InstructOCR2: a lightweight and efficient multi-modal language model for document understanding

Anonymous ACL submission

Abstract

In recent years, there has been significant interest in using Multi-modal Large Language Models (MLLMs) for OCR tasks, leading to the development of MLLMs specifically designed for the OCR domain. The majority of existing approaches focus on developing larger and more sophisticated models, which demand substantial computational resources for training and deployment. Furthermore, these methods often fail to achieve effective alignment between text and its corresponding positions within the image. Some approaches merely feed all text directly into the model, while others, despite incorporating coordinate information, still struggle to accurately capture the precise location and contextual relationships of text within images. In this paper, we propose a lightweight multi-modal language model called InstructOCR2, which achieves multi-scene and multi-task OCR recognition with fewer parameters. InstructOCR2 enhances the model's comprehension of global and local text through fine-grained alignment of text and images, thereby improving the performance of downstream tasks such as Visual Question Answering (VQA) and Key Information Extraction (KIE).

1 Introduction

Optical Character Recognition (OCR) is a crucial technology in the fields of computer vision and natural language processing, with widespread applications in document digitization, automated data entry, and information retrieval. Traditional OCR methods typically focus on single tasks, each presenting unique challenges. For instance, Text Spotting (TS) requires handling complex backgrounds and diverse fonts, Visual Question Answering (VQA) necessitates understanding text content and answering related questions, while Key Information Extraction (KIE) demands extracting specific information. However, these single-task approaches often fall short in addressing the complexities of real-world applications.

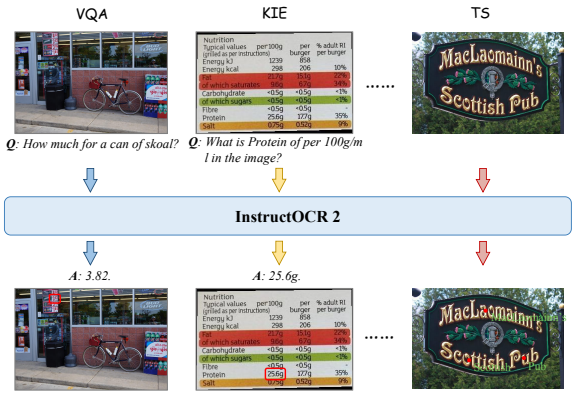

Figure 1: An overview of the capabilities of InstructOCR2 across various image understanding tasks is provided. The accompanying figure illustrates the application of our proposed lightweight multi-modal language model in Visual Question Answering (VQA), Key Information Extraction (KIE), and Text Spotting (TS).

With the rapid development of Large Language Models (LLMs) (Achiam et al., 2023; Bai et al., 2023; Yang et al., 2023; Touvron et al., 2023; Brown et al., 2020; Zhang et al., 2022), a series of Multi-modal Large Language Models (MLLMs) have emerged (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2024b; Zhu et al., 2023; Zhang et al., 2023). These MLLMs, which integrate visual and linguistic information, are better equipped to understand and process textual content within images, examples of which include (Liu et al., 2024b; Chen et al., 2023; Ye et al., 2023d; Li et al., 2024a). By pretraining on large-scale image-text data, these models can capture the complex relationships between images and text, enabling them to excel in a wide range of general vision tasks. General-purpose MLLMs emphasize task generalization, whereas OCR tasks place greater importance on resolution and corresponding training data. Consequently, some MLLMs (Liu et al., 2024c; Feng

et al., 2023b; Liao et al., 2024) specifically tailored for OCR tasks have emerged. These OCR-specific MLLMs enhance their performance through methods such as expanding the input resolutions and utilizing MLLMs instruction tuning datasets.

Despite the powerful capabilities of MLLMs in OCR tasks, their large parameter sizes and high demands for extensive image-text data pose significant computational and resource challenges. To address these challenges, lightweight multi-modal models have gradually gained attention. These models (Wei et al., 2024b; Xiao et al., 2024; Wei et al., 2024a) aim to reduce computational and storage requirements while maintaining high performance by decreasing the number of parameters and optimizing architectural design. However, current lightweight multi-modal models often lag behind MLLMs in terms of accuracy and robustness when addressing complex OCR tasks. Moreover, there remains significant potential for further reduction in parameter sizes.

In this paper, we propose InstructOCR2, a novel training framework for lightweight multi-modal language models, featuring 284M parameters. We enhance the model's perception of OCR text by emphasizing alignment mechanisms, which is fundamental to various downstream OCR tasks. The training of the InstructOCR2 consists of two stages. In the first stage, we use scene text spotting as a pretraining task. This task requires the model not only to recognize text in images but also to perceive the specific locations of the text within the images. Through this approach, the model learns the transformation relationship from image to sequence, i.e., by extracting serialized text data from visual information, thereby better understanding the alignment between images and text.

