SCIENCE CHINA
Information Sciences

December 2024, Vol. 67, Iss. 12, 220102:1–220102:13
https://doi.org/10.1007/s11432-024-4235-6

• RESEARCH PAPER •
Special Topic: Large Multimodal Models

OCRBench: on the hidden mystery of OCR in large multimodal models

Yuliang LIU[1], Zhang LI[1], Mingxin HUANG[6], Biao YANG[1], Wenwen YU[1], Chunyuan LI[3], Xu-Cheng YIN[4], Cheng-Lin LIU[5], Lianwen JIN[6] & Xiang BAI[2*]

[1]*School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan* 430074*, China;*
[2]*School of Software Engineering, Huazhong University of Science and Technology, Wuhan* 430074*, China;*
[3]*Microsoft Research, Washington* 20237*, USA;*
[4]*School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing* 100083*, China;*
[5]*Institute of Automation, Chinese Academy of Sciences, Beijing* 101408*, China;*
[6]*School of Electronic and Information Engineering, South China University of Technology, Guangzhou* 510641*, China*

Received 12 August 2024/Accepted 23 November 2024/Published online 11 December 2024

Abstract Large models have recently played a dominant role in natural language processing and multimodal vision-language learning. However, their effectiveness in text-related visual tasks remains relatively unexplored. In this paper, we conducted a comprehensive evaluation of large multimodal models, such as GPT4V and Gemini, in various text-related visual tasks including text recognition, scene text-centric visual question answering (VQA), document-oriented VQA, key information extraction (KIE), and handwritten mathematical expression recognition (HMER). To facilitate the assessment of optical character recognition (OCR) capabilities in large multimodal models, we propose OCRBench, a comprehensive evaluation benchmark. OCRBench contains 29 datasets, making it the most comprehensive OCR evaluation benchmark available. Furthermore, our study reveals both the strengths and weaknesses of these models, particularly in handling multilingual text, handwritten text, non-semantic text, and mathematical expression recognition. Most importantly, the baseline results presented in this study could provide a foundational framework for the conception and assessment of innovative strategies targeted at enhancing zero-shot multimodal techniques. The evaluation pipeline and benchmark are available at https://github.com/Yuliang-Liu/MultimodalOCR.

Keywords large multimodal model, OCR, text recognition, scene text-centric VQA, document-oriented VQA, key information extraction, handwritten mathematical expression recognition

1 Introduction

The advent of large models has unlocked a wealth of potential in the realm of advanced computing. In recent years, there has been an explosion in the development of large language models (LLM) such as ChatGPT [1] and GPT-4 [2], giving rise to extraordinary applications in zero-shot task transfer to many new real-world scenarios. The success of proprietary LLMs has stimulated tremendous interest in developing open-source LLMs. Among them, LLaMA [3] is an open-source LLM that matches the performance of GPT-3, followed by Alpaca [4], Vicuna [5], GPT-4-LLM [6] to improve the LLM's alignment ability to follow human instruction, reporting impressive performance compared with proprietary LLMs.

The success of large models has also been extended to the multimodal vision-language space [7], leading to a line of research on large multimodal models (LMMs), including contrastive learning [8–11] and generative modeling [2,12–15]. Surprisingly, Liu et al. [15] showed that LMM exhibits excellent zero-shot optical character recognition (OCR) performance in the wild, without explicitly training on the OCR domain-specific data. Understanding the efficacy of LMM in handling text-related visual tasks is pivotal, given their potential to infer context from multiple data sources, such as text and images. Despite this advantage, these models may face challenges when dealing with complex relationships between different data types due to their general training on web-scale data. Recognizing these limitations could guide improvements in multimodal methodologies and inspire the creation of more robust models that can

* Corresponding author (email: xbai@hust.edu.cn)

handle text-related tasks more efficiently. Additionally, this knowledge can open up novel applications in areas such as digital marketing or social media analysis, where understanding the interplay between textual and visual content in the images is crucial.

To this end, we conduct a comprehensive study on 14 LMMs by evaluating their OCR ability on five representative tasks: text recognition, scene text-centric visual question answering (VQA), document-oriented VQA, key information extraction, and handwritten mathematical expression recognition. The results indicate that even state-of-the-art large multimodal models, such as Gemini [16] and GPT4V [17], still encounter challenges in recognizing blurry text images, handwritten text, multilingual text, and handwritten mathematical expressions. Moreover, we observe that they heavily rely on semantic understanding to recognize words, often favoring common words over random letter sequences. The above findings demonstrate that even the most powerful LMM still exhibits significant gaps compared to the domain-specific methods in various text-related tasks. Consequently, there exists a promising opportunity to enhance the OCR capabilities of LMMs through domain-specific adaptations and optimizations.

2 Related work

2.1 Large multimodal models

The remarkable success of LLMs has paved the way for the development of large multimodal models, which combine pretrained visual models with LLMs to enable their visual capabilities. BLIP-2 [18] introduces the querying transformer (Q-Former) as a means to bridge the gap between vision and language models. Flamingo [13] and OpenFlamingo [19] enhance a frozen pretrained LLM by incorporating novel gated cross-attention-dense layers, enabling conditioning on visual inputs. LLaVA [15] pioneers the use of GPT-4 to generate multimodal instruction-following data. Other studies, such as [20, 21], also focused on aligning the vision module and LLM for improved multimodal understanding. Additionally, Refs. [22, 23] emphasized modality collaboration for image and text. LLaVAR [24] collects training data with rich text and uses a higher-resolution CLIP as the visual encoder to enhance LLaVA's OCR ability. BLIVA [25] combines instruction-aware and global visual features to capture richer image information. MiniGPT4-v2 [26] uses unique identifiers for different tasks when training the model to better distinguish each task instruction effortlessly. UniDoc [27] performs unified multimodal instruction tuning on large-scale instruction following datasets and leverages the beneficial interactions among tasks to enhance the performance of each individual task. Docpedia [28] directly processes visual input in the frequency domain rather than the pixel space. Monkey [29] enhanced the LMM's ability to perceive details at a low cost by the generated detailed caption and a high-resolution model architecture. TextMonkey [30] introduced window attention to strengthen the correlation between different patches and introduced a token resampler to reduce the length of image tokens.

2.2 Benchmarks for large multimodal models

With the ongoing advancements in large multimodal models, the question of how to effectively evaluate their performance has emerged. Refs. [31, 32] have developed effective systems to assess LMMs. Refs. [15,22,33] introduced GPT-4 or human evaluation to assess the output of LMMs. MMBench [34] evaluates LMM using multiple-choice questions across various dimensions of ability. MME [35] measures both perception and cognition abilities on True or False questions. However, their testing on OCR data is limited. Additionally, the True/False question or multiple-choice question cannot accurately assess the OCR's ability to recognize words in an image. In this paper, we conducted an extensive study on various LMMs for five prominent text-related tasks. To facilitate the evaluation of LMMs' OCR ability, we also present OCRBench, a collection of 1000 manually filtered and corrected question-answer pairs on five representative text-related tasks.

3 Experiments

3.1 Evaluation metric and evaluation dataset

The responses generated by LMM often include many explanatory terms, so the exact matching approach or average normalized levenshtein similarity (ANLS) [36] used in the original dataset is not suitable for

evaluating LMM in zero-shot scenarios. We have defined a unified and simple evaluation criterion for all datasets, which is to determine whether the ground truth (GT) is present in the output of the LMM. To reduce false positives, we filter out questions that have answers containing fewer than 4 symbols from all datasets. Additionally, we choose 3000 question instances from some large datasets for testing purposes.

