B_1, B_2) = \frac{Intersect(B_1, B_2)}{Union(B_1, B_2)}, \tag{2}$$

where $Intersect(B_1, B_2)$ refers to the overlap area of bounding box B_1 and B_2, while $Union(B_1, B_2)$ refers to their union area.

Extraction Type. The F1 score is used to evaluate LMMs' relation extraction capability. Given the predicted and ground truth Key-Value pairs, the F1 score is formulated as follows:

$$Precision = \frac{N_3}{N_2}, \tag{3}$$

$$Recall = \frac{N_3}{N_1}, \tag{4}$$

$$Fmean = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{5}$$

where N_1, N_2, and N_3 denote the number of ground-truth Key-Value pairs, predicted Key-Value pairs, and correctly matched Key-Value pairs, respectively.

Long Reading Type. To evaluate LMMs' ability to recognize text across entire paragraphs or pages, BLEU [90], METEOR [91], F1 score, and normalized edit distance are employed. And the final score is the average value of these metrics.

BLEU evaluates prediction quality by comparing n-gram match rates between the prediction and ground truth sequences. For each n-gram type, precision is calculated as the ratio of matching n-grams to total predicted n-grams. The final BLEU score is the geometric mean of these precision values multiplied by a penalty BP, which is defined as:

$$BLEU = BP * exp(\sum_{n=1}^{N} w_n \log p_n), \tag{6}$$

$$BP = \begin{cases} 1 & L_p \geq L_g \\ e^{(1-\frac{L_p}{L_g})} & L_p < L_g \end{cases}, \tag{7}$$

where p_n represents the precision of n-grams, L_p represents the length of prediction sequence, L_g represents the length of ground truth sequence, w_n is weight factor, usually evenly distributed ($w_n = \frac{1}{N}$). Typically, N is set to 4.

METEOR employs a semantic-aware matching strategy with four levels. 1) Exact Match: words in the prediction that are identical to the ground truth. 2) Stem match: matching words that have the same word stem. 3) Synonym Match: matching words based on synonymous relationships. 4) Paraphrase Match: Matching similar phrases at the phrase level. These matches are combined to calculate precision and recall, from which a weighted harmonic mean F1 score is derived as:

$$P_{meteor} = \frac{N_{match}}{N_{pred}}, \tag{8}$$

$$R_{meteor} = \frac{N_{match}}{N_{gt}}, \tag{9}$$

$$F_{meteor} = \frac{10 * P_{meteor} * R_{meteor}}{P_{meteor} + 9 * R_{meteor}}, \tag{10}$$

where N_{match}, N_{pred}, and N_{gt} represent the number of matched items, words in prediction, and words in ground truth, respectively. The final METEOR score is obtained by multiplying the F_{meteor} by the penalty adjustment factor. The calculation is formulated as follows:

$$METEOR = F_{meteor} * (1 - BP_{meteor}), \tag{11}$$

$$BP_{meteor} = 0.5 * \frac{N_{chunk}}{N_{match}}, \tag{12}$$

where N_{chunk} refers to the number of contiguous matching phrases. More chunks indicate greater word order differences, resulting in a heavier penalty.

The calculation method of the F1 score in long reading metrics follows the same approach as discussed in extraction metrics, as shown in Equations 3, 4, 5.

Normalized Edit Distance (NED) measures string similarity by computing the minimum number of operations needed to transform one string into another. And then NED is normalized by the length of the longer string. The calculation is formulated as follows:

$$NED(S_1, S_2) = \frac{ED(S_1, S_2)}{\max(len(S_1), len(S_2))} \tag{13}$$

where $ED(S_1, S_2)$ represents the edit distance between the prediction string S_1 and the ground truth S_2. The NED value of 0 indicates identical strings, while 1 indicates completely different strings.

Counting Type. In *OCRBench v2*, character frequency counting and word counting tasks are included. For character frequency, we use exact match evaluation since the answers are typically single-digit integers. For word counting, we evaluate using the L1 distance between predicted and ground truth counts, normalized to $[0, 1]$ based on the ground truth. This can be formulated as follows:

$$score = \begin{cases} 0 & C_{pred} \leq 0 \\ 1 - \frac{|C_{pred} - C_{gt}|}{C_{gt}}) & 0 < C_{pred} < 2 * C_{gt} \\ 0 & C_{pred} \geq 2 * C_{gt} \end{cases}, \tag{14}$$

where C_{pred} and C_{gt} denote the predicted count and ground truth count, respectively.

Basic VQA Type. The remaining tasks in *OCRBench v2* are basic VQA types, and we employ different evaluation metrics based on question types. For multiple-choice questions, we use exact matching between predictions and answer options. In other cases, we check whether the ground truth is contained in the prediction for answers shorter than 5 words, and use ANLS for longer answers.

A.6 Experimental setting

The detailed public data construction are shown in Sec. A.2 and Sec. A.5. Private data consists of unlabeled images collected manually from websites and real life. At the same time, we annotated and checked the private test set to ensure the quality. The environment configuration of each open-source model experiment strictly complies with the official version and uses the official pre-trained model and inference code. The model parameters of the open-source model and the API parameters of the closed-source model use the official default parameters for fair. Specifically, we use the official API versions: GPT-4o (gpt-4o-2024-08-06), GPT-4o-mini (gpt-4o-mini-2024-07-18), and Gemini 1.5 Pro (gemini1.5-pro-002).

A.7 Compute resources

Evaluations of open-source models were conducted on 8×NVIDIA GeForce RTX 4090 (24GB) and a NVIDIA H800 Tensor Core GPU (80GB). The closed-source experiments obtained the results by calling the official API.

A.8 Results and Discussions

Tab. 5, Tab. 6, Tab. 7, and Tab. 8 exhibit the results of 39 open-source models and 5 closed-source models on the public and private test sets of *OCRBench v2*

Evaluation results on public data are shown in Tab. 5 and Tab. 6. Most LMMs performed well in tasks such as Understanding, Recognition, Extraction, which shows that current models have basic OCR capabilities. However, they performed poorly in tasks such as Referring, Spotting, Parsing, and Calculation. The scores of all models are basically below 50 points, which shows that the models still lack the ability in text localization, logical reasoning, and understanding complex elements.

Evaluation results on private data are shown in Tab. 7 and Tab. 8. The performance trends of the models on private and public datasets are consistent. In addition, most models perform worse on

Method	Recognition	Referring	Spotting	Extraction	Parsing	Calculation	Understanding	Reasoning	Average
Open-source LMMs									
LLaVA-Next-8B [92]	41.3	18.8	0	49.5	21.2	17.3	55.2	48.9	31.5
LLaVA-OV-7B [93]	46.0	20.8	0.1	58.3	25.3	23.3	64.4	53.0	36.4
Monkey [94]	35.2	0	0	16.6	16.3	14.4	59.8	42.3	23.1
TextMonkey [95]	39.1	0.7	0	19.0	12.2	19.0	61.1	40.2	23.9
XComposer2-4KHD [96]	45.1	21.8	0.1	15.9	11.7	15.7	66.8	45.9	27.9
Molmo-7B [97]	52.4	21.3	0.1	45.5	7.6	28.5	65.3	55.0	34.5
Cambrian-1-8B [98]	45.3	21.5	0	53.6	19.2	19.5	63.5	55.5	34.7
Pixtral-12B [99]	48.9	21.6	0	66.3	35.5	29.8	66.9	53.7	40.3
EMU2-chat [100]	42.1	0.2	0	12.5	8.1	11.2	42.7	33.4	18.8
mPLUG-Owl3 [101]	41.6	14.0	0.6	24.4	10.9	11.1	52.2	46.0	25.1
CogVLM-chat [102]	50.9	0	0	0.2	8.4	15.0	58.1	41.7	21.8
Qwen-VL [103]	34.6	7.5	0	18.2	20.0	8.1	57.2	41.1	23.3
Qwen-VL-chat [103]	34.5	4.1	0	25.9	14.0	13.8	55.7	39.5	23.4
Qwen2-Vl-7B [6]	<u>72.1</u>	**47.9**	**17.5**	82.5	25.5	25.4	78.4	61.5	51.4
Qwen2.5-VL-7B [104]	68.8	25.7	1.2	80.2	30.4	38.2	73.2	56.2	46.7
InternVL2-8B [105]	49.9	23.1	0.5	65.2	24.8	26.7	73.5	52.9	39.6
InternVL2-26B [105]	63.4	26.1	0	76.8	37.8	32.3	**79.4**	58.9	46.8
InternVL2.5-8B [7]	59.0	25.0	1.4	77.5	35.1	29.4	75.3	57.2	45.0
InternVL2.5-26B [7]	65.6	26.1	1.6	<u>86.9</u>	36.2	37.4	78.3	**62.9**	49.4
InternVL3-8B [7]	68.6	30.4	8.8	85.3	34.0	27.1	77.5	60.3	49.0
InternVL3-14B [7]	67.3	36.9	11.2	**89.0**	38.4	38.4	<u>79.2</u>	60.5	**52.6**
Deepseek-VL-7B [106]	37.1	15.4	0	23.5	14.6	20.8	53.3	52.9	27.2
Deepseek-VL2-Small [107]	62.7	28.0	0.1	77.5	32.7	14.3	77.1	53.9	43.3
MiniCPM-V-2.6 [108]	66.8	6.0	0.8	62.0	28.8	32.4	73.7	52.1	40.3
MiniCPM-o-2.6 [108]	66.9	29.5	0.5	70.8	33.4	31.9	69.9	57.9	45.1
GLM-4V-9B [109]	61.8	22.6	0	71.7	31.6	22.6	72.1	58.4	42.6
VILA1.5-8B [110]	35.3	15.5	0	21.1	12.7	17.3	46.3	40.3	23.6
LLaVAR [70]	37.3	0	0	1.0	9.9	12.3	34.6	27.0	15.3
UReader [111]	22.4	0.1	0	0	9.2	7.9	41.0	29.1	13.7
DocOwl2 [112]	24.0	9.7	0	13.4	13.5	8.8	53.7	32.0	19.4
Yi-VL-6B [113]	28.9	2.9	0	9.7	12.9	15.8	36.1	32.0	17.3
Janus-1.3B [114]	46.1	0	0	0.2	14.5	13.5	36.0	39.1	18.7
Eagle-X5-7B [115]	34.7	17.8	0	21.7	20.6	21.5	61.0	42.6	27.5
Idefics3-8B [116]	23.8	13.2	0	63.2	23.8	23.0	65.8	44.9	32.2
Phi-4-MultiModal [117]	63.7	16.4	0	40.4	19.1	18.3	69.8	53.9	35.2
SAIL-VL-1.6-8B [118]	67.7	28.6	2.8	70.5	25.9	29.5	73.9	59.7	44.8
Kimi-VL-A3B-16B [119]	56.5	13.8	0	59.2	33.8	32.9	75.5	56.7	41.1
Ovis1.6-3B [120]	59.2	14.3	0	65.0	32.1	29.0	69.8	56.8	40.8
Ovis2-8B [120]	**73.2**	24.6	0.7	62.4	**44.8**	40.6	72.7	<u>62.6</u>	47.7
Closed-source LMMs									
GPT-4o [8]	61.2	26.7	0	77.5	36.3	<u>43.4</u>	71.1	55.5	46.5
GPT-4o-mini [121]	57.9	23.3	0.6	70.8	31.5	38.8	65.9	55.1	43.0
Gemini-Pro [9]	61.2	<u>39.5</u>	<u>13.5</u>	79.3	<u>39.2</u>	**47.7**	75.5	59.3	<u>51.9</u>
Claude3.5-sonnet [122]	62.2	28.4	1.3	56.6	37.8	40.8	73.5	60.9	45.2
Step-1V [10]	67.8	31.3	7.2	73.6	37.2	27.8	69.8	58.6	46.7

Table 5: **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.

private datasets than on public datasets, which shows that private data may be more challenging for LMMs due to the lack of training, and also reflects the importance of private data construction.