In the second stage, we train the model with a large amount of instruction data, enabling it to understand and execute various downstream tasks. This data includes instructions for different tasks along with corresponding input-output examples. Through this method, the model can not only recognize text in images but also complete specific tasks based on the instructions, such as TS, VQA, as shown in Figure 1. This stage of training endows the model with greater task generalization and flexibility. Through the aforementioned two-stage pretraining, our InstructOCR2 framework significantly improves the accuracy and robustness of the model in OCR tasks while maintaining a small parameter size.

In summary, the main contributions are threefold:

1. We propose a lightweight and efficient multi-modal framework called InstructOCR2, featuring only 284M parameters and supporting a maximum output length of 4096 tokens. This framework can accomplish various tasks of multi-modal models, such as Text Spotting (TS), Visual Question Answering (VQA), and Key Information Extraction (KIE).

2. We propose a local-global alignment approach. By performing an image-to-sequence generation task that simultaneously predicts the text and its corresponding position within the image, our approach achieves precise alignment, and through full-document recognition, enables the model to possess contextual capabilities.

3. Experimental results on public datasets demonstrate that InstructOCR2 exhibits outstanding performance and surpasses existing methods in a series of downstream tasks. It is even competitive when compared to the results of MLLMs.

2 Related Work

2.1 Multi-modal Large Language Models

The rapid development and exceptional performance of MLLMs have inspired researchers to explore the potential and applications of MLLMs in Optical Character Recognition (OCR) tasks, thereby driving a series of related works.

UniDoc (Feng et al., 2023b) begins with the data, performing unified multi-modal instruction tuning on the contributed large-scale instruction-following datasets. Monkey (Li et al., 2024a) divides input images into uniform patches and supports resolutions up to 1344×896 pixels, which allows for a more detailed capture of visuals. Textmonkey (Liu et al., 2024c) adopts Shifted Window Attention to incorporate cross-window connectivity while expanding the input resolutions, and reduces the token length through token compression. UReader (Ye et al., 2023b) designs a shape-adaptive cropping module to process high-resolution images and develops auxiliary tasks for text reading and key points generation to enhance text recognition and semantic understanding capabilities. To tackle the challenge of resolution, DocPedia (Feng

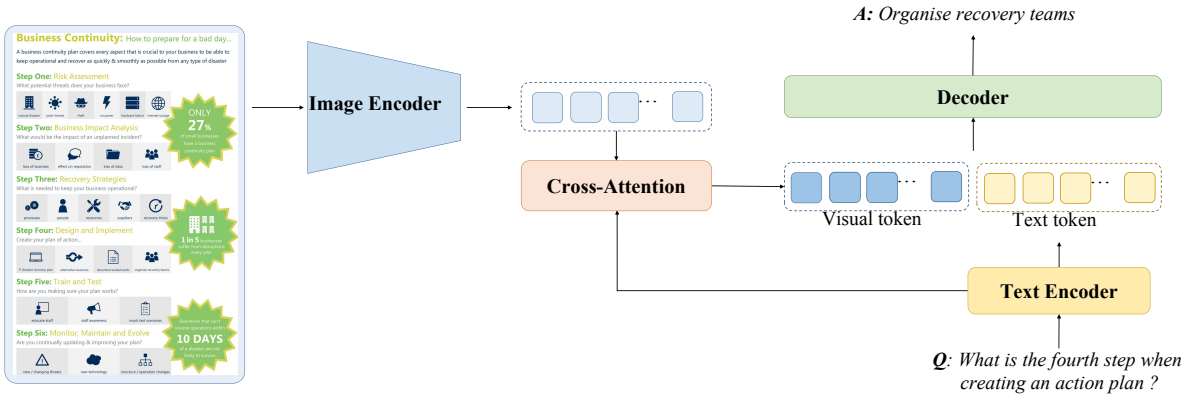

Figure 2: The main framework of InstructOCR2 consists of two distinct input branches: an image encoder and a text encoder. The image encoder is responsible for processing visual features, while the text encoder handles textual features. These extracted features are then fed into the decoder to produce the final results.

et al., 2023a) processes visual input in the frequency domain rather than the pixel space to capture a greater amount of visual and textual information. mPLUG-DocOwl (Ye et al., 2023a) and mPLUG-DocOwl1.5 (Hu et al., 2024a) are based on mPLUG-Owl (Ye et al., 2023d) and further strengthens the ability to understand OCR-free documents.

2.2 Document Understanding

Document understanding methods can be broadly categorized into two types based on whether they use OCR systems for text extraction: OCR-dependent methods (Appalaraju et al., 2021; Huang et al., 2022; Powalski et al., 2021; Wang et al., 2023a; Xu et al., 2020) and OCR-free methods (Davis et al., 2022; Lee et al., 2023). OCR-dependent methods achieve document understanding by inputting pre-extracted OCR text, layout, and other information into language models. For example, UDOP (Tang et al., 2023) inputs text, image, and layout modalities into the decoder, establishing aligned representations of spatial and textual embeddings. However, this approach relies on OCR systems and is susceptible to errors from these OCR systems. Additionally, processing the entire document may lead to unnecessary computation, as some tasks are only related to specific regions of the document.

