This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Marten: Visual Question Answering with Mask Generation for Multi-modal Document Understanding

Zining Wang[1*], Tongkun Guan[2*], Pei Fu[1(✉)], Chen Duan[1], Qianyi Jiang[1], Zhentao Guo[3], Shan Guo[1], Junfeng Luo[1], Wei Shen[2(✉)], Xiaokang Yang[2]
[1] Meituan [2] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[3] Beijing Institute of Technology
{wangzining03,fupei}@meituan.com, {gtk0615,wei.shen}@sjtu.edu.cn

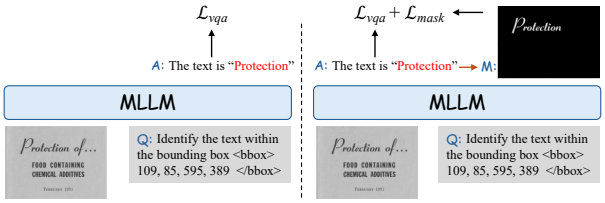
Abstract

*Multi-modal Large Language Models (MLLMs) have introduced a novel dimension to document understanding, i.e., they endow large language models with visual comprehension capabilities; however, how to design a suitable image-text pre-training task for bridging the visual and language modality in document-level MLLMs remains underexplored. In this study, we introduce a novel visual-language alignment method that casts the key issue as a **V**isual **Q**uestion **A**nswering with **Mask** generation (**VQAMask**) task, optimizing two tasks simultaneously: VQA-based text parsing and mask generation. The former allows the model to implicitly align images and text at the semantic level. The latter introduces an additional mask generator (discarded during inference) to explicitly ensure alignment between visual texts within images and their corresponding image regions at a spatially-aware level. Together, they can prevent model hallucinations when parsing visual text and effectively promote spatially-aware feature representation learning. To support the proposed VQAMask task, we construct a comprehensive image-mask generation pipeline and provide a large-scale dataset with 6M data (MTMask6M). Subsequently, we demonstrate that introducing the proposed mask generation task yields competitive document-level understanding performance. Leveraging the proposed VQAMask, we introduce Marten, a training-efficient MLLM tailored for document-level understanding. Extensive experiments show that our Marten consistently achieves significant improvements among 8B-MLLMs in document-centric tasks. Code and datasets are available at https://github.com/PriNing/Marten.*

1. Introduction

Large Language Models (LLMs) have shown a comprehensive generalization ability across a wide range of language-related tasks [1, 2]. These successful experiences have inspired researchers to explore Multi-modal Large Language Models (MLLMs) in the context of Visual Question Answering (VQA), *i.e.,* empower the LLMs with visual comprehension capabilities. However, a significant challenge arises in understanding text within document images, possibly due to high resolution, densely packed, small visual texts, and diverse image forms.

To enhance visual comprehension, several studies [9, 28, 29, 36, 41, 44, 45, 51, 74, 87, 97] have focused on designing pre-training tasks specifically tailored for document images to achieve visual-language alignment. These tasks include full-text recognition or transcription, text spotting, and visual text grounding, *etc*, following various prompts. For instance, KOSMOS-2.5 [52] proposes a visual text grounding task, which inputs the texts within images and produces the corresponding bounding boxes. Vary [80] and mPLUG-DocOWL [27] introduce the learning of struct-aware document parsing, table parsing, chart parsing and natural image parsing for different image forms to enhance fine-grained



(a) Visual Question Answering (b) VQA with Mask Generation

Figure 1. Different pre-training paradigms of MLLMs for document understanding: (a) *Visual Question Answering* (VQA) paradigm that implicitly aligns visual and language modality at the semantic level; (b) our proposed *Visual Question Answering with Mask generation* (VQAMask) paradigm. Building on VQA, we introduce an additional Mask Generator during training to explicitly align visual texts and their corresponding image regions at a spatially-aware level. During the inference stage, the mask generator is discarded.

*These authors contributed equally. ✉Corresponding Author.

textual perception.

Although existing methods demonstrate promising capabilities, we argue that these training tasks predominantly emphasize *semantic alignment* and only implicitly capture the spatial location of text within document images. However, *spatial alignment* is also a crucial factor for accurately interpreting document images. Without spatially-aware supervision, the outputs may disproportionately rely on the powerful semantic context capabilities of large language models (LLMs) rather than optimizing image features from visual encoders, potentially leading to model hallucinations.

To address this issue, we propose a novel vision-language alignment method for visual document understanding, **V**isual **Q**uestion **A**nswering with **Mask** generation (**VQAMask**), to explicitly facilitate spatially-aware feature representation learning. As illustrated in Figure 1 (b), visual tokens and language tokens are input into the LLM to jointly optimize two tasks: VQA-based text parsing and mask generation. For the task of VQA-based text parsing, the model predicts the corresponding answer, following different OCR-related prompts. This task can facilitate the model to align images and text at the semantic level. For the task of mask generation, we introduce an additional **M**ask **G**enerator **M**odule (**MGM**) to explicitly align images and text at the spatially-aware level. Specifically, in the intermediate layer of the LLM, we take the cross-attention interaction between the part of the visual modality (query) and the part of the language modality (key) to obtain attention maps. These attention maps, followed by several deconvolution layers, are restored to the original image resolution. Subsequently, we constrain them to ensure spatial alignment between visual texts within images and their corresponding image regions, under the groundtruth mask supervision constructed by our established mask acquisition pipeline. Additionally, it is important to note that this mask generation task is discarded during the inference stage, and it does not add any additional cost to the inference process. Experiments demonstrate the proposed VQAMask works well in various visual encoders and language models.

Utilizing the proposed VQAMask, we introduce a training-efficient MLLM, Marten, which consistently achieves significant improvements among 8B-MLLMs in document-centric tasks. Our contributions are as follows:

1) We introduce a novel Visual Question Answering with Mask generation (VQAMask) task to facilitate spatially-aware and semantic-aware feature representation learning for visual language alignment.
2) We establish a mask acquisition pipeline to generate mask labels without manual annotation, and provide a large-scale dataset (MTMask6M) with 6M image-mask pairs.
3) Extensive experiments demonstrate the effectiveness of the VQAMask task and outperform the previous state-of-the-art method by 0.4%, 0.4%, 6.2%, 1.8%, 6.2%, 4.0%, 1.5%, and 10.1% on DocVQA, InfoVQA, DeepForm, KLC, WTQ, TabFact, FUNSD, and SROIE datasets.