A.9 Potential Factors Affecting OCR Capabilities

High-Res Visual Encoders. Since text often appears small in images, the resolution setting of the visual encoder could be a key factor affecting the text perception ability [94]. Here we change the input resolution of the LMMs and observe the performance changes. In particular, InternVL2-8B is chosen, and the resolution setting includes 448, 896, and dynamic. Tab. 9 lists the results. Indeed, when the input resolution increases from 448 to 896, the performance increases by 4.1%.

Pre-provided OCR Information. To study the impact of OCR information, we use PaddleOCR[15] to pre-extract OCR results and incorporate them with prompts. Tab. 10 shows the results. We observe that adding OCR information does not help much. This suggests that *OCRBench v2* evaluates LMMs capabilities across multiple dimensions, rather than solely focusing on text recognition abilities.

[15] `https://github.com/PaddlePaddle/PaddleOCR`

Method	LLM Size	Recognition	Extraction	Parsing	Understanding	Reasoning	Average
Open-source LMMs							
LLaVA-Next-8B [92]	8B	5.7	2.9	12.2	7.5	17.2	9.1
LLaVA-OV-7B [93]	8B	14.8	15.7	13.7	16.0	28.7	17.8
Monkey [94]	8B	4.6	11.2	8.4	21.5	20.0	13.1
TextMonkey [95]	8B	23.5	14.8	8.4	19.9	12.2	15.8
XComposer2-4KHD [96]	7B	16.7	18.8	12.1	27.5	2.3	15.5
Molmo-7B [97]	8B	7.1	15.0	9.2	9.0	23.7	12.8
Cambrian-1-8B [98]	8B	5.3	14.9	12.6	8.5	8.1	9.9
Pixtral-12B [99]	12B	13.4	10.9	21.0	7.0	20.7	14.6
EMU2-chat [100]	37B	2.3	0.5	8.5	1.0	7.3	3.9
mPLUG-Owl3 [101]	8B	6.6	17.9	9.7	6.0	26.1	13.3
CogVLM-chat [102]	7B	5.5	10.0	9.8	1.5	2.5	5.9
Qwen-VL [103]	8B	7.2	5.3	10.7	11.5	11.2	9.2
Qwen-VL-chat [103]	8B	9.5	8.2	9.3	11.0	21.1	11.8
Qwen2-Vl-7B [6]	7B	51.3	51.4	21.6	52.5	37.5	42.9
Qwen2.5-VL-7B [104]	7B	**75.3**	61.4	**41.8**	59.3	40.4	55.6
InternVL2-8B [105]	8B	20.6	45.2	23.2	54.4	38.1	36.3
InternVL2-26B [105]	26B	21.9	46.0	34.8	50.9	34.8	37.7
InternVL2.5-8B [7]	8B	52.8	52.8	28.6	56.4	40.5	46.2
InternVL2.5-26B [7]	26B	32.4	56.1	32.6	56.3	43.6	44.2
InternVL3-8B [7]	8B	68.9	62.0	31.6	57.9	47.3	53.5
InternVL3-14B [7]	14B	66.2	**64.8**	33.5	**63.4**	**50.6**	**55.7**
Deepseek-VL-7B [106]	7B	8.0	13.3	15.7	5.5	18.5	12.2
Deepseek-VL2-Small [107]	16B	60.9	50.6	28.3	53.0	20.5	42.7
MiniCPM-V-2.6 [108]	8B	51.0	29.9	21.2	34.0	33.6	33.9
MiniCPM-o-2.6 [108]	7B	53.0	49.4	27.1	43.5	32.7	41.1
GLM-4V-9B [109]	9B	24.4	60.6	20.4	52.8	25.2	36.6
VILA1.5-8B [110]	8B	5.4	8.8	8.5	3.0	15.5	8.2
LLaVAR [70]	13B	2.3	1.7	8.9	0	2.5	3.1
UReader [111]	7B	6.8	2.7	8.4	2.5	7.2	5.5
DocOwl2 [112]	7B	4.2	10.3	8.6	4.0	9.6	7.3
Yi-VL-6B [113]	6B	4.8	4.4	8.5	4.0	25.0	9.4
Janus-1.3B [114]	1.3B	7.6	8.7	11.4	4.5	10.7	8.6
Eagle-X5-7B [115]	8B	7.5	12.0	11.6	5.0	19.2	11.1
Idefics3-8B [116]	8B	7.0	15.5	15.9	9.0	18.1	13.1
Phi-4-MultiModal [117]	5.6B	51.5	32.3	12.1	34.4	23.0	30.7
SAIL-VL-1.6-8B [118]	8B	31.2	40.0	23.9	42.3	35.0	34.5
Kimi-VL-A3B-16B [119]	16B	57.2	54.7	31.5	52.5	31.4	45.5
Ovis1.6-3B [120]	3B	11.5	23.7	22.8	28.8	18.9	21.1
Ovis2-8B [120]	7B	72.2	50.8	37.7	47.9	37.4	49.2
Closed-source LMMs							
GPT-4o [8]	-	21.6	53.0	29.8	38.5	18.2	32.2
GPT-4o-mini [121]	-	13.1	38.9	27.2	28.8	16.9	25.0
Gemini-Pro [9]	-	52.5	47.3	30.9	51.5	33.4	43.1
Claude3.5-sonnet [122]	-	21.0	56.2	35.2	55.0	30.5	39.6
Step-1V [10]	-	56.7	41.1	37.6	38.3	39.2	42.6

Table 6: **Evaluation of existing LMMs on Chinese tasks of OCRBench v2**. "LLM Size" indicates the number of parameters of the language model employed in each method.

Connection Between OCR and LLMs. We further explore a direct pipeline by first extracting OCR information and then by feeding it directly into Qwen2.5. Unlike LMMs, this pipeline separates OCR and language modeling into distinct stages. The results shown in Tab. 10 suggest that Qwen2-VL-7B outperforms Qwen2.5 with OCR information, demonstrating LMMs' remarkable ability to incorporate both textual and visual features efficiently.

A.10 Samples for Each Task

As show in Fig. 6 to Fig. 14 , there are 23 OCR tasks included in *OCRBench v2*. Among them, Fig. 6 to Fig. 12 present examples of English tasks, including text recognition, diagram QA, text counting, formula recognition, math QA, VQA with position, ASCII art classification, reasoning VQA, text translation, APP agent, table parsing, cognition VQA, document classification, science QA, chart parsing, key information extraction, full-page OCR, text spotting, fine-grained text recognition, text grounding, key information mapping, and document parsing. These figures show corresponding images and QA pairs for each of the 23 tasks. Fig. 13 to Fig. 14 provide examples of Chinese tasks, including key information extraction, text translation, formula recognition, reasoning VQA, cognition VQA, handwritten content extraction, document parsing, full-page OCR, and table parsing, along with their associated images and QA pairs.

Method	Recognition	Referring	Spotting	Extraction	Parsing	Calculation	Understanding	Reasoning	Average
Open-source LMMs									
LLaVA-Next-8B [92]	41.4	17.0	0	49.0	12.9	16.1	60.9	30.5	28.5
LLaVA-OV-7B [93]	45.4	18.5	0	60.0	15.5	32.0	59.0	39.3	33.7
Monkey [94]	31.5	0.1	0	34.4	26.3	17.7	61.4	22.4	24.2
TextMonkey [95]	39.8	1.6	0	27.6	24.8	10.2	62.3	21.2	23.4
XComposer2-4KHD [96]	39.5	12.0	0	69.7	26.0	20.2	68.2	35.8	33.9
Molmo-7B [97]	40.8	19.5	0	51.7	10.0	33.9	67.0	48.0	33.9
Cambrian-1-8B [98]	44.0	19.0	0	52.3	19.0	20.7	64.0	39.3	32.3
Pixtral-12B [99]	45.1	21.8	0	71.6	21.7	30.4	77.3	39.5	38.4
EMU2-chat [100]	34.3	0	0	20.4	21.3	20.3	47.1	18.3	20.2
mPLUG-Owl3 [101]	34.9	17.0	0	12.0	14.9	24.1	50.7	25.5	22.4
CogVLM-chat [102]	40.8	0	0	1.6	18.6	10.9	60.2	26.8	19.9
Qwen-VL [103]	35.9	4.2	0	38.7	28.5	13.8	60.1	16.9	24.8
Qwen-VL-chat [103]	34.1	12.6	0.1	42.6	19.5	18.4	58.3	20.3	25.7
Qwen2-Vl-7B [6]	47.0	**42.0**	1.5	**90.2**	13.7	36.4	71.1	36.6	42.3
Qwen2.5-VL-7B [6]	51.5	24.5	3.1	64.8	13.1	53.3	78.6	45.5	41.8
InternVL2-8B [105]	43.0	21.6	0	70.2	19.2	35.6	65.9	33.6	36.1
InternVL2-26B [105]	56.0	21.2	0	80.5	23.9	40.3	72.1	40.7	41.8
InternVL2.5-8B [7]	48.9	21.2	0	82.1	20.3	41.2	67.8	42.3	40.5
InternVL2.5-26B [7]	53.5	21.4	0	84.0	21.4	51.5	67.5	41.5	42.6
InternVL3-8B [7]	49.7	22.3	0.2	86.8	22.4	57.0	70.7	53.0	45.3
InternVL3-14B [7]	55.8	24.5	2.1	89.3	21.0	59.5	72.0	50.0	46.8
Deepseek-VL-7B [106]	33.5	13.7	0	19.1	11.7	24.8	60.5	32.5	24.5
Deepseek-VL2-Small [107]	56.6	23.7	0	86.4	18.9	30.6	72.2	39.5	41.0
MiniCPM-V-2.6 [108]	52.2	18.6	0.3	45.8	19.6	20.9	68.9	37.3	33.0
MiniCPM-o-2.6 [108]	54.1	24.7	0.3	74.4	17.6	39.2	75.7	47.0	41.6
GLM-4v-9B [109]	52.7	20.6	0	79.4	15.9	21.5	74.7	32.0	37.1
VILA1.5-8B [110]	36.0	14.5	0	26.0	17.4	20.3	44.7	27.0	23.2
LLaVAR [70]	13.8	0	0	8.3	15.2	4.4	42.4	15.0	12.4
UReader [111]	20.9	0	0	0	20.7	11.3	39.0	20.8	14.1
DocOwl2 [112]	25.4	7.5	0	47.1	26.2	8.3	52.8	19.5	23.4
Yi-VL-6B [113]	31.1	4.0	0	23.4	22.5	18.1	43.0	15.5	19.7
Janus-1.3B [114]	32.6	0	0	0.3	13.0	18.4	32.1	17.9	14.3
Eagle-X5-7B [115]	34.6	18.5	0	9.7	18.5	24.0	63.1	37.0	25.7
Idefics3-8B [116]	37.4	13.0	0	28.9	19.4	21.1	65.4	21.8	26.0
Phi-4-MultiModal [117]	58.4	19.0	0	53.5	**38.7**	28.7	66.8	39.8	38.1
SAIL-VL-1.6-8B [118]	56.7	24.1	2.2	79.3	22.8	45.4	69.2	45.3	43.1
Kimi-VL-A3B-16B [119]	49.1	13.5	0	28.8	21.9	37.6	69.4	36.2	32.1
Ovis1.6-3B [120]	48.5	19.5	0	69.2	20.7	22.1	74.6	49.5	38.0
Ovis2-8B [120]	54.2	20.9	0	83.6	24.2	54.7	74.1	57.3	46.1
Closed-source LMMs									
GPT-4o [8]	58.6	23.4	0	87.4	23.1	51.6	74.4	**62.3**	47.6
GPT-4o-mini [121]	55.3	21.8	0	85.4	20.6	45.2	75.5	49.0	44.1
Gemini1.5-Pro [9]	**59.1**	41.2	**6.6**	89.5	22.4	54.7	**78.8**	60.3	**51.6**
Claude3.5-sonnet [122]	52.9	24.9	2.5	86.9	23.8	**61.4**	74.4	53.0	47.5
Step-1V [10]	56.7	27.4	2.6	86.3	33.3	42.6	76.6	48.7	46.8

Table 7: **Evaluation of existing LMMs on English tasks of OCRBench v2's private data**.