OCR-free methods do not require OCR input and perform document understanding tasks in an end-to-end manner. Donut (Kim et al., 2022) directly maps an input document image into a desired structured output and can be trained in an end-to-end manner. VisFocus (Abramovich et al., 2024) proposes an OCR-free method to better exploit the vision encoder's capacity by coupling it directly with the language prompt. These OCR-free methods are parameter-efficient; for example, VisFocus has a total of 408M parameters, yet they exhibit limited generalization and can lack the capability to handle more downstream tasks. Although MLLMs is versatile and can handle various downstream tasks, the large number of parameters presents challenges in both training and deployment. Our proposed InstructOCR2 possesses the capabilities of MLLMs while maintaining a reduced parameter count. And InstructOCR2 supports a maximum output token length of 4096, surpassing that of other models, such as SCOB (Kim et al., 2023), which only outputs 512 tokens.

3 Method

We introduce InstructOCR2, an end-to-end document understanding framework. The overall structure is shown in Figure 2. In the following sections, we will detail the structure of our model.

3.1 Architecture

Text Encoder. The text encoder of InstructOCR2 adopts the T5 small model (Raffel et al., 2020), with a maximum input length of 512 tokens. We use the encoder of T5 to encode text features, consisting of only 6 layers of transformers.

Image Encoder. The image encoder utilizes the ResNet50 (He et al., 2016) architecture to extract features from the input image, initialized with the ODM (Duan et al., 2024a) weights. By applying cross-attention between the extracted visual and textual features, the model can better comprehend and capture the contextual relationships between

them. This enhancement of context leads to improved performance in multi-modal tasks.

Decoder. The language model of InstructOCR2 adopts LongT5 base (Guo et al., 2021), which is an extension of the T5 model that handles long sequence inputs more efficiently, with a maximum processing length of 4096 tokens. To achieve precise alignment of text and position through the image-to-sequence generation task, we introduce a custom token vocabulary. This includes a special separator token $< sep >$ and 1000 position tokens. These additions enable the model to better manage and differentiate text and position within the input sequences.

3.2 Training Strategy

The training strategy of InstructOCR2 consists of two components: alignment and instruction tuning.

Alignment. The image-to-sequence generation task is used to achieve precise alignment of text and position. In the sequence representation, each text instance is represented by a sequence consisting of three parts: $[x, y, t]$, where (x, y) denotes the coordinates of the center point, and t represents the transcription text. Text instances are separated by the token $< sep >$. Additionally, the tokens $< SOS >$ and $< EOS >$ are inserted at the beginning and end of the sequence, respectively, to indicate the start and end of the sequence. In this stage of training, the input prompt to the text encoder remains as *"Recognize text in the image, provide text coordinates and text recognition results"*.

By employing the training strategy of image-to-sequence generation, InstructOCR2 is equipped with the ability to perform the text spotting task, capable of predicting all the text and corresponding positions in an image. While this training strategy provides the model with word-level sensitivity, it inherently lacks comprehensive contextual predictive capabilities due to the absence of document-level context training. By incorporating document-level recognition, the model gains enhanced contextual capabilities. The dataset used for this purpose is DocGenome (Xia et al., 2024). During this training stage, the input prompt to the text encoder is *"Recognize all text in the image"*.

Instruction tuning. In this stage, instruction tuning enables the model with VQA capabilities. InstructOCR (Duan et al., 2024b) proposes a set of instructions meticulously designed based on text attributes. This method facilitates the efficient acquisition of large amounts of VQA data without requiring manual annotation. We first apply this method to train the model's instruction-tuning capability. Then, we train the model using collected public VQA datasets. Additionally, we consider the text spotting task as a type of VQA task, with the input prompt being *"recognize text in the image, provide text coordinates and text recognition results"*.

3.3 Loss Function

In InstructOCR2, the training objective is to predict tokens, and we utilize the standard cross-entropy loss for model training. This loss function aims to maximize the likelihood of the correct tokens during training. The mathematical expression of the cross-entropy loss is as follows:

$$\mathcal{L}_{seq} = \text{maximize} \sum_{i=1}^{L} w_i \log P(\tilde{s}_i | I, s_{1:i}) \quad (1)$$

where I is the input image, s is the input sequence, $\tilde{s}$ is the output sequence, L is the length of the sequence, and w_i is the weight of the likelihood of the $i-th$ token, which is empirically set to 1.