2. Related Work

2.1. Multi-modal Document Understanding

Multi-modal Document Understanding aims to extract meaningful information from text images of various types, such as charts, tables, documents, and other scene texts, through a question-driven image-to-sequence task. Some early studies [85] have explored end-to-end solutions within a specialist model, which may not provide broad robustness and generality for various scenarios. The recent emergence of Multi-modal Large Language Models (MLLMs) has introduced a novel dimension to the field by linking visual image tokens and language tokens in a sequence-to-sequence format, thereby facilitating task unification. This structure seamlessly integrates computer vision with natural language processing, allowing MLLMs to significantly enhance text reading capabilities, supported by large-scale data and GPU resources. These methods can be roughly categorized into two types: OCR-dependent MLLMs [36, 41, 45, 51, 74] and OCR-free MLLMs [9, 28, 29, 44, 87, 97].

OCR-dependent MLLMs enhance document understanding by integrating text, layout, and other data extracted from external OCR tools [43] into large language models. LayTextLLM [51] and DocLayLLM [45] both utilize an external OCR engine to extract layout and text, integrating them into a LLM for document understanding. However, this integration complicates the workflow and leads to an excess of auxiliary tokens, particularly in images with dense texts.

OCR-free MLLMs perform the multi-modal document understanding task by directly producing question-driven outputs in an end-to-end manner. These methods typically focus on high-resolution image processing [15, 27, 44, 87], efficient token compression [28, 91, 95], and refined attention mechanisms [29, 67]. In the study, we focus on exploring suitable pre-training tasks, tailored for document images.

2.2. Vision Language Pre-training

Inspired by recent advancements [6, 18, 37, 66, 92] in pre-training techniques, the integration of image and text multi-modal information into OCR-related tasks has gained increasing attention. Using cross-modal visual-language priors, early works focused on endowing visual foundation models with semantic knowledge [13, 19, 22, 70, 76, 84, 88, 89] for applications such as text spotting, detection [17**?**], recognition [20, 21, 53**?**], removal, and super-resolution. As MLLMs rapidly develop, researchers are further capitalizing on these visual-language priors to bridge visual and language modalities through diverse pre-training tasks [9, 15, 27, 36, 45, 49, 51, 78, 87, 97]. For instance, UReader [87] introduces the Read Full Text

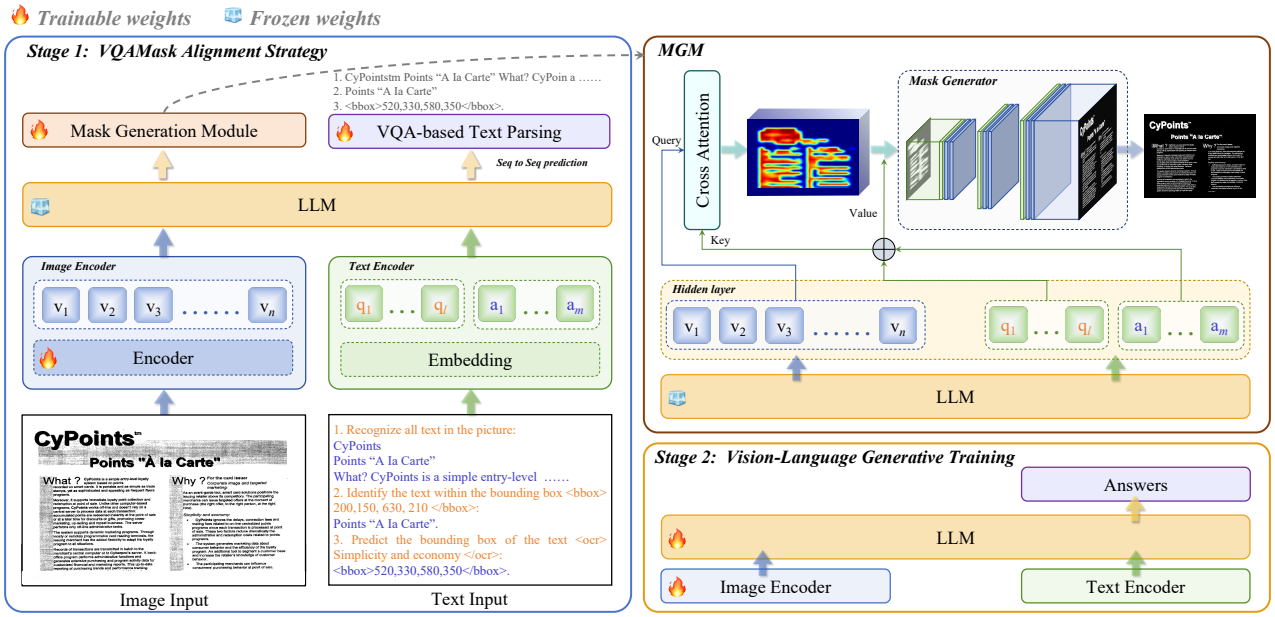


Figure 2. Overview of our proposed Marten architecture. The training of the model is divided into two stages: **1) VQAMask Alignment Training**: the proposed vision-language alignment method, VQAMask, includes two pre-training tasks: VQA-based text parsing and mask generation. By integrating these two tasks, VQAMask not only effectively enables the Marten model to implicitly learn the visual text within images at the semantic level but also explicitly aligns images and text at the spatially-aware level; **2) Vision-Language Generative Training**: In the stage, we discard the mask generation task. A wide range of high-quality instruction data is collected to conduct VQA tasks for general document-level understanding.

(RFT) task in VQA form for enhancing document-level understanding. Park *et al.* [64] propose two new pretext tasks: Reading Partial Text (RPT) and Predicting Text Position (PTP). Similarly, KOSMOS-2.5 [52] designs a Visual Text Grounding (VTG) task, which inputs the texts within images and produces the corresponding bounding boxes. mPLUG-DocOWL [27] integrates multiple tasks to conduct the struct-aware parsing in documents, tables, charts, and natural scenes. However, these question-driven image-to-sequence tasks predominantly emphasize *semantic alignment*, and may rely on the powerful semantic context capabilities of LLMs when responding. Following these VQA forms, we further introduce an additional mask generation pre-text task (VQAMask) to explicitly facilitate spatially-aware visual-language alignment.

3. Methodology

In this section, we first review the representative MLLM method that connects the visual modality and language modality into LLM to generate responses. Building on this foundation, we present our proposed pre-training method, Visual Question Answering with Mask generation (VQAMask), designed specifically for Multi-modal Document Understanding.