Text recognition. We evaluate LMM using widely-adopted OCR text recognition datasets, including (1) regular text recognition: IIIT5K [37], SVT [38], IC13 [39]; (2) irregular text recognition: IC15 [40], SVTP [41], CT80 [42], COCO-Text (COCO) [43], SCUT-CTW1500 (CTW) [44], Total-Text (TT) [45]; (3) occlusion scene text [46], encompassing weakly occluded scene text (WOST) and heavily occluded scene text (HOST); (4) artistic text recognition: WordArt [47]; (5) handwritten text recognition: IAM [48]; (6) Chinese text recognition: ReCTS [49]; (7) handwritten digit string recognition: ORAND-CAR-2014 (CAR-A) [50]; (8) non-semantic text (NST) and semantic text (ST): LMMs primarily relying on semantic understanding to recognize words. In the experiments, we found that the LMMs have poor recognition performance on character combinations that lacked semantics. To confirm this, we create two datasets: ST and NST using the IIIT5K dictionary. The ST dataset consists of 3000 images with words from the IIIT5K dictionary, while the NST dataset contains the same words but with shuffled characters without semantics. For English words, we employ the consistent prompt "What is written in the image?". For Chinese words in the ReCTS dataset, we adapt the prompt to "What are the Chinese characters in the image?". For handwritten digit strings, we utilize the prompt "What is the number in the image?".

Scene text-centric VQA. We test LMMs on STVQA [51], TextVQA [52], OCR-VQA [53], and ESTVQA [54]. Scene text visual question answering (STVQA) consists of 31000+ questions across 23000+ images collected from various public datasets. TextVQA dataset comprises 45000+ questions on 28000+ images sampled from specific OpenImages dataset categories expected to contain text. OCR-VQA features over 1 million question-answer pairs spanning 207000+ book cover images. ESTVQA contains 20757 images along with 15056 English questions and 13006 Chinese questions. We have divided the ESTVQA dataset into ESTVQA(CN) and ESTVQA(EN), which specifically include questions and answers in Chinese or English, respectively.

Document-oriented VQA. We assess these LMMs on DocVQA [55], InfographicVQA (InfoVQA) [56] and ChartQA [57]. DocVQA is a large-scale dataset with 12767 document images of diverse types and content, and over 50000 questions and answers. InfographicVQA is a diverse collection of infographics that includes 5485 images and a total of 30035 questions. ChartQA includes a total of 9608 human-written questions covering 4804 charts, as well as 23111 questions generated from human-written chart summaries on 17141 charts.

Key information extraction. We conduct experiments on SROIE [58], FUNSD [59], and POIE [60]. SROIE contains 1000 complete scanned receipt images for OCR and key information extraction competitions. In this competition, the company, date, address, and total expenditure information must be extracted based on the receipts. FUNSD dataset consists of 199 real, fully annotated, scanned forms that may contain noise. POIE consists of camera images from the nutrition facts label of products in English and 3000 images with 111155 text instances are collected. The KIE dataset requires the extraction of key-value pairs in the image. To enable LMMs to accurately extract the correct value for a given key in the KIE dataset, we employ manual prompt design. For the SROIE dataset, we utilize the following prompts to assist LMMs in generating the respective values for "company", "date", "address", and "total": "What is the name of the company that issued this receipt?", "When was this receipt issued?", "Where was this receipt issued?", and "What is the total amount of this receipt?". Additionally, to retrieve the corresponding value for a given key in FUNSD and POIE, we utilize the prompt "What is the value for '{key}'?".

Handwritten mathematical expression recognition (HMER). We evaluate on HME100K [61], which consists of 74502 images for training and 24607 images for testing, with 245 symbol classes. During evaluation, we use the prompt "Please write out the expression of the formula in the image using LaTeX format."

3.2 Results

The results of text recognition are shown in Table 1 [61–64]. Details of the supervised SOTA can be found in Appendix A. LMMs achieve comparable performance to state-of-the-art supervised models in recognizing regular text, irregular text, occluded text, and artistic text. Particularly in the WordArt dataset, which predominantly comprises challenging artistic text, InstructBLIP2 and BLIVA even outperform the

Table 1 Text recognition results[a)]

Method	Regular			Irregular						Occluded		Artistic	Handwritten	Chinese	Digits	Semantic	
	IIIT5K	SVT	IC13	IC15	SVTP	CT80	COCO	CTW	TT	HOST	WOST	WordArt	IAM	ReCTS	ORAND	NST	ST
BLIP2-6.7B	79.1	86.7	83.4	71.2	78.8	76.5	51.4	62.4	68.7	59.8	69.4	68.4	32.9	0	1.2	13.0	82.6
mPLUG-Owl	81.1	84.3	85.9	67.5	73.9	81.4	52.3	69.2	74.4	49.7	62.7	72.1	34.8	0	13.5	44.7	92.3
InstructBLIP	86.3	92.0	86.8	80.9	85.6	86.3	62.6	70.8	77.9	67.8	78.8	**73.7**	42.6	0	18.4	25.3	89.7
LLaVAR	84.0	87.6	87.7	79.4	84.0	84.5	61.9	69.5	75.6	61.1	71.9	67.1	49.4	0	9.8	36.2	86.5
BLIVA	86.5	90.6	87.3	80.9	87.7	86.7	**64.8**	71.2	78.1	67.7	77.6	**73.7**	45.1	0	13.8	20.4	89.4
mPLUG-Owl2	80.9	69.6	79.8	53.9	53.5	74.8	52	59.1	60.9	32.5	50.6	60.6	23.8	0	9.9	48.2	93.9
LLaVA1.5-7B	84.2	85.7	86.4	71.9	79.8	82.7	55.6	66.8	73.2	61.4	70.6	68.7	55.4	0	10.4	15.2	85.3
UniDoc	91.9	89.2	90.9	78	80.3	88.2	64.1	75.3	78.2	52.4	68.5	–	–	–	–	–	–
Monkey	83.7	75.1	85.4	53.4	58.4	73.9	43.5	64.5	64.6	43.3	54.9	67.7	30.3	13.1	29.1	56.5	95.5
Supervised-SOTA	PARSeq [61]	PARSeq [61]	PARSeq [61]	PARSeq [61]	PARSeq [61]	PARSeq [61]	PARSeq [61]	PARSeq [61]	PARSeq [61]	PARSeq [61]	PARSeq [61]	PARSeq [61]	AttentionHTR [62]	TPS-ResNet [63]	Yu et al. [64]	PARSeq [61]	PARSeq [61]
	96.6	**93.0**	**96.7**	**85.7**	**89.3**	**89.9**	64.4	**78.6**	**80.1**	**73.1**	**81.6**	72.5	**91.2**	**94.8**	**95.5**	**95.4**	**100.0**

a) Bold black digits indicate the best result, while underline signifies the second best.

Table 2 Results of scene text-centric VQA, document-oriented VQA, KIE, and HMER[a)]. Since the Supervised-SOTA on the ESTVQA dataset does not provide separate results on Chinese and English data, the average performance of 42.3 is used as a reference

