Docopilot: Improving Multimodal Models for Document-Level Understanding

Yuchen Duan[1,2*], Zhe Chen[3,1*], Yusong Hu[4,1*], Weiyun Wang[*5,1], Shenglong Ye[1], Botian Shi[1], Lewei Lu[7], Qibin Hou[4], Tong Lu[3,1], Hongsheng Li[2,1], Jifeng Dai[6,1], Wenhai Wang[2,1✉]

[1]Shanghai AI Laboratory, [2]The Chinese University of Hong Kong, [3]Nanjing University, [4]Nankai University, [5]Fudan University, [6]Tsinghua University, [7]SenseTime Research

Abstract

Despite significant progress in multimodal large language models (MLLMs), their performance on complex, multi-page document comprehension remains inadequate, largely due to the lack of high-quality, document-level datasets. While current retrieval-augmented generation (RAG) methods offer partial solutions, they suffer from issues, such as fragmented retrieval contexts, multi-stage error accumulation, and extra time costs of retrieval. In this work, we present a high-quality document-level dataset, Doc-750K, designed to support in-depth understanding of multimodal documents. This dataset includes diverse document structures, extensive cross-page dependencies, and real question-answer pairs derived from the original documents. Building on the dataset, we develop a native multimodal model—Docopilot, which can accurately handle document-level dependencies without relying on RAG. Experiments demonstrate that Docopilot achieves superior coherence, accuracy, and efficiency in document understanding tasks and multi-turn interactions, setting a new baseline for document-level multimodal understanding. Data, code, and models are released at https://github.com/OpenGVLab/Docopilot.

1. Introduction

In recent years, multimodal large language models (MLLMs) [6, 12, 41, 56, 60, 72, 81, 82, 84, 92] have rapidly developed, achieving remarkable performance in various visual understanding tasks [30, 77], particularly image-level tasks, such as image captioning [10, 42], optical character recognition (OCR) [45, 65], and visual question answering (VQA) [24, 44]. Despite these advances, current MLLMs still face significant challenges in document-level understanding [51, 76, 89], where models are required to identify and integrate key information across multi-page documents, setting high expectations for their long-context processing capabilities of MLLMs.

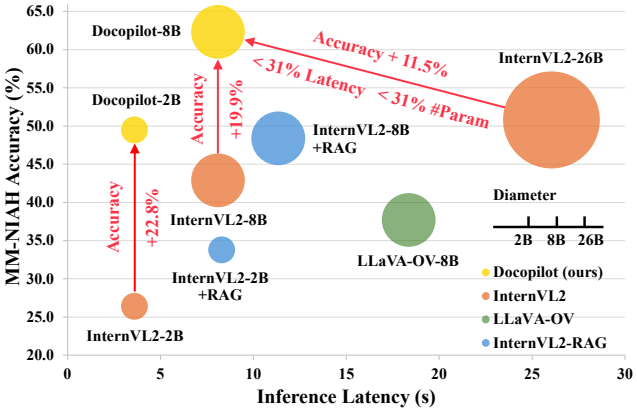


Figure 1. **Accuracy *v.s* inference latency on MM-NIAH.** The proposed Docopilot-8B shows a notable improvement over baseline models [73], achieving a +19.9% accuracy gain compared to InternVL2-8B and surpassing InternVL2-26B with less than 31% of the inference latency. Additionally, Docopilot-2B uses fewer parameters (less than 10%) while exhibiting comparable performance to the 10× larger InternVL2-26B. These results suggest that our Docopilot strikes a reasonable balance between latency, model size, and performance.

Current research on long-content understanding primarily focuses on text-only models [5, 9, 74], targeting specific retrieval tasks such as "Needle in a Haystack" (NIAH) [2, 31]. However, existing open-source MLLMs [12, 13, 36, 41, 62, 81, 83] are primarily trained on image-level data, lacking the long-context understanding capacity required for document-level understanding. Retrieval-augmented generation (RAG) methods [14, 22, 50, 61, 63, 95] attempt to address this by retrieving key information to fit within the limited context windows of MLLMs, but they still encounter the following challenges in document-level tasks. (1) *Fragmented Retrieval Contexts*. Retrieved information is fragmented, lacking the overall structure of the document; (2) *Multi-Stage Error Accumulation*. Incorrect retrieval results can affect subsequent responses, leading to errors or omissions of critical details, especially in multi-turn or complex tasks; (3) *Extra Time Costs*. The retrieval

* Equal contribution;
✉ Corresponding author: wangwenhai@pjlab.org.cn

step increases the latency of the QA system, limiting the scalability of RAG in time-sensitive scenarios.

To address these problems, two primary challenges need to be considered. (1) *High-Quality Multimodal Document Dataset*. While extensive datasets [11, 91, 96, 97] exist for long-context, text-only tasks, high-quality document-level question-answering datasets remain scarce. This shortage is largely attributed to the high costs associated with annotation and the lack of streamlined construction pipelines. (2) *Native Document-Level MLLMs*. Although RAG-based methods [14, 86, 95] provide some relief, native multimodal models with long-context processing abilities are crucial. However, training native MLLMs specifically for document-level understanding is constrained by current hardware limitations.