A.11 Samples for LMMs' Limitations

Fig. 15 to Fig. 17 provide examples corresponding to the findings discussed in Sec. 5.3 of the main text, which show error results of GPT-4o [8], Monkey [94], and Qwen2VL-8B on various tasks in *OCRBench v2*. These examples highlight the current limitations of LLMs on OCR tasks. For instance, LLMs exhibit poor recognition of less frequently encountered texts, struggle to accurately locate text in tasks involving text and coordinates, and demonstrate insufficient perception of text in complex layouts such as rotated texts. Additionally, their logical reasoning abilities are limited when addressing mathematical problems, and their analysis of complex elements in charts remains weak. These are the capabilities of LLMs in OCR tasks that require further improvement.

A.12 Broader Impacts

Our benchmark aims to enhance the evaluation of LMMs in text-oriented visual comprehension tasks. By establishing comprehensive benchmarks that reveal deficiencies in models' OCR capabilities, we provide insights for improving model performance. This advancement will elevate processing efficiency across scenarios such as document automation, assisted reading tools, and complex layout analysis, thereby benefiting applications in domains like healthcare and education. However, enhanced OCR functionality also introduces risks of misuse, including unauthorized extraction of sensitive information from images, surveillance-related applications, or generation of forged documents. To mitigate these risks, we restrict the use of this benchmark solely to research purposes and urge the community to prioritize privacy and fairness considerations in future model development.

Method	LLM Size	Recognition	Extraction	Parsing	Understanding	Reasoning	Average
Open-source LMMs							
LLaVA-Next-8B [92]	8B	2.8	0.9	14.9	20.0	7.4	9.2
LLaVA-OV-7B [93]	8B	5.4	13.6	20.3	34.0	13.6	17.4
Monkey [94]	8B	1.5	28.4	29.1	40.0	8.3	21.5
TextMonkey [95]	8B	10.5	15.2	30.2	44.0	7.6	21.5
XComposer2-4KHD [96]	7B	12.9	38.6	37.5	60.0	13.1	32.4
Molmo-7B [97]	8B	3.4	29.8	6.6	24.0	11.1	15.0
Cambrian-1-8B [98]	8B	2.4	19.8	26.7	36.0	7.6	18.5
Pixtral-12B [99]	12B	6.2	22.3	11.4	26.0	14.0	16.0
EMU2-chat [100]	37B	1.2	3.0	29.3	4.0	3.6	8.2
mPLUG-Owl3 [101]	8B	1.6	27.4	27.3	16.0	10.0	16.5
CogVLM-chat [102]	7B	2.4	16.2	22.5	20.0	3.1	12.8
Qwen-VL [103]	8B	4.3	0	30.6	38.0	5.1	15.6
Qwen-VL-chat [103]	8B	9.1	3.6	18.9	44.0	7.1	16.5
Qwen2-Vl-7B [6]	7B	23.7	63.5	27.9	80.0	28.5	44.7
Qwen2.5-VL-7B [6]	8B	24.4	**78.9**	33.1	82.0	29.0	49.5
InternVL2-8B [105]	8B	35.2	42.8	26.1	78.0	24.4	41.3
InternVL2-26B [105]	26B	20.4	50.7	29.0	76.0	14.5	38.1
InternVL2.5-8B [7]	8B	42.8	47.9	27.3	80.0	23.5	44.3
InternVL2.5-26B [7]	26B	40.2	42.7	25.6	74.0	27.0	41.9
InternVL3-8B [7]	8B	57.7	55.8	29.9	72.0	29.4	49.0
InternVL3-14B [7]	14B	62.1	59.5	33.2	80.0	29.2	52.8
Deepseek-VL-7B [106]	7B	3.2	14.7	10.7	30.0	9.8	13.7
DeepSeek-VL2-Small [107]	16B	51.6	56.3	27.8	79.6	25.3	48.1
MiniCPM-V-2.6 [108]	8B	53.1	53.2	32.8	76.0	23.4	47.7
MiniCPM-o-2.6 [108]	7B	54.0	62.4	24.1	68.0	29.8	47.7
GLM-4v-9B [109]	9B	60.6	65.2	32.4	82.0	18.2	51.7
VILA1.5-8B [110]	8B	1.4	9.1	22.2	16.0	6.4	11.0
LLaVAR [70]	13B	2.2	2.0	27.1	10.0	1.9	8.6
UReader [111]	7B	0.3	2.0	28.1	12.0	2.4	9.0
DocOwl2 [112]	7B	1.0	17.8	29.4	20.0	3.9	14.4
Yi-VL-6B [113]	6B	1.6	6.4	28.8	10.0	5.3	10.4
Janus-1.3B [114]	1.3B	4.1	2.2	10.4	14.0	6.7	7.5
Eagle-X5-7B [115]	8B	1.9	16.1	13.6	22.0	8.1	12.3
Idefics3-8B [116]	8B	2.9	29.0	12.3	26.0	7.9	15.6
Phi-4-MultiModal [117]	5.6B	30.5	40.5	42.7	56.0	16.9	37.3
SAIL-VL-1.6-8B [118]	8B	35.8	41.5	35.7	76.0	23.9	42.6
Kimi-VL-A3B-16B [119]	16B	54.0	71.1	32.5	**84.0**	28.7	54.1
Ovis1.6-3B [120]	3B	22.5	33.3	31.5	54.0	17.0	31.7
Ovis2-8B [120]	7B	61.0	67.7	**43.6**	82.0	25.6	**56.0**
Closed-source LMMs							
GPT-4o [8]	-	41.7	52.1	29.0	76.0	29.4	45.7
GPT-4o-mini [121]	-	20.0	53.6	27.9	66.0	19.6	37.4
Gemini1.5-Pro [9]	-	**71.4**	63.8	30.5	82.0	29.9	55.5
Claude3.5-sonnet [122]	-	34.2	62.5	35.2	78.0	**32.2**	48.4
Step-1V [10]	-	65.2	64.9	33.1	78.0	25.5	53.4

Table 8: **Evaluation of existing LMMs on Chinese tasks of OCRBench v2's private data**.

Method	Resolition	Recognition	Referring	Spotting	Extraction	Parsing	Calculation	Understanding	Reasoning	Average
InternVL2-8B [105]	448	47.3	19.1	0.1	52.8	**27.3**	25.4	61.1	49.1	35.3
	896	48.7	23.0	**0.5**	**66.2**	26.2	25.9	73.2	51.9	39.4
	dynamic	**49.9**	**23.1**	**0.5**	65.2	24.8	**26.7**	**73.5**	**52.9**	**39.6**

Table 9: Evaluation of InternVL2-8B with different resolution settings on the English tasks of OCRBench v2's public data.

Method	Recognition	Referring	Spotting	Extraction	Parsing	Calculation	Understanding	Reasoning	Average
Qwen2-VL-7B [6]	**72.1**	47.9	17.5	**82.5**	25.5	25.4	**78.4**	**61.5**	51.4
Qwen2-VL-7B+OCR	69.8	**50.4**	**20.1**	79.1	**29.4**	28.0	77.7	60.0	**51.8**
Qwen2.5-8B+OCR	28.6	13.8	0	45.9	24.2	**31.3**	61.1	40.5	30.7

Table 10: Evaluation of Qwen2-VL-7B and Qwen2.5-7B with pre-provided OCR information on English tasks of OCRBench v2's public data.

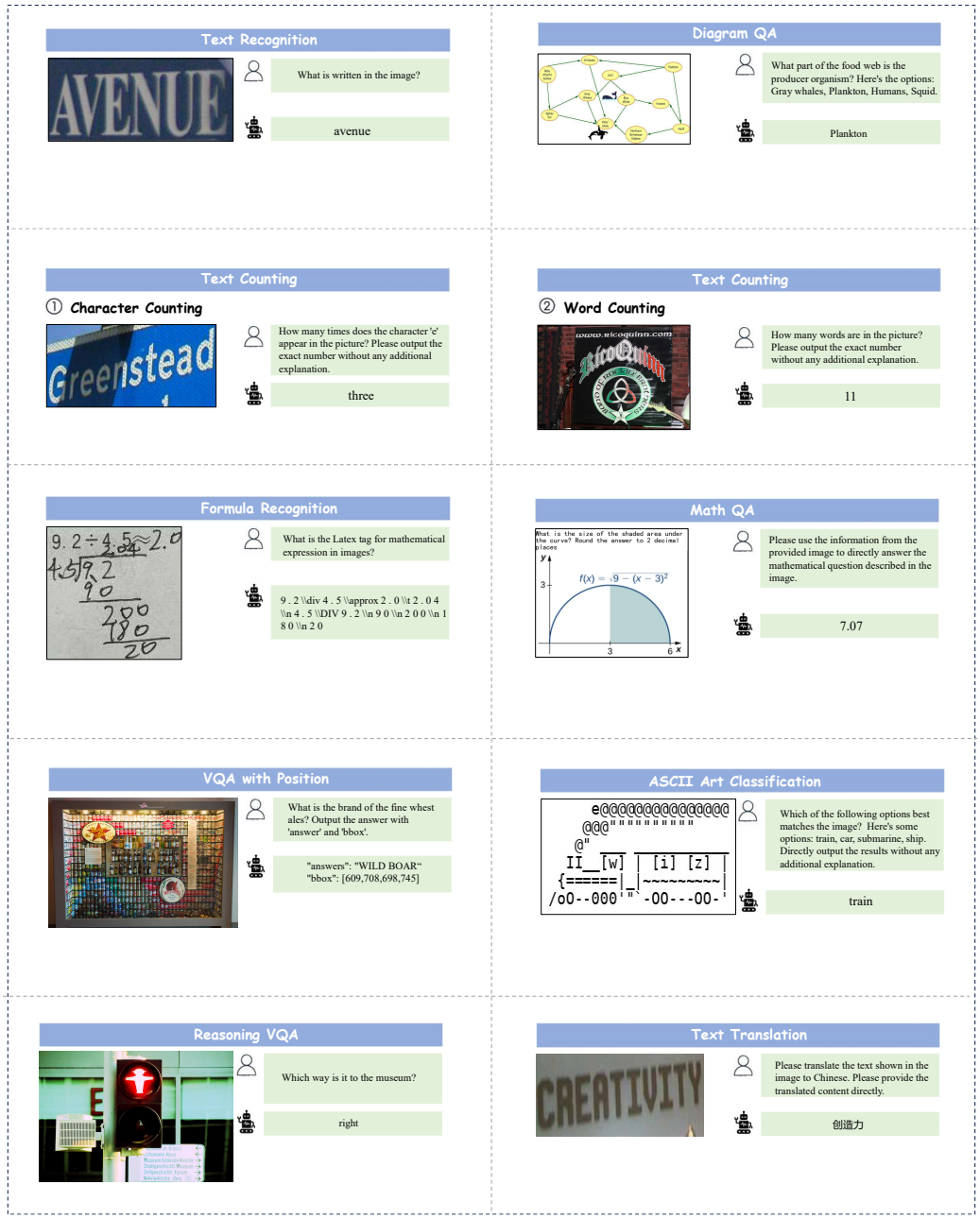


Figure 6: Samples for each task.

A.13 Limitations

One challenge we encountered is that LMMs sometimes produce responses that deviate from the given instructions, making it difficult to extract the desired answers. In future work, we plan to develop a more objective assessment framework to address this issue.

Another limitation arises when evaluating commercial LMMs, as some models occasionally refuse to answer certain questions due to safety filters or unclear content policies. This can lead to incomplete or biased performance assessments compared to open-source models that do not exhibit such behavior.

