DocQIR-Emb: Document Image Retrieval with Multi-lingual Question Query

Chih-Hui Ho Giovanna Marinho Felipe Viana
Varad Pimpalkhute* Rodolfo Tonoli Andre Von Zuben
Articul8.ai

Abstract

Document image retrieval is a fundamental task for improving document understanding, where the goal is to retrieve relevant images in the document and to answer the question from the user. Unlike other text-to-image tasks, which mainly focus on the alignment between image caption and natural image, document image retrieval requires the model to understand the question from user and return related table image or scientific image. The significant domain difference between image caption and user question, as well as natural image and scientific images, prevents the off-the-shelf retrieval model from becoming applicable. To systematically study the degradation, we curate a novel multi-lingual Document Question-Image Retrieval benchmark, DocQIR, that covers questions in 5 different languages. Our preliminary study shows that off-the-shelf retrieval models fail to retrieve documents images when questions in various languages are presented. To address this issue, we proposed a novel architecture, DocQIR-Emb, that leverages a multi-lingual text embedder and a VLM to encode a question and an image into a shared feature space. Since the multi-lingual embedder is trained to align text in different languages, the text embedder is frozen and only the VLM is optimized. Experiments show that DocQIR-Emb outperforms the baseline by at least 40% on the proposed DocQIR dataset and the gain is consistent across table image and scientific image. Different architecture designs are also ablated to demonstrate the effectiveness of DocQIR-Emb.

1 Introduction

Answering questions from complex documents has been a long-established research topic, which requires a system to first comprehend the query from a user and then leverages the information in the document to answer the query. A popular approach is to first ingest the documents and extract entities like text chunks, tables, images and other information from the documents, as shown in Figure 1. Then, relevant entities are retrieved as a supporting context to answer the query. This pipeline has led to a series of document understanding tasks in the literature, including document layout analysis [62, 56, 49, 6], document retrieval [16, 57, 58, 64], structure table parsing [28, 45, 55, 39] and optical character recognition (OCR) [34, 51]. Despite the advance of these techniques, retrieving relevant images from the document that contains the answer to an user's query remains challenging and under-studied. In this work, we refer to this problem as Question-Image Retrieval (QIR), which is a fundamental task to allow the user to ask question about an image.

While the problem of QIR shares the same fundamental concept of text-to-image (T2I) retrieval [43, 7, 24, 41], where the goal is to retrieve images that matches the input text query, we hypothesize most of the techniques developed for T2I retrieval are not applicable. This is due to the manner that existing T2I methods are trained and how the T2I datasets are curated. First, despite the large

*Work done in Articul8.ai

Preprint.

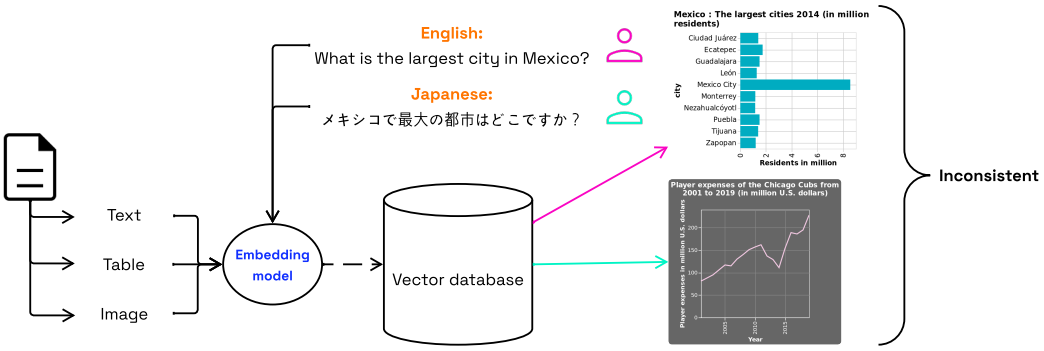

Figure 1: Illustration of Question-Image Retrieval (QIR) task. A QIR system aims to retrieve relevant scientific images or table images that accurately answer a user's question, while also being robust to queries in multiple languages, enabling effective information retrieval across linguistic boundaries.

number of T2I datasets, most of these datasets uses image description as text query and the models [43, 48, 37, 8, 10] trained on these datasets only have decent performance when the input query is an image description. For example, it is known that the performance of CLIP-[43] based methods degrades when the query contains questions or negations [38, 42, 35]. This limitation hinders the practical use cases where the user inserts a question (e.g. "What is the largest city in Mexico?"), as shown in Figure 1. In addition, since most of the existing T2I datasets contain less scientific images (e.g. bar charts and flow charts), existing methods trained on these datasets underperform on scientific images, which have a significant visual difference than natural images. This indicates that these existing methods cannot be used in tasks related to document analysis, which contains many scientific images. Furthermore, in practice, the user can ask questions in different languages about the same scientific image and the QIR system should be multi-lingual as in Figure 1. This poses a more challenging problem to the existing T2I solutions as most methods are not multi-lingual.