In this work, we introduce a new multimodal document dataset that supports document-level understanding tasks. Compared to counterparts [33, 76, 78], this dataset has the following features: (1) *Large Scale*. It includes a total of 758K question-answer samples, containing 5.2B text tokens and 3.1M images. It encompasses content from various sources, such as Sci-Hub, Arxiv, and OpenReview, covering a wide range of topics and document layouts. (2) *High Quality*. Unlike existing datasets that insert irrelevant questions into documents, we collect real, in-depth question-answer pairs and construct single-page and cross-page questions based on document structure. Such high-quality question-answer data accounts for 31.6% of the dataset. (3) *Multimodal*. For the document content, we provide not only the conventional interleaved text-image context but also purely rendered image inputs, catering to the needs of different models.

Building upon this dataset, we developed a native baseline model for document-level multimodal understanding–Docopilot. Unlike existing approaches [14, 86, 95] that rely on RAG, our model achieves efficient document-level training and testing through simple engineering optimizations, such as multimodal data packing, Ring Attention [43], and Liger Kernel [16]. Leveraging the proxy tasks carefully designed within the dataset, Docopilot can directly handle long-distance dependencies and cross-page information integration without external retrieval support. As shown in Figure 1, this approach not only enhances coherence and accuracy compared to RAG methods but also significantly reduces the response time of the entire question-answering system, delivering superior real-time performance in multi-turn interactions.

The main contributions are summarized as follows:

(1) We develop the first large-scale, high-quality dataset for document-level multimodal understanding, consisting of 758K QA pairs from 3 sources, supporting 9 types of proxy tasks. This dataset includes 31.6% real QA pairs directly extracted from documents.

(2) Based on the dataset, we implement Docopilot, a native MLLM designed for document-level understanding without relying on retrieval mechanisms. This approach greatly improves its ability to integrate and comprehend information across multi-page documents.

(3) Through extensive experiments on multiple document-level benchmarks, our method demonstrates performance significantly superior to existing approaches, proving its effectiveness and generality. As shown in Figure 1, Docopilot-8B achieves a score of 61.8 on MM-NIAH [86], outperforming InternVL2-8B by 19.9 points and surpassing InternVL2-26B with less than 31% of the latency. We hope this work could provide a baseline for future advancements in MLLMs for document-level tasks.

2. Related Work

2.1. Multimodal Large Language Models

Multimodal large language models (MLLMs) have demonstrated impressive capabilities in processing image and text information, opening up new directions for applications such as visual question answering and image captioning. Early models [12, 29, 37, 59] trained with contrastive learning methods excelled in recognizing and understanding open-world semantics within an image-text matching framework. However, their limited generative abilities restricted their applicability. To leverage the powerful generation abilities of large language models (LLMs), subsequent works [13, 38, 42, 47, 83, 84] introduced a connector to align the embedding spaces of vision encoders and LLMs, allowing encoded image embeddings to serve as soft prompts for LLMs. Another series of works [1, 34, 39, 99] extended LLMs by integrating additional visual experts, reducing reliance on standalone vision encoders. More recently, models capable of both understanding and generating images have also made notable progress [20, 35, 67, 75], leveraging the insight that image generation can enhance image understanding. Despite these advancements, current MLLMs still face challenges with long-context multimodal inputs. For instance, InternVL 2.0 [13, 25] performs optimally within a token range of up to 8192, constraining its effectiveness in document-level applications.

2.2. Document Understanding Models

Extracting key information from documents is crucial for industries and academic research. OCR-model-driven methods [3, 4, 71, 80] represent one of the primary technical approaches. These methods extract text, layout, and bounding box information from external systems and integrate it with another model. However, they are prone to error propagation and high processing times due to their reliance on multiple components. Benefitting from the rapid advancements in LLMs, OCR-free methods have also achieved

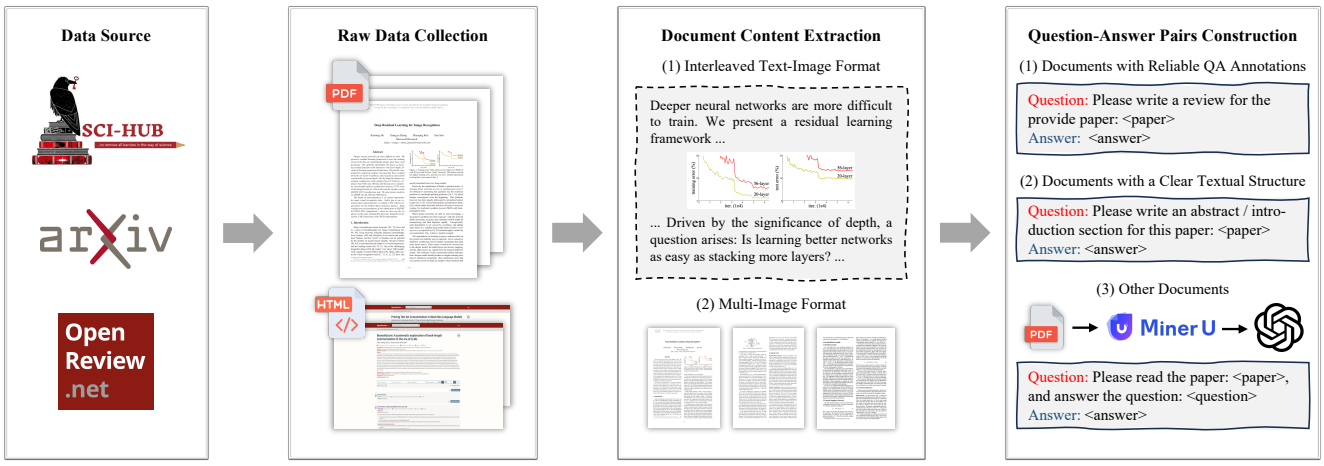


Figure 2. **Multimodal document dataset generation pipeline.** This pipeline involves three main stages: (1) Raw Data Collection: Documents are gathered from sources like Sci-Hub, arXiv, and OpenReview, available in PDF and HTML formats. (2) Document Content Extraction: Multimodal content is processed in two formats: interleaved text-image format and multi-image format. (3) Question-Answer Pairs Construction: QA pairs are generated based on the document structure or constructed using GPT-4o.

great progress. Donut [32] is the first end-to-end training framework based on a Transformer without requiring OCR engines or APIs. Subsequent works [23, 49, 87, 88, 94] propose diverse modifications in model architectures and training algorithms. However, these models are designed for specific tasks and lack general abilities.

