M3DOCVQA: Multi-modal Multi-page Multi-document Understanding

Jaemin Cho[1*] Debanjan Mahata[2] Ozan İrsoy[2] Yujie He[2] Mohit Bansal[1]
[1]UNC Chapel Hill [2]Bloomberg
{jmincho,mbansal}@cs.unc.edu {dmahata,oirsoy,yhe247}@bloomberg.net

Abstract

*Document Visual Question Answering (DocVQA) offers a promising approach to extracting insights from large document corpora. However, existing benchmarks focus on evaluating multi-modal understanding within a single document. This gap hinders the development of methods integrating scattered information across pages and documents. To address this, we introduce **M3DocVQA**, the first benchmark designed for multi-modal, multi-page, and multi-document understanding. M3DocVQA comprises over 3,000 PDF documents with more than 40,000 pages, offering a challenging environment where evidence is distributed across diverse sources and modalities. Alongside the dataset, we introduce M3DocRAG, a baseline method based on multi-modal retrieval-augmented generation. M3DocRAG flexibly handles both single and multiple document settings while preserving critical visual information, establishing a useful starting point for future work in open-domain multi-modal document understanding. Our experiments across three benchmarks (M3DocVQA, MMLongBench-Doc, and MP-DocVQA) show that existing methods struggle with open-domain question answering over extensive, multi-modal documents. Although M3DocRAG has shown promising performance, there is large room for future improvement. We provide comprehensive ablation studies of different indexing, multi-modal language models, and multi-modal retrieval models, along with qualitative examples to guide future research.*

1. Introduction

Document visual question answering (DocVQA) [16, 32, 42, 44, 58] is a multi-modal task that answers textual questions by interpreting information contained within document images. The capability of accurately and efficiently answering questions across numerous, lengthy documents with intricate layouts would greatly benefit many domains such as finance, healthcare, and law, where document AI assistants can streamline the daily processing of large volumes of documents, improving productivity and enabling faster, more informed decision-making. However, existing DocVQA benchmarks focus on evaluating question answering (QA) capabilities within a single document, so their questions assume that a QA model already knows the context of that specific document. For example, they have questions given a single-page CV and *"In which year did the author publish their first journal article?"* as shown in Fig. 1 (left). This gap hinders the development of methods integrating scattered information across pages and documents.

To address this limitation, we introduce **M3DOCVQA** (**M**ulti-modal **M**ulti-page **M**ulti-**Doc**ument **V**isual **Q**uestion **A**nswering), an open-domain dataset that significantly raises the challenge of DocVQA to answering questions from a large document corpus (Sec. 2). As exemplified in Fig. 1 (right), M3DOCVQA supports scenarios like reviewing a corpus of thousands of multi-page CVs and answering a question like *"Which candidate has published in ICCV on document understanding?"* By extending the MultimodalQA dataset's [55] closed-domain context to an open-domain setting, M3DOCVQA introduces 2,441 questions spanning 3,368 PDF documents, which collectively contain over 41,005 pages of diverse multi-modal content, including text, images, and tables. This dataset presents real-world challenges by requiring models to navigate complex reasoning paths across pages and within various types of document elements, better reflecting the intricacies of document understanding.

As a useful starting point for M3DOCVQA, we introduce **M3DOCRAG**, a baseline method based on multi-modal retrieval-augmented generation (Sec. 3). M3DOCRAG retrieves relevant document pages using a multi-modal retrieval model, such as ColPali [19], and generates answers to questions from the retrieved pages using a multi-modal language model (MLM), such as Qwen2-VL [60]. M3DOCRAG operates in three stages: In (1) document embedding (Sec. 3.1), we convert all document pages into RGB images and extract visual embeddings (*e.g.*, via ColPali) from the page images. In (2) page retrieval (Sec. 3.2), we retrieve the top-K pages of high similarity

Work done during an internship at Bloomberg as a recipient of the Bloomberg Data Science Ph.D. Fellowship.

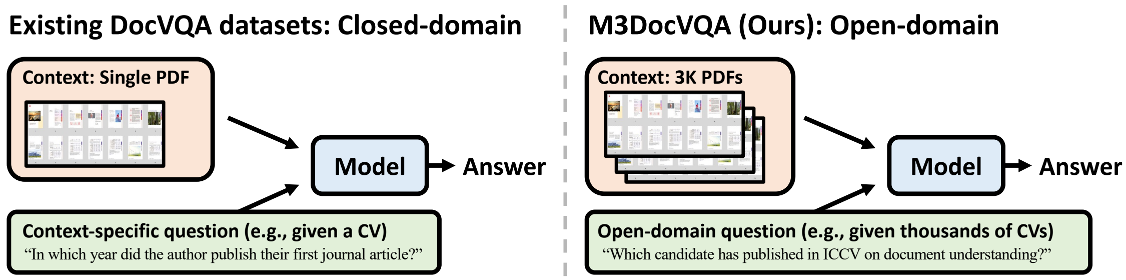

Figure 1. Comparison of existing DocVQA datasets (left; *e.g.*, DocVQA [44]) and our **M3DocVQA** dataset (right). In contrast to previous DocVQA datasets that have questions that are specific to a single provided PDF, M3DocVQA has information-seeking questions that benchmark open-domain question answering capabilities across more than 3,000 PDF documents (*i.e.*, 40,000+ pages).

with text queries (*e.g.*, MaxSim operator for ColPali). For the open-domain setting, we create approximate page indices, such as inverted file index (IVF) [53, 68], for faster search. In (3) question answering (Sec. 3.3), we conduct visual question answering with MLM to obtain the final answer. M3DocRAG can flexibly handle DocVQA in both closed domain (*i.e.*, a single document) and open-domain (*i.e.*, a large corpus of documents) settings.