In this work, we tackle the challenges of QIR in document understanding by introducing a novel multi-lingual QIR dataset, dubbed DocQIR. This dataset comprises 500K question-image pairs for training and 2,444 question-image pairs for evaluation, with a unique characteristic: all images are scientific images and table images, and all text queries in the test split are questions in 5 different languages. To support queries in multiple languages, we propose a novel multi-modal embedder, DocQIR-Emb, which consists of a visual embedder and a multi-lingual text embedder. Specifically, we leverage an off-the-shelf multi-lingual text embedder to encode the query into a text embedding, allowing us to decouple the problem of mapping positive question-image pairs into the same embedding space from the challenge of mapping text in different languages into a shared embedding space. For image encoding, we adopt a visual language model (VLM) as the backbone pre-trained on numerous visual language tasks. We pass the image and a fixed prompt ("What is shown in the image?") to the VLM and utilize the output embedding associated with the last input text token to present the image content. The resulting embedding is then projected to ensure dimensionality alignment with the text embedding, enabling effective multi-modal retrieval.

Our experiments demonstrate that the proposed DocQIR-Emb model achieves significant improvements over the baseline, with a minimum gain of 40% on the multi-lingual question image retrieval task. More importantly, this performance gain is consistent across both the scientific image subset and the table image subset of the DocQIR dataset, highlighting the model's robustness and generalizability. To further validate the design choices underlying DocQIR-Emb, we conduct an extensive ablation, examining the impact of its architectural components and training dataset on its performance.

Overall, this work makes three key contributions to the field of document understanding. Firstly, we formalize the Question-Image Retrieval (QIR) setting, a crucial component of document understanding that has been largely overlooked, accompanied by an analysis of current limitations of existing datasets and models - namely the visual context mismatch between natural images and document-specific images, as well as the lack of support for multi-lingual user queries. Secondly, to address these gaps, we propose a novel DocQIR dataset and a specially designed multi-modal embedder, DocQIR-Emb, tailored to the QIR task. Finally, our extensive experiments demonstrate that DocQIR-Emb significantly outperforms existing baselines by a substantial margin, showcasing

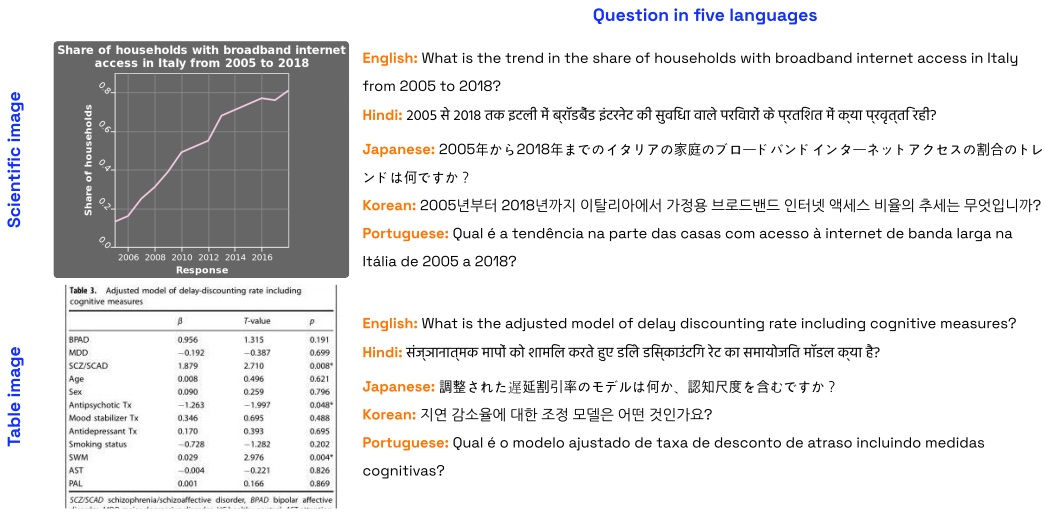

Figure 2: Example question in 5 different languages associated with (top) scientific image and (bottom) table image. Please refer to appendix for more examples.

its effectiveness in retrieving relevant images for queries in various languages and establishing a new state-of-the-art in QIR for document understanding.

2 Related Work

2.1 Document Question Answering

Document Question Answering is a specialized task that involves answering natural language questions about document pages. Unlike traditional text-based QA systems, document QA models typically leverage multi-modal features, combining textual content, positional information of words, and visual elements of documents. This approach enables systems to understand both the semantic content and the visual structure of documents, making it particularly effective for complex document types. A popular pipeline is to first extract text [34, 51] and layout [62, 56, 49, 6], search for relevant context related to the question and pass the retrieved context and the question to the model for generating the response. Another line of research streamlines the pipeline by treating each document page as an image and retrieve the relevant document page related to the query [16]. The retrieved document page and the query are then passed to the visual language model (VLM) [11, 3, 5, 61] to perform the Document Question Answering. Despite the latter approach reduce the effort of parsing the document, this pipeline cannot generalize to other tasks, such as extracting images from documents. Please refer to the Document Question Answer survey paper [4, 14] for more details. In this paper, we focus on the question-to-image retrieval in the first pipeline.