4 Experiment

4.1 Datasets

Alignment. In this training stage, we use text spotting data from both documents and natural scenes, with a total training dataset of 2.44M. Specifically, for the document data, we randomly sample 1.33M images from the IIT-CDIP (Lewis et al., 2006) dataset and employ PPOCRv3 (Li et al., 2022) to generate pseudo labels (i.e., text and position in the image). We also utilize training sets from the following document datasets: DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), and ChartQA (Masry et al., 2022). The natural scene data includes the following datasets: Total-Text (Ch'ng and Chan, 2017), SCUT-CTW1500 (Yuliang et al., 2017), ICDAR2015 (Karatzas et al., 2015), ICDAR2013 (Karatzas et al., 2013), ICDAR2017 MLT (Nayef et al., 2017b), Curved Synthetic Dataset 150k (Liu et al., 2020), TextOCR (Singh et al., 2021), HierText (Long et al., 2022), and OpenVINO (Krylov et al., 2021). For context alignment training, we randomly sample 0.69M images from the DocGenome (Xia et al., 2024) dataset to enhance document-level recognition capabilities.

Model	Size	DocVQA	InfoVQA	DeepForm	KLC	ChartQA	WTQ	TabFact
DocPeida	7.1B	47.1	15.2	-	-	46.9	-	-
DocOwl	7.3B	62.2	38.2	42.6	30.3	57.4	26.9	67.6
UReader	7.1B	65.4	42.2	49.5	32.8	59.3	29.4	67.6
DocKylin	7.1B	77.3	46.6	-	-	66.8	32.4	-
Qwen-vl	9.6B	62.6	-	-	-	66.3	-	-
Monkey	9.8B	66.5	36.1	40.6	-	65.1	25.3	-
TextMonkey	9.7B	66.7	28.6	61.6	37.8	66.9	31.9	-
HRVDA	7.1B	72.1	43.5	63.2	37.5	67.6	31.2	72.3
DocLayLLM	8B	86.5	58.4	77.1	40.7	-	58.6	83.4
KOSMOS-2.5	1.3B	81.1	41.3	65.8	35.1	62.3	32.4	49.9
TextHawk2	7.4B	89.6	67.8	-	-	81.4	46.2	78.1
InternVL2	8.1B	91.6	74.8	-	-	83.3	-	-
Dessurt	127M	63.2	-	-	-	-	-	-
Donut	176M	67.5	11.6	61.6	30.0	41.8	18.8	54.6
Pix2Struct	282M	72.1	38.2	-	-	56.0	-	-
VisFocus	408M	72.9	31.9	-	-	57.1	-	-
InstructOCR2	284M	64.8	26.0	67.7	35.0	57.9	19.9	55.7

Table 1: Comparison with Multi-modal Large Language Models(MLLMs) and OCR-free document understanding methods on various types of document image understanding tasks. All evaluation benchmarks use the officially designated metrics. “size" refers to the number of parameters in the model. The MLLMs public benchmark includes DocPeida (Feng et al., 2023a), DocOwl (Ye et al., 2023d), UReader (Ye et al., 2023c), DocKylin (Zhang et al., 2024), Qwen-vl (Bai et al., 2023), Monkey (Li et al., 2024b), TextMonkey (Liu et al., 2024c), HRVDA (Liu et al., 2024a), DocLayLLM (Liao et al., 2024), KOSMOS-2.5 (Lv et al., 2023), TextHawk2 (Yu et al., 2024), InternVL2 (Chen et al., 2024). The OCR-free document understanding methods include Dessurt (Davis et al., 2022), Donut (Kim et al., 2022), Pix2Struct (Lee et al., 2023), VisFocus (Abramovich et al., 2024)

Instruction tuning. In this training stage, we utilize a diverse set of datasets to enhance the model's ability to understand and execute instructions across various domains, with a total training dataset of 9.2M. These include Docmatix (Laurençon et al., 2024), DocReason25k (Hu et al., 2024a), Sujet-Finance (AI, 2025), ai2d (Hiippala et al., 2021), figqa (Liu et al., 2022), HME100k (Yuan et al., 2022), CROHME 2014 (Mouchere et al., 2014), CROHME 2016 (Mouchère et al., 2016), CROHME 2019 (Mahdavi et al., 2019), UniMER-1M (Wang et al., 2024), SPE, CPE, SCE, Latex-OCR (Blecher, 2022), IAM Handwriting (Marti and Bunke, 2002), HCTR (Stamatopoulos et al., 2013), Synthdog-en (Kim et al., 2022), TableBench (Wu et al., 2024), TableVQA (Kim et al., 2024), TabMWP (Lu et al., 2023) and UniChart (Masry et al., 2023).