Method	Scene text-centric VQA					Document-oriented VQA			KIE			HMER
	STVQA	TextVQA	OCR-VQA	ESTVQA(EN)	ESTVQA(CN)	DocVQA	InfoVQA	ChartQA	FUNSD	SROIE	POIE	HME100K
BLIP2-6.7B	20.9	23.5	9.7	40.7	0	3.2	11.3	3.4	0.2	0.1	0.3	0
mPLUG-Owl	30.5	34	21.1	52.7	0	7.4	20	7.9	0.5	1.7	2.5	0.1
InstructBLIP	27.4	29.1	41.3	48.6	0.1	4.5	16.4	5.3	0.2	0.6	1	0
LLaVAR	39.2	41.8	24	58.2	0	12.3	16.5	12.2	0.5	5.2	5.9	0
BLIVA	32.1	33.3	50.7	51.2	0.2	5.8	23.6	8.7	0.2	0.7	2.1	0.1
mPLUG-Owl2	49.8	53.9	58.7	68.6	4.9	17.9	18.9	19.4	1.4	3.2	9.9	0
LLaVA1.5-7B	38.1	38.7	58.1	52.3	0	8.5	14.7	9.3	0.2	1.7	2.5	0
UniDoc	35.2	46.2	36.8	–	–	7.7	14.7	10.9	1	2.9	5.1	0.4
DocPedia	45.5	60.2	57.2	–	–	47.1	15.2	46.9	29.9	21.4	39.9	–
Monkey	54.7	64.3	64.4	**71**	42.6	50.1	25.8	54.0	24.1	41.9	19.9	0.2
Supervised-SOTA	GIT [14]	Mia [65]	GIT [14]	Fang et al. [66]	Fang et al. [66]	AIDU-DI [67]	DUBLIN [68]	DePlot [69]	ERNIE-Layout [67]	StrucTexT [70]	Kuang et al. [60]	Li et al. [71]
	69.6	**73.7**	**68.1**	43.3	**43.3**	**90.16**	**36.8**	**70.5**	**93.1**	**98.7**	**79.5**	**64.3**

a) Bold black digits indicate the best result, while underline signifies the second best.

supervised state-of-the-art model. However, LMMs exhibit poor performance in recognizing handwritten text, Chinese text, handwritten strings, and non-semantic text. In tasks, such as scene text-centric VQA, document-oriented VQA, and KIE, LMMs with smaller input resolutions consistently produce poorer results compared to those with larger input sizes. This is due to the intricate structures and varying sizes of texts, requiring LMMs to capture fine details. The results are shown in Table 2 [14,65–71]. LMMs face challenges in accurately recognizing handwritten mathematical expressions. By extensively analyzing the results, we provide a qualitative overview of the limitations of LMMs in text-related tasks.

• **Semantic-reliance.** LMMs primarily rely on semantic understanding to recognize words. In our experiments, we observed that LMMs exhibit poor recognition performance when it comes to character combinations that lack semantic meaning. Specifically, when we altered the order of letters in each word, the accuracy of LMMs on the NST dataset decreased by an average of 57.0% compared to the ST dataset, while the SOTA method for scene text recognition only dropped by around 4.6%. We believe this is because the SOTA method for scene text recognition directly recognizes each character, and semantic information is just used to assist the recognition process, while LMMs primarily rely on semantic understanding to recognize words. This finding is further supported by the low accuracy observed on the ORAND dataset. As shown in Figure 1, LMM successfully identified the word "Message", but incorrectly recognized "egaesMs", which is a rearranged version of the word "Message".

• **Handwritten text.** LMMs may encounter difficulties in accurately recognizing handwritten text due to various reasons, including the resemblances in shapes between handwritten letters and numbers. Handwritten text often appears incomplete or blurry due to factors like fast writing speed, irregular handwriting, or low-quality paper. On average, LMMs exhibit a 51.9% lower performance compared to the supervised state-of-the-art model in this particular task.

• **Multilingual text.** As indicated in Tables 1 and 2, there exists a notable performance gap between ESTVQA(CN) and ESTVQA(EN). The LMMs achieve limited proficiency in the Chinese language. Accurately recognizing Chinese words or responding to Chinese questions poses a considerable challenge for LMMs. Training LMMs with Chinese data emerges as a viable solution. For example, Monkey surpasses other LMMs in Chinese scenarios due to the substantial training of its LLM and visual encoder on Chinese data.

• **Fine-grain perception.** As shown in Appendix B, the resolution of most LMMs is currently limited to 224 × 224 due to the visual encoder used in their architecture. However, supporting high-resolution input is essential for LMMs to capture finer details within images. The restricted input resolution of LMMs like BLIP-2 hinders their ability to extract detailed information in tasks such as scene text-centric VQA, document-oriented VQA, and KIE. Conversely, LMMs like Monkey, which can handle a resolution of 1344 × 896, demonstrate enhanced performance in these specific tasks.

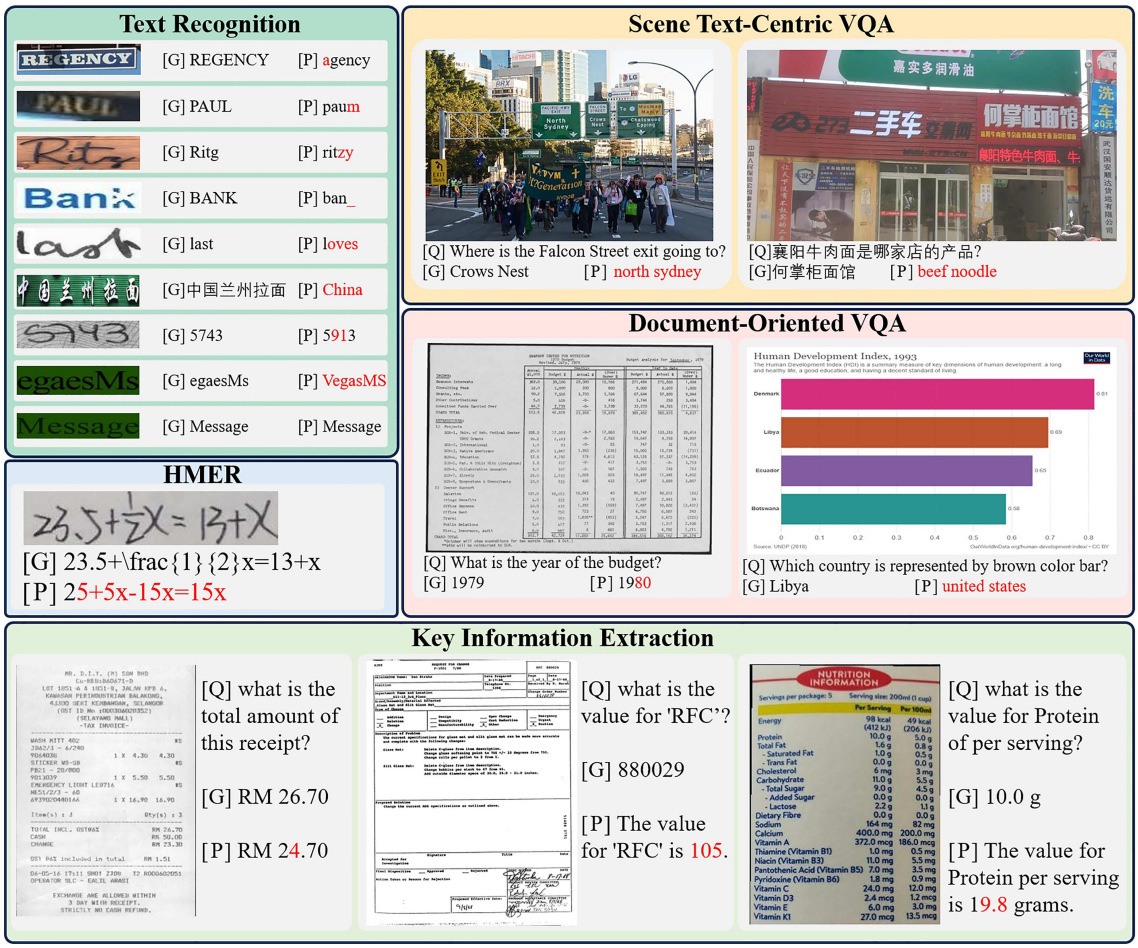

Figure 1 (Color online) Visualization results of the five tasks. '[Q]' represents question, '[G]' represents ground truth, and '[P]' represents prediction generated by LMM.

- **HMER.** Recognizing handwritten mathematical expressions poses a challenge for LMMs due to the presence of messy handwritten characters, complex spatial structures, and the indirect LaTeX representation. Additionally, the scarcity of training data for this specific task further complicates the recognition process for LMMs.