2.3. Long-Context Large Language Models

With advancements in engineering, architecture, and algorithms, long-context large language models have made substantial progress. Techniques such as Flash Attention [17, 18] and Ring Attention [43] have notably reduced GPU memory usage for training on extended contexts. Additionally, various sparse attention mechanisms—including Shifted Sparse Attention [11], Dilated Attention [19], and Attention Sinks [26, 90]—have enabled efficient scaling to handle larger contexts. New positional embedding methods, like ALiBi [58], xPOS [68], and RoPE [66], further enhance the models' generalization capabilities in length extrapolation. However, these advancements remain largely confined to natural language processing, and methods to extend the context size of MLLMs are still under-explored. Another research approach aims to reduce context size by leveraging retrieval augmented generation (RAG) [14, 22, 95], where only the most relevant passages are retrieved and fed into the generation model. However, this retrieval-based approach can disrupt the coherence of the semantic chain, particularly in complex reasoning tasks, due to fragmented information flow. In this work, we integrate the above engineering techniques into MLLMs and demonstrate that a model fine-tuned on a high-quality, long-context training corpus is a strong baseline, achieving superior performance compared to its RAG counterpart.

3. Multimodal Document Dataset Generation

In this section, we begin by introducing the details of the data engine. Following this, we provide a comprehensive overview of the dataset—Doc-750K.

3.1. Data Engine

The data engine operates primarily through two steps: document content extraction and question-answer (QA) pair construction. Specifically, we first extract multimodal content, including both text and images, from the documents. Based on this extracted content, we then create question-answer pairs. The document content and these constructed pairs are combined to produce conversational-style training data. The format is outlined as:

```
Please read the paper: <paper>, and
answer the question: <question> Answer:
<answer>
```

Here, the `<paper>`, `<question>`, and `<answer>` are the placeholder for extracted document content, the generated questions and answer, respectively. In the following, we will provide a detailed explanation of the two key steps: document content extraction and QA pair construction.

Document Content Extraction. In practical applications, different documents have varying page layouts and content types, which poses significant challenges for content extraction. To enhance the efficiency of multimodal models, it is necessary to organize documents into a unified format for streamlined processing. In this work, we process each document into two formats as follows:

(1) *Interleaved Text-Image Format.* Using the document content extractor MinerU [79], we segment the document content into interleaved text and image annotations, for

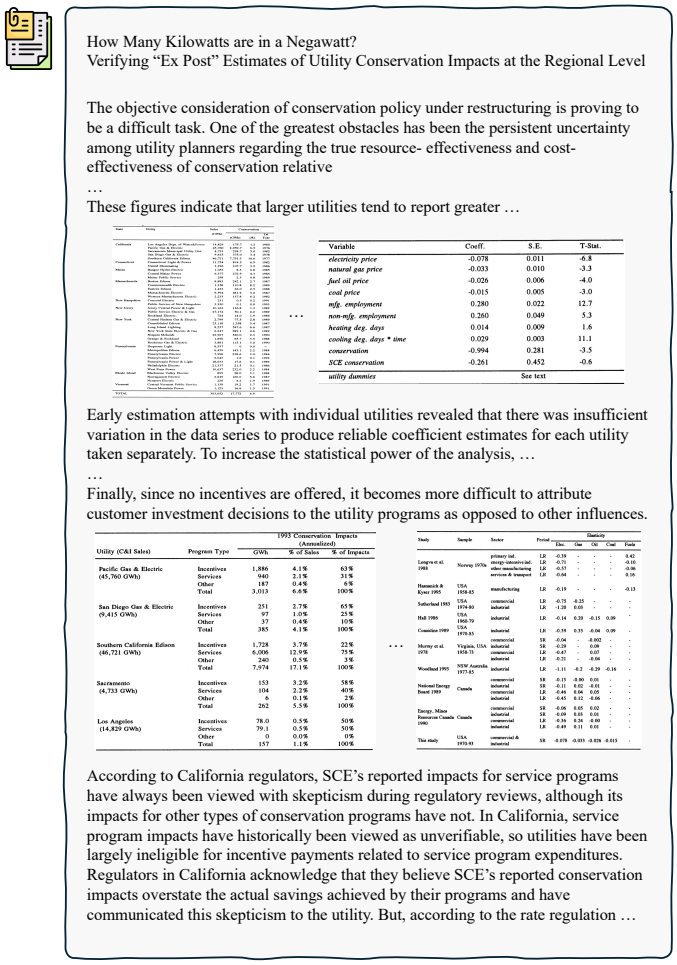


Figure 3. **Visualization of an example from Doc-750K.** The left side presents the interleaved text-image format data obtained through Document Content Extraction, while the right side showcases the annotations generated via Question-Answer Pairs Construction.

example, “<text>\n<image>\n<text>\n<image>” This format captures the document’s textual content, making it easier to construct question-answer pairs.

