Visually Rich Document Understanding and Intelligence

YIHAO DING

Doctor of Philosophy

Supervisor: Dr. Caren Han
Associate Supervisor: Dr. Josiah Poon

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

4 November 2024

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Name: Yihao Ding

Signature:

Abstract

Visually Rich Documents (VRDs) are potent carriers of multimodal information widely used in academia, finance, medical fields, and marketing. Traditional approaches to extracting information from VRDs rely on expert knowledge and manual labour based on predefined rules, leading to high costs. However, significant gaps still exist between current research focuses and the real-world nature of VRDs. Most existing research focuses on simple or well-structured domains, such as receipts and academic papers, often limited to single-page scenarios.

This research introduces three research directions to address these gaps: 1) Neglect of Complex VRD Structures, 2) Lack of Multi-Page Scenarios Exploration, and 3) Lack of Exploration in Large Model Integration. For complex format VRD understanding, a new dataset for complex form structure and key information extraction is introduced to address complex form structures and multi-party(authors) information gaps. A new key information extraction baseline is introduced with multi-aspect features. To address the challenges in multi-page VRD understanding, a new dataset for multimodal, multi-page VRD question answering is proposed to retrieve multimodal document semantic entities across several pages based on a question. A set of frameworks are proposed to leverage various pretrained models to enhance the representative of multi-page document representations. Finally, to address the gaps in knowledge transfer from large-scale models for VRD understanding, a new knowledge distillation-based framework is proposed to distil knowledge from multiple large-scale models by introducing multi-task losses to transfer multi-teacher knowledge to lightweight student models, achieving more comprehensive document representation. These proposed datasets and experiments are analysed through qualitative and quantitative analysis, demonstrating the insights and effectiveness of the introduced datasets and proposed models in advancing VRD understanding and bridging the gaps between VRD research and practical applications.

Acknowledgements

First and foremost, I want to express my deepest gratitude to my family—my parents, my grandmother, my aunt, and other relatives—for their unwavering love and support throughout my life. "Family" has always been my greatest pillar, accompanying me on every step of my journey. Even though I rarely say it out loud, I hope they know how much I appreciate their endless care. My grandmother and aunt always remind me not to overwork myself, my mother urges me to take care of my health, and my father worries if I have enough money (not only money of course). Although our conversations often seem repetitive, they are a source of immense comfort and strength to me.

Special thanks to Darren, who walked with me through journey after journey.

I am profoundly grateful to my supervisor, Caren, for her guidance, encouragement, and passion for research, which have inspired me and will continue to influence many others. I also want to extend my heartfelt thanks to Josiah, who has been an invaluable mentor, providing support both in research and in my personal life. I am especially grateful to Caren and Josiah for cultivating a supportive research group that transcends the boundaries of different cities.

To my friends, I am thankful for your companionship, laughter, and support during these fleeting yet enduring years. Every casual chat, unexpected message, brief reunion, and long-term connection has been precious. From shared meals, ice cream after dinner, beers on the way home, and indecisive restaurant choices, to udon and Lanzhou noodles—your presence has made all the difference.

I would like to express my heartfelt gratitude to all the co-authors from our group who have been a part of this journey: Dr. Caren Han, Dr. Siwen Luo, Dr. Siqu Long, Dr. Kunze Wang, Dr Jie Yang, Dr Jean Lee, Feiqi Cao, Rina Cabral, Tianyi Chen, David Chung, Lorenzo Vaiani. And meal buddies across different cities and stages of life including Dr. Dr. Henry Weld, Chen Chen, Xionghong Guo, Teajun Lim, Zhihao Zhang, Yue Dai, Yan Li, Shan

Ng, Tianyi Zhang, Kaixuan Ren, Zechuan Li, Yanbei Jiang, Shuo Yang. Thank you for the companionship, whether long or short.

This journey feels like a train ride with countless stops in different cities, each with its unique sights and experiences. Though my journey has not ended, I am forever grateful for the family, friends, and mentors who have made it so remarkable.

Authorship attribution statement

(1) **Chapter 3 of this thesis is published as [26].** (SIGIR 2023)
I was responsible for formulating the research aim, designing the methodology, conducting the dataset annotation and analysis, implementing the experiments, and writing the entire paper.

(2) **Chapter 4 of this thesis is published as [28].** (IJCAI 2024)
I formulated the research aim, designed the dataset structure and construction, analysed and visualised the dataset, designed the methodology and implementation, and wrote the whole paper.

(3) **Chapter 5 of this thesis is published as [29].** (ACL 2024)
I was responsible for formulating the research aim, designing the methodology, implementing the experiments, and writing the entire paper.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Name: Yihao Ding
Signature:
Date: 27 September 2024

Name: Caren Han

Signature:

Date: 23 October 2024

Name: Josiah Poon

Signature:

Date: 23 October 2024

Publication List

Ding, Y., Ren, K., Huang, J., Luo, S., & Han, C. (April 2024). MMVQA: A Dataset for Multimodal Information Retrieval in PDF-based Visual Question Answering. *In Proceedings of the 33rd International Joint Conference on Artificial Intelligence* (Rank A* Top-tier Conference)

Ding, Y., Vaiani, L., Han, C., Lee, J., Garza, P., Poon, J., & Cagliero, L. (May 2024). 3MVRD: Multimodal Multi-task Multi-teacher Visually-Rich Form Document Understanding. *Findings of the Association for Computational Linguistics: ACL 2024* (Rank A* Top-tier Conference)

Ding, Y., Long. S., Huang, J., Ren, K., Luo, X., Chung, H. & Han, C. (July 2023). Form-NLU: Dataset for the Form Natural Language Understanding. *In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Rank A* Top-tier Conference)

Ding, Y., Luo, S., Chung, H. & Han, C. (September 2023). PDF-VQA: A Dataset for VQA on PDF Documents. *In Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (Rank A Top-tier Conference)

Ding, Y., Zhe, H., Wang, R., Zhang, Y., Chen, X., Ma, Y., Chung, H. & Han, C. (June 2022). V-Doc: Visual questions answered with Documents. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Rank A* Top-tier Conference)

Ding, Y., Luo. S, Long, S., Poon, J & Han, C. (October 2022). Doc-GCN: Heterogeneous Graph Convolutional Networks for Document Layout Analysis. *In Proceedings of the 29th*

International Conference on Computational Linguistics (Rank A Top-tier NLP Conference)

Wang, K.,**Ding, Y.** & Han, C (May 2024). Graph Neural Networks for Text Classification: A Survey. Accepted at *Artificial Intelligence Review* (Rank A Top-tier Journal)

Chen, T., Cao, F., **Ding, Y.** & Han, C (Feb 2024). The Language Model Can Have the Personality: Joint Learning for Personality Enhanced Language Model. *In Proceedings of The 38th Annual AAAI Conference on Artificial Intelligence* (Rank A* Top-tier Conference)

Han, C, **Ding, Y.**, Luo. S & Poon, J. (October 2023). Workshop on Document Intelligence Understanding. *In In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Rank A Top-tier Conference)

Yang, J., **Ding, Y.**, Long, S., Poon, J., & Han, C. (March 2023). DDI-MuG: Multi-Aspect Graphs for Drug-Drug Interaction Extraction. *Frontiers in Digital Health Volume 5*

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

Visually Rich Documents (VRDs) are documents that contain a diverse mix of visual and textual elements designed to convey information in a comprehensive and visually engaging manner, which are pervasive in our daily lives, appearing in domains such as finance, medicine, and academia. These documents integrate various types of visually rich elements, including *paragraphs*, *tables*, *charts*, *diagrams*, and *photos*. Both textual elements (such as paragraphs, lists, and captions) and visually rich elements (like tables and figures), collectively referred to as document semantic entities, are essential for illustrating, explaining, and summarising information. Common formats for VRDs include *PDF* (Portable Document Format), *DOC/DOCX* (Microsoft Word Document), and *image files* (JPEG, PNG). These documents are typically structured in a semi-structured or even unstructured manner, making their understanding and information extraction challenging.

The VRD Understanding (VRDU) field is dedicated to comprehending the structure of VRDs and extracting pertinent information, thus transforming unstructured or semi-structured content into a machine-readable format. VRDU mainly encompasses two tasks: Key Information Extraction (KIE) and Question Answering (QA). **Key Information Extraction** (KIE) aims to identify and extract values based on predefined keys. Depending on the model used, KIE can be approached as an entity retrieval task (KIE Task Formulation 1) or a sequence tagging task (Task Formulation 2), as shown in Figure 1.1. **Visually Rich Document Question Answering** (VRD QA) involves answering natural language questions based on the contextual information in VRDs. As explained in Figure 1.1, the model must locate the answer within

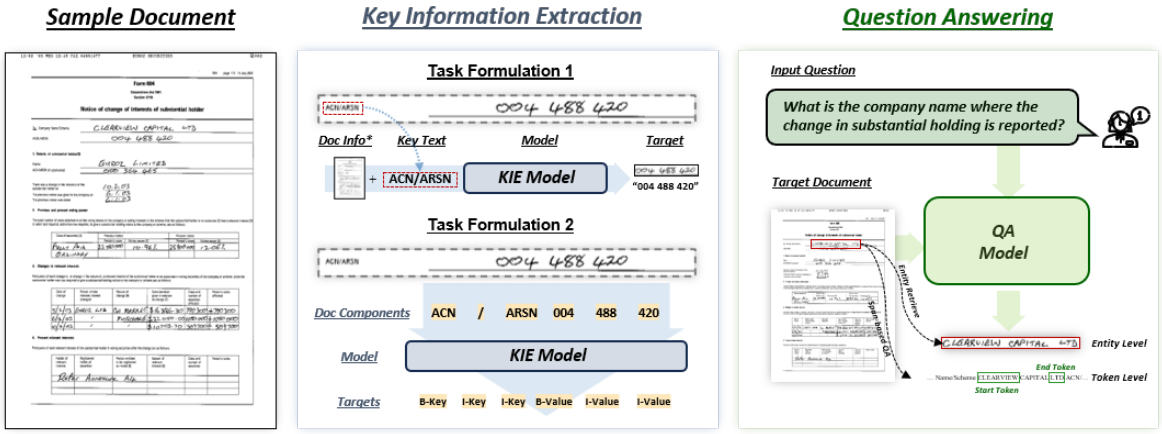


FIGURE 1.1. A sample with two main VRD tasks: Key Information Extraction and Question Answering.

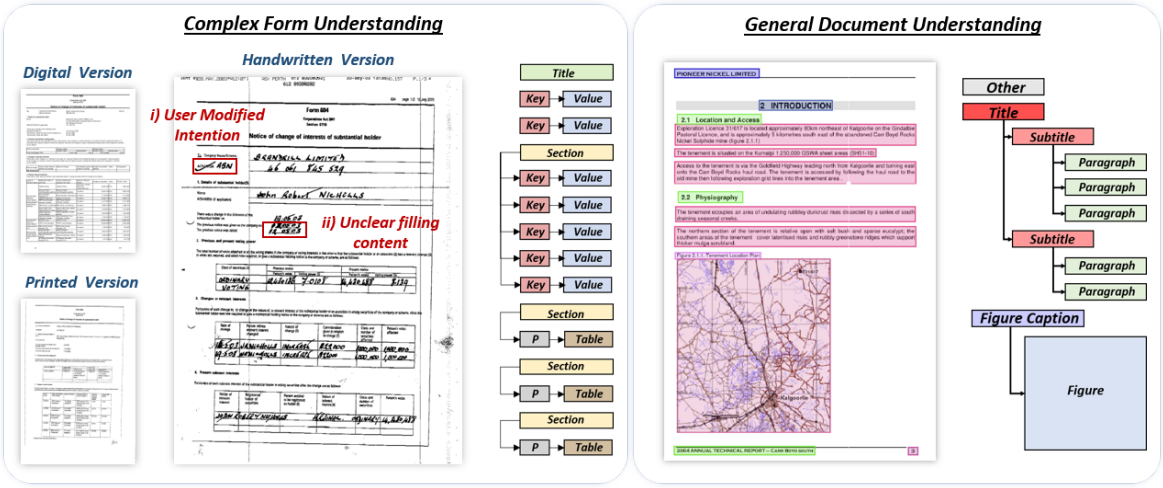


FIGURE 1.2. Structural comparison between forms and general VRDs.

the document to respond to the input question. Beyond extractive methods, KIE and VRD QA can also be formulated as generative tasks, where the model generates the desired value or answer auto-regressively based on the query or question and the document images.

Existing frameworks are mainly based on deep learning models and have achieved state-of-the-art performance on KIE (see Section 2.2) and QA tasks (see Section 2.3), as well as in learning comprehensive document representations for multiple VRDU downstream tasks as introduced in Section 2.4. Despite these advancements, significant gaps remain between current research focuses and the real-world nature of VRD:

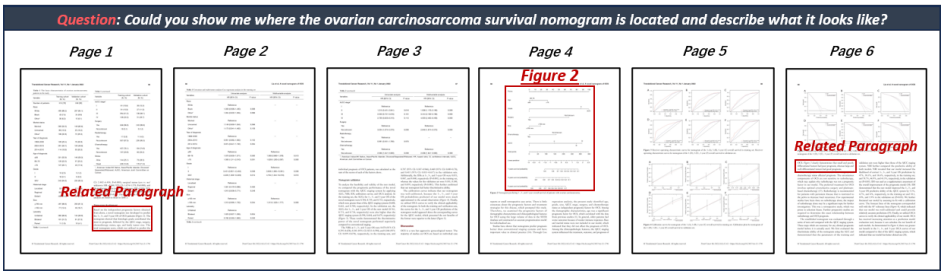


FIGURE 1.3. An illustrative scenario demonstrating the multi-page and multimodal information for answering a question.

(1) **Neglect of Complex VRD Structures**: Existing approaches predominantly concentrate on well-organised document domains, often overlooking the complexities inherent in structures like VRD forms, which involve multiple stakeholders, such as designers and users. **Visually Rich Form Document Understanding** is crucial since forms play a critical role across diverse sectors, including political, financial, and business domains. Their complex layout structures involve multiple stakeholders throughout the design, data collection, transmission, and storage phases, posing challenges in comprehending form structures and extracting valuable information from form images. As Figure 1.2, forms exhibit a more intricate logical structure than general documents such as technical reports or academic papers. They are collected in various formats—including digital, printed, and handwritten—based on user preferences and logistical considerations, thereby introducing complexities in downstream processing. Moreover, forms are completed by multiple stakeholders, leading to inherent information gaps between form designers and users. For instance, while designers may specify collecting '*ACN*' from holders, users might interchangeably use '*ABN*' due to availability. Inconsistently empty fields and unclear handwriting further compound these challenges. Addressing these complexities through automated methods for comprehending and extracting information from forms holds promise for cost-effectiveness and operational efficiency improvements.

(2) **Lack of Multi-Page Scenarios Exploration**: There is a lack of exploration in developing approaches that address multi-page scenarios, which are more common in practical applications. Since DocVQA [84] introduced the VRD QA task, significant advancements [45, 139] have been made in handling single-page VRD scenarios.

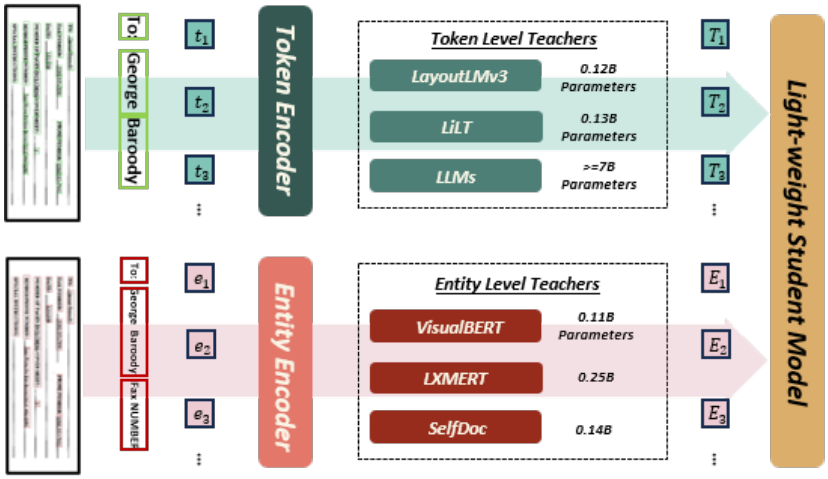


FIGURE 1.4. A sample workflow to distil implicit knowledge from multiple large-scale pretrained models.

However, the real-world has more Multi-page VRDs with interconnected content. Moreover, VRDs frequently feature visually rich elements like tables and charts (see Figure 1.3), which are typically overlooked in current VRD QA datasets [27, 84]. For instance, addressing a query like, "*Could you show me where the ovarian carcinosarcoma survival nomogram is located and describe what it looks like?*" requires locating *Figure 2* and extracting relevant descriptions and captions across multiple pages. Addressing these challenges requires developing effective approaches for **Multi-page VRD Question Answering**.

(3) **Lack of Exploration in Large Models Integration**: Most existing state-of-the-art VRDU models are large-scale pretrained models. However, limited research is focused on leveraging knowledge from various existing large-scale pretrained models. Many approaches opt to train new models directly instead. Addressing these gaps is crucial for advancing the field and enhancing the applicability of VRDU tasks in real-world settings. Inspired by pretrained language models like BERT and RoBERTa [24, 75], numerous pretrained VRDU models [68, 140] have been developed. The choice of pretraining corpus and the nature of the pretraining tasks can influence the types of implicit knowledge these models acquire, contributing to more comprehensive and robust document representations. Form understanding encompasses various challenges, including complex structures, content that can easily confuse

both designers and users and diverse formats. Leveraging pretrained backbones to integrate this implicit knowledge has proven to be an effective strategy. However, these pretrained models often focus on different granularities of information with heavy model architecture. For instance, as Figure 1.4 represented, fine-grained models like LayoutLMv3 and LiLT [45, 127] concentrate on the sequence of text tokens, while coarse-grained models like VisualBERT and LXMERT [66, 114] focus on broader document semantic entities. Exploring effective methods to **transfer knowledge from multiple large-scale pretrained teachers** to a lightweight student model is a significant yet largely unexplored area in VRDU, particularly for understanding complex forms.

These substantial gaps persist between current research focuses and VRDs' real-world nature. Neglect of complex structures, such as forms involving multiple stakeholders, a lack of exploration into multi-page scenarios common in practical applications, and insufficient utilisation of large-scale pretrained models are notable deficiencies. Addressing these gaps is crucial for advancing the field and enhancing the applicability of VRDU tasks in real-world settings. In the following section, I will discuss the research aims and contributions aligned with addressing these stated problems.

1.2 Research Aims

In this thesis, the focus is on building practical solutions for understanding real-world visually rich documents. Aligned with the aforementioned three research problems in the previous section, the thesis pioneered three new research directions: 1) how to understand visually-rich form documents with a complex structure and multiple stakeholders, 2) how to deal with multi-page VRD understanding and question answering, and 3) how to leverage VRD understanding expertise from multiple large-scale pretrained models.

To understand the complex VRD form structure, a novel dataset is introduced specifically designed for form structure understanding and key information extraction, which includes a variety of form types, such as digital, printed, and handwritten. To complement this dataset,

a proposed baseline model leverages robust positional and logical relationships to capture hierarchical document structures spanning from individual words to document semantic entities. The effectiveness of this proposed model is evaluated using off-the-shelf PDF layout extraction tools in real-world scenarios, ensuring its practical applicability and performance in typical use cases **To address the challenges in multi-page VRD QA**, a new dataset is introduced, specifically designed for multi-page, multimodal document entity retrieval, aiming to significantly enhance information retrieval systems in knowledge-intensive environments. By utilizing multimodal document entities, the effectiveness of retrieval-based models is refined, enabling better handling of complex queries and diverse document structures. Furthermore, methods to accurately capture and utilise cross-page layouts and logical correlations are being explored, which are essential for advancing document understanding and improving the accuracy of entity retrieval. The practical applicability of existing Visual Language Pretrained Models (VLPMs) and other pretrained language models is assessed through their demonstrated ability to precisely locate target entities, underlining their utility in real-world scenarios. This initiative seeks to push the boundaries of document VQA and establishes a foundation for future innovations in the field. **To integrate knowledge from multiple large-scale pretrained models in VRD**, a novel joint-grained document understanding approach has been introduced, integrating multimodal multi-teacher knowledge distillation. This approach utilises knowledge from task-specific pre-trained models (teachers) to enhance the inclusiveness and representation of multimodal and joint-grained document representations. The proposed model can comprehensively capture and analyse complex multimodal relationships within form documents, thereby improving overall understanding and usability for both designers and users.

1.3 Contributions

In summary, this research explores effective methods to address three pivotal document understanding tasks: complex form understanding, multi-page document visual question answering, and knowledge distillation for VRDU. These initiatives are designed to expand the scope and accelerate the development of VRDU, significantly boosting its practical

applications and impact across the domain. Specifically, for complex form understanding, a new dataset focusing on form structure and content understanding has been proposed, along with a multi-aspect feature-enhanced baseline model designed to tackle complicated structures and mitigate information gaps between multiple form stakeholders during key information extraction. To extend the scope of VRD QA to conduct multi-page and multimodal information retrieval, a new dataset and a set of frameworks have been introduced to mitigate the limitations of existing VRDU models and leverage knowledge from VLPMs for multi-page multimodal information retrieval. Finally, to better utilise knowledge from pretrained teacher models, a novel framework has been introduced that uses multi-task learning to distil knowledge from multiple teachers, enhancing the robust document representation, especially for form understanding.

The detailed contributions of each work are summarised as follows:

(1) **Visually Rich Form Document Understanding:**
   (a) introduces Form-NLU, a new form structure understanding and key information extraction dataset that covers specific form designers' intentions and aligns with the user's values;
   (b) proposes a new baseline model that handles form documents' positional and logical relations and hierarchical structure. The proposed model has outperformed other SOTA form key information extraction models on Form-NLU;
   (c) applies the proposed dataset and the model with the off-the-shelf pdf layout extraction tool and proves the feasibility in real-world form document cases.

(2) **Multi-page Visually Rich Document Question Answering:**
   (a) introduces MMVQA, a new VRD QA dataset for retrieving multimodal document semantic entities in multi-page VRDs, accompanied by versatile metrics for diverse scenarios;
   (b) proposes a set of frameworks for multi-page document entity retrieval, leveraging the implicit knowledge from VLPMs and fine-grained level information to enhance model effectiveness and robustness;

(c) performer a series of experiments combining quantitative and qualitative analyses to provide deeper insights into MMVQA and demonstrate the effectiveness of the proposed techniques for multimodal multi-page document entity retrieval.

(3) **VRD Understanding Knowledge Transfer from multiple large-scale pretrained models:**

(a) presents a groundbreaking multimodal, multi-task, multi-teacher joint-grained knowledge distillation model designed explicitly to understand visually rich form documents;

(b) the proposed model outperforms across publicly available form document datasets;

(c) to the best of my knowledge, this research marks the first in adopting multi-task knowledge distillation, focusing on incorporating multimodal form document components;

1.4 Thesis Outline

Chapter 2 presents a comprehensive literature review on visually rich document understanding (VRDU), specifically focusing on key information extraction (KIE), visually rich document question answering (VRD QA), and multi-task VRDU frameworks. It systematically reviews benchmark datasets in these areas, providing an in-depth analysis of current research trends.

Chapter 3 is focused on the first aim and contribution, 'Visually Rich Form Document Understanding', and is based on the work ***Form-NLU: Dataset for Form Natural Language Understanding*** [26], published at **The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)**. This work focuses on deep learning based form intelligent understanding. To mitigate the gaps in the lack of a complex form understanding dataset, a novel dataset, Form-NLU, for form structure understanding and key information extraction is introduced for interpreting the form designer's intent and alignment of user-written values. A novel form key information extraction model is introduced by integrating multi-aspect features and entity-token joint-grained information

to get more comprehensive form representations and logical-spatial structure understanding. Extensive experiments and case studies are conducted and analysed in this work to reveal more insight into the dataset and effectiveness of proposed models.

Chapter 4 is focused on the second aim and contribution, 'Multi-page Visually Rich Document Question Answering', and is based on the work ***MMVQA: A Comprehensive Dataset for Investigating Multipage Multimodal Information Retrieval in PDF-based Visual Question Answering*** [28] accepted at **The 33rd International Joint Conference on Artificial Intelligence** (**IJCAI 2024**). This work focuses on multi-page retrieval-based visual question answering on VRD. To mitigate the gaps in multi-modal multi-page VRD QA, a new dataset, MMVQA, is proposed, which is the first VRD QA dataset for retrieving multimodal document semantic entities in multi-page VRDs. Diverse metrics such as Exact Matching Accuracy, Partial Matching Accuracy, and Multi-label Recall are employed to evaluate the datasets accurately based on real-world application scenarios. The chapter also introduces several multi-page retrieval-based QA frameworks that utilise the implicit knowledge of Vision-Language Pretrained Models (VLPMs). Extensive quantitative and qualitative analyses are conducted to provide deeper insights into the dataset's structure and demonstrate the proposed frameworks' effectiveness.

Chapter 5 is focused on the third aim and contribution, 'VRD Understanding Knowledge Transfer from multiple large-scale pretrained models', and is based on the work ***3MVRD: Multimodal Multi-task Multi-teacher Visually-Rich Form Document Understanding*** [29], accepted at **The 62nd Annual Meeting of the Association for Computational Linguistics** (**ACL 2024**). This work focuses on knowledge distillation for visually rich document understanding (VRDU). This work introduces a multimodal, multi-task, multi-teacher joint-grained knowledge distillation model to mitigate the limitations of fine-grained or coarse-grained models without heavily pretraining a joint-grained framework. This framework, named 3MVRD, introduces various inter/intra-grained knowledge distillation losses to effectively distil multimodal implicit knowledge from multiple teacher models. The lightweight student model can leverage the knowledge distilled from multiple fine-tuned teachers to generate more comprehensive representations and achieve better performance.

Chapter 6 summarises the research findings and outlines future directions, focusing on robust document representation, relation-aware VRD visual question answering, and domain adaptation in document understanding.

CHAPTER 2

Literature Review

2.1 Overview

Visually rich document content understanding encompasses several independent downstream tasks tailored to different application scenarios and user demands. Firstly, methods focused on specific document understanding tasks are summarised. Many of these models aim to provide comprehensive document representation and integrate target-oriented modules or techniques to improve performance and efficiency across various VRDU tasks. The models for two VRDU downstream tasks—Key Information Extraction (KIE) (Section 2.2) and Question Answering (QA) (Section 2.3), as shown in Figure 2.1 are introduced and summarised with insights into current trends. Then, multi-task document understanding frameworks (Section 2.4, listed in Figure 2.2, are introduced, focusing on generating more comprehensive document representations to achieve delicate performance on several document understanding tasks. Before delving into these frameworks in detail, I will provide a general overview of the development of VRDU.

Heuristic methods [104, 105, 135] and statistical machine learning techniques [86] have been utilised in domain-specific document applications, requiring expert customization and proving challenging to update. Recent advances in deep learning have introduced promising alternatives. LSTM and CNN-based models [23, 53, 153], feature-driven approaches [128, 146, 150], and layout-aware pretrained frameworks [42, 127, 140], along with visual-integrated pretrained frameworks [45], have significantly improved document representation and achieved state-of-the-art performance in various downstream tasks.However, these models primarily focus on fine-grained features, such as grids and word/word pieces (textual tokens), and often

struggle with computational complexity and the ability to capture global logical and layout correlations. In contrast, coarse-grained models, which operate at the entity level, address some limitations of fine-grained models but may omit detailed information, resulting in less comprehensive representations. To bridge these gaps, joint-grained frameworks [3, 69, 147] and LLM-based frameworks [71, 79] have been developed.

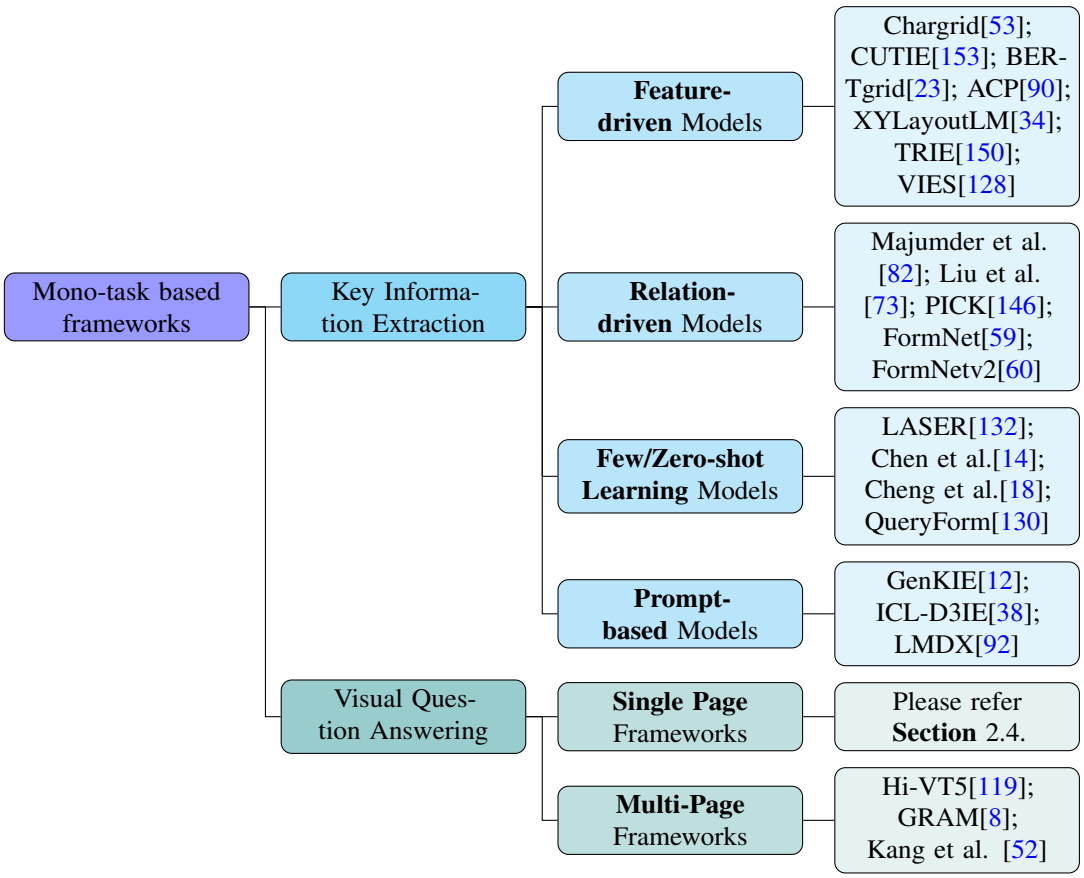


FIGURE 2.1. Mono-task visually rich document understanding frameworks.

2.2 Visually Rich Document Key Information Extraction

Key Information Extraction (KIE), a typical natural language processing task, refers to the task of identifying and extracting crucial pieces of information from textual data. Unlike typical name entity recognition methods for addressing plain text, VRDs contain visually rich entities like *tables* and *charts*, as well as spatial and logical layout arrangements to enhance the challenge of extracting crucial information. Although plain-text pretrained language models,

such as BERT [24], RoBERTa [75], and ALBERT [58], are widely used as solid baselines on many benchmark datasets, more recent works introduce layout-aware pretrained models such as LayoutLM families [45, 139, 140], LiLT [127], Bros [42] to enhance the document representation by leveraging visual and layout information and achieving SoTA performance on several downstream tasks (see Section 2.4). This section will mainly focus on models specifically proposed for the document KIE models or only evaluated on KIE benchmark datasets including FUNSD [49], CORD [91], SROIE [47], XFUND [141], etc. Based on innovative aspects, we categorise KIE frameworks into five types: *Feature-driven models* use multimodal cues for rich feature representation. *Joint Learning frameworks* integrate auxiliary tasks to enhance document representation. *Relation-aware models* leverage spatial or logical relations via graphs or masked attention mechanisms. *Few/Zero-shot learning frameworks* explore methods for extracting key information with minimal labelled data, often using transfer learning. *Prompt-based frameworks* use structured prompts to guide specific information extraction from pretrained models or LLM/MLLMs. These categories represent diverse approaches to improving Key Information Extraction (KIE) by leveraging specific model designs and learning strategies tailored to document understanding challenges.

2.2.1 Feature-driven Models

In the initial phase, certain Recurrent Neural Network (RNN)-based models [57] were introduced, primarily focusing on key information extraction tasks from plain text. However, these approaches overlook the importance of visual cues and layout information. Hence, several multimodal frameworks with feature-driven designs have been proposed to generate more representative document representations.

Chargrid [53] first mentioned the significance of 2D structure for document KIE and designed a character-box-based chargrid to convert the textual and 2D layout structure into coloured **visual cues** feed into CNN. Instead of using fine-grained character information, ***CUTIE*** [153] and ***BERTgrid*** [23] utilise various word embedding methods with bounding box coordinates to persevere the layout structure. ***ACP*** [90] leverages attention mechanism and multi-aspect

features, including visual, semantic (both character and word) and spatial, into dilated CNN for capturing both short and long-term dependencies of each word piece.

XYLayoutLM [34] introduces an Augmented XY Cut module to correct the improper reading order generated by OCR engines and a Dialted Conditional Position Encoding module to handle the variable lengths of input text and image tokens. ***LayoutXLM*** [141] [1] is adopted as a basic framework to process multimodal inputs. The Augmented XY Cut module enhances traditional XY Cut [85] by incorporating thresholds (λ_x, λ_y) and a shift factor (θ) to adjust box positions on the x and y axis based on randomly generated values. It recursively divides token boxes using valleys in horizontal and vertical projection profiles, forming clusters in descending order. Each cluster's reading order is determined recursively, prioritizing divisions until no significant valleys remain, ensuring a proper reading order is derived from an XY Tree structure. DCPE (Dilated Conditional Position Encoding) addresses limitations of CPE [20] in multimodal tasks by separately processing textual and visual features. It employs 1D convolutions for textual tokens to extract 1D local layouts, accommodating their inherent 1D relationships. Additionally, DCPE utilises dilated convolutions [145] to capture long-range dependencies effectively without increasing model complexity, thereby enhancing performance in multimodal document understanding tasks.

Joint learning KIE frameworks are proposed to leverage multi-aspect features to mitigate the information gap between various focused tasks. ***TRIE*** [150] firstly offers an end-to-end framework for simultaneously conducting Optical Character Recognition (OCR) and KIE. OCR module will generate multi-aspect features, including positional, visual and textual aspects. Adaptively trainable weighting mechanisms are adopted to generate the fused embedding, followed by a Bi-LSTM-based Entity Extraction Module to conduct the final prediction. ***VIES*** [128] is introduced to use the multi-level cues of vision, position and text to generate more comprehensive representations. Both token and segment (entity) level positional and visual features are acquired from text detection modules, and dual-level textual features are gathered by text recognition brunch, fused by a self-attention-based fusion module for sequence labelling.

[1] Please refer to Section 2.4.1 to see a detailed description of the LayoutXLM model.

2.2.2 Relation-driven Models

Instead of only leveraging multimodal information, leveraging the inherent spatial or logical relations between multi-level components may lead to more robust and comprehensive document representations. ***Majumder et al.*** [82] propose a model which starts by generating candidates for each field using type-specific detectors and key phrases. The neural model scores these candidates by learning dense representations that consider both textual content and spatial positioning, allowing for accurate information extraction across different document templates by focusing on the candidates' relevance to the fields. Graph-based frameworks are increasingly favoured for modelling document elements' spatial and logical relationships. This trend involves meticulously defining distinct graph structures and employing graph convolution techniques to incorporate these relationships into feature representations seamlessly.

Liu et al. [73] first utilise a graph convolution module on top of the BiLSTM-CRF for key information extraction. Each document is a fully connected graph where the node is the textual representation T of the document entity $E_i \in \mathbb{E}$, and the edge is related to the spatial relation between E_i and another entity E_j. It is defined as $e_{ij} = [x_{ij}, y_{ij}, \frac{w_i}{h_i}, \frac{h_j}{h_i}, \frac{w_j}{h_i}]$, where x_{ij} and y_{ij} are horizontal and vertical distance between E_i and E_j. The self-attention-based graph convolution is performed on node-edge-node triplets (concatenated nodes and edge embeddings) to learn contextually with all other nodes. ***PICK*** [146] adopts a similar way to construct the document graph but utilising multimodal node representations, including transformer-encoded textual embedding and CNN-encoded visual embedding, as well as the edge embedding is updated to $e_{ij} = [x_{ij}, y_{ij}, \frac{w_i}{h_i}, \frac{h_j}{h_i}, \frac{w_j}{h_i}, \frac{S_j}{S_i}]$ where S is the sentence length of corresponding entity. Additionally, to get the node embedding for the downstream tagging task, a soft adjacent matrix-based graph learning layer is applied to obtain the task-specific node representations.

FormNet [59] introduces an end-to-end framework with a new attention mechanism, *Rich Attention*, and graph framework to make an order/distance sensitive long sequence transformer. In Rich Attention, the model introduces two order-based and pixel distance scores along the x/y

axis. These are summed up with usual self-attention scores to enable the model order/distance awareness. Additionally, a GCN is applied to contextually learn the neighbouring token embedding before serialising the token sequence feeding into the transformer encoders. The graph nodes are text embedding of tokens, and the edges are relative positions between nodes. This could mitigate the imperfect serialization issue, enabling token representation capturing in more contexts. ***FormNetv2*** [60] integrates image modality as the additional GCN edge feature to capture more visual cues. To use contrastive loss to learn the multimodal graph representation, they first perform stochastic graph corruption to sample topology-corrupted and feature-corrupted graphs. Topology corruption randomly removes edges in the original graph, while feature corruption drops all modality features from nodes and edges. Then, the contrastive objective is to maximise agreement between a pair of tokens between corrupted and original graphs under a *standard normalised temperature-scaled cross-entropy* [109] loss.

2.2.3 Few-shot Learning Models

Extracting key information from VRDs based on deep learning models typically involves more manual work to annotate the training data. However, acquiring large-scale, high-quality annotations in urgent and labouring-limited application scenarios is challenging. Thus, few-shot or one-shot models emerged to extract the key information with less labouring annotations.

LASER [132] leverage the architecture from ***LayoutReader*** [133] to reformulate the entity recognition task to sequence-labelling to generative model by embedding the entity type information (named label surface name) into the target sequence to enable the model label semantic-aware. ***LASER*** depends on a "partially triangle" mask to use one encoder to encode source text and generate the prediction. The n source text sequence, $\{t_1, ..., t_n\}$, is the text sequence by summing the relative word, spatial, and positional embedding, feeding into the $\mathcal{E}$ with full self-attention. The target sequence only attends previous tokens. The generative formulation is defined as:

$$t_{i-1}, [B], t_i, ..., t_j, [E], e_1, ..., e_k, [T], t_{j+1} \tag{2.1}$$

where $[B]$ and $[E]$ denote start and end of entity, e_i are the label surface name; $[T]$ denotes the end of the label surface name. As label surface names and functional tokens do not exist, the learnable embeddings are applied to them to ensure the generative-progress layout-aware. A binary classification module is applied to classify the next generated token type of current generated token hidden states: *from source* or not, effectively controlling the generative series.

Chen et al. [14] introduce a novel framework for ***entity-level, N-way soft-K-shot*** extracting key information from VRDs, focusing on the challenges of extracting rare or unseen entities from documents with few examples. It leverages a meta-learning approach [88, 108], utilizing a hierarchical decoder and contrastive learning (***ContrastProtoNet***) for task personalization and improved adaptation to new entity types. Additionally, it introduces ***FewVEX***, a dataset tailored for entity-level few-shot VDER, to facilitate research and benchmarking in this area. The framework's effectiveness is demonstrated through significant improvements in robustness and performance over existing meta-learning baselines.

In the domain of **one-shot** scenarios, Cheng et al. [18] presents a novel application of graphs. They utilise attention mechanisms to seamlessly transfer spatial relationships between static text regions (key/landmark) and dynamic text regions (value/field) from support documents to query documents. This transfer enables the acquisition of label probability distributions for files. Additionally, a self-attention-based module is integrated to exploit relationships between field entities, facilitating the derivation of label transition probability distributions. Finally, belief propagation is harnessed for inference within a pairwise Conditional Random Field (CRF), resulting in an assessable end-to-end trainable pipeline.

2.2.4 Prompt-learning Models

Prompt learning is a technique in natural language processing where models are guided by specific prompts to produce or interpret targeted responses. With the rise of large-scale models, prompt learning can leverage contextual representation and implicit knowledge to enhance performance on targeted tasks. Since most LLMs [87, 120] and MLLMs [72] are

trained on plain text or natural images, layout-aware prompting and in-context learning [9] methods have been introduced to improve understanding in VRDs.

QueryForm [130] introduces a query-based framework for zero-shot document key information extraction. A dual prompting-based mechanism, entity query (E-prompt) and schema query (S-prompt), is introduced for pretraining and transferring knowledge from large-scale weakly annotated pretrained webpages to target domains. The pretraining target is highly aligned with fine-tuning to ensure the proposed framework can consistently make query-conditional predictions at both stages. The HTML tags acquire the E-prompt ($\boldsymbol{e_p}$) during pretraining, and the S-prompt ($\tilde{\boldsymbol{s}}_{\boldsymbol{p}}$) is generated from webpage domains, while during the fine-tuning, $\boldsymbol{e_p}$ is predefined and S-prompt $\boldsymbol{s_p}$ is learnable vectors. Supposing $\boldsymbol{t}$ is the serialised text of the input document, the pretraining and fine-tuning target $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ can be represented as:

$$\hat{\boldsymbol{y}} = \mathcal{F}([\boldsymbol{s_p}; \mathcal{E}[\boldsymbol{e_p}, \boldsymbol{x}]]), \tag{2.2}$$

$$\boldsymbol{y} = \mathcal{F}([\mathcal{E}[\tilde{\boldsymbol{s}}_{\boldsymbol{p}}; \boldsymbol{e_p}, \boldsymbol{x}]]) \tag{2.3}$$

where $\mathcal{E}$ and $\mathcal{F}$ in here are the feature encoder and the rest of the language model, respectively. The training objective is to minimise the cross-entropy loss between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$.

GenKIE [12] proposes an encoder-decoder-based multimodal KIE framework to leverage prompt to adapt various datasets and better leverage multimodal information. Following [139, 140], different encoding methods are adopted to acquire textual, layout and visual embeddings. Byte Pair Encoding is used as language pretrained backbones, and the OCR extracted document content is concatenated with predefined prompts split by the "[SEP]" token between OCR tokens and each prompt. The 2D-positional encoding introduced by [140] is used to acquire the layout embedding of each OCR extracted token, and ResNet [40] is used to extract the visual representations following [129]. The multimodal representations are fed into an encoder to learn interactively between modalities. Prompts, inserted at the end of the encoder's textual inputs, are either template-style or question-style. For the entity extraction task, the prompt specifies the target entity type, and the decoder generates the entity value (e.g., for "Company is?", the decoder outputs the company name). For the entity-labeling

task, the prompt includes the value, and the decoder provides the entity type (e.g., for "Es Kopi Rupa is [SEP]", the decoder identifies the entity type).

ICL-D3IE [38] is the first framework to employ LLMs with in-context learning to extract key information from VRDs using the iterative updated diverse demonstrates. Before designing the initial diverse demonstrations, the most similar n training documents to the n test samples need to be selected by calculating the cosine-similarity of document representations encoded by Sentence-BERT [102]. Then, different types of demonstrations are introduced to integrate multiple-view context information into LLMs. Hard Demonstrations highlight the most challenging cases, are initially designed based on the incorrectly predicted cases from GPT-3 [9] predictions and are updated based on prediction results during the training process. Layout-aware demonstrations are created by selecting the adjacent hard segments to understand the positional relations. Formatting demonstrations are designed to guide LLMs in formatting the outputs for easy post-processing.

LMDX [92] designs a pipeline to use arbitrary LLMs to extract singular, repeated and hierarchical entities from VRDs. The document images are first fed into off-the-shelf OCR tools and are divided into smaller document chunks to be processed by the LLMs with accessible input length. Then, prompts are generated under XML-like tags to control the LLM's responses and mitigate hallucination. Document Representation is a prompt contains the chunk content with the coordinates of OCR lines to bring layout modality to LLMs, where a text segment extracted by OCR tools is represented <*text*> [x_{centre}, y_{centre}]. After that, the task description and scheme representation prompts are designed to explain the task to accomplish and determine the output format. During inference, N prompts with K LLMs completions are generated to sample the correct answers.

2.2.5 Summary

Several models like Chargrid [53] and ACP [90] have enhanced VRDU by integrating visual and textual information. Additionally, auxiliary tasks such as OCR, utilised by [128, 150], aid in improving multimodal feature representations through joint training. However, these

frameworks often rely on smaller, randomly initialised models, which produce less representative features compared to those generated by large-scale pretrained models like LayoutLM [140] and SelfDoc [68]. Documents typically exhibit specific layouts and logical structures [80], which has prompted many models [59, 60, 146] to adopt graph-based approaches. These methods capture spatial and logical correlations among document elements, such as key-value pairs, leading to a more comprehensive document representation. While these frameworks have achieved improvements in document representation, their effectiveness hinges on having sufficient well-annotated training samples, which are time-consuming to acquire. This limitation has escalated the demand for few-shot [132] and zero-shot [130] frameworks, which leverage contrastive learning and innovative attention mechanisms. Additionally, prompt learning has been applied to distill implicit knowledge from large-scale layout-aware pretrained models [42, 121] and large language models (LLMs/MLLMs) [38, 92]. Despite these advances, a performance gap remains between well-fine-tuned models and few/zero-shot frameworks, highlighting the ongoing challenges in VRDU optimization.

2.3 Visually Rich Document Question Answering

Unlike key information extraction, which targets specific details within document images, answering natural language questions involves interpreting more complex intentions and requires models to facilitate interactive understanding between the queries and document representations. The introduction of DocVQA [84] marked a significant shift in focus from natural scene images to text-dense, layout-aware single-page document images, establishing a benchmark in the field. As advancements have continued, demands have recently emerged for models capable of addressing more complex, multi-page scenarios [27, 28, 119]. These emerging requirements highlight the need for models to process multimodal inputs and navigate through extensive documents, reflecting user inquiries' evolving complexity and naturalness in document-based question-answering systems. This section will briefly review the SoTAs in single-page document VQA models and introduce some recently proposed multi-page document understanding solutions [8, 119].

2.3.1 Single-page VRD QA

Similar to key information extraction from visually rich documents, single-page question answering (QA) often utilises classical pretrained language models [24, 75] as baselines. These models conduct span-based question answering to extract sequences of text tokens. Additionally, general domain visual language pretrained models [55, 66, 114] are employed to identify the target document's semantic entities [27]. Beyond these plain text or general domain vision-language pretrained models, numerous layout-aware models specifically pretrained in the document domain (see Section 2.4) have achieved state-of-the-art (SoTA) performance on single-page document tasks. This approach underscores the importance of integrating layout awareness in the preprocessing stages to enhance performance on these specific document-based tasks.

2.3.2 Multi-page VRD QA

As demand grows for retrieving answers from multi-page documents [28, 119], most SoTA models [42, 127, 140], which are typically designed for single-page inputs, encounter a significant challenge due to their input length limitation of 512 tokens. To overcome this limitation, recent innovations have introduced solutions such as transformers capable of handling longer sequences and page-locating modules. These advancements are specifically designed to address the requirements of multi-page document question answering (QA), enabling more efficient processing.

Hi-VT5 [119] proposes a multimodal hierarchical encoder-decoder architecture for multi-page generative question-answering. A T5-based [97] multimodal transformer is applied to encode the single-page level information, including questions, OCR-extracted page content, image patches and a set of page tokens. Question and OCR extracted sequence of tokens is following [5] to acquire the layout-aware initial textual representations. Document Image Trasnfoer (DIT) [65] is leveraged to acquire initial patch representations. Then, the concatenated question (Q) and page OCR tokens (T), image patch tokens (I), and randomly initialised page tokens (P) are fed into T5-based page encoder to enhance the multi-level and multimodal

contextual learning contextually. The enhanced page token embeddings P' of each input document image are fed into a T5-based decoder to generate the predicted answer, as well auto-regressively, and the target answer located page needs to output the page number based on P'. As T5 is not a layout-aware language model, masked language modelling is applied to predicted masked tokens by leveraging non-based visual and layout information to boost multimodal understanding. As T5 can accept input sequence lengths up to 20,480 tokens, it dramatically increases the standard VRDU pretrained models that apply Hi-VT5 to multi-page scenarios.

GRAM [8]: proposes a framework to extend single-page models to tackle multi-page document VQA scenarios. Following the way of pretrained DocFormerv2 [2], the framework encodes single-page inputs, including questions, OCR-extracted content, and visual features. For multi-page scenarios, a slim global encoder follows each single-page encoder layer, allowing the new learnable page tokens to interact with other pages contextually. This enables page tokens to capture page-level information through self-attention and enhance document-level understanding with sparse attention. As the global layer is a new stream applied to the pretrained model, an attention bias is applied following ALiBi [95] to prevent the model from possibly disregarding the page tokens, enabling them to capture more fine-grained information from pretrained weights. During the decoding stage, unlike Hi-VT5 [119], which only feeds page tokens to the decoder, GRAM uses all fine-trained information for the decoder. C-Former [98] is applied to alleviate the high computation consumption, which could revise the information across all pages and distil only important details.

Kang et al. [52] introduce a multi-page document VQA framework to use a scoring self-attention mechanism to select and identify the related pages for generating the answer to the input question. The training processes include single-page Document VQA training and then training a self-attention scoring module based on the frozen trained encoder to feed the most relevant page information into the decoder for answer generating. Pix2Struct [61] is used as the single-page model fine-tuned on the DocVQA [84] dataset. The output from the Pix2Struct is fed into the self-attention scoring module to extract the first token for predicting

a question-page matching score. The page with the highest matching score will be fed into the decoder for answer generation.

2.3.3 Summary

Document Visual Question Answering (Document VQA) is a relatively new field, pioneered by DocVQA [84], which aims to generate or extract answers to natural language questions based on document images. This differs from key information extraction, which focuses on recognizing or extracting predefined key-value pairs. Document VQA requires a more comprehensive representation of the document and an understanding of correlations between the document and questions. Pretrained VRDU models [45, 139] demonstrate robust performance in single-page document understanding tasks. However, these models encounter challenges when applied to more typical and natural multi-page scenarios due to input length limitations. Recent solutions, such as [8, 119], mainly address these challenges by identifying the page where a possible answer may be located and then applying SoTA single-page techniques to retrieve the answer. Despite these advancements, real-world applications often present more complex situations, such as long-term dependencies and cross-page entity relationships, which still need further exploration in the document VQA field.

2.4 Multi-Task Visually Rich Document Understanding Frameworks

Models designed for specific VRDU tasks often incorporate task-oriented techniques. Drawing inspiration from pretrained models in the vision [30, 103] and language [24, 75] domains, enhancing document representations may significantly improve the performance of various downstream tasks. Consequently, a range of models have been developed to extract knowledge from extensive document collections. These models employ different pretrained tasks based on their architecture, the modalities they process, and the granularity of the information they extract from documents. Moreover, some models introduce specialised techniques to boost the effectiveness of pretrained model representations, achieving better performance without

heavy pretraining. This section explores various document understanding frameworks focused on enhancing document representation for robustness and comprehensiveness in addressing multiple VRDU downstream tasks.

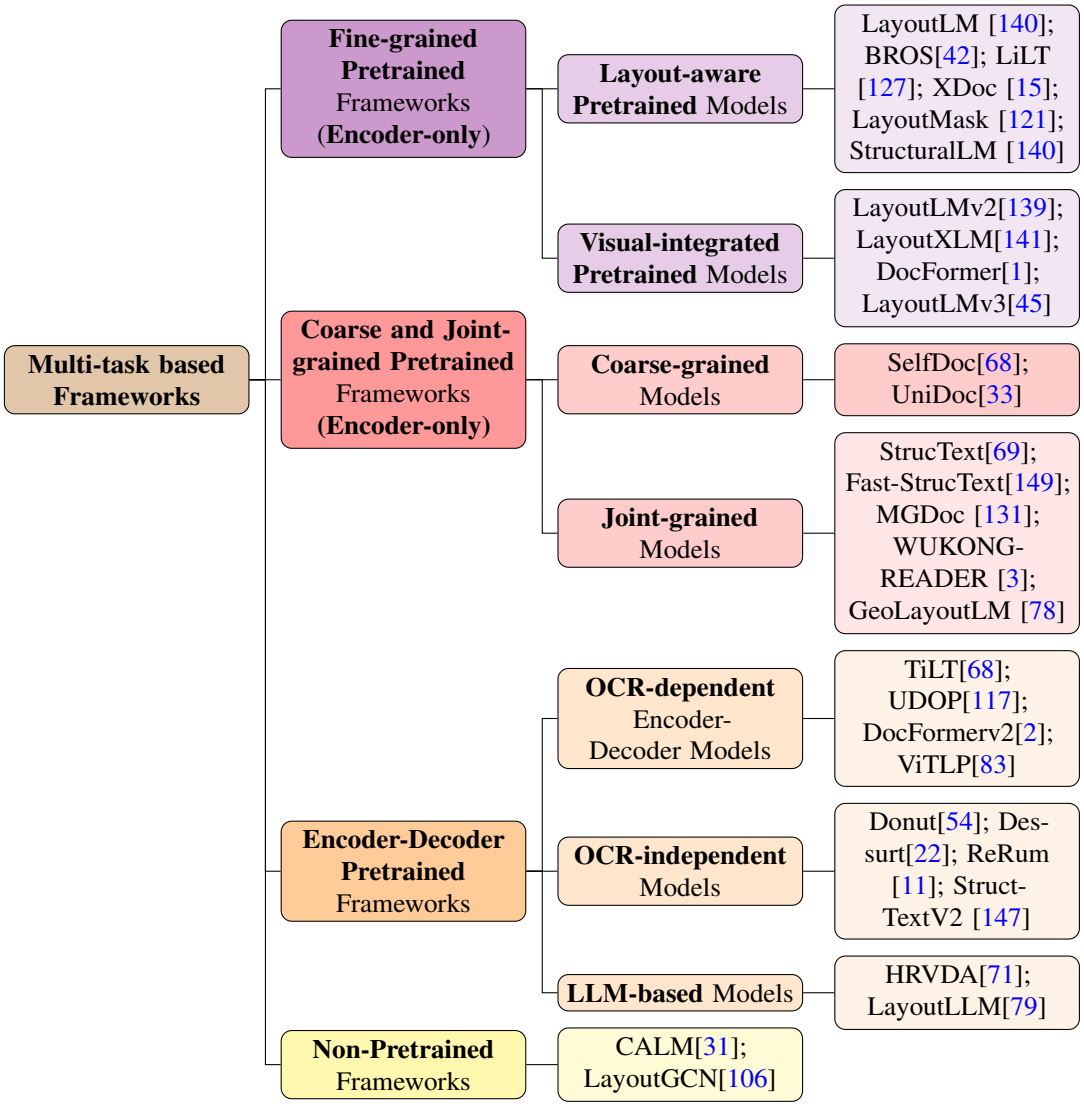


FIGURE 2.2. Overview of Multi-task document understanding frameworks.

2.4.1 Fine-grained Pretrained Frameworks

Inspired by BERT-style pretrained models, many researchers have proposed effective methods to integrate layout and visual information into models, aiming to enhance the comprehensiveness of textual token representations.

(1) Layout-aware Pretrained Models

Understanding layout structure and spatial correlations between textual tokens can yield a more comprehensive representation of documents, going beyond what plain-text input offers. To this end, various methods have been proposed to encode layout features. These methods, combined with tailored pretraining tasks, enable models to capture layout-aware information better and effectively fuse textual and layout features.

LayoutLM [140] is the first pretrained document understanding model by leveraging textual and layout information in the pretraining stage. BERT architecture is the backbone and 2-D positional embedding with textual information is used to pretrain on IIT-CDIP Test Collection 1.0 [62]. Two specific pretraining tasks are first introduced, named *Masked Visual-Language Model* (MVLM) and *Multi-label Document Classification*, to generate layout-aware textual representation and more comprehensive document representation, respectively. Like Masked Language Modeling adopted by most pretrained language models, MVLM randomly masks some input tokens but keeps the corresponding 2-D position embeddings to predict masked tokens to ensure the pretrained model is aware of the spatial relations between input tokens. MDC is a supervised pretraining task to predict the input document types (e.g. forms, exam papers, academic papers) that generate more comprehensive document-level representations. The fine-tuned LayoutLM could perform much better than textual-only frameworks on key information extraction [47, 49] and document classification [37].

BROS [42] aims to propose a pretrained VRDU model to represent the continuous property of 2D space by introducing a new 2-D positional encoding and a textual-spatial correlation aware attention score to replace the vanilla self-attention. Supposing $[x1, y1, x2, y2, x3, y3, x4, y4]$ is the bounding box coordinates of input token t, $p^1 = [\mathcal{F}_{sin}(x1) \oplus \mathcal{F}_{sin}(y1)], p^1 \rightarrow \mathbb{R}^{d_{pos}}$. The final positional representation of t is $pos_t = W_{p1}p_1 + W_{p2}p_2 + W_{p3}p_3 + W_{p4}p_4$, where all $W_p \in \mathbb{R}^{2d_{pos} \times d}$. The new 2-D positional encoding intends to provide a more natural way to encode the continuous 2-D coordinates. Additionally, instead of simply summing textual and positional representations, a novel attention score is introduced to consider both textual and spatial features and their correlations. To calculate the attention score α_{ij} between t_i and t_j,

they consider correlation intra and inter-correlation between textual and positional modalities.

$$\alpha_{ij} = (W^{q_{t_i}} T_i)^\top (W^{q_{t_j}} T_j) + (W^{q_{t_i}} T_i \circ W^{pos_{t_i}} pos_{t_i})^\top (W^{pos_{t_j}} pos_{t_j}) + (W'^{pos_{t_i}} pos_{t_i})^\top (W'^{pos_{t_j}} pos_{t_j}) \tag{2.4}$$

the terms $(W^{q_{t_i}} T_i)^\top (W^{q_{t_j}} T_j)$ and $(W'^{pos_{t_i}} pos_{t_i})^\top (W'^{pos_{t_j}} pos_{t_j})$ calculate the attention scores between tokens based on textual and positional data, respectively. The term $(W^{q_{t_i}} T_i \circ W^{pos_{t_i}} pos_{t_i})^\top (W^{pos_{t_j}} pos_{t_j})$ captures spatial dependencies based on the combined text and position of the tokens. Inspired by SpanBERT [51], the model also employs an area-masked language model that masks tokens within random-sized rectangular regions to improve text comprehension in complex layouts.

StructuralLM [64] is the first VRUD model using image patches (named "cells" in the paper) to group input tokens for conducting various pretraining tasks. They use BERT as the backbone and generate multimodal representations P of each patch, which will be used for two pretraining tasks, MVLM and Cell (Patch) Position Classification (CPC). The bounding box $(x0, y0, x1, y1)$ of an image patch is encoded by the 2-D positional encoding introduced by ***LayoutLM*** [140] and the n tokens inside p are represented as $\{t_1, t_2, \ldots, t_n\}$ which share the same 2-D positional encoding as for tokens belongs to one patch. A token t_i could be represented as summing up token representation T_i, 2-D positional encoding $pos^{2d}_{t_i}$ and 1-D position embedding pos_i. Two pretraining tasks are used to understand the patch-level spatial correlations, including MVLM and CPC. MVLM is similar to LayoutLM but uses patch-level layout embeddings instead of token-level. Another novel cell position classification task is applied to predict the area index of randomly selected tokens from N evenly split areas. Two pretraining tasks are performed simultaneously to capture the patch-level spatial dependencies between input tokens.

LiLT [127] introduces a language-independent layout Transformer for mono/multi-lingual document understanding. The text and layout information are first separately encoded and jointly learned during pretraining. In the fine-tuning stage, two modality representations are concatenated to perform downstream tasks. The input token representations follow BERT, where j-th token is $T_j = T_j + pos_{t_j} + pos^{2d}_{t_j}$. The layout representations are slightly different

from LayoutLM, normalising the bbox coordinates in the $[0, 1000]$ and four linear layers encode the x-axis, y-axis, width and height features. The normalised bbox of a token t is $[x0, y0, x1, y1, w, h]$ is encoded to get the final layout representation L:

$$L = W_L(W_x x0 \oplus W_y x0 \oplus W_x x1 \oplus W_y y1 \oplus W_w w \oplus W_h h) + pos_L \tag{2.5}$$

where $W_L \in \mathbb{R}^{6d'_L \times d_L}$ and $W_x, W_y, W_w, W_h \in \mathbb{R}^{1 \times d_L}$. The text and layout embedding are fed into two sub-models to generate high-level representations. A bi-directional attention complementation mechanism (BiACM) is proposed to augment the cross-modality interaction. Three self-supervised learning methods are proposed to enable the model to understand the multimodal document representation, including MVLM and Key Point Location, which is similar to the CPC proposed by StructureLM to predict the area index of masked tokens based on the layout features. Additionally, a text and layout alignment task is proposed to enhance further the cross-modality representation to predict whether each pair is aligned. LiLT could achieve promising performance on multi-lingual document understanding benchmarks [128, 141].

XDoc [15] propose a unified framework to deal with text inputs from multi-format inputs, including plain text, document and web text. Various encoding methods are proposed to tackle diverse text formats. For plain text token t_{pln}, the token representation follows BERT to get T_{pln}. For document text token t_{doc}, they adopted similar strategies [42, 140] to get T_{doc} by summing up initial token representation $E_{t_{doc}}$, 1-D position embedding $pos_{t_{doc}}$ and 2-D position embedding $pos^{2d}_{t_{doc}}$. $pos^{2d}_{t_{doc}}$ uses different bbox formats $[x_0, x_1, y_0, y_1, w, h]$ following a *Linear-ReLu-Linear* adaptor to make the document format text embedding more representative. To encode web text inputs, they follow the way introduced by MarkupLM to encode the source file tags and subscript information to get the XPath embedding. The same structured adaptor is leveraged for better pretraining. All three text formats are pretrained by MLM and fine-tuned on plain text [99, 125], document [49, 84], and web text [16] benchmark datasets, respectively.

LayoutMask [121] aims to improve the text-layout interactions by using local 1D positional encoding and new pretraining tasks. OCR-dependent document understanding frameworks

are suffered from improper reading orders. It uses LayoutLMv2 as the backbone but removes the visual module, and the local 1D positional encoding is adopted to replace the global 1D positional encoding, which only encodes the token ordered within each segment detected by OCR tools and always restarts with one from each segment. Moreover, the 2D positional encoding also uses segment bbox instead of individual words. The first task pertains to traditional masked language modelling, but two different masking strategies are used. Firstly, they set the mask at the word level instead of masking the token itself to enable the model to capture more contextual information to predict the masked word. Additionally, to boost cross-segment understanding, the probability of masking the first and last word of a segment is higher than that of others. Another Masked Position Modelling is proposed to ask the model to predict the masked word-level 2D positions to promote the layout information representation.

(2) Visual Integrated Pretrained Models

Integrating visual cues alongside text and layout information during the pretraining stage can significantly enhance model capabilities, capturing more comprehensive document insights than text and layout alone. Various frameworks and pretraining tasks focusing solely on text and layout have been expanded to include visual-text matching tasks to strengthen cross-modality alignment. This integration enables models to interpret better the complex interplay of visual, textual, and layout features within documents. Existing models that utilise these comprehensive inputs can be categorised based on their approach to visual feature acquisition: *feature map*-based models, which generate comprehensive visual representations, and *patch pixel*-based models, which focus on granular visual details. This classification helps understand how different models leverage visual information to enhance document understanding.

LayoutLMv2 [139] is the first pretrained model integrating text, layout and visual aspects during the pretraining stage. A single framework multimodal transformer is applied to get the visual, text and layout features simultaneously, where each input textual and visual tokens are assigned a segment ID to distinguish the modality or semantic type. The textual and layout representations generally follow LayoutLM, except the linear projected segment ID, seg_t,

embeddings are addicted to original textual embeddings. The input document image is firstly fed into a trainable visual encoder (based on ResNeXt-FPN), of which output is evenly split into m flattened visual tokens, v_{token} following a linear layer to project into the same dimension as textual embedding. The j-th visual token embedding is $V_j = W_{v_j} + pos_j + W_{seg_v}$. Moreover, to capture the relative position information of inter/intra-modality features, spatial-aware self-attention is introduced by adding bias terms of 1-D (b^{1D}) and 2-D (b^{2D}) relative position. The 1-D relative positional bias between input visual or textual tokens t_i and t_j is $b^{1D} = W_{b_{1D}}(j-i)$ and $b^{2D} = b_x^{2D} + b_y^{2D}$, where $b_x^{2D} = W_{b_x^{2D}}(x0_i - x0_j)$, $b_y^{2D} = W_{b_y^{2D}}(y0_i - y0_j)$. ***LayoutLMv2*** is pretrained on MVLM and the other two new proposed tasks, Text-Image Alignment (TIA) and Text-Image Matching (TIM) to capture more cross-modality alignment.

LayoutXLM [141] extends the ***LayoutLMv2*** architecture to a multilingual setup where the MVLM is extended to Multilingual Mased Visual-Language Modeling pretrained on 22 million self-collected digital-born PDF files and 8 million scanned English documents from IIT-CDIP processed by off-the-shelf PDF parsers.

DocFormer [1] is a multimodal transformer encoder-based architecture which also uses a trainable CNN-backbone (ResNet50) to extract the visual cues of input document image to get visual representations $\mathbb{V} \in \mathbb{R}^{(d \times N)}$, where $d = 768$ is the transformer hidden state and $N = 512$ is the number of visual tokens. The initial textual representations $\mathbb{T}$ are acquired by feeding the OCR extracted text with bbox into LayoutLM where $\mathbb{T} \in \mathbb{R}^{(d \times N)}$. 2D positional encoding is also used by using W_v^x, W_v^y and W_t^x, W_t^y to encode x, y axis, as well as, visual and textual aspects to acquire pos_v^{2d} and pos_t^{2d}. More positional features are adopted such as width, height and relative distance between neighbours. Moreover, certain inductive biases are applied to get a new self-attention score α for paying more attention to local features. Supposing α_{ij}^v is attention score between visual feature V_i and V_j, it could have $\alpha_{ij}^v = (W_v^K V_j)^\top (W_v^Q V_i) + (pos_{ij})^\top W_v^Q V_i + (pos_{ij})^\top W_v^K V_j + (W_s^Q pos_{ij}^{2d})^\top (W_s^K pos_{ij}^{2d})$, where pos_{ij} is the 1D relative positional encoding. The same procedures are applied to textual features as well, and two modalities share the same W_s^Q and W_s^K to help the model correlate features across modalities. The multimodal token representations $\mathbb{M}$ of l encoder layer is $\mathbb{M}_l = \mathbb{T}_l + \mathbb{V}_l$.

LayoutLMv3 [45] is the first pretrained VRDU model to encode visual features without using heavy CNN-backbones. ***LayoutLMv3*** uses the exact same way as ***LayoutLMv2*** to encode the textual and layout information except replacing the word-level bbox to segment-level for 2-D positional embedding. For visual feature representation, they follow ViT and ViLT to linearly project the evenly split image patches with learnable 1-D positional encoding to the transformer encoder. For the pretraining setting, they use masked language modelling to mask 30% tokens drawn from a Poisson distribution. To augment the visual representation by contextually learning with multimodal features, a Masked Image Modelling (MIM) [4] task is used to mask 40% image tokens with clockwise masking randomly. Each image token will be converted into discrete tokens following [100], and a cross-entropy loss is adopted to reconstruct the masked image tokens. Moreover, to explicitly learn the correlation between text and image modalities, a Word-Patch Alignment (WPA) pretraining objective is applied. WPA is a binary classification task to predict whether the unmasked token-patch pair is "*aligned*" or "*unaligned*".

2.4.2 Coarse and Joint-grained Pretrained Frameworks

Fine-grained models achieve state-of-the-art performance on many downstream tasks but face challenges with input length limitations and capturing document image layout and logical arrangement. To address these issues, coarse-grained or joint-grained frameworks have been introduced. To mitigate these limitations, these frameworks leverage multimodal information from document semantic entities such as paragraphs, tables, and textlines.

(1) Coarse-grained Models

SelfDoc [68] is the first pretrained VRDU model to leverage coarse-grained document elements for various document understanding tasks. Unlike fine-grained models that rely on OCR-extracted text sequences, ***SelfDoc*** uses Faster-RCNN to extract Regions of Interest (RoIs) from document entities. This approach reduces the input sequence length for text-dense and long documents with improved time and space complexities. To acquire the initial multimodal representations, Faster-RCNN extracts visual embeddings from these RoIs, while

Sentence-BERT [102] provides textual embeddings based on the OCR-extracted text of each entity. The textual and visual representations are fed into two single-modality BERT-like encoders to learn the intra-modality correlations between entities contextually. A cross-modality encoder is introduced to enhance inter-modality learning, mirroring the structure of single-modality encoders but incorporating cross-attention layers. It randomly masks entities in document images within either language or vision branches during pretraining. Furthermore, a modality-adaptive attention mechanism dynamically adjusts weights for visual and textual embeddings based on the input and task, creating robust entity representations for fine-tuning.

UniDoc [33] is an entity-level document understanding model which contains a trainable image encoder with RoI-Align [39] to extract entity visual representation and novel cross-attention mechanisms to fuse multimodal information. The initial sentence and textual embeddings are acquired by RoI features and averaged word embeddings summing up with the linear projected bbox coordinates. To enable the trainable visual representation to predict meaningful visual cues effectively, product quantization [50] is applied to discretise a RoI feature into a finite set of visual representations and mapped to a new embedding. A novel Gated Cross-Attention is also introduced to improve the interactive learning between various modalities. After a multi-head cross-attention to get the enhanced visual and textual representations, a gating mechanism is applied to dynamically weight the visual (V) and textual features (T) by feeding concatenated $[V : T]$ to a non-linear network to acquire a modality-aware attention bias β_v, β_t for visual and textual respectively. Three pretraining tasks are conducted simultaneously to enhance multimodal feature representatives. Masked Sentence Modelling and Visual Contrastive Learning are required to predict the masked sentence representation and quantize visual representation, respectively. Another Vision-Language Alignment introduced by LayoutLMv2 [139] to enforce contextual learning between multi-modalities. But instead of using split image regions, UniDoc aligns the image and text to belonging to the same entity.

(2) Joint-grained Models

StrucText [69] is a multimodal pretrained VRDU model using fine-grained textual and coarse-grained vision information to capture richer geometric and semantic information from different levels and modalities. The layout embedding of each text token and segment is encoded by $L = W_l[x0, y0, x1, y1, w, h], W_1 \in \mathbb{R}^{6\times N}$. They follow token-level models to encode the sequence of tokens and use pretrained ResNet50 with FPN to extract the entity visual features. A segment-ID embedding is allocated to each token's textual and entity visual representations to boost the alignment learning. Three self-supervised tasks are introduced to enhance inter-modality learning: MVLM, Segment length Prediction (SLP), and Paired Box Direction (PBD). As a new self-supervised learning task, SLP asks the model to predict the entity's length to leverage the entity's visual embeddings and textual information from the same segment ID. Another self-supervised learning task aims to learn more comprehensive pair-wise spatial correlations between entities by clarifying their spatial correlation.

Fast-StructText [149] is built on StrucText [69] to improve model efficiency and enhance the feature expressiveness by introducing an hourglass transformer and Symmetry Cross-Attention mechanisms (SCA). Similar feature encoding methods to StructText are adopted, but the bbox formats of layout encoding are $[x0, y0, x1, y1]$. To encode the module and progressively reduce the redundancy tokens, an hourglass transformer is proposed consisting of several Merging blocks and Extension blocks to down-sampling and up-sampling the number of input tokens. Merging blocks suggest merging neighbouring k tokens with the weighted 1D average pooling to shorten the sequence length. The extension block is required to transform the shortened sequence back to the entire length by simply applying a repeat up-sampling method. In addition, to enhance the interaction between textual and vision modalities, SCA consists of two dual cross-attention to handle text and visual features. Different self-supervised tasks are conducted simultaneously, including MVLM [69, 139], Graph-based Token Relation (GTR), Sentence Order Prediction (SOP) and Text-Image Alignment [45]. For the two new proposed tasks, GTR is a task similar to Paired-Box Direction to predict the positional relations between paired entity visual features and SOP is used to predict whether two sentence pairs are normal-order adjacent to learning semantic knowledge.

MGDoc [131] is the first multimodal multi-granular pretrained framework that introduces multi-granular and cross-modal attention to increase the inter-grained learning and better cross-modality fusion. For acquiring initial features of different modalities, pretrained language models [102] and vision encoders [40] are used to encode different-grained inputs from word to whole page. The encoded modalities are associated with positional features [149] and modality-type embeddings [69] to acquire the final input representation of various modalities. To encode the hierarchical relation between multi-grained features, including pages, entities and words, two attention biases are added to the original self-attention weights. A hierarchical bias is a binary value of 0 or 1 to examine the inside or outside relation between two inputs, while the relation bias is the relative positive between two bboxes. Apart from multi-grained learning, cross-attention is applied to fuse cross-modality information better. Three tasks are proposed during pretraining: Mask Text Modeling (MTM), Mask Vision Modeling (MVM) and Multi-Granularity Modeling. MTM and MVM randomly mask the multi-grained textual and visual representations to predict the masked features under the mean absolute error. To boost the interactive understanding of multi-granularity, a token-entity linking prediction is proposed by computing the dot-product between token and entity features.

WUKONG-READER [3] uses fine-grained level inputs but leverages coarse-grained level self-supervised tasks to enhance the fine-grained level information. The model inputs contain a document image and OCR extracted sequence of tokens with their bboxes. A Mask-RCNN is used as a visual backbone to acquire several visual tokens and extract visual features of textlines from a RoIHead layer. For textual information encoding, the first is layers of RoBERTa[136] are applied to extract the token representation and sum up with additional features following [139]. The concatenated visual and textual features are fed into a multimodal modal encoder based on the rest six layers of RoBERTa. Different pretraining tasks are proposed, including MLM, Textline-Region Contrastive Learning (TRC), Masked Region Modeling (MRM) and Textline Grid Matching (TCM). TRC aims to enhance cross-modality interactive learning at the entity (textline) level by following a text-vision-aligned contrastive learning approach [142]. To enhance the visual representation, MRM is applied to mask 15% of textlines randomly and to predict the masked visual embeddings following [68]. A text-grid alignment task is introduced to enhance layout understanding by dividing the

document image into N-grids and predicting the tokens in 15% selected textlines belonging to which grid. The scaling parameters are applied to each loss of pretraining tasks.

GeoLayoutLM [78] is a sophisticated multimodal framework that distinctively incorporates geometric information through specialised pretraining tasks and the development of innovative relation heads. Inspired by the dual-stream structure of METER and ***SelfDoc*** [68], GeoLayoutLM features separate vision and text-layout modules coupled with interactive co-attention layers that enhance the integration of visual and textual data. The model introduces two advanced relation heads—the Coarse Relation Prediction (CRP) head and the Relation Feature Enhancement (RFE) head—which refine relation feature representation crucial for both pretraining and fine-tuning phases. The pretraining regimen includes tasks designed to understand geometric relationships, such as GeoPair, GeoMPair, and GeoTriplet, aiding the model in grasping the complex dynamics of document layouts. During fine-tuning, the model utilises pretrained parameters to optimise both semantic entity recognition and relation extraction tasks, employing a novel inference technique that enhances relation pair selection accuracy by focusing on the most probable relationships and minimizing variance among potential options.

2.4.3 Encoder-Decoder Pretrained Frameworks

In addition to the above encoder-only frameworks, researchers have proposed encoder-decoder pretrained models that often approach tasks like Key Information Extraction (KIE) or Visual Question Answering (VQA) in a generative style. Addressing limitations of OCR-dependent frameworks, such as accumulated OCR errors and incorrect reading orders, OCR-independent models have been introduced for end-to-end VRDU.

(1) OCR-dependent Encoder-Decoder Models

TILT [94] is a T5-based transformer encoder-decoder architecture enhanced with a relative spatial bias in the self-attention mechanism to acquire fine-grained token representations. It incorporates encoding methods that apply relative sequence input bias and capture horizontal and vertical distance biases in the attention scores. A U-Net-based framework is applied to

extract the fixed-size feature maps fed into the encoder together. They follow T5 pretraining strategies on RVL-CIDP dataset [37] but use a salient span masking scheme adopted by [36, 98]. As the first encoder-decoder framework, ***TILT*** requires off-the-shelf OCR tools to acquire the textual token sequence. To conduct VRD content understanding end-to-end, some OCR-free frameworks are proposed to solve the limitations of OCR-dependent models.

UDOP [117] proposes an encoder-decoder pretrained document understanding framework that leverages a ViT-based model [30], following the principles of LayoutLMv3 [45], to process multimodal information. To enhance the comprehensiveness of textual token embeddings, the framework sums the text token embeddings with token-aligned image patch embeddings when the centre of the bounding box falls within the image patch. Additionally, the positional biases introduced by TILT are added, but ID positional encoding is not used. The decoding stage is designed to generate all vision, text and layout modalities consisting of a bi-directional Transformer text-layout decoder and an MAE vision decoder. Two decoders are cross-attended with each other. The pertaining tasks contain self-supervised on unlabeled documents to learn robust document representation and supervised pertaining tasks for fine-grained model supervision. Masked text layout and image reconstruction are used to predict the masked information using unmasked multimodal information. A layout moulding task is applied to predict the positions of groups of text tokens, and a visual text recognition task is proposed to identify text at the given location in the image. Supervised tasks utilise a training set of publicly available benchmark datasets of different downstream tasks, including document classification [37], KIE [112], VQA [84, 116] and layout analysing [157].

DocFormerv2 [2] is an encoder-decoder transformer architecture that uses multimodal (visual, textual and positional) to enhance the multimodal understanding and layout-aware language decoder to predict the predictions. The patched image pixels are fed into the convolutional and linear layers to get down-sampled patch embedding. The textual embeddings are acquired by linear projected token one-hot encoding. Both visual and textual embeddings are summed with the 2D-positional encoding [140] of Patches and linear project bbox embedding ($[x0, y0, x1, y1]$) [127], respectively. Two encoder-based Token to Line (T2L), Token to Grid (T2G), and one decoder-based self-supervised learning task, MLM, are proposed to

enable multimodal feature interactive learning. T2L aims to improve the relative position understanding between tokens by predicting the number of textlines between two randomly selected tokens. For improving the layout and structure, understanding needs to split the image into $m \times n$ grids to predict the located grid number of each OCR token. For the decoder MLM, the spatial feature of each masked text token is masked as well, and the other setup follows T5 [97].

ViTLP [83] introduces an encoder-decoder architecture for OCR and document understanding using a ViT-based vision encoder to obtain image patch representations, which are fed into a decoder to generate text and layout sequences autoregressively. A special "[LOC]" token, encoding bounding box coordinates $[x0, y0, x1, y1]$, reduces the layout token sequence length. To control generation flow, "[BOS]" and "[CONT]" tokens are added, representing input sequences of two tokens, t1,t2, as "[BOS], t1, [LOC], t2, [LOC]". The decoder has hierarchical heads: the text head uses all tokens to generate the next text token, while the layout head uses "[LOC]" tokens to predict bounding box coordinates. The "[CONT]" token handles sequences of arbitrary length by continuing generation until "[END]", based on the prefix token ratio.

(2) OCR-free Pretrained Models

Donut [54] is the first OCR-free VRD understanding model to understand and extract key information from input document images. Donut contains a Swin Transformer-based visual encoder to encode the input document image into image patches, which are then fed into a BART-based [63] decoder pretrained on multi-lingual scenarios. During model training, teacher forcing is applied, and in the test stage, inspired by GPT-3 [9], prompts with special identify tokens are fed into the model for different downstream tasks. The output token sequence contains the special tokens $< START\_* >$ and $< END\_* >$ to identify the type of tasks and struct predict entities. The wrongly structured entity will be treated as an empty prediction. The model is pretrained on next-token prediction on the IIT-CDIP dataset and a Synthetic Dataset, which can be interpreted as a pseudo-OCR task. Similar to Donut, ***Dessurt*** [22] also proposes an encoder-decoder architecture but a different decoding process. Instead of using BART, the cross-attention used in Dessurt attends to all visual, query and previously

generated textual information and is pretrained on more synthetic datasets with different font sizes and handwritten content.

ReRum [11] introduce an end-to-end architecture in which the decoding process focuses on local interest and visual cues. Like other OCR-free frameworks, a Swin-Transformer extracts image patch representations. The extracted visual features are fed into a transformer-based Query-Decoder to acquire the vision-enhanced query representation. The enhanced query representation with visual features is fed into a Text Decoder to generate the predicted text tokens auto-regressively. Notably, a Content-aware Token Merge technique is proposed to dynamically focus on more relevant foreground parts of visual features, which selects top-K visual tokens based on the averaged correlation scores between query and visual token representations. Additionally, the unselected visual tokens (called Background Area) contain rich global features that could be used to enhance the Top-K foreground visual tokens through basic attention mechanisms. Three pretraining tasks are applied: query to segmentation (Q2S), text to segmentation (T2S), and segmentation to text (S2T). Q2S follows DETR setup for a token generation but alters to an instance segmentation task to predict N mask for the target text area to improve text detection ability. T2S acquire the Text Decoder output to conduct another instance segmentation task after feeding the image text snippet into the Text Decoder to boost the layout-aware textual information understanding. The last S2T is to use the outputs from the Text Decoder to generate the token like an OCR task auto-regressively.

StrucTextV2 [147] is an end-to-end structure that uses image-only input to conduct several downstream tasks. It contains a CNN-based visual extractor with FPN strategies [70] and follows ViT [30] to get linear projected flattened patch-level representations. The patch token embeddings serve as the input to the Transformer encoder to enhance the contextually semantic representations. Then, the lightweight fusion network is applied to generate the final representations and fed into two branches during pretraining: made language Modeling (MLM) and Masked Image Modeling (MIM). Instead of using text inputs when MLM is used by other models [24], a portion of the text regions are masked with RGB values $[255, 255, 255]$ randomly with a 2-layer MLP decoder to predict the masked token. MIM masks the rectangular text regions and predicts the RGB values of the missing pixels to

improve the document representations. Except for the global average pooled FPN fused visual representations, the MLM-generated hidden state of each text region is concatenated and fed into a Fully Convolutional New York to get the regressed masked missing pixel values.

(3) LLM-based Frameworks

HRVDA [71] aims to propose a MLLM accepting high-resolution image inputs to conduct fine-grained information extraction from VRDs. A swin-transformer [76] is used to encode document images into image patch tokens. A pluggable content detector then identifies visual tokens that contain relevant document content information. Following this, a content filtering mechanism performs token pruning to remove irrelevant tokens. The remaining encoded visual tokens are processed through an MLP to ensure consistency with the LLM embedding space dimensions. These pruned tokens are then fused with instruction features, allowing further filtering of tokens irrelevant to the instructions. The final streamlined set of visual tokens and instructions is fed into the LLM, which generates the corresponding responses.

LayoutLLM [79] introduces an LLM/MLLM-based approach integrated with a pretrained document understanding model to address the challenges of applying LLMs to zero-shot document understanding tasks. The input document's visual, textual, and layout information and any question text are encoded by a pretrained LayoutLMv3 [45] encoder and projected into the same embedding space as the adopted LLM, Vicuna-7B-v1.5 [154]. The method incorporates layout-aware pretraining tasks at three levels: document-level (e.g., document summarization), region-level (e.g., layout analysis), and segment-level (e.g., MVLM). These tasks enable the model to achieve comprehensive document understanding. A novel module called LayoutCoT is also designed to help LayoutLLM focus on question-relevant regions and generate accurate answers through intermediate steps. GPT-3.5-turbo [87] is used to prepare the dataset for document summarization training and to construct LayoutCoT training data.

2.4.4 Non-Pretrained Frameworks

CALM [31] introduces a common-sense augment document understanding framework to understand the query and extrapolate answers not contained in the context of the input

document image. They follow LayoutLMv2 [139] to encode input document multimodal representations. The textual token embeddings are fed into a Document Purifier component to merge the tokens $\{t_1, \ldots, t_n\}$ belonging to one entity type N to one Upper Layer token $\hat{c}$ by applying average pooling of $\hat{c} = AvePool(t_1, \ldots, t_n)$. Each Upper Layer token is concatenated with the commence augmented on ConceptNet NumberBatch [110] entity word vector c' to get the final entity representation $c = concat(\hat{c}, c')$. A similar Question-Purifier is applied to use common-sense knowledge to enhance the question representation. Then, with the assistance of ConceptNet, relevant common-sense knowledge is recalled based on the common-sense representation of both documents and queries. By considering the predicted question-answer relationship, a final self-attentive graph convolutional network following [160] is proposed to address document reasoning tasks more effectively.

LayoutGCN [106] proposes a lightweight and effective model which contains a fully connected graph where text blocks are nodes and edges connect every two blocks. The model architecture includes a TextCNN-based [56] encoder to encode N-gram textual embeddings, a linear trainable layout encoder to project the normalised bbox coordinates into hyperspace following other layout-aware models, and a visual encoder (CSP-Darknet [126] for document image features). These features are integrated using a Graph Convolution Network (GCN) to capture relationships between nodes. The final node representation combines text, layout, and visual information, benefiting various VRDU tasks.

2.4.5 Summary

Various models are proposed to enhance document representations for VRDU tasks by leveraging pretrained language models to enrich text token sequences with layout information through positional encoding [140], attention mechanisms [127], and layout-aware tasks [121]. However, VRDs contain rich visual details like font, texture, and colour and visually complex entities such as tables, charts, and photos. Many models [1, 45, 139] integrate visual cues to enhance fine-grained document features, but their quadratic time and space complexity pose challenges for handling long sequences in multi-page document understanding [28]. Fine-grained models excel but struggle with capturing layout and structural details from

document images. Coarse-grained frameworks [33, 68] mitigate fine-grained limitations by leveraging entity-level multimodal information, yet compressing diverse entity aspects into a single dense vector risks losing information [29]. Joint-grained frameworks [3, 69, 78, 131, 149] integrate multi-grained information to produce comprehensive representations. Non-pretrained models leverage external knowledge [31] or lightweight networks [106] to rival large-scale pretrained frameworks' performance. Most document understanding models [45, 68, 127, 140] rely on off-the-shelf OCR tools for text extraction, which can be susceptible to OCR quality issues and incorrect reading orders. OCR-free frameworks directly process document images to mitigate these limitations; However, these frameworks may exhibit sub-optimal performance compared to methods using established OCR tools with additional resource consumption.

2.5 Datasets for Visually Rich Document Understanding

Considering the variations in downstream tasks, we offer a detailed summary of the Key Information Extraction (KIE) and Entity Linking datasets, as well as the Visually Rich Document Question Answering (VRD QA) dataset, in Table 2.1 and 2.2. This section provides an in-depth comparison of the datasets, highlighting their unique characteristics, data structures, and the specific requirements they pose for effective task performance.

2.5.1 VRD Key Information Extraction Datasets

Name	Conf./J.	Year	Domain	# Docs	# Images	# Keys	MP.	Language	Metrics	Format
FUNSD [49]	ICDAR-w	2019	Multi-source	N/A	199	4	N	English	F1	P.& H.
SROIE [47]	ICDAR-c	2019	Scanned Receipts	N/A	973	4	N	English	F1*	P.
CORD [91]	Neurips-w	2019	Scanned Receipts	N/A	1,000	54	N	English	F1	P.
Payment-Invoice [82]	ACL	2020	Invoice Form	N/A	14,237+595	7	N	English	F1	D.
Payment-Receipts [82]	ACL	2020	Scanned Receipts	N/A	478	2	N	English	F1	P.
Kleister-NDA [112]	ICDAR	2021	Private Agreements	540	3,229	4	Y	English	F1	D.
Kleister-Charity [112]	ICDAR	2021	AFR	2,778	61,643	8	Y	English	F1	D.& P.
EPHOIE [128]	AAAI	2021	Exam Paper	N/A	1,494	10	N	Chinese	F1	P.& H.
XFUND [141]	ACL	2022	Synthetic Forms	N/A	1,393	4	N	Multilingual	F1	D.& P.& H.
Form-NLU [26]	SIGIR	2023	Financial Form	N/A	857	12	N	English	F1	D.& P.& H.
VRDU-Regist. Form [134]	KDD	2023	Registration Form	N/A	1,915	6	N	English	F1	D.
VRDU-Ad-buy Form [134]	KDD	2023	Political Invoice Form	N/A	641	9+1(5)	N	English	F1	D.&P.
DocILE [107]	ICDAR	2023	Invoice Form	6,680	106,680	55	Y	English	AP, CLEval	D.&P.

TABLE 2.1. Summary of visually rich document key information extraction datasets.

(1) Scanned Receipt Datasets

SROIE [47] is a widely used dataset for text localization, OCR, and key information extraction from scanned receipts, introduced in the ICDAR 2019 Challenge on "*Scanned Receipts OCR and Key Information Extraction*" The Key Information Extraction (KIE) task focuses on four key types: *Address*, *Date*, *Company*, and *Total*, with corresponding values provided in the annotation file for each receipt. The F1 score for this task is calculated based on Mean Average Precision (MAP) and recall. Note that entity-level annotations are not provided in the official dataset, requiring the use of external tools to obtain the information for coarse-grained or joint-grained models.

Payment-Receipts [82] is a subset of SROIE, created by sampling up to 5 documents from each template in the original SROIE dataset. The template of each receipt is decided by the Company annotation. The target schema focuses on extracting only *Date* and *Total*. This subset is used to evaluate the model's ability to handle unseen templates.

CORD [91] is a widely used dataset for post-OCR receipt understanding, featuring two-level labels annotated by crowdsourcing workers. It includes eight superclasses, such as *Store*, *Payment*, *Menu*, *Subtotal*, and *Total*, each with several subclasses. For example, *Store* contains subclasses like *Name*, *Address*, and *Telephone*. CORD provides both textline-level and word-level annotations for both fine-grained and coarse-grained frameworks, with some sensitive information blurred. All models are evaluated on the released first 1,000 samples.

(2) Form-style Datasets

FUNSD [49] is derived from the RVL-CDIP dataset [37] by manually selecting 199 readable and diverse template form images. The dataset is annotated using the GuiZero library to provide both entity and word-level annotations, including manual text recognition. Semantic links indicate relationships between entities, such as Question-Answer or Header-Question pairs. Consequently, FUNSD supports key information extraction, OCR, and entity linking tasks.

XFUND [141] is the first multilingual dataset following the FUNSD format. It collects form templates in seven languages (Chinese, Japanese, Spanish, French, Italian, German, and Portuguese) from the internet. Human annotators fill these templates with synthetic information by typing or handwriting, ensuring each template is used only once. The filled forms are then scanned into document images, processed with OCR, and annotated with key-value pairs. Each language has 199 annotated forms, supporting multilingual key information extraction and entity linking tasks.

Payment-Invoice [82] contains two corpora of invoices from different sources. The first corpus, Invoice 1, includes 14,273 invoices from various vendors with different template styles used for training and validation. The second corpus, Invoice 2, comprises 595 documents with distinct templates not found in Invoice 1, serving as the test set. Human annotators extract six required keys from each single-page invoice, such as *Invoice Date*, *Total Amount*, and *Tax Amount*. This dataset is suitable for evaluating generative-style models and fine-grained sequence labelling models. For coarse-grained models, additional text line or entity-level information can be extracted using off-the-shelf tools.

VRDU-Registration Form [134] is a dataset of registration forms about foreign agents registering with the US government collected from the Federal Communications Commission. Commercial OCR tools extract the text content of the forms. Annotators draw bounding boxes around six unrepeated entities (each entity appears only once per document) per document: *File Date*, *Foreign Principal Name*, *Registrant Name*, *Registration ID*, *Signer Name*, and *Signer Title*. The dataset provides entity-level annotations, which can be easily preprocessed to acquire token-level annotations, supporting any granularity of Key Information Extraction (KIE) models.

VRDU-Ad-buy Form [134] consists of 641 invoices or receipts signed between TV stations and campaign groups for political advertisements. It follows the same annotation procedure as the VRDU-Registration Form but involves a more complex schema. This includes nine unique entities (e.g., *Advertiser*, *Agency*, *Contract ID*), four repeated entities (e.g., *Item Description*, *Sub Prices*), and hierarchical entities (e.g., *Line Item*). Repeated entities may contain different

values within a single document, while hierarchical entities comprise several repeated entities as components.

Form-NLU [26] is a visual-linguistics dataset designed to support researchers in interpreting specific designer intentions amidst various types of noise from different form carriers, including digital, printed, and handwritten forms. Fine-grained key-value pairs, such as *Company Name*, *Previous Notice Date*, and *Previous Shares*, are manually annotated. The training and validation set comprises 535 digital-born forms, with 76 reserved for validation. Additionally, three test sets are provided, containing 146 digital, 50 printed, and 50 handwritten form images, respectively. Form-NLU can be used to evaluate form layout analysis and Key Information Extraction (KIE) models of any granularity. With proper processing, it can also be used to evaluate entity linking frameworks, thanks to the well-annotated key-value pairs.

(3) Multi-page Datasets

Kleister-NDA [112] is a dataset collected from the Electronic Data Gathering, Analysis, and Retrieval System (EDGAR) focusing on Non-disclosure Agreements (NDAs). During preprocessing, the collected 540 HTML files are converted into digital multi-page PDF files (totalling 3,229 pages) using the Puppeteer library. Four key items, *Effective Date*, *Party*, *Jurisdiction*, and *Term*, are manually annotated by three annotators to extract the corresponding values. The NDA dataset is widely used by fine-grained level models but may require additional processing for frameworks with limited sequence length due to the multi-page inputs.

Kleister-Charity [112] contains 2,778 annual financial reports from the Charity Commission, which lack strict formatting rules. The Charity Commission website provides eight key pieces of information, such as Postcode, Charity Name, and Report Date. Annotators manually correct minor errors to ensure accuracy. Compared to Kleister-NDA, the Charity dataset has longer document inputs, totalling 61,643 pages, requiring models to handle long sequence outputs. Both Charity and NDA datasets provide only key-value pair annotations, making them suitable for the generation and fine-grained sequence labelling tasks but requiring additional processing to acquire entity-level annotations.

DocILE [107] comprises three subsets: an annotated set of 6,680 real business documents, an unlabeled set of 932,000 real business documents for unsupervised pretraining, and a synthetic set of 100,000 documents generated with full task labels. Documents come from public sources like the UCSF Industry Documents Library and Public Inspection Files, with annotations for Key Information Localization and Extraction and Line Item Recognition. Synthetic documents were created using annotated templates and a rule-based synthesiser. DocILE provides entity-level annotations that can be easily post-processed to acquire token-level annotations.

2.5.2 VRD Question Answering Datasets

Name	Conf./J.	Year	Domain	Lang.	# Docs	# Images	# Questions	Answer Type	MP	Format	Metrics	Annotation
DocVQA [84]	WACV	2021	Industrial Reports	English	N/A	12,767	50,000	Text	N	D./P./H.	ANLS	Human
VisualMRC [116]	AAAI	2021	Website	English	N/A	10,197	30,562	Text	N	D.	BLUE, etc	Human
TAT-DQA [158]	MM	2022	Financial Reports	English	2,758	3,067	16,558	Text/RS-Gen.	Y	D.	EM, F1	Human
RDVQA [137]	MM	2022	Data Analysis Report	N/A	8,362	8,514	41,378	Text	N	D.	ANLS, ACC	Human
CS-DVQA [31]	MM	2022	Industry Documents	English	N/A	600	1,000	Text and Nodes	N	D./P./H.	ANLS	Human
PDFVQA-Task A [27]	ECML-PKDD	2023	Academic Paper	English	N/A	12,337	81,085	Num or Yes/No	N	D.	F1	Template
PDFVQA-Task B [27]	ECML-PKDD	2023	Academic Paper	English	N/A	12,337	53,872	Entity	N	D.	F1	Template
PDFVQA-Task C [27]	ECML-PKDD	2023	Academic Paper	English	1,147	12,337	5,653	Entity	Y	D.	EM	Template
MPDocVQA [119]	PR	2023	Industrial Reports	English	6,000	48,000	46,000	Text	Y	D./P./H.	ANLS	Human
DUDE [122]	ICCV	2023	Cross-domain	English	5,019	28,709	41,541	Text, Yes/No	Y	D.	ANLS	Human
MMVQA [28]	IJCAI	2024	Academic Paper	English	3,146	30,239	262,928	Entity	Y	D.	EM, PM, MR	LLM + Human

TABLE 2.2. Summary of visually rich document question answering datasets.

(1) Datasets for Single Page VRD QA

DocVQA [84] is a pioneering dataset in visually rich document question answering, sourced from the UCSF Industry Document Library. It comprises 50,000 manually generated questions framed on 12,767 document images, encompassing digital, printed, and handwritten formats. The dataset follows an extractive-style QA format similar to benchmarks like SQuAD [99] and VQA [7]. Evaluation typically involves fine-grained models such as LayoutLM variants [45, 139, 140], LiLT [127], and generative models [54, 117], using metrics like Average Normalised Levenshtein Similarity (ANLS) [6]. However, it requires additional processing for coarse-grained models like SelfDoc [68]. ***CS-DVQA*** [31] builds upon DocVQA by enhancing QA pairs to better reflect real-world requirements. It extracts 600 images from the DocVQA dataset and generates 1,000 QA pairs under human supervision. During question generation, it incorporates common-sense knowledge from real life, expanding answers beyond extractive in-line text to include question-related nodes (Nodes) sourced from ConceptNet [110].

VisualMRC [116] is compiled from website screenshots across 35 domains, carefully selected to exclude pages with handwritten content and to prefer pages containing short text (no more than 2 to 3 paragraphs). Unlike other datasets that might only provide question-answer annotations [84] or automatically acquire document semantic entities [27], VisualMRC includes manually annotated layout structures with fine-grained semantic entity types such as *Heading*, *Paragraph*, *Subtitle*, *Picture*, and *Caption*. Question-answer pairs are generated through crowdsourcing. Consequently, VisualMRC is well-suited for evaluating both fine-grained and coarse-grained-based QA frameworks, providing a rich resource for assessing the effectiveness of models in understanding and interpreting detailed document layouts and semantic entities.

PDFVQA-Task A and Task B [27] form part of the first VRD QA dataset from PubMed Central, focusing on content and structural understanding. This dataset includes three tasks: two for single-page documents (Tasks A and B) and one for multi-page documents (Task C). Task A evaluates the structural and spatial relationships within document images, with answers being either counts or Yes/No. Task B focuses on extracting document entities based on their logical and spatial configurations. The PDFVQA dataset provides only coarse-grained, entity-level annotations, necessitating further processing for models that require fine-grained analysis. This setup is ideal for testing models' capabilities in understanding the logical and spatial structures of document images.

(2) Datasets for Multi-Page VRD QA

TAT-DQA [158], an extension of the TAT-QA [159] dataset, is developed with more complex natural document structures and an expanded set of manually corrected and generated question-answer pairs derived from business financial reports. Unlike other datasets primarily focusing on extractive or simple abstractive answers (such as counting or yes/no), TAT-DQA includes questions requiring arithmetic reasoning, where values must be extracted from tabular data and textual content for discrete calculations. This dataset adopts the evaluation metrics of TAT-QA, including Exact Matching and a numeracy-focused F1 score. These metrics are particularly tailored to assess the accuracy of arithmetic reasoning and data extraction capabilities of the models tested with TAT-DQA.

RDVQA dataset [137] compiles a large collection of conversational chats and associated images from an E-commerce platform. It employs standard OCR and Named Entity Recognition (NER) techniques to extract text and redact sensitive information, ensuring privacy protection through masking. The dataset includes question-answer pairs within the images, which are manually verified to confirm image clarity and the presence of at least one question-answer pair per image. Although some documents span multiple pages, this dataset is structured such that it can be processed relatively easily by single-page VRD QA models.

PDFVQA-Task C [27] is a distinct sub-task within the PDFVQA dataset that expands document VQA to encompass entire long documents, moving beyond the single-page focus of Tasks A and B. In Task C, to answer questions, the model often needs to retrieve information from multiple document entities. Thus, Task C employs Exact Matching for its ground truth annotations. Similar to Tasks A and B, additional processing is required for evaluating models at a fine-grained level.

MP-DocVQA [119] extends the original DocVQA [84] dataset to accommodate multi-page document analysis. This version includes adjacent pages from the same documents, expanding the dataset from 12,767 to 64,057 document images. In adapting to a multi-page format, some inappropriate questions were removed. However, it's important to note that while the dataset allows for questions across multiple pages, the answers remain confined to individual pages; there are no cross-page answers in the MP-DocVQA dataset.

DUDE [122] is the first cross-domain, multi-page VRD QA dataset, featuring a diverse collection of documents from various fields such as medical, legal, technical, and financial, and different document types, including CVs, reports, and papers. It comprises 5,019 documents, 28,709 document pages, and 41,541 manually annotated questions. Question types vary from extractive in-line text and Yes/No answers to multi-hop reasoning and structural understanding, similar to those in PDFVQA [27]. To evaluate model performance, DUDE uses the ANLS metric [84] for assessing answer prediction accuracy. Additionally, it employs two other metrics: Expected Calibration Error [35] and Area-Under-Risk-Coverage-Curve (CURC) [32, 48] to gauge the overconfidence and miscalibration in document understanding models.

These features make DUDE a comprehensive tool for evaluating cross-domain document understanding models.

MMVQA [28] is a dataset sourced from PubMed Central, designed for the retrieval of multimodal semantic entities from multi-page documents. The questions are generated using ChatGPT [87] and subsequently verified manually. Unlike other datasets that focus solely on in-line text or text-dense entities, MMVQA also considers entire tables and figures as potential answers to the given questions. This dataset introduces various evaluation metrics to cater to different application scenarios: Exact Matching and Partial Matching Accuracy assess the precision of responses, while Multi-label Recall evaluates how well the model identifies all relevant answers across the document. This diverse set of metrics makes MMVQA suitable for comprehensive performance evaluation in complex, multimodal document understanding tasks.

2.5.3 Summary

The common characteristics of KIE and VRD QA datasets are represented in Table 2.1 and Table 2.2. Most existing benchmark datasets for key information extraction are designed for single-page scenarios and predominantly cater to English documents. However, real-world applications often involve more complex multi-page forms. Even forms typically require input from multiple parties and contain multiple languages, such as customs or import/export declaration forms, presenting unique challenges not addressed by current datasets. Similarly, in the domain of VRD QA, while multi-page VRD QA has gained increased attention, the integration and exploration of multimodal information remain insufficiently developed. Specifically, structure and relation-aware VRD QA, which is crucial for interpreting and understanding cross-page relationships, is still a largely unexplored area.

CHAPTER 3

Visually Rich Form Document Understanding

This chapter is based on the work ***Form-NLU: Dataset for Form Natural Language Understanding***, published at **SIGIR 2023** [26]. I was responsible for formulating the research aim, designing the methodology, conducting the dataset annotation and analysis, implementing the experiments, and writing the entire paper.

This work introduces a novel natural language understanding dataset to address existing gaps in form understanding, accompanied by a detailed analysis. It also presents a novel framework for extracting key information from complex form layouts and formats, including digital, handwritten and printed.

Understanding and retrieving the form document structure is challenging compared to general document analysis tasks. Form documents are typically made by two types of authors: A form designer, who develops the form structure and keys, and a form user, who fills out form values based on the provided keys. Hence, the form values may not be aligned with the form designer's intention (structure and keys) if a form user gets confused. In this paper, Form-NLU is introduced, the first novel dataset for form structure understanding and its key and value information extraction, interpreting the form designer's intent and the alignment of user-written value on it. Form-NLU also includes three form types: digital, printed, and handwritten, which cover diverse form appearances and layouts. A robust positional and logical relation-based form key-value information extraction framework is proposed. Using this dataset, Form-NLU first examines strong object detection models for the form layout understanding and then evaluates the key information extraction task on the dataset, providing fine-grained results for different types of forms and keys. Furthermore, I examine it with the off-the-shelf PDF layout extraction tool and prove its feasibility in real-world cases.

3.1 Introduction

3.1.1 Background

The structural information and its value extraction in a document can be a valuable source for Natural Language Processing (NLP) tasks, especially information extraction and retrieval. Recently, NLP communities and industries, like IBM and Microsoft, have proposed a range of techniques to *understand* positional and/or logical structure of documents [80, 101, 151, 157], and *extract* the essential information[25]. Those researches have mainly focused on Visually Rich Document(VRD)-based tasks, such as academic papers[101, 157], receipts[46, 91], and forms [49, 141], and many benchmark problems have been solved, including layout analysing [67, 157], table structure recognition [19, 155, 156], document question answering [84, 116].

Most VRD-based problems have been successfully solved. However, the form-understanding task is relatively challenging. This is mainly because of two reasons: two types of authors in a form and the combination of diverse visual cues. First, unlike general VRDs, the main aim of forms is very clear, *collecting data from the form users*, and this aim is applied from the medical domain to administrative data collection. According to this aim, it can be assumed that there are two main authors: a form designer and a form user. A form designer focuses on developing a form structure to collect the required information by defining the clear key point. The developed form would be used as a user interface so a form user can supply the form value based on their understanding. Unfortunately, not every form is clear and easy to understand. To collect the diverse required information, several form designers tend to make forms with diverse layouts/structures, which have complex logical and positional relationships between semantic entities. The form user can easily get confused with the designer's intention, and this derives a wrong alignment of key-value pairs. The confusion about the form designer's intention and the uncertainty of the form user would raise the difficulty of understanding and extracting information from the form document. Secondly, due to the involvement of *form developers-to-users* relationship, the form has a high possibility of having a combination of different natures, such as digital, printed, or handwritten. For example, a designer may provide users with electrical paper forms, while users may submit filled forms via various

carriers, such as digital, printed or handwritten versions. It also commonly happens that users provide various types of noise (low-resolution, uneven scanning, or bad handwriting) in the submitted forms. This causes huge difficulties in understanding form document structure and extracting the essential key-value pairs. Regarding form understanding, several datasets have been released in recent years, collected from scanned receipts [46, 91], contracts[113], and cross-domain forms [49, 141] (shown in Table 3.1). However, those datasets produce the form developer's intention as relatively simple and general, which does not deal with the confusion about the form designer's intention and the form user's uncertainty. Moreover, most datasets do not cover the various carriers of document versions and their noises. This would worsen understanding of the form structure and extracting the key information.

3.1.2 Research Aims

In this paper, a new dataset is introduced for form structure understanding and key information extraction. The dataset would enable the interpretation of the form designer's specific intention and the alignment of user-written value on it. Our dataset also includes three form types: digital, printed, and handwritten, which cover diverse form appearances/layouts and deal with their noises. In addition to this, a novel model is proposed for form structure understanding and key-value information extraction, which applies robust positional and logical relations. To do this, I cover user-caused form diversities to precisely extract key information from forms, which is currently an emerging industrial demand. The model covers the hierarchical structure of documents, from words to sentences, sentences to semantic entities like a paragraph, and entities to a document page. Note that exiting transformer-based models mainly focus on token [45, 139, 140, 141] or entity level [66, 114, 151] independently, ignoring the contextual dependency between different level elements. In the evaluation, I first examine strong object detection models for understanding the form layout and then test the proposed model for the key information extraction task. Moreover, we also examine our key information extraction dataset and proposed model with the off-the-shelf pdf layout extraction tool and prove their feasibility in real-world cases.

Name	Source	Type			Features		Designer
		D	***P***	***H***	**B.Box**	**Text**	**Intention**
FUNSD [49]	Noise Form	×	◯	◯	◯	◯	General
XFUND [141]	Synthetic Form	◯	×	◯	◯	◯	General
EPHOIE [128]	Exam Paper	×	×	◯	◯	◯	General
Charity [113]	Annual Report	◯	◯	◯	×	◯	Specific
NDA [113]	Agreements	◯	×	×	×	◯	Specific
Form-NLU **(Ours)**	Financial Form	◯	◯	◯	◯	◯	Specific

TABLE 3.1. Summary of form understanding datasets.

3.1.3 Contributions

The main contribution of this research can be summarised as follows: 1) I introduce Form-NLU, a new form structure understanding and key information extraction dataset that covers specific form designers' intentions and aligns with the user's values. 2) I propose a novel model that handles form documents' positional and logical relations and hierarchical structure. The proposed model has outperformed other SOTA form key information extraction models on Form-NLU. 3) I apply the proposed dataset and the model with the off-the-shelf pdf layout extraction tool and prove the feasibility in real-world form document cases.

3.2 Related Work

There are established benchmarks in document understanding, such as those introduced in [101, 116, 157]. Form-NLU focuses on form document understanding and information extraction, which entail multi-party interactions leading to intricate positional and logical relationships, as illustrated in Table 3.1. There are three major points that warrant discussion and comparison with previous benchmarks. First, most of benchmarks [49, 128, 141] do not cover the various carriers of document versions and their noises. For example, a form designer may provide the form with digital forms, while the form user may submit the filled form with various carriers, such as digital (***D***), printed (***P***) or handwritten(***H***) version. This trend commonly affects the quality of form structure understanding and information extraction due to the various types of noise, including resolution or scanning issues. Hence, the successful benchmark should cover real-world cases with various carriers. Secondly, in order to conduct

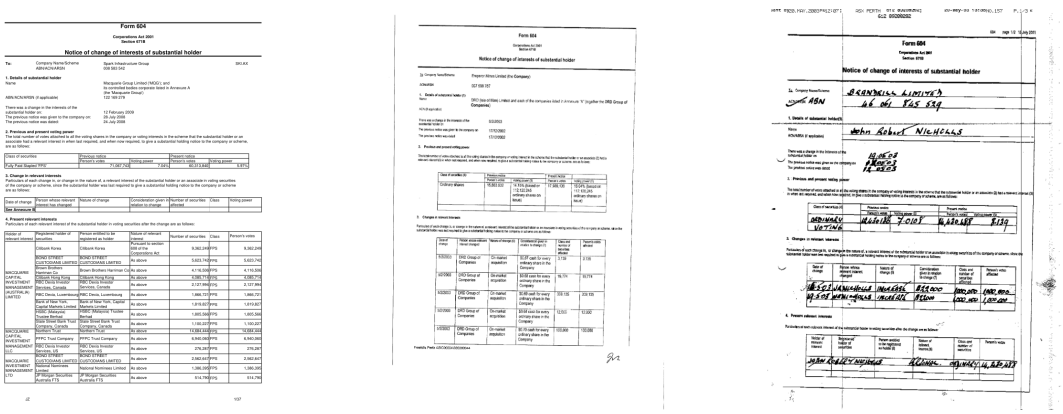
the form structure understanding and information extraction, it is crucial to interpret the positional and logical relationships between form components. Most benchmarks enable understanding positional and logical/semantic relations by using bounding boxes (B.Box) and Textual information. However, [113] releases two form information extraction datasets, NDA and Charity, without providing bounding box coordinates of semantic entities. Finally, the form document has two main authors: form designers and form users. The form designers design the form structure to collect the required information (designer's intent), and the form users try to understand the designer's intention but easily get confused. Hence, handling the designer's intention is crucial to understanding the form structure and extracting key information. Several popular benchmarks [49, 141] produce the form developer's intention as relatively simple and general, which does not deal with the confusion about the form designer's intention and the form user's uncertainty. For example, those datasets cover simple form component types, keys and values. Scanned exam paper [128] datasets have a relatively simple layout with horizontal key-value pair structures, which do not include any dynamic layout components, such as tables, paragraphs, or complex key-value pairs. **Form-NLU** is the first visual-linguistics form language understanding dataset for supporting researchers in interpreting specific designer intentions under noises from user's input with various types of form carriers.

3.3 Dataset

3.3.1 Data Collection

Form-NLU is a subset of the publicly available financial form data source for Form 604 (notice of change of interests of the substantial holder)[1] collected by SIRCA. It provides the text records of each substantial shareholder notice form submitted to the Australian Stock Exchange (ASX)[2] from 2003 to 2015. To better comprehend the form designer's intentions, the twelve most essential form fields [13] are included. Each form field expects a **Value** from

[1] https://asic.gov.au/regulatory-resources/forms/forms-folder/604-notice-of-change-of-interests-of-substantial-holder/
[2] https://www2.asx.com.au/



(A) Digital (***D***) (B) Printed (***P***) (C) Handwritten (***H***)

FIGURE 3.1. Digital, printed and handwritten form samples.

the user, which should contain the specific information being asked by the corresponding **Key** that expresses the form designer's intention. For example, the key of *"company name"* anticipates a string value that gives the name of a company, whereas *"voting percentage"* asks for a number that tells the portion of the voting. In order to address the aforementioned developer-to-user relationships in real-world scenarios, three variants of format are covered depending on how the forms were filled and submitted by different users: (1) Digital (***D***), (2) Printed ($\mathcal{P}$) and (3) Handwritten (***H***). The Digital forms are filled in digitally and directly submitted as PDF files. The Printed and Handwritten forms are filled in digitally and by hand, respectively, and then scanned first before being saved as PDF files for submission. A sample of each type of the form is provided in Figure 3.1, which demonstrates the potential diversity in terms of the form format, e.g., font styles and rotations.

3.3.2 Annotation Guideline

Based on the task goals (see Section 3.5), the annotation schema and guidelines are designed. For Layout Analysis (as Task A[3]), seven distinct form layout component categories (*Title*, *Section*, *Form Key*, *Form Value*, *Table Key*, *Table Value* and *Others*) are created that encompass all possible components from the collected forms. Each form layout component represents a semantic segment of the text area. The *Title* and *Section* correspond to the title of the form

[3]See Section 3.5.1 Task A - Form Layout Analysing

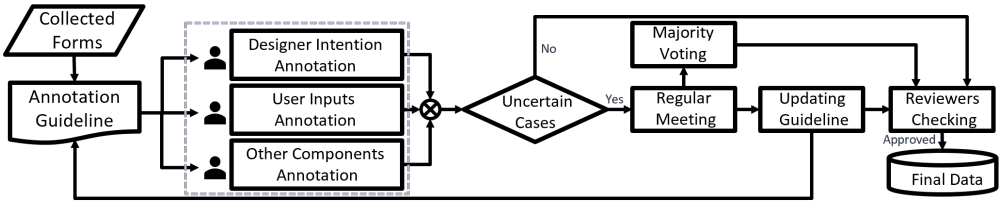


FIGURE 3