MMVQA: A Comprehensive Dataset for Investigating Multipage Multimodal Information Retrieval in PDF-based Visual Question Answering

Yihao Ding[1,2], **Kaixuan Ren**[2], **Jiabin Huang**[2], **Siwen Luo**[3] and **Soyeon Caren Han**[1,2*]
[1]The University of Melbourne
[2]The University of Sydney
[3]The University of Western Australia
yihao.ding@sydney.edu.au, kren4925@uni.sydney.edu.au,
jiabin.eta@gmail.com, siwen.luo@uwa.edu.au, caren.han@unimelb.edu.au

Abstract

Document Question Answering (QA) presents a challenge in understanding visually-rich documents (VRD), particularly with lengthy textual content. Existing studies primarily focus on real-world documents with sparse text, while challenges persist in comprehending the hierarchical semantic relations among multiple pages to locate multimodal components. The paper introduces MMVQA, a dataset tailored for research journal articles, encompassing multiple pages and multimodal retrieval. Our approach aims to retrieve entire paragraphs containing answers or visually rich document entities like tables and figures. The main contribution is introducing a comprehensive PDF Document VQA dataset, allowing the examination of semantically hierarchical layout structures in text-dominant documents. We also present new VRD-QA frameworks to grasp textual contents and relations among document layouts simultaneously, extending page-level understanding to the entire multi-page document. We aim to enhance the capabilities of existing vision-and-language models in handling challenges posed by text-dominant documents in VRD-QA. Code and Appendix are in https://github.com/adlnlp/pdfmvqa.

1 Introduction

The growing demands for visually rich document (VRD) question-answering (QA) areas are becoming increasingly evident, especially in specialised fields such as finance and medicine. VRDs, including forms [Ding *et al.*, 2023a], academic papers [Ding *et al.*, 2023b], and industrial reports [Mathew *et al.*, 2021a], typically comprise text-dense and visually rich components such as *titles*, *paragraphs*, *tables*, and *charts*. These components, ***document semantic entities***, are not only knowledge-intensive but are also organised in a predefined layout that maintains a logical and semantic correlation, usually extending across multiple pages. This complexity requires a more grounded and fact-dependent approach to QA. It is essential to comprehend the layout and logical structure of VRDs, especially in multi-page documents, to accurately locate and use these document entities as reliable evidence for answering knowledge-intensive questions. Recent generative models [Ouyang *et al.*, 2022; Touvron *et al.*, 2023; Liu *et al.*, 2023a] have made impressive progress in providing interactive human-like responses by memorising vast knowledge [Zhao *et al.*, 2023]. These models rely on plain text to learn textual content [Touvron *et al.*, 2023] and use image patches to encode visual cues [Yasunaga *et al.*, 2022]. This approach makes understanding document entities' layout and logical relationships in VRDs difficult. Generative models are suffered from hallucinations [Ye *et al.*, 2023], high costs [Hofstätter *et al.*, 2023], and updating knowledge difficulties[Hu *et al.*, 2023]. Retrieval-based QA [Liu *et al.*, 2023b] addresses these limitations when applying generative models to VRD-QA. This approach helps locate answers or supporting evidence precisely, offering more grounded and factually dependent information. While recent retrieval-based applications mainly focus on web-crowded domains like Wikipedia[Hu *et al.*, 2023], VRD-QA requires a deep understanding of domain-specific multimodal knowledge.

A few VRD-QA datasets [Mathew *et al.*, 2021a; Tanaka *et al.*, 2021] have been devised to extract in-line text from input document pages but often overlook prevalent multi-page scenarios. Recent multi-page datasets focus on extracting short phrases or sentences [Tito *et al.*, 2023], causing recently proposed models [Huang *et al.*, 2022; Yu *et al.*, 2022] to excel at retrieving annotated in-line text but disregarding the logical and layout connections among document entities. Moreover, they are limited in handling the entire lengthy document. To address these limitations, entity-level document understanding tasks have been introduced by [Ding *et al.*, 2023a] and [Ding *et al.*, 2023b]. A common issue with these datasets is their text-dense mono-modal information extraction, overlooking visually rich entities such as *tables* and *figures*.

This paper proposes a new multi-page, multimodal document entity retrieval dataset, MMVQA, for knowledge-intensive domain. MMVQA addresses the limitations of generative models and expands upon the benefits of retrieval-based models by incorporating multimodal document entities like paragraphs, tables and figures and exploring the cross-page layout and logical correlation between them. This expansion supports the models to navigate and interpret real-

*Corresponding author

world documents at a multi-page or entire document level by leveraging joint-grained and multimodal information. The proposed models demonstrates how to effectively use existing VLPMs and pretrained language models with long sequence support to locate target entities from MMVQA.

The contributions are summarised as follows: We introduce MMVQA, a new VQA dataset for retrieving multimodal document semantic entities in multi-page VRDs, accompanied by versatile metrics for diverse scenarios. A set of frameworks for multi-page document entity retrieval is proposed by leveraging the implicit knowledge from VLPMs and fine-grained level information. A series of experiments are performed to provide deeper insights into MMVQA and demonstrate the effectiveness of our proposed techniques for multi-modal multi-page document entity retrieval.