3.3 OCRBench

Evaluating the model on multiple datasets is time-consuming, and the presence of inaccurate annotations in some datasets diminishes the precision of accuracy-based evaluations. To solve these issues, we develop OCRBench to facilitate the accurate and convenient evaluation of LMMs' OCR capabilities. OCRBench consists of five components: text recognition, scene text-centric VQA, document-oriented VQA, KIE, and HMER. Detailed descriptions can be found in Appendix C. It includes 1000 question-answer pairs, and for the KIE task, we added the prompt "Answer this question using the text in the image directly" to restrict the model's response format. The specific composition of OCRBench is shown in Appendix C. To ensure a more accurate evaluation, we manually verified and corrected incorrect answers for the 1000 question-answer pairs, providing alternative correct answer candidates. The evaluation results on OCRBench are presented in Table 3, where Gemini achieves the highest score, followed by GPT4V in the second position. It is important to note that due to rigorous safety reviews by OPEN AI, GPT4V refused to provide results for 84 images in OCRBench. Monkey exhibited OCR capabilities that trailed only behind GPT4V and Gemini. From Table 3, we can observe that even state-of-the-art models like GPT4V and Gemini still struggle with the HMER task. Moreover, they also face challenges in processing unclear images, handwritten text, non-semantic text, and adhering to task instructions. As shown in Figure 2(g), even when explicitly requested to answer using the text found in the image, Gemini consistently interprets

Table 3 Results of LMMs on OCRBench[a)]

Method	Recog.	VQA^{S}	VQA^{D}	KIE	HMER	Final score
gemini-pro-vision	**215**	**174**	128	134	8	**659**
gpt-4-vision-preview	167	163	**146**	**160**	**9**	645
Monkey	174	161	91	88	0	514
mPLUG-Owl2	153	153	41	19	0	366
LLaVAR	186	122	25	13	0	346
LLaVA1.5-13B	176	129	19	7	0	331
LLaVA1.5-7B	160	117	15	5	0	297
mPLUG-Owl	172	104	18	3	0	297
BLIVA	165	103	22	1	0	291
InstructBLIP	168	93	14	1	0	276
BLIP-2	154	71	10	0	0	235
MiniGPT4-v2	124	29	4	0	0	157

a) Recog. represents text recognition, VQA^{S} represents scene text-centric VQA, and VQA^{D} represents document-oriented VQA. Bold black digits indicate the best result.

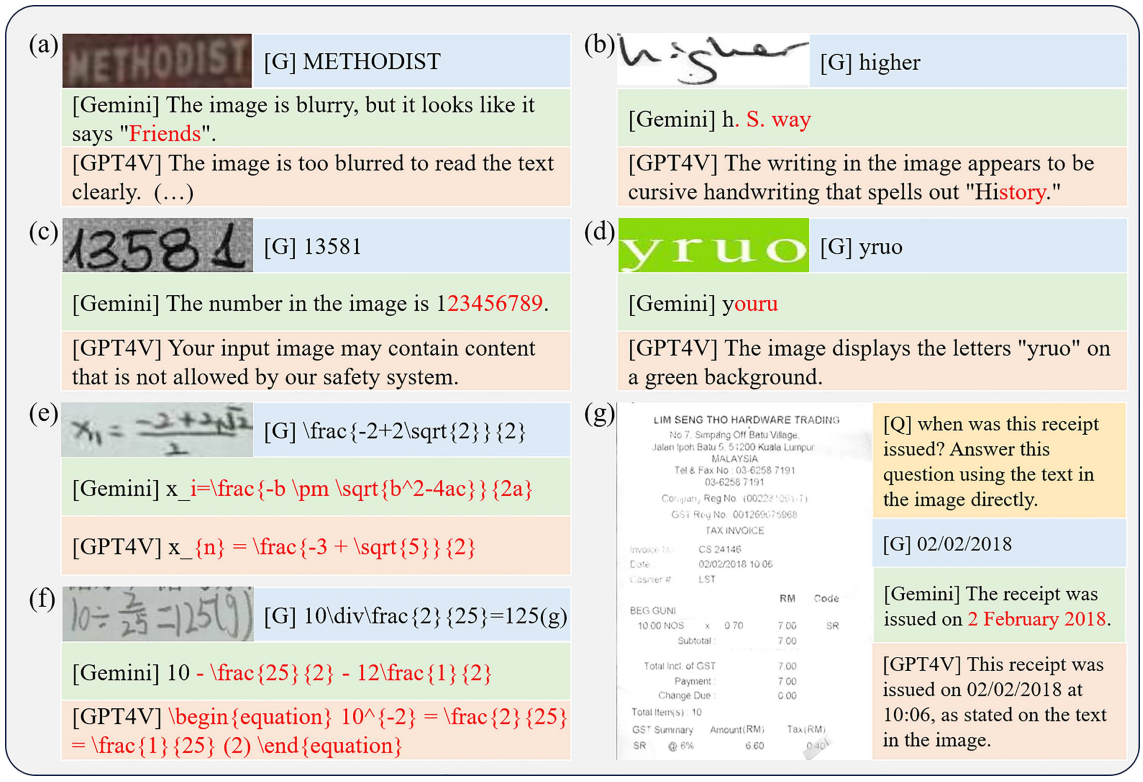

Figure 2 (Color online) Some erroneous results from Gemini and GPT-4V. '[Q]' represents the question while '[G]' represents the ground truth. (a) The results on the blurry image; (b) the results on handwritten text; (c) and (d) the results on non-semantic text; (e) and (f) the results on HMER; (g) instances where the task instructions are not adhered to.

"02/02/2018" as "2 February 2018".

4 Discussion

Why LMMs work for OCR? While we offer some analysis based on the results, the question as to why these multimodal models can deliver acceptable performance on OCR tasks remains challenging to conclusively explain. One plausible explanation for the success of multimodal models in OCR tasks lies in the training data of multimodal models (similar to CLIP), which we believe includes some OCR data. However, unsupervised text-image pairs in the training data cannot compete with fully supervised data.

From the perspective of architecture, the pre-trained visual encoder and LLM already demonstrate a

solid understanding of their respective domain data, each working within their designated feature spaces. These LLMs connect visual and language data through elements like a linear projection layer, which acts as a visual tokenization step. By aligning visual tokens within the word embedding space of the pre-trained language model, visual embeddings closely resemble their corresponding text embeddings. This alignment facilitates text recognition, allowing the LLM to then present this OCR data to users in a generative manner. Future research could benefit from conducting an ablation analysis to better understand how the volume of multimodal training data impacts OCR performance.

Comparing computational costs and efficiency between LMM and specialized OCR model. We further conduct experiments to compare the computational costs and efficiency of LMM and the specialized OCR model. The experiments are performed on text recognition tasks in OCRBench, which includes 300 images. We choose Monkey as the representative for LMM and PARSeq as the representative for a specialized OCR model. Specifically, Monkey demonstrates a computational cost of 38.57 TFLOPS and an inference time of 0.46 s per image, achieving an accuracy of 58.0%. In contrast, PARSeq operates with significantly lower computational requirements, consuming only 6.2 GFLOPS and achieving an inference time of 0.07 s per image, with an accuracy of 87.3%. The results indicate that while LMMs demonstrate promising potential across diverse tasks, specialized models remain indispensable for specific tasks due to their significantly lower computational resource requirements.

Future directions. Although current LMMs have shown promising results, they still face challenges when dealing with complex tasks and are unable to match domain-specific methods in traditional text-related tasks. To enhance their performance, these models should concentrate on improving their ability to detect intricate features in images, enhancing their recognition of individual character shapes, and developing multilingual datasets for training.