Preliminary. Typically, Multi-modal Large Language Models (MLLMs) include a visual foundation model (VFM), a modality connector, and a large language model (LLM). Initially, following the prevalent multi-scale adaptive cropping strategy, the input high-resolution image $\mathbf{X} \in \mathbb{R}^{H \times W}$ is first cropped into several non-overlapping sub-images. H and W represent the image height and width. These sub-images are then processed by the visual foundation model to obtain image patches, concretely represented by $[\mathbf{x}_1, ..., \mathbf{x}_n]$, along with their corresponding visual embeddings $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_n]$. Here, n denotes the number of image patches. For language input, the question and the answer (option for training) is tokenized using the BPE tokenizer, resulting in l question tokens embedded as $\mathbf{Q} = [\mathbf{q}_1, ..., \mathbf{q}_l]$ and m answer tokens embedded as $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_m]$. Subsequently, the modality connector acts a bridge between the visual embeddings and language (question and answer) embeddings. Finally, the visual embeddings $\mathbf{V}$ and language embeddings $\mathbf{Q}$ and $\mathbf{A}$ are fed into the LLM to generate more precise and comprehensive answers. Specifically, the LLM process can be described as:

$$\mathbf{V}^{k+1}, \mathbf{Q}^{k+1}, \mathbf{A}^{k+1} = \text{Layer}_{\text{LLM}}([\mathbf{V}^k, \mathbf{Q}^k, \mathbf{A}^k, \mathbf{E}]) \quad (1)$$

where $k \in \{0, 1, ..., K\}$ and $\mathbf{V}^{k+1}$, $\mathbf{Q}^{k+1}$, and $\mathbf{A}^{k+1}$ refers to the outputs of k-th layer of LLM. $\mathbf{E}$ refers to the attention mask, which is usually a lower triangular matrix used to prevent attention to certain positions. Note that we omit the superscript for $k = 0$, because these vectors are the initial value fed into the LLM. During inference, the input answer tokens $\mathbf{A}$ are replaced as the previous predicted tokens.

Notably, some works introduce the pixel shuffle operation [9] to reduce the number of visual tokens, a strategy also adopted in our proposed method. However, the global operation changes the original spatial structure of these visual tokens. Differently, inspired by the Swin Transformer [50], we conduct pixel shuffle in each local window (4×4 by default) to adapt to our subsequent VQAMask.

3.1. VQA with Mask Generation

To bridge the gap between visual and language modality for multi-modal document understanding, previous MLLMs have formulated various pre-training tasks, including text transcription and visual text grounding, *i.e.*, given customized task prompts, these methods generate prompt-related text responses. The pre-training paradigm lacks spatially-aware supervision, which may result in model hallucinations. To address this, we introduce a novel pre-training method, Visual Question Answering with Mask generation (VQAMask). This method incorporates an additional mask generation task to ensure spatial alignment between visual texts within images and their corresponding image regions, as illustrated in Figure 2. Specifically, the proposed VQAMask includes two tasks as follows:

VQA-based Text Parsing. Following existing works [27, 28, 49, 80], we introduce the text parsing task to implicitly align images and text at the semantic level. The specific task prompts are presented in Figure 3. The outputs of the last layer of LLM are utilized to predict these answers, and the optimization loss is formulated as follows:

$$\mathcal{L}_{vqa} = \mathbf{A} \log p(\mathbf{A}^{K+1} | \mathbf{V}, \mathbf{Q}) \tag{2}$$

Mask Generation. In the subsection, we integrate a mask generation module (MGM) into the hidden layers of the LLM to explicitly enhance vision-language alignment at a spatial-aware level. Specifically, we first feed the hidden states ($\mathbf{V}^k$, $\mathbf{Q}^k$, and $\mathbf{A}^k$) of the selected layer $k-1$ into a four-layer transformer module, with each layer including two sub-layers: a multi-head cross-attention mechanism, and a positionwise fully connected feed-forward network. The specific implementation is as follows:

$$\begin{cases} \mathbf{H}^k = [\mathbf{Q}^k, \mathbf{A}^k] \\ \mathbf{Attn} = \sigma\left(\dfrac{\mathbf{V}^k\mathbf{W}_{query} \cdot (\mathbf{H}^k\mathbf{W}_{key})^\intercal}{\sqrt{d}}\right)\mathbf{H}^k\mathbf{W}_{value}, \\ \mathbf{V}_{attn} = \max(0, \mathbf{Attn} \cdot \mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \end{cases}$$

where $[\cdot]$ denotes the concatenation operation and $\sigma(\cdot)$ refers to the softmax activate function. The projections are parameter matrices $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_{query}, \mathbf{W}_{key}, \mathbf{W}_{value}$ and d denotes the dimension.

By fully interacting with the answer tokens, the visual tokens corresponding to the visual text regions are highlighted. Subsequently, these one-dimensional visual tokens are re-organized into two-dimensional image space. Followed by several transposed convolutions ϕ, we then restore these visual tokens to the resolution of the input image. The specific process is as follows:

$$\tilde{\mathbf{M}} = \phi(\mathbf{V}_{attn}) \tag{3}$$

where $\tilde{\mathbf{M}}$ refers to the final predicted mask. Finally, a Dice loss [60] and Cross-Entropy loss are employed to optimize the segmentation network:

$$\mathcal{L}_{mask} = l_{\text{DICE}}(\tilde{\mathbf{M}}, \mathbf{M}) + l_{\text{CE}}(\tilde{\mathbf{M}}, \mathbf{M}) \tag{4}$$

where $\mathbf{M}$ denotes the groundtruth mask of the input image, which will be introduced in Sec. 3.2.

3.2. Mask Acquisition Pipeline

We note that in document scenarios, the boundary between text and background is typically distinct, allowing for easy separation of text from the entire image using a threshold. Previous research, such as CCD [19], has explored and confirmed this observation. Inspired by CCD [19], we propose a clustering-based binarization method for foreground construction, comprising three stages: preparation, clustering, and generation. The specific process is as follows:

Preparation. We utilize PaddleOCR [43] to detect all visual text regions within an image and obtain corresponding cropped text instance images based on the bounding boxes.

Clustering. For each cropped text instance image, we employ a simple yet effective clustering model (K-means) to classify image pixels into two clusters. Given that visual text tends to be concentrated in the center region of an image, we calculate the distance of the pixels in each cluster from the center position of the cropped image. The cluster with pixels closer to the center is identified as the foreground (with a pixel value of 1), while the other is identified as the background (with a pixel value of 0). Subsequently, a secondary calibration is conducted to verify the correctness of the obtained foreground cluster. Specifically, we compare the average pixel value of the edge regions of the cropped text instance image with the overall average pixel value. If the former is higher, a 0-1 inversion is implemented.