After being trained on large-scale VQA data, the model gains the ability to accept instructions in natural language. We then further fine-tune the model using the training sets of downstream tasks. Additionally, we utilize another dataset of 1.66M samples for training during this stage. These include document datasets such as DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), DeepForm (Svetlichnaya, 2020), OCR-VQA (Mishra et al., 2019), KLC (Stanisławek et al., 2021), DocGenome (Xia et al., 2024) and VisualMRC (Tanaka et al., 2021). Table datasets such as TableFact (Chen et al., 2019) and WikiTableQuestions (Pasupat and Liang, 2015). Chart datasets include ChartQA (Masry et al., 2022), ChartBench (Xu et al., 2023) and DVQA (Kafle et al., 2018). Natural scene datasets include TextVQA (Singh et al., 2019), ST-VQA (Biten et al., 2019), ic13 (Karatzas et al., 2013), ic15 (Karatzas et al., 2015), Total-Text (Ch'ng and Chan, 2017), TextOCR (Singh et al., 2021), Curved Synthetic Dataset 150k (Liu et al., 2020), MLT-2017 (Nayef et al., 2017a), HierText (Long et al., 2022) and TextCaps (Sidorov et al., 2020). KIE datasets include FUNSD (Jaume et al., 2019), POIE (Kuang et al., 2023) and

SROIE (Huang et al., 2019).

4.2 Implementation Details

The entire model is distributively trained on 32 NVIDIA A100-80G GPUs. During the training process in the alignment stage, to enhance training efficiency, the short side of the input image is randomly resized to a range from 704 to 1024 (intervals of 32), and the maximum length of the image is set to 1024. The batch size per GPU is 5, and the model is trained for 150 epochs, with an initial 5-epoch warm-up phase. We use the AdamW optimizer with a learning rate of 4.6×10^{-4}. Subsequently, the model is trained for another 50 epochs, with a fixed learning rate of 6×10^{-5}, and the maximum length of image is set as 1920. Then, the model's text reading capability is refined using the DocGenome dataset. And the model is further trained for another 30 epochs. For instruction tuning, we first fine-tune the model for 10 epochs on the text spotting data using the instructions from InstructOCR. Then, we use 9.2 million samples to equip the model with interaction capabilities, training for 15 epochs in this stage. The model is then fine-tuned using downstream data, with training conducted for 40 epochs during this phase.

4.3 Comparison with Results on Document Benchmarks

InstructOCR2 can perform VQA tasks in scenarios such as documents, charts, and tables. Compared to previous OCR-free methods, our approach is more comprehensive. We compare our method with recent MLLMs and OCR-free document understanding methods. At the inference stage, the maximum input size is set to 1920 pixels, and the minimum input size is set to 1280 pixels. As shown in Table 1, our method achieved 64.8% on the DocVQA dataset, surpassing MLLMs such as DocPeida, DocOwl, and Qwen-vl, as well as document understanding methods like Dessurt. The metrics on the InfoVQA dataset surpass those of DocPeida and Donut. The DeepForm dataset achieves state-of-the-art (SOTA) performance among OCR-free document understanding methods, achieving a position just below DocLayLLM in comparison to MLLMs. The metrics for the KLC, ChartQA, WTQ, and TabFact datasets also surpass those of previous OCR-free document understanding methods.

Table 1 presents the results of the KIE task on the DeepForm and KLC datasets. Our method achieves state-of-the-art (SOTA) performance among OCR-free document understanding methods and surpasses several MLLMs, such as UReader and HRVDA. To further demonstrate the effectiveness of our approach in document understanding, we evaluate the model on the FUNSD, SROIE, and POIE datasets. As shown in Table 2, our method is only slightly lower than Mini-Monkey on the FUNSD and SROIE datasets, while achieving SOTA performance on the POIE dataset, demonstrating the effectiveness of our proposed method for KIE tasks.

Model	Size	FUNSD	SROIE	POIE
DocOwl	7.3B	0.5	1.7	2.5
LLaVA1.5	7.3B	0.2	1.7	2.5
TGDoc	7B	1.4	3.0	22.2
InternVL	13B	6.5	26.4	25.9
DocPeida	7.1B	29.9	21.4	39.9
Monkey	9.8B	24.1	41.9	19.9
TextMonkey	9.7B	32.3	47.0	27.9
Mini-Monkey	2B	42.9	70.3	69.9
InstructOCR2	284M	37.2	73.2	**78.8**

Table 2: The results of our proposed method for Key Information Extraction(KIE) are presented alongside the public benchmark of MLLMs, which includes DocOwl (Ye et al., 2023d), LLaVA1.5 (Liu et al., 2024b), TGDoc (Wang et al., 2023b), InternVL (Chen et al., 2024), DocPeida (Feng et al., 2023a), Monkey (Li et al., 2024b), TextMonkey (Liu et al., 2024c), and Mini-Monkey (Huang et al., 2024).