Potential of OCR+LLM in visual document understanding. Recent work such as the ICL-D3IE [72], introduces an in-context learning framework leveraging large language models like GPT-3 and ChatGPT for document information extraction. This approach effectively navigates the modality and task gaps in this field, extracting challenging and distinct segments from difficult training documents, designing relationship demonstrations for enhanced positional understanding, and incorporating formatting demonstrations for answer extraction. Despite its notable success across various datasets and settings, ICL-D3IE encounters challenges in the in-domain setting on CORD [73]. This result points to the potential of large language models in managing tasks involving visually intricate documents and encourages the development of novel, minimally supervised OCR techniques. Moreover, a study by Li et al. [74], evaluates ChatGPT across seven detailed information extraction tasks. This evaluation reveals that ChatGPT performs well in OpenIE settings, generates high-quality responses, shows a tendency towards overconfidence, and maintains strong fidelity to original texts. However, this study focused exclusively on pure text, without venturing into visually rich document texts within the OCR domain. Future research should therefore explore ChatGPT's performance in these visually rich contexts.

Potential application of LMMs in specialized domains. The application of LMMs in specialized domains is also a crucial field, and recent developments have shown their vast potential. One such model, developed by Google, is Med-PaLM2 [75], which is fine-tuned on medical knowledge from PaLM2. It is the first language model to perform an expert level on US Medical Licensing Examination (USMLE) style questions and can analyze patients' conditions through medical images, including plain films and mammograms. Google claims that it has approached the performance of clinician experts. Additionally, LLaVA-Med [76] attempts to extend multimodal instruction-tuning to the biomedical domain and demonstrates excellent chat abilities with domain knowledge. Furthermore, in the field of automated office, LMMs can significantly enhance the efficiency of reading and processing word documents and PDFs by providing comprehensive summaries, accurately answering questions based on both textual and visual content, and improving accessibility. They can automatically cross-reference information, offer translations for multilingual documents, summarize key information, thereby streamlining the way users interact with, and extract value from these documents. This makes them powerful tools for boosting productivity and effectiveness in various professional and educational contexts. Besides, in the field of education, LMMs significantly lighten teachers' workloads by automatically evaluating student essays, diagrams, and projects, thereby providing immediate and detailed feedback. For assessments, these models can analyze test papers to pinpoint areas of weakness and generate targeted practice questions, reinforcing students' learning. Furthermore, LMMs enhance the educational experience by analyzing written assignments and offering personalized learning plans. Overall, LMMs contribute to elevating both the quality and efficiency of teaching.

Table 4 Results of OCRBench across various models are sourced from [31] and their papers[a)]

Model	Affiliation	OCRBench	Model	Affiliation	OCRBench
MiniCPM-V2.6 [77]	OpenBMB	852	Cambrian-8B [78]	NYU	614
InternVL2-Llama3-76B [79]	Shanghai AI Lab	842	PaliGemma-3B-mix-448 [80]	Google	614
CongRong [81]	CloudWalk	827	Cambrian-13B [78]	NYU	610
GLM-4v [82]	Zhipu AI	814	MiniCPM-V-2 [83]	OpenBMB	605
MiniMonkey-2B [84]	HUST	802	Cambrian-34B [78]	NYU	591
InternVL2-8B [79]	Shanghai AI Lab	794	CogVLM-17B-Chat [82]	Zhipu AI	590
Claude3.5-Sonnet [85]	Anthropic	788	LLaVA-Next-Yi-34B [86]	UW-Madison	574
GPT-4o-mini-20240718 [87]	OpenAI	785	TextMonkey [30]	HUST	561
InternVL2-4B [79]	Shanghai AI Lab	784	Monkey-Chat [29]	HUST	534
InternVL2-2B [79]	Shanghai AI Lab	781	InternLM-XComposer2 [88]	Shanghai AI Lab	532
GLM-4v-9B [82]	Zhipu AI	776	LLaVA-Next-Vicuna-7B [86]	UW-Madison	532
CogVLM2-19B-Chat [82]	Zhipu AI	757	LLaVA-Next-Mistral-7B [86]	UW-Madison	531
InternVL2-1B [79]	Shanghai AI Lab	755	RekaEdge [89]	Reka AI	506
Gemini-1.5-Pro [90]	Google	754	XVERSE-V-13B [91]	XVERSE	489
Ovis1.5-Llama3-8B [92]	Alibaba	744	Qwen-VL-Chat [93]	Alibaba	488
Qwen-VL-Plus [93]	Alibaba	726	InternLM-XComposer2-1.8B [88]	Shanghai AI Lab	447
MiniCPM-Llama3-V2.5 [94]	OpenBMB	725	Emu2_chat [95]	BAAI	436
InternVL-Chat-V1.5 [79]	Shanghai AI Lab	720	DeepSeek-VL-7B [96]	DeepSeek	435
Claude3-Opus [97]	Anthropic	694	OmniLMM-12B [98]	OpenBMB	420
RekaFlash [89]	Reka AI	692	TransCore-M [99]	PCI Research	405
InternLM-XComposer2.5 [100]	Shanghai AI Lab	686	LLaVA-InternLM2-7B [101]	Shanghai AI Lab	402
Qwen-VL-Max [93]	Alibaba	684	ShareGPT4V-13B [102]	Shanghai AI Lab	398
Gemini-1.0-Pro [90]	Google	680	360VL-70B [103]	QiHoo360	397
InternLM-XComposer2-4KHD [104]	Shanghai AI Lab	675	MiniCPM-V [105]	OpenBMB	366
Claude3-Haiku [97]	Anthropic	658	Yi-VL-34B [106]	01-AI	290
Mini-InternVL-Chat-2B-V1.5 [79]	Shanghai AI Lab	652	IDEFICS-80B-Instruct [107]	Hugging Face	283
Claude3-Sonnet [97]	Anthropic	646	LLaVA-Next-Llama3 [86]	ByteDance	252
Mini-InternVL-Chat-4B-V1.5 [79]	Shanghai AI Lab	639	MMAlaya [108]	DataCanvas	223
Phi-3-Vision [109]	Microsoft	637	VisualGLM [110]	Tsinghua	170
WeMM [109]	WechatCV	628	OpenFlamingo v2 [111]	UW	149
IDEFICS2-8B [112]	Hugging Face	626	PandaGPT-13B [113]	Tencent AI Lab	36
Step-1V [114]	StepFun	625	Chameleon-30B [115]	Meta	27

a) Closed-source models are indicated by underlines.

These studies highlight the potential of LMMs in vertical domain tasks. Further exploration of their applications in other domains, such as gaming and education, is also warranted. Ultimately, these developments and future research directions could potentially pave the way for multimodal models that can more efficiently handle complex tasks like OCR, expanding the application range of LMM.

5 Future work

With the advancement of large multimodal models, numerous research institutions have made significant strides in OCR capabilities. For instance, MiniCPMV-2.6 achieved a score of 852 on OCRBench, while the Mini-Monkey scored 802 using just 2B parameters. These results underscore the immense potential of large multimodal models in the OCR domain, as highlighted in Table 4 [29–31, 77–115]. OCRBench has been instrumental in advancing this field. However, it is important to note that the data utilized by most models is not entirely open-source, and many models remain proprietary. This demonstrates an existing opportunity for the open-source community to further explore and enhance the OCR capabilities of large multimodal models.

Furthermore, OCRBench currently lacks a comprehensive coverage of image data types and tasks. For instance, it falls short in encompassing more challenging data types like multilingual documents and texts captured in diverse scenarios, as well as tasks related to text detection. To address these gaps, our future work will focus on expanding OCRBench to include these elements, thereby fostering the ongoing advancement and utilization of large multimodal models in the realm of OCR.