(2) *Multi-Image Format.* In this format, a document with n pages is rendered as n images, with each image corresponding to a single page. The structure follows the pattern “<image>\n<image>\n<image>”. This format preserves the original layout, enabling the model to learn the overall pagination and visual layout of the document.

After processing the document into contexts in interleaved text-image and paginated image formats, we can not only use these contexts for next-token prediction training but also leverage the document’s content, hierarchical structure, and layout features to flexibly and precisely generate high-quality question-answer pairs.

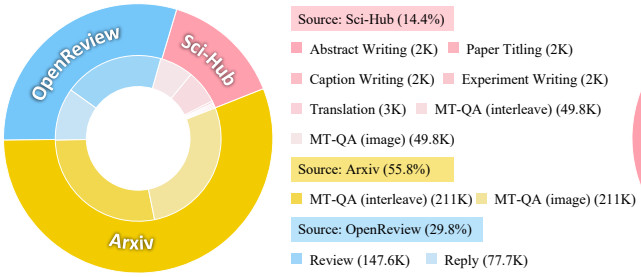
Question-Answer Pairs Construction. In this step, we create question-and-answer pairs tailored to the source, content features, and formatting structure of each document. This process is divided into the following main categories:

(1) *For documents with reliable QA annotations,* like the review and reply in OpenReview, we extract the QA pairs and organize them into conversation format.

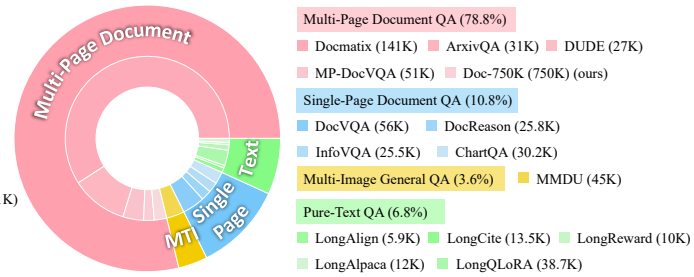
(2) *For documents with a clear textual structure*, such as well-structured papers from Sci-Hub and Arxiv, we convert them to text and segment them, while the model is instructed to generate contents for each segment, including abstracts, experiment descriptions, and captions for figures and tables. The details of each task for structural papers are illustrated in Table 1.

(3) *For other documents*, in addition to using them directly for NTP pretraining, we can input text interspersed with images into MLLMs to obtain QA pairs. To ensure high-quality generated data, we use the state-of-the-art model GPT-4o [56].

Through our pipeline, most data has been processed into high-quality document-level question-answering data, while the remaining data is converted to plain text and used for next-token prediction tasks. Our pipelines are meticu-



(a) Data Distribution of Doc-750K



(b) Data Recipe for Supervised Fine-tuning

Figure 4. **Data distribution of our dataset.** The outer circle shows the distribution of all data categories and the inner circle shows the distribution of data subsets. **Left:** Data distribution of Doc-750K. **Right:** Data distribution of our complete SFT training dataset. Note that the number reported in the figure represents the number of samples. "MT" is short for multi-turn.

Tasks	Questions
Abstract Writing	Read the full text of the paper and provide a concise summary in the form of an abstract.
Paper Titling	Based on the provided abstract or introduction of the research paper, please generate a concise and informative title
Caption Writing	Give the relative texts of the images or tables, please write a caption for each image or table based on the relative texts provided.
Experiment Writing	Please write the "Experiments" section based on the incomplete research paper provided.
Translation	Please read the full text of the following research paper and translate the Experiments section into Chinese.

Table 1. **Questions format for different tasks.** For documents with a clear textual structure, we design several proxy tasks. All tasks leverage the inherent structure of the documents, with answers directly sourced from the original text.

Statistics	Number
Total Questions	758K
Total Images	3.1M
Total Conversations	251K
Multi-Turn Questions	87K
Single-Turn Questions	164K
Average Text Tokens	11245
Average Image Tokens	6178

Table 2. **Key statistics of the Doc-750K datasets.** It comprises 758K questions, 3.1M images, and 251K conversations, including 87K multi-turn and 164K single-turn questions. With an average of 11,245 text tokens and 6,178 image tokens, it highlights the dataset's richness and diversity for multimodal research.

lously designed to ensure high data quality across all generated context. Each LLM-generated sample is explicitly marked in the metadata as model-generated. Across the entire dataset, only 4.8% of the data is LLM-generated, reinforcing the overall reliability and quality of the dataset.

3.2. Multimodal Document Dataset

Data Source. The composition and distribution of our training data are detailed in Figure 4. Specifically, our dataset predominantly consists of academic papers, which constitute approximately 32.6% of the total data. The multimodal data, carefully selected to augment our model's learning dimensions, makes up about 88.8% of our dataset. This strategic distribution is designed to optimize the training process and improve the model's ability to generalize across different types of data inputs.

Dataset Statistics. In our Doc-750K dataset, the majority of the data consists of reliably annotated entries, with OpenReview and Arxiv collectively accounting for 75.4%. The remaining data, sourced from Sci-Hub, is processed using our designed tasks. The overall distribution and number of tasks are shown in Figure 4(a). Our dataset ultimately consists of 251K conversations, comprising a total of 758K questions. Additional statistical details are provided in Table 2. Compared to previous datasets, Doc-750K contains a larger number of images, with an average of four images per conversation segment. Further comparisons with other datasets are shown in Table 3.