2 Related Work

The first document image-baased QA dataset, DocVQA [Mathew *et al.*, 2021b], includes scanned industrial documents. Questions in the DocVQA dataset are designed as in-line questions where the single-span answers and the keywords in questions are in the same line of text. Based on the DocVQA dataset document images, CS-DVQA [Du *et al.*, 2022] proposed new questions requiring commonsense knowledge. Unlike extracting in-line answers on document pages, answers to CS-DVQA dataset questions could be the node of ConceptNet. RDVQA dataset [Wu *et al.*, 2022], on the other hand, focuses on the question answering over coupon and promotion vouchers. Unlike the in-line questions, the RDVQA dataset proposed the in-region questions, which require the answer to be inferences from the information in the related region. In contrast to the single document page processing, DocCVQA [Tito *et al.*, 2021] and SlideVQA [Tanaka *et al.*, 2023] datasets proposed the question answering over the document collections. DocCVQA specifically focuses on a single document source, the US Candidate Registration Form. Due to the similar form layout and form fields, this dataset only proposed a limited number of in-line questions. However, multiple answer values could be extracted from multiple independent document images for answering one question. SlideVQA collects the set of slides, and there will be multiple answers to one question from different slide pages. Although DocCVQA and SlideVQA improve document VQA tasks to a multi-page level from the ordinary single page, their documents are not consecutive pages with dense texts. On the other hand, VisualMRC [Tanaka *et al.*, 2021] collected the text-dense webpage screenshots, and questions are formed like in the machine reading comprehension task that requires the contextual understanding of textual paragraphs. However, VisualMRC limits the task scope to the single-page level. Existing datasets primarily extract text on MRC style and overlook visually rich elements like *tables* and *figures*. Current multi-page datasets mainly use sparse text sources, such as slides, while the demand is growing for text-dense documents. Our proposed MMVQA dataset aims to bridge these gaps by creating a multi-modal VRD-QA dataset that retrieves target document entities across multiple pages.[1]

[1]Please refer to Appendix A to check dataset comparison table.

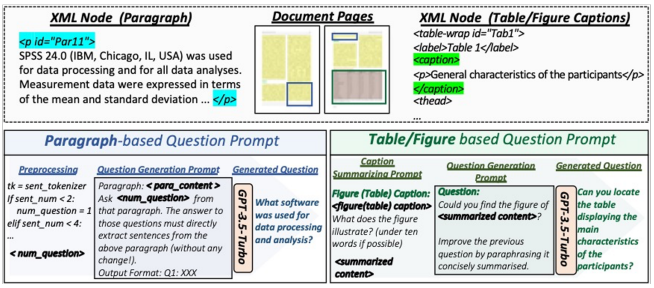


Figure 1: A sample Question generation progress.

3 MMVQA

Dataset Collection The documents are collected from PubMed [2], a biomedical and life science journal literature archive. The subset contains millions of open-access articles in machine-readable formats, including PDF and XML. We randomly downloaded 10K articles in both PDF and XML and then filtered out 3146 documents, including research articles, review articles, and systematic review articles, based on the metadata in XML.

Dataset Preprocessing The dataset includes both PDF images and segmented document components, categorised into predefined semantic categories such as *Title*, *Section*, *Paragraph*, *List*, *Figure*, *Table*, *Figure Caption*, and *Table Caption*. We refer to those segmented document components as ***document semantic entities***, which contain associated text within its bounding box. We follow the way that [Zhong *et al.*, 2019] uses PDFminer [3] to extract the bounding box coordinates and text of each document page's textbox, textline, image, and geometric shapes. We match the exact texts in XML files for the segmented bounding boxes by applying fuzzy string matching for XML texts and the detected texts.

Question Generation We focus on generating a large number of diverse types of content-related questions that are associated with different multi-modal document entities of journal articles. To do so, we use ChatGPT[4] to automatically generate 1-3 questions based on the contents of each paragraph of these main sections. As shown in Figure 1, for paragraph-based questions, the number of sentences in the paragraph determines the number of questions (n_q) to be generated. The paragraph text (P_t) is then used as a prompt for ChatGPT (GPT-3.5-turbo) with n_q. For questions based on tables or figures, the caption content is first summarised (S_c) using ChatGPT, and then questions are generated based on the summarised content. Then, the questions are filtered by predefined rules to ensure quality and evaluated by raters. [5]

Dataset Format The MMVQA is divided into three sets: training, validation, and testing, with the statistics in Table 1[6]. Each set comprises a DataFrame (CSV file) with attributes such as "*question*", "*answer*", and "*document_id*". Extra annotations for "*context*" and "*page_range*" are included, presenting the text content and the covered page range (in the

[2]https://www.ncbi.nlm.nih.gov/pmc/

[3]https://pypi.org/project/pdfminer/

[4]Any LLM can be usable to generate diverse types of questions

[5]Please refer to Appendix J [McHugh, 2012].

[6]More attribute examples are in Appendix B

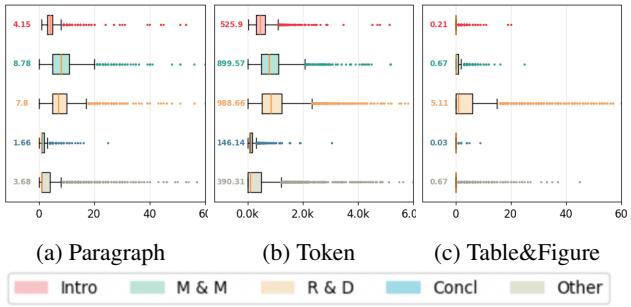


Figure 2: Distribution of various document components and semantic entities of each Super-Section type.

first-level/top-level section) of the answer for the question. For each set, we provided the metadata information (in an additional JSON file), respectively. It contains annotated features, including document entity *bounding box*, *text content*, *category*, etc, which are essential for model implementation.

4 Dataset Analysis

Document Components Statistics Our dataset includes only documents that contain multiple pages with numbers of *tables/figures* or includes complex structures of contents with multiple different sections and subsections. Based on the statistics[7], we found the number of document components is quite consistent. Most documents contain around ten pages and have 10-20 different sections with around 20-40 paragraphs of 2000-4000 tokens. Hence, with the analysis, we can ensure that the collected documents are mostly lengthy and have a complex structure enough to evaluate the model's feasibility to contextualise understanding over multiple consecutive pages. In addition to this, each document contains enough tables and figures to ensure the possible questions asked over these components. Most documents have around five *tables* and *figures* or more.