6 Conclusion

This paper has presented an extensive study on the performance of LMM on OCR tasks, including text recognition, scene text-centric, document-oriented VQA, KIE, and HMER. Our quantitative assessment reveals that LMM can achieve promising results, especially in text recognition, even attaining SOTA performance in some datasets. However, significant gaps persist compared to domain-specific supervised methods, suggesting that specialized techniques tailored to each task are still essential, as the latter uses much less computational resources and data. The proposed OCRBench has served as the evaluation benchmark for the OCR capabilities of multimodal large models, driving the development of LMMs and

demonstrating their immense potential in the OCR field. In the future, we hope to further explore the potential of LMMs across more scenarios, more complex tasks, and multiple languages, ultimately paving the way for more intelligent and versatile OCR solutions for large multimodal models.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62225603, 62226104). Thanks for the help from Hongliang LI, Yang LIU, Dezhi PENG, Mingyu LIU, and Mingrui CHEN.

References

1 OpenAI. ChatGPT. 2023. https://openai.com/blog/chatgpt
2 Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. 2023. ArXiv:2303.08774
3 Touvron H, Lavril T, Izacard G, et al. Llama: open and efficient foundation language models. 2023. ArXiv:2302.13971
4 Taori R, Gulrajani I, Zhang T Y, et al. Stanford Alpaca: an instruction-following LLaMA model. 2023. https://github.com/tatsu-lab/stanford_alpaca
5 Vicuna. Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. 2023. https://vicuna.lmsys.org.
6 Peng B L, Li C Y, He P C, et al. Instruction tuning with GPT-4. 2023. ArXiv:2304.03277
7 Gan Z, Li L J, Li C Y, et al. Vision-language pre-training: basics, recent advances, and future trends. FNT Comput Graph Vision, 2022, 14: 163–352
8 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. 2021. ArXiv:2103.00020
9 Yuan L, Chen D D, Chen Y L, et al. Florence: a new foundation model for computer vision. 2021. ArXiv:2111.11432
10 Jia C, Yang Y F, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. 2021. ArXiv:2102.05918
11 Li C Y, Liu H T, Li L N, et al. ELEVATER: a benchmark and toolkit for evaluating language-augmented visual models. In: Proceedings of the Advances in Neural Information Processing Systems, 2022. 9287–9301
12 Driess D, Xia F, Sajjadi S M, et al. PaLM-E: an embodied multimodal language model. 2023. ArXiv:2303.03378
13 Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. In: Proceedings of the Advances in Neural Information Processing Systems, 2022. 23716–23736
14 Wang J F, Yang Z Y, Hu X W, et al. GIT: a generative image-to-text transformer for vision and language. 2022. ArXiv:2205.14100
15 Liu H T, Li C Y, Wu Q Y, et al. Visual instruction tuning. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
16 Google Gemini Team. Gemini: a family of highly capable multimodal models. 2023. ArXiv:2312.11805
17 OpenAI. GPT-4V(vision) system card. 2023. https://openai.com/index/gpt-4v-system-card
18 Li J N, Li D X, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the International Conference on Machine Learning, 2023. 19730–19742
19 Awadalla A, Gao I, Gardner J, et al. OpenFlamingo: an open-source framework for training large autoregressive vision-language models. 2023. ArXiv:2308.01390
20 Liu H T, Li C Y, Li Y H, et al. Improved baselines with visual instruction tuning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2024. 26296–26306
21 Zhu D Y, Chen J, Shen X Q, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models. 2023. ArXiv:2304.10592
22 Ye Q H, Xu H Y, Ye J, et al. mPLUG-Owl: modularization empowers large language models with multimodality. 2023. ArXiv:2304.14178
23 Ye Q H, Xu H Y, Ye J B, et al. mPLUG-Owl2: revolutionizing multi-modal large language model with modality collaboration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2024. 13040–13051
24 Zhang Y Z, Zhang R Y, Gu J B, et al. LLaVAR: enhanced visual instruction tuning for text-rich image understanding. 2023. ArXiv:2306.17107
25 Hu W B, Xu Y F, Li Y, et al. BLIVA: a simple multimodal llm for better handling of text-rich visual questions. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 2256–2264
26 Chen J, Zhu D Y, Shen X Q, et al. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. 2023. ArXiv:2310.09478
27 Feng H, Wang Z J, Tang J Q, et al. UniDoc: a universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. 2023. ArXiv:2308.11592
28 Feng H, Liu Q, Liu H, et al. DocPedia: unleashing the power of large multimodal model in the frequency domain for versatile document understanding. 2023. ArXiv:2311.11810
29 Li Z, Yang B, Liu Q, et al. Monkey: image resolution and text label are important things for large multi-modal models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2024. 26763–26773
30 Liu Y L, Yang B, Liu Q, et al. TextMonkey: an OCR-free large multimodal model for understanding document. 2024. ArXiv:2403.04473
31 Duan H D, Yang J M, Qiao Y X, et al. VLMEvalKit: an open-source toolkit for evaluating large multi-modality models. In: Proceedings of the 32nd ACM International Conference on Multimedia, 2024. 11198–11201
32 Zhang K C, Li B, Zhang P Y, et al. LMMs-eval: reality check on the evaluation of large multimodal models. 2024. ArXiv:2407.12772
33 Liu F X, Lin K, Li L J, et al. Mitigating hallucination in large multi-modal models via robust instruction tuning. In: Proceedings of the 12th International Conference on Learning Representations, 2023
34 Liu Y, Duan H D, Zhang Y H, et al. MMBench: is your multi-modal model an all-around player? In: Proceedings of the European Conference on Computer Vision, 2025. 216–233
35 Fu C Y, Chen P X, Shen Y H, et al. MME: a comprehensive evaluation benchmark for multimodal large language models. 2024. ArXiv:2306.13394
36 Biten A F, Tito R P, Mafla A, et al. Scene text visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 4291–4301
37 Mishra A, Alahari K, Jawahar C V. Top-down and bottom-up cues for scene text recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012. 2687–2694