We benchmark state-of-the-art methods and M3DocRAG baseline in three datasets: M3DocVQA, MMLongBench-Doc [42], and MP-DocVQA [58], which cover both open-domain (Sec. 5.1) and closed-domain (Sec. 5.2) DocVQA settings. We find existing methods struggle with open-domain document understanding in M3DocVQA, and M3DocRAG achieves the text-only RAG baseline, but there remains large room for future improvement. We also provide a comprehensive analysis (Sec. 5.3) about different indexing, MLMs, and retrieval components and qualitative examples where M3DocRAG can successfully handle various scenarios, such as when the relevant information exists across multiple pages and when answer evidence only exists in images.

2. M3DocVQA: A New Benchmark for Multi-modal, Multi-page, Multi-document Understanding

We present **M3DocVQA** (**M**ulti-modal **M**ulti-page **M**ulti-**Doc**ument **V**isual **Q**uestion **A**nswering), a new open-domain DocVQA benchmark designed to evaluate the ability to answer questions using multi-modal information from a large corpus of documents.

As illustrated in Fig. 1 and Table 1, existing DocVQA datasets [16, 32, 42, 44, 58] primarily focus on evaluating question answering within the context of a single document (*i.e.*, closed-domain). These datasets are not well-suited for benchmarking open-domain visual question answering, where relevant information, often in multiple modalities such as text, images, and tables, must be retrieved from multiple documents. This limitation stems from their questions being designed around specific content on certain pages within a single document. In real-world scenarios, users often seek answers that span across multiple documents and modalities, making open-domain settings critical. However, the questions in the existing DocVQA datasets are not applicable in such an open-domain setting. For example, a question from MP-DocVQA, such as *"What was the gross profit in the year 2009?"* assumes that the model already has access to specific information within the document. M3DocVQA challenges models in an open-domain DocVQA setting, where they must navigate a large 'haystack' of multi-modal documents and re-

Table 1. Comparison of recent DocVQA datasets with the proposed M3DocVQA dataset in terms of context size.

Datasets	Multi-page	Multi-document	Avg. # pages per question
DocVQA [44]	✗	✗	1
MP-DocVQA [58]	✓	✗	8.3
DUDE [32]	✓	✗	5.7
MMVQA [16]	✓	✗	9.6
MMLongBench-Doc [42]	✓	✗	47.5
M3DocVQA (ours)	✓	✓	41,005

Table 2. M3DocVQA statistics.

# Documents	3,368
# Pages	41,005
- Avg. # pages per document	12.2
# Questions	2,441
(Answer modalities)	
- Text	1,048 (35.2%)
- Table	860 (42.9%)
- Image	533 (21.8%)
(Question hops)	
- Single-hop	1,461 (59.9%)
- Multi-hop	980 (40.1%)
Avg. # characters for question	100.8
Avg. # characters for answer	7.1

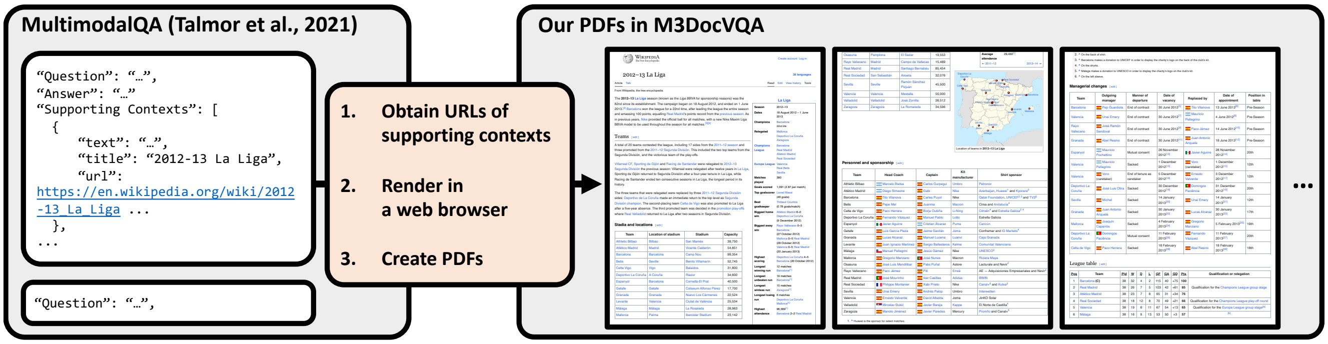

Figure 2. Illustration of PDF collections in M3DocVQA. We first collect the URLs of all supporting contexts (Wikipedia documents) of individual questions of MultimodalQA [55]. Then, we create PDF versions from their URLs by rendering them in a web browser.

trieve relevant information to generate the final answer. The dataset consists of 2,441 questions spread across 3,368 PDF documents, totaling 41,005 pages. Each question is supported by evidence found in one or more documents, spanning multiple modalities such as text, images, and tables, capturing the complexity and diversity typical of real-world documents. In Table 2, we provide detailed statistics of M3DocVQA. Additionally, we provide the training split, consisting of 24,162 Wikipedia PDFs. Although the documents in the training split were not utilized in our experiments, they offer future researchers the opportunity to explore even larger-scale retrieval tasks or use the documents for training models, further expanding the potential applications of M3DocVQA.