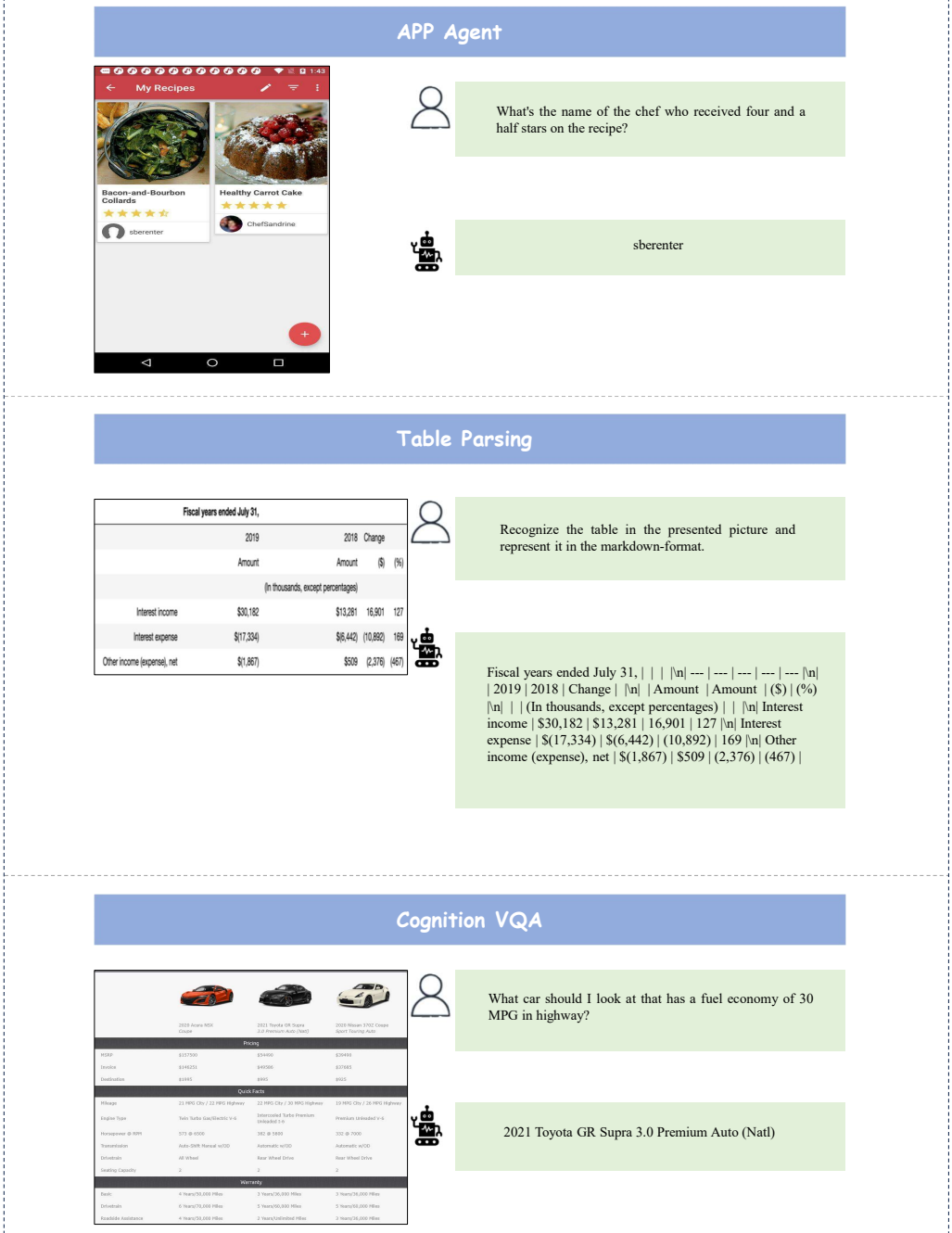
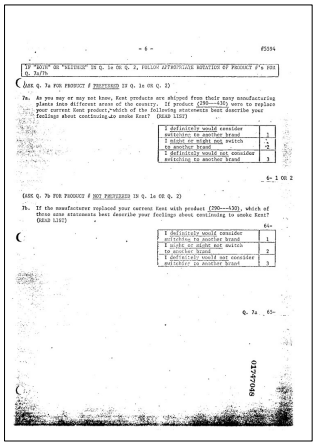


Figure 7: Samples for each task.

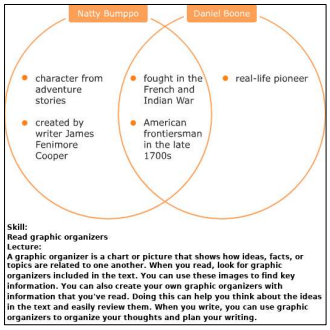
Document Classification



What type of document is the image? Here is the options: letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, memo. Directly output the results without any additional explanation.

Questionnaire

Science QA

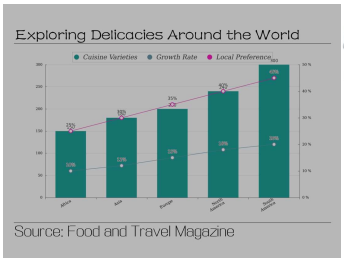


Based on the information from the provided image, answer the following question: Based on the Venn diagram, what do Natty Bumppo and Daniel Boone have in common? Here are the choices: 'Both fought in the French and Indian War.', 'Both were created by writer James Fenimore Cooper.'. Please output the answer directly.

Both fought in the French and Indian War.

Figure 8: Samples for each task.

Chart Parsing



Convert the key information in the chart into a nested python dict.

```
{'title': 'Exploring Delicacies Around the World',
  'source': 'Source: Food and Travel Magazine',
  'x_title': 'None',
  'y_title': ['None', 'None'],
  'values': {
    'Cuisine Varieties': {
      'Africa': '150',
      'Asia': '180',
      'Europe': '200',
      'North America': '240',
      'South America': '300'
    },
    'Growth Rate': {
      'Africa': '10.0%',
       'Asia': '12.0%',
      'Europe': '15.0%',
      'North America': '18.0%',
      'South America': '20.0%'
    },
    'Local Preference': {
      'Africa': '25.0%',
      'Asia': '30.0%',
      'Europe': '35.0%',
      'North America': '40.0%',
      'South America': '45.0%'}}}
```

Key Information Extraction

Find out the value of 'Serving Size', 'Calories/Energy per serving', 'Total Fat per serving', 'Total Fat % daily value/intake', 'Sodium per serving', 'Sodium % daily value/intake', 'Carbohydrate per serving', 'Carbohydrate % daily value/intake', 'Protein per serving' stated in the image. Extract the required information using the text in the image directly, and return the result in a dict with keys 'SS', 'CE-PS', 'TF-PS', 'TF-D', 'SO-PS', 'SO-D', 'CAR-PS', 'CAR-D', 'PRO-PS'.

```
{
  'SS': ['1/2 cup (122g)'],
  'CE-PS': ['90'],
  'TF-PS': ['0g'],
  'TF-D': ['0%'],
  'SO-PS': ['0mg'],
  'SO-D': ['0%'],
  'CAR-PS': ['21g'],
  'CAR-D': ['7%'],
  'PRO-PS': ['less than 1g']
}
```

Figure 9: Samples for each task.

Full Page OCR

Read all the text in the image. Directly output the content and split the texts with space.

SURREALISM, ABSTRACTION\n& THE UNCONSCIOUS\nIn the 1920s, while a student at Yale, Rothko took\nphilosophy and psychology classes that led to his\ninterest in the unconscious and eventually to Sigmund\nFreud's The Interpretation of Dreams. Then, in the '40s,\nlike many American artists, he came under the influence\nof European Surrealist artists and writers living in New\nYork. Many artists also found inspiration in the Museum\nof Modern Art's influential 1936 exhibition, Fantastic Art,\nDada and Surrealism.\nBy the mid-1940s, Rothko loosened up his technique,\ninspired by the Surrealist method of automatism\n(\"automatic\" drawing or writing). Unlike some\nSurrealist artists, such as Salvador Dali, who \"pictured\"\nunconscious dreams in paintings, Surrealists using\nautomatism tried to access the unconscious by letting\nthe brush meander freely without planning or control.\nRothko experimented with the fluidity of watercolor and\nsoon realized he could achieve similar luminous effects\nin oil paint by diluting the pigment and applying color in\nthin washes, one on top of another. Rothko's imagery also\nchanged. Many works suggest paleontology and geology\nand evoke a vision of primordial life. Water seems to be\na primal element in which biomorphic shapes proliferate.\nSome compositions include stacked horizontal zones that\nmay stand for layers of the unconscious.

Text Spotting

Spotting all the text in the image with word-level. Output the normalized coordinates of the left-top and right-bottom corners of the bounding box and the text content. The coordinates should be normalized ranging from 0 to 1000 by the image width and height.\nYour answer should be in the following format: [(x1, y1, x2, y2, text content), (x1, y1, x2, y2, text content)...] # The normalized coordinates and the content of the text in the image.

543, 770, 589, 794, 49-0223A,
545, 731, 580, 760, 502,
309, 594, 666, 641, YELLOWSTONE,
417, 160, 554, 198, TOUR

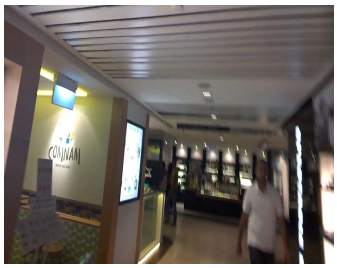
Figure 10: Samples for each task.

Fine-grained Text Recognition

Recognize the text within the [192, 223, 332, 346] of the image. The coordinates have been normalized ranging from 0 to 1000 by the image width and height.

DIOS LE ABRE CAMINO\\n
AL HOMBRE\\n
QUE SABE A DONDE VA

Text Grounding



Where is the region of the text 'COMNAM'? Output the normalized coordinates of the left-top and right-bottom corners of the bounding box. The coordinates should be normalized ranging from 0 to 1000 by the image width and height.
Your answer should be in the following format:
(x1, y1, x2, y2) # x1, y1, x2, y2 are the normalized coordinates of the bounding box.

[126,537,248,624]

Key Information Mapping

NUTRITION INFORMATION

Servings per package: 5 Serving size: 200ml (1 cup)

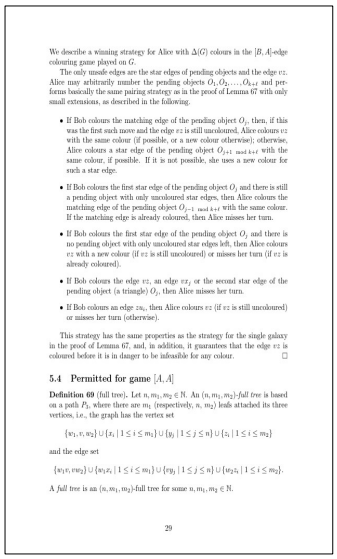
	Per Serving	Per 100ml
Energy	98 kcal (412 kJ)	49 kcal (206 kJ)
Protein	10.0 g	5.0 g
Total Fat	1.6 g	0.8 g
- Saturated Fat	1.0 g	0.5 g
- Trans Fat	0.0 g	0.0 g
Cholesterol	6 mg	3 mg
Carbohydrate	11.0 g	5.5 g
- Total Sugar	9.0 g	4.5 g
- Added Sugar	0.0 g	0.0 g
- Lactose	2.2 g	1.1 g
Dietary Fibre	0.0 g	0.0 g
Sodium	164 mg	82 mg
Calcium	400.0 mg	200.0 mg
Vitamin A	372.0 mcg	186.0 mcg
Thiamine (Vitamin B1)	1.0 mg	0.5 mg
Niacin (Vitamin B3)	11.0 mg	5.5 mg
Pantothenic Acid (Vitamin B5)	7.0 mg	3.5 mg
Pyridoxine (Vitamin B6)	1.8 mg	0.9 mg
Vitamin C	24.0 mg	12.0 mg
Vitamin D3	2.4 mcg	1.2 mcg
Vitamin E	6.0 mg	3.0 mg
Vitamin K1	27.0 mcg	13.5 mcg

According to the information in the image, please pair the corresponding keys and values below: Keys that need to be paired are 'Serving Size', 'Calories/Energy per 100g/ml', 'Carbohydrate per serving', 'Protein per 100g/ml', 'Total Fat per serving', 'Carbohydrate per 100g/ml', 'Total Fat per 100g/ml', 'Protein per serving'. Values that need to be paired are '0.8 g', '11.0 g', '200ml (1 cup)', '10.0 g', '1.6 g', '49 kcal(206 kJ)', '5.0 g', '5.5 g'.