Generation. These foreground masks from all cropped text instance images are reassembled according to their original coordinates to obtain a complete mask image.

3.3. Training Strategy

As shown in Figure 2, we divide the training process into two stages: our proposed VQAMask vision-language alignment training and vision-language generative training.

Stage 1: VQAMask Alignment Training. Currently, most MLLMs for document understanding implement image-text alignment to bridge the visual foundation model with the LLM as the first training stage task. Although such alignment methods endow the MLLMs with basic text recognition capabilities, they lack spatial awareness of the visual text within images. As a result, they struggle to accurately locate complex text within text-rich images and understand

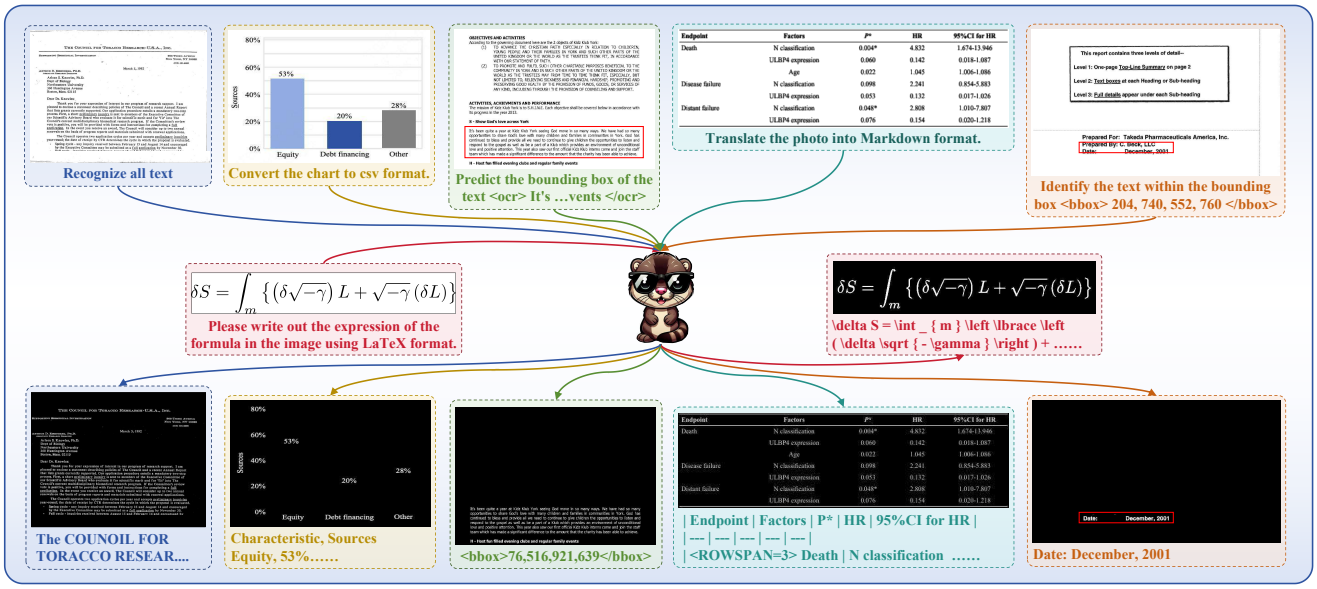


Figure 3. Illustration of the VQAMask alignment training for document parsing question answering. We introduced a total of six tasks, which can be broadly categorized into 1) Read Full Text, Reading Partial Text within Localization, and Visual Text Grounding; 2) Transcription involves converting formulas into LaTeX, tables into markdown or LaTeX, and charts into CSV and markdown formats.

the structural information of documents. To enhance the spatial awareness of visual text in documents, we propose a VQAMask vision-language alignment training to bridge the visual foundation model with the language model.

Regarding data usage, as displayed in Table 1, we utilized the pipeline proposed in Section 3.2 to construct 6 million samples, referred to as MTMask6M, including DocStruct4M [27], IIT-CDIP [42] and DocGenome [82], Scene Text Datasets [23, 34, 35, 69]. Each sample includes the image, question, answer, and the corresponding mask. DocStruct4M is provided by DocOwl 1.5 [27], which includes five categories: natural images, documents (CCPdf [75], DUE [4], RVL-CDIP [24]), tables (TURL [12], PubTabNet [96]), charts (ChartQA [56], FigureQA [33], DVQA [32], PlotQA [59]), and web pages (VisualMRC [73]). DocGenome [82] includes three types of scenarios: documents, formulas, and tables. We name them DocGenome-D, DocGenome-F, and DocGenome-T, respectively. Additionally, we also introduce ICDAR13 [34], ICDAR15 [35], SynthText [23], TextOCR [69], and OpenVINO [38] as the scene text datasets.

For VQA-based text parsing tasks, we construct six types of QA pairs, as illustrated in Figure 3. These include reading full text, visual text recognition with coordinates, visual text grounding, and markdown, LaTeX and CSV format transcription.

During pre-training, the weights of vision foundation model, MLP, and MGM are updated, while the LLM remains frozen. The goal is to preserve the inherent semantic context capability of the LLM while specifically enhancing the overall MLLM's spatial awareness for visual texts within images.

Stage 2: Vision-Language Generative Training. In this stage, we collect existing VQA datasets related to document understanding scenarios, as shown in Table 2. A generative training strategy is then employed to enhance general document-level comprehension. Specifically, our visual foundation model and MLP inherit the weights from stage 1. The proposed MGM is discarded in this stage. Additionally, we unfreeze the weights of LLM and all parameters are updated to conduct supervised fine-tuning (SFT), which includes full data training and high-quality data fine-tuning. First, all data are included in the full data training. Then we collect a batch of high-quality instruction data (one-tenth) from the full dataset to fine-tune the model. The detailed data usage is summarized in Table 2. Through the combined training of these two phases, our model is able to extract more powerful visual representations and exhibits significantly enhanced document understanding capabilities.

4. Experiments

4.1. Implementation Details

Stage 1. In practical implementation, Marten selects InternViT-300M [8] as the visual foundation model, and InternLM2, a 7B large language model [5], as the language decoder. We employ a dynamic image-slicing strategy in which each image is cropped into a maximum of six sub-images based on the aspect ratio and resolution, with a fixed resolution of 448×448 for each sub-image. Subsequently, we employ the Pixel Shuffle module, compatible

Table 1. Details of MTMask6M used in VQAMask Alignment Training (Stage 1). "Text R/G" refers to visual text recognition and visual text grounding, with data sourced from the Multi-Grained Text Localization section of DocStruct4M.