2.2 Text-Image Retrieval

Text to image retrieval has wide applications where the goal is to retrieve images relevant to the input text. One of the most prevalent research direction is the CLIP-based method [43]. CLIP based methods have dual encoder that encode the text and visual input, respectively, and produces the text and image embedding. For the visual towers, CNN based encoder and ViT are commonly adopted. The former (i.e. ResNet [18]) converts the image into grid latent representation and a pooling layer is applied on the grid representation to obtain the image embedding, while the later use the classification token of the vision transformer as the image embedding. For the text encoder, encoder based model such as BERT [13] and T5 [44] are used and the embedding associated to the classification token is used to represent the input text. Both the visual encoder and the text encoder are usually pretrained on the large multi-modal corpus (i.e. LAION-5B [46] and Datacomp [17]). To train the text and visual encoder, contrastive loss function (e.g. InfoNCE loss) are optimized, where text and image embeddings from positive pairs are pull together and vice versa. Despite the

Table 1: Comparison between the proposed DocQIR dataset with existing datasets. The first block shows dominant T2I datasets with natural images and captions. The second block features datasets with questions, but none focus on scientific or table images, making DocQIR unique in this domain.

Dataset	Task	Text Type	Image Type	Language
CC3M [48]	Retrieval	Caption	Natural Image	English
LAION-5B [47]	Retrieval	Caption	Natural Image	English
Encyclopedic VQA [33]	VQA & Retrieval	Question	Natural Image	English
ViQuAE [25]	VQA & Retrieval	Question	Object Entity	En
OVEN [19]	VQA	Question	Natural Image	En
OK-VQA [31]	VQA	Question	Natural Image	En
Vidore [16]	VQA & Retrieval	Question	PDF page	En, Fr
M-LONGDOC [12]	VQA & Retrieval	Question	PDF page	En
REAL-MM-RAG [59]	VQA & Retrieval	Question	PDF page	En
Wiki-SS [29]	Retrieval	Question	Web screenshots	En
SlideVQA [53]	VQA	Question	Slide	En
VISA [30]	Retrieval	Question	PDF page	En
DocQIR	Retrieval	Question	Scientific Image & Table Image	En, Ja, Ko, Pt, Hi

zero-shot generalization of CLIP based methods on different visual domains, it is known that the performance of these methods have short context length (typically 77 tokens for CLIP) and tends to degrade when the input text contains negation [38, 42, 35]. To address these issue, LongCLIP [60] increases the maximum input length of CLIP from 77 to 248, by performing non-linear positional embedding interpolation. JinaCLIPv2 [22] retrained a text embedder that support long context, while LLM2CLIP [21] adopts the existing LLM models with large context length. Both the text embedding extracted from JinaCLIPv2 and LLM2CLIP then aligned with visual encoder. Unlike prior works, we proposed a novel architecture that leverages VLM and multi-lingual text embedder, and show that such design outperform prior works.

2.3 Visual Language Model

Visual language model endows the large language model (LLMs) comprehending the visual input (i.e. image, video and etc.) and performing visual question answer (VQA). One of earliest works is LLavA [27], which connects the visual domain and the language domain with a projector. More specifically, the projector takes the input image and converts it to a set of visual embeddings that lie on the shared space as the text embeddings. Over the past few years, series of VLM are presented in the literature, such as Idefics3 [23], InternVL [11], QwenVL [3] and etc. These proposed methods advances the VLM from different perspectives. For example, Llama4 proposed to use a visual projector with a mixture of expert LLM as the VLM design. Idefics3 [23] and Llama2 proposed to use cross attention to connect the visual input and the LLM. InternVL and QwenVL series advance the visual encoder dynamic resolution design, so there are less limitation on the image resolution and supports multi-image computation. Please refer to the recent VLM survey [61, 26] for more details.

While VLM are mostly used for VQA, DSE [29] is one of the pioneer works that proposed to use VLM to extract visual embeddings. Unlike traditional visual encoder, such as ResNet [18] or ViT [15], using VLM to extract image embedding tends to preserve the semantic representation of the image, as it is already well aligned with the LLM. Unlike DSE, which uses the VLM to encode both the text and image, we decouple the text encoder from the VLM and show that our framework is able to support multi-lingual question image retrieval and outperforms prior works.

3 Question Image Retrieval Dataset

Table 1 presents a comprehensive overview of existing text-to-image datasets, highlighting their characteristics and limitations. The first block of the table illustrates the dominant type of text-to-image pretraining datasets, which predominantly consist of natural images as the visual source and image captions as the textual source. This category encompasses a plethora of datasets, including but not limited to [8, 48, 47, 17, 26], which have been widely utilized for pretraining text-to-image

models. Due to space constraints, we only provide a selection of popular datasets from this category, and refer the reader to a recent survey paper [26] for an exhaustive review.

In contrast, the second block of the table highlights datasets characterized by the use of questions as the textual source. However, upon closer inspection, it becomes apparent that none of these datasets specifically focus on scientific images and table images, which are the primary targets of our dataset. While there exist some document-related datasets, such as SlideVQA [53] and Vidore [16], our experiments demonstrate that multi-modal embedders, including ColPali [16], trained on these datasets do not generalize well to scientific images and table images. Specifically, as detailed in Section 5, our experiments reveal that the performance of these embedders degrades significantly when applied to our target domain, underscoring the need for a specialized dataset.