{"Calories/Energy per 100g/ml": "49 kcal(206 kJ)"
"Protein per serving": "10.0 g"
"Protein per 100g/ml": "5.0 g"
"Total Fat per serving": "1.6 g"
"Total Fat per 100g/ml": "0.8 g"
"Carbohydrate per serving": "11.0 g"
"Carbohydrate per 100g/ml": "5.5 g"
"Serving Size": "200ml (1 cup)"}

Figure 11: Samples for each task.

Document Parsing



We describe a winning strategy for Alice with $\Delta(G)$ colours in the $[B, A]$-edge colouring game played on G.

The only unsafe edges are the star edges of pending objects and the edge vz. Alice may arbitrarily number the pending objects $O_1, O_2, \ldots, O_{k+\ell}$ and performs basically the same pairing strategy as in the proof of Lemma 67 with only small extensions, as described in the following.

- If Bob colours the matching edge of the pending object O_j, then, if this was the first such move and the edge vz is still uncoloured, Alice colours vz with the same colour (if possible, or a new colour otherwise); otherwise, Alice colours a star edge of the pending object $O_{j+1 \bmod k+\ell}$ with the same colour, if possible. If it is not possible, she uses a new colour for such a star edge.
- If Bob colours the first star edge of the pending object O_j and there is still a pending object with only uncoloured star edges, then Alice colours the matching edge of the pending object $O_{j-1 \bmod k+\ell}$ with the same colour. If the matching edge is already coloured, then Alice misses her turn.
- If Bob colours the first star edge of the pending object O_j and there is no pending object with only uncoloured star edges left, then Alice colours vz with a new colour (if vz is still uncoloured) or misses her turn (if vz is already coloured).
- If Bob colours the edge vz, an edge vx_j or the second star edge of the pending object (a triangle) O_j, then Alice misses her turn.
- If Bob colours an edge zu_i, then Alice colours vz (if vz is still uncoloured) or misses her turn (otherwise).

This strategy has the same properties as the strategy for the single galaxy in the proof of Lemma 67, and, in addition, it guarantees that the edge vz is coloured before it is in danger to be infeasible for any colour. □

5.4 Permitted for game $[A, A]$

Definition 69 (full tree)**.** Let $n, m_1, m_2 \in \mathbb{N}$. An (n, m_1, m_2)-*full tree* is based on a path P_3, where there are m_1 (respectively, n, m_2) leafs attached its three vertices, i.e., the graph has the vertex set

$$\{w_1, v, w_2\} \cup \{x_i \mid 1 \leq i \leq m_1\} \cup \{y_j \mid 1 \leq j \leq n\} \cup \{z_i \mid 1 \leq i \leq m_2\}$$

and the edge set

$$\{w_1 v, v w_2\} \cup \{w_1 x_i \mid 1 \leq i \leq m_1\} \cup \{v y_j \mid 1 \leq j \leq n\} \cup \{w_2 z_i \mid 1 \leq i \leq m_2\}.$$

A *full tree* is an (n, m_1, m_2)-full tree for some $n, m_1, m_2 \in \mathbb{N}$.

29

Convert the privided document into markdown format.

We describe a winning strategy for Alice with \\(\\Delta(G)\\) colours in the \\([B,A]\\)-edge colouring game played on \\(G\\).\n\nThe only unsafe edges are the star edges of pending objects and the edge \\(vz\\). Alice may arbitrarily number the pending objects \\(O_{1},O_{2},\\ldots,O_{k+\\ell}\\) and performs basically the same pairing strategy as in the proof of Lemma 67 with only small extensions, as described in the following.\n\n* If Bob colours the matching edge of the pending object \\(O_{j}\\), then, if this was the first such move and the edge \\(vz\\) is still uncoloured, Alice colours \\(vz\\) with the same colour (if possible, or a new colour otherwise); otherwise, Alice colours a star edge of the pending object \\(O_{j+1\\mod{k+\\ell}}\\) with the same colour, if possible. If it is not possible, she uses a new colour for such a star edge.\n* If Bob colours the first star edge of the pending object \\(O_{j}\\) and there is still a pending object with only uncoloured star edges, then Alice colours the matching edge of the pending object \\(O_{j-1\\mod{k+\\ell}}\\) with the same colour. If the matching edge is already coloured, then Alice misses her turn.\n* If Bob colours the first star edge of the pending object \\(O_{j}\\) and there is no pending object with only uncoloured star edges left, then Alice colours \\(vz\\) with a new colour (if \\(vz\\) is still uncoloured) or misses her turn (if \\(vz\\) is already coloured).\n* If Bob colours the edge \\(vz\\), an edge \\(vx_{j}\\) or the second star edge of the pending object (a triangle) \\(O_{j}\\), then Alice misses her turn.\n* If Bob colours an edge \\(zu_{i}\\), then Alice colours \\(vz\\) (if \\(vz\\) is still uncoloured) or misses her turn (otherwise).\n\nThis strategy has the same properties as the strategy for the single galaxy in the proof of Lemma 67, and, in addition, it guarantees that the edge \\(vz\\) is coloured before it is in danger to be infeasible for any colour. \n\n### Permitted for game \\([A,a]\\)\n\n**Definition 69** (full tree).: Let \\(n,m_{1},m_{2}\\in\\mathbb{N}\\). An _\\((n,m_{1},m_{2})\\)-full tree_ is based on a path \\(P_{3}\\), where there are \\(m_{1}\\) (respectively, \\(n\\), \\(m_{2}\\)) leafs attached its three vertices, i.e., the graph has the vertex set\n\n\\[\\{w_{1},v,w_{2}\\}\\cup\\{x_{i}\\mid 1\\leq i\\leq m_{1}\\}\\cup\\{y_{j}\\mid 1\\leq j\\leq n\\}\\cup\\{z_{i}\\mid 1\\leq i\\leq m_{2}\\}\\]\n\nand the edge set\n\n\\[\\{w_{1}v,vw_{2}\\}\\cup\\{w_{1}x_{i}\\mid 1\\leq i\\leq m_{1}\\}\\cup\\{vy_{j}\\mid 1\\leq j\\leq n\\}\\cup\\{w_{2}z_{i}\\mid 1\\leq i\\leq m_{2}\\}.\\]\n\nA _full tree_ is an \\((n,m_{1},m_{2})\\)-full tree for some \\(n,m_{1},m_{2}\\in\\mathbb{N}\\).“

Figure 12: Samples for each task.

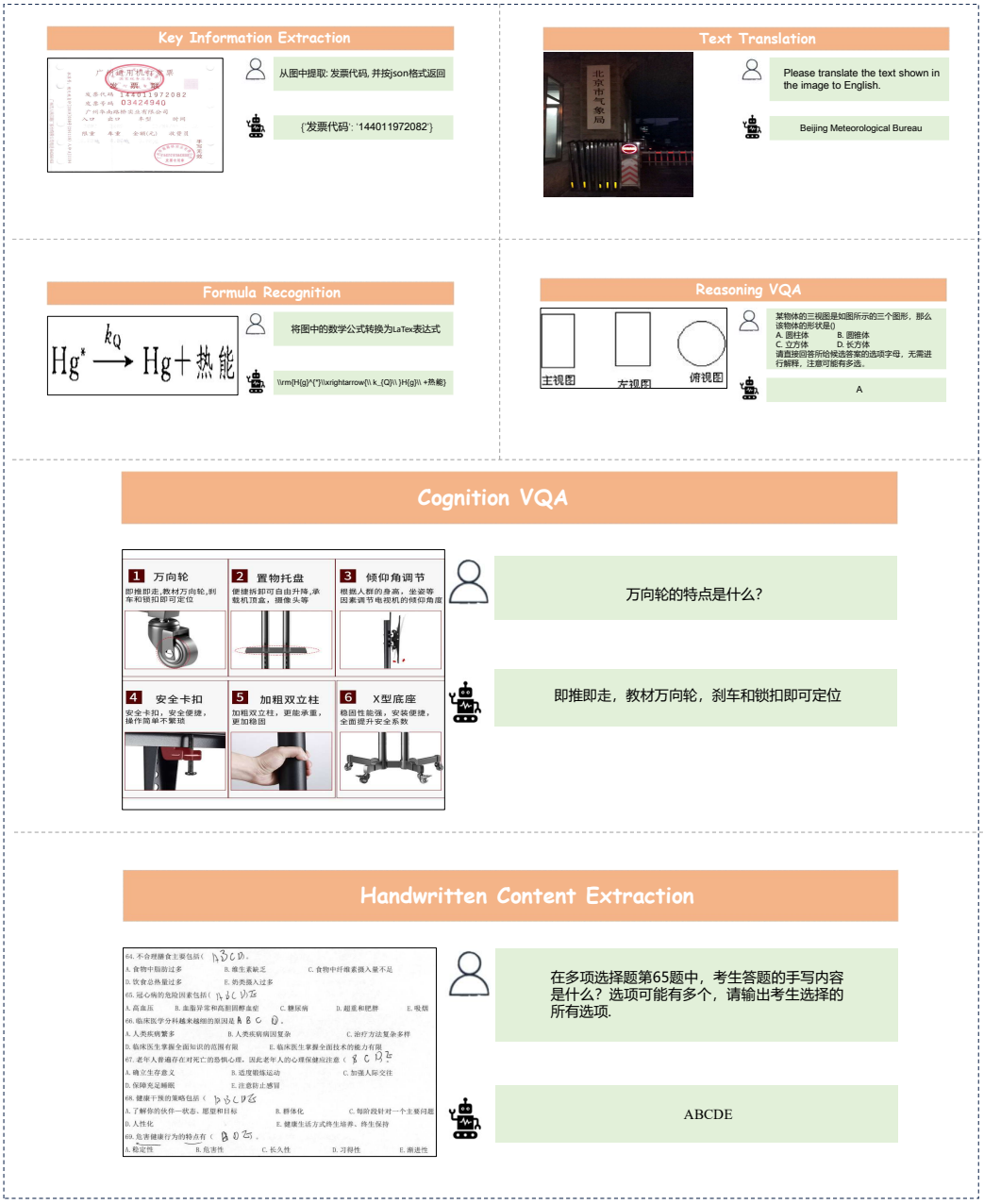
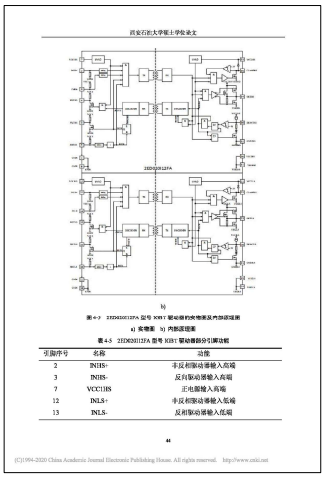


Figure 13: Samples for each task.

Document Parsing



Parse the document image in Markdown format.

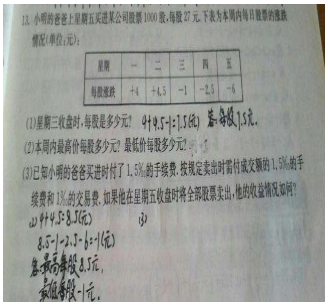
西安石油大学硕士学位论文 \n图4-5 2ED020I12FA型号IGBT驱动器的实物图及内部原理图 \na)实物图 b)内部原理图 \n表4-5 2ED020I12FA型号IGBT驱动器部分引脚功能 \n<table><tr><td>引脚序号</td><td>名称</td><td>功能</td></tr><tr><td>2</td><td>INHS+</td><td>非反相驱动器输入高端</td></tr><tr><td>3</td><td>INHS-</td><td>反向驱动器输入高端</td></tr><tr><td>7</td><td>VCC1HS</td><td>正电源输入高端</td></tr><tr><td>12</td><td>INLS+</td><td>非反相驱动器输入低端</td></tr><tr><td>13</td><td>INLS-</td><td>反相驱动器输入低端</td></tr></table> \n\n44 \n(C)1994-2020 China Academic Journal Electronic Publishing House.All rights reserved.http://www.cnki.net

Full Page OCR

Read all the text in the image .