Task	Samples	Datasets
Document parsing	3361.3k	IIT-CDIP [42], CCPdf [75] DUE [4], VisualMRC [73] RVL-CDIP [24], DocGenome-D [82]
Table parsing	600k	TURL [12], PubTabNet [96] DocGenome-T [82]
Chart parsing	475.1k	ChartQA [56], FigureQA [33] DVQA [32], PlotQA [59]
Formula parsing	200k	DocGenome-F [82]
Scene text parsing	395.6k	ICDAR13 [34], ICDAR15 [35] SynthText [23], Textocr [69] OpenVINO [38]
Text R/G	1000k	DocStruct4M-subset [27]
Total	6032k	

Table 2. Details of the training datasets used in Vision-Language Generative Training (Stage 2). † denotes the selected high-quality instruction data, utilized for supervised fine-tuning again.

Task	Samples	Datasets
Document VQA	2301.5k	DocVQA† [57], InfoVQA† [58] DeepForm† [72], KLC† [71] DocMatix [40]
Table VQA	107.6k	TableFact† [7], WTQ† [65] TableBench [81]
Chart VQA	318.1k	ChartQA† [56], FigureQA [33] DVQA† [32]
Formula VQA	274.5k	UniMER [77], CROHME† [54, 62, 63]
Sence Text VQA	289.3k	TextVQA† [68], ST-VQA† [3] OCR-VQA [61], IAM [55]† EST-VQA [79]
KIE	6.2k	FUNSD† [31], SROIE† [30]
Total	3297.2k	

with VQAM, to reduce the number of tokens to 256. We perform one epoch on MTMask6M in Table 1. The learning rate for the MGM module is set to 2e-4, while for other parameters, it is set to 2e-5. The batch size on each GPU is 64, and the training is conducted on 24 GPUs for two days.

Stage 2. The dataset of Table 2 is used in the stage. The learning rate and batch size are 2e-5 and 64, respectively. The training phase is conducted on 24 H800 GPUs over 56 hours. More details are introduced in Section 3.3.

4.2. Results

Text-rich Result. We compared Marten with OCR-free multimodal large language models on 11 text-rich image benchmarks, which cover documents (DocVQA [57], InfoVQA [58], DeepForm [72], KLC [71]), tables (WTQ [65], TabFact [7]), charts (ChartQA [56]), sence text (TextVQA [68]), and KIE (FUNSD [31], SROIE [30], POIE [39]). The evaluation metrics used are derived from the official metrics provided. It is important to note that TextVQA is evaluated using the validation set, while the other datasets are evaluated using their respective test sets. As shown in Table 3, Marten demonstrates superior performance compared to existing MLLMs, particularly excelling in text-dense and smaller document scenarios. Marten achieve consistently and significantly performence improvements on multiple benchmarks, leading in datasets such as DocVQA, InfoVQA, DeepForm, KLC, WTQ, TabFact, FUNSD, and SROIE, indicating a more comprehensive capability in visual document understanding. Compared to the existing best methods under each benchmark, Marten achieves an average improvement of 1.97% in document benchmarks, 5.09% in table benchmarks, and 3.73% in key information extraction benchmarks. This demonstrates that our alignment strategy aids Marten in better locating the position of visual texts and accurately finding the answers. However, in the chart and sence text benchmarks, Marten's performance is lower than that of InternVL2, which is trained on hundreds of millions of samples. This indicates that Marten still lacks understanding in charts and perception abilities in natural scenes, which will be a focus for future optimization efforts.

OCRBench. To comprehensively evaluate the performance of Marten, Table 4 presents a comparison of Marten with existing MLLMs on OCRBench [48]. OCRBench is a recently developed benchmark designed to assess the optical character recognition (OCR) capabilities of MLLMs. It encompasses a wide range of text-related visual tasks, divided into five subtasks: Text Recognition, Scene Text-centric VQA, Doc-oriented VQA, Key Information Extraction (KIE), and Handwritten Mathematical Expression Recognition (HMER). In total, it includes 29 datasets and aims to produce an overall score. Specifically, Marten achieved a score of 820 on OCRBench, which is 26 points higher than InternVL2 and 18 points higher than MiniMonkey, demonstrating Marten's efficient performance across a broad spectrum of text-related visual tasks. Additionally, Figure 4 illustrates Marten's scores compared to recent MLLMs in the five subtasks. It is observed that by employing the VQAMask vision-language alignment method, Marten demonstrates superior performance in both VQA tasks and transcription tasks. It is noteworthy that since the Text Recognition task lacks layout information, our method does not provide effective improvements in this area.

4.3. Ablation Study

Extensive ablation experiments are conducted to verify the effectiveness of the module. The results of both the first and second training stages are validated separately. To assess the effectiveness of the MGM, different model combi-

Table 3. Comparison with OCR-free methods on various types of text-rich image understanding tasks. All evaluation benchmarks use the officially designated metrics. "size" refers to the number of parameters in the model, and "Val" refers to the validation set.