To create M3DocVQA, we extend the question-answer pairs from a short-context VQA dataset to a more complex setting that includes 1) PDF documents and 2) open-domain contexts. Specifically, we use the question-answer pairs from the development split[1] of MultimodalQA [55], where models answer multi-hop questions based on short multimodal contexts (*e.g.*, short text passages, 1-2 images, a table) sourced from Wikipedia. We retrieved the URLs of all Wikipedia documents used as context in any of the MultimodalQA development split questions. Then we generated PDF versions of the Wikipedia pages by rendering them in a Chromium web browser [57], using the Playwright Python package [46]. These PDFs retain all vector graphics and metadata, ensuring zoom-in functionality and maintaining operational hyperlinks. In addition, no objects are split between different pages in the resulting PDFs.

While both M3DocVQA and MultimodalQA [55] share the goal of evaluating question answering given multimodal context, M3DocVQA introduces a more demanding scenario by requiring models to retrieve relevant information from a large set of documents, as opposed to being provided with a short context. In MultimodalQA, models are given short, curated context (*e.g.*, a paragraph from a Wikipedia document) that directly contains the information needed to answer the questions, simplifying the task to reasoning within the provided material. In contrast, M3DocVQA presents an open-domain setting, where models must retrieve information from a diverse collection of 3,368 PDF documents before attempting to answer any question. This not only requires handling large-scale document retrieval but also dealing with multi-modal content–text, images, and tables–distributed across multiple documents. This key distinction highlights M3DocVQA's ability to simulate real-world challenges, where the relevant data is often spread across multiple sources. Consequently, M3DocVQA serves as a robust benchmark for retrieval-augmented generation tasks in document understanding, pushing the boundaries of models to deal with large-scale, multi-modal, and multi-document settings.

[1]The test split of MultimodalQA [55] is unavailable, and previous works have used the development split for comparison.

3. M3DocRAG: A New Baseline for Open-domain Document Understanding

As a useful starting point for M3DocVQA, we propose **M3DocRAG**, a baseline method based on multi-modal retrieval-augmented generation. As illustrated in Fig. 3, M3DocRAG operates in three stages: (1) encoding document images into visual embeddings (Sec. 3.1), (2) retrieving relevant document pages (Sec. 3.2), and (3) generating answers to questions based on the retrieved pages (Sec. 3.3). Below, we explain the problem definition and the details of each stage.

Problem definition. We define a corpus of documents as $C = \{D_1, D_2, \ldots, D_M\}$, where M is the total number of documents, and each document D_i consists of a set of pages, P_i, represented as RGB images. From the documents in C, we construct a global set of page images $P = \bigcup_{i=1}^{M} P_i = \{p_1, p_2, \ldots, p_N\}$, where each p_j represents an individual page image, and N is the total number of page

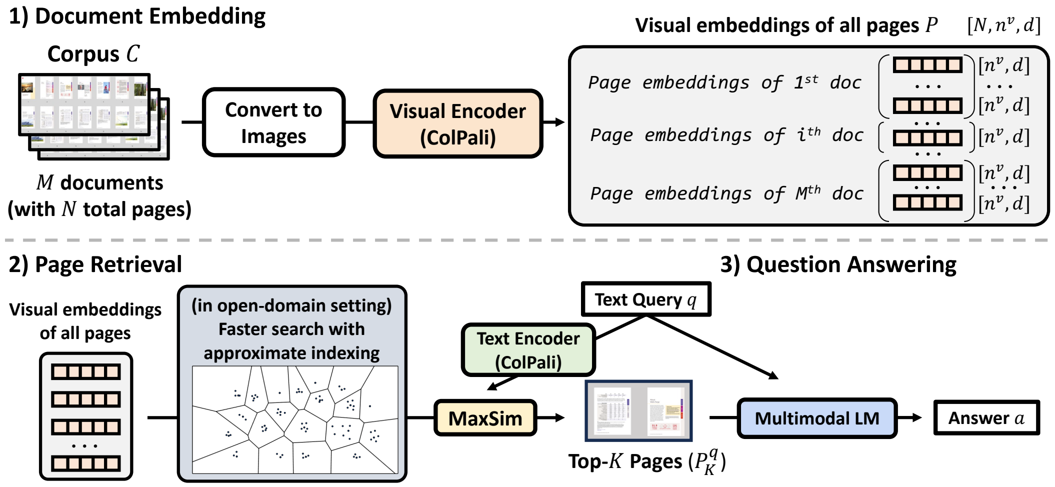

Figure 3. Our M3DocRAG framework (Sec. 3) consists of three stages: (1) document embedding (Sec. 3.1), (2) page retrieval (Sec. 3.2), and (3) question answering (Sec. 3.3). In **(1) document embedding**, we extract visual embedding (with ColPali) to represent each page from all PDF documents. In **(2) page retrieval**, we retrieve the top-K pages of high relevance (MaxSim scores) with text queries. In an open-domain setting, we create approximate page indices for faster search. In **(3) question answering**, we conduct visual question answering with multi-modal LM (*e.g.* Qwen2-VL) to obtain the final answer.

images across all documents in C (*i.e.*, $N = \sum_{i=1}^{M} |P_i|$). The objective of M3DocRAG is to accurately answer a given question q using the multi-modal information available in the corpus of documents C. First, we identify P_K^q, the top K ($\ll N$) pages that are most relevant to answering the query q from the global page set P. Then, we obtain the final answer with a question answering model that takes retrieved page images P_K^q and query q as inputs. The problem of question answering can be categorized into two settings with different document context sizes:

Closed-domain question answering – The query q should be answerable from a given single document D_i. The retrieval model outputs the top K relevant page images P_K^q, from the page images P_i of the document D_i.