Super-Section Component Analysis We refer first-level section of each document as **Super-Section**, where the sections under the same Super-Section play similar structural roles in a medical domain academic paper, including *Introduction* (*Intro*), *Material and Method* (*M&M*), *Result and Discussion* (*R&D*), *Conclusion* (*Concl*) and *Other* [8]. Sections are categorised into *Other* Super-Section in documents, like *Conflict of Interest*, *Funding*, *Ethical Approval*, and *Supplementary* are less common but contain critical information.

The document layout statistics across Super-Sections are in Figure 2. The *Materials and Methods* (*M&M*) and *Results and Discussion* (*R&D*) sections are normally more complex, with multiple subsections, paragraphs, and most tables and figures. In contrast, the *Introduction* (*Intro*) and *Conclusion* (*Concl*) sections are simpler, with fewer subsections. The *Other* Super-Section, encompassing diverse contents like *Supplementary* or *Fundings*, has a larger interquartile range and more outliers, reflecting its varied nature.

Number of Question Distribution MMVQA contains 3,146 documents, which are a total of 30,239 pages. Each document is averagely associated with 84 questions, resulting in 262,928 question-answer pairs in MMVQA. The detailed Training/Validation/Test set size and the question number of each document Super-Section can be found in Table 1.

Splits	# Docs	# Pages	Number of Questions							
			Overall	Intro.	M&M	R&D	Concl.	Others	Figure	Table
Train	2,209	21,495	180,797	21,749	39,484	78,240	4,886	36,438	7,645	4,920
Val	314	2,862	27,588	3,301	6,047	12,274	1,004	4,962	996	755
Test	623	5,882	54,543	6,669	12,906	26,007	1,825	7,136	2,115	1,513
Total	3,146	30,239	262,928	31,719	58,437	116,521	7,715	48,536	10,756	7,188

Table 1: Dataset distribution across different splits with question count by Super-Section category.

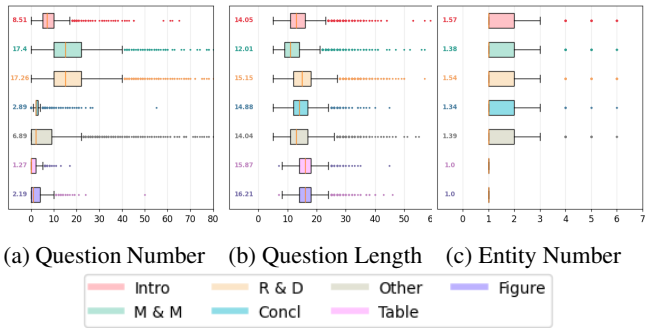


Figure 3: Question pattern analysis of each Super-Section type.

Super-Section-oriented Question-Answer Distribution The distribution of questions over each Super-Section is shown in Figure 3a. Most questions are asked over *M&M* and *R&D* sections, each having an average of around 17 questions. The average question length is in Figure 3b. *Table/figure*-related questions are longer, and the average question length of *M&M* sections is the shortest. For *table/figure*-related questions, answers to questions can be recognised from one document entity (segmented by a bounding box). For other Super-Section questions, answers may located in more than one document entity.

5 Task Definitions and Metrics

We introduce our main task as ***Multimodal Document Information Retrieval (DIR)*** aimed at **retrieving semantic entities**, such as *paragraphs*, *tables*, and *figures*, from the input entity sequence across ***multiple pages***. As demonstrated by [Ding *et al.*, 2023b; Gu *et al.*, 2021], the document entity-level task encourages the exploration of logical and spatial relationships between semantic entities, and it is more straightforward to extend to the multi-page level compared to fine-grained token-level inputs. For instance, as shown in Figure 4, utilising document-entity sequences as input enhances both logical aspects (e.g., linking *Table* E_t with its corresponding *Table Caption*) and semantic understanding (e.g., handling split *Paragraph* entities E_{p1} and E_{p2})[9]. Additionally, to address diverse application scenarios and effectively meet specific requirements, we introduced a set of distinct evaluation metrics for more adaptive performance assessment, including ***Exact Matching*** (*EM*), ***Partial Matching*** (*PM*), and ***Multi-Label Recall*** (*MR*). More details can be articulated in Section 5.2.

[7]Please refer to Appendix C.2 to check the statistics chart.

[8]Please check Appendix C.3 for more Super-Section analysis.

[9]Token-level models struggle to capture entity-level correlations.

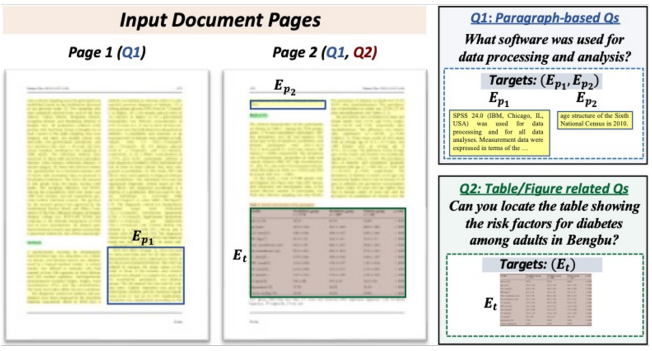


Figure 4: Defining tasks of multi-modal cross-page information retrieval with illustrative examples.