着力打造上海国际旅游度假区，
提升上海旅游业的国际竞争力和影响力，
推动世界著名旅游城市建设。

Table Parsing



Parse the HTML-formatted structural information of all the tables in the image.

<table><tr><td> 星期</td><td> 一</td><td>二</td><td> 三</td><td> 四</td><td> 五</td></tr> <tr><td> 每股涨跌</td><td> +4</td><td> +4.5</td><td> -1</td><td> -2.5</td><td> -6</td> </tr></table>

Figure 14: Samples for each task.

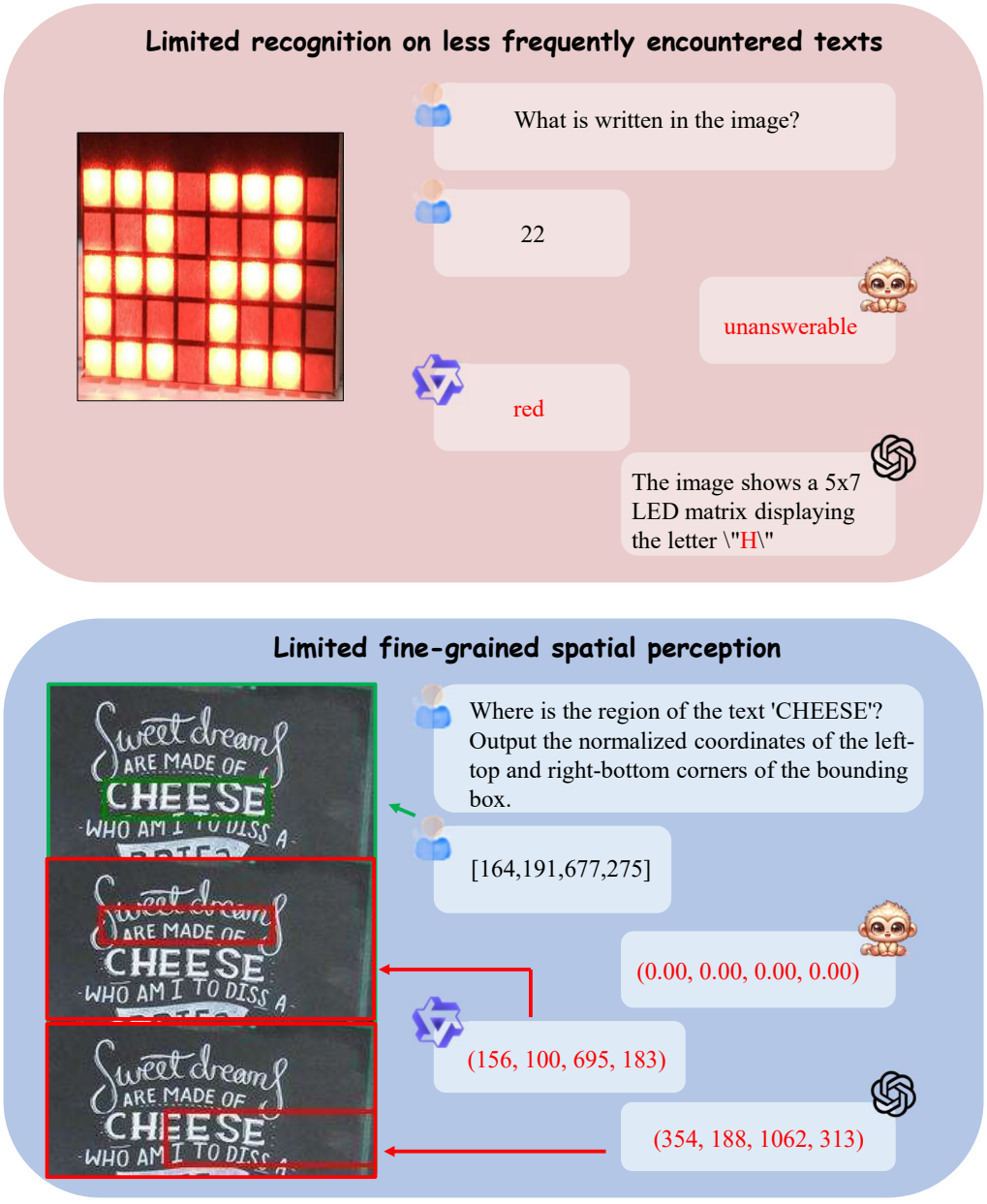


Figure 15: **Samples for LMM'S Limitations**. The portion of LLM's response marked in red is incorrect content, and the content in the red dashed box is missing information.

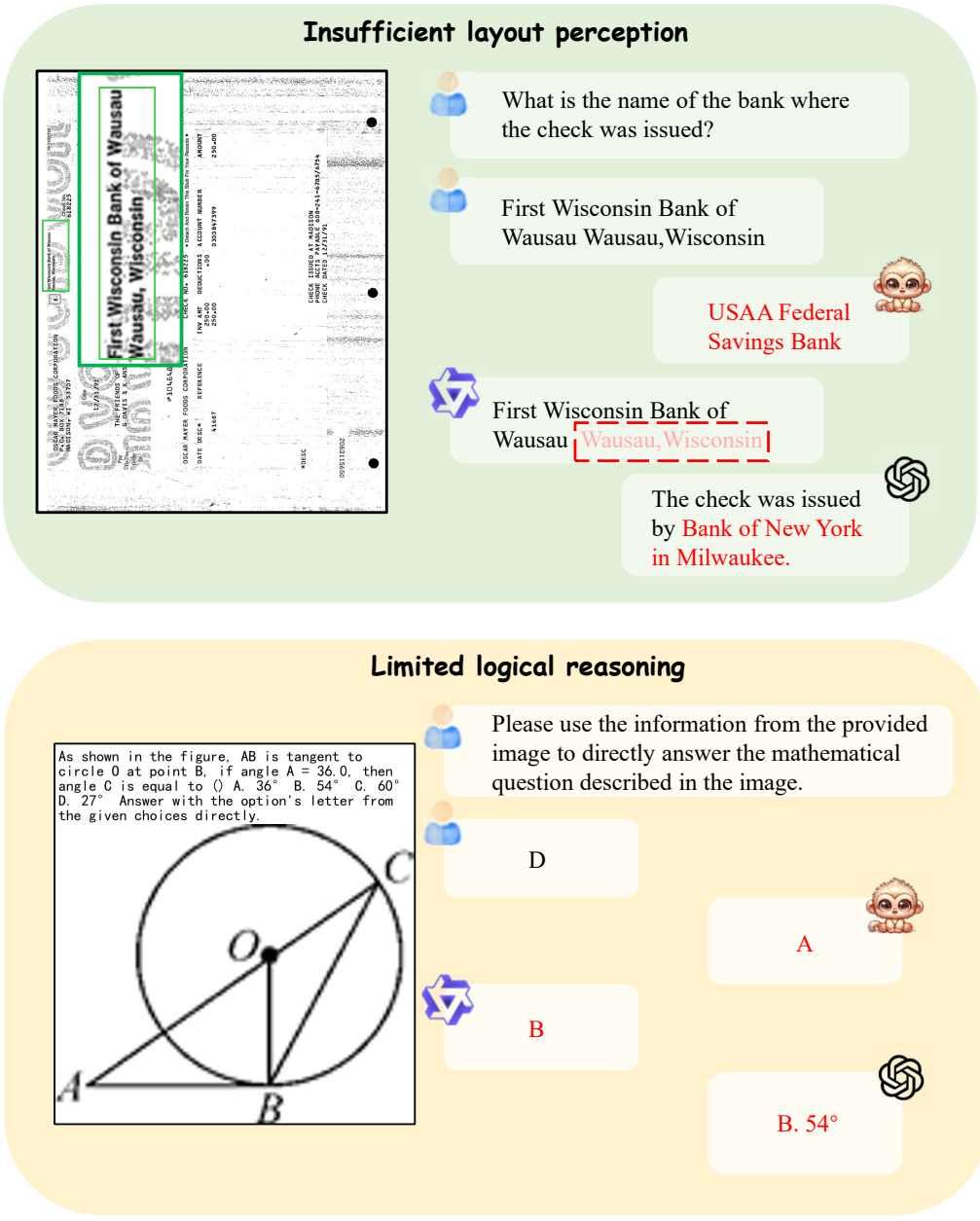


Figure 16: **Samples for LMM'S Limitations**. The portion of LLM's response marked in red is incorrect content, and the content in the red dashed box is missing information.

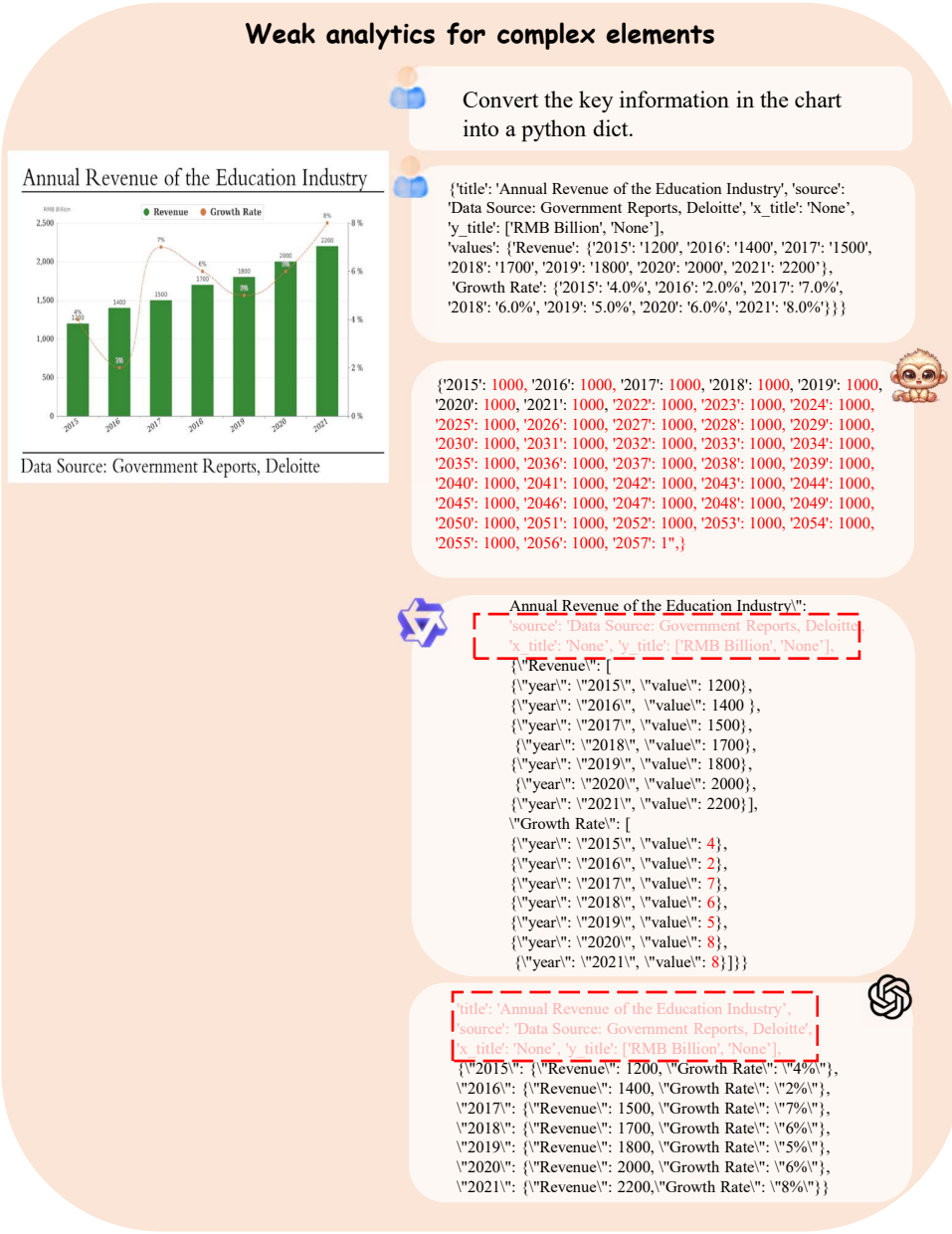


Figure 17: **Samples for LMM'S Limitations**. The portion of LLM's response marked in red is incorrect content, and the content in the red dashed box is missing information.