Model	size	Venue	DocVQA	InfoVQA	DeepForm	KLC	ChartQA	$\text{TextVQA}_{\text{Val}}$	WTQ	TabFact	FUNSD	SROIE	POIE
DocPeida [16]	7.1B	arxiv'23	47.1	15.2	-	-	46.9	60.2	-	-	29.9	21.4	39.9
DocOwl [86]	7.3B	arxiv'23	62.2	38.2	42.6	30.3	57.4	52.6	26.9	67.6	0.5	1.7	2.5
LLaVA1.5 [47]	7.3B	NeurIPS'23	-	-	-	-	9.3	-	-	-	0.2	1.7	2.5
UReader [87]	7.1B	EMNLP'23	65.4	42.2	49.5	32.8	59.3	57.6	29.4	67.6	-	-	-
CHOPINLLM [14]	7B	arxiv'24	-	-	-	-	69.98	-	-	-	-	-	-
TextHawk [90]	7.4B	arxiv'24	76.4	50.6	-	-	66.6	-	34.7	71.1	-	-	-
DocKylin [94]	7.1B	arxiv'24	77.3	46.6	-	-	66.8	-	32.4	-	-	-	-
MM1.5 [93]	7.3B	arxiv'24	88.1	59.5	-	-	78.6	76.8	46.0	75.9	-	-	-
Mini-Monkey [29]	2B	arxiv'24	87.4	60.1	-	-	76.5	75.7	-	-	42.9	70.3	**69.9**
DocOwl-1.5 [27]	8.1B	EMNLP'24	81.6	50.4	68.8	37.9	70.5	68.8	39.8	80.4	-	-	-
DocOwl-1.5-Chat [27]	8.1B	EMNLP'24	82.2	50.7	68.8	38.7	70.2	68.6	40.6	80.2	-	-	-
CogAgent [26]	17.3B	CVPR'24	81.6	44.5	-	-	68.4	76.1	-	-	-	-	-
Monkey [44]	9.8B	CVPR'24	66.5	36.1	40.6	-	65.1	67.6	25.3	-	-	-	-
TextMonkey [49]	7.7B	arxiv'24	73.0	28.6	-	-	66.9	65.6	-	-	32.3	47.0	27.9
HRVDA [46]	7.1B	CVPR'24	72.1	43.5	63.2	37.5	67.6	73.3	31.2	72.3	-	-	-
InternVL2 [9]	8.1B	CVPR'24	91.6	74.8	-	-	**83.3**	**77.4**	-	-	-	-	-
Park et al. [64]	7.2B	NeurIPS'24	72.7	45.9	53.0	36.7	63.3	59.2	34.5	68.2	-	-	-
MOAI [41]	7B	ECCV'24	-	-	-	-	-	67.8	-	-	-	-	-
Vary [80]	7.4B	ECCV'24	76.3	-	-	-	66.1	-	-	-	-	-	-
TextHawk2 [91]	7.4B	arxiv'24	89.6	67.8	-	-	81.4	75.1	46.2	78.1	-	-	-
PDF-WuKong [83]	8.5B	arxiv'24	76.9	-	-	-	-	-	-	-	-	-	-
Zhang et al. [95]	8.1B	arxiv'24	78.3	50.2	65.7	35.9	68.9	66.6	38.6	79.3	-	-	-
Marten	8.1B	-	**92.0**	**75.2**	**75.1**	**39.5**	81.7	74.4	**52.4**	**84.4**	**44.4**	**80.4**	69.5

Table 4. Comparison of Marten with existing OCR-free multimodal large language models on OCRBench.

dataset \ model	Monkey [44]	TextMonkey [49]	DocOwl-1.5 [27]	MM1.5 [93]	TextHawk2 [91]	GLM-4v [25]	InternVL2 [9]	MiniMonkey [29]	Marten(ours)
OCRBench	514	561	599	635	784	786	794	802	**820**

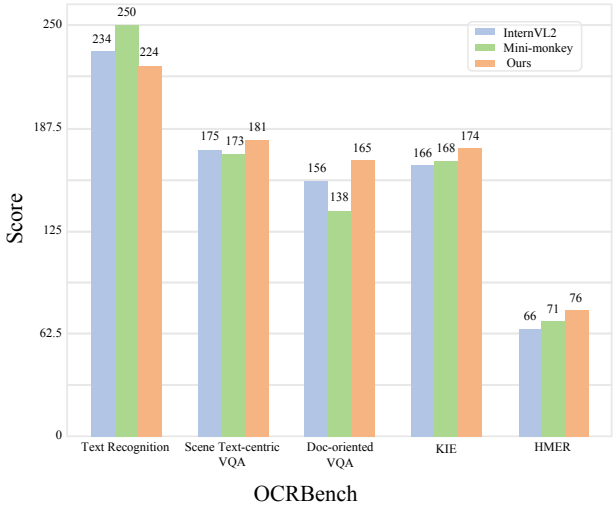


Figure 4. Bar chart of scores for each subtask in OCRBench. "KIE" stands for Key Information Extraction, and "HMER" stands for Handwritten Mathematical Expression Recognition.

nations are integrated for verification.

Stage 1. In the first training stage, we compared Marten's performance with and without the MGM, as shown in Table 5. The enhancement of Marten's vision-language alignment capability by MGM is verified through recognition results in both natural and document scenarios. For natural scenes, we use the ICDAR15 [35] and TotalText [11] datasets. For document scenarios, one thousand images from IIT-CDIP [42], not involved in training, are selected, and PaddleOCR is used to recognize the visual texts, constructing the recognition results for evaluation. Additionally, we extract one thousand latex-formatted tables and equations from DocGenome [82], which are also not used during training, to assess Marten's transcription performance. We discuss the impact of MGM on vision-language alignment under different model combinations. The visual foundation model options include Swin-Transformer [50] and InternViT [8], while the LLM choices are Vicuna1.5 [10] and InternLM2 [5]. Since bounding box information is not included during the training phase, the recognition output is evaluated using Edit Distance. Experimental results indicate that after adding MGM, Marten's average Edit Distance in both natural and document scenarios decreases by 0.06. In transcription tasks, the average edit distance decreases by approximately 0.1, showing a more significant improvement. This indicates that MGM helps align the visual foundation model with the LLM, thereby enhancing the model's ability to recognize and parse visual texts.

In Figure 5, we present the visualization output results of Marten during VQAMask alignment training across three tasks: full-image parsing, transcription, and partial text recognition. The binary masks of the outputs for these

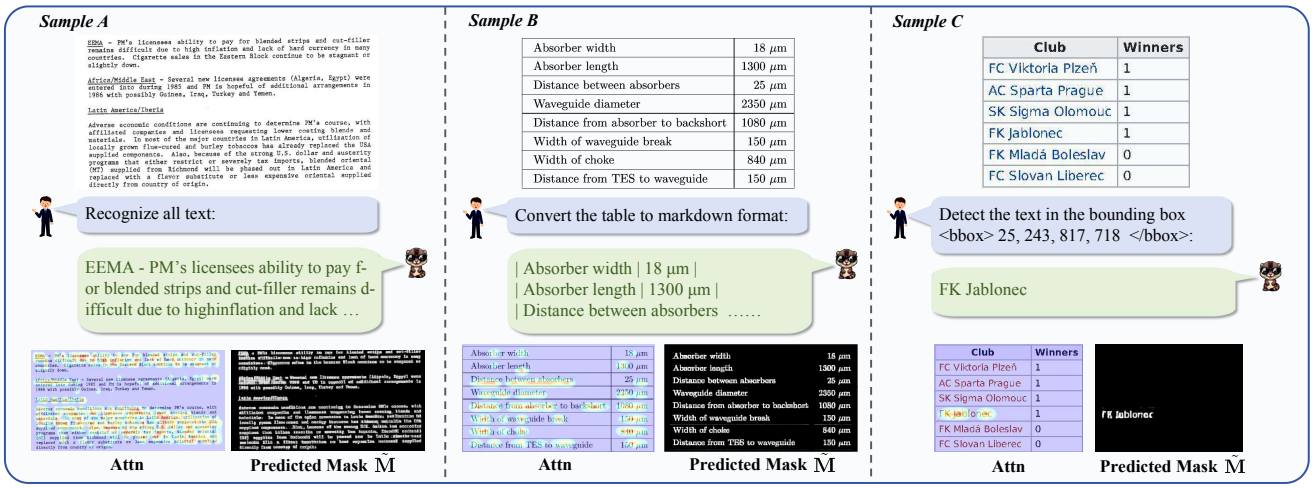


Figure 5. Visualization of output results in VQAMask alignment training. We present samples for three different tasks: 1) Sample A represents full-image visual text recognition, 2) Sample B represents Markdown-style transcription, and 3) Sample C represents reading partial text guided by the bounding box.