3.3. Data Recipe for Supervised Fine-Tuning

Although Doc-750K effectively covers multimodal document QA scenarios, using it directly may lead to model over-fitting on a specific document domain. Therefore, we combine it with several open-source datasets to create a mixed dataset for SFT training. As shown in Figure 4(b), these datasets are organized into 4 categories as follows:

(1) *For multi-page document QA*, Doc-750K serves

Dataset	#Images	#QA Pairs	#Tokens
Docmatix [33]	2,444,750	9,500,000	390,000,000
DocVQA [15]	10,189	39,463	337,829
TextCaps [64]	21,953	21,953	389,658
TextVQA [65]	21,953	34,602	181,918
ST-VQA [8]	17,247	23,121	127,846
OCR-VQA [55]	165,746	801,579	6,073,824
VisualMRC [70]	3,027	11,988	168,828
DUDE [78]	147,597	23,716	11,341,228
Doc-750K (ours)	3,103,494	758,000	5,200,000,000

Table 3. **Comparison with popular VQA datasets.**

as the core dataset, specifically curated to address complex, multi-page document comprehension. Additional datasets such as MP-Docmatix [33], MP-DocVQA [53], DUDE [78], and Taesiri-ArxivQA [69] offer valuable multi-page scenarios requiring inter-page reasoning and contextual retention across sequences.

(2) *For multi-image general QA*, MMDU-45K [48] offers a comprehensive dataset encompassing diverse real-world scenarios, such as natural environments and everyday contexts. It emphasizes multi-turn dialogues and integration of multiple images, supporting the development of systems capable of generating coherent and accurate responses from complex, lengthy inputs.

(3) *For single-page document QA*, We introduce DocVQA [53], DocReason [93], InfoVQA [54], and ChartQA [52] to further enhance the diversity of the SFT dataset. These datasets focus on individual pages with complex layouts, rich textual information, and, in some cases, graphical data interpretation.

(4) *For pure-text QA*, we add datasets including LongAlpaca [11], LongAlpaca-16K-Length [11], LongQLoRA [91], LongCite [96], LongAlign [7], and LongReward [97] to support the assessment of the model's capabilities in QA tasks requiring long-range dependencies.

This expanded dataset provides a balanced foundation for training and evaluating multimodal document understanding models, enhancing robustness and adaptability across diverse document-related VQA tasks.

4. Enhanced Baseline for Document-Level Multimodal Understanding

4.1. Model Architecture

Our model architecture leverages the widely-adopted ViT-MLP-LLM structure [41, 42, 73], consisting of a pre-trained Vision Transformer (ViT), a two-layer MLP projector, and a pre-trained Language Model (LLM). This combination provides a strong baseline for multimodal document analysis, effectively integrating visual and textual information within a unified framework.

4.2. Optimizing Training Efficiency

The training efficiency of MLLMs is hindered by two key challenges: (1) *Inconsistent Sample Lengths.* Samples with different context lengths will result in excessive padding and lower training throughput; and (2) *Limited GPU Memory.* As the model scale and context length increase, GPU memory consumption becomes increasingly unsustainable. To address these issues, we have implemented the following strategies:

(1) *Multimodal Data Packing.* To balance the computational load between the vision model (ViT) and the language model (LLM) while minimizing resource waste caused by padding, we implement a multimodal data-packing strategy. The key idea is to concatenate multiple samples into long sequences to fully utilize the model's input capacity. Specifically, thresholds T_{img} and T_{tok} are set for the number of images and tokens, respectively. Samples are managed using a priority queue, sorted in descending order by the number of images and total tokens. A new sample s attempts to combine with the sample at the front of the priority queue. If the combination meets the thresholds (*i.e.*, T_{img}, T_{tok}), the combined sample is pushed back into the priority queue. If s cannot match with any existing sample in the queue, it is directly added to the queue. When the image number and total token number of the front sample reach one of the thresholds or the number of samples exceeds the maximum limit M, the front sample is dequeued, padded as needed, and sent for training. This strategy optimizes resource utilization and ensures balanced computational workloads. The detailed pseudo-code can be found in the supplementary materials.

(2) *Ring Attention.* We implement the Ring Attention mechanism [43] to alleviate memory constraints associated with processing long sequences. By partitioning sequences into blocks and distributing computation across multiple devices, Ring Attention allows the model to accommodate larger contexts. This approach enables overlapping communication between key-value blocks and attention computations, thereby enhancing parallel processing efficiency. Consequently, Ring Attention improves the model's capacity to handle extended context lengths without exceeding memory limits.

(3) *Liger Kernel.* To further improve memory and computational efficiency, we integrate the Liger Kernel [16], a specialized kernel library optimized for large-scale model training. The Liger Kernel enhances throughput and reduces memory consumption by employing techniques like kernel fusion, in-place operations, and input chunking. Leveraging the Liger Kernel thus enables higher training throughput and addresses memory limitations, allowing for efficient scaling of large multimodal models.