Open-domain question answering – The query q may require information from single or multiple documents within the entire document corpus C. The retrieval model outputs the top K relevant page images P_K^q from the entire set of page images P.

3.1. Document Embedding

In M3DocRAG, both textual query q and page images P are projected into a shared multi-modal embedding space using ColPali [19]. ColPali is a multi-modal retrieval model based on a late interaction mechanism, which encodes the text and image inputs into unified vector representations and retrieves the top K most relevant images. ColPali adopts both training objective and similarity scoring from ColBERT [30, 51], which utilizes a shared architecture to encode either textual or visual inputs. In our framework, each page $p \subseteq P_i$ of a document D_i is treated as a single image with fixed dimensions (width $\times$ height).

From an image of a page, we extract a dense visual embedding $E^p \in \mathbb{R}^{n^v \times d}$, where n^v represents the number of visual tokens per page (which remains constant across all pages), and d denotes the embedding dimension (*e.g.*, 128). For a textual query q, we similarly obtain an embedding $E^q \in \mathbb{R}^{n^q \times d}$, where n^q is the number of text tokens.

For efficiency, we treat each page of a document independently. This allows us to flatten all pages in the document corpus C into a single page-level embedding tensor: $E^{\mathrm{C}} \in \mathbb{R}^{N \times n^v \times d}$, where N represents the total number of pages in the entire document corpus, n^v is the number of visual tokens per page, and d is the embedding dimension. M3DocRAG can flexibly adapt to different retrieval settings, such as a single-page document ($N = 1$), a single document with multiple pages (*e.g.* $N = 100$), and a large corpus of multi-page documents (*e.g.* $N > 1,000$).

3.2. Page Retrieval

The relevance between the query q and the page p is computed using the `MaxSim` score $s(q, p)$:

$$s(q,p) = \sum_{i=1}^{n^q} \max_{j \in [n^v]} E^q_{i,\cdot} \cdot E^p_{j,\cdot}$$

where $\cdot$ denotes the dot product, and $E_{i,\cdot} \in \mathbb{R}^d$ denotes the i-th row (vector) of the embedding matrix $E \in \mathbb{R}^{n \times d}$. We then identify P_K^q, the top K ($\ll N$) pages that are most relevant to answering the query q; *i.e.* we search K pages scoring highest $s(q, p)$. That is,

$$P_K^q = \{p_1^q, p_2^q, \ldots, p_K^q\} = \text{argtop-k}_{p \in P}\, s(q,p)$$

Approximate indexing for open-domain page retrieval. Searching pages over in a large document corpus can be time-consuming and computationally expensive. When a faster search is desired, we create page indices offline by applying approximate nearest neighborhood search, based on Faiss [18, 27]. We use exact search for closed-domain page retrieval and employ inverted file index (IVF) [53, 68] (`IVFFlat` in Faiss) for an open-domain setting, which could reduce page retrieval latency from 20s/query to less than 2s/query when searching across 40K pages. See Sec. 5.3 for a detailed comparison of speed-accuracy trade-offs across different indexing methods.

3.3. Question Answering

We run visual question answering by giving the text query q and retrieved page images P_K^q to a multi-modal language model to obtain the final answer. For this, we employ multi-modal language models (*e.g*. Qwen2-VL [60]) that consist of a visual encoder $\texttt{Enc}^{\texttt{Vis}}$ and a language model $\texttt{LM}$. The visual encoder takes K retrieved page images P_K^q as inputs and outputs visual embeddings (different from ColPali encoder's outputs). The language model takes the visual embeddings and text embeddings of query q as inputs and outputs the final answer a in an autoregressive manner:

$$a = \texttt{LM}(\texttt{Enc}^{\texttt{Vis}}(P_K^q), q).$$

4. Experiment Setup

Datasets. We benchmark M3DocRAG on three PDF document understanding datasets that represent different scenarios: (1) M3DocVQA (Open-domain DocVQA); (2) MMLongBench-Doc [42] (Closed-domain DocVQA); (3) MP-DocVQA [58] (Closed-domain DocVQA). In M3DocVQA, M3DocRAG processes over 3,000 PDFs, totaling more than 40,000 pages. For MP-DocVQA, models handle a single PDF with up to 20 pages for each question. For MMLongBench-Doc, models handle a single PDF with up to 120 pages for each question.

Evaluation Metrics. For M3DocVQA, we follow the evaluation setup of MultimodalQA [55]. For MMLongBench-Doc [42] and MP-DocVQA [58], we follow their official evaluation setups. For M3DocVQA, we evaluate answer accuracy with exact match (EM) and F1. For MMLongBench-Doc, we extract short answers with GPT4o [47] from the model outputs and report answer accuracy with generalized accuracy (based on a rule-based evaluation script covering different answer types) and F1 score. For MP-DocVQA, we report answer accuracy with ANLS [8] and page retrieval with accuracy (same as recall@1, as there is a single page annotation for each question) by submitting the generation results to the test server.[2]

[2] https://rrc.cvc.uab.es/?ch=17&com=tasks

Models. We mainly experiment with the ColPali v1 [19][3] retrieval model and various recent open source multi-modal LMs with <10B parameters, including Idefics 2 [34], Idefics 3 [33], InternVL 2 [12], and Qwen2-VL [60]. We also experiment with a text-based RAG pipeline by combining recent widely used text retrieval and language models: ColBERT v2 [51] and Llama 3.1 [38]. For reproducible evaluation, we use deterministic greedy decoding for answer generation. We compare these multi-modal and text-based RAG pipelines with recent top entries with comparable parameters (<10B) reported on the leaderboards.