Table 5. Ablation study in stage 1. Edit Distance (ED) is used as the evaluation metric. "BB" refers to the backbone.

BB	LLM	MGM	Edit Distance (ED) IC15↓	TT↓	IIT↓	DocG↓
Swin	Vicuna1.5	×	0.344	0.484	0.289	0.346
		✓	**0.262**$_{(-0.082)}$	**0.397**$_{(-0.087)}$	**0.234**$_{(-0.055)}$	**0.242**$_{(-0.112)}$
Swin	InternLM2	×	0.325	0.478	0.286	0.315
		✓	**0.254**$_{(-0.071)}$	**0.396**$_{(-0.082)}$	**0.218**$_{(-0.068)}$	**0.225**$_{(-0.09)}$
ViT	InternLM2	×	0.339	0.451	0.271	0.319
		✓	**0.278**$_{(-0.061)}$	**0.381**$_{(-0.07)}$	**0.193**$_{(-0.078)}$	**0.211**$_{(-0.108)}$

tasks indicate that Marten generates relatively accurate visual texts, demonstrating the feasibility of the method. Additionally, heatmaps are included to show the regions of the image that Marten focuses on. It is observed that after applying VQAMask alignment training, the LLM's perception of the image is concentrated on areas associated with the QA content, confirming that VQAMask enhances the model's spatial awareness of visual texts.

Stage 2. In Table 6, we discuss the improvement in visual document understanding performance brought by MGM under different model combinations. The model configurations remain consistent with those in Table 5. We conduct comparisons on four text-rich image benchmarks, including DocVQA, InfoVQA, ChartQA, and TextVQA. MGM improves performance across all four benchmarks, with a particularly noticeable enhancement in DocVQA. Specifically, in the combination of Swin-Transformer and InternLM2, DocVQA shows an improvement of 4.73%. However, when the Swin-Transformer is used as the visual foundation model, its performance on InfoVQA is inferior to that of InternViT. This is mainly because InfoVQA consists images with super high aspect ratio, which makes it challenging for Swin-Transformer, without employing a crop strategy,

Table 6. Ablation study in stage 2. "BB" refers to the backbone, and "Val" refers to the validation set.

BB	LLM	MGM	DocVQA	InfoVQA	ChartQA	TextVQA$_{Val}$
Swin	Vicuna1.5	×	78.45	43.55	69.15	71.63
		✓	**81.89**$_{(+3.44)}$	**47.19**$_{(+3.64)}$	**72.01**$_{(+2.86)}$	**74.97**$_{(+3.34)}$
Swin	InternLM2	×	81.12	48.50	73.75	71.34
		✓	**85.85**$_{(+4.73)}$	**52.21**$_{(+3.71)}$	**76.77**$_{(+3.02)}$	**74.92**$_{(+3.58)}$
ViT	InternLM2	×	89.52	71.65	79.26	71.25
		✓	**92.01**$_{(+2.49)}$	**75.21**$_{(+3.56)}$	**81.72**$_{(+2.46)}$	**74.38**$_{(+3.13)}$

to effectively extract visual texts.

5. Conclusion

In this study, we introduce a novel visual language alignment method, Visual Question Answering with Mask generation (VQAMask), during the pre-training stage to bridge the gap between visual and language modalities. While keeping LLM weights frozen, VQAMask assists the MLLM in simultaneously conducting VQA-based text parsing and mask generation tasks. This optimization process not only leverages the contextual capabilities of the powerful large language model but also promotes the learning of spatially-aware and semantic-aware feature representations for the image encoder. To achieve this, we establish a comprehensive image-mask generation pipeline, and provide MTMask6M with 6M data. Extensive ablation experiments validate the effectiveness and significance of the proposed VQAMask. Finally, leveraging the proposed VQAMask, we introduce Marten, a training-efficient MLLM tailored for general document-level understanding. In future work, we aim to further explore more fine-grained and robust visual language alignment methods to enhance visual document understanding.

6. Acknowledgements

This work was supported by NSFC 62322604, NSFC 62176159 and Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 6

[4] Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. Due: End-to-end document understanding benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 5, 6

[5] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, and et al. Internlm2 technical report, 2024. 5, 7

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[7] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019. 6

[8] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 5, 7

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 2, 4, 7

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 7

[11] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, pages 935–942. IEEE, 2017. 7

[12] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022. 5, 6

[13] Chen Duan, Pei Fu, Shan Guo, Qianyi Jiang, and Xiaoming Wei. 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, 2024. 2

[14] Wan-Cyuan Fan, Yen-Chun Chen, Mengchen Liu, Lu Yuan, and Leonid Sigal. On pre-training of multimodal language models customized for chart understanding. *arXiv preprint arXiv:2407.14506*, 2024. 7

[15] Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*, 2023. 2

[16] Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*, 2023. 7

[17] Tongkun Guan, Chaochen Gu, Changsheng Lu, Jingzheng Tu, Qi Feng, Kaijie Wu, and Xinping Guan. Industrial scene text detection with refined feature-attentive network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6073–6085, 2022. 2

[18] Tongkun Guan, Chaochen Gu, Jingzheng Tu, Xue Yang, Qi Feng, Yudi Zhao, and Wei Shen. Self-supervised implicit glyph attention for text recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15285–15294, 2023. 2

[19] Tongkun Guan, Wei Shen, Xue Yang, Qi Feng, Zekun Jiang, and Xiaokang Yang. Self-supervised character-to-character distillation for text recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19473–19484, 2023. 2, 4

[20] Tongkun Guan, Chengyu Lin, Wei Shen, and Xiaokang Yang. Posformer: recognizing complex handwritten mathematical expression with position forest transformer. In *European Conference on Computer Vision*, pages 130–147. Springer, 2024. 2