To overcome these limitation, we propose the DocQIR dataset by leveraging the existing datasets on scientific images and table images, which provide a rich source of visual content for our task. More specifically, we sample 280,000 images from Pubtables1M [50], 32,719 images from ChartQA [32], 44,285 from ChartVE [20] and 147,796 images from SPIQA [40], to curate around 500K training examples. For each image, we utilize the existing visual language models [3, 5, 1], to generate a question about the image. This approach allows us to create a large-scale dataset of image-question pairs, where each question is designed to probe the visual content of the corresponding image.

Furthermore, we only include questions in English in the training split of DocQIR. This design choice is motivated by our findings, which suggest that training with multi-lingual questions does not provide significant benefits in terms of performance (see Section 5 for a detailed analysis). By focusing on a single language, we can simplify the training process of our multi-modal embedder, as we can decouple the challenging task of aligning multiple languages with the already complex task of aligning visual and textual modalities. This simplification enables us to focus on optimizing the performance of multi-modal retrieval, rather than on handling linguistic variations.

For the testing split of DocQIR, we have created a multi-lingual testing split that allows for a more nuanced analysis of model generalizability and robustness. Specifically, we use LLM [54] to translate the English version of the questions into four additional languages, namely Hindi, Japanese, Korean, and Portuguese. The testing split of DocQIR is further divided into two distinct subsets, each containing a specific type of visual content. The first subset consists of 1,444 scientific images and the second subset comprises 1,000 table images. Figure 2 shows an example for each subset with questions from different languages. By evaluating model performance on these two subsets separately, we can gain insights into the performance of different models on different types of visual data.

4 Method

In this section, we proposed a multi-lingual multi-modal embedder for the task of document QIR. We refer the proposed framework as DocQIR-Emb in the following.

4.1 Architecture

Given a dataset $D = \{(q_i, v_i)\}_{i=1}^{N}$ comprising pairs of question q_i and image v_i, our proposed framework, DocQIR-Emb, aims to generate a dense text embedding and a corresponding image embedding for each (q_i, v_i) pair. To facilitate multi-lingual text input and enable the model to effectively process queries in various languages, we leverage an off-the-shelf multi-lingual text embedder $\mathcal{T}$, specifically the BGE-m3 model [9], to encode the question q_i, resulting in the text embedding $e_i^q = \mathcal{T}(q_i)$. This design choice is motivated by the fact that pre-trained multi-lingual text embedders have already learned to map semantically equivalent text in different languages to a shared feature space, thereby reducing the complexity of aligning multi-modal and multi-lingual embeddings during training. Furthermore, our framework is designed to be flexible, allowing for seamless integration with existing multi-lingual text embedder, making it a plug-and-play solution that can be easily adapted. By doing so, we can focus on optimizing the image embedding and the overall framework, while leveraging the strengths of pre-trained language models, and avoiding the need for extensive retraining or fine-tuning of the text embedder.

To effectively encode the visual information present in an image v_i, we leverage a pretrained vision-language model (VLM) $\mathcal{V}$ (i.e. Qwen2-VL). Most VLM has been extensively pretrained on a large-scale image-text corpus, encompassing a diverse range of downstream tasks such as OCR, VQA

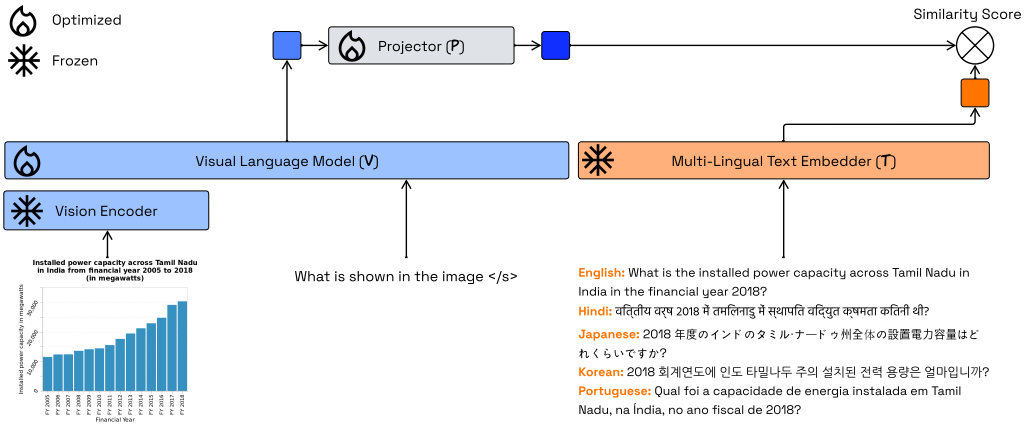

Figure 3: Architecture of the proposed DocQIR-Emb.

and image captioning. The pretraining process enables $\mathcal{V}$ to develop a robust understanding of the intricate relationships between visual and text modalities.