References

[1] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[2] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7098–7107.

[3] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2018.

[4] N. Lu, W. Yu, X. Qi, Y. Chen, P. Gong, R. Xiao, and X. Bai, "Master: Multi-aspect non-local network for scene text recognition," *Pattern Recognition*, vol. 117, p. 107980, 2021.

[5] Y. Du, Z. Chen, C. Jia, X. Yin, T. Zheng, C. Li, Y. Du, and Y. Jiang, "SVTR: scene text recognition with a single visual model," in *Proceedings of the International Joint Conference on Artificial Intelligence*, L. D. Raedt, Ed. ijcai.org, 2022, pp. 884–890. [Online]. Available: https://doi.org/10.24963/ijcai.2022/124

[6] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[7] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[8] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[9] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[10] StepFun, "Step-1V," https://www.stepfun.com/#step1v, 2024, accessed: 2024-12-29.

[11] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9809–9818.

[12] Y. Liu, C. Shen, L. Jin, T. He, P. Chen, C. Liu, and H. Chen, "Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8048–8064, 2021.

[13] X. Zhang, Y. Su, S. Tripathi, and Z. Tu, "Text spotting transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9519–9528.

[14] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proceedings of International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 2017, pp. 935–942.

[15] H. Wei, C. Liu, J. Chen, J. Wang, L. Kong, Y. Xu, Z. Ge, L. Zhao, J. Sun, Y. Peng *et al.*, "General ocr theory: Towards ocr-2.0 via a unified end-to-end model," *arXiv preprint arXiv:2409.01704*, 2024.

[16] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazàn, and L. P. de las Heras, "Icdar 2013 robust reading competition," in *Proceedings of International Conference on Document Analysis and Recognition*, 2013, pp. 1484–1493.

[17] C. Shi, C. Wang, B. Xiao, S. Gao, and J. Hu, "End-to-end scene text recognition using tree-structured models," *Pattern Recognition*, vol. 47, pp. 2853–2866, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:30201169

[18] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in *British Machine Vision Conference*, 2012.

[19] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *Proceedings of International Conference on Document Analysis and Recognition*. IEEE, 2015, pp. 1156–1160.

[20] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognition*, vol. 90, no. C, p. 337–345, Jun. 2019. [Online]. Available: https://doi.org/10.1016/j.patcog.2019.02.002

[21] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.

[22] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Systems with Applications*, vol. 41, pp. 8027–8048, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:15559857

[23] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," in *Proceedings of IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, 2013, pp. 569–576. [Online]. Available: https://doi.org/10.1109/ICCV.2013.76

[24] X. Xie, L. Fu, Z. Zhang, Z. Wang, and X. Bai, "Toward understanding wordart: Corner-guided transformer for scene text recognition," in *Proceedings of European Conference on Computer Vision*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 303–321.

[25] Y. Liu, Z. Li, B. Yang, C. Li, X. Yin, C.-l. Liu, L. Jin, and X. Bai, "On the hidden mystery of ocr in large multimodal models," *arXiv preprint arXiv:2305.07895*, 2023.

[26] U.-V. Marti and H. Bunke, "The iam-database: an english sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition*, vol. 5, pp. 39–46, 2002.

[27] M. Diem, S. Fiel, F. Kleber, R. Sablatnig, J. M. Saavedra, D. Contreras, J. M. Barrios, and L. S. Oliveira, "Proceedings of ieee international conference on frontiers in handwriting recognition," in *2014 14th International Conference on Frontiers in Handwriting Recognition*, 2014, pp. 779–784.

[28] Y. Wang, H. Xie, S. Fang, J. Wang, S. Zhu, and Y. Zhang, "From two to one: A new scene text recognizer with visual language modeling network," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 194–14 203.

[29] C. Liu, H. Wei, J. Chen, L. Kong, Z. Ge, Z. Zhu, L. Zhao, J. Sun, C. Han, and X. Zhang, "Focus Anywhere for Fine-grained Multi-page Document Understanding," *arXiv preprint arXiv:2405.14295*, 2024.

[30] L. Yuliang, J. Lianwen, Z. Shuaitao, and Z. Sheng, "Detecting curve text in the wild: New dataset and new solution," *arXiv preprint arXiv:1712.02170*, 2017.

[31] S. Long, S. Qin, D. Panteleev, A. Bissacco, Y. Fujii, and M. Raptis, "Towards end-to-end unified scene text detection and layout analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1049–1059.

[32] T.-L. Yuan, Z. Zhu, K. Xu, C.-J. Li, T.-J. Mu, and S.-M. Hu, "A large chinese text dataset in the wild," *Journal of Computer Science and Technology*, vol. 34, no. 3, pp. 509–521, 2019. [Online]. Available: https://jcst.ict.ac.cn/en/article/doi/10.1007/s11390-019-1923-y

[33] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, "Icdar2017 competition on reading chinese text in the wild (rctw-17)," in *Proceedings of International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 2017, pp. 1429–1434.

[34] X. Liu, R. Zhang, Y. Zhou, Q. Jiang, Q. Song, N. Li, K. Zhou, L. Wang, D. Wang, M. Liao, M. Yang, X. Bai, B. Shi, D. Karatzas, S. Lu, and C. V. Jawahar, "Icdar 2019 robust reading challenge on reading chinese text on signboard," 2019. [Online]. Available: https://arxiv.org/abs/1912.09641

[35] Y. Sun, J. Liu, W. Liu, J. Han, E. Ding, and J. Liu, "Chinese street view text: Large-scale chinese text reading with partially supervised learning," 2020. [Online]. Available: https://arxiv.org/abs/1909.07808

[36] H. Cheng, P. Zhang, S. Wu, J. Zhang, Q. Zhu, Z. Xie, J. Li, K. Ding, and L. Jin, "M^{6}doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis," 2023. [Online]. Available: https://arxiv.org/abs/2305.08719

[37] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8317–8326.

[38] B. Deka, Z. Huang, C. Franzen, J. Hibschman, D. Afergan, Y. Li, J. Nichols, and R. Kumar, "Rico: A mobile app dataset for building data-driven design applications," in *Proceedings of the 30th annual ACM symposium on user interface software and technology*, 2017, pp. 845–854.

[39] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "Funsd: A dataset for form understanding in noisy scanned documents," in *Proceedings of International Conference on Document Analysis and Recognition Workshops*, vol. 2. IEEE, 2019, pp. 1–6.

[40] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. Jawahar, "Icdar2019 competition on scanned receipt ocr and information extraction," in *Proceedings of International Conference on Document Analysis and Recognition*. IEEE, 2019, pp. 1516–1520.

[41] J. Kuang, W. Hua, D. Liang, M. Yang, D. Jiang, B. Ren, and X. Bai, "Visual information extraction in the wild: practical dataset and end-to-end solution," in *Proceedings of International Conference on Document Analysis and Recognition*. Springer, 2023, pp. 36–53.

[42] Y. Xu, T. Lv, L. Cui, G. Wang, Y. Lu, D. Florencio, C. Zhang, and F. Wei, "XFUND: A benchmark dataset for multilingual visually rich form understanding," in *Proceedings of Annual Meeting of the Association for Computational Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3224. [Online]. Available: https://aclanthology.org/2022.findings-acl.253

[43] W. Yu, C. Zhang, H. Cao, W. Hua, B. Li, H. Chen, M. Liu, M. Chen, J. Kuang, M. Cheng, Y. Du, S. Feng, X. Hu, P. Lyu, K. Yao, Y. Yu, Y. Liu, W. Che, E. Ding, C.-L. Liu, J. Luo, S. Yan, M. Zhang, D. Karatzas, X. Sun, J. Wang, and X. Bai, "Icdar 2023 competition on structured text extraction from visually-rich document images," 2023. [Online]. Available: https://arxiv.org/abs/2306.03287

[44] M. Zheng, X. Feng, Q. Si, Q. She, Z. Lin, W. Jiang, and W. Wang, "Multimodal Table Understanding," in *Proceedings of Annual Meeting of the Association for Computational Linguistics*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 9102–9124. [Online]. Available: https://doi.org/10.18653/v1/2024.acl-long.493

[45] L. Rujiao, W. Wen, X. Nan, G. Feiyu, Y. Zhibo, W. Yongpan, and X. Gui-Song, "Parsing table structures in the wild," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, October 2021.

[46] F. Yang, L. Hu, X. Liu, S. Huang, and Z. Gu, "A large-scale dataset for end-to-end table recognition in the wild," *Scientific Data*, vol. 10, no. 1, p. 110, 2023.

[47] J. Chen, L. Kong, H. Wei, C. Liu, Z. Ge, L. Zhao, J. Sun, C. Han, and X. Zhang, "Onechart: Purify the chart structural extraction via one auxiliary token," in *Proceedings of the ACM International Conference on Multimedia*, 2024, pp. 147–155.

[48] F. Liu, X. Wang, W. Yao, J. Chen, K. Song, S. Cho, Y. Yacoob, and D. Yu, "MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024, pp. 1287–1310.

[49] Y. Liang, Y. Zhang, C. Ma, Z. Zhang, Y. Zhao, L. Xiang, C. Zong, and Y. Zhou, "Document image machine translation with dynamic multi-pre-trained models assembling," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024, pp. 7077–7088.

[50] M. Mathew, D. Karatzas, and C. V. Jawahar, "Docvqa: A dataset for VQA on document images," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2021, pp. 2199–2208.

[51] Y. Yuan, X. Liu, W. Dikubab, H. Liu, Z. Ji, Z. Wu, and X. Bai, "Syntax-aware network for handwritten mathematical expression recognition," *arXiv preprint arXiv:2203.01601*, 2022.

[52] K. Wang, J. Pan, W. Shi, Z. Lu, M. Zhan, and H. Li, "Measuring multimodal mathematical reasoning with math-vision dataset," *CoRR*, vol. abs/2402.14804, 2024.

[53] W. Yang, Z. Li, D. Peng, L. Jin, M. He, and C. Yao, "Read ten lines at one glance: Line-aware semi-autoregressive transformer for multi-line handwritten mathematical expression recognition," in *Proceedings of the ACM International Conference on Multimedia*, A. El-Saddik, T. Mei, R. Cucchiara, M. Bertini, D. P. T. Vallejo, P. K. Atrey, and M. S. Hossain, Eds. ACM, 2023, pp. 2066–2077.

[54] P. Gervais, A. Fadeeva, and A. Maksai, "Mathwriting: A dataset for handwritten mathematical expression recognition," *CoRR*, vol. abs/2404.10690, 2024.

[55] L. Ding, M. Zhao, F. Yin, S. Zeng, and C.-L. Liu, "A large-scale database for chemical structure recognition and preliminary evaluation," in *Proceedings of the International Conference on Pattern Recognition*, 2022, pp. 1464–1470.

[56] D. Saxton, E. Grefenstette, F. Hill, and P. Kohli, "Analysing mathematical reasoning abilities of neural models," in *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2019.

[57] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K. Chang, Y. Qiao, P. Gao, and H. Li, "MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems?" in *Proceedings of European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., vol. 15066. Springer, 2024, pp. 169–186.