38 Shi C Z, Wang C H, Xiao B H, et al. End-to-end scene text recognition using tree-structured models. Pattern Recogn, 2014, 47: 2853–2866
39 Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 robust reading competition. In: Proceedings of the 12th International Conference on Document Analysis and Recognition, 2013. 1484–1493
40 Karatzas D, Gomez-Bigorda L, Nicolaou A, et al. ICDAR 2015 competition on robust reading. In: Proceedings of the 13th International Conference on Document Analysis and Recognition, 2015. 1156–1160
41 Phan T Q, Shivakumara P, Tian S X, et al. Recognizing text with perspective distortion in natural scenes. In: Proceedings of the IEEE International Conference on Computer Vision, 2013. 569–576
42 Risnumawan A, Shivakumara P, Chan C S, et al. A robust arbitrary text detection system for natural scene images. Expert Syst Appl, 2014, 41: 8027–8048
43 Veit A, Matera T, Neumann L, et al. COCO-Text: dataset and benchmark for text detection and recognition in natural images. 2016. ArXiv:1601.07140
44 Liu Y L, Jin L W, Zhang S T, et al. Curved scene text detection via transverse and longitudinal sequence connection. Pattern Recogn, 2019, 90: 337–345
45 Chng C-K, Chan C S. Total-Text: a comprehensive dataset for scene text detection and recognition. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition, 2017. 935–942
46 Wang Y X, Xie H T, Fang S C, et al. From two to one: a new scene text recognizer with visual language modeling network. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 14194–14203
47 Xie X D, Fu L, Zhang Z F, et al. Toward understanding WordArt: corner-guided transformer for scene text recognition. In: Proceedings of the European conference on computer vision, 2022. 303–321
48 Marti U V, Bunke H. The IAM-database: an English sentence database for offline handwriting recognition. Int J Document Anal Recogn, 2002, 5: 39–46
49 Zhang R, Zhou Y S, Jiang Q Y, et al. ICDAR 2019 robust reading challenge on reading Chinese text on signboard. In: Proceedings of the International Conference on Document Analysis and Recognition, 2019. 1577–1581
50 Diem M, Fiel S, Kleber F, et al. ICFHR 2014 competition on handwritten digit string recognition in challenging datasets (HDSRC 2014). In: Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition, 2014. 779–784
51 Biten A F, Tito R, Mafla A, et al. ICDAR 2019 competition on scene text visual question answering. In: Proceedings of the International Conference on Document Analysis and Recognition, 2019. 1563–1570
52 Singh A, Natarajan V, Shah M, et al. Towards VQA models that can read. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 8317–8326
53 Mishra A, Shekhar S, Singh A K, et al. OCR-VQA: visual question answering by reading text in images. In: Proceedings of the International Conference on Document Analysis and Recognition, 2019. 947–952
54 Wang X Y, Liu Y L, Shen C H, et al. On the general value of evidence, and bilingual scene-text visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 10126–10135
55 Mathew M, Karatzas D, Jawahar C V. DocVQA: a dataset for VQA on document images. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2021. 2200–2209
56 Mathew M, Bagal V, Tito R, et al. InfographicVQA. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2022. 1697–1706
57 Masry A, Long D X, Tan J Q, et al. ChartQA: a benchmark for question answering about charts with visual and logical reasoning. 2022. ArXiv:2203.10244
58 Huang Z, Chen K, He J H, et al. ICDAR2019 competition on scanned receipt OCR and information extraction. In: Proceedings of the International Conference on Document Analysis and Recognition, 2019. 1516–1520
59 Jaume G, Ekenel H K, Thiran J-P. FUNSD: a dataset for form understanding in noisy scanned documents. In: Proceedings of the International Conference on Document Analysis and Recognition Workshops, 2019. 1–6
60 Kuang J F, Hua W, Liang D K, et al. Visual information extraction in the wild: practical dataset and end-to-end solution. In: Proceedings of the International Conference on Document Analysis and Recognition, 2023. 36–53
61 Bautista D, Atienza R. Scene text recognition with permuted autoregressive sequence models. In: Proceedings of the European Conference on Computer Vision, 2022. 178–196
62 Kass D, Vats E. AttentionHTR: handwritten text recognition based on attention encoder-decoder networks. In: Proceedings of the International Workshop on Document Analysis Systems, 2022. 507–522
63 Baek J, Kim G, Lee J, et al. What is wrong with scene text recognition model comparisons? Dataset and model analysis. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 4715–4723
64 Yu M-M, Zhang H, Yin F, et al. An efficient prototype-based model for handwritten text recognition with multi-loss fusion. In: Proceedings of the International Conference on Frontiers in Handwriting Recognition, 2022. 404–418
65 Qiao Y X, Chen H, Wang J, et al. Winner Team Mia at TextVQA challenge 2021: vision-and-language representation learning with pre-trained sequence-to-sequence model. 2021. ArXiv:2106.15332
66 Fang Z Q, Li L, Xie Z W. Cross-modal attention networks with modality disentanglement for scene-text VQA. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 2022. 1–6
67 Peng Q M, Pan Y X, Wang W J, et al. ERNIE-Layout: layout knowledge enhanced pre-training for visually-rich document understanding. In: Proceedings of Findings of the Association for Computational Linguistics, 2022. 3744–3756
68 Aggarwal K, Khandelwal A, Tanmay K, et al. DUBLIN: visual document understanding by language-image network. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023. 693–706
69 Liu F Y, Eisenschlos J M, Piccinno F, et al. DePlot: one-shot visual language reasoning by plot-to-table translation. In: Proceedings of the Findings of the Association for Computational Linguistics, 2023. 10381–10399
70 Li Y L, Qian Y X, Yu Y C, et al. StrucTexT: structured text understanding with multi-modal transformers. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021. 1912–1920
71 Li B H, Yuan Y, Liang D K, et al. When counting meets HMER: counting-aware network for handwritten mathematical expression recognition. In: Proceedings of the European Conference on Computer Vision, 2022. 197–214
72 He J B, Wang L, Hu Y P, et al. ICL-D3IE: in-context learning with diverse demonstrations updating for document information extraction. In: Proceedings of the IEEE International Conference on Computer Vision, 2023. 19485–19494
73 Park S, Shin S, Lee B, et al. CORD: a consolidated receipt dataset for post-ocr parsing. In: Proceedings of the Document Intelligence Workshop at Neural Information Processing Systems, 2019
74 Li B, Fang G X, Yang Y, et al. Evaluating ChatGPT's information extraction capabilities: an assessment of performance,

explainability, calibration, and faithfulness. 2023. ArXiv:2304.11633
75 Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. 2023. ArXiv:2305.09617
76 Li C Y, Wong C, Zhang S, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
77 OpenBMB. MiniCPM-V 2.6. 2024. https://huggingface.co/openbmb/MiniCPM-V-2_6
78 Tong S, Brown E, Wu P, et al. Cambrian-1: a fully open, vision-centric exploration of multimodal LLMs. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems, 2024
79 Chen Z, Wang W Y, Tian H, et al. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. 2024. ArXiv:2404.16821
80 Beyer L, Steiner A, Pinto A S, et al. PaliGemma: a versatile 3B VLM for transfer. 2024. ArXiv:2407.07726
81 CloudWalk. Congrong. 2024. https://www.cloudwalk.com
82 Wang W H, Lv Q, Yu W M, et al. CogVLM: visual expert for pretrained language models. 2023. ArXiv:2311.03079
83 OpenBMB. MiniCPM-V-2. 2024. https://huggingface.co/openbmb/MiniCPM-V-2
84 Huang M X, Liu Y L, Liang D K, et al. Mini-Monkey: alleviate the sawtooth effect by multi-scale adaptive cropping. 2024. ArXiv:2408.02034
85 Anthropic. Claude3.5-sonnet. 2024. https://docs.anthropic.com/en/docs/build-with-claude/vision
86 Liu H T, Li C Y, Li Y H, et al. LLaVa-NeXT: improved reasoning, OCR, and world knowledge. 2024. https://llava-vl.github.io/blog/2024-01-30-llava-next
87 OpenAI. GPT-4o-mini-20240718. 2024. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence
88 Dong X Y, Zhang P, Zang Y H, et al. InternLM-XComposer2: mastering free-form text-image composition and comprehension in vision-language large model. 2024. ArXiv:2401.16420
89 Reka AI. Rekaflash. 2024. https://www.reka.ai
90 Google. Gemini models. 2024. https://deepmind.google/technologies/gemini
91 XVERSE. XVERSE-V. 2024. https://github.com/xverse-ai/XVERSE-V-13B
92 Lu S Y, Li Y, Chen Q-G, et al. Ovis: structural embedding alignment for multimodal large language model. 2024. ArXiv:2405.20797
93 Bai J Z, Bai S, Tan S S, et al. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond. 2023. ArXiv:2308.12966
94 OpenBMB. MiniCPM-Llama3-V2.5. 2024. https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5
95 Sun Q, Cui Y F, Zhang X S, et al. Generative multimodal models are in-context learners. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2024. 14398–14409
96 Lu H Y, Liu W, Zhang B, et al. DeepSeek-VL: towards real-world vision-language understanding. 2024. ArXiv:2403.05525
97 Anthropic. Claude3. 2024. https://docs.anthropic.com/en/docs/build-with-claude/vision
98 OpenBMB. OmniLMM-12B. 2024. https://huggingface.co/openbmb/OmniLMM-12B
99 PCI Research. TransCore-M. 2024. https://github.com/PCIResearch/TransCore-M
100 Zhang P, Dong X Y, Zang Y H, et al. InternLM-XComposer-2.5: a versatile large vision language model supporting long-contextual input and output. 2024. ArXiv:2407.03320
101 XTuner Contributors. XTuner: a toolkit for efficiently fine-tuning LLM. 2023. https://github.com/InternLM/xtuner
102 Chen L, Li J S, Dong X Y, et al. ShareGPT4V: improving large multi-modal models with better captions. In: Proceedings of the European Conference on Computer Vision, 2024
103 QiHoo360. 360VL. 2024. https://github.com/360CVGroup/360VL
104 Dong X Y, Zhang P, Zang Y H, et al. InternLM-XComposer2-4KHD: a pioneering large vision-language model handling resolutions from 336 pixels to 4K HD. 2024. ArXiv:2404.06512
105 OpenBMB. MiniCPM-V. 2024. https://huggingface.co/openbmb/MiniCPM-V
106 AI 01. Yi-VL-34B. 2024. https://huggingface.co/01-ai/Yi-VL-34B
107 Laurencon H, Saulnier L, Tronchon L, et al. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
108 DataCanvas Ltd. MMAlaya. 2024. https://github.com/DataCanvasIO/MMAlaya
109 Microsoft. Phi-3-vision. 2024. https://huggingface.co/microsoft/Phi-3-vision-128k-instruct
110 Du Z X, Qian Y J, Liu X, et al. GLM: general language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. 320–335
111 Awadalla A, Gao I. OpenFlamingo v2: new models and enhanced training setup. https://laion.ai/blog/open-flamingo-v2, 2023
112 Laurencon H, Tronchon L, Cord M, et al. What matters when building vision-language models? 2024. ArXiv:2405.02246
113 Su Y X, Lan T, Li H Y, et al. PandaGPT: one model to instruction-follow them all. 2023. ArXiv:2305.16355
114 StepFun. Step-1v. https://platform.stepfun.com, 2024
115 Team Chameleon. Chameleon: mixed-modal early-fusion foundation models. 2024. ArXiv:2405.09818