Model	Size	Overall
BLIP2-6.7B	6.7B	235
InstructBLIP	7B	276
mPLUG-Owl	7B	297
BLIVA	7B	291
InternLM-XComposer	7B	303
LLaVA1.5-13B	13B	331
TextMonkey	9.7B	561
MiniCPM-V2.6	7B	<u>852</u>
InstructOCR2	284M	**357**

Table 3: The results of our proposed method on OCRBench are compared with the following methods: BLIP2-6.7B (Li et al., 2023), InstructBLIP (Dai et al., 2023), mPLUG-Owl (Ye et al., 2023d), BLIVA (Hu et al., 2024b), InternLM-XComposer (Dong et al., 2024), LLaVA1.5-13B (Liu et al., 2023a), TextMonkey (Liu et al., 2024c), and MiniCPM-V2.6 (Yao et al., 2024).

Model	Size	ST-VQA	$\text{TextVQA}_{\text{Val}}$
BLIP2-OPT	6.7B	20.9	23.5
mPLUG-Owl	7.3B	30.5	34.0
DocPeida	7.1B	45.5	60.2
DocOwl	7.3B	-	52.6
UReader	7.1B	-	57.6
KOSMOS-2.5	1.3B	-	40.7
Monkey	9.8B	67.7	67.6
TextMonkey	9.7B	61.8	65.6
Dessurt	127M	63.2	-
Donut	176M	-	43.5
InstructOCR	78M	45.8	42.0
InstructOCR2	284M	51.5	41.8

Table 4: The results of our proposed method on scene text VQA. The MLLMs public benchmark includes BLIP2-OPT (Li et al., 2023), mPLUG-Owl (Ye et al., 2023d), DocPeida (Feng et al., 2023a), DocOwl (Ye et al., 2023d), UReader (Ye et al., 2023c), KOSMOS-2.5 (Lv et al., 2023), Monkey (Li et al., 2024b), TextMonkey (Liu et al., 2024c). The OCR-free document understanding methods include Dessurt (Davis et al., 2022), Donut (Kim et al., 2022), InstructOCR (Duan et al., 2024b).

4.4 Comparison with OCRBench Results

To further evaluate the performance of our method in document understanding, we assess the results on OCRBench (a comprehensive benchmark encompassing 29 OCR-related evaluations). This represents a capability that prior OCR-free methods, including Dessurt, Pix2Struct, and VisFocus, have been unable to achieve. As shown in Table 3, our method even surpasses LLaVA1.5-13B, which has 13 B parameters.

4.5 Comparison with Scene Text Visual Question Answering Results

InstructOCR2 is capable of comprehending both documents and natural scene images. Table 4 presents the results on the ST-VQA and TextVQA datasets. As observed in the table, InstructOCR2 surpasses MLLMs such as mPLUG-Owl, KOSMOS-2.5 and BLIP2-OPT.

4.6 Comparison with Text Spotting Results on the VQA Task

To demonstrate the extensive capabilities of InstructOCR2, we evaluate its performance on text spotting datasets without fine-tuning. During the inference stage, the prompt input to the text encoder is *"Recognize text in the image, provide text coordinates and text recognition results"*. The maximum length of the image is shorter than 1920 pixels, and the minimum is 1024 pixels. We evaluate the model using the point-based metric proposed in SPTS. Specifically, ICDAR2015 is a multi-oriented text dataset, while Total-Text is an arbitrarily shaped text dataset.

Table 5 shows the results of the text spotting task. Compared to TextMonkey, we surpass it by 10.3% on the Total-Text and by 17.8% on the ICDAR2015, achieving better performance with our lightweight model compared to the 9.7B model. This demonstrates the superiority of our method in position awareness.

Methods	Total-Text		ICDAR2015		
	None	Full	S	W	G
TextMonkey	61.4	-	-	-	45.1
InstructOCR2	**71.7**	75.7	65.8	64.5	**62.9**

Table 5: Text spotting results on Total-Text and ICDAR2015 in the VQA task. 'None' means lexicon-free. 'Full' indicates that we use all the words that appeared in the test set. 'S', 'W', and 'G' represent recognition with 'Strong', 'Weak', and 'Generic' lexicons, respectively. And we use the TextMonkey (Liu et al., 2024c) for comparison.

4.7 Ablation Study

Ablation study on text spotting. We propose an image-to-sequence generation task to achieve precise alignment of text and its corresponding position within the image, which enables the model to effectively execute the text spotting task. In this section, we explore the effectiveness of the text spotting task. Table 6 presents the performance of the model on the text spotting task after the first stage of pre-training. Following the training and evaluation protocols of the scene text spotting task, we fine-tuned the model for 170 epochs separately on the Total-Text and ICDAR2015 datasets, and subsequently evaluated its performance on these datasets.

