MMDocIR: Benchmarking Multimodal Retrieval for Long Documents

Anonymous ACL submission

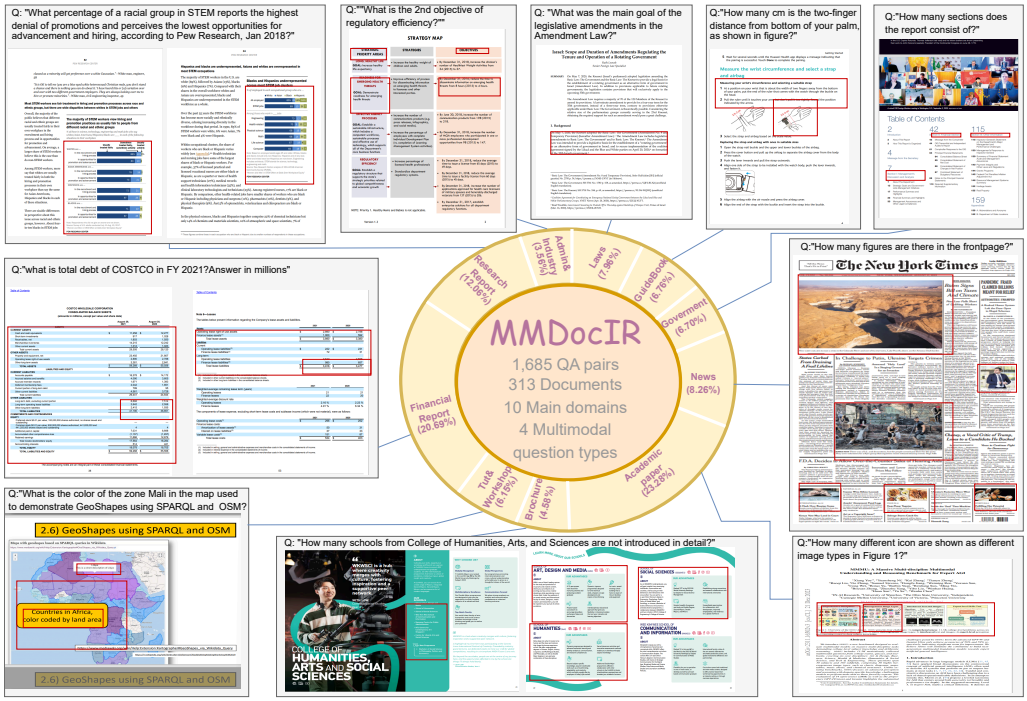


Figure 1: MMDocIR evaluation set comprises 313 long documents and 1,658 queries across 10 domains. For each query, page-level labels are provided via selected screenshots. Red boundary boxes represent layout-level labels.

Abstract

Multimodal document retrieval aims to identify and retrieve various forms of multimodal content, such as figures, tables, charts, and layout information from extensive documents. Despite its increasing popularity, there is a notable lack of a comprehensive and robust benchmark to effectively evaluate the performance of systems in such tasks. To address this gap, this work introduces a new benchmark, named MMDocIR, that encompasses two distinct tasks: page-level and layout-level retrieval. The former evaluates the performance of identifying the most relevant pages within a long document, while the later assesses the ability of detecting specific layouts, providing a more fine-grained measure than whole-page analysis. A layout refers to a variety of elements, including textual paragraphs, equations, figures, tables, or charts. The MMDocIR benchmark comprises a rich dataset featuring 1,685 questions annotated by experts and 173,843 questions with bootstrapped labels, making it a valuable resource in multimodal document retrieval for both training and evaluation. Through rigorous experiments, we demonstrate that (i) visual retrievers significantly outperform their text counterparts, (ii) MMDocIR training set effectively enhances the performance of multimodal document retrieval and (iii) text retrievers leveraging VLM-text significantly outperforms retrievers relying on OCR-text. Our dataset is available at https://anonymous.4open.science/r/MMDocIR.

1 Introduction

Multimodal document retrieval (Hassan et al., 2013; Lee et al., 2024) aims to retrieve information from visually rich documents based on user queries. Unlike traditional document retrieval (Zhang et al., 2022; Chen et al., 2023) which primarily deals with textual data, multimodal document retrieval imposes substantially greater demands on understanding multimodal elements such as images, tables, charts, and layout designs. Such elements often carry significant information that plain text fails to convey (Cui et al., 2021; Sassioui et al., 2023):

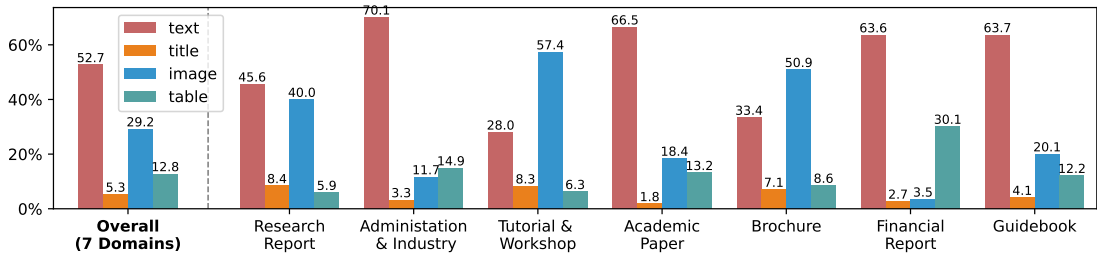


Figure 2: Area ratio of different modalities (1) in overall and (2) by domains in MMLongBench-Doc benchmark.

Benchmarks	**Question**					**Document**			**Label**	
	Type	Expert?	IR?	#Num	Evidence Type	Domain	#Pages	Source	Page	Layout
DocCVQA	VQA question	✓	✓	20	TXT/L	Finance	1.0	✓	✓	✗
SciMMIR	Image caption	✗	✗	530k	TAB/I	Science	1.0	✗	✗	✗
ViDoRe	VQA question	✓	✗	3,810	TXT/C/TAB/I	Multiple	1.0	✗	✓	✗
PDF-MVQA	Search query	✗	✓	260k	TXT/TAB/I	Biomedical	9.6	✓	✓	✓
MMLongBench-Doc	VQA question	✓	✗	1,082	TXT/C/TAB/I	Multiple	47.5	✓	✓	✗
Wiki-SS	Natural question	✗	✓	3,610	TXT	Wikipedia	1.0	✗	✓	✗
DocMatix-IR	VQA question	✗	✗	5.61m	TXT/C/TAB/I	Multiple	4.2	✓	✓	✗
MMDocIR (eval)	VQA question	✓	✓	1,658	TXT/C/TAB/I	Multiple	65.1	✓	✓	✓
MMDocIR (train)	VQA question	✓	✓	73.8k	TXT/C/TAB/I	Multiple	49.3	✓	✓	✓

Table 1: MMDocIR versus existing document IR datasets. **TXT/C/TAB/I** refers to text/chart/table/image.

tables reveal structured data patterns, charts visualize trends or correlations, images offer contextual and semantic cues, etc. Combining these visual elements enriches the quality of retrieved content. Our analysis of MMLongBench-Doc benchmark (Ma et al., 2024b) in Figure 2 shows that: text occupies only 52.7% of content area, while images and tables account for 29.2% and 12.8% respectively. This highlights the need for retrieval systems that effectively handle multimodal information.

However, as shown on (see Table 1), existing benchmarks exhibit several critical limitations that undermine comprehensive evaluation of multimodal retrieval systems. The **key limitations** include: **1. Question Quality:** Many questions used in existing benchmarks are directly sourced from datasets for Visual Question Answering (VQA) tasks. Some questions often assume the input is already relevant, making it not suited for meaningful evaluation of retrieval capabilities. **2. Document Completeness and Diversity**: Existing benchmarks often provide only partial documents, limiting the ability to evaluate within full document context. Additionally, the narrow range of document domains further restricts their applicability across diverse use-cases in real-world. **3. Retrieval Granularity:** Most benchmarks support only page-level retrieval. Such granularity is often insufficient, as user queries frequently target specific elements, such as figures or tables, rather than entire pages.

To address these gaps, we introduce **MMDocIR**, a multimodal document information retrieval benchmark. MMDocIR is designed for **two key** tasks: ***page-level*** and ***layout-level*** retrieval. **(1)** The page-level retrieval identifies the most relevant pages within a document to answer user query. **(2)** The layout-level retrieval targets the most relevant fine-grained elements such as paragraphs, equations, figures, tables, and charts (see Appendix E.2 for more examples). Such task supports more precise and context-aware retrieval that pinpoint specific elements to address user queries. To support both tasks, we develop **MMDocIR evaluation set** that comprises 313 documents, each averaging 65.1 pages, along with 1,658 modified queries derived from MMLongBench-Doc and DocBench (Zou et al., 2024). The queries are annotated with 2,107 page-level and 2,638 layout-level labels. The page labels are specific pages that contain the evidence needed to answer the query.[1] The layout labels consist of precisely drawn bounding boxes around the key evidence within the identified pages. In addition, we introduce the **MMDocIR training set**, designed to support retriever training. It contains 73,843 questions sourced from 7 DocQA datasets. To construct this set, we manually collect 6,878 documents and apply a semi-automatic pipeline to annotate the ground truth labels.

By leveraging MMDocIR, we conduct a comprehensive evaluation on multimodal document retrieval across two retriever types: visual-driven and text-driven. ***Visual-driven retrievers*** (Ma et al., 2024a; Faysse et al., 2024), leverage vision lan-

[1]While MMLongBench-Doc provided initial page labels, our meticulous review lead to corrections in 21.3% of them.

guage models (VLMs) to capture rich multimodal cues and generate embeddings for both queries and documents. In contrast, ***text-driven retrievers*** (Karpukhin et al., 2020; Khattab and Zaharia, 2020; Xiao et al., 2023) rely on OCR or VLM to first convert the multimodal content into text, subsequently employing language models (LMs) to generate embeddings for both queries and documents. Our extensive experiments reveal that visual-driven retrievers consistently outperform their text-driven counterparts, often by a significant margin. In summary, our contributions are threefold:

- **Dual-task Retrieval Framework**: We propose a dual-task retrieval framework (§ 2) that supports page-level and fine-grained layout-level multimodal document retrieval.
- **MMDocIR Benchmark**: We introduce Multimodal Document Information Retrieval benchmark. The evaluation set (§ 3) consists of 313 documents with expert-annotated labels for 1,658 questions. The training set (§ 4) consists of 6,878 documents and labels for 73,843 questions.
- We conduct extensive experiments and comparisons of both text and visual retrievers (§ 5), demonstrating clear advantage of incorporating visual content in multimodal retrieval tasks.

2 Dual-Task Retrieval Definition

Let $\mathcal{D}$ be a document corpora consisting of document pages: $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, and layouts: $\mathcal{L} = \{l_1, l_2, \ldots, l_m\}$ extracted via layout detection. The objective is to perform document retrieval at both page-level and layout-level. Specifically, given query Q, the task is to retrieve the top k pages and layouts most relevant to Q, where $k << n$ and $k << m$. The relevance of pages (p) and layouts (l) to Q is measured by similarity scores, $\text{Sim}(Q, p)$ and $\text{Sim}(Q, l)$ respectively. The retrieval system consists of two phases: (1) an offline indexing phase, where pages and layouts from $\mathcal{P}$ and $\mathcal{L}$ are encoded into vectors, and (2) an online querying phase, in which a query Q is encoded into a vector, which is then compared against the offline-indexed vectors using similarity scores $\text{Sim}(Q, p)$ for pages and $\text{Sim}(Q, l)$ for layouts.

3 MMDocIR: Evaluation Set

3.1 Document Corpora Collection

After a comprehensive review of existing DocVQA datasets, we select MMLongBench-Doc (Ma et al., 2024b) and DocBench (Zou et al., 2024) to facilitate our benchmark construction (see Appendix B.2 for our selection criteria). MMLongBench-Doc is a long-context, multimodal benchmark comprising 1,091 questions across 135 documents with 47.5 pages on average. DocBench emphasizes long document understanding, consisting of 1,102 questions across 229 documents, each with an average length of 77.5 pages. Both datasets offer corpora from diverse domains with expert-annotated questions that require evidence from various modalities. Consequently, we curate a set of 364 documents and 2,193 questions for our subsequent annotation.

3.2 Annotation Process

Question Filtering and Revision. To ensure that the questions in MMDocIR are optimally suited for document retrieval tasks, we identify four specific types of questions (see Appendix B.3) that do not align well with the objectives of IR. By filtering and refining these questions, we ensure the integrity and relevance of MMDocIR, resulting in 1,658 questions for subsequent annotation.

Page-level Annotation. We annotate page labels that precisely identify the exact pages containing ground truth evidence. Given that documents in MMDocIR contain 65.1 pages on average, pinpointing relevant pages is highly non-trivial, akin to finding a needle in haystack, which demands careful inspection and document understanding.

- For **DocBench**: we manually annotate page labels for all 864 questions from scratch, by carefully reviewing each document and locating the pages containing answer evidence.
- For **MMLongBench-Doc**: we rigorously review and validate the answers and page labels of 794 questions. This effort results in corrections to 10 answers and 169 page labels[2].

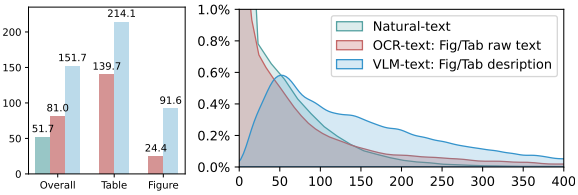
Layout-level Annotation. To enhance the granularity of our benchmark, we extend our annotations to include layout-level labels, identifying specific layout elements as evidence. Compared to page annotation, layout-level labeling is significantly more complex and labor-intensive.

- **Layout Detection**. We begin by utilizing MinerU (Wang et al., 2024) to automatically parse all documents and detect all layouts (*e.g.,* layout type and bounding boxes).

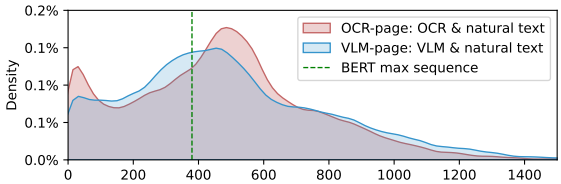
[2]Common errors in page labeling: annotators starting page indexing at 1 rather than 0, missing labels for questions spanning multiple pages, and incorrect or absent page labels.

Consistency	Page Labels			Layout Labels		
	Prec.	Recall	F1	Prec.	Recall	F1
A←B	95.7	96.1	95.9	88.1	86.8	87.4
B←A	94.3	94.6	94.4	85.9	87.5	86.7
Average	95.0	95.4	95.2	87.0	87.2	87.1

Table 2: Annotation consistency between group A & B.



(a) #words/layout (b) Distribution density: #words/layout



(c) Distribution density: #words/page

Figure 3: Distribution of OCR/VLM-text by length.

- **Evidence Identification**. We identify the layouts that contain necessary answer evidence. In case where MinerU fails to detect evidentiary element, we manually annotate the the bounding boxes, accounting for 7% of the total layout-level labels.

3.3 Quality Control

To ensure annotation quality and reliability in MMDocIR, we have adopted a rigorous 3-stage quality control process. We split questions into two parts. Each group is responsible for annotating approximately 1,000 questions, with an overlap of 400 questions serving the need for cross-validation.

- **Overlap Scoring**: For the 400 overlapping questions, A←B evaluates A's labels with B's labels as ground truth, and vice versa for B←A.
- **Cross-Evaluation**: We cross-evaluate and achieve F_1 score of 95.2 and 87.1 for page and layout labels, as shown in Table 2. We then identify and fix the discrepancies.
- **Random Cross-Validation**: We randomly cross-validate 50% of the remaining annotations. In the cases where we have different opinions, we discuss to achieve mutually-agreed annotations.

3.4 Multimodal content as OCR/VLM-text

To apply multimodal retrieval to text retrievers, we convert multimodal layouts (*e.g.,* tables or figures) into text. Specifically, we extract text using OCR (Smith, 2007) ("OCR-text") and generate detailed descriptions using VLMs (OpenAI, 2024; Qwen-Team, 2024) ("VLM-text"). As a result, each image layout is represented in three formats: original image, OCR-text, and VLM-text.

For **layouts**, the average word length and distribution of OCR-text and VLM-text of MMDocIR are shown in Figure 3a and 3b. Notably, the length of VLM-text is 1.5 and 3.8 times of OCR-text for table (with more structured numbers) and figure (mostly with visual elaboration), respectively.

For **pages**, we construct two variants by combining the natural text with either OCR-text and VLM-text for each page, resulting in OCR-page and VLM-page representations. The average word length is 477 and 505 for OCR-page and VLM-page respectively, with their distribution shown in Figure 3c.

3.5 Statistics and Analysis

Document Analysis. As shown in Table 3, MMDocIR evaluation set includes 313 long documents, averaging 65.1 pages, categorized into 10 domains. Different domains feature distinct multimodal distribution. The overall modality distribution is as follows: text (60.4%), image (18.8%), table (16.7%), and others (4.1%), with fine-grained distribution shown in Figure 4a.

Question and Annotation Analysis. MMDocIR includes 1,658 questions, and 2,107 page and 2,638 layout labels. The evidence spans 4 modalities: text (44.7%), image (21.7%), table (37.4%), and layout/meta (11.5%). Notably, MMDocIR presents several challenges: 254 questions require cross-modal understanding, 313 questions require evidence across multiple pages, and 637 questions require reasoning over multiple layouts.

4 MMDocIR: Training Set

4.1 Document Corpus Collection

After screening related DocVQA datasets, we collect our training set corpora from 7 datasets, namely MP-DocVQA (Tito et al., 2023), SlideVQA (Tanaka et al., 2023), TAT-DQA (Zhu et al., 2022), SciQAG (Wan et al., 2024), DUDE (Landeghem et al., 2023), and CUAD (Hendrycks et al., 2021). Since most of these datasets do not provide original document, we invest significant efforts in tracing and recovering the original document, as detailed in Appendix B.4.