Models	MP-Doc	MMLong-Doc		DocGenome					MM-NIAH			
	ANSL↑	Acc↑	F1↑	Class Acc↑	Title ED↓	Abstract ED↓	SP Acc↑	MP Acc↑	Short	Medium	Long	Overall
Proprietary Models												
Gemini-1.5-Pro [60]	–	28.2	20.6	–	–	–	–	–	73.8	65.2	60.8	67.1
GPT-4o [56]	–	42.8	44.9	97.6	9.5	6.5	71.8	67.6	–	–	–	–
Open-Source Models												
MiniMonkey-2B [28]	70.3	10.3	8.6	57.4	16.5	55.0	40.3	28.9	40.9	26.9	23.5	31.0
InternVL2-2B [13]	71.8	10.5	10.8	60.8	18.4	54.3	39.4	28.9	36.6	21.2	19.4	26.4
InternVL2-2B + RAG [85]	72.6	17.2	16.7	60.8	18.4	54.3	39.4	28.4	36.8	30.2	34.8	33.8
Llama3.2-3B-Instruct† [21]	–	23.7	21.2	85.3	194.7	51.0	40.2	34.9	15.5	2.2	0.5	6.6
Docopilot-2B (ours)	76.2	21.8	16.0	56.2	4.5	43.6	45.1	37.4	58.0	46.7	40.9	49.2
MiniCPM-V2.6-8B [92]	–	16.9	15.4	92.8	10.2	32.6	60.0	54.2	49.0	15.3	0.0	23.4
LLaVA-OneVision-8B [36]	–	10.8	9.6	85.6	49.9	77.5	9.8	7.1	65.7	38.0	0.0	37.7
mPLUG-DocOwl2-8B [27]	69.4	13.4	8.9	–	–	–	–	–	17.9	0.1	0.0	6.6
M3DocRAG [14]	84.4	21.0	22.6	–	–	–	–	–	–	–	–	–
VisRAG-8B [95]	–	18.8	18.3	92.8	10.2	32.6	60.0	50.7	47.1	29.2	29.5	35.8
InternLM2.5-7B-1M† [9]	–	28.7	25.6	92.7	77.6	59.3	42.7	42.5	40.5	37.2	35.1	37.8
InternVL2-8B [13]	79.3	17.4	16.5	90.6	8.2	39.6	56.0	46.1	56.4	37.3	32.4	42.9
InternVL2-8B + RAG [85]	78.7	24.2	24.5	90.6	8.2	39.6	56.0	46.0	55.7	43.4	45.2	48.4
InternVL2-26B [13]	–	15.5	15.4	87.5	16.9	23.3	49.7	42.7	65.0	48.7	41.9	52.8
Docopilot-8B (ours)	81.3	28.8	23.0	93.8	2.0	19.7	53.9	51.9	71.2	57.4	55.3	61.8

Table 4. **Evaluation on multi-page and interleaved VQA benchmarks.** We report the metrics on MP-DocVQA [76] (MP-Doc), MMLongbench-Doc [51] (MMLong-Doc), DocGenome [89], and MM-NIAH [85]. Our model outperforms document-level MLLMs and multimodal RAG methods on multi-page, medium, and long-context QA. The "Short", "Medium", and "Long" in MM-NIAH refer to input length in $[0, 8k]$, $(8k, 32k]$, $(32k, 64k]$, respectively. "†" denotes input documents are parsed by OCR models.

5. Experiments

5.1. Experimental Setup

Training Details. Our model is available in two sizes: Docopilot-2B and Docopilot-8B, both of which are based on the InternVL2 [73] and fine-tuned for one epoch using the data recipe that includes Doc-750K. The training uses a batch size of 128, the AdamW optimizer with a learning rate of 1e-5, weight decay of 0.01 for the 2B variant, and 0.05 for the 8B variant, along with a cosine learning rate schedule. To speed up training, we apply multimodal data packing to reduce padding and a dynamic high-resolution strategy [13] to enhance OCR for document understanding. The maximum number of tiles for multimodal data is limited to 24, and the maximum sequence length is set to 32k tokens.

Baselines. We compare our Docopilot with a series of open-source document-level MLLMs [27, 28, 36, 40, 46, 92, 98] that supports multi-image input and proprietary MLLMs, including Gemini-1.5-pro [60], GPT-4o [57]. For comparison with the commonly used RAG method for handling long documents, we selected the latest multimodal RAG methods VisRAG [95], InternVL + RAG [85], and M3DocRAG [14]. To compare with more long-context large language models, we use InternVL2-8B as the OCR model to extract texts from the documents and images and feed the parsed documents to long-context LLMs [9, 21].

5.2. Multi-Page VQA

Benchmarks. For the multi-page VQA task, we evaluate our model on three benchmarks: (1) **MP-DocVQA** [76], which is designed to evaluate the ability to handle complex questions across multiple scanned document pages. (2) **MMLongbench-Doc** [51], a benchmark for evaluating the performance of MLLMs on multi-modal documents. (3) **DocGenome** [89], a large-scale benchmark for the evaluation of scientific document comprehension.

Models	DocVQA	ChartQA	InfoVQA
Gemini-1.5-Pro [60]	93.1	87.2	81.0
GPT-4o [56]	92.8	85.7	–
MiniMonkey-2B [28]	87.4	76.5	60.1
InternVL2-2B [13]	86.9	76.2	58.9
Docopilot-2B (ours)	87.3	76.4	58.5
Monkey-8B [40]	66.5	65.1	36.1
TextMonkey-9B [46]	73.0	66.9	28.6
mPLUG-DocOwl2-8B [27]	80.7	70.0	46.4
IXC2.5-7B [98]	90.9	82.2	69.9
InternVL2-8B [13]	91.6	83.3	74.8
Docopilot-8B (ours)	92.0	83.3	73.3

Table 5. **Results on single-page VQA benchmarks.** Our Docopilot models perform comparably to baselines [73], demonstrating enhanced long-context modeling without loss on shorter tasks.