Appendix A Supervised SOTA

State-of-the-art (SOTA) in widely-adopted OCR text recognition datasets is achieved by PARSeq [61], which utilizes permutation language modeling for enhanced contextual information and unifies context-free non-AR and context-aware AR inference. We train PARSeq on the ST[1)] and MJ[2)] synthetic datasets, and test it directly on real datasets. The ReCTS dataset is not commonly used, therefore, we chose TPS-ResNet [63], which has the highest ranking on the ICDAR ranking table along with a published article, as the SOTA. The model employs thin-plate-spline (TPS)-based spatial transformer network (STN) to normalize the input text images, followed by a ResNet-based feature extractor and BiLSTM for text recognition. Although the evaluation metric for ReCTS is normalized edit distance, we still use it as a reference. AttentionHTR [62] proposes an attention-based sequence-to-sequence model to achieve SOTA on IAM. The word error rate of this method on IAM is 8.76, so we use 91.24 as the corresponding word

1) Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2315–2324.
2) Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic data and artificial neural networks for natural scene text recognition. 2014. ArXiv:1406.2227.

Table B1 Details of the models

Model	Language model	Input resolution
BLIP2-OPT-6.7B	OPT-6.7B[a)]	224
mPLUG-Owl	LLaMA-2 7B[b)]	224
InstructBLIP	Vicuna-7B [5]	224
LLaVAR	Vicuna-7B [5]	336
BLIVA	Vicuna-7B [5]	224
mPLUG-Owl2	LLaMA-2 7B[b)]	448
LLaVA1.5-7B	Vicuna-7B [5]	336
LLaVA1.5-13B	Vicuna-13B [5]	336
MiniGPT4-v2	Llama2-Chat-7B[b)]	224
Monkey	Qwen-7B[c)]	896
Unidoc	Vicuna [5]	336
DocPedia	Vicuna [5]	2560
GPT4V	gpt-4-vision-preview	
Gemini	gemini-pro-vision	

a) Zhang S, Roller S, Goyal N, et al. OPT: open pre-trained transformer language models. 2022. ArXiv:2205.01068.
b) Touvron H, Martin L, Stone K, et al. LLaMA 2: open foundation and fine-tuned chat models. 2023. ArXiv:2307.09288.
c) Bai J Z, Bai S, Chu Y F, et al. Qwen technical report. 2023. ArXiv:2309.16609.

accuracy. Yu et al. [64] achieves SOTA on ORAND-CAR-2014 by proposing an efficient text line recognition method based on prototype learning with feature-level sliding windows for classification.

For scene text visual question answering (STVQA) and optical character recognition visual question answering (OCR-VQA), SOTA is reached by GIT [14]. GIT trains a generative image-to-text transformer which contains one image encoder and one text decoder under a single language modeling task. For TextVQA, Mia [65] achieves SOTA by using T5[3)] for TextVQA task. To align object feature and scene text, Mia is pretrained by masked language modeling (MLM) and relative position prediction (RPP) task. For the DocVQA dataset, BAIDU-DI assembles ERNIE-Layout [67] and DocPrompt (a few-shot model using multi-stage training based on ERNIE-Layout) to achieve SOTA. ESTVQA dataset has not been explored well, Fang et al. [66] achieved SOTA using a cross-modal attention network to guide the expression of visual and textual features simultaneously while using a transformer decoder to output the results. However, since Fang et al. [66] did not provide separate results on Chinese and English data, the average performance is used as a reference. For InfographicVQA, SOTA is reached by DUBLIN [68]. For ChartQA, Liu et al. [69] achieved SOTA by designing a modality conversion module DePlot.

The SOTA on the FUNSD dataset is achieved by ERNIE-Layout [67], which enhances layout knowledge by correcting the reading order in the pretraining phase. For SROIE, StrucTexT [70] achieves the SOTA. StrucTexT introduces a segment-token aligned encoder for the transformer and is pretrained by masked visual language modeling tasks and the new sentence length prediction and paired boxes direction tasks. For POIE, Kuang et al. [60] achieved SOTA by adopting contrastive learning to effectively establish the connections between the tasks of OCR and information extraction. For HME100K, Li et al. [71] proposed a counting-aware network which jointly optimizes HMER and symbol counting tasks to achieve SOTA.

Appendix B Summary of the models

Table B1 displays the details of the testing models, with most LMMs having their resolution limited to an input size of 224 or 336 due to the visual module. As a result, they perform poorly on Doc-oriented VQA and KIE tasks. Conversely, models with higher resolutions exhibit better performance on these tasks.

Appendix C Summary of the evaluation benchmarks

Table C1 presents the detailed composition of OCRBench. The text recognition task consists of a total of 300 images, including Regular Text Recognition, Irregular Text Recognition, Artistic Text Recognition, Handwriting Recognition, Digit String Recognition, and Non-Semantic Text Recognition, with 50 images each. The scene text-centric VQA task includes 200 questions, with 50 questions each from STVQA, TextVQA, ESTVQA(EN), and OCR-VQA datasets. The Doc-oriented VQA task comprises 200 questions, with 50 questions each from DocVQA, ChartQA(Aug.), ChartQA(Hum.), and InfoVQA datasets. The key information extraction task consists of 200 questions, with 67 questions from SROIE, 66 questions from FUNSD, and 67 questions from POIE. The Handwritten mathematical expression recognition task includes 100 images from the HME100K dataset. OCRBench encompasses a total of 1000 questions across these five tasks, and all answers have been manually filtered and corrected.

3) Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res, 2020, 21: 1–67.

Table C1 Details of the OCRBench

Task	Dataset	Smaples
Text recognition	Regular Text Recognition	50
	Irregular Text Recognition	50
	Artistic Text Recognition	50
	Handwriting Recognition	50
	Digit String Recognition	50
	Non-Semantic Text Recognition	50
Scene text-centric VQA	STVQA	50
	TextVQA	50
	ESTVQA(EN)	50
	OCR-VQA	50
Doc-oriented VQA	DocVQA	50
	ChartQA(Aug.)	50
	ChartQA(Hum.)	50
	InfoVQA	50
Key information extraction	SROIE	67
	FUNSD	66
	POIE	67
HMER	HME100K	100
All	–	1000