Other implementation details. We use PyTorch [48, 49], Transformers [61], and FlashAttention-2 [14] libraries for running models. We use Tesseract [54] for OCR in text RAG baselines, following Ma et al. [42]. We use Faiss [18, 27] for document indexing. We use the pdf2image [6] library to convert each PDF page into an RGB image with a resolution of DPI=144. While all PDF pages in M3DocVQA have the same size – 8.5 (width) × 11 (height) in inches (*i.e*. US letter size) and 1224 (width) × 1584 (height) in pixels, in MP-DocVQA and MMLongBench-Doc datasets, pages have slightly different sizes. To handle this, we resize page images to the most common image size within the dataset – 1700 (width) × 2200 (height) for MP-DocVQA, and to the most common image size within each PDF document for MMLongBench-Doc. All experiments are conducted with a single H100 80GB GPU. We provide up to 4 pages as visual inputs to our multi-modal LMs, the maximum number of images we could fit in the single GPU.

5. Results and Key Findings

In the following, we describe experiment results of M3DocRAG and baselines in both open-domain (Sec. 5.1) and closed-domain settings (Sec. 5.2). Next, we provide ablation studies (Sec. 5.3) about different page indexing strategies, multi-modal LMs, and retrieval models. Lastly, we show a qualitative example (Sec. 5.4) where M3DocRAG can tackle M3DocVQA questions whose answer source exists in the visual modality. Please also see the appendix for additional qualitative examples.

5.1. Open-domain DocVQA

Multi-modal RAG outperforms text RAG, especially on non-text evidence sources. Table 3 shows the evaluation results on M3DocVQA. As a model needs to find relevant documents from 3,000+ PDFs for each question, we focus solely on RAG pipelines. We observe that our M3DocRAG (ColPali + Qwen2-VL 7B) outperforms text RAG (ColBERT v2 + Llama 3.1 8B), across all different

[3] https://huggingface.co/vidore/colpali

Table 3. Open-domain DocVQA evaluation results on M3DocVQA. The scores are based on F1, unless otherwise noted.

Method	# Pages	Evidence Modalities			Question Hops		Overall	
		Image	Table	Text	Single-hop	Multi-hop	EM	F1
Text RAG (w/ ColBERT v2)								
Llama 3.1 8B	1	8.3	15.7	29.6	25.3	12.3	15.4	20.0
Llama 3.1 8B	2	7.7	16.8	31.7	27.4	12.1	15.8	21.2
Llama 3.1 8B	4	7.8	21.0	34.1	29.4	15.2	17.8	23.7
M3DocRAG *(w/ ColPali)*								
Qwen2-VL 7B (Ours)	1	25.1	27.8	39.6	37.2	25.0	27.9	32.3
Qwen2-VL 7B (Ours)	2	**26.8**	**30.4**	**42.1**	41.0	25.2	29.9	34.6
Qwen2-VL 7B (Ours)	4	24.7	**30.4**	41.2	**43.2**	**26.6**	**31.4**	**36.5**

Table 4. Closed-domain DocVQA evaluation results on MMLongBench-Doc. We report the generalized accuracy (ACC) across five evidence source modalities: text (TXT), layout (LAY), chart (CHA), table (TAB), and image (IMG), and three evidence locations: single-page (SIN), cross-page (MUL), and unanswerable (UNA). The scores from non-RAG methods are from Ma et al. [42].

Method	# Pages	Evidence Modalities					Evidence Locations			Overall	
		TXT	LAY	CHA	TAB	IMG	SIN	MUL	UNA	ACC	F1
Text Pipeline											
LMs											
ChatGLM-128k [5]	up to 120	23.4	12.7	9.7	10.2	12.2	18.8	11.5	18.1	16.3	14.9
Mistral-Instruct-v0.2 [26]	up to 120	19.9	13.4	10.2	10.1	11.0	16.9	11.3	24.1	16.4	13.8
Text RAG											
ColBERT v2 + Llama 3.1	1	20.1	14.8	12.7	17.4	7.4	21.8	7.8	**41.3**	21.0	16.1
ColBERT v2 + Llama 3.1	4	23.7	17.7	14.9	**24.0**	11.9	25.7	12.2	38.1	**23.5**	19.7
Multi-modal Pipeline											
Multi-modal LMs											
DeepSeek-VL-Chat [39]	up to 120	7.2	6.5	1.6	5.2	7.6	5.2	7.0	**12.8**	7.4	5.4
Idefics2 [34]	up to 120	9.0	10.6	4.8	4.1	8.7	7.7	7.2	5.0	7.0	6.8
MiniCPM-Llama3-V2.5 [62, 66]	up to 120	11.9	10.8	5.1	5.9	12.2	9.5	9.5	4.5	8.5	8.6
InternLM-XC2-4KHD [17]	up to 120	9.9	14.3	7.7	6.3	13.0	12.6	7.6	9.6	10.3	9.8
mPLUG-DocOwl 1.5 [23]	up to 120	8.2	8.4	2.0	3.4	9.9	7.4	6.4	6.2	6.9	6.3
Qwen-VL-Chat [4]	up to 120	5.5	9.0	5.4	2.2	6.9	5.2	7.1	6.2	6.1	5.4
Monkey-Chat [37]	up to 120	6.8	7.2	3.6	6.7	9.4	6.6	6.2	6.2	6.2	5.6
M3DocRAG											
ColPali + Idefics2 (Ours)	1	10.9	11.1	6.0	7.7	15.7	15.4	7.2	8.1	11.2	11.0
ColPali + Qwen2-VL 7B (Ours)	1	25.7	21.0	18.5	16.4	19.7	30.4	10.6	5.8	18.8	20.1
ColPali + Qwen2-VL 7B (Ours)	4	**30.0**	**23.5**	**18.9**	20.1	**20.8**	**32.4**	**14.8**	5.8	21.0	**22.6**

evidence modalities / question hops / # pages. The performance gap is especially big when the evidence involves images, underscoring that M3DocRAG addresses the information loss over non-textual content by text-only pipelines. We also notice that providing more retrieved pages as context generally increases the performance of both text RAG and M3DocRAG (using the top 4 pages yields higher performance than using only the top 1 or 2 pages).