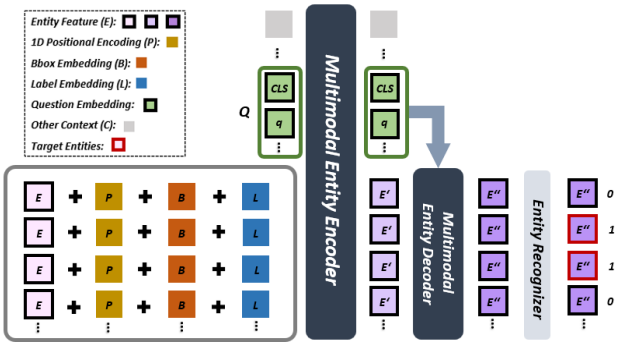


Figure 5: Multimodal Multi-page Retriever Framework

5.1 Task Definition

How is our multimodal DIR task conducted? Assuming Q is a natural language question and $S_E = \{E_1, E_2, \ldots, E_m\}$ is a set of document entities comprising m semantic entities of the target multiple document pages. $S_{E_{gt}} = \{E_1, \ldots, E_j\}$ represents the ground truth entity set for Q. If a paragraph is divided into several regions, $S_{E_{gt}}$ may include more than one entity (as in Figure 4). The task involves proposing a model $\mathcal{F}ir$ with inputs Q and S_E to predict an entity set $S_{E_{Q_{pre}}}$. As in Figure 4, for a **paragraph-based** question $Q1$, the ground truth set $S_{E_{Q1_{gt}}} = \{E_{p_1}, E_{p_2}\}$, where E_{p_1}, E_{p_2} belong to the same paragraph but are split into two regions. For a **table/figure-based** question $Q2$ in Figure 4, the ground truth set only contains the table entity E_t.

5.2 Evaluation Metrics

Distinct evaluation metrics cater to the varied application scenarios of retrieved entities. These metrics encompass stringent exact-match accuracy to more lenient measures, allowing partial retrieval and multi-label recall and providing a comprehensive performance assessment. **Exact Matching Accuracy***(EM)* is a stringent metric suitable for scenarios requiring precise, unambiguous information retrieval, particularly when used as supporting evidence or reliable references. We also introduced **Partial Matching Accuracy***(PM)* with tolerance for partial matches. It is especially beneficial when capturing every relevant entity is less crucial than ensuring the correctness of the predicted entities, such as ensuring the correct identification of the primary entity E_{p1} in a target paragraph. **Multi-Label Recall***(MR)* is applied to assess the proportion of correctly identified actual positives in situations where identifying all positive instances is critical. We provide the detailed definitions of each metric in Appendix D.

6 Methodology

6.1 Multimodal Multi-Page Retriever

Existing document understanding models [Huang *et al.*, 2022; Kim *et al.*, 2022; Wang *et al.*, 2022; Li *et al.*, 2021] and datasets [Mathew *et al.*, 2021a; Tanaka *et al.*, 2021] are designed for single-page document comprehension, relying on token-level representations. However, the fine-grained token-level information suffers from the limited length. It neglects the correlations between document entities, particularly in capturing long contextual dependencies in more prevalent multi-page scenarios. Instead of employing sequences of tokens that lead to significant memory consumption, we introduce a multimodal entity-level retrieval framework $\mathcal{R}$ to identify the target entity set S_Q from the cross-page entity sequence in a given question Q, as illustrated in Figure 5.

The input, comprising multiple pages, consists of a set of document entity embeddings $\mathbb{E} = E_1, E_2, ..., E_n$. These embeddings, elaborated in Section 6.2, are combined with 1D positional encoding $\mathbb{P}$, bounding box embedding $\mathbb{B}$, and label embedding $\mathbb{L}$[10]. The combined representation, $\mathbb{E}+\mathbb{P}+\mathbb{B}+\mathbb{L}$, is fed into the ***multimodal Entity Encoder*** $\mathcal{E}$, alongside the question token embeddings $Q = q_1, q_2, ..., q_m$ and additional context elements like image patch embeddings P. The encoder $\mathcal{E}$ models the correlations among these entities, the question, and other contexts. The enhanced entity representation $\mathbb{E}'$ from $\mathcal{E}$, along with Q, serves as input for a transformer-based ***Multimodal Entity Decoder*** $\mathcal{D}$, producing the final representation $\mathbb{E}''$. Each entity in $\mathbb{E}''$ is linearly projected by a ***Entity Recogniser*** $\mathcal{L}_{er}$ for binary classification, distinguishing target entities (label 1) from non-target entities (label 0) in the context of the question Q and Entity Set $\mathbb{E}$.

6.2 VLPM Augmented Retriever

Existing Vision Language Pre-training Models (VLPM)s can be classified into two categories based on their focus on visual cues: Region-of-Interest (RoI)-based and Image Patch-based [Long *et al.*, 2022]. RoI-based models utilise features from ground truth or predicted regions, while Patch-based models process segmented image patches. Even though these VLPMs are initially pretrained on general photo-like image-related tasks rather than visually-rich documents, previous studies have illustrated the feasibility of employing VLPMs such as [Li *et al.*, 2019; Tan and Bansal, 2019; Kim *et al.*, 2021] in tasks related to understanding documents. Thus, we propose methods to harness the implicit information embedded in pretrained VLPMs for obtaining more comprehensive and robust representations of multimodal entities.