To encode the image v_i, we employ a prompt p with the text "What is shown in the image </s>" and pass it along with the image to $\mathcal{V}$. Rather than generating a descriptive text for the image, we extract the output embedding of the last token of the prompt, specifically the </s> token of the prompt p, to represent the image description. The image embedding can then be computed as $e_i^v = \mathcal{P} \circ \mathcal{V}(p, v_i)$, where $\mathcal{P}$ is a projector that maps the embedding output from VLM to the same dimension as the text embedding. After embedding the input image v_i and question q_i, the dataset then becomes $D = \{D_v, D_q\}$, where $D_v = \{(e_i^v)\}_{i=1}^N$ and $D_q = \{(e_i^q)\}_{i=1}^N$. Note that the prompt p remain fixed for all images and e_i^v and e_i^q are normalized.

4.2 Optimization

To train the VLM and the projector in DocQIR-Emb, we sample batch size of B from the dataset D for each gradient update, where $\hat{D}_v = \{(e_i^v)\}_{i=1}^B$ and $\hat{D}_q = \{(e_i^q)\}_{i=1}^B$ denotes the sampled batch. For the sampled image embeddings $\hat{D}_v$, we compute the InfoNCE loss with respect to the image embeddings as shown in Equation 1, where τ is the is a temperature hyperparameter that controls the sharpness of the distribution and is set to 0.01 in our experiments.

$$L(\hat{D}_v, \hat{D}_q) = -log \sum_{i=1}^{B} \frac{exp(e_i^{vT} e_i^q)/\tau}{\sum_{j=1...B, j\neq i} exp(e_i^{vT} e_j^q)/\tau}, \tag{1}$$

Similarly, we can compute the InfoNCE loss with respect to the question embeddings as shown in Equation 2.

$$L(\hat{D}_q, \hat{D}_v) = -log \sum_{i=1}^{B} \frac{exp(e_i^{qT} e_i^v)/\tau}{\sum_{j=1...B, j\neq i} exp(e_i^{qT} e_j^v)/\tau}. \tag{2}$$

The Equation 3 shows the overall loss function.

$$L = \frac{1}{2} * (L(\hat{D}_v, \hat{D}_q) + L(\hat{D}_q, \hat{D}_v)). \tag{3}$$

During training, we only optimize the VLM and the projector, as shown in Figure 3. During inference, the proposed DocQIR-Emb supports question in different languages and produce a text embedding, which aligns with image embedding associated with the targeted retrieved image.

5 Experiments

In this section, we discuss experiments for evaluating the effectiveness of DocQIR-Emb. All experiments are conducted on a single node 8 Nvidia A100 GPU, using Pytorch [2].

Table 2: QIR result on the **table image** subset of the DocQIR test set.

Method	Mean RR@1					Mean RR@5				
	En	Hi	Ja	Ko	Pt	En	Hi	Ja	Ko	Pt
MegaPairs [63]	0.089	0.009	0.012	0.009	0.060	0.121	0.013	0.021	0.018	0.076
Nomic-v1.5 [36]	0.030	0.008	0.008	0.008	0.017	0.048	0.011	0.013	0.015	0.027
JinaClip-v2 [22]	0.226	0.133	0.170	0.159	0.210	0.273	0.176	0.223	0.213	0.261
LongCLIP [60]	0.334	0.032	0.053	0.039	0.116	0.388	0.040	0.072	0.056	0.143
LLM2CLIP [21]	0.278	0.119	0.205	0.213	0.247	0.339	0.157	0.269	0.264	0.316
ColQwen2 [16]	0.556	0.128	0.362	0.290	0.409	0.620	0.163	0.439	0.353	0.476
ColPali-v1.3 [16]	0.498	0.167	0.338	0.268	0.390	0.558	0.211	0.405	0.337	0.461
DocQIR-Emb	**0.845**	**0.580**	**0.803**	**0.755**	**0.822**	**0.877**	**0.643**	**0.843**	**0.803**	**0.863**

Training Details All experiments use AdamW optimizer and the learning rate is set to 5e-4 and weight decay 0.05. The model is warmup for 2000 steps and trained for 2 epochs with batch size of 64. For the text encoder, BGE-m3 [9] is used and kept frozen through the training. For the visual encoder, the weight of the VLM is initialized with the pretrained weight from DSE [29] and the projector is randomly initialized.

Metric To evaluate the performance of DocQIR-Emb model and existing baselines, we employ the reciprocal rank (RR)@K as our primary evaluation metric. This metric is particularly suited for assessing the ranking quality of image retrieval systems, such as the one proposed in this work. Given a question query, the RR@K metric considers the top K retrieved images, where the images are sorted in descending order based on their similarity scores with respect to the query. The reciprocal ranking is then computed on these top K images by measuring the position of the ground truth image within the ranked list. The reciprocal rank is defined as the multiplicative inverse of the rank of the ground truth image, i.e. $\frac{1}{rank}$, where the rank is the position of the ground truth image in the sorted list. In our experimental setup, we consider multiple values of K, specifically $K = 1, 3, 5, 10$, to provide a comprehensive understanding of our model's performance across different retrieval scenarios. By evaluating the RR@K at these various K values, we can assess the ability of our DocQIR-Emb model to retrieve the ground truth image at different positions within the ranked list.