5.2. Closed-domain DocVQA

Multi-modal RAG boosts long document understanding of MLMs. In MMLongBench-Doc, the models must handle a long PDF document (up to 120 pages) for each question. Since many multi-modal LMs have limited context length, Ma et al. [42] employed a concatenation strategy that combines all screenshot pages into either 1 or 5 images and inputs these concatenated images to multi-modal LMs. Table 4 shows that M3DocRAG with Idefics2 surpass Idefics2 without RAG, as well as all previous multi-modal entries. In addition, M3DocRAG with Qwen2-VL achieves the best scores in overall F1 and most evidence modality/page settings. This demonstrates the effectiveness of multi-modal retrieval over handling many pages by concatenating low-resolution images. As observed in M3DocVQA experiments, we also notice that providing more retrieved pages as context generally increases the performance of both text RAG and M3DocRAG (using the top 4 pages yields higher performance than using only the top 1 page).

Table 5. Closed-domain DocVQA evaluation results on MP-DocVQA. The RAG methods retrieve a single page to the downstream QA models.

Method	Answer Accuracy ANLS	Page Retrieval R@1
Multi-modal LMs		
Arctic-TILT 0.8B [10]	0.8122	50.79
GRAM [9]	0.8032	19.98
GRAM C-Former [9]	0.7812	19.98
ScreenAI 5B [3]	0.7711	77.88
Text RAG		
ColBERT v2 + Llama 3.1 8B	0.5603	75.33
M3DocRAG		
ColPali + Qwen2-VL 7B (Ours)	**0.8444**	**81.05**

M3DocRAG achieves the state-of-the-art performance in MP-DocVQA. In MP-DocVQA, the models must handle a PDF document of up to 20 pages for each question. Table 5 presents the top-performing entries in the MP-DocVQA test split leaderboard, comparing text-based and multi-modal RAG pipelines. While the text RAG (ColBERT v2 + Llama 3.1) falls short compared to existing approaches, all multi-modal RAG pipelines outperform their text-based counterpart. Notably, the M3DocRAG delivers the state-of-the-art results on MP-DocVQA.

5.3. Additional Analysis of M3DocRAG

Different page indexing: speed and accuracy. In Table 6, we analyze the speed and accuracy of M3DocRAG pipeline with different document embedding indexing methods. While the naive indexing with exact search (`FlatIP`) is slow (21s per query), we find that using approximate indexing such as inverted file [53, 68] (`IVFFlat`) and product quantization [28] (`IVFPQ`) can retain most of the accuracy, while making the search significantly faster (< 2s per query). We use `FlatIP+IVFFlat` indexing by default, and users can choose appropriate indexing methods depending on their requirements.

Different QA models. In Table 7, we compare four different QA models in the M3DocRAG framework: Idefics2 8B [34], Idefics3 8B [33], InternVL2 8B [12], InternVL2.5 [13] and Qwen2-VL 7B [60]. The Qwen2-VL 7B model outperforms other MLMs in all three benchmarks. Thus, we use the model as the default MLM component for M3DocRAG.

Different retrieval models. In Table 8, we compare different text-only (ColBERTv2 [51]) and multi-modal (CLIP [50], DSE [41], VisRAG [65], and ColPali) retrieval models on M3DocVQA. ColBERTv2 and ColPali use late-interaction [30] score calculation, while DSE and CLIP use dot-product scores. We find that ColPali achieves the best performance in M3DocVQA, even outperforming DSE trained on Wikipedia document screenshots, showing the effectiveness of late-interaction approaches for multi-modal document retrieval. Thus, we use ColPali as the default retrieval model for M3DocRAG. Users should note that late-interaction approaches (ColBERTv2 and ColPali) requires more storage requirements than dot-product approaches (CLIP, DSE, and VisRAG), as they store multiple vectors instead of a single vector per document.

Table 6. Speed-accuracy tradeoff with different indexing strategies on M3DocVQA. Backbones: ColPali + Qwen2-VL 7B.

# Pages	Indexing	Latency (s) (↓)		Accuracy (↑)	
		Retrieval	VQA	EM	F1
1	FlatIP	21.0	1.1	28.9	33.7
1	FlatIP + IVFFlat	1.8	1.1	27.9	32.3
1	FlatIP + IVFPQ	0.2	1.1	25.9	30.3
2	FlatIP + IVFFlat	1.8	2.4	29.9	34.6
2	FlatIP + IVFPQ	0.2	2.4	29.0	33.5
4	FlatIP + IVFFlat	1.8	4.8	31.4	36.5
4	FlatIP + IVFPQ	0.2	4.8	29.9	34.7

Table 7. Comparison of different QA models within RAG pipelines, evaluated on M3DocVQA.