RoI-Based Frameworks

RoI-based VLPMs focus on learning the contextual entity relationships and correlation between textual content and as-

[10]Appendix F includes details of the input representation.

sociated visual cues of each RoI, which in our scenario are document-semantic entities (e.g. *section*, *paragraph*, *table*, etc.). $\mathcal{F}_{roi}$ donates a ***RoI-based VLPM*** backbone. This backbone takes a question token sequence Q and a set of visual representations $\mathbb{V}$ as input, where $\mathbb{V} = \{V_1, V_2, \ldots, V_n\}$ signifies the initial visual representations of each entity in the document D. Our objective is to generate an improved visual embedding set $\mathbb{V}'$, capturing the contextual relationships among entities and their correlation with the question. Then, $\mathbb{V}'$ is concatenated with textual embedding $\mathbb{T}$ and fed into a linear ***Vision-Textual Projector*** $\mathcal{L}_{vl}$ to produce the entity representation set $\mathbb{E}$ for input into the retriever $\mathcal{R}$. We employ vanilla **Transformer** as a foundational benchmark for evaluating the impact of various pretrained techniques in comparative studies [Ding *et al.*, 2023a]. Additionally, we introduce **VisualBERT** [Li *et al.*, 2019] and **LXMERT** [Tan and Bansal, 2019] to enhance the initial visual embedding of each document entity [11]. The improved visual embeddings are concatenated with $\mathbb{T}$ to obtain $\mathbb{E}$.

Image Patch-Based Frameworks

Recently emerged VLPMs commonly employ image patches without prior RoI bounding box information, a practice also observed in document understanding frameworks designed for single-page scenarios [Xu *et al.*, 2021; Huang *et al.*, 2022]. Despite these advancements, the demands of cross-page document understanding remain insufficiently addressed. Consequently, our research investigates the effectiveness of image-patch-based VLPMs in the general domain in cross-page information retrieval tasks. Extensive experiments and analyses are conducted to evaluate the effectiveness of patch-based methods in enhancing entity representation in cross-page document information.

To apply a vision-language model for cross-page document understanding, we first merge multiple document pages $\mathbb{I} = \{I_1, I_2, ..., I_m\}$ into a composite image I. After that, the resized image and question are fed into VLPM processors to produce image patch pixel and question token sequences, which are the inputs of corresponding ***Patch-based VLPM encoders***. The generated patch embedding $P = \{p_1, p_2, ..., p_t\}$ and the question token embedding Q are combined with the entity embedding $\mathbb{E}$ and fed into a ***Multimodal Entity Encoder*** $\mathcal{E}$ within the retriever $\mathcal{R}$, facilitating contextual learning between them. Then, we can get $[Q', P', \mathbb{E}'] = \mathcal{E}([Q, P, \mathbb{E}])$, where $\mathbb{E} = \mathcal{L}_{vt}(\mathbb{V} \oplus \mathbb{T})$. $\mathbb{E}'$ and Q are fed into the ***Multimodal Decoder Entity Decoder*** $\mathcal{D}$ within $\mathcal{R}$ as target embedding and memory embedding for the retrieval process. We introduce patch-based VLPMs to obtain contextual patch embedding P, including models such as **CLIP** [Radford *et al.*, 2021], **ViLT** [Kim *et al.*, 2021], **BridgeTower** [Xu *et al.*, 2023] [12].

6.3 Joint-Grained Retriever

Entity-level document understanding models can gain advantages by incorporating logical and layout relationships to improve entity representations. However, overlooking fine-grained details, such as crucial phrases and sentences within text-dense document entities, diminishes robustness in semantic comprehension for lengthy VRDs. Inspired by [Ding *et al.*, 2024], we introduce a **Joint-grained Retriever** (Jg) architecture, shown in Figure 6[13], designed to enrich ***coarse-grained*** document entity representations with ***fine-grained*** token-level textual content. These augmented textual representations are subsequently utilised as input for retriever $\mathcal{R}$ to obtain final predictions. Supposing the input multi-pages contain n document entities, each entity has an initial textual representation, denoted as $\mathbb{T} = \{T_1, T_2, ..., T_n\}$. In addition, for each document page, text token sequences can be extracted using various approaches (e.g., OCR tools, PDF parsers, and source files) based on different application scenarios. These text token sequences are then processed by a pre-trained language model $\mathcal{F}_{lm}$ to obtain token representations $t = \{t_1, t_2, ..., t_p\}$, where p represents the number of input tokens. Since p is typically greater than 512 tokens in the case of multiple input pages, models capable of handling long sequences are required to acquire token representations t, e.g. BigBird [Zaheer *et al.*, 2020]. Then, the fine-grained token representation t and the coarse-grained entity representation $\mathbb{T}$ are utilised as memory and source inputs, respectively, for a Joint-grained decoder $\mathcal{D}_{jg}$, resulting in an enhanced entity representation $\mathbb{T}$. $\mathbb{T}$ is then fed into the retriever $\mathcal{R}$ (RoI-based or Patch-based), along with the entity visual embedding $\mathbb{V}$, to obtain the entity representation $\mathbb{E}$ for final prediction.

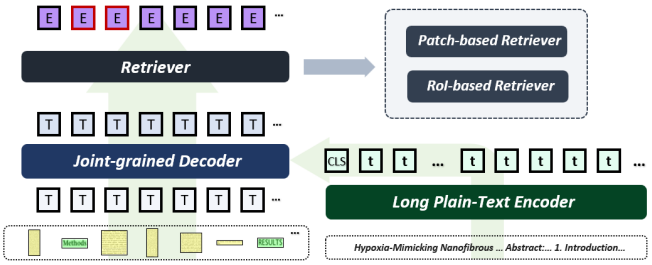


Figure 6: Joint-grained(coarse-and-fine grained) Retriever

7 Experiments and Discussions

7.1 Baseline Framework Results

Type	Model	EM		PM		MR	
		Val	Test	Val	Test	Val	Test
RoI-based	Transformer	17.92	19.46	22.48	23.96	25.68	27.50
	VisualBERT	15.39	17.80	21.92	23.86	**26.72**	**28.70**
	LXMERT	17.81	19.77	23.37	25.07	25.38	26.86
Patch-based	CLIP	20.71	22.55	25.70	27.59	24.79	26.56
	ViLT	**21.71**	**23.47**	**27.56**	**29.14**	25.71	27.40
	BridgeTower	19.88	22.37	23.99	26.30	25.37	27.64
Joint-grained BridgeTower	*w/ PDFMiner*	21.62	<u>23.56</u>	26.63	28.50	27.50	<u>29.22</u>
	w/ OCR	21.53	23.25	26.90	28.56	26.75	28.45

Table 2: Overall performance under various evaluation metrics.