Results. As illustrated in Table 4, our model achieves consistent improvements on multi-page QA benchmarks, outperforming previous document-level MLLMs. Notably, our Docopilot-8B surpasses Gemini-1.5-Pro [60] on MMLongBench-Doc, positioning it as the closest open-source model to GPT-4o. In comparison to RAG-based methods [14, 86, 95], our model demonstrates advantages in multi-page scenarios. For example, in the Multi-Page QA of DocGenome benchmark, the RAG method shows a perfor-

mance decline due to the disruption of document continuity while our Docopilot exhibits a significantly stable improvement compared to the baseline, with Docopilot-8B showing an increase of 12.6% over InternVL2-8B.

5.3. Interleaved Long-Context QA

Benchmarks. For the interleaved long-context QA task, we evaluate our models on MM-NIAH [86], a benchmark designed for long multimodal document comprehension.

Results. The right side of Table 4 presents the results of MM-NIAH across context lengths ranging from 1K to 64K. We categorize the context lengths into "Short," "Medium," and "Long" based on the context window of InternVL2 (8K) and Docopilot (32K). Our Docopilot demonstrates exceptional performance in both medium- and long-context scenarios, while maintaining high accuracy in short-context situations. Notably, for QA tasks with context lengths in the range of $(32K, 64K]$, Docopilot-2B outperforms InternVL2-2B by 110%, and Docopilot-8B surpasses InternVL2-8B by 70%. Furthermore, our model performs comparably to the state-of-the-art multimodal long-context model Gemini-1.5-Pro in contexts longer than 8K, establishing a new state-of-the-art performance among open-source long-context MLLMs.

5.4. Single-Page VQA

Benchmarks. For single-page VQA tasks, we evaluate our model on three benchmarks: (1) DocVQA [53], a benchmark for the evaluation of extracting key information from an image of the given document. (2) ChartQA [52], a benchmark for evaluating the reasoning abilities for chart images. (3) InfoVQA [54], a benchmark for infographic image comprehension.

Results. As shown in Table 5, our model achieves comparable performance to baseline models. Across the three benchmarks, Docopilot-2B and InternVL2-2B exhibit comparable results, while Docopilot-8B outperforms InternVL2-8B by 0.4 points in DocVQA. These results demonstrate that Doc-750K effectively enhances the model's long-context modeling capabilities without compromising its performance on shorter documents.

5.5. Ablation Study

Effect of Doc-750K. We conducted ablation studies on MMLongBench-Doc [51] to analyze the impact of our Doc-750K. We divided Doc-750K into 3 parts according to the source of the data: (1) Sci-Hub data; (2) Arxiv data; and (3) OpenReview data. We demonstrate the effects of incorporating each part of the data into the SFT process, reported in Table 6. We observed that with the inclusion of different parts of Doc-750K, the model's performance improves continuously. Utilizing only open-source data results in an inferior F1 score.

Models	Acc	F1
Baseline (InternVL2-2B [73])	10.5	10.8
– Variant1: SFT using data recipe w/o Doc-750K	18.4	9.4
– Variant2: Variant1 + our Sci-Hub data	18.5	15.2
– Variant3: Variant2 + our Arxiv data	20.5	15.5
Docopilot-2B: Variant3 + our OpenReview data	21.8	16.0

Table 6. **Ablation study on the data recipe.** We evaluate the effectiveness of the training data from different sources on MMLongBench-Doc. Our Doc-750K can consistently enhance the ability of the model to understand multi-page documents.

Models	Latency	Acc	F1
MiniCPM-V2.6 [92]	225.4ms	16.9	15.4
VisRAG-12B [95]	288.3ms	18.8	18.3
InternVL2-2B [13]	35.9ms	10.5	10.8
InternVL2-2B + RAG [85]	82.9ms	17.2	16.7
Docopilot-2B (ours)	35.9ms	21.8	16.0
InternVL2-8B [13]	81.0ms	17.4	16.5
InternVL2-8B + RAG [85]	113.4ms	24.2	24.5
Docopilot-8B (ours)	81.0ms	28.8	23.0

Table 7. **Latency analysis.** We evaluate the average token output latency of model outputs on MMLongBench-Doc [51]. RAG-based methods [85, 95] exhibit slower processing speeds due to their two-stage inference process, making them less efficient than document MLLMs for handling multimodal long documents.

Latency Analysis. To compare the latency in inference between RAG methods and our Docopilot, we conducted a latency analysis on MMLongBench-Doc [51], as reported in Table 7. While RAG reduces the document length input to the MLLM, its own time cost remains non-negligible. For instance, InternVL2-2B + RAG is 130% slower than InternVL2-2B, and VisRAG is 28% slower than MiniCPM-V2.6. Our Docopilot does not require additional processes and therefore has the same inference time as baseline models, making it more suitable for analyzing long documents.

6. Conclusions

This work introduced a diverse document-level question-answering dataset that covers complex structures and cross-page dependencies, providing a robust foundation for training and evaluating document understanding models. We also proposed a retrieval-free long-document understanding model that effectively integrates multi-page information, reducing reliance on external retrieval systems. Experimental results show that our model achieves state-of-the-art performance across several document-level QA benchmarks, underscoring its strength in multi-page integration and complex reasoning. Future work will focus on improving computational efficiency, extending the model to larger multimodal tasks, and adapting it to broader applications for enhanced practicality and generalization.