[21] Tongkun Guan, Wei Shen, and Xiaokang Yang. Ccdplus: Towards accurate character to character distillation for text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[22] Tongkun Guan, Wei Shen, Xue Yang, Xuehui Wang, and Xiaokang Yang. Bridging synthetic and real worlds for pre-training scene text detectors. In *European Conference on Computer Vision*, pages 428–446. Springer, 2025. 2

[23] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2315–2324, 2016. 5, 6

[24] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE, 2015. 5, 6

[25] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 7

[26] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024. 7

[27] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 1, 2, 3, 4, 5, 6, 7

[28] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*, 2024. 1, 2, 4

[29] Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Alleviate the sawtooth effect by multi-scale adaptive cropping. *arXiv preprint arXiv:2408.02034*, 2024. 1, 2, 7

[30] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019. 6

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

[32] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. 5, 6

[33] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 5, 6

[34] 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. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE, 2013. 5, 6

[35] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 5, 6, 7

[36] Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. Visually-situated natural language understanding with contrastive reading model and frozen large language models. *arXiv preprint arXiv:2305.15080*, 2023. 1, 2

[37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 2

[38] Ilya Krylov, Sergei Nosov, and Vladislav Sovrasov. Open images v5 text annotation and yet another mask text spotter. In *Asian Conference on Machine Learning*, pages 379–389. PMLR, 2021. 5, 6

[39] Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. 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, 2023. 6

[40] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions., 2024. 6

[41] Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Moai: Mixture of all intelligence for large language and vision models. *ECCV*, 2024. 1, 2, 7

[42] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 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, 2006. 5, 6, 7

[43] Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*, 2022. 2, 4

[44] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024. 1, 2, 7

[45] Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient and effective multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*, 2024. 1, 2

[46] Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. Hrvda: High-resolution visual document assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15534–15545, 2024. 7

[47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 7

[48] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 6

[49] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024. 2, 4, 7

[50] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4, 7

[51] Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, et al. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. *arXiv preprint arXiv:2407.01976*, 2024. 1, 2

[52] Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, et al. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*, 2023. 1, 3

[53] Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Maskocr: Text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311*, 2022. 2

[54] Mahshad Mahdavi, Richard Zanibbi, Harold Mouchere, Christian Viard-Gaudin, and Utpal Garain. Icdar 2019 crohme+ tfd: Competition on recognition of handwritten mathematical expressions and typeset formula detection. pages 1533–1538, 2019. 6

[55] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5: 39–46, 2002. 6

[56] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 5, 6

[57] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 6

[58] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 6

[59] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020. 5, 6

[60] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 4

[61] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 6

[62] Harold Mouchere, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. Icfhr 2014 competition on recognition of on-line handwritten mathematical expressions (crohme 2014). pages 791–796, 2014. 6

[63] Harold Mouchère, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. Icfhr2016 crohme: Competition on recognition of online handwritten mathematical expressions. pages 607–612, 2016. 6

[64] Jaeyoo Park, Jin Young Choi, Jeonghyung Park, and Bohyung Han. Hierarchical visual feature aggregation for ocr-free document understanding. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3, 7

[65] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 6

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pages 8748–8763, 2021. 2

[67] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. 2024. 2

[68] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 6

[69] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 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, 2021. 5, 6

[70] Sibo Song, Jianqiang Wan, Zhibo Yang, Jun Tang, Wenqing Cheng, Xiang Bai, and Cong Yao. Vision-language pre-training for boosting scene text detectors. In *CVPR*, pages 15681–15691, 2022. 2

[71] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021. 6

[72] S Svetlichnaya. Deepform: Understand structured documents at scale. 2020. 6

[73] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document im-

ages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13878–13888, 2021. 5, 6

[74] Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *AAAI*, pages 19071–19079, 2024. 1, 2

[75] Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. Ccpdf: Building a high quality corpus for visually rich documents from web crawl data. In *International Conference on Document Analysis and Recognition*, pages 348–365. Springer, 2023. 5, 6

[76] Qi Wan, Haoqin Ji, and Linlin Shen. Self-attention based text knowledge mining for text detection. In *CVPR*, pages 5983–5992, 2021. 2

[77] Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. Unimernet: A universal network for real-world mathematical expression recognition. *arXiv preprint arXiv:2404.15254*, 2024. 6

[78] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding. page 8529–8548, 2024. 2

[79] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135, 2020. 6

[80] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *ECCV*, pages 408–424. Springer, 2025. 1, 4, 7

[81] Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, et al. Tablebench: A comprehensive and complex benchmark for table question answering. *arXiv preprint arXiv:2408.09174*, 2024. 6

[82] Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*, 2024. 5, 6, 7

[83] Xudong Xie, Liang Yin, Hao Yan, Yang Liu, Jing Ding, Minghui Liao, Yuliang Liu, Wei Chen, and Xiang Bai. Wukong: A large multimodal model for efficient long pdf reading with end-to-end sparse sampling. *arXiv preprint arXiv:2410.05970*, 2024. 7

[84] Chuhui Xue, Wenqing Zhang, Yu Hao, Shijian Lu, Philip HS Torr, and Song Bai. Language matters: A weakly supervised vision-language pre-training approach for scene text detection and spotting. In *ECCV*, pages 284–302. Springer, 2022. 2

[85] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8751–8761, 2021. 2

[86] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023. 7

[87] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023. 1, 2, 7

[88] Wenwen Yu, Yuliang Liu, Wei Hua, Deqiang Jiang, Bo Ren, and Xiang Bai. Turning a clip model into a scene text detector. In *CVPR*, pages 6978–6988, 2023. 2

[89] Wenwen Yu, Yuliang Liu, Xingkui Zhu, Haoyu Cao, Xing Sun, and Xiang Bai. Turning a clip model into a scene text spotter. *IEEE TPAMI*, 2024. 2

[90] Ya-Qi Yu, Minghui Liao, Jihao Wu, Yongxin Liao, Xiaoyu Zheng, and Wei Zeng. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *arXiv preprint arXiv:2404.09204*, 2024. 7

[91] Ya-Qi Yu, Minghui Liao, Jiwen Zhang, and Jihao Wu. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. *arXiv preprint arXiv:2410.05261*, 2024. 2, 7

[92] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 2

[93] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. 7

[94] Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. *arXiv preprint arXiv:2406.19101*, 2024. 7

[95] Renshan Zhang, Yibo Lyu, Rui Shao, Gongwei Chen, Weili Guan, and Liqiang Nie. Token-level correlation-guided compression for efficient multimodal document understanding. *arXiv preprint arXiv:2407.14439*, 2024. 2, 7

[96] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer, 2020. 5, 6

[97] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2