To assess the effectiveness of RoI-based and Patch-based frameworks in retrieving entities from multi-page documents under different scenarios, performance metrics (*EM*, *PM* and

[11] For detailed model configurations, please refer to Appendix E.1.

[12] For further configuration details, please refer to Appendix E.2.

[13] Please refer to Appendix E.3 to see more detailed RoI-based and Patched-based retriever architectures.

MR) were used. Overall, Patch-based frameworks outperform others on *EM* and PM, with ViLT achieving 23.47% in *EM* and 29.14% in *PM* on the test set. However, for *MR*, there is no apparent difference among the applied models. VisualBERT achieved the highest result at 28.70%, indicating its robustness in retrieving target entities but sensitivity to noise, leading to the lowest *EM* (17.80%) in the test set. Notably, Patch-based surpassed all RoI-based models in *EM*. This indicates the document image patches, even pre-trained on the general domains, possibly lead to more representative question and entity representations, thereby boosting the comprehensive cross-page question-oriented retrieving. For ***RoI-based models***, no significant performance discrepancies are observed in *EM* and *PM* across three frameworks, where LXMERT (19.77%) shows slightly superior performance than pretrained VisualBERT (17.8%) and vanilla Transformer (19.46%) in the test set. This may be attributed to pre-trained RoI-based VLPMs not significantly augmenting entity vision representations. For ***Patch-based frameworks***, ViLT demonstrates approximately 1% higher performance than CLIP and BridgeTower, respectively, in terms of *EM*. This trend is more apparent in *PM* as well. The possible reason might demonstrate the proficiency of uni-encoder frameworks (ViLT) for text-vision alignment under text-dense domains. Table 2 demonstrates the superiority of Joint-grained models, exceeding vanilla models and even achieving the highest *EM* (23.56%) and *MR* (29.22%) in the test set. Further Joint-grained model results are discussed in Section 7.2 and 7.3. We also analyse the breakdown performance of each model from views of the Super-Section and the number of input pages, as articulated in Appendix G.1.

7.2 Joint-Grained Framework Results

Overall and Super-Section Breakdown Performance

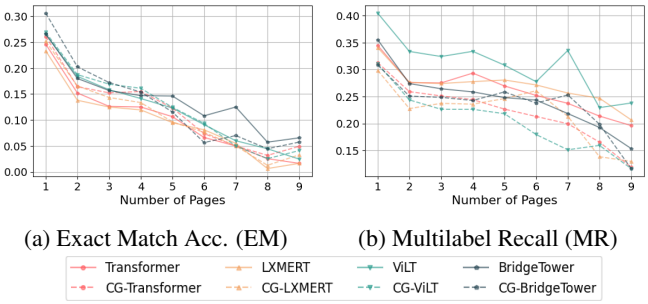
To illustrate the effectiveness of the proposed Joint-grained framework (Figure 6), we conducted a performance comparison between the top two vanilla frameworks on paragraph-based questions from both the **RoI-based** (Transformer and LXMERT) and **Patch-based** (ViLT and BridgeTower) groups and their respective Joint-grained architectures by feeding the provided *context* attribute of each question. Overall, Joint-grained models consistently improve performance, with LXMERT and BridgeTower showing more than a 2% increase. Regarding Super-Sections, complex Super-Sections like *M&M* and *R&D* benefit notably, especially BridgeTower, which improves by around 4% in *M&M* and 3.5% in *R&D*. Super-Sections with simple complexity (*Intro* and *Concl*) see less improvement, and the *Conclusion* (*Concl*) even performance decreases, especially in Patch-based frameworks (around 6% decrease). These trends suggest that fine-grained information enhances the understanding of text-dense entity textual representations by capturing important words or phrases missed at the entity level.

Page Range-Based Breakdown Analysis

To assess the Joint-grained framework's robustness across different input page numbers, we conducted a comparative analysis, shown in Figure 7. Figure 7a indicates that the Joint-grained framework enhances performance with smaller page gaps but experiences a decrease in performance with larger input page numbers. This suggests that fine-grained information may improve document entity representations. But, with the number of input pages increasing, textual tokens may introduce more noise that adversely affects document entity representations. Exploring additional Joint-grained mechanisms may help enhance entity representations. However, as shown in Figure 7b, Joint-grained frameworks notably enhance robustness in *MR*-oriented scenarios, from smaller to larger numbers of pages. This highlights that incorporating fine-grained textual information can aid the model in locating target entities even in long, visually rich document scenarios.

Model	Overall	Intro	M&M	R&D	Conl	Other
Transformer	17.32	24.19	12.36	15.71	44.82	15.97
Jg-Transformer	18.97	25.14	15.06	17.35	44.38	17.36
LXMERT	16.29	21.00	12.04	14.49	47.95	15.81
Jg-LXMERT	18.33	22.41	15.52	16.53	45.68	17.42
ViLT	19.87	26.06	15.67	18.03	46.76	19.10
Jg-ViLT	20.44	26.36	16.11	19.25	40.93	19.44
BridgeTower	19.95	33.02	14.47	16.46	51.62	18.59
Jg-BridgeTower	22.20	31.47	18.31	19.95	46.98	19.63

Table 3: Overall and paragraph-based exact matching performance between Joint-grained(Jg) models and vanillas on the Test set.