4.2 Label Construction and statistics

We use semi-automated construction pipeline to generate page-level and layout-level labels for

MMDocIR	Document Statistics							Questions (%)				Modality Distribution %			
	#Doc	#QA	#Page Label	#Lay Label	#Page /Doc	#Lay /Page	%Lay	Text	Image	Table	Lay/ Meta	Text	Image	Table	Title
Eval Domains	313	1,658	2,107	2,638	65.1	8.3	41.8	44.7	21.7	37.4	11.5	60.4	18.8	16.7	4.1
- Research Report	34	200	318	400	39.4	6.0	39.1	45.0	17.5	74.5	13.5	45.6	40.0	5.9	8.4
- Admin & Industry	10	59	82	113	16.8	9.1	45.1	78.0	20.3	13.5	13.5	70.1	11.7	14.9	3.2
- Tut & Workshop	17	102	165	225	57.5	4.1	43.8	37.2	61.7	24.5	9.8	28.0	57.3	6.3	8.3
- Academic Paper	75	386	473	571	19.5	10.1	48.4	28.8	25.7	50.0	10.4	74.6	12.8	11.1	1.5
- Brochure	15	76	121	178	30.3	9.7	41.1	60.5	52.6	18.4	36.8	33.3	50.8	8.5	7.0
- Financial Report	51	343	394	477	169.5	9.2	44.8	28.0	13.1	54.5	5.3	60.3	7.9	29.2	2.6
- Guidebook	22	112	168	223	78.4	10.0	33.6	51.8	54.4	26.8	17.8	63.7	20.0	12.1	4.1
- Government	44	111	116	132	68.9	6.9	45.4	69.37	2.7	0	7.6	88.2	3.7	5.7	2.4
- Laws	44	132	133	149	58.5	6.0	31.2	62.1	0	10.6	27.3	83.8	1.6	12.3	2.2
- News	1	137	137	170	50.0	73.6	72.3	70.1	1.5	0	28.5	48.5	39.8	0.0	11.6

Table 3: Detailed statistics of MMDocIR evaluation set. "#Lay/Page" is averaging layouts per page, reflecting layout complexity. "%Lay" is the area ratio of useful layouts (excluding white spaces, headers, and footers).

MMDocIR	Domain	Document Statistics					Evidence Modality (%)				Labels	
		#Doc	#QA	#Page /Doc	#Lay /Page	%Lay	Text	Image	Table	Title	Page	Lay
Train Subsets	assorted docs	6,878	73,843	32.6	6.32	42.6	49.3	34.3	10.8	4.9	✓	✓
- MP-DocVQA	health/ind. docs	875	15,266	46.8	6.9	38.8	57.3	18.0	22.7	1.9	✓	✗
- SlideVQA	diverse slides	2,011	11,066	49.3	4.4	42.3	30.1	56.2	4.7	8.8	✓	✗
- TAT-DQA	annual reports	163	15,814	147.3	9.2	42.2	66.4	4.4	26.5	2.7	✓	✓
- arXivQA	arXiv papers	1,579	12,314	18.4	7.9	50.0	70.4	22.3	2.8	1.0	✓	✓
- SciQAG	science papers	1,197	4,976	9.0	9.1	53.7	61.8	28.0	6.7	1.5	✓	✓
- DUDE	assorted docs	779	3,173	15.6	7.4	42.5	57.1	24.7	15.2	2.9	✓	✓
- CUAD	legal contracts	274	11,234	29.6	7.4	24.7	89.3	2.5	6.4	1.1	✓	✗

Table 4: MMDocIR training set statistics about our collected documents, questions, and constructed labels.

datasets that lack them, referring to Appendix B.5 and B.6 for more details of construction process. Notably, layout annotations are missing from most existing datasets, as we manage to obtain or construct layout-level labels for only 4 datasets. The overall statistics (*e.g.,* document information, modality distribution, domain, etc) of MMDocIR training set are summarized in Table 4.

5 Experiment

5.1 Evaluation Metric

The retriever scores each page or layout in the document based on its relevance to the question, and returns the top k candidates with the highest scores. Recall@k is defined as the proportion of ground truth page/layout evidence successfully retrieved. For layout matching, we calculate recall based on the overlaps between the bounding boxes of retrieved layouts and gold-standard layouts.

5.2 Baseline Models and Setting

We evaluate 6 state-of-the-art text retrievers: namely DPR, ColBERT, BGE, E5, Contriever, and GTE (see Appendix D.1). Additionally, we evaluate 5 VLM-based retrievers: 3 off-the-shelf models, namely $\text{DSE}_{\text{wiki-ss}}$, $\text{DSE}_{\text{docmatix}}$, and ColPali (see Appendix D.2), and 2 models trained using MMDocIR training set (see Appendix C). Among all retrievers, ColBERT, ColPali, and $\text{Col-Phi3}_{\text{ours}}$ represent query/document as a list of token-level embeddings, while the other retrievers represent query/document as a single dense embedding. All retrievers are adapted to a dual-task setting:

- **Page Retrieval**: For text retrievers, we use the text from OCR-page or VLM page as described in Section 3.4. For visual retrievers, we directly utilize document page screenshots.
- **Layout Retrieval**: Text retrievers process multimodal layouts using OCR or VLM text (see Section 3.4). Visual retrievers process textual layouts using either **Image** input (cropped image of textual area) or **Hybrid** input (original text, as VLM can directly encode text).

5.3 Main Results for Page-level Retrieval

Table 5 presents the main results for page-level retrieval. Our key findings are as follows:

- **Superiority of Visual Retrievers**: Visual retrievers consistently outperform text retrievers across various domains and retrieval metrics, highlighting the advantage of using screenshots to retain multimodal cues often lost in text conversion.
- **Effectiveness of MMDocIR**: The visual retrievers trained on the MMDocIR training set demonstrate superior performance, demonstrating the

		Method \ Domain	Resear. Report	Admin &Indu.	Tutori.& Worksh.	Acade. Paper	Broch-ure	Finance Report	Guide-book	Govern-ment	Laws	News	Average Macro	Average Micro
Recall@$k=1$	VLM-text	DPR	32.3	25.5	27.0	31.0	28.4	18.8	23.5	31.2	38.3	16.1	27.2	26.9
		ColBERT	48.6	42.8	51.1	46.2	36.0	36.8	49.6	60.9	59.5	26.3	45.8	44.9
		BGE	48.8	30.9	47.1	40.8	37.6	28.4	43.4	51.9	48.9	28.5	40.6	39.6
		E5	48.1	30.0	50.4	39.4	41.1	29.7	40.9	52.8	51.1	24.1	40.8	39.5
		Contriever	45.5	31.2	49.8	41.5	39.4	29.4	45.2	55.3	51.1	20.4	40.9	39.7
		GTE	46.5	26.3	48.7	38.9	35.9	27.0	46.2	50.1	45.8	24.1	38.9	37.9
	Image	$\text{DSE}_{\text{wiki-ss}}$	53.0	50.0	54.0	48.7	45.1	43.0	51.5	46.9	54.2	33.6	48.0	47.5
		$\text{DSE}_{\text{docmatix}}$	52.3	40.4	56.1	51.7	45.8	43.5	53.8	53.7	58.3	46.7	50.2	50.1
		ColPali	56.0	**51.8**	**58.6**	55.9	52.0	47.2	57.9	53.9	64.0	32.8	53.0	52.7
		DPR-Phi3$_{\text{ours}}$	**58.9**	50.4	57.4	59.0	**57.3**	44.6	**63.8**	50.5	**64.4**	35.0	54.1	53.7
		Col-Phi3$_{\text{ours}}$	56.7	50.4	56.9	**61.3**	54.8	**50.7**	60.8	**61.3**	63.6	**54.0**	**57.0**	**57.1**
Recall@$k=3$	VLM-text	DPR	52.2	44.2	43.5	54.6	52.0	35.1	44.4	53.9	57.2	25.5	46.3	46.2
		ColBERT	70.1	64.4	70.3	72.3	59.1	55.3	71.1	**81.3**	70.8	34.3	64.9	64.8
		BGE	71.5	48.2	68.8	65.7	56.2	46.5	66.1	69.9	72.0	32.1	59.7	59.6
		E5	68.4	45.7	68.1	63.7	60.1	44.0	69.3	72.3	78.8	32.8	60.3	59.3
		Contriever	69.4	55.3	68.3	64.9	56.9	46.2	69.9	71.1	72.0	32.1	60.6	59.7
		GTE	71.1	44.5	67.2	64.4	54.3	43.0	70.6	71.9	68.2	31.4	58.7	58.3
	Image	$\text{DSE}_{\text{wiki-ss}}$	75.4	65.0	73.9	79.8	69.5	63.5	75.4	71.5	81.4	50.4	70.6	71.4
		$\text{DSE}_{\text{docmatix}}$	75.4	67.5	73.3	80.0	66.3	61.6	72.8	76.4	82.6	57.7	71.4	71.8
		ColPali	77.6	71.8	**79.4**	83.4	**72.6**	66.1	**80.0**	80.4	**86.4**	49.6	74.7	75.0
		DPR-Phi3$_{\text{ours}}$	**80.3**	66.5	77.6	83.9	71.9	63.8	79.8	71.4	84.5	55.5	73.5	74.3
		Col-Phi3$_{\text{ours}}$	80.2	**74.1**	77.4	**84.8**	69.1	**67.7**	78.7	79.5	81.8	**69.3**	**76.3**	**76.8**
Recall@$k=5$	VLM-text	DPR	66.5	60.1	56.0	68.9	58.8	43.8	57.1	68.6	64.8	33.6	57.8	57.8
		ColBERT	78.8	74.0	78.7	82.3	66.1	60.8	77.0	**88.5**	78.0	38.7	72.3	72.3
		BGE	79.5	65.8	71.3	76.8	62.4	56.0	77.2	77.4	79.5	38.0	68.4	68.5
		E5	76.9	64.2	75.3	74.4	67.4	52.0	78.5	78.6	82.6	40.9	69.1	67.9
		Contriever	77.2	67.1	76.7	75.2	65.1	53.7	75.4	79.2	83.3	39.4	69.2	68.3
		GTE	77.4	62.6	74.7	75.8	62.0	51.8	77.8	80.0	75.0	39.4	67.6	67.2
	Image	$\text{DSE}_{\text{wiki-ss}}$	84.0	**80.2**	78.7	87.0	75.7	73.0	82.0	77.3	88.3	58.4	78.5	79.2
		$\text{DSE}_{\text{docmatix}}$	82.1	77.2	79.6	87.8	73.9	72.4	81.7	83.1	89.4	67.9	79.5	80.1
		ColPali	84.6	79.3	82.3	89.0	79.8	72.1	86.7	84.9	**92.4**	56.9	80.8	81.0
		DPR-Phi3$_{\text{ours}}$	**86.9**	76.2	**85.3**	91.9	**80.0**	71.2	**87.1**	79.5	92.0	61.3	81.1	81.8
		Col-Phi3$_{\text{ours}}$	86.3	78.8	81.2	**92.4**	79.0	**73.8**	85.3	**85.1**	87.1	**73.0**	**82.2**	**83.0**

Table 5: Main results for page-level retrieval, with the best results in **boldface** and second best results underlined. For clarity, we omit results using VLM-text (Refer to Table 10 for full results).

value of high-quality training data.

- **Necessity of Token-level Embeddings**: Compared to dense-level retrievers (*e.g.,* BGE, DSE, DPR-Phi3$_{\text{ours}}$), token-level retrievers (*e.g.,* ColBERT, ColPali, Col-Phi3$_{\text{ours}}$) achieve more advantageous results in Recall@1 and have marginal performance improvement in Recall@3/5. However, token-level embedding can incur storage costs of 10 times more than a single embedding (DSE requires 0.24GB for indexing MMDOCIR while ColPali requires 10.0GB).
- **Top 5 Coverage**: Retrieving top 5 pages provides substantial coverage of relevant information.

5.4 Main Results for Layout-level Retrieval

Table 6 shows the main results for layout-level retrieval. Our key findings are as follows:

- **Effectiveness of VLM-Text**: Interestingly, VLM-text approaches perform comparably to visual retrievers, demonstrating the promising image description capabilities of state-of-the-art VLM. This greatly benefits textual retrievers in multimodal understanding.
- **Necessity of Token-level Embeddings**: For layout retrieval tasks, token-level retrievers marginally outperform dense-level retrievers, demonstrating its importance of in such task.
- **Top 10 Coverage**: For layout retrieval tasks, retrieving top 10 layouts does not guarantee comprehensive coverage of the ground truth layout labels, emphasizing the complexity of the tasks.

5.5 Text Retrieval: OCR-text vs VLM-Text

Text retrievers leveraging VLM-text significantly outperform those using OCR-text in both tasks. Based on results, OCR-text is insufficient for multimodal retrieval, while VLM-text retains richer multimodal information. Although VLM-text offers much more comprehensive text information than OCR-text, it also introduces higher computational overhead and longer inference time.

Most text retrievers based on on BERT (Devlin et al., 2019), truncate input that exceed 512 tokens (approximately 380 english words). As shown in Figure 3c, there are many pages containing more than 380 words (62.9% for OCR-page and 61.1% for VLM-page). Those pages suffer from critical information loss during page retrieval if the ground truth evidence is in the truncated part. In contrast, only a small fraction of layouts contain more than

		Method \ Domain	Resear. Report	Admin &Indu.	Tutori.& Worksh.	Acade. Paper	Broch-ure	Finance Report	Guide-book	Govern-ment	Laws	News	Average Macro	Average Micro
Recall@$k=1$	VLM-text	DPR	11.6	9.5	19.2	19.2	14.9	15.9	15.8	25.6	34.7	27.0	19.3	19.2
		ColBERT	22.0	14.9	28.0	28.3	17.9	29.7	21.1	**52.6**	**54.5**	**44.5**	31.3	31.4
		BGE	19.2	15.2	24.6	28.7	12.8	27.6	19.7	47.0	52.3	35.8	28.3	29.0
		E5	15.9	8.8	27.7	24.3	14.6	21.8	14.7	45.6	53.0	40.5	26.7	26.4
		Contriever	**23.4**	7.5	28.2	26.8	17.1	25.7	16.1	43.6	51.5	42.9	28.3	28.9
		GTE	17.5	10.5	23.0	27.2	14.5	26.3	14.4	39.8	49.2	38.3	26.1	27.1
	Image	$DSE_{wiki-ss}$	20.6	15.1	31.0	31.1	20.1	29.2	22.0	39.3	37.5	35.8	28.2	29.2
		$DSE_{docmatix}$	19.9	11.4	31.5	30.1	17.8	30.0	20.8	46.5	39.4	31.4	27.9	29.1
		ColPali	22.5	21.3	36.6	30.9	26.8	**32.1**	19.3	52.5	51.8	33.6	**32.7**	**32.5**
		DPR-Phi3$_{ours}$	21.1	**22.1**	36.8	**35.2**	25.6	28.7	**24.1**	38.3	35.4	27.4	29.5	30.2
		Col-Phi3$_{ours}$	22.6	22.0	**37.5**	34.9	**28.9**	30.3	22.7	50.2	45.1	26.3	31.1	31.6
Recall@$k=5$	VLM-text	DPR	31.0	25.7	36.7	44.9	33.0	34.1	34.9	49.9	56.3	51.1	39.8	40.4
		ColBERT	41.8	37.7	53.7	61.8	35.1	52.4	46.1	**83.2**	70.1	62.5	**54.4**	**56.0**
		BGE	41.0	28.1	52.7	59.2	36.7	46.0	**50.7**	72.0	71.5	59.9	51.8	53.2
		E5	35.4	28.1	51.7	58.5	33.2	41.2	40.2	79.7	**77.9**	**64.6**	51.1	51.8
		Contriever	40.2	29.2	54.1	57.9	36.8	47.1	44.6	68.8	76.2	61.9	51.7	53.0
		GTE	36.7	25.6	51.2	56.6	39.7	46.6	46.4	72.2	74.3	63.2	51.3	52.3
	Image	$DSE_{wiki-ss}$	42.4	32.9	56.3	58.5	39.8	50.6	41.6	68.6	60.9	50.0	50.2	52.1
		$DSE_{docmatix}$	39.6	36.2	53.9	57.5	33.7	52.5	42.8	69.4	63.1	48.9	49.8	51.9
		ColPali	40.7	**45.9**	54.9	58.5	42.6	51.2	45.7	76.8	74.5	48.9	54.0	54.3
		DPR-Phi3$_{ours}$	45.5	37.7	**57.0**	**62.9**	41.4	51.1	45.5	65.1	60.8	49.3	51.6	53.9
		Col-Phi3$_{ours}$	**46.4**	38.2	53.1	61.8	**45.0**	**54.6**	45.7	68.8	65.7	43.8	52.3	54.5
Recall@$k=10$	VLM-text	DPR	42.2	33.1	52.1	56.2	39.9	43.5	44.0	62.8	61.7	59.7	49.5	50.5
		ColBERT	51.0	48.7	60.6	69.8	43.9	**61.6**	53.7	**88.4**	74.8	66.4	61.9	**63.7**
		BGE	51.1	38.7	62.1	71.5	41.9	55.6	**58.7**	80.8	78.7	63.5	60.3	62.4
		E5	45.3	38.6	62.0	70.5	45.6	50.0	55.3	87.1	82.4	66.8	60.4	61.2
		Contriever	49.9	41.3	62.0	70.5	44.8	56.5	54.5	81.3	78.0	64.9	60.4	62.2
		GTE	48.6	41.1	61.5	68.8	44.3	56.9	58.0	83.0	77.5	**66.9**	60.7	62.2
	Image	$DSE_{wiki-ss}$	55.9	41.3	61.5	68.1	47.8	60.7	54.2	72.9	68.3	54.4	58.5	61.1
		$DSE_{docmatix}$	53.7	43.3	59.6	66.5	44.7	59.1	50.3	75.4	69.2	53.7	57.5	59.9
		ColPali	53.6	**54.1**	64.4	69.5	48.8	60.7	54.0	81.9	**82.5**	50.4	**62.0**	63.2
		DPR-Phi3$_{ours}$	**58.1**	49.1	**67.0**	**74.7**	48.4	57.9	57.8	68.7	66.2	54.4	60.2	62.8
		Col-Phi3$_{ours}$	57.7	50.5	66.6	72.3	**50.7**	59.3	53.6	68.5	74.8	57.5	61.1	63.3

Table 6: Main results for layout-level retrieval (Refer to Table 11 for full results with VLM-text and Hybrid inputs).

	Method	Page recall OCR	Page recall VLM	Page recall Δ	Layout recall OCR	Layout recall VLM	Layout recall Δ
$k=1$	DPR	22.3	27.2	+4.9	12.6	19.3	+6.7
	ColBERT	40.3	45.8	+5.5	19.8	31.3	+11.5
	BGE	35.7	40.6	+4.9	19.0	28.3	+9.3
	E5	35.0	40.8	+5.8	18.4	26.7	+8.3
	Contriever	35.3	40.9	+5.6	18.8	28.3	+9.5
	GTE	35.4	38.9	+3.5	18.2	26.1	+7.9
$k=3$ or 5	DPR	39.4	46.3	+6.9	23.7	39.8	+16.1
	ColBERT	58.8	64.9	+6.1	33.2	54.4	+21.2
	BGE	55.4	59.7	+4.3	32.7	51.8	+19.1
	E5	54.8	60.3	+5.5	33.3	51.1	+17.8
	Contriever	54.9	60.6	+5.7	31.7	51.7	+20.0
	GTE	54.9	58.7	+3.8	33.5	51.3	+17.8
$k=5$ or 10	DPR	49.0	57.8	+8.8	29.9	49.5	+19.6
	ColBERT	66.0	72.3	+6.3	37.6	61.9	+24.3
	BGE	62.7	68.4	+5.7	37.8	60.3	+22.5
	E5	64.1	69.1	+5.0	39.0	60.4	+21.4
	Contriever	63.1	69.2	+6.1	37.3	60.4	+23.1
	GTE	63.2	67.6	+4.4	40.9	60.7	+19.8

Table 7: Results of text retrievers using OCR/VLM-text.

	Method	Layout recall Hybrid	Layout recall Image	Layout recall Δ
$k=1$	$DSE_{wiki-ss}$	24.6	28.2	+3.6
	$DSE_{docmatix}$	27.5	27.9	+0.4
	ColPali	28.5	32.7	+4.2
	DPR-Phi3$_{ours}$	28.9	29.5	+0.6
	Col-Phi3$_{ours}$	29.8	31.1	+1.3
$k=5$	$DSE_{wiki-ss}$	46.7	50.2	+3.5
	$DSE_{docmatix}$	48.2	49.8	+1.6
	ColPali	52.2	54.0	+1.8
	DPR-Phi3$_{ours}$	50.1	51.6	+1.5
	Col-Phi3$_{ours}$	50.0	52.3	+2.3
$k=10$	$DSE_{wiki-ss}$	55.8	58.5	+2.7
	$DSE_{docmatix}$	57.4	57.5	+0.1
	ColPali	60.0	62.0	+2.0
	DPR-Phi3$_{ours}$	55.5	60.2	+4.7
	Col-Phi3$_{ours}$	58.7	61.1	+2.4

Table 8: Results of visual retrievers: image vs hybrid.

380 tokens (3.9% for OCR-text, 4.8% for VLM-text, 0.5% for natural-text). Hence, as reflected in Table 5 and 6, text retriever demonstrates stronger performance on layout-level retrieval than on page-level retrieval.

5.6 Visual Retrieval: Image vs Hybrid input

Visual retrievers tend to perform better when encoding text as images via visual encoders, rather than processing native textual input with LLM backbones. This advantage largely stems from their training setup: visual retrievers are typically optimized using text queries paired with image-based passages or documents, but are not fine-tuned directly on purely textual passages. However, encoding text as images incurs substantial computational overhead. Representing text as image tokens requires significantly more resources than native text encoding. To address this inefficiency and promote balanced retrieval capabilities, we advocate for future visual retrievers to be jointly trained or fine-tuned on both text and visual retrieval tasks.

		Model	Store (GB)	Index (MM:SS)	Search (MM:SS)
Page Processing	Text	DPR	**0.06**	**6:53**	**00:02**
		ColBERT	3.45	14:12	00:04
		BGE	0.08	7:31	00:03
		E5	0.08	8:26	00:03
	Image	DPR-Phi3	0.24	101:20	00:04
		ColPali	10.00	47:14	00:05
		Col-Phi3	24.56	106:23	00:07
Layout Processing	Text	DPR	**0.51**	**41:33**	**00:15**
		ColBERT	26.72	94:56	02:25
		BGE	0.66	55:05	00:18
		E5	0.66	60:21	00:18
	Image	DPR-Phi3	1.99	735:51	00:44
		ColPali	83.50	262:29	09:06
		Col-Phi3	204.32	784:07	10:56
	Hybrid	DPR-Phi3	1.84	130:38	01:09
		ColPali	12.06	73:50	04:24
		Col-Phi3	22.72	140:44	02:41

Table 9: Efficiency analysis different retrievers.

Such hybrid training would enable models to efficiently process text when appropriate, without compromising performance on visual inputs.

5.7 Efficiency Analysis

We evaluate the inference efficiency by measuring three key metrics : storage consumption, indexing time and, retrieval latency, as shown in Table 9. Experiments are conducted on a NVIDIA A100 40GB GPU with batch size of 4 for image and 256 for text. DPR-styled retrievers which generate single vector embeddings, demonstrates higher efficiency and lower computation across all metrics, compared to ColBERT-styled retrievers that produce token-level embeddings. Although DPR-styled retrievers slightly underperform in retrieval accuracy, their smaller embeddings size provide a significant advantage in the inference stage when storage space and inference time are concerned.

Another key finding is that textual inputs are significantly more efficient than the visual inputs across all metrics. Meanwhile, hybrid retrieval system, which processes text in the image through LLM rather than visual encoders, further reduces memory and time consumption. Hence, future works on training hybrid retrieval system are encouraged as it offers a strong balance between computational efficiency and retrieval performance.

6 Related Work

DocCVQA (Tito et al., 2021) proposes extracting information from a document image collection. However, it is limited by its small question set (20 questions). While **PDF-MVQA** (Ding et al., 2024) is tailored for multimodal retrieval in biomedical articles, it is annotated by GPT-3.5-turbo rather than experts. **SciMMIR** (Wu et al., 2024) also investigates multimodal retrieval but only provides image-caption pairs, lacking user queries paired with the corresponding document pages. Ma et al. (2024a) introduces two relevant datasets, namely Wiki-SS and DocMatix-IR. **Wiki-SS** is derived from natural questions (Kwiatkowski et al., 2019) , wherein evidence passages are screenshots of Wikipedia pages. However, natural questions are primarily designed for text retrieval, and the provided screenshots may not consistently capture the ground-truth evidence, as only the front page is considered. **DocMatix-IR** is constructed from the large-scale DocMatix (Laurençon et al., 2024) dataset using filtering and hard negative mining. However, the questions are generated by `Phi-3-small` (Abdin et al., 2024) rather than human experts, and are not de-contextualized for retrieval task. **ViDoRe** (Faysse et al., 2024) is the most relevant benchmark to MMDocIR. It integrates multiple DocVQA datasets and provides new industrial documents. Upon a thorough examination of the 2,400 questions, we find that over 80% questions exhibit notable limitations in terms of their complexity, contextual clarity, and the absence of complete document corpora. Refer to Appendix F for detailed quantitative and qualitative analysis of ViDoRe.

7 Conclusion

In conclusion, multimodal document retrieval presents a complex challenge that requires both understanding and integrating diverse data modalities beyond plain text. To more effectively evaluate these capabilities, we introduce the MMDocIR benchmark, which features the innovative dual-task retrieval capabilities targeting page-level and layout-level document granularity. The MMDocIR includes a rich dataset featuring expertly annotated labels for 1,685 questions and bootstrapped labels for 73,843 questions, serving as a valuable resource for both training and evaluation of multimodal document retrieval. Our comprehensive empirical studies show that visual-driven retrievers significantly outperform text-driven ones, underscoring the importance of visual information in improving retrieval performance. Future work can expand upon these findings by optimizing retrieval algorithms to enhance both accuracy and efficiency of multimodal document retrieval systems.

Limitations

The limitations of MMDOCIR are summarized as follows:

- **Incomplete layout label annotations for training set**: For 3 out of 7 training subsets, our semi-automated pipelines could not extract layout labels. These pipelines are optimized for datasets with single text or image layouts and cannot handle complex or cross-modal layouts. Future work should explore leveraging advanced vision-language models (VLMs) to facilitate annotation of layout labels for these subsets.

- **Lack of joint text and visual training**: As demonstrated in Section 5.6, all visual retrievers are suboptimal at modeling text passages, compared to modeling text as image screenshots. Our current visual retrievers do not explicitly utilize text query-document pairs to address this limitation. Future research should consider integrating both text and visual passages for joint training or finetuning to improve performance on both retrieval tasks.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. 2018. Convolutional image captioning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5561–5570, Salt Lake City, UT, USA. Computer Vision Foundation / IEEE Computer Society.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

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 A. Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, and 16 others. 2024. Paligemma: A versatile 3b VLM for transfer. *CoRR*, arXiv:2407.07726.

Eugene Borovikov. 2014. A survey of modern optical character recognition techniques. *CoRR*, arXiv:1412.4183.

Hui Chao and Jian Fan. 2004. Layout and content extraction for PDF documents. In *Document Analysis Systems VI, 6th International Workshop, DAS 2004, Florence, Italy, September 8-10, 2004, Proceedings*, volume 3163 of *Lecture Notes in Computer Science*, pages 213–224, Florence, Italy. Springer.

Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, and Soumya K. Ghosh. 2017. *Optical Character Recognition Systems*, pages 9–41. Springer International Publishing, Cham.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023. Walking down the memory maze: Beyond context limit through interactive reading. *Preprint*, arXiv:2310.05029.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024a. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications. *Preprint*, arXiv:2111.08609.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024. MMVQA: A comprehensive dataset for investigating multipage multimodal information retrieval in pdf-based visual question answering. In *Proceedings of the Thirty-Third*

International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024, pages 6243–6251. ijcai.org.

Kuicai Dong, Derrick Goh Xin Deik, Yi Quan Lee, Hao Zhang, Xiangyang Li, Cong Zhang, and Yong Liu. 2024. MC-indexing: Effective long document retrieval via multi-view content-aware indexing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2673–2691, Miami, Florida, USA. Association for Computational Linguistics.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *Preprint*, arXiv:2407.01449.

Ehtesham Hassan, Santanu Chaudhury, and Madan Gopal. 2013. Multi-modal information integration for document retrieval. In *12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, August 25-28, 2013*, pages 1200–1204, Washington, DC, USA. IEEE Computer Society.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: an expert-annotated NLP dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, virtual. Advances in Neural Information Processing Systems.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Jordy Van Landeghem, Rafal Powalski, Rubèn Tito, Dawid Jurkiewicz, Matthew B. Blaschko, Lukasz Borchmann, Mickaël Coustaty, Sien Moens, Michal Pietruszka, Bertrand Anckaert, Tomasz Stanislawek, Pawel Józiak, and Ernest Valveny. 2023. Document understanding dataset and evaluation (DUDE). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19471–19483, Paris, France. IEEE.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

Jaewoo Lee, Joonho Ko, Jinheon Baek, Soyeong Jeong, and Sung Ju Hwang. 2024. Unified multimodal interleaved document representation for retrieval. *Preprint*, arXiv:2410.02729.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. Association for Computational Linguistics.

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. 2024b. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Preprint*, arXiv:2407.01523.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208, Waikoloa, HI, USA. IEEE.

Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada. 1999. *Optical character recognition*. John Wiley & Sons, Inc., USA.

OpenAI. 2024. Hello gpt-4o: We're announcing gpt-4o, our new flagship model that can reason across audio, vision, and text in real time.

Qwen-Team. 2024. Introducing qwen-vl.

Vatsal Raina and Mark Gales. 2024. Question-based retrieval using atomic units for enterprise rag. *Preprint*, arXiv:2405.12363.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126, Maryland, USA. National Institute of Standards and Technology (NIST).

Gerard Salton, Edward A. Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Commun. ACM*, 26(11):1022–1036.

Abdellatif Sassioui, Rachid Benouini, Yasser El Ouargui, Mohamed El-Kamili, Meriyem Chergui, and Mohammed Ouzzif. 2023. Visually-rich document understanding: Concepts, taxonomy and challenges. In *10th International Conference on Wireless Networks and Mobile Communications, WINCOM 2023, Istanbul, Turkey, October 26-28, 2023*, pages 1–7, Istanbul, Turkey. IEEE.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13636–13645, Washington, DC, USA. AAAI Press.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2021. Document collection visual question answering. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 778–792. Springer.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multi-page docvqa. *Preprint*, arXiv:2212.05935.

Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. 2024. Sciqag: A framework for auto-generated science question answering dataset with fine-grained evaluation. *Preprint*, arXiv:2405.09939.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. Mineru: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. SciMMIR: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12560–12574, Bangkok, Thailand. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2579–2591, Virtual. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training

of text and layout for document image understanding. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1192–1200, CA, USA. ACM.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4651–4659, Las Vegas, NV, USA. IEEE Computer Society.

Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Multi-view document representation learning for open-domain dense retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5990–6000, Dublin, Ireland. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4857–4866, Lisboa Portugal. ACM.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. 2024. Docbench: A benchmark for evaluating llm-based document reading systems. *Preprint*, arXiv:2407.10701.

Appendix Table of Contents

Due to page limits, we provide meaningful descriptions, results, implementation details, case studies, and other supplementary materials in Appendix. Our 21-page appendix is organized as follows:

- **Appendix A**: Supplementary experimental results for page and layout retrieval.
- **Appendix B**: process and implementation for MMDocIR curation.
  - Related DocVQA benchmarks (§B.1)
  - Multimodal corpora selection (§B.2)
  - Question filtering for IR context (§B.3)
  - Collection of training multimodal documents (§B.4)
  - Curation of ground-truth training labels (§B.5)
  - Selection of hard negatives (§B.6)
  - Modality distribution (§B.7)
  - Resource and Artifact URLs (§B.8)
- **Appendix C**: Training visual retrievers with MMDocIR training set.
  - Encoding text queries and images (§C.1)
  - Similarity calculation (§C.2)
  - Training objectives (§C.3)
  - Implementation details (§C.4)
- **Appendix D**: Text and visual retriever details.
  - Text-centric retrieval (§D.1)
  - Vision-driven retrieval (§D.2)
  - Retriever implementation (§D.3)
- **Appendix E**: Examples from MMDocIR.
  - Document pages (10 domains) (§E.1)
  - Document layouts (§E.2)
  - Page and layout annotations (§E.3)
- **Appendix F**: Analyses on ViDoRe benchmark.
  - Search query and new annotation (§F.1)
  - Corpora analysis (§F.2)
- **Appendix G**: License agreements.
- **Appendix H**: Ethical considerations.

A Supplementary Experimental Results

In this section, we provide the full results of page-level and layout-level retrievals, which supplement our main results discussion in Section 5.3 and Section 5.4 respectively.

Specifically, Table 10 extends the main page-level results shown in Table 5 with the results of text retrievers using OCR-text. Table 11 extends the main layout-level results shown in Table 6 with the results of (i) text retrievers using OCR-text and (ii) visual retrievers using hybrid inputs.

	Method \ Domain		Resear. Report	Admin &Indu.	Tutori.& Worksh.	Acade. Paper	Broch-ure	Finance Report	Guide-book	Govern-ment	Laws	News	Average Macro	Average Micro
Recall@$k=1$	OCR-text	DPR	21.2	22.1	27.7	23.3	24.4	16.7	21.1	20.7	31.0	15.1	22.3	21.7
		ColBERT	43.8	39.8	42.4	39.3	39.2	38.7	46.3	50.6	46.1	17.1	40.3	40.0
		BGE	45.5	29.0	41.5	33.6	40.8	32.7	40.0	42.8	36.4	15.1	35.7	35.2
		E5	44.2	30.8	39.9	33.2	33.0	32.3	40.4	41.7	38.9	15.8	35.0	34.7
		Contriever	39.1	33.3	44.0	34.2	43.9	26.4	40.6	39.4	37.0	15.1	35.3	33.6
		GTE	44.6	32.6	45.0	33.2	37.2	31.8	40.0	39.9	35.2	14.5	35.4	34.6
	VLM-text	DPR	32.3	25.5	27.0	31.0	28.4	18.8	23.5	31.2	38.3	16.1	27.2	26.9
		ColBERT	48.6	42.8	51.1	46.2	36.0	36.8	49.6	60.9	59.5	26.3	45.8	44.9
		BGE	48.8	30.9	47.1	40.8	37.6	28.4	43.4	51.9	48.9	28.5	40.6	39.6
		E5	48.1	30.0	50.4	39.4	41.1	29.7	40.9	52.8	51.1	24.1	40.8	39.5
		Contriever	45.5	31.2	49.8	41.5	39.4	29.4	45.2	55.3	51.1	20.4	40.9	39.7
		GTE	46.5	26.3	48.7	38.9	35.9	27.0	46.2	50.1	45.8	24.1	38.9	37.9
	Image	$\text{DSE}_{\text{wiki-ss}}$	53.0	50.0	54.0	48.7	45.1	43.0	51.5	46.9	54.2	33.6	48.0	47.5
		$\text{DSE}_{\text{docmatix}}$	52.3	40.4	56.1	51.7	45.8	43.5	53.8	53.7	58.3	46.7	50.2	50.1
		ColPali	56.0	**51.8**	**58.6**	55.9	52.0	47.2	57.9	53.9	64.0	32.8	53.0	52.7
		DPR-$\text{Phi3}_{\text{ours}}$	**58.9**	50.4	57.4	59.0	**57.3**	44.6	**63.8**	50.5	**64.4**	35.0	54.1	53.7
		Col-$\text{Phi3}_{\text{ours}}$	56.7	50.4	56.9	**61.3**	54.8	**50.7**	60.8	**61.3**	63.6	**54.0**	**57.0**	**57.1**
Recall@$k=3$	OCR-text	DPR	46.1	40.6	38.9	46.7	43.9	32.4	38.4	37.0	50.0	20.4	39.4	39.8
		ColBERT	72.6	59.7	57.8	66.7	60.0	53.7	63.8	68.5	61.4	23.7	58.8	59.5
		BGE	69.8	57.7	56.3	58.6	60.7	48.5	57.9	60.9	62.7	20.4	55.4	55.0
		E5	66.6	48.7	59.0	58.0	60.9	48.8	63.7	61.4	60.8	20.4	54.8	54.6
		Contriever	70.2	55.8	60.4	56.6	62.1	43.0	60.0	56.8	61.4	22.4	54.9	53.6
		GTE	69.2	47.0	58.7	59.5	61.8	46.6	65.5	59.1	61.4	19.7	54.9	54.7
	VLM-text	DPR	52.2	44.2	43.5	54.6	52.0	35.1	44.4	53.9	57.2	25.5	46.3	46.2
		ColBERT	70.1	64.4	70.3	72.3	59.1	55.3	71.1	**81.3**	70.8	34.3	64.9	64.8
		BGE	71.5	48.2	68.8	65.7	56.2	46.5	66.1	69.9	72.0	32.1	59.7	59.6
		E5	68.4	45.7	68.1	63.7	60.1	44.0	69.3	72.3	78.8	32.8	60.3	59.3
		Contriever	69.4	55.3	68.3	64.9	56.9	46.2	69.9	71.1	72.0	32.1	60.6	59.7
		GTE	71.1	44.5	67.2	64.4	54.3	43.0	70.6	71.9	68.2	31.4	58.7	58.3
	Image	$\text{DSE}_{\text{wiki-ss}}$	75.4	65.0	73.9	79.8	69.5	63.5	75.4	71.5	81.4	50.4	70.6	71.4
		$\text{DSE}_{\text{docmatix}}$	75.4	67.5	73.3	80.0	66.3	61.6	72.8	76.4	82.6	57.7	71.4	71.8
		ColPali	77.6	71.8	**79.4**	83.4	**72.6**	66.1	**80.0**	80.4	**86.4**	49.6	74.7	75.0
		DPR-$\text{Phi3}_{\text{ours}}$	**80.3**	66.5	77.6	83.9	71.9	63.8	79.8	71.4	84.5	55.5	73.5	74.3
		Col-$\text{Phi3}_{\text{ours}}$	80.2	**74.1**	77.4	**84.8**	69.1	**67.7**	78.7	79.5	81.8	**69.3**	**76.3**	**76.8**
Recall@$k=5$	OCR-text	DPR	59.5	55.8	43.4	59.1	56.2	41.2	50.7	45.5	56.0	23.0	49.0	49.4
		ColBERT	78.4	71.1	63.3	75.2	68.8	60.5	72.0	72.7	67.5	30.3	66.0	66.5
		BGE	79.3	65.9	62.1	69.7	69.8	56.5	68.0	62.8	66.3	26.3	62.7	62.9
		E5	79.3	62.4	67.0	70.3	71.8	57.6	72.5	67.1	67.5	25.7	64.1	64.2
		Contriever	79.9	62.7	64.7	71.7	71.1	48.8	72.4	65.8	67.5	26.3	63.1	62.5
		GTE	78.3	61.9	67.3	72.3	68.7	55.2	72.6	64.0	67.5	24.3	63.2	63.5
	VLM-text	DPR	66.5	60.1	56.0	68.9	58.8	43.8	57.1	68.6	64.8	33.6	57.8	57.8
		ColBERT	78.8	74.0	78.7	82.3	66.1	60.8	77.0	**88.5**	78.0	38.7	72.3	72.3
		BGE	79.5	65.8	71.3	76.8	62.4	56.0	77.2	77.4	79.5	38.0	68.4	68.5
		E5	76.9	64.2	75.3	74.4	67.4	52.0	78.5	78.6	82.6	40.9	69.1	67.9
		Contriever	77.2	67.1	76.7	75.2	65.1	53.7	75.4	79.2	83.3	39.4	69.2	68.3
		GTE	77.4	62.6	74.7	75.8	62.0	51.8	77.8	80.0	75.0	39.4	67.6	67.2
	Image	$\text{DSE}_{\text{wiki-ss}}$	84.0	**80.2**	78.7	87.0	75.7	73.0	82.0	77.3	88.3	58.4	78.5	79.2
		$\text{DSE}_{\text{docmatix}}$	82.1	77.2	79.6	87.8	73.9	72.4	81.7	83.1	89.4	67.9	79.5	80.1
		ColPali	84.6	79.3	82.3	89.0	79.8	72.1	86.7	84.9	**92.4**	56.9	80.8	81.0
		DPR-$\text{Phi3}_{\text{ours}}$	**86.9**	76.2	**85.3**	91.9	**80.0**	71.2	**87.1**	79.5	92.0	61.3	81.1	81.8
		Col-$\text{Phi3}_{\text{ours}}$	86.3	78.8	81.2	**92.4**	79.0	**73.8**	85.3	**85.1**	87.1	**73.0**	**82.2**	**83.0**

Table 10: Main results for page-level retrieval. "*OCR-text*" and "*VLM-text*" refer to converting multi-modal content using OCR and VLM respectively. "*Image*" refers to processing document page as screenshot image.

		Method \ Domain	Resear. Report	Admin &Indu.	Tutori.& Worksh.	Acade. Paper	Broch-ure	Finance Report	Guide-book	Govern-ment	Laws	News	Average Macro	Average Micro
Recall@$k=1$	Textual Retrieval OCR-text	DPR	3.4	7.2	1.2	11.3	3.0	9.8	8.2	24.7	30.9	26.3	12.6	12.4
		ColBERT	5.0	8.8	4.7	16.4	2.0	13.2	4.6	50.8	47.7	**45.3**	19.8	19.1
		BGE	7.0	10.9	3.7	14.3	2.3	16.3	8.4	46.1	45.5	35.8	19.0	18.5
		E5	6.3	6.0	2.6	14.0	3.5	14.4	6.1	44.7	45.4	40.5	18.4	17.9
		Contriever	6.7	7.0	3.8	14.3	4.1	13.3	8.3	43.6	43.9	42.9	18.8	18.1
		GTE	5.7	7.0	3.9	17.2	2.8	15.8	9.4	40.7	41.6	38.3	18.2	18.4
	Textual Retrieval VLM-text	DPR	11.6	9.5	19.2	19.2	14.9	15.9	15.8	25.6	34.7	27.0	19.3	19.2
		ColBERT	22.0	14.9	28.0	28.3	17.9	29.7	21.1	**52.6**	**54.5**	44.5	31.3	31.4
		BGE	19.2	15.2	24.6	28.7	12.8	27.6	19.7	47.0	52.3	35.8	28.3	29.0
		E5	15.9	8.8	27.7	24.3	14.6	21.8	14.7	45.6	53.0	40.5	26.7	26.4
		Contriever	23.4	7.5	28.2	26.8	17.1	25.7	16.1	43.6	51.5	42.9	28.3	28.9
		GTE	17.5	10.5	23.0	27.2	14.5	26.3	14.4	39.8	49.2	38.3	26.1	27.1
	Visual Retrieval Pure-Image	$\text{DSE}_{\text{wiki-ss}}$	20.6	15.1	31.0	31.1	20.1	29.2	22.0	39.3	37.5	35.8	28.2	29.2
		$\text{DSE}_{\text{docmatix}}$	19.9	11.4	31.5	30.1	17.8	30.0	20.8	46.5	39.4	31.4	27.9	29.1
		ColPali	22.5	21.3	36.6	30.9	26.8	**32.1**	19.3	52.5	51.8	33.6	**32.7**	**32.5**
		$\text{DPR-Phi3}_{\text{ours}}$	21.1	**22.1**	36.8	**35.2**	25.6	28.7	24.1	38.3	35.4	27.4	29.5	30.2
		$\text{Col-Phi3}_{\text{ours}}$	22.6	22.0	**37.5**	34.9	28.9	30.3	22.7	50.2	45.1	26.3	31.1	31.6
	Visual Retrieval Hybrid	$\text{DSE}_{\text{wiki-ss}}$	14.0	10.4	29.8	18.0	13.7	20.4	13.5	46.0	45.1	34.7	24.6	23.4
		$\text{DSE}_{\text{docmatix}}$	18.2	11.6	32.7	24.0	17.7	27.2	16.7	48.1	45.5	33.0	27.5	27.4
		ColPali	17.7	12.3	30.0	18.4	19.0	25.5	20.6	49.7	51.2	40.9	28.5	27.1
		$\text{DPR-Phi3}_{\text{ours}}$	**28.3**	11.1	35.5	18.8	29.3	24.0	**27.4**	38.0	41.9	34.5	28.9	27.3
		$\text{Col-Phi3}_{\text{ours}}$	26.4	12.6	33.7	27.3	**30.1**	27.9	24.6	46.2	47.4	21.9	29.8	29.6
Recall@$k=5$	Textual Retrieval OCR-text	DPR	7.3	12.5	5.6	24.0	8.9	16.9	13.6	47.2	50.2	51.1	23.7	23.5
		ColBERT	10.9	23.8	10.2	32.2	6.8	25.5	17.0	78.7	63.4	63.2	33.2	32.6
		BGE	11.9	20.3	13.6	30.0	11.7	27.7	18.8	68.5	65.4	59.1	32.7	32.2
		E5	12.8	16.2	8.9	31.9	10.9	23.6	19.9	76.1	68.8	63.9	33.3	32.7
		Contriever	11.9	17.9	11.9	28.8	9.3	24.6	18.1	64.3	68.0	62.7	31.7	31.2
		GTE	10.0	18.2	12.8	32.9	15.2	29.4	21.0	67.7	65.3	62.4	33.5	33.5
	Textual Retrieval VLM-text	DPR	31.0	25.7	36.7	44.9	33.0	34.1	34.9	49.9	56.3	51.1	39.8	40.4
		ColBERT	41.8	37.7	53.7	61.8	35.1	52.4	46.1	83.2	70.1	62.5	**54.4**	**56.0**
		BGE	41.0	28.1	52.7	59.2	36.7	46.0	**50.7**	72.0	71.5	59.9	51.8	53.2
		E5	35.4	28.1	51.7	58.5	33.2	41.2	40.2	**79.7**	**77.9**	**64.6**	51.1	51.8
		Contriever	40.2	29.2	54.1	57.9	36.8	47.1	44.6	68.8	76.2	61.9	51.7	53.0
		GTE	36.7	25.6	51.2	56.6	39.7	46.6	46.4	72.2	74.3	63.2	51.3	52.3
	Visual Retrieval Pure-Image	$\text{DSE}_{\text{wiki-ss}}$	42.4	32.9	56.3	58.5	39.8	50.6	41.6	68.6	60.9	50.0	50.2	52.1
		$\text{DSE}_{\text{docmatix}}$	39.6	36.2	53.9	57.5	33.7	52.5	42.8	69.4	63.1	48.9	49.8	51.9
		ColPali	40.7	**45.9**	54.9	58.5	42.6	51.2	45.7	76.8	74.5	48.9	54.0	54.3
		$\text{DPR-Phi3}_{\text{ours}}$	45.5	37.7	**57.0**	**62.9**	41.4	51.1	45.5	65.1	60.8	49.3	51.6	53.9
		$\text{Col-Phi3}_{\text{ours}}$	46.4	38.2	53.1	61.8	**45.0**	**54.6**	45.7	68.8	65.7	43.8	52.3	54.5
	Visual Retrieval Hybrid	$\text{DSE}_{\text{wiki-ss}}$	31.8	29.5	51.1	43.0	34.5	39.3	38.3	71.3	71.3	57.1	46.7	45.6
		$\text{DSE}_{\text{docmatix}}$	37.3	26.7	48.2	49.7	34.5	48.6	41.0	72.2	69.4	54.2	48.2	49.3
		ColPali	40.1	38.3	55.2	49.2	42.7	47.6	40.8	78.6	68.7	61.0	52.2	51.4
		$\text{DPR-Phi3}_{\text{ours}}$	**54.2**	27.4	53.8	39.9	36.6	45.4	49.6	67.2	66.8	59.8	50.1	49.3
		$\text{Col-Phi3}_{\text{ours}}$	50.9	25.5	49.1	58.3	41.9	48.1	49.2	62.3	60.6	48.5	50.0	51.8
Recall@$k=10$	Textual Retrieval OCR-text	DPR	10.5	21.1	8.8	32.0	14.9	19.6	17.0	59.7	56.4	58.9	29.9	29.3
		ColBERT	14.1	31.1	13.1	38.4	9.1	31.0	22.4	83.9	67.9	65.4	37.6	37.4
		BGE	15.7	24.3	15.9	35.9	17.9	31.6	25.3	76.3	73.4	62.1	37.8	37.3
		E5	16.9	24.5	13.8	40.1	15.8	26.7	24.7	83.5	77.9	66.1	39.0	38.3
		Contriever	15.1	25.7	14.2	36.8	15.9	27.9	24.2	76.8	71.8	64.9	37.3	36.6
		GTE	19.1	29.0	21.4	39.0	19.2	32.9	29.2	78.8	74.5	66.2	40.9	40.1
	Textual Retrieval VLM-text	DPR	42.2	33.1	52.1	56.2	39.9	43.5	44.0	62.8	61.7	59.7	49.5	50.5
		ColBERT	51.0	48.7	60.6	69.8	43.9	**61.6**	53.7	**88.4**	74.8	66.4	61.9	**63.7**
		BGE	51.1	38.7	62.1	71.5	41.9	55.6	**58.7**	80.8	78.7	63.5	60.3	62.4
		E5	45.3	38.6	62.0	70.5	45.6	50.0	55.3	87.1	82.4	**66.8**	60.4	61.2
		Contriever	49.9	41.3	62.0	70.5	44.8	56.5	54.5	81.3	78.0	64.9	60.4	62.2
		GTE	48.6	41.1	61.5	68.8	44.3	56.9	58.0	83.0	77.5	66.9	60.7	62.2
	Visual Retrieval Pure-Image	$\text{DSE}_{\text{wiki-ss}}$	55.9	41.3	61.5	68.1	47.8	60.7	54.2	72.9	68.3	54.4	58.5	61.1
		$\text{DSE}_{\text{docmatix}}$	53.7	43.3	59.6	66.5	44.7	59.1	50.3	75.4	69.2	53.7	57.5	59.9
		ColPali	53.6	**54.1**	64.4	69.5	48.8	60.7	54.0	81.9	**82.5**	50.4	**62.0**	63.2
		$\text{DPR-Phi3}_{\text{ours}}$	58.1	49.1	**67.0**	74.7	48.4	57.9	57.8	68.7	66.2	54.4	60.2	62.8
		$\text{Col-Phi3}_{\text{ours}}$	57.7	50.5	66.6	72.3	**50.7**	59.3	53.6	68.5	74.8	57.5	61.1	63.3
	Visual Retrieval Hybrid	$\text{DSE}_{\text{wiki-ss}}$	44.1	34.3	57.6	56.3	42.6	50.7	48.6	81.1	79.1	63.7	55.8	56.0
		$\text{DSE}_{\text{docmatix}}$	49.9	37.3	57.3	61.3	45.9	57.9	50.1	77.9	74.9	61.5	57.4	58.9
		ColPali	52.1	46.4	65.0	64.4	**50.7**	53.8	51.0	82.7	71.4	62.5	60.0	60.2
		$\text{DPR-Phi3}_{\text{ours}}$	**65.2**	33.7	60.3	51.3	42.4	52.4	52.9	79.1	72.5	65.6	55.5	53.4
		$\text{Col-Phi3}_{\text{ours}}$	59.1	38.7	57.0	**77.7**	43.9	57.7	51.7	72.4	68.0	60.7	58.7	62.8

Table 11: Main results for layout-level retrieval. "*Pure-Image*" and "*Hybrid*" refer to reading textual layouts in image and text format respectively.

B Dataset Construction

B.1 Related DocVQA Benchmarks

Early DocVQA benchmarks primarily address single-page visual question answering (VQA), exemplified by datasets such as DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), and TAT-DQA (Zhu et al., 2022). To overcome the limitations of single-page inputs, subsequent datasets like DUDE (Landeghem et al., 2023), MP-DocVQA (Tito et al., 2023), and Slide-VQA (Tanaka et al., 2023) have extended the context length to an average of 5.7, 8.3, and 20 pages, respectively. More recent benchmarks, including MMLongBench-Doc (Ma et al., 2024b) and DocBench (Zou et al., 2024), treat DocVQA as a long-context task, accommodating entire documents that average between 50–70 pages in length. As document lengths increase, retrieval becomes essential. Relevant pages must first be identified, followed by answer generation based on the retrieved multimodal evidence.

B.2 Document Corpora Collection Criteria

To facilitate the development of MMDocIR, we leverage resources from existing DocVQA benchmarks described in Appendix B.1. Despite not being curated for IR, they offer valuable document corpora and questions that can be adapted for IR tasks. We select relevant DocVQA datasets based on the following criteria:

- **Document Source**: The dataset must include accessible original documents or sources for these documents. We need to access and enrich them to support more complex retrieval tasks.
- **Diverse Domain/Modality**: The document collections must (1) encompass diverse domains suitable for generalized evaluation, and (2) contain multiple modalities, such as text, figures, tables, charts, and layouts.
- **Long Document**: We choose documents with extensive texts as longer texts pose more significant challenges. This criterion can evaluate models in handling complex and lengthy documents.
- **Question Diversity and Comprehensiveness:** The questions included in the dataset should be diverse and challenging. For example, cross-modal questions require reasoning across both text and visual tables/figures; multi-hop questions require reasoning over multiple steps; multi-page questions require combining information from multiple pages.

Considering these criteria, we utilize document corpora and questions from datasets as follows:

- **Evaluation**: MMLongBench-Doc (Ma et al., 2024b) and DocBench (Zou et al., 2024).
- **Training**: MP-DocVQA (Tito et al., 2023), Slide-VQA (Tanaka et al., 2023), TAT-DQA (Zhu et al., 2022), SciQAG (Wan et al., 2024), DUDE (Landeghem et al., 2023), and CUAD (Hendrycks et al., 2021).

B.3 Question Filtering Guidelines

We filter questions based on the following criteria:

- Summarization Questions: Questions such as "*What does this book mainly illustrate?*""*What does this story mainly tell?*" require understanding of large sections or even the entire document. The broad scope makes it hard to pinpoint specific content and contradicts the IR nature of our task.
- Overwhelm Statistical Questions: Questions that demand extensive data computation or collation, such as "*How many words are there in total in the paper?*""*How many pictures are there in total in the document?*" are also excluded from our scope.
- Online Search Questions: Questions like "*What is the Google Scholar citation count of the author?*" rely on information from external online resources. We focus only on retrieving information within the documents, and therefore exclude these questions.
- Unanswerable Questions: These are designed to test if models generate answers based on non-existent information (model hallucinations). Since they do not facilitate the retrieval of factual document-based information, these questions are excluded.

B.4 Training Document Collection

We collect the training datasets as follows:

- **MP-DocVQA** (Tito et al., 2023) contains 47,952 images collected from Industry Documents Library (IDL) [3]. IDL is a crucial resource for public health research, containing millions of documents produced by industries such as tobacco, drug, chemical, and food, which have had significant impacts on public health. We group the 47,952 document images into separate document files, and obtain 875 long documents (46.8 pages on average) with 15,266 QA pairs.

[3] https://www.industrydocuments.ucsf.edu/

- **SlideVQA** (Tanaka et al., 2023) contains 2,619 slide documents collected from slideshare [4] and covering 39 topics. SlideVQA hosts a wide variety of slide presentations across various categories such as business, mobile, social media, marketing, technology, arts, career, design, education, and government & nonprofit, among others, which can enrich the diversity of our corpus. Note that SlideVQA contains only the first 20 pages for each slide deck. In our research, we manually collect the remaining missing pages, and obtain 2,011 long documents (averaging 49.3 pages) with 11,066 QA pairs. SlideVQA requires complex reasoning, including single-hop, multi-hop, and numerical reasoning, and also provides annotated arithmetic expressions of numerical answers for enhancing the ability of numerical reasoning.
- **TAT-DQA** (Zhu et al., 2022) consists of 3,067 document pages from financial reports [5], dated between 2018 and 2020. AnnualReports.com provides access to a comprehensive collection of corporate annual reports from over 10,320 companies worldwide. Note that neither original documents nor links are provided. We use OCR to extract text in the pages, and use search engine to find relevant documents. After careful tracing and recognition, we identify 163 original documents (averaging 147.3 pages) with 15,814 QA pairs.
- **arXivQA** (Li et al., 2024) comprises 32k figures cropped from academic pages [6]. The papers on arXiv cover a wide range of disciplines including physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, engineering, and systems science, and economics, etc. We use the arXiv DOIs provided to collect the academic papers. Due to the missing of paper versions, extra efforts are made to identify paper versions. After careful tracing, recognition, and document length filtering, we identify 1,579 documents averaging 18.4 pages.
- **SciQAG** (Wan et al., 2024) consists of 22,728 papers and 188,042 QA pairs in 24 scientific disciplines, collected from Web of Science (WoS) Core Collection database. WoS provides comprehensive scientific literature in natural sciences, social sciences, arts, and humanities. We sample 50 documents from each discipline, and manually collect 1,197 papers using the DOIs provided.
- **DUDE** (Landeghem et al., 2023) provides 5,019 documents from aggregator websites[7]. It covers a broad range of domains, including medical, legal, technical, and financial, among others, to evaluate models' ability to handle diverse topics and the specific knowledge each requires. We filter out short documents and obtain 779 relatively long documents (averaging 15.6 pages) with 3,173 QA pairs.
- **CUAD** (Hendrycks et al., 2021) provides 510 commercial legal contracts, collected from Electronic Data Gathering, Analysis, and Retrieval (EDGAR)[8]. EDGAR contracts are usually more complex and heavily negotiated than the general population of all legal contracts. We filter out short documents in CUAD and obtain 274 long documents (29.6 pages on average) with 11,234 QA pairs.

B.5 Training Dataset Label Construction

The page labels can be directly obtained in the MP-DocVQA, SlideVQA, and DUDE datasets. Among these, only DUDE provides layout labels.

SciQAG provides only question and answer in texts. We use these information to infer the page-level and layout-level labels. Specifically, we first use MinerU to obtain layout-level passage chunks. For each QA pair, we deploy E5 and BGE retrievers to obtain question-passage and answer-passage similarity scores against all extracted passage chunks. If both scores rank within top 3 for a specific passage chunk, we assign this layout as the layout-level labels for the given QA pair.

Similarly, **arXivQA** provides only cropped images, without document page/layout labels. We first use MinerU to obtain layout-level images. For each cropped image, we calculate its similarity against all extracted images using brute-force matcher[9], and select the most similar one. Subsequently, we manually examine if the selected image matches the cropped image. In this way, we filter around 20% unmatched images, resulting 1,579 questions with page and layout level labels.

For **TAT-DQA**, layout-level labels are provided for each sampled page. To localize the page index

[4] https://www.slideshare.net/
[5] https://www.annualreports.com/
[6] https://arxiv.org/
[7] 1: archive.org, 2: http://commons.wikimedia.org/, 3: http://documentcloud.org/
[8] https://www.sec.gov/search-filings
[9] https://opencv.org/

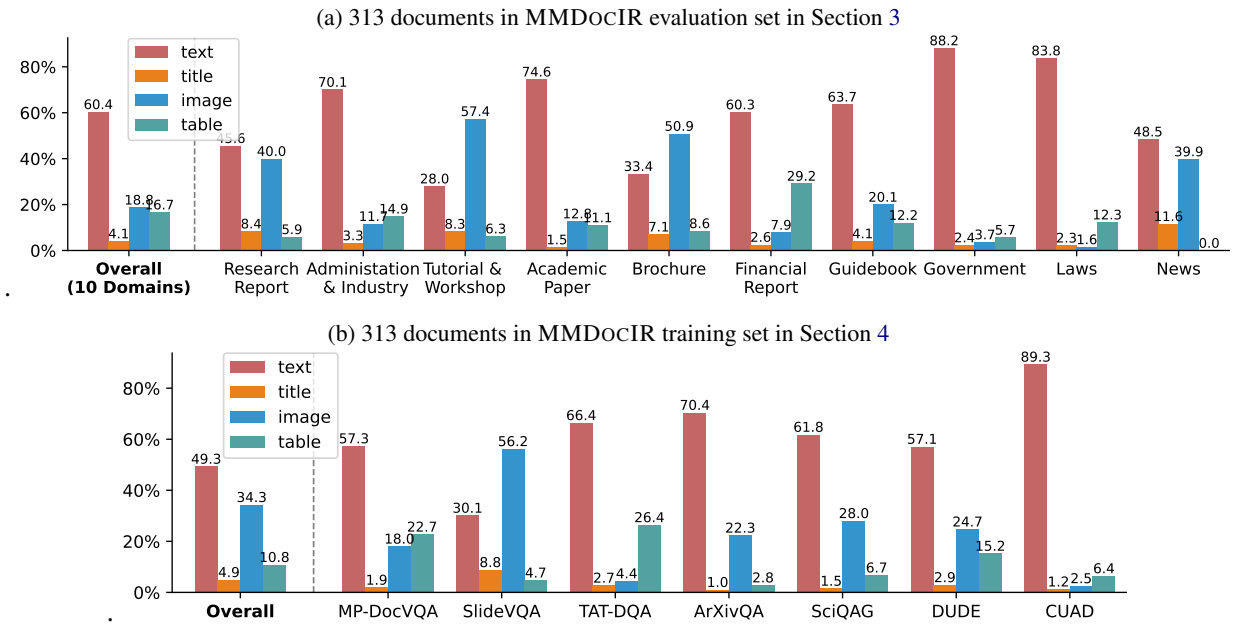


Figure 4: Area ratio of different modalities (1) in overall and (2) by domains/datasets in MMDocIR evaluation and training set. Note that white spaces, headers, and footers are excluded from the area calculations.

of the sampled pages, we first utilize PDF mapping tool[10] to retrieve best matched page in the document. Then, we manually verify whether the retrieved page matches the given page, and correct the labels if there were any errors.

B.6 Hard Negative Sampling

In addition to annotating ground truth (positive) page labels, we enhance our training data with negative labels. In the context of retrieval, *hard negatives* are particularly informative non-relevant documents that closely resemble true positives according to the model's current scoring function. Unlike randomly selected negatives, hard negatives are challenging for the model to distinguish from relevant documents, thus providing stronger supervision.

In our framework, hard negatives are crucial for improving retrieval performance. By training the model on these challenging examples, we encourage it to learn more discriminative representations, ultimately enhancing its robustness and reducing false positives during retrieval.

As described in Appendix C.3, training is conducted using a contrastive loss, where the model aims to separate relevant documents from irrelevant ones. Specifically, we obtain hard negatives using the ColPali retriever (Faysse et al., 2024), which scores all document pages for a given query. The irrelevant pages with the highest top-k scores (*i.e.,* those most likely to be confused with positives) are selected as hard negatives for training.

B.7 Fine-grained Modality Distribution

MMDocIR evaluation set includes 313 long documents with an average length of 65.1 pages, categorized into ten main domains: research reports, administration&industry, tutorials&workshops, academic papers, brochures, financial reports, guidebooks, government documents, laws, and news articles. Overall, the modality distribution is: Text (60.4%), Image (18.8%), Table (16.7%), and other modalities (4.1%), as shown in Figure 4a Different domains exhibit different distributions of multimodal information. For instance, research reports, tutorials, workshops, and brochures predominantly contain images, whereas financial and industry documents are table-rich. In contrast, government and legal documents primarily comprise text.

MMDocIR training set includes 6,878 long documents with an average length of 32.6 pages, categorized into seven Document VQA or QA datasets: MP-DocVQA (Tito et al., 2023), SlideVQA (Tanaka et al., 2023), TAT-DQA (Zhu et al., 2022), arXivQA (Li et al., 2024), SciQAG (Wan et al., 2024), and DUDE (Landeghem et al., 2023). Overall, the modality distribution is: Text (49.3%), Image (34.3%), Table (10.8%), and other modalities (4.9%), as shown in Figure 4b. Each dataset

[10] https://github.com/pymupdf/PyMuPDF

Artifacts	Purpose	Referred Section	Resource URL
MMLong'-Doc	Eval-set curation	Section 3	https://github.com/mayubo2333/MMLongBench-Doc
DocBench	Eval-set curation		https://github.com/Anni-Zou/DocBench
MP-DocVQA	Train-set curation	Section 4 Appendix B.4	https://rrc.cvc.uab.es/?ch=17&com=tasks
SlideVQA	Train-set curation		https://github.com/nttmdlab-nlp/SlideVQA
TAT-DQA	Train-set curation		https://github.com/NExTplusplus/TAT-DQA
arXivQA	Train-set curation		https://huggingface.co/datasets/taesiri/arxiv_qa
SciQAG	Train-set curation		https://github.com/MasterAI-EAM/SciQAG
DUDE	Train-set curation		https://github.com/duchallenge-team/dude
CUAD	Train-set curation		https://www.atticusprojectai.org/cuad
MinerU	Doc Parsing	Section 3.2	https://github.com/opendatalab/MinerU
Tesseract OCR	OCR-text	Section 3.4	https://github.com/tesseract-ocr/tesseract
GPT-4o	VLM-text		https://openai.com/index/hello-gpt-4o/
QwenVL2.5	VLM-text		https://github.com/QwenLM/Qwen2.5-VL
PyMuPDF	Page matching	Appendix B.5	https://github.com/pymupdf/PyMuPDF
OpenCV	Image matching		https://opencv.org/
Text Retriever	Layout location		BGE and E5 (see Table 13)
Visual Retriever	Hard negatives	Appendix B.6	Colpali (see Table 13)

Table 12: Artifacts used to facilitate construction of MMDOCIR evaluation & train set.

features unique distributions of multimodal content. The legal documents and academic papers are text-intensive. The slides consist mostly of visual features. Industrial documents and financial reports are table-intensive.

B.8 Resource URL of Artifacts

In this section, we summarize the artifacts used to facilitate the construction of MMDOCIR's evaluation and train set, as shown in Table 12. These artifacts mainly includes: datasets used for curating MMDOCIR evaluation and training sets, tools for parsing documents, packages for locating evidence, and etc.

C Model Training: DPR-Phi3&Col-Phi3

To evaluate the effectiveness of the MMDOCIR training set, we train two visual retrievers based on `Phi3-Vision` (Abdin et al., 2024). `Phi3-Vision` ($\mathbf{M_{phi3v}}$) reuses the image tokenizer from `clip-vit-large`[11] ($\mathbf{M_{vit}}$). It can deal with high-resolution images by cropping them into sub-images, where each sub-image has 336×336 pixels.

C.1 Document/Query Encoding

`DPR-Phi3` and `Col-Phi3` represent document page or query using a single dense embedding (following DPR (Karpukhin et al., 2020)) and a list of token-level embeddings (following ColBERT (Khattab and Zaharia, 2020)), respectively. Specifically, we follow Ma et al. (2024a) to concatenate document image with a text prompt: "*<s><d> What is shown in this image?</s>*". Here, the <d> token is a special placeholder token and is replaced by the sequence of patch latent embeddings from the vision encoder. We consider only text queries and use text prompt: "*<s> query: <q> </s>*". Similarly, the placeholder <q> token is replaced by input query. We encode query q and document d in two ways:

$$\begin{aligned} \mathrm{E}_{\mathrm{d}}^{\mathrm{dpr}} &= \mathbf{M_{phi3v}}\big(\mathbf{M_{vit}}(d), \text{prompt}\big)[-1], \in \mathbb{R}^{D_1} \\ \mathrm{E}_{\mathrm{q}}^{\mathrm{dpr}} &= \mathbf{M_{phi3v}}\big(q, \text{prompt}\big)[-1], \in \mathbb{R}^{D_1} \end{aligned} \tag{1}$$

where the end-of-sequence token </s> from the last hidden state ($D_1 = 3072$) of $\mathbf{M_{phi3v}}$ is used to represent $\mathrm{E}_{\mathrm{d}}^{\mathrm{dpr}}$ and $\mathrm{E}_{\mathrm{q}}^{\mathrm{dpr}}$.

$$\begin{aligned} \mathrm{E}_{\mathrm{d}}^{\mathrm{col}} &= \mathbf{M_{proj}} \cdot \mathbf{M_{phi3v}}\big(\mathbf{M_{vit}}(d), \text{prompt}\big) \\ \mathrm{E}_{\mathrm{q}}^{\mathrm{col}} &= \mathbf{M_{proj}} \cdot \mathbf{M_{phi3v}}\big(q, \text{prompt}\big) \end{aligned} \tag{2}$$

where $\mathrm{E}_{\mathrm{d}}^{\mathrm{col}} \in \mathbb{R}^{N_d \times D_2}$ and $\mathrm{E}_{\mathrm{q}}^{\mathrm{col}} \in \mathbb{R}^{N_q \times D_2}$, and $\mathbf{M_{proj}}$ is projection layer to map the last hidden states of $\mathbf{M_{phi3v}}$ into reduced dimension $D_2 = 128$. $N_d \approx 2500$ for a typical high-resolution page and N_q is the number of query tokens.

C.2 Query-Doc Similarity

The similarity between the query and the document is computed as follows:

$$\mathrm{Sim}(q, d)_{dpr} = \frac{\langle \mathrm{E}_{\mathrm{q}}^{\mathrm{dpr}} | \mathrm{E}_{\mathrm{d}}^{\mathrm{dpr}} \rangle}{\left\| \mathrm{E}_{\mathrm{q}}^{\mathrm{dpr}} \right\| \cdot \left\| \mathrm{E}_{\mathrm{d}}^{\mathrm{dpr}} \right\|} \tag{3}$$

[11] ViT-Large: https://huggingface.co/openai/clip-vit-large-patch14-336

where $Sim(q,d)_{dpr}$ is computed as the cosine similarity between their embeddings. and $\langle\cdot|\cdot\rangle$ is the dot product.

$$\mathrm{Sim}(q,d)_{col} = \sum_{i\in[1,N_q]} \max_{j\in[1,N_d]} \langle \mathrm{E}_{\mathrm{q}}^{\mathrm{col}(i)} | \mathrm{E}_{\mathrm{d}}^{\mathrm{col}(j)} \rangle \tag{4}$$

where $Sim(q,d)_{col}$ is the sum over all query vectors $\mathrm{E}_{\mathrm{q}}^{\mathrm{col}(i)}$, of its maximum dot product $\langle\cdot|\cdot\rangle$ with each of the N_d document embedding vectors $\mathrm{E}_{\mathrm{d}}^{\mathrm{col}(j)}$.

C.3 Contrastive Loss

Given the query q, we have the positive document d^+ and a set of negative documents d^- including hard negatives and in-batch negatives. The hard negatives are negative pages within the document with highest $Sim(q,d^-)$ scored by ColPali (Faysse et al., 2024) retriever, refer to Appendix B.6 for more details on hard negative selection. We calculate the loss as:

$$\mathcal{L}_{(q,d^+,d^-)}^{dpr} = -\log \frac{\exp(\mathrm{Sim}_{(q,d^+)}^{dpr}/\tau)}{\sum_{d_i\in d^+\cup d^-} \exp(\mathrm{Sim}_{(q,d_i)}^{dpr}/\tau)} \tag{5}$$

where `DPR-Phi3` is trained on the InfoNCE loss, and the temperature parameter $\tau = 0.02$ in our experiments.

$$\mathcal{L}_{(q,d^+,d^-)}^{col} = \log\Big(1+\exp\big(\max_{d_i\in d^-}(\mathrm{Sim}_{(q,d_i)}^{col}) - \mathrm{Sim}_{(q,d^+)}^{col}\big)\Big) \tag{6}$$

where `Col-Phi3` is trained via the softplus loss based on the positive scores w.r.t. to the maximal negative scores.

C.4 Training Implementation Details

In summary, we train two visual retrievers based on `Phi3-Vision` (Abdin et al., 2024). `DPR-Phi3` and `Col-Phi3` represent document page or query using a single dense embedding (following DPR (Karpukhin et al., 2020)) and a list of token-level embeddings (following ColBERT (Khattab and Zaharia, 2020)), respectively. To train the model, we employ memory-efficient techniques such as LoRA (Hu et al., 2022), FlashAttention (Dao, 2024), and DeepSpeed (Rasley et al., 2020).

The model is trained with a batch size of 64 for one epoch on MMDocIR training set. The model weights are shared between the language models for document screenshot and query encoding. In both tasks, each training query is paired with one positive document and one hard negative document. The document screenshots are resized to $1,344 \times 1,344$ pixels and cropped into 4×4 sub-images. The training process is performed on four A100 80GB GPUs.

D Retrievers: Introduction and Implementation Details

D.1 Text-Centric Document Retrieval

For text retrieval, the first step is to convert multimodal document into text using techniques, *e.g.,* Document Parsing (Chao and Fan, 2004; Wang et al., 2024), Optical Character Recognition (OCR) (Chaudhuri et al., 2017; Borovikov, 2014; Mori et al., 1999), Layout Detection (Sassioui et al., 2023; Xu et al., 2020, 2021), Chunking (Chen et al., 2024a; Raina and Gales, 2024; Dong et al., 2024), and Image Captioning (You et al., 2016; Aneja et al., 2018). These steps are time-consuming and can introduce errors that impact the overall retrieval performance. Current text retrieval are primarily categorized as sparse or dense retrieval. For two widely-used sparse retrievers: TF-IDF (Salton et al., 1983) calculates the relevance via word frequency with the inverse document frequency, and BM25 (Robertson et al., 1994) introduces nonlinear word frequency saturation and length normalization. Dense retrievers encode content into vector representations. DPR (Karpukhin et al., 2020) is the pioneering work of dense vector representations for QA tasks. Similarly, ColBERT (Khattab and Zaharia, 2020) introduces an efficient question-document interaction model with late fine-grained term matching. Contriever (Izacard et al., 2022) leverages contrastive learning to improve content dense encoding. E5 (Wang et al., 2022) and BGE (Xiao et al., 2023) propose novel training and data preparation techniques to enhance retrieval performance. Moreover, GTE (Li et al., 2023) integrates graph-based techniques to enhance dense embedding. However, most text retrieval systems overlook valuable visual information present in documents.

D.2 Vision-Driven Document Retrieval

Vision Language Models (VLMs) (Abdin et al., 2024; Beyer et al., 2024; Bai et al., 2023; Chen et al., 2024b) can understand and generate text based on combined text and visual in-

	Model	Dimension	Base Model	HuggingFace Checkpoint
Text	DPR	768	BERT-base	facebook/dpr-ctx_encoder-multiset-base facebook/dpr-question_encoder-multiset-base
	ColBERT	$N_{\text{tok}} \times 768$	BERT-base	colbert-ir/colbertv2.0
	Contriever	768	BERT-base	facebook/contriever-msmarco
	E5	1,024	BERT-large	intfloat/e5-large-v2
	BGE	1,024	RetroMAE	BAAI/bge-large-en-v1.5
	GTE	1,024	BERT-large	thenlper/gte-large
Visual	$\text{DSE}_{\text{wiki-ss}}$	3,072	Phi-3-Vision	Tevatron/dse-phi3-v1.0
	$\text{DSE}_{\text{docmatix}}$	3,072	Phi-3-Vision	Tevatron/dse-phi3-docmatix-v2
	ColPali	$N_{\text{tok}} \times 1{,}024$	PaliGemma	vidore/colpali

Table 13: Implementation details for Text and Vision Retrieval Models

puts. This advancement has led to the development of cutting-edge visual-driven retrievers, such as ColPali (Faysse et al., 2024) and DSE (Ma et al., 2024a). These models specifically leverage PaliGemma (Beyer et al., 2024) and Phi3-Vision (Abdin et al., 2024) to directly encode document page screenshots for multimodal document retrieval. ColPali adopts a similar question-document interaction as ColBERT, and represents each document page in token-level embeddings. By contrast, DSE is similar to DPR in that it encodes each page with a single dense embedding. Visual retrievers are capable of modeling useful visual information, allowing direct utilization of multimodal content without first converting it into text first. Despite these advancements, visual retrievers face challenges, particularly in dealing with text details when document page resolutions are high. The high resolution of document pages substantially increases the computational cost and complexity of the embedding process, which may hinder the model's performance.

D.3 Implementation Details

In our experiments (refer to Section 5.2), we implement 9 off-the-shelf retrievers including 6 text retrievers and 3 visual retrievers. The text retrieval models deployed are namely DPR, ColBERT, Contriever, E5, BGE and GTE. These models use the WordPiece tokenizer from BERT and also inherit the maximum input length of 512 tokens from BERT (Devlin et al., 2019). Additionally, we make use of the sentence-transformer library[12] when deploying E5, BGE and GTE. The visual retrieval models deployed are namely $\text{DSE}_{\text{wiki-ss}}$, $\text{DSE}_{\text{docmatix}}$, and ColPali. We use pre-trained checkpoints available on HuggingFace [13]; the specific checkpoint information can be found in Table 13 alongside other configuration details.

E Dataset Demonstration

E.1 Document Pages by Domains

The documents in MMDocIR can be categorized into 10 types. We provide examples of each type as below.

- **Admin & Industry**: These documents primarily consist of instructional and overview content on industry, reflected by the dominance of text-based questions (78.0%) and a smaller reliance on visual evidence (image questions only 20.3%), which shows a text-heavy structure (70.1%). Some detailed examples are shown in Figure 5b.
- **Tut & Workshop**: Documents in this category focus on slides or tutorials, which exhibit a balanced question modality: 61.7% text, 24.5% image, and 9.5% table questions. Strong visual components are present, with 57.4% of its content being images—the highest among all categories. Some detailed examples are shown in Figure 5c.
- **Academic Paper**: These documents are formal publications with structured layouts, citations, and academic pictures. The questions span multiple modalities: 28.8% text, 25.7% image, and 50.0% table. Text modality dominates content distribution (74.6%), with the presence of tables (11.1%) and images (12.8%) demonstrating rich multimodal alignment and explicit questions with answers. Some detailed examples are shown in Figure 6a.

[12] https://www.sbert.net/

[13] https://huggingface.co/

- **Brochure**: Designed for promotional purposes, the brochure category contains highly visual documents. Over 52.6% of questions are image-based—the highest among all domains—while text-based questions account for only 60.5%. Modality distribution is similarly diverse: 50.8% image, showcasing their visually complex layout. Some detailed examples are shown in Figure 6b.
- **Financial Report**: These documents involve massive numerical and quantitative data, reflected in a high proportion of table questions (54.5%) and strong table content distribution (29.2%). While text remains significant (60.3%), the inclusion of tabular and numerical analysis is essential for understanding these documents. Some detailed examples are shown in Figure 6c.
- **Guidebook**: Instruction manuals for electronics and tools, guidebooks exhibit the most balanced question modality: 51.8% text, 54.4% image, and 26.8% table, indicating multimodal instructional designs. Some detailed examples are shown in Figure 7a.
- **Government**: This category covers policy files and governmental reports. It is highly text-centric with 69.9% text questions and 88.2% text content. This reflects the formal and regulatory nature of such documents. Some detailed examples are shown in Figure 7b.
- **Laws**: Legal documents exhibit strong textual dominance both in questions (62.1%) and content (83.8%), with very limited visual presence (image content only 1.6%). They often maintain specific formats and focus on linguistic interpretation rather than visual layout. Some detailed examples are shown in Figure 7c.
- **News**: Although based on only one document, the "News" domain shows notable multimodal richness. It includes a significant image portion (39.8%), high text presence (48.5%), and 11.6% titles. This reflects the use of images and headlines typical of news articles. Some detailed examples are shown in Figure 7d.

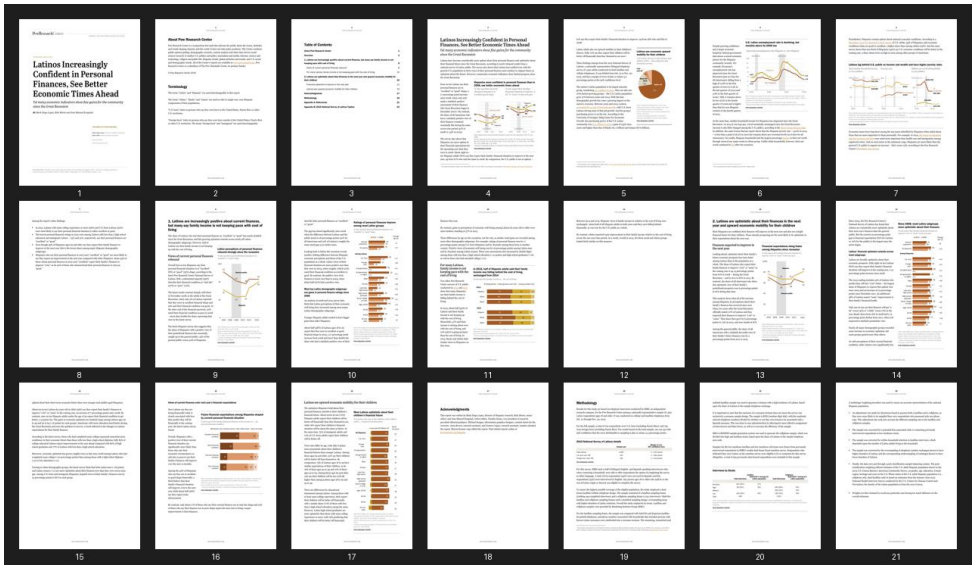
E.2 Document Layouts

In this section, we present 9 pages along with their detected layouts, which are highlighted for better visualizations, as shown in Figure 8, 9, and 10. Specifically, layout detection identifies the spatial location of different content types, such as images, tables, and text within a document. With the help of layout detection, we can precisely locate the specific position of an answer, whether it is an image, a text paragraph, or a table. This enables a more fine-grained layout-level evaluation of multimodal retrieval capabilities.

E.3 Annotation Examples

In this section, we present 4 annotation examples that illustrate typical multimodal retrieval and reasoning patterns, which help explain the construction and retrieval process. Each annotation includes the following primary components: question, answer, page-level labels, and layout-level labels. The page-level labels show the selected pages that contain ground truth evidence. Based on these selected pages, layout-level labels further display the specific layout box detection of evidence. These examples frequently require reasoning across multiple pages and modalities. The evidence encompasses diverse formats such as figures, charts, tables, and texts, highlighting the complexity and richness of the multimodal retrieval tasks.

(a) Page screenshots in Research Report domain.



(b) Page screenshots in Administration & Industry domain.

(c) Page screenshots in Tutorial & Workshop domain.

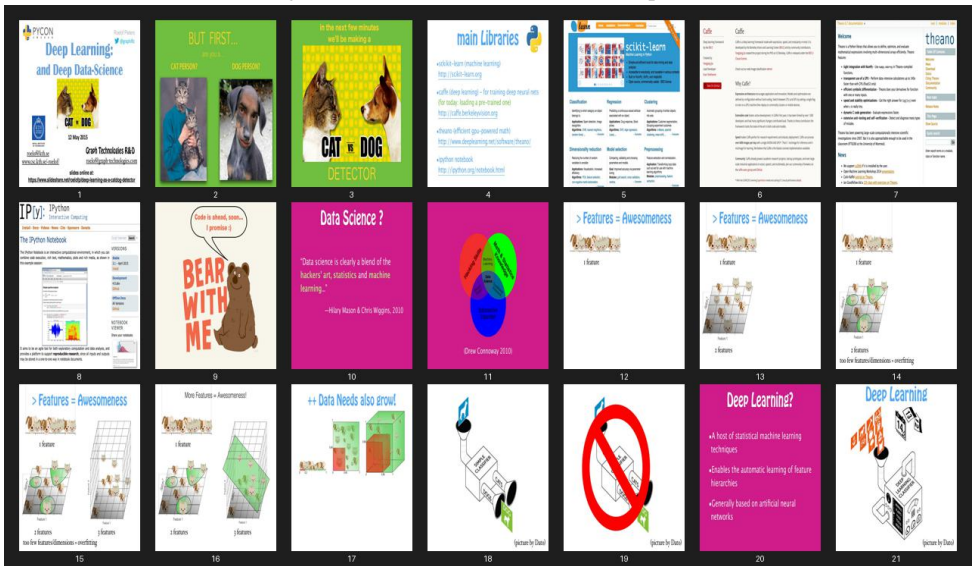
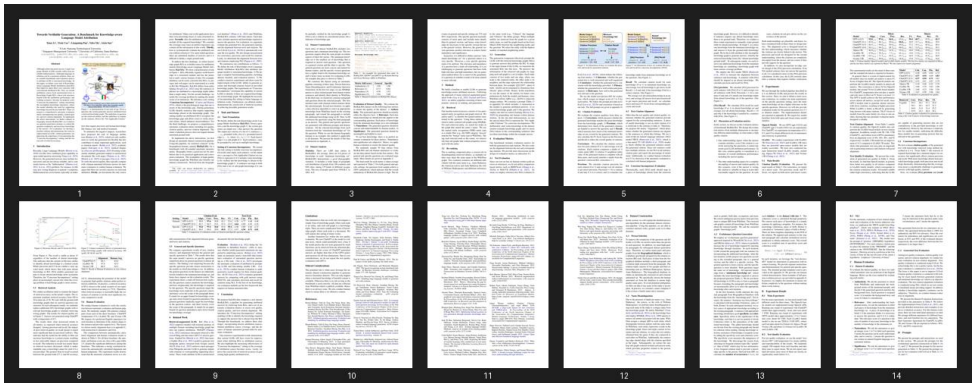
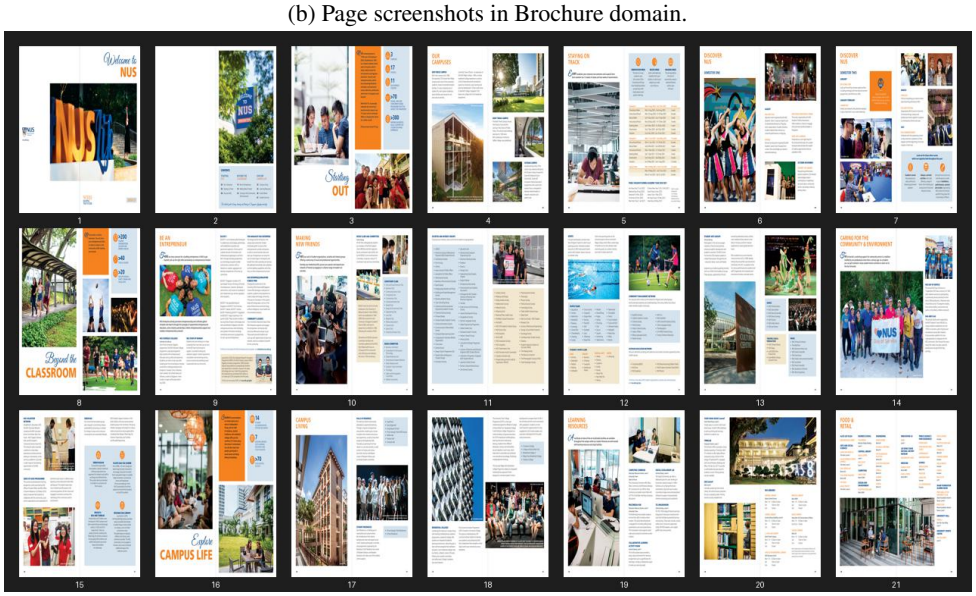


Figure 5: The screenshot examples of typical document pages for (a) Research Report, (b) Administration & Industry, and (c) Tutorial & Workshop domain.

(a) Page screenshots in Academic Paper domain.



(b) Page screenshots in Brochure domain.



(c) Page screenshots in Financial Report domain.

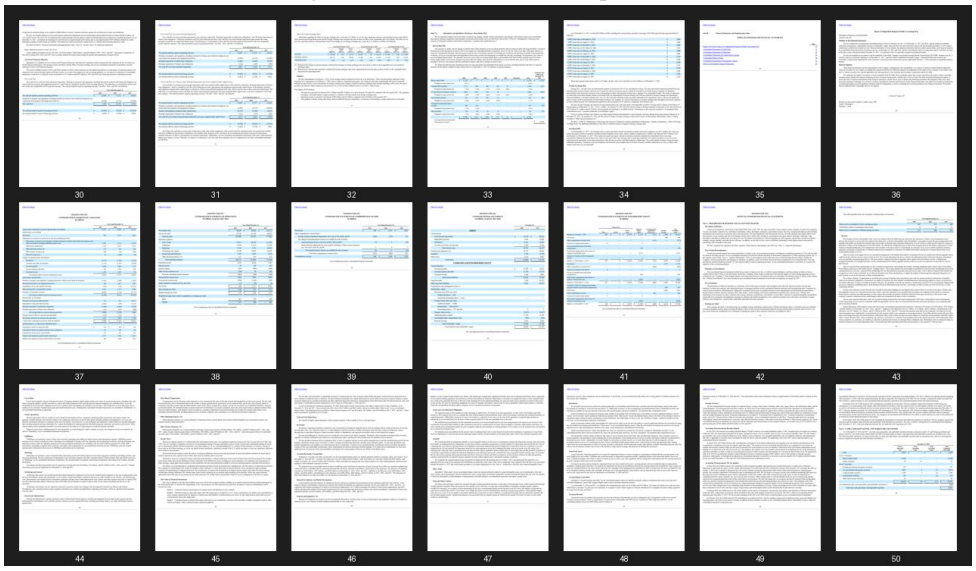
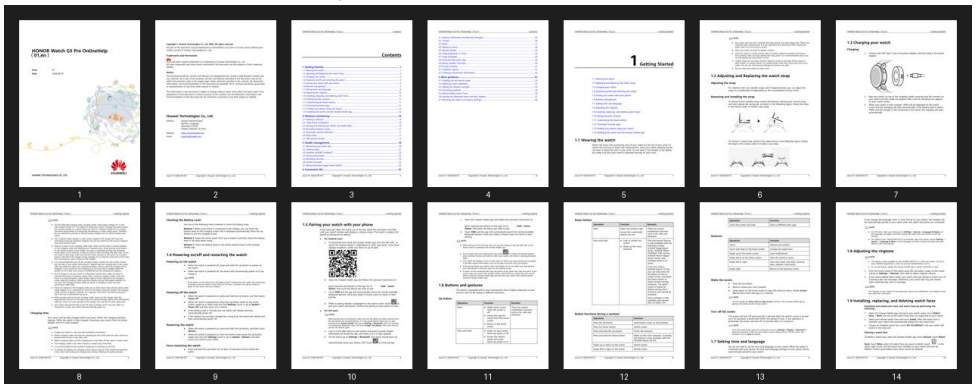
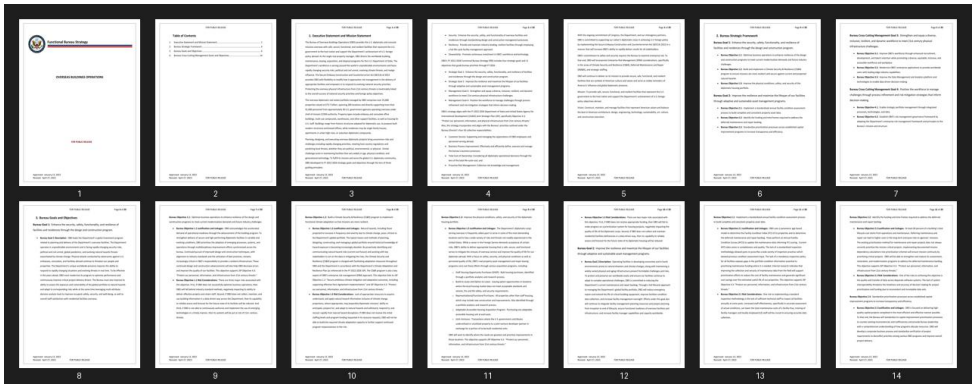


Figure 6: The screenshot examples of typical document pages for (a) Academic Paper, (b) Brochure, and (c) Financial Report domain.

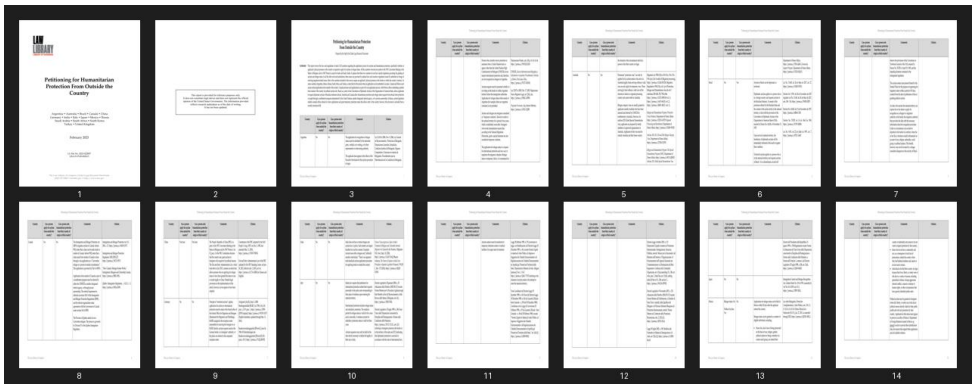
(a) Page screenshots in Guidebook domain.



(b) Page screenshots in Government domain.



(c) Page screenshots in Laws domain.



(d) Page screenshots in News domain.

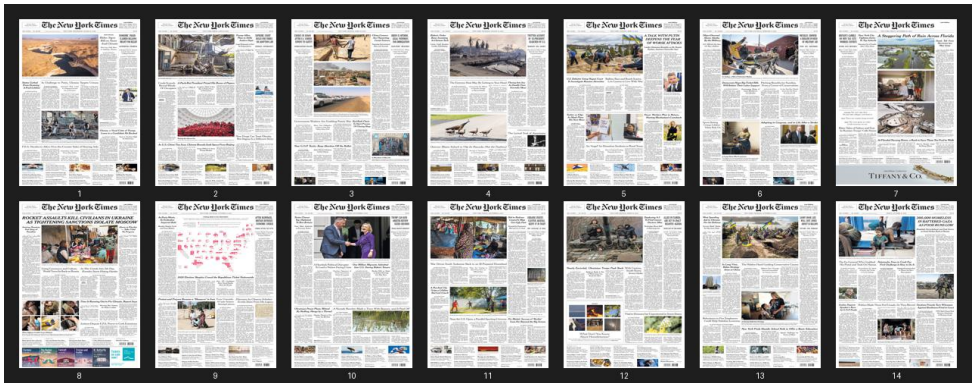
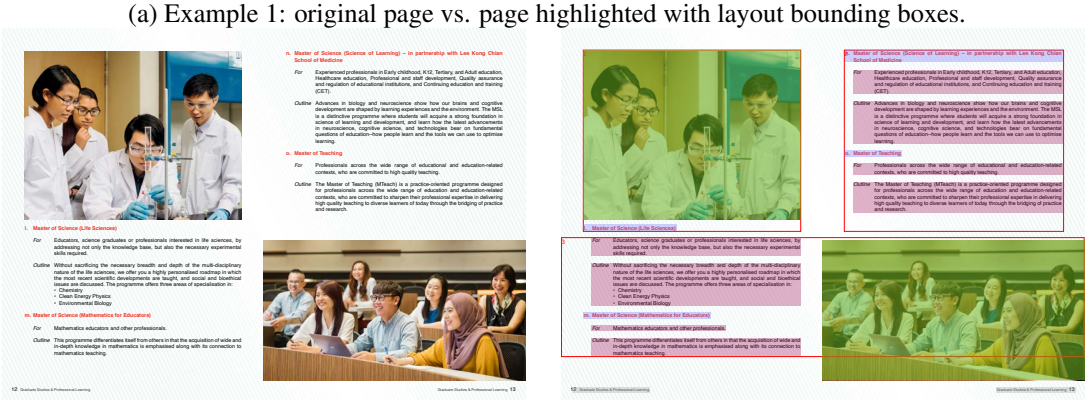
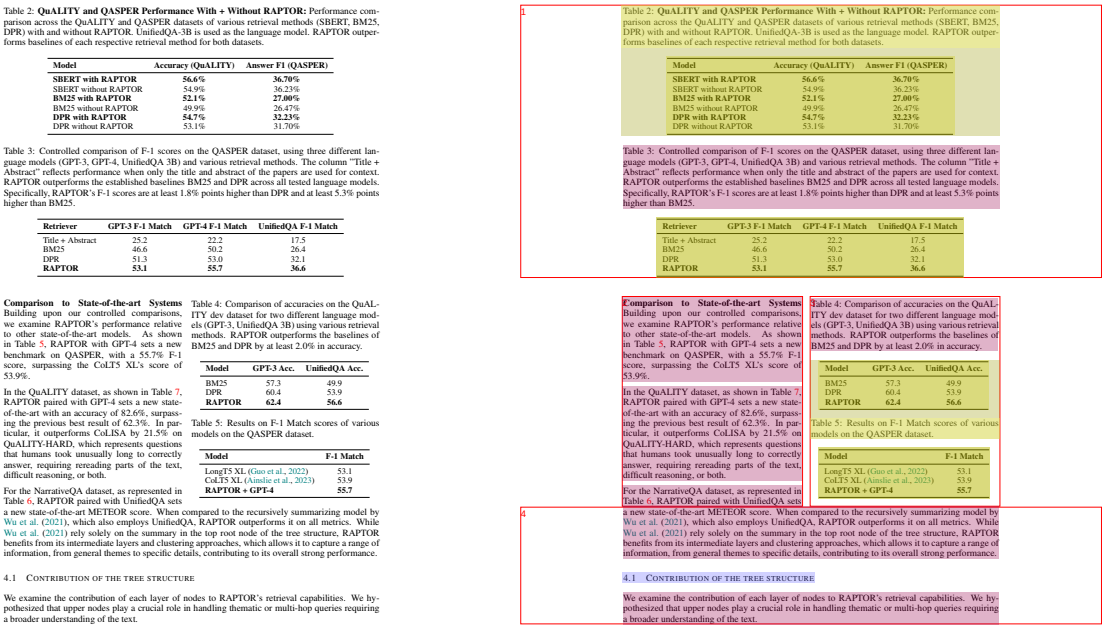


Figure 7: The screenshot examples of typical document pages for (a) Guidebook, (b) Government, (c) Laws domain, and (d) News domain.

(a) Example 1: original page vs. page highlighted with layout bounding boxes.



(b) Example 2: original page vs. page highlighted with layout bounding boxes.



(c) Example 3: original page vs. page highlighted with layout bounding boxes.

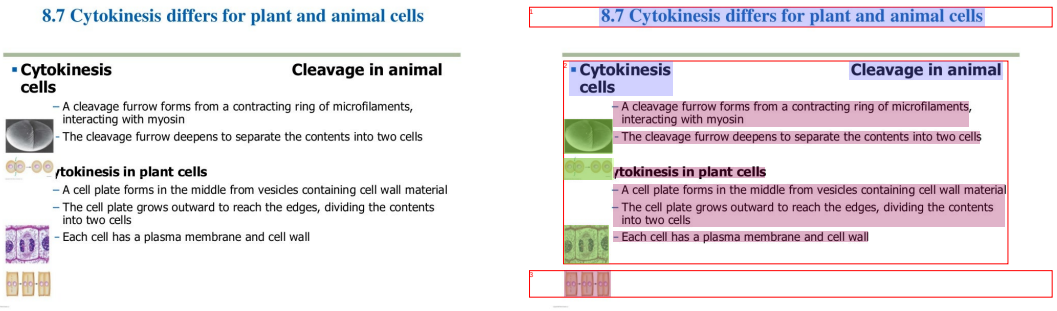
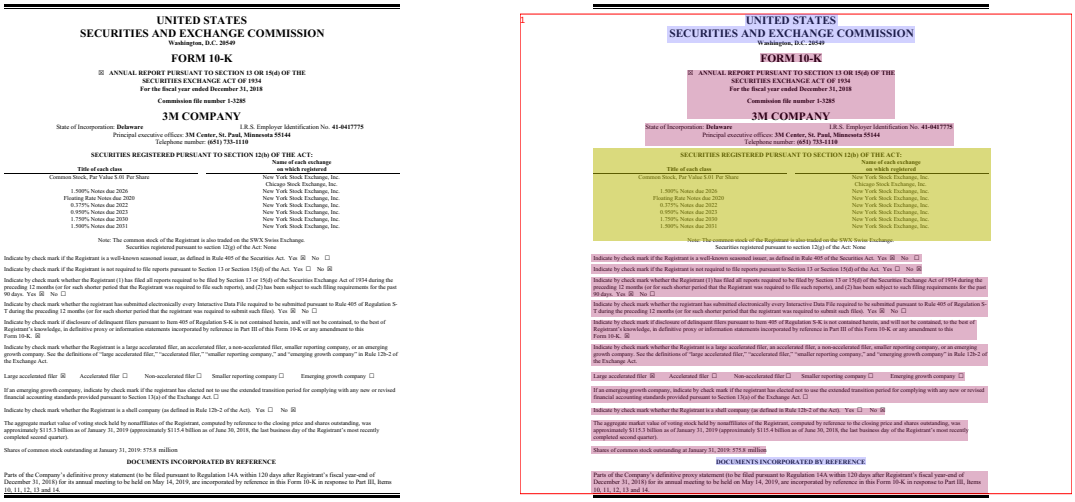
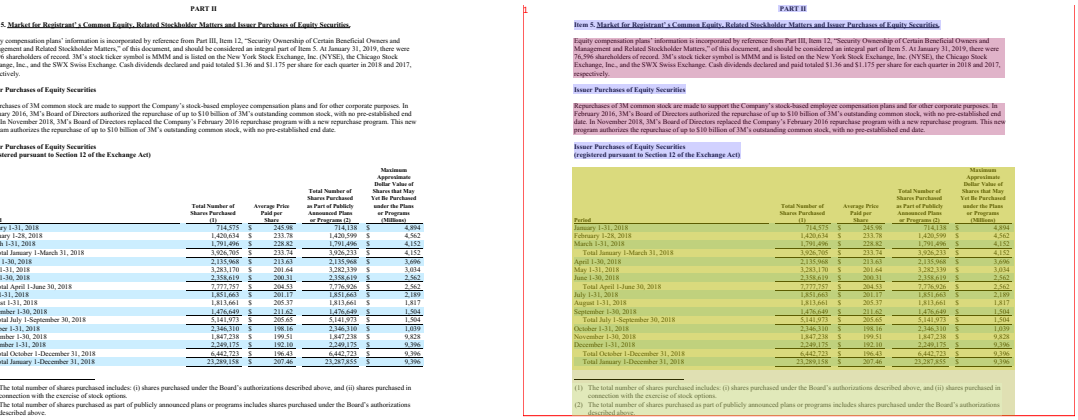


Figure 8: The 3 examples illustrate the function and effectiveness of layout detection on document pages.

(a) Example 1: original page vs. page highlighted with layout bounding boxes.



(b) Example 2: original page vs. page highlighted with layout bounding boxes.



(c) Example 3: original page vs. page highlighted with layout bounding boxes.

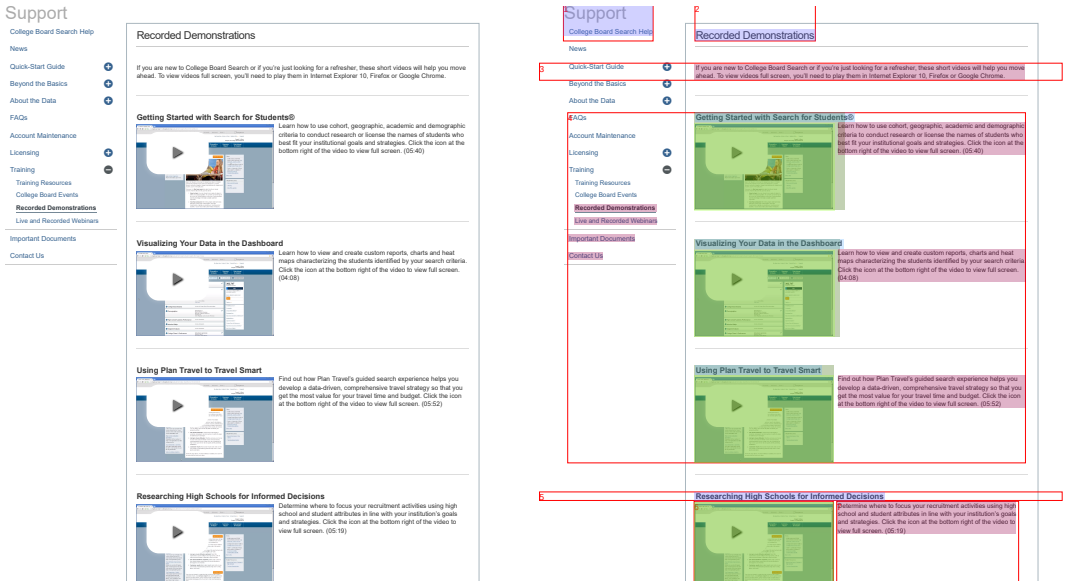
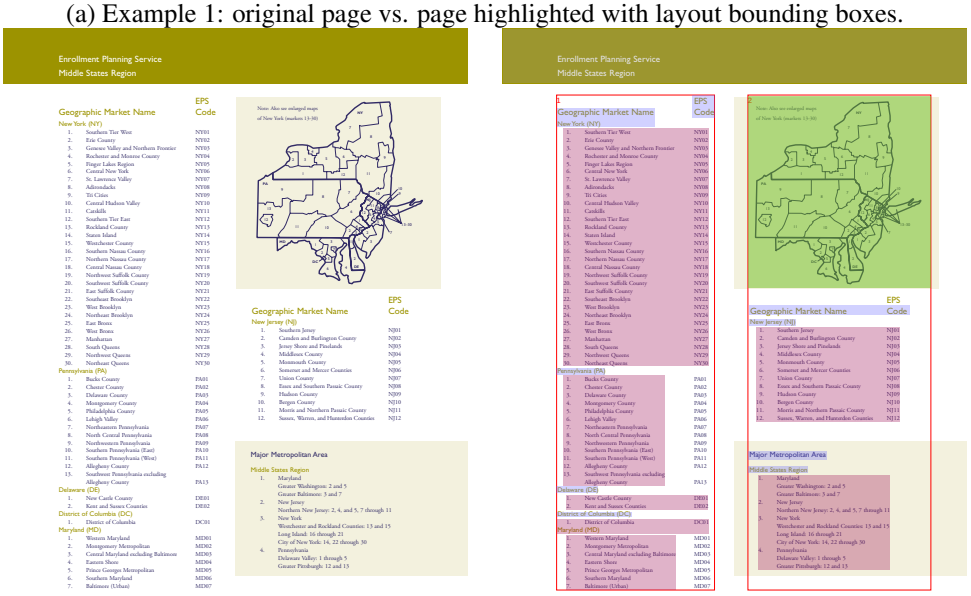
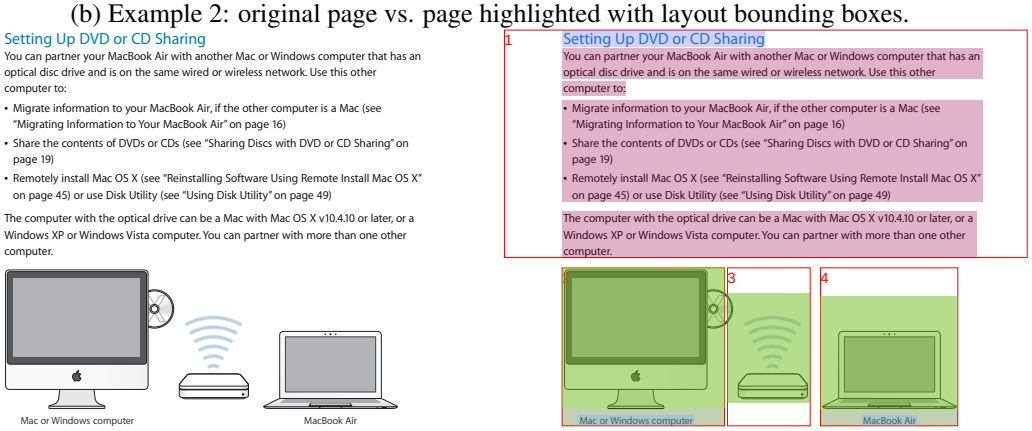


Figure 9: The 3 examples illustrate the function and effectiveness of layout detection on document pages.

(a) Example 1: original page vs. page highlighted with layout bounding boxes.



(b) Example 2: original page vs. page highlighted with layout bounding boxes.



(c) Example 3: original page vs. page highlighted with layout bounding boxes.

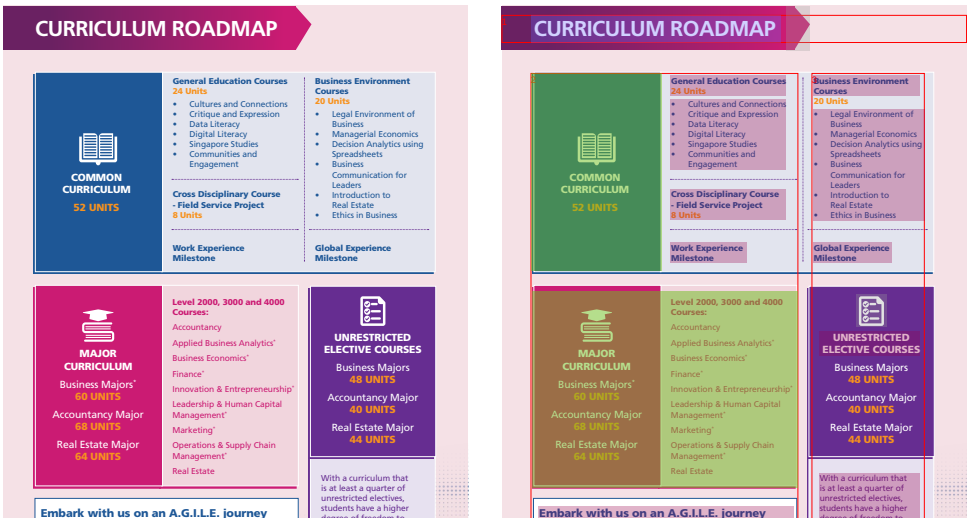


Figure 10: The 3 examples illustrate the function and effectiveness of layout detection on document pages.

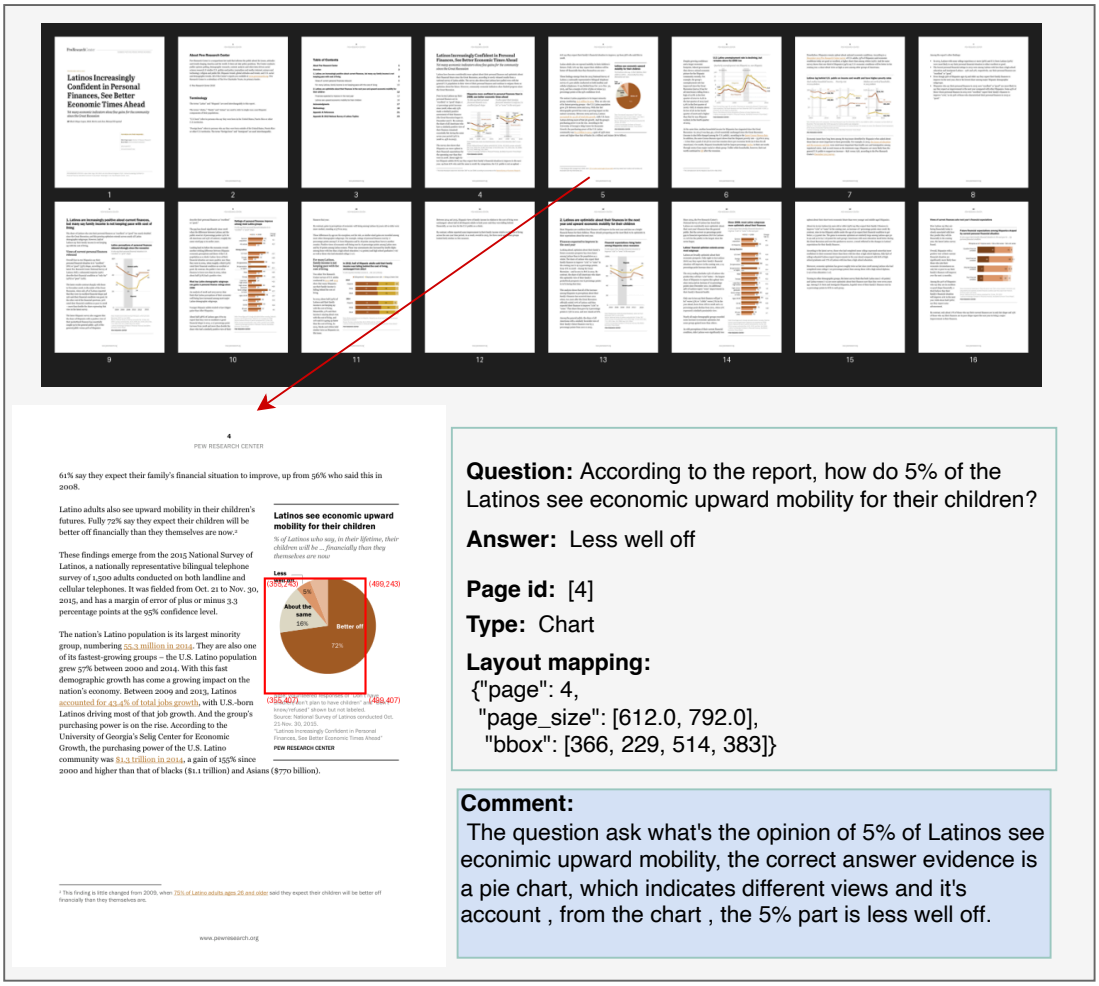
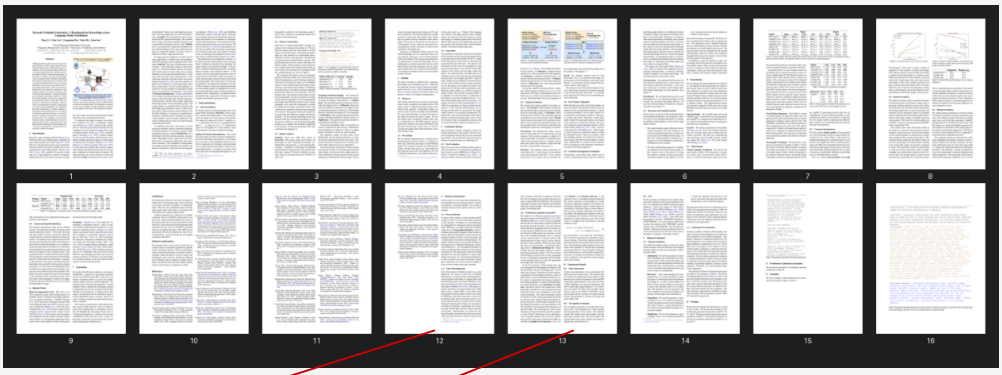
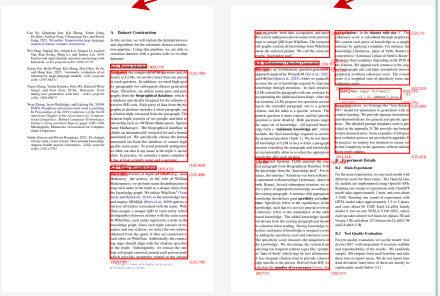


Figure 11: This example shows a typical image retrieval and reasoning task that requires synthesizing information from pie chart.





Question: For dataset construction, which step takes the most word to describe than the others

Answer: Evolutionary Question Generation

Page id: [11, 12]

Type: Text

Layout mapping:

[{"page": 11, "page_size": [595.2760009765625, 841.8900146484375], "bbox": [306, 171, 410, 184]},

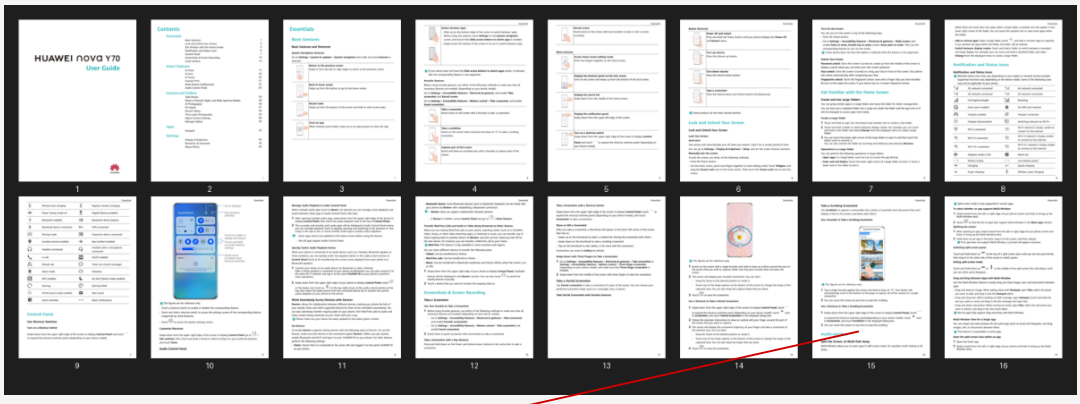
.....

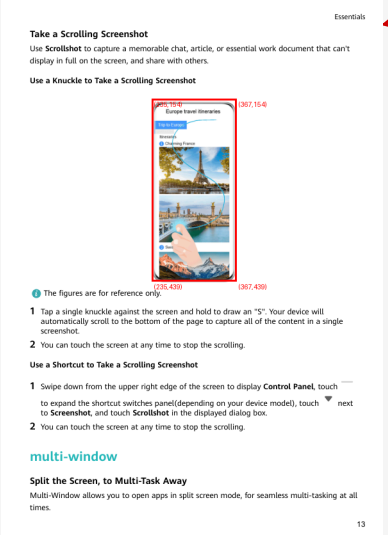
{"page": 12, "page_size": [595.2760009765625, 841.8900146484375], "bbox": [318, 242, 501, 276]}, {"page": 12, "page_size": [595.2760009765625, 841.8900146484375], "bbox": [305, 286.197021484375, 526, 434.83447265625]}]

Comment:

The question ask in data construction part, which part have most words,, first dataset construction is in the appendix , and in page12 and page13, so after contrast, the dataset Evolutionary Question Generation part have most words. From the picture ,through the layout mapping we can see every paragraph is located explicitly and compared.

Figure 12: This example shows a typical multi-page retrieval task that requires synthesizing information from text passages across multiple pages.





Essentials

Take a Scrolling Screenshot

Use **Scrollshot** to capture a memorable chat, article, or essential work document that can't display in full on the screen, and share with others.

Use a Knuckle to Take a Scrolling Screenshot

The figures are for reference only.

1 Tap a single knuckle against the screen and hold to draw an "S". Your device will automatically scroll to the bottom of the page to capture all of the content in a single screenshot.

2 You can touch the screen at any time to stop the scrolling.

Use a Shortcut to Take a Scrolling Screenshot

1 Swipe down from the upper right edge of the screen to display **Control Panel**, touch to expand the shortcut switches panel(depending on your device model), touch next to **Screenshot**, and touch **Scrollshot** in the displayed dialog box.

2 You can touch the screen at any time to stop the scrolling.

multi-window

Split the Screen, to Multi-Task Away

Multi-Window allows you to open apps in split screen mode, for seamless multi-tasking at all times.

13

Question: In the demostration of how to use a Knuckle to Take a Scrolling Screenshot, what buildings appear in the first picture?
Answer: Eiffel tower

Page id: [14]

Type: Figure

Layout mapping:
{"page": 14,
"page_size": [595.275634765625, 841.8897705078125],
"bbox": [235, 154, 367, 439]}

Comment:
The question ask in the how to use a Knuckle to take a screen shot part, what buildings appear in the first picture, in the image we ca see the first image is Eiffel tower, from the layout mapping we can see the evidence location is the right picture

Figure 13: This example shows a typical image reasoning task that requires synthesizing information from specific image.

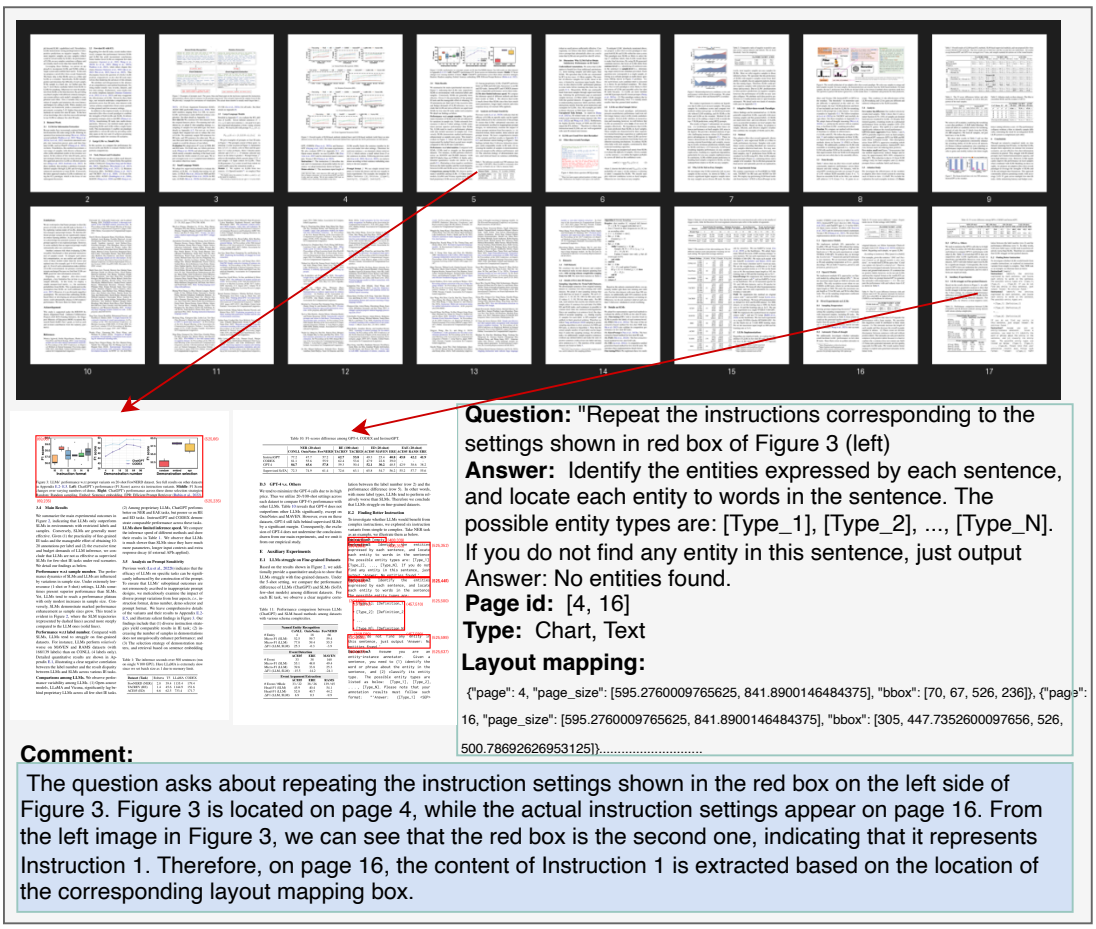


Figure 14: This example shows a typical multi-page image and text reasoning task that requires synthesizing cross-modal information from image and text.

F Detailed Analysis of ViDoRe Benchmark

F.1 Query and Annotation Analysis

As mentioned in Section 6, **ViDoRe** (Faysse et al., 2024) is the most relevant benchmark to MMDocIR. It integrates several datasets such as DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), TAT-DQA (Zhu et al., 2022), arXiVQA (Li et al., 2024), and providing new documents in scientific, medical, administrative, and environment domains. In this Appendix, we elaborate our analysis of the sampled 2,400 questions sampled from ViDoRe. The statistics are shown in Table 14. ViDoRe test set contains questions in either English or French. In our work, we examine only the English questions. For academic subsets, we examine all 1,500 questions: 500 questions from DocVQA, 500 questions from InfoVQA, and 500 questions from arXiVQA. In TAT-DQA, we sample and examine the first 500 questions. For the industrial documents, we select 100 questions from each domain (*i.e.,* energy, healthcare, government, and artificial intelligence). We examine sampled questions and summarize these questions into 3 categories:

- **Unsuitable Queries.** Queries that are not well-suited for IR systems can often burden these systems by generating numerous irrelevant results. For example, a query such as "*What's the x-axis of the figure*" is likely to prompt matches from multiple passages within document corpora that mention figures with an x-axis. This tends to happen because the query is overly broad and lacks contextual specificity. When such queries stem from Document Visual Question Answering (DocVQA) tasks targeting a single image, the challenge is exacerbated, as the reliance on precise context increases while the target remains too vague, undermining the fundamental principles of effective IR.

- **Barely Suitable Queries.** Queries that fall into this category provide some guidance towards locating useful passages, yet suffer from a lack of precise detail. These queries often fetch moderate number of passages, where both relevance and focus may not be as sharp. For example, the query "*What was the total assets from AMER in 2018?*" is meant for Visual Question Answering (VQA) focused on a specific financial topic. Although this seems specific, the issue arises when multiple sections within an annual report discuss AMER's total assets for the year. This causes significant confusion since ViDoRe is set to acknowledge only a single passage as the verified answer. This lack of uniqueness in the ground truth makes it hard to evaluate the actual performance of IR system.

- **Suitable Queries.** The most effective queries for IR systems are characterized by their specificity and ability to distinguish between different sections of texts. These queries often involve precise facts or detailed inquiries that facilitate pinpointing exact passages. For instance, the question "*What was the magnitude of the earthquake that occurred in Maule on 2/27/2010?*" incorporates significant keywords and details that guide the retrieval system directly to the necessary data. Such queries align perfectly with the objectives of IR, leveraging specificity and detailed context to efficiently retrieve most relevant information.

The comprehensive analysis of our queries, as presented in Table 14, reveals a significant challenge in adapting questions from Visual Question Answering (VQA) datasets (such as DocVQA, InfoVQA, TAT-DQA, and arXiVQA) for Information Retrieval (IR) purposes. Only 8% of these queries prove suitable for effective IR usage. In comparison, queries derived from industrial documents perform slightly better, with 15.5% deemed suitable. A common issue identified is that these queries are either excessively simplistic or highly specific to a particular context. Our findings suggest that the primary difficulty stems from the inherent differences between DocIR and DocVQA. VQA queries are typically crafted to address content on a specific page or within a particular image, inherently limiting their scope and specificity. This specificity and simplism are functional within the confines of the intended VQA context but pose substantial limitations when such queries are repurposed for IR tasks. Due to this gap, we exclude ViDoRe benchmark from our experiments.

F.2 Document Corpora Analysis

ViDoRe bootstrap document corpora directly from existing DocVQA benchmarks (*i.e.,* DocVQA, InfoVQA, TAT-DQA, arXiVQA) that perform single-page VQA. In the DocVQA setting, only selected pages are provided for VQA, rather than the entire document pages. For arXivQA, the retrieved passages are not document pages, but are cropped

Sub dataset	#Not suitable	#Barely suitable	#Suitable	#Total	HuggingFace Resource
arXiVQA	245	203	52	500	vidore/arxivqa_test_subsampled
DocVQA	345	130	25	500	vidore/docvqa_test_subsampled
InfoVQA	139	284	77	500	vidore/infovqa_test_subsampled
TAT-DQA	373	121	6	500	vidore/tatdqa_test
Industrial	78	260	62	400	vidore/syntheticDocQA_energy_test vidore/syntheticDocQA_healthcare_industry_test vidore/syntheticDocQA_government_reports_test vidore/syntheticDocQA_artificial_intelligence_test
Sum	1180	998	222	2,400	-
Percentage	49.1%	41.5%	9.25%	-	-

Table 14: Document statistics for ViDoRe Benchmark.

images (*e.g.,* figures, tables, and charts). In our experiments, we need to rely on the entire documents pages to evaluate retrieval on long documents. To bridge the gap of missing complete document corpora, we put in considerable efforts to collect the original documents of existing DocVQA datasets, as mentioned in Section 4.1.

G License Agreements

We ensure that the distribution of each dataset complies with the corresponding licenses, all of which are listed below:

- MMLongBench-Doc: is under Apache-2.0 license agreement for academic research purposes.
- DocBench: we achieved the agreement of usage as academic research from the dataset's author.
- MP-DocVQA: is under "MIT License" license agreement for academic research purposes.
- SlideVQA: is under "NTT License" license agreement for academic research purposes.
- TAT-DQA: is under "CC-BY-4.0" license agreement for academic research purposes.
- ArXivQA: is under "CC-BY-SA-4.0" license agreement for academic research purposes.
- SciQAG: is under "CC-BY-4.0" license agreement for academic research purposes.
- DUDE: is under "GPL-3.0" license agreement for academic research purposes.
- CUAD: is under "CC-BY-4.0" license agreement for academic research purposes.

For the new annotations contributed in MMDOCIR, including but not limited to the questions, page and layout annotations, we make them available solely for research purposes. Users are permitted to use, modify, and share these annotations for academic and non-commercial research activities. Any other use, including commercial exploitation, is not permitted without explicit written permission from the authors.

H Ethical Considerations

The introduction and broader adoption of MMDOCIR may have potential ethical impacts spanning both positive and negative dimensions. Below, we outline possible negative consequences and discuss potential mitigation strategies:

Privacy Risks: MMDOCIR enables models to retrieve relevant information over lengthy, multimodal documents, which may include sensitive personal, financial, or health information. There is a risk that such technologies could be leveraged for large-scale surveillance, unauthorized extraction of personal data, or other privacy violations.

Fairness and Bias: If benchmarked models are trained or evaluated on data that does not reflect diverse demographic, linguistic, and backgrounds, outputs may exhibit biases. This may lead to unfair decision-making or stereotypes.

Mitigation Strategies: To mitigate these risks, we make sure that: (i) Benchmark development uses only publicly available, carefully vetted datasets, with sensitive information anonymized or removed; (ii) Retrieval outputs are monitored for bias and fairness.

We encourage researchers and practitioners employing MMDOCIR to be mindful of these factors and to actively work toward responsible development and deployment, including transparency about limitations and proactive safeguards where needed. We welcome community feedback and collaboration on best practices to further reduce risks as this technology evolves.