QA models	M3DocVQA F1 ↑
M3DocRAG w/ ColPali	
Idefics2 8B	27.8
Idefics3 8B	31.8
InternVL 2 8B	30.9
InternVL 2.5 8B	32.1
Qwen2 VL 7B	**32.3**
Qwen2.5 VL 7B	29.1
Text RAG w/ ColBERTv2	
Llama 3.1	18.8
InternVL 2.5 (text only)	17.5
Qwen2 VL 7B	19.5
Qwen2.5 VL 7B	20.9

Table 8. Comparison of different retrieval models on M3DocVQA. QA model=InternVL 2.5. Batch size=1. Precision=bfloat16. Measured on a single A6000 48GB GPU.

Retrievers	Embedding Time (s/page) ↓	Embed/Index Storage (KB/page) ↓	Accuracy F1 ↑
Text RAG w/ OCR			
ColBERTv2	2.21	60.6/118.7	17.5
M3DocRAG			
CLIP (ViT-L/14)	**0.04**	**1.7/3.1**	23.8
DSE (Qwen2-2B)	0.15	3.2/6.2	26.6
VisRAG	0.37	4.7/9.2	24.7
ColPali	0.08	227.8/530.4	**32.1**

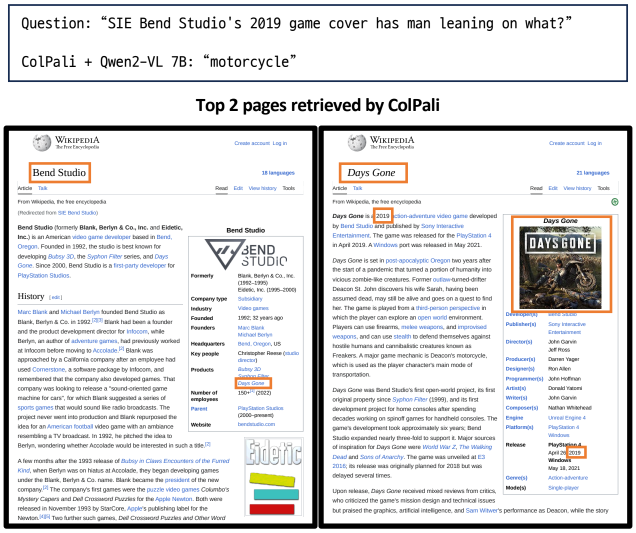

Figure 4. Qualitative example of M3DocRAG on M3DocVQA. Image regions relevant to the question/answer are highlighted with orange boxes. The answer is only stored visually within the game logo, where a man is leaning on a motorcycle. Best viewed by zooming in for details. See additional examples in appendix.

Document as pixels vs. text. We compare pixel-based and text-based representations of documents using the same MLM, InternVL 2.5, which can optionally take image input. The blue rows of Table 7 show that the pixel-based representation outperforms the text-based representation of documents. Table 8 also shows that text embedding (based on OCR) from PDFs is much slower than visual embedding due to additional costs incurred by the OCR model.

5.4. Qualitative Examples

we provide qualitative examples of M3DocRAG (ColPali + Qwen2-VL 7B)'s question answering results on several M3DocVQA examples. In Fig. 4, the answer information is only visually stored within the game logo (‘man is leaning on a motorcycle’), and M3DocRAG could find the information. Please see the appendix for additional qualitative examples where M3DocRAG can tackle M3DocVQA questions whose answer source exists in various modalities.

6. Related Work

Document visual question answering. Mathew et al. [44] proposed document visual question answering (DocVQA) task, where a model extracts information from documents by treating them as images, as in generic visual question answering [1]. Most research on DocVQA focuses on handling a single-page document [23, 24, 31, 35, 43, 44, 56, 59, 64], and it has been now a common practice to include the single-page DocVQA [44] as a part of the image understanding evaluation suite among recent MLMs [7, 12, 21, 33, 47, 60]. Several recent works propose DocVQA benchmarks with multi-page documents [15, 16, 32, 42, 58]. However, all previous DocVQA benchmarks have focused on handling questions in the context of a specific document, such as “What was the gross profit in the year 2009?”. While this is probably due to the limited context length of the backbone multi-modal LMs, this does not reflect real-world scenarios, where users often ask questions that require information across different pages/documents. We address the limitation by proposing M3DocVQA that benchmarks open-domain document understanding capabilities from over 3,000 documents.

Retrieval-augmented generation. Retrieval-augmented generation (RAG) [36] has emerged as a hybrid approach combining retrieval systems with generative models to improve the quality and relevance of generated content [20]. RAG has been widely studied for open-domain question answering [2, 22, 25, 29, 40, 67], where the community has well-established practices for text-based pipelines. A line of work in VQA studies RAG on visual questions that require world knowledge [11, 45, 52, 63], but their retrieval context is usually generic images and/or short text snippets and does not cover DocVQA settings. To the best of our knowledge, no prior work has explored RAG for multi-modal document understanding only with multi-modal models (instead of using OCR methods). Our M3DocRAG tackles open-domain question answering over documents with complex multi-modal contexts.