(a) Exact Match Acc. (EM) (b) Multilabel Recall (MR)

Figure 7: Visualised breakdown performance of each model across different input page ranges.

7.3 Real-World Scenarios

Model	Overall	Intro	M&M	R&D	Conl	Other	Table	Figure
Vanilla BridgeTower	22.37	33.02	14.47	16.46	51.62	18.59	50.03	46.15
Jg-BridgeTower	22.20*	31.47	18.31	19.95	46.98	19.63	N/A	N/A
Jg-BridgeTower-***PDFMiner***	23.56	31.94	15.80	19.11	52.59	19.10	44.93	46.86
Jg-BridgeTower-***OCR***	23.25	29.50	16.61	17.82	51.08	17.68	55.07	53.14

* *Note*: Jg-BridgeTower exclusively handles paragraph-based questions, rendering its results non-comparable with others directly.

Table 4: Comprehensive Breakdown Performance: BridgeTower Joint-grained frameworks based on various sourced textual token sequences, overall and super-Section based breakdown.

To demonstrate the real-world efficacy of our proposed Joint-grained framework, we evaluated its performance using text extracted from off-the-shelf tools. Because BridgeTower, highlighted in Table 3, exhibits significant improvements, we present the performance of BridgeTower-based Joint-grained frameworks on various text token sequences from the MMVQA dataset (Jg-BridgeTower), PDF parser (Jg-BridgeTower-PDFMiner), and OCR tools (Jg-BridgeTower-OCR). As shown in Table 4, incorporating fine-grained tex-

tual information results in performance enhancements, increasing from 22.37% to 23.56% (PDFMiner) and 23.25% (OCR) in overall. In addition, high structural complexity sections (e.g., *M&M*, *R&D*) show notable improvements, particularly in MMVQA, reaching around 4.5% in *M&M* and 3.5% in *R&D*. This may be attributed to the "*context*" provided by the MMVQA dataset, extracted from XML nodes containing prior knowledge. Despite inherent noise raised by off-the-shelf tools, they still yield substantial improvements. Notably, OCR, while facing challenges with mis-detected characters, demonstrates considerable increases in retrieving *Table* (about 5%) and *Figure* (7%) based questions. However, *Introduction* (*Intro*) shows a decreasing trend after the incorporation of fine-grained information. This could be due to the introduction covering the entire document content, making learning the relations between tokens and entities more challenging. Future work may explore more refined Joint-grained aligning methods.

7.4 Category-Oriented Entity Representation

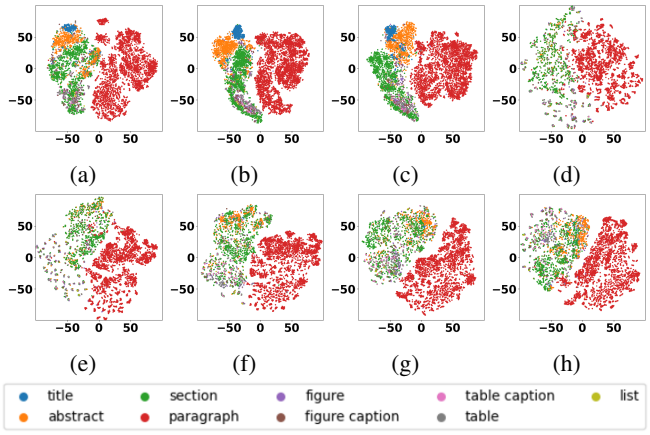


Figure 8: Category-oriented entity representation T-SNE analysis of various frameworks including (a) Transformer, (b) VisualBERT, (c) LXMERT, (d) CLIP, (e) ViLT, (f) BridgeTower, (g) Jg-BridgeTower-PDFMiner, (h) Jg-BridgeTower-OCR.

To understand the insight of document entity representations of each framework, two-dimensional *T-SNE* analysis is performed on final entity embeddings extracted from decoder $\mathcal{D}$, as shown in Figure 8. In general, RoI-based frameworks tend to have more representative feature embedding in understanding the semantic roles of each document entity. Especially compared with unclear boundaries between various text-dense entities such as *Abstract*, *Title*, *Paragraph*, RoI-based models can effectively distinguish them. However, RoI-based models underperform compared to Patch-based models, as shown in Table 2. The possible reason is although they benefit from pre-trained backbones and are good at learning visual cues within document entity RoIs, they lack in addressing the broader document layout and the relationships between question and target entities, crucial for understanding multi-page documents[14].

[14]We conducted an additional question-answering embedding correlation analysis in Appendix G.2.

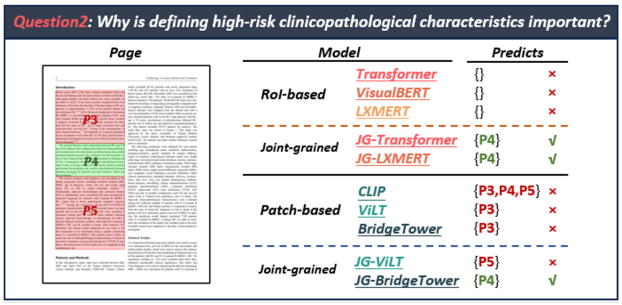


Figure 9: Qualitative analysis of various model performance on two sample questions.