As observed in Table 6, our model achieves SOTA performance on ICDAR2015 datasets using a generic lexicon, demonstrating the robustness of our pre-training stage, surpassing dedicated models for scene text spotting tasks. However, when evaluated using a lexicon, the performance falls

short of that achieved by scene text spotting methods, suggesting that our model exhibits a reduced frequency of recognition errors, thereby offering limited scope for correction through the lexicon. This indicates that the model tends to accurately recognize entire words, as opposed to the internal character errors often observed in traditional scene text spotting methods.

Methods	Total-Text		ICDAR2015		
	None	Full	S	W	G
TOSS	65.1	74.8	65.9	59.6	52.4
SPTS	74.2	82.4	77.5	70.2	65.8
SPTS-v2	75.5	84.0	82.3	77.7	72.6
InstructOCR	77.1	84.1	82.5	77.1	72.1
InstructOCR2	**76.1**	80.1	77.9	76.1	**74.0**

Table 6: Text spotting results on Total-Text and ICDAR2015. 'None' means lexicon-free. 'Full' indicates that we use all the words that appeared in the test set. 'S', 'W', and 'G' represent recognition with 'Strong', 'Weak', and 'Generic' lexicons, respectively. And we use the following models for comparison: TOSS (Tang et al., 2022), SPTS (Peng et al., 2022), SPTS-V2 (Liu et al., 2023b), and InstructOCR (Duan et al., 2024b).

Ablation study on input resolution. The text within document images is often densely packed, and the images typically have a high resolution. Our model supports a maximum input size of 1920 pixels; thus, we examine the impact of various input resolutions on the metrics. The results are presented in Table 7, with the minimum size set to 1024 and the maximum size increased from 1024 to 1920.

The table indicates that as resolution increases, the metrics improve as well. However, different datasets have distinct requirements for high resolution; for instance, the POIE dataset achieves optimal performance at a resolution of 1760 pixels, whereas TabFact necessitates a resolution of 1920 pixels. This finding provides valuable guidance on input resolution for applications across various scenarios.

Resolution	DF	CQA	TF	TVQA	POIE
1024	51.5	51.4	51.7	38.0	69.1
1280	62.1	53.2	53.6	38.0	72.6
1440	64.4	53.2	54.4	40.0	72.5
1760	64.2	53.7	54.9	40.5	74.1
1920	64.2	53.7	55.3	40.3	73.6

Table 7: Comparison of different input image sizes in various types of document image understanding tasks. The datasets used in this comparison include DF(DeepForm), CQA(ChartQA), TF(TabFact), TVQA(TextVQA), and POIE.

5 Conclusion

In this paper, we introduce InstructOCR2, a lightweight multi-modal language model specifically designed for Optical Character Recognition (OCR) tasks. Our model effectively addresses the limitations of existing multi-modal large language models (MLLMs), which often require extensive computational resources and struggle with aligning text to its corresponding positions within images. By focusing on local-global alignment mechanisms, InstructOCR2 enhances its positional awareness, resulting in improved performance on various downstream tasks. With only 284 million parameters, it demonstrates that high performance can be achieved with efficiency. The model employs a two-stage training process that enables it to recognize text in images while understanding the spatial relationships between text and visual content, which is essential for real-world applications requiring both accuracy and contextual understanding.

6 Limitation

Multi-modal models require large quantities and diverse data during training. Generally, increasing the amount and variety of data significantly enhances model performance. In our pre-training phase, we utilize data from natural scenes; however, this dataset still has room for expansion. Increasing the amount of training data, may lead to further improvements in model performance. We employ the IIT-CDIP dataset in the document domain. Although this dataset offers a certain degree of diversity, we believe that the diversity of document data still requires enhancement. Future research will explore incorporating a broader range of document types to enhance the model's generalization capabilities.

References

Ofir Abramovich, Niv Nayman, Sharon Fogel, Inbal Lavi, Ron Litman, Shahar Tsiper, Royee Tichauer, Srikar Appalaraju, Shai Mazor, and R Manmatha.

2024. Visfocus: Prompt-guided vision encoders for ocr-free dense document understanding. In *European Conference on Computer Vision*, pages 241–259. Springer.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sujet AI. 2025. Sujet finance instruct 177k dataset. https://huggingface.co/datasets/sujet-ai/Sujet-Finance-Instruct-177k. Accessed: 2025-02-13.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Lukas Blecher. 2022. Latex-ocr. https://github.com/LinXueyuanStdio/Data-for-LaTeX_OCR. Accessed: 2022-05-13.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. 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*, pages 24185–24198.

Chee Kheng Ch'ng and Chee Seng Chan. 2017. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 935–942. IEEE.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven CH Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv abs/2305.06500 (2023).

Brian Davis, Bryan Morse, Brian Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. 2022. End-to-end document recognition and understanding with dessurt. In *European Conference on Computer Vision*, pages 280–296. Springer.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*.