Dataset To evaluate our proposed approach, we utilize the training and testing split of the DocQIR dataset. Specifically, we employ around 500K question-image pairs for training, unless otherwise specified. Notably, our experiments reveal that training with multi-lingual questions does not provide significant benefits, likely due to the fact that the pre-trained text encoder already possesses robust language-agnostic capabilities. Consequently, we keep the pre-trained text encoder frozen during training, thereby avoiding unnecessary complexity and computational overhead. For evaluation, we assess the performance of all models on two distinct subsets of the DocQIR dataset. The first subset consists of 1,444 visual question answering (VQA) pairs on scientific images and the second subset comprises 1,000 VQA pairs on table images. Performance for both subsets are reported.

5.1 QIR Performance

We conduct a comprehensive evaluation of the proposed DocQIR-Emb against the baseline models on two subsets of the DocQIR testing split (table and chart images). Table 2 summarized the result of both RR@1 and RR@5, when the proposed model and the baselines are evaluated on the table image subset of DocQIR test set. Our results indicate that while most baseline models exhibit satisfactory performance when the text query is a descriptive caption and the image is a natural image, they struggle to retrieve the correct table image when the text query is a question. This discrepancy highlights the limitations of standard training pipelines and datasets that primarily focus on image captioning and natural images. The inability of these models to effectively handle the table images and question-based text queries underscores the need for specialized models like DocQIR-Emb. As show in the table, DocQIR-Emb outperforms all the baseline by at least 40%, under both RR@1 and RR@5 across all different languages.

Table 3 further compares the second subset of the DocQIR test split on chart images. Our evaluation reveals that DocQIR-Emb outperforms all baseline methods by at least 30%. This finding is consistent with our previous observation on the table image split, where DocQIR-Emb demonstrated superior performance. However, a notable difference emerges when comparing the overall performance across

Table 3: QIR result on the **chart image** subset of the DocQIR test set.

Method	Mean RR@1					Mean RR@5				
	En	Hi	Ja	Ko	Pt	En	Hi	Ja	Ko	Pt
MegaPairs [63]	0.107	0.008	0.012	0.005	0.044	0.161	0.010	0.020	0.009	0.070
Nomic-v1.5 [36]	0.391	0.015	0.036	0.021	0.211	0.449	0.022	0.052	0.035	0.264
JinaClip-v2 [22]	0.602	0.404	0.519	0.490	0.563	0.668	0.475	0.591	0.556	0.631
LongCLIP [60]	0.631	0.019	0.116	0.041	0.265	0.681	0.028	0.154	0.059	0.326
LLM2CLIP [21]	0.578	0.349	0.458	0.470	0.501	0.644	0.419	0.543	0.551	0.571
ColQwen2 [16]	0.684	0.141	0.480	0.349	0.497	0.760	0.192	0.566	0.438	0.581
ColPali-v1.3 [16]	0.663	0.296	0.506	0.436	0.537	0.751	0.377	0.601	0.531	0.634
DocQIR-Emb	**0.909**	**0.664**	**0.861**	**0.835**	**0.882**	**0.932**	**0.732**	**0.898**	**0.876**	**0.915**

Table 4: Ablation of different vision embedder and training datasets. For all the experiments in this study, we fix the text embedder as BGE-m3 [9]. We adopt the pretrained weight from EVA-CLIP [52] to represented CLIP style vision encoder.

Vision Embedder	Training Dataset	Table Image - Mean RR@1					Scientific Image - Mean RR@1				
		En	Hi	Ja	Ko	Pt	En	Hi	Ja	Ko	Pt
EVA-CLIP	CC3M [48]	0.019	0.010	0.019	0.012	0.014	0.211	0.113	0.152	0.148	0.172
EVA-CLIP	CC12M [8]	0.020	0.011	0.021	0.015	0.016	0.197	0.117	0.159	0.130	0.180
EVA-CLIP	CC3M [48] (Translated)	0.014	0.010	0.019	0.009	0.013	0.198	0.109	0.154	0.149	0.167
EVA-CLIP	DocQIR	0.209	0.116	0.166	0.156	0.187	0.364	0.202	0.292	0.283	0.311
VLM	DocQIR	**0.845**	**0.580**	**0.803**	**0.755**	**0.822**	**0.909**	**0.664**	**0.861**	**0.835**	**0.882**

the two splits. Specifically, the results in Table 3 show higher performance for almost all methods and languages, indicating that retrieving scientific charts is a relatively simpler task compared to retrieving table images. We attribute this difference to the inherent complexity of table images, which often contain heavy contextual text and intricate structures that require more sophisticated reasoning. In contrast, scientific charts typically exhibit more straightforward visual structures, making it easier for models to capture relevant features and establish relationships between text and image elements.

When comparing the performance of Table 2 and Table 3 across different languages, we observe that question written in English achieves the best performance, followed by Portuguese . Among all 5 languages, Hindi is the most challenging one and have approximately 0.2 performance drops on average compared to other languages.

Furthermore, our analysis highlights the importance of leveraging VLM in document image retrieval tasks. Baseline models that utilize VLM, such as ColQwen2 and ColPali, consistently outperform those that rely on encoding images with Vision Transformers (ViT) [15] or convolutional neural networks. The superior performance of VLM-based models can be attributed to their ability to jointly learn visual and linguistic representations, enabling more effective fusion of text and image features. By learning to represent both modalities in a shared embedding space, VLMs can capture subtle relationships and nuances that might be missed by models that process text and images separately.