[58] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K. Chang, M. Galley, and J. Gao, "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts," in *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2024.

[59] X. Wang, Y. Liu, C. Shen, C. C. Ng, C. Luo, L. Jin, C. S. Chan, A. v. d. Hengel, and L. Wang, "On the general value of evidence, and bilingual scene-text visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 126–10 135.

[60] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, "Ocr-vqa: Visual question answering by reading text in images," in *Proceedings of International Conference on Document Analysis and Recognition*. IEEE, 2019, pp. 947–952.

[61] A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C. Jawahar, and D. Karatzas, "Scene text visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4291–4301.

[62] H. Feng, W. Zhou, J. Deng, Y. Wang, and H. Li, "Geometric representation learning for document image rectification," in *Proceedings of European Conference on Computer Vision*. Springer, 2022, pp. 475–492.

[63] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "Chartqa: A benchmark for question answering about charts with visual and logical reasoning," *arXiv preprint arXiv:2203.10244*, 2022.

[64] K. Kafle, S. Cohen, B. Price, and C. Kanan, "Dvqa: Understanding data visualizations via question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[65] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar, "Plotqa: Reasoning over scientific plots," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, March 2020.

[66] M. Mathew, V. Bagal, R. P. Tito, D. Karatzas, E. Valveny, and C. V. Jawahar, "Infographicvqa," *CoRR*, vol. abs/2104.12756, 2021. [Online]. Available: https://arxiv.org/abs/2104.12756

[67] X. Zhong, E. ShafieiBavani, and A. J. Yepes, "Image-based table recognition: data, model, and evaluation," *arXiv preprint arXiv:1911.10683*, 2019.

[68] P. Pasupat and P. Liang, "Compositional semantic parsing on semi-structured tables," in *Proceedings of Annual Meeting of the Association for Computational Linguistics*, C. Zong and M. Strube, Eds. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1470–1480. [Online]. Available: https://aclanthology.org/P15-1142

[69] S. Park, S. Shin, B. Lee, J. Lee, J. Surh, M. Seo, and H. Lee, "Cord: a consolidated receipt dataset for post-ocr parsing," in *Advances in Neural Information Processing Systems Workshop*, 2019.

[70] Y. Zhang, R. Zhang, J. Gu, Y. Zhou, N. Lipka, D. Yang, and T. Sun, "Llavar: Enhanced visual instruction tuning for text-rich image understanding," *arXiv preprint arXiv:2306.17107*, 2023.

[71] X. Chen, Z. Zhao, L. Chen, D. Zhang, J. Ji, A. Luo, Y. Xiong, and K. Yu, "Websrc: A dataset for web-based structural reading comprehension," *arXiv preprint arXiv:2101.09465*, 2021.

[72] X. Zhong, J. Tang, and A. Jimeno-Yepes, "Publaynet: Largest dataset ever for document layout analysis," in *Proceedings of International Conference on Document Analysis and Recognition*. IEEE, 2019, pp. 1015–1022.

[73] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *Proceedings of International Conference on Document Analysis and Recognition*, 2015.

[74] G. Baechler, S. Sunkara, M. Wang, F. Zubach, H. Mansoor, V. Etter, V. Cărbune, J. Lin, J. Chen, and A. Sharma, "Screenai: A vision-language model for ui and infographics understanding," 2024.

[75] R. Tanaka, K. Nishida, K. Nishida, T. Hasegawa, I. Saito, and K. Saito, "Slidevqa: A dataset for document visual question answering on multiple images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[76] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi, "A diagram is worth a dozen images," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 235–251.

[77] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi, "Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4999–5007.

[78] S. Lee, B. Lai, F. Ryan, B. Boote, and J. M. Rehg, "Modeling multimodal social interactions: New challenges and baselines with densely aligned representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 585–14 595.

[79] G. Zhang, X. Du, B. Chen, Y. Liang, T. Luo, T. Zheng, K. Zhu, Y. Cheng, C. Xu, S. Guo, H. Zhang, X. Qu, J. Wang, R. Yuan, Y. Li, Z. Wang, Y. Liu, Y.-H. Tsai, F. Zhang, C. Lin, W. Huang, and J. Fu, "Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark," 2024. [Online]. Available: https://arxiv.org/abs/2401.11944

[80] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022.

[81] X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun *et al.*, "Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark," *arXiv preprint arXiv:2409.02813*, 2024.

[82] Q. Jia, X. Yue, S. Huang, Z. Qin, Y. Liu, B. Y. Lin, and Y. You, "Visual perception in text strings," *arXiv preprint arXiv:2410.01733*, 2024.

[83] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1083–1090.

[84] M. He, Y. Liu, Z. Yang, S. Zhang, C. Luo, F. Gao, Q. Zheng, Y. Wang, X. Zhang, and L. Jin, "Icpr2018 contest on robust reading for multi-type web images," in *Proceedings of the International Conference on Pattern Recognition*, 2018, pp. 7–12.

[85] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, "Icdar2017 competition on reading chinese text in the wild (rctw-17)," 2018. [Online]. Available: https://arxiv.org/abs/1708.09585

[86] J. Tang, Z. Yang, Y. Wang, Q. Zheng, Y. Xu, and X. Bai, "Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping," *Pattern Recognition*, vol. 96, p. 106954, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320319302511

[87] C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo, Z. Ni, C. Fang, S. Zhang, J. Han, E. Ding *et al.*, "Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art," in *Proceedings of International Conference on Document Analysis and Recognition*. IEEE, 2019, pp. 1571–1576.

[88] X. Zheng, D. Burdick, L. Popa, X. Zhong, and N. X. R. Wang, "Global table extractor (GTE): A framework for joint table identification and cell structure recognition using visual context," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2021, pp. 697–706. [Online]. Available: https://doi.org/10.1109/WACV48630.2021.00074

[89] X. Zhong, E. ShafieiBavani, and A. Jimeno Yepes, "Image-based table recognition: data, model, and evaluation," in *Proceedings of European Conference on Computer Vision*. Springer, 2020, pp. 564–580.

[90] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[91] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[92] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," 2024.

[93] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[94] Z. Li, B. Yang, Q. Liu, Z. Ma, S. Zhang, J. Yang, Y. Sun, Y. Liu, and X. Bai, "Monkey: Image resolution and text label are important things for large multi-modal models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 763–26 773.

[95] Y. Liu, B. Yang, Q. Liu, Z. Li, Z. Ma, S. Zhang, and X. Bai, "Textmonkey: An ocr-free large multimodal model for understanding document," *arXiv preprint arXiv:2403.04473*, 2024.

[96] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, S. Zhang, H. Duan, W. Zhang, Y. Li *et al.*, "Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd," *arXiv preprint arXiv:2404.06512*, 2024.

[97] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini *et al.*, "Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models," *arXiv preprint arXiv:2409.17146*, 2024.

[98] S. Tong, E. L. Brown II, P. Wu, S. Woo, A. J. IYER, S. C. Akula, S. Yang, J. Yang, M. Middepogu, Z. Wang *et al.*, "Cambrian-1: A fully open, vision-centric exploration of multimodal llms," in *Advances in Neural Information Processing Systems*, 2024.

[99] P. Agrawal, S. Antoniak, E. B. Hanna, B. Bout, D. Chaplot, J. Chudnovsky, D. Costa, B. De Monicault, S. Garg, T. Gervet *et al.*, "Pixtral 12b," *arXiv preprint arXiv:2410.07073*, 2024.

[100] Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Y. Wang, Y. Rao, J. Liu, T. Huang, and X. Wang, "Generative multimodal models are in-context learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 398–14 409.

[101] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou, "mplug-owl3: Towards long image-sequence understanding in multi-modal large language models," *arXiv preprint arXiv:2408.04840*, 2024.

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

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

[104] Q. Team, "Qwen2.5-vl," January 2025. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5-vl/

[105] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.

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

[107] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang *et al.*, "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," *arXiv preprint arXiv:2412.10302*, 2024.

[108] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.

[109] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai *et al.*, "ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools," *arXiv preprint arXiv:2406.12793*, 2024.

[110] Z. Liu, L. Zhu, B. Shi, Z. Zhang, Y. Lou, S. Yang, H. Xi, S. Cao, Y. Gu, D. Li *et al.*, "Nvila: Efficient frontier visual language models," *arXiv preprint arXiv:2412.04468*, 2024.

[111] J. Ye, A. Hu, H. Xu, Q. Ye, M. Yan, G. Xu, C. Li, J. Tian, Q. Qian, J. Zhang *et al.*, "Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model," *arXiv preprint arXiv:2310.05126*, 2023.

[112] A. Hu, H. Xu, L. Zhang, J. Ye, M. Yan, J. Zhang, Q. Jin, F. Huang, and J. Zhou, "mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding," *arXiv preprint arXiv:2409.03420*, 2024.

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

[114] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan *et al.*, "Janus: Decoupling visual encoding for unified multimodal understanding and generation," *arXiv preprint arXiv:2410.13848*, 2024.

[115] M. Shi, F. Liu, S. Wang, S. Liao, S. Radhakrishnan, D.-A. Huang, H. Yin, K. Sapra, Y. Yacoob, H. Shi *et al.*, "Eagle: Exploring the design space for multimodal llms with mixture of encoders," *arXiv preprint arXiv:2408.15998*, 2024.

[116] H. Laurençon, A. Marafioti, V. Sanh, and L. Tronchon, "Building and better understanding vision-language models: insights and future directions," in *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024.

[117] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen *et al.*, "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," *arXiv preprint arXiv:2503.01743*, 2025.

[118] H. Duan, J. Yang, Y. Qiao, X. Fang, L. Chen, Y. Liu, X. Dong, Y. Zang, P. Zhang, J. Wang *et al.*, "Vlmevalkit: An open-source toolkit for evaluating large multi-modality models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 11 198–11 201.

[119] K. Team, A. Du, B. Yin, B. Xing, B. Qu, B. Wang, C. Chen, C. Zhang, C. Du, C. Wei, C. Wang, D. Zhang, D. Du, D. Wang, E. Yuan, E. Lu, F. Li, F. Sung, G. Wei, G. Lai, H. Zhu, H. Ding, H. Hu, H. Yang, H. Zhang, H. Wu, H. Yao, H. Lu, H. Wang, H. Gao, H. Zheng, J. Li, J. Su, J. Wang, J. Deng, J. Qiu, J. Xie, J. Wang, J. Liu, J. Yan, K. Ouyang, L. Chen, L. Sui, L. Yu, M. Dong, M. Dong, N. Xu, P. Cheng, Q. Gu, R. Zhou, S. Liu, S. Cao, T. Yu, T. Song, T. Bai, W. Song, W. He, W. Huang, W. Xu, X. Yuan, X. Yao, X. Wu, X. Zu, X. Zhou, X. Wang, Y. Charles, Y. Zhong, Y. Li, Y. Hu, Y. Chen, Y. Wang, Y. Liu, Y. Miao, Y. Qin, Y. Chen, Y. Bao, Y. Wang, Y. Kang, Y. Liu, Y. Du, Y. Wu, Y. Wang, Y. Yan, Z. Zhou, Z. Li, Z. Jiang, Z. Zhang, Z. Yang, Z. Huang, Z. Huang, Z. Zhao, and Z. Chen, "Kimi-VL technical report," 2025. [Online]. Available: https://arxiv.org/abs/2504.07491

[120] S. Lu, Y. Li, Q.-G. Chen, Z. Xu, W. Luo, K. Zhang, and H.-J. Ye, "Ovis: Structural embedding alignment for multimodal large language model," *arXiv preprint arXiv:2405.20797*, 2024.

[121] OpenAI, "GPT-4o mini: advancing cost-efficient intelligence," https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence, 2024, accessed: 2024-12-29.

[122] Anthropic, "Claude 3.5 Sonnet," https://www.anthropic.com/news/claude-3-5-sonnet, 2024, accessed: 2024-12-29.