Chen Duan, Pei Fu, Shan Guo, Qianyi Jiang, and Xiaoming Wei. 2024a. Odm: A text-image further alignment pre-training approach for scene text detection and spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597.

Chen Duan, Qianyi Jiang, Pei Fu, Jiamin Chen, Shengxi Li, Zining Wang, Shan Guo, and Junfeng Luo. 2024b. Instructocr: Instruction boosting scene text spotting. *arXiv preprint arXiv:2412.15523*.

Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2023a. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*.

Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. 2023b. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. 2021. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. 2024a. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024b. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264.

Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. 2024. Mini-monkey: Alleviate the sawtooth effect by multi-scale adaptive cropping. *arXiv preprint arXiv:2408.02034*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. 2015. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE.

Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. 2013. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE.

Daehee Kim, Yoonsik Kim, DongHyun Kim, Yumin Lim, Geewook Kim, and Taeho Kil. 2023. Scob: Universal text understanding via character-wise supervised contrastive learning with online text rendering for bridging domain gap. In *International Conference on Computer Vision (ICCV2023)*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Ilya Krylov, Sergei Nosov, and Vladislav Sovrasov. 2021. Open images v5 text annotation and yet another mask text spotter. In *Asian Conference on Machine Learning*, pages 379–389. PMLR.

Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. 2023. Visual information extraction in the wild: practical dataset and end-to-end solution. In *International Conference on Document Analysis and Recognition*, pages 36–53. Springer.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 665–666.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al. 2022. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*.

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

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024a. 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*, pages 26763–26773.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024b. 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*, pages 26763–26773.

Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. 2024. Doclayllm: An efficient and effective multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*.

Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. 2024a. Hrvda: High-resolution visual document assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15534–15545.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. 2020. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9809–9818.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024c. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

Yuliang Liu, Jiaxin Zhang, Dezhi Peng, Mingxin Huang, Xinyu Wang, Jingqun Tang, Can Huang, Dahua Lin, Chunhua Shen, Xiang Bai, et al. 2023b. Spts v2: single-point scene text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. 2022. Towards end-to-end unified scene text detection and layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*.

Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, et al. 2023. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*.

Mahshad Mahdavi, Richard Zanibbi, Harold Mouchere, Christian Viard-Gaudin, and Utpal Garain. 2019. Icdar 2019 crohme+ tfd: Competition on recognition of handwritten mathematical expressions and typeset formula detection. pages 1533–1538.

U-V Marti and Horst Bunke. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5:39–46.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *Preprint*, arXiv:2305.14761.

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

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

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

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952.

Harold Mouchere, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. 2014. Icfhr 2014 competition on recognition of on-line handwritten mathematical expressions (crohme 2014). pages 791–796.

Harold Mouchère, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. 2016. Icfhr2016 crohme: Competition on recognition of online handwritten mathematical expressions. pages 607–612.

Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, Wafa Khlif, Muhammad Muzzamil Luqman, Jean-Christophe Burie, Cheng-lin Liu, and Jean-Marc Ogier. 2017a. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification - rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1454–1459.

Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. 2017b. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1454–1459. IEEE.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*.

Dezhi Peng, Xinyu Wang, Yuliang Liu, Jiaxin Zhang, Mingxin Huang, Songxuan Lai, Jing Li, Shenggao Zhu, Dahua Lin, Chunhua Shen, et al. 2022. Spts: single-point text spotting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4272–4281.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 732–747. Springer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

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

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812.

Nikolaos Stamatopoulos, Basilis Gatos, Georgios Louloudis, Umapada Pal, and Alireza Alaei. 2013. Icdar 2013 handwriting segmentation contest. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1402–1406. IEEE.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2021. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer.

S Svetlichnaya. 2020. Deepform: Understand structured documents at scale.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888.

Jingqun Tang, Su Qiao, Benlei Cui, Yuhang Ma, Sheng Zhang, and Dimitrios Kanoulas. 2022. You can even annotate text with voice: Transcription-only-supervised text spotting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4154–4163.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19254–19264.

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

Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. 2024. Unimernet: A universal network for real-world mathematical expression recognition. *arXiv preprint arXiv:2404.15254*.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2023a.

Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*.

Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. 2023b. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv preprint arXiv:2311.13194*.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. 2024b. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, et al. 2024. Tablebench: A comprehensive and complex benchmark for table question answering. *arXiv preprint arXiv:2408.09174*.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. 2024. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2023. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023a. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. 2023b. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. 2023c. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023d. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Ya-Qi Yu, Minghui Liao, Jiwen Zhang, and Jihao Wu. 2024. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. *arXiv preprint arXiv:2410.05261*.

Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. 2022. Syntax-aware network for handwritten mathematical expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4553–4562.

Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. 2017. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*.

Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. 2024. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. *arXiv preprint arXiv:2406.19101*.

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

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

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