Acknowledgments

This project was supported by the National Key R&D Program of China (No. 2022ZD0161300, 2022ZD0160101), the National Natural Science Foundation of China (No. 62376134, 62372223). Zhe Chen is supported by the Youth PhD Student Research Project under the National Natural Science Foundation (No. 623B2050).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35: 23716–23736, 2022. 2

[2] Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023. 1

[3] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. Docformerv2: Local features for document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 709–718, 2024. 2

[4] Haoli Bai, Zhiguang Liu, Xiaojun Meng, Wentao Li, Shuang Liu, Nian Xie, Rongfu Zheng, Liangwei Wang, Lu Hou, Jiansheng Wei, et al. Wukong-reader: Multi-modal pre-training for fine-grained visual document understanding. *arXiv preprint arXiv:2212.09621*, 2022. 2

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[6] 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

[7] Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*, 2024. 6

[8] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019. 6

[9] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 1, 7

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1

[11] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023. 2, 3, 6

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 1, 2

[13] 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. 1, 2, 7, 8

[14] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024. 1, 2, 3, 7

[15] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *ACL*, pages 845–855, 2018. 6

[16] Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, Yanning Chen, et al. Liger kernel: Efficient triton kernels for llm training. *arXiv preprint arXiv:2410.10989*, 2024. 2, 6

[17] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 3

[18] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*, 35:16344–16359, 2022. 3

[19] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023. 3

[20] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024. 2

[21] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 7

[22] Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024. 1, 3

[23] 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. 3

[24] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1

[25] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*, 2024. 2

[26] Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Simple on-the-fly length generalization for large language models. *arXiv preprint arXiv:2308.16137*, 2023. 3

[27] 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. 7, 2

[28] 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. 7

[29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. Zenodo. Version 0.1. https://doi.org/10.5281/zenodo.5143773, 2021. DOI: 10.5281/zenodo.5143773. 2

[30] Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2(1):17, 2024. 1

[31] Greg Kamradt. Llmtest_needleinahaystack. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2024. Accessed: 2024-11-11. 1

[32] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 3

[33] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 2, 6

[34] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *NIPS*, 36, 2024. 2

[35] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2

[36] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 7

[37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 2

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 2

[39] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: An unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024. 2

[40] 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. *arXiv preprint arXiv:2311.06607*, 2023. 7

[41] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 6

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2023. 1, 2, 6

[43] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023. 2, 3, 6

[44] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1

[45] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 1

[46] 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. 7

[47] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 2

[48] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024. 6

[49] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15630–15640, 2024. 3

[50] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*, 2024. 1

[51] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations, 2024. 1, 7, 8, 5

[52] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, pages 2263–2279, 2022. 6, 8, 1, 5

[53] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021. 6, 8, 1, 5

[54] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 6, 8, 1, 5

[55] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952, 2019. 6

[56] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 1, 4, 7

[57] OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024. 7

[58] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021. 3

[59] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *NeurIPS*, 30, 2017. 2

[60] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 7

[61] Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. Unirag: Universal retrieval augmentation for multi-modal large language models. *arXiv preprint arXiv:2405.10311*, 2024. 1

[62] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 1

[63] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023. 1

[64] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020. 6

[65] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 1, 6

[66] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3

[67] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. In *ICLR*, 2024. 2

[68] Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*, 2022. 3

[69] Mohammad Reza Taesiri. Arxivqa. https://github.com/taesiri/ArXivQA, 2024. 6

[70] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13878–13888, 2021. 6

[71] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 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, 2023. 2

[72] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[73] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. 1, 6, 7, 8

[74] Qwen Team. Qwen2.5: A party of foundation models, 2024. 1

[75] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024. 2

[76] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023. 1, 2, 7, 5

[77] Xiaoguang Tu, Zhi He, Yi Huang, Zhi-Hao Zhang, Ming Yang, and Jian Zhao. An overview of large ai models and their applications. *Visual Intelligence*, 2(1):1–22, 2024. 1

[78] Jordy Van Landeghem, Ruben Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Joziak, Rafał Powalski, Dawid Jurkiewicz, Mickael Coustaty, Bertrand Ackaert, Ernest Val-

veny, et al. Document understanding dataset and evaluation (dude). In *Proceedings IEEE/CVF international conference on computer vision-ICCV 2023*, pages 19528–19540. IEEE/CVF, 2023. 2, 6

[79] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024. 3

[80] 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. *arXiv preprint arXiv:2401.00908*, 2023. 2

[81] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[82] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[83] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024. 1, 2

[84] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *ICLR*, 2024. 1, 2

[85] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024. 7, 8, 3

[86] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024. 2, 7, 8, 5

[87] Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv preprint arXiv:2311.13194*, 2023. 3

[88] 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 models. *arXiv preprint arXiv:2312.06109*, 2023. 3

[89] 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. 1, 7, 2, 5

[90] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 3

[91] Jianxin Yang. Longqlora: Efficient and effective method to extend context length of large language models. *arXiv preprint arXiv:2311.04879*, 2023. 2, 6

[92] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 1, 7, 8

[93] 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. 6

[94] 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. 3

[95] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024. 1, 2, 3, 7, 8

[96] Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, et al. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv e-prints*, pages arXiv–2409, 2024. 2, 6

[97] Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. Longreward: Improving long-context large language models with ai feedback. *arXiv preprint arXiv:2410.21252*, 2024. 2, 6

[98] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 7

[99] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *NIPS*, 36, 2024. 2