5.2 Ablation Study

We analyze the performance of DocQIR-Emb from different perspective, including the architecture design, the affect of pretraining dataset to the performance across different languages and metrics.

Training dataset Table 4 compares the use of different training dataset. The first row shows the configuration where the CLIP-style vision encoder is used and optimized using the CC3M [48] datasets. Unfortunately, due to the large visual domain gap between the image content in CC3M and the DocQIR dataset, this configuration does not perform well for both table image and scientific image split. The second row further extends the training dataset by nearly 4 times with the use of CC12M [8] and, again, suffers from the large domain gain between natural image and images in the documents. In addition, we also translate the CC3M into 5 languages using off-the-shelf LLM (i.e. Gemma [54]) and perform training on the CC3M(Translated). During training, question of different languages are randomly sampled to align with the associated image. Row 3 shows the performance and it shows that training with question of multiple languages does not improve the performance

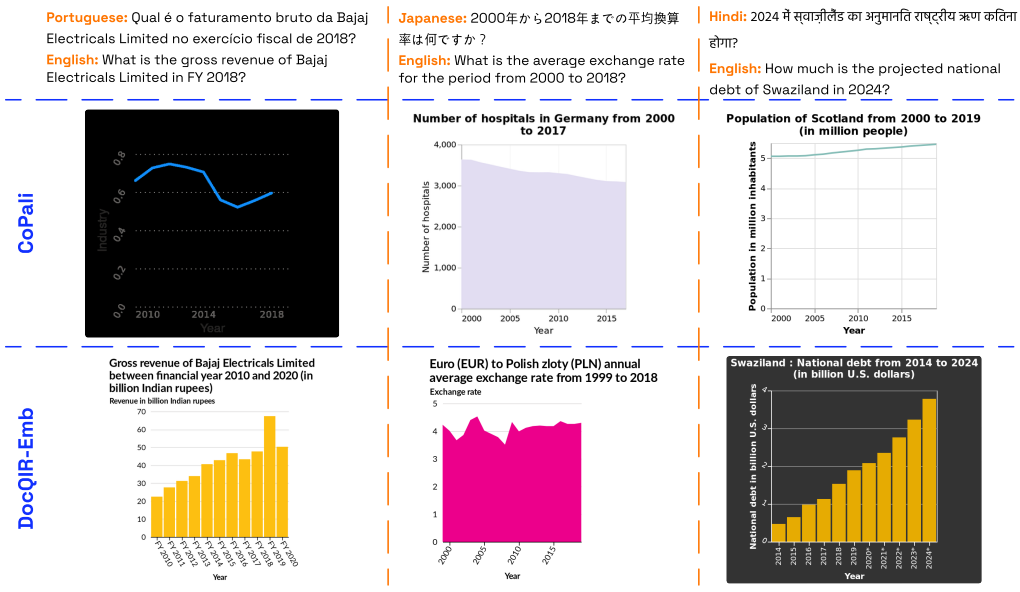

Figure 4: We demonstrate the effectiveness of our proposed DocQIR-Emb in retrieving relevant images for questions in multiple languages. The top row displays questions in Portuguese (left), Japanese (middle), and Hindi (right), along with their English translations. The middle row shows the most relevant images retrieved by the CoPali model [16], which fails to retrieve the correct images. In contrast, the bottom row shows the images retrieved by our DocQIR-Emb model, which successfully retrieves the correct images associated with each question, highlighting its superior performance.

over that trained on vanilla CC3M. This also highlights the importance of using a multi-lingual text embedder, which is already capable of mapping text in different languages into a similar text embedding. Finally, row 4 and row 5 adopt the DocQIR training set and dramatically improve the performance over the counterpart that uses CC3M or CC12 as training dataset. This highlights the challenges of generalizing existing multi-modal embedder to documents, which are mostly trained on natural image datasets with image caption.

Architecture While row 4 and row 5 of Table 4 are trained on the DocQIR training set, they adopt different variants of the DocQIR-Emb design. Row 4 uses CLIP-style vision encoder (ViT based), while Row 5 adopts the VLM to embedding the image. Given that VLM is pretrained on large multi-modal dataset with various downstream tasks, we can see that using VLM beats CLIP-style vision encoder by at least 0.4 points for both testing subset and languages.

Qualitative Results Figure 4 illustrates three examples of our DocQIR-Emb model's performance in retrieving relevant images for query questions in multiple languages, including Portuguese (left), Japanese (middle), and Hindi (right). The objective is to retrieve the most relevant image that can serve as supporting context to answer the input query. To evaluate the effectiveness of our approach, we compare the quality of the retrieved images between DocQIR-Emb and the CoPali model, which ranked second in our experiments (see Table 2 and Table 3). As shown in Figure 4, CoPali retrieves images with content that is unrelated to the query, whereas DocQIR-Emb successfully retrieves the correct images associated with each query, demonstrating the superior performance of our model.