7. Conclusion

We introduce M3DocVQA, the first benchmark that evaluates open-domain multi-modal document understanding capabilities. In contrast to previous DocVQA datasets that evaluate question answering within the context of single document, M3DocVQA offers a challenging question answering task where the answers exist among 3,000+ PDF documents, totaling more than 40,000 pages, containing various modalities such as images, text, and tables. We also introduce M3DocRAG, a multi-modal RAG baseline that flexibly accommodates various document contexts, question hops and evidence modalities. We benchmark state-of-the-art methods and M3DocRAG baseline in three datasets: M3DocVQA, MP-DocVQA, and MMLongBench-Doc. Existing methods struggle with open-domain document understanding in M3DocVQA, and M3DocRAG achieves the text-only RAG baseline, but there remains large room for future improvement. We hope our work encourages future advancements in multi-modal frameworks for document understanding, paving the way for more robust, scalable, and practical solutions in real-world applications.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 8

[2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023. 8

[3] Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. ScreenAI: A Vision-Language Model for UI and Infographics Understanding, 2024. 7

[4] 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*, abs/2308.12966, 2023. 6

[5] 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*, abs/2401.18058, 2024. 6

[6] Edouard Belval. pdf2image, 2017. 5

[7] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer, 2024. 8

[8] Ali Furkan Biten, Andres Mafla, Lluis Gomez, Valveny C V Jawahar, and Dimosthenis Karatzas. Scene Text Visual Question Answering. In *ICCV*, 2019. 5

[9] Tsachi Blau, Sharon Fogel, Roi Ronen, Alona Golts, Roy Ganz, Elad Ben Avraham, Aviad Aberdam, Shahar Tsiper, and Ron Litman. Gram: Global reasoning for multi-page vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2024. 7

[10] Łukasz Borchmann, Michał Pietruszka, Wojciech Jaśkowski, Dawid Jurkiewicz, Piotr Halama, Paweł Józiak, Łukasz Garncarek, Paweł Liskowski, Karolina Szyndler, Andrzej Gretkowski, Julita Ołtusek, Gabriela Nowakowska, Artur Zawłocki, Łukasz Duhr, Paweł Dyda, and Michał Turski. Arctic-TILT. Business Document Understanding at Sub-Billion Scale, 2024. 7

[11] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022. 8

[12] 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, 8

[13] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. 7

[14] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024. 5

[15] Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. Pdfvqa: A new dataset for real-world vqa on pdf documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 585–601. Springer, 2023. 8

[16] Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. Mmvqa: A comprehensive dataset for investigating multipage multimodal information retrieval in pdf-based visual question answering. In *the 33rd International Joint Conference on Artificial Intelligence*, 2024. 1, 2, 8

[17] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-Xcomposer2-4KHD: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint*, abs/2404.06512, 2024. 6

[18] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2024. 5

[19] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient Document Retrieval with Vision Language Models, 2024. 1, 4, 5

[20] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. 8

[21] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 8

[22] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. In *ICML*, 2020. 8

[23] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding, 2024. 6, 8

[24] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified

text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 4083–4091, New York, NY, USA, 2022. Association for Computing Machinery. 8

[25] Gautier Izacard and Edouard Grave. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *EACL*, 2021. 8

[26] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 6

[27] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. 5

[28] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011. 7

[29] Vladimir Karpukhin, O Barlas, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP*, pages 6769–6781, 2020. 8

[30] Omar Khattab and Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48, 2020. 4, 7

[31] 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 (ECCV)*, 2022. 8

[32] Jordy Van Landeghem, Rafał Powalski, Rubèn Tito, Dawid Jurkiewicz, Matthew Blaschko, Łukasz Borchmann, Mickaël Coustaty, Sien Moens, Michał Pietruszka, Bertrand Ackaert, Tomasz Stanisławek, Paweł Józiak, and Ernest Valveny. Document Understanding Dataset and Evaluation (DUDE). In *ICCV*, 2023. 1, 2, 8

[33] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024. 5, 7, 8

[34] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 5, 6, 7

[35] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 8

[36] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen Tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020. 8

[37] 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*, abs/2311.06607, 2023. 6

[38] Llama Team. The llama 3 herd of models, 2024. 5

[39] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. DeepSeek-VL: towards real-world vision-language understanding. *arXiv preprint*, abs/2403.05525, 2024. 6

[40] Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. Sail: Search-augmented instruction learning, 2023. 8

[41] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *EMNLP*, 2024. 7

[42] 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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 1, 2, 5, 6, 8

[43] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2582–2591, 2021. 8

[44] 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. 1, 2, 8

[45] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3090–3101, 2023. 8

[46] Microsoft. Playwright for python, 2021. 3

[47] OpenAI. Hello gpt-4o, 2024. 5, 8

[48] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chana, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Workshop*, 2017. 5

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32 (NeurIPS), 2019. 5

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7

[51] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 3715–3734, 2022. 4, 5, 7

[52] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. 8

[53] Sivic and Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, 2003. 2, 5, 7

[54] Ray Smith. An overview of the tesseract ocr engine. In *ICDAR*, 2007. 5

[55] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021. 1, 3, 5

[56] 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, 2023. 8

[57] The Chromium Project Authors. The chromium projects, 2024. 3

[58] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023. 1, 2, 5, 8

[59] 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, 2023. 8

[60] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 5, 7, 8

[61] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. HuggingFace's Transformers: State-of-the-art Natural Language Processing. In *EMNLP*, 2020. 5

[62] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. LLaVA-UHD: An LMM perceiving any aspect ratio and high-resolution images. *arXiv preprint*, abs/2403.11703, 2024. 6

[63] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-Augmented Multimodal Language Modeling. In *ICML*, 2023. 8

[64] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Mingshi Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Feiyan Huang. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *Conference on Empirical Methods in Natural Language Processing*, 2023. 8

[65] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. Visrag: Vision-based retrieval-augmented generation on multi-modality documents, 2025. 7

[66] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint*, abs/2405.17220, 2024. 6

[67] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021. 8

[68] Justin Zobel and Alistair Moffat. Inverted files for text search engines. *ACM Comput. Surv.*, 38(2):6–es, 2006. 2, 5, 7