For RoI-based frameworks, Transformer underperforms VisualBERT and LXERMT in *Table* and *Figure* question types (refer to Appendix G.1.). This performance gap can be attributed to the distinctiveness of entity embeddings for *Figure* and *Table*, as shown in Figures 8b and 8c for VisualBERT and LXMERT, respectively, compared to Transformer (Figure 8a). Additionally, for Patch-based models, BridgeTower outperforms other counterparts on paragraph-based questions. This may be linked to BridgeTower's focused pre-training on textual content and clearer clustering of text-dense entities as illustrated in Figure 8f. Moreover, compared to the vanilla BridgeTower framework (Figure 8f), Joint-grained information-augmented models (Figure 8g, 8h) tend to have more representative entity representations, especially for text-dense document entities, e.g. *Abstract*, *Section*.

7.5 Qualitative Analysis

To demonstrate the effectiveness of proposed frameworks, especially the benefits of joint-grained frameworks, we represent the predictions of various architectures and analyse them qualitatively. As shown in Figure 9, all RoI-only frameworks failed to identify the correct answer paragraph (***P4***); however, integrating patch embeddings enables the models to locate the surrounding entities (***P3***, ***P5***) of the target (***P4***), which demonstrates patch information could bring more comprehensive layout understanding. After joint-grained frameworks incorporate fine-grained information achieve correct predictions, underlining the effectiveness of fine-grained data in improving entity representation robustness. [15]

8 Conclusion

This paper presents a contribution by introducing the MMVQA dataset and a novel joint-grained architecture. The MMVQA from PubMed Central showcases diverse document types, complex structures, and extensive content-related questions in multi-page documents. We also introduce the strong benchmark, Joint-grained retrieval architecture, which consistently enhances model performance, particularly in complex document sections. We hope this research could not only advance the understanding of multi-page document comprehension but also set a foundation for future exploration and refinement of models in this domain, marking a significant step forward in document understanding research.

[15]Please refer to Appendix H to check more qualitative samples.

Acknowledgments

We express our profound gratitude to all the authors—Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han—for their critical contributions to this project. Their combined expertise, associated with The University of Melbourne, The University of Sydney, and The University of Western Australia, has been essential in advancing this research. We are grateful for the unwavering support from these institutions, which provided the necessary resources and conducive environments for our studies. Furthermore, we appreciate the insightful feedback from our peers and reviewers, which has greatly enhanced the quality of our work. We hope that our research will make a meaningful impact in the field.

References

[Ding *et al.*, 2023a] Yihao Ding, Siqu Long, Jiabin Huang, Kaixuan Ren, Xingxiang Luo, Hyunsuk Chung, and Soyeon Caren Han. Form-nlu: Dataset for the form natural language understanding. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2807–2816, 2023.

[Ding *et al.*, 2023b] Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. Pdf-vqa: A new dataset for real-world vqa on pdf documents. In Gianmarco De Francisci Morales, Claudia Perlich, Natali Ruchansky, Nicolas Kourtellis, Elena Baralis, and Francesco Bonchi, editors, *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*, pages 585–601, Cham, 2023. Springer Nature Switzerland.

[Ding *et al.*, 2024] Yihao Ding, Lorenzo Vaiani, Caren Han, Jean Lee, Paolo Garza, Josiah Poon, and Luca Cagliero. M3-vrd: Multimodal multi-task multi-teacher visually-rich form document understanding. *arXiv preprint arXiv:2402.17983*, 2024.

[Du *et al.*, 2022] Qinyi Du, Qingqing Wang, Keqian Li, Jidong Tian, Liqiang Xiao, and Yaohui Jin. Calm: Commensense knowledge augmentation for document image understanding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3282–3290, 2022.

[Gu *et al.*, 2021] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50, 2021.

[Hofstätter *et al.*, 2023] Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. Fid-light: Efficient and effective retrieval-augmented text generation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1437–1447, 2023.

[Hu *et al.*, 2023] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23369–23379, 2023.

[Huang *et al.*, 2022] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

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

[Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[Li *et al.*, 2021] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660, 2021.

[Liu *et al.*, 2023a] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[Liu *et al.*, 2023b] Xuejing Liu, Wei Tang, Xinzhe Ni, Jinghui Lu, Rui Zhao, Zechao Li, and Fei Tan. What large language models bring to text-rich vqa? *arXiv preprint arXiv:2311.07306*, 2023.

[Long *et al.*, 2022] Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. Vision-and-language pretrained models: A survey. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5530–5537. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Survey Track.

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

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

[McHugh, 2012] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin,

Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

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

[Tan and Bansal, 2019] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019.

[Tanaka *et al.*, 2021] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888, 2021.

[Tanaka *et al.*, 2023] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645, 2023.

[Tito *et al.*, 2021] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 778–792. Springer, 2021.

[Tito *et al.*, 2023] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wang *et al.*, 2022] Jiapeng Wang, Lianwen Jin, and Kai Ding. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7747–7757, 2022.

[Wu *et al.*, 2022] Xinya Wu, Duo Zheng, Ruonan Wang, Jiashen Sun, Minzhen Hu, Fangxiang Feng, Xiaojie Wang, Huixing Jiang, and Fan Yang. A region-based document vqa. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4909–4920, 2022.

[Xu *et al.*, 2021] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multimodal 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 (Volume 1: Long Papers)*, pages 2579–2591, 2021.

[Xu *et al.*, 2023] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10637–10647, 2023.

[Yasunaga *et al.*, 2022] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022.

[Ye *et al.*, 2023] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 174–184, 2023.

[Yu *et al.*, 2022] Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Structextv2: Masked visual-textual prediction for document image pre-training. In *The Eleventh International Conference on Learning Representations*, 2022.

[Zaheer *et al.*, 2020] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

[Zhao *et al.*, 2023] Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4736–4756, 2023.

[Zhong *et al.*, 2019] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019.