6 Discussion, Societal Impact and Limitations

In this work, we introduce a novel dataset and a multi-modal embedder for question-image retrieval (QIR). Our proposed dataset is motivated by the limitations of existing multi-modal datasets, which are dominated by natural images and their corresponding captions. As a result, models trained on these datasets often struggle to retrieve relevant images from documents. To address this challenge, we propose the DocQIR-Emb, which achieves a significant improvement of over 40% compared to prior models. This substantial gain demonstrates the effectiveness of our approach in retrieving

correct images from documents, and underscores the importance of developing datasets and models that are tailored to the specific requirements of question-image retrieval in document understanding.

Limitations: While our multi-modal embedder is trained on visual content that does not contain unsafe material, we do not have safeguards to prevent its potential misuse. The investigation of unsafe visual content, including the methods to detect potential misuses, is outside the scope of this work.

Societal Impact: We envision the concept of DocQIR will stimulate further research of multi-modal datasets for document understanding. Furthermore, our proposed DocQIR-Emb introduces a novel design multi-modal embedders for QIR task, offering a foundation for future research in this field.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Singh Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C'esar Teodoro Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. *ArXiv*, abs/2412.08905, 2024.

[2] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024.

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

[4] Camille Barboule, Benjamin Piwowarski, and Yoan Chabot. Survey on question answering over visually rich documents: Methods, challenges, and trends. *arXiv preprint arXiv:2501.02235*, 2025.

[5] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[6] Galal M. Binmakhashen and Sabri A. Mahmoud. Document layout analysis: A comprehensive survey. *ACM Comput. Surv.*, 52(6), October 2019.

[7] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[9] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

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

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[12] Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. *arXiv preprint arXiv:2411.06176*, 2024.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics.

[14] Yihao Ding, Jean Lee, and Soyeon Caren Han. Deep learning based visually rich document content understanding: A survey. *arXiv preprint arXiv:2408.01287*, 2024.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models, 2024.

[17] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

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

[19] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *arXiv preprint arXiv:2302.11154*, 2023.

[20] Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Ren Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. Do lvlms understand charts? analyzing and correcting factual errors in chart captioning. *ArXiv*, abs/2312.10160, 2023.

[21] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, and Lili Qiu. Llm2clip: Powerful language model unlock richer visual representation, 2024.

[22] Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images, 2024.

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

[24] Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. *arXiv preprint arXiv:2406.03411*, 2024.

[25] Paul Lerner, Olivier Ferret, and Camille Guinaudeau. Cross-modal Retrieval for Knowledge-based Visual Question Answering. Accepted at ECIR 2024, January 2024.

[26] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[28] Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. Parsing table structures in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 944–952, 2021.

[29] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

[30] Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. Visa: Retrieval augmented generation with visual source attribution. *arXiv preprint arXiv:2412.14457*, 2024.

[31] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[32] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[33] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124, 2023.

[34] Mindee. doctr: Document text recognition. `https://github.com/mindee/doctr`, 2021.

[35] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *The Twelfth International Conference on Learning Representations*, 2024.

[36] Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder, 2024.

[37] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011.

[38] Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know" no" better: A data-driven approach for enhancing negation awareness in clip. *arXiv preprint arXiv:2501.10913*, 2025.

[39] ShengYun Peng, Seongmin Lee, Xiaojing Wang, Rajarajeswari Balasubramaniyan, and Duen Horng Chau. High-performance transformers for table structure recognition need early convolutions. *arXiv preprint arXiv:2311.05565*, 2023.

[40] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqa: A dataset for multimodal question answering on scientific papers. *ArXiv*, abs/2407.09413, 2024.

[41] Leigang Qu, Haochuan Li, Tan Wang, Wenjie Wang, Yongqi Li, Liqiang Nie, and Tat-Seng Chua. Unified text-to-image generation and retrieval. *arXiv preprint arXiv:2406.05814*, 2024.

[42] Vincent Quantmeyer, Pablo Mosteiro, and Albert Gatt. How and where does CLIP process negation? In Jing Gu, Tsu-Jui (Ray) Fu, Drew Hudson, Asli Celikyilmaz, and William Wang, editors, *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 59–72, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

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

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

[45] Susie Xi Rao, Johannes Rausch, Peter Egger, and Ce Zhang. Tableparser: Automatic table parsing with weak supervision from spreadsheets. *arXiv preprint arXiv:2201.01654*, 2022.

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[48] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[49] Tahira Shehzadi, Didier Stricker, and Muhammad Zeshan Afzal. A hybrid approach for document layout analysis in document images. In *International Conference on Document Analysis and Recognition*, pages 21–39. Springer, 2024.

[50] Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642, 2022.

[51] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534*, 2020.

[52] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[53] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*, 2023.

[54] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[55] Jianqiang Wan, Sibo Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. Omniparser: A unified framework for text spotting key information extraction and table recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15641–15653, 2024.

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

[57] Jianyou Andre Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. Scientific document retrieval using multi-level aspect-based queries. *Advances in Neural Information Processing Systems*, 36:38404–38419, 2023.

[58] Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, et al. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems*, 35:25600–25614, 2022.

[59] Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. Real-mm-rag: A real-world multi-modal retrieval benchmark. *arXiv preprint arXiv:2502.12342*, 2025.

[60] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024.

[61] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[62] Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception, 2024.

[63] Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*, 2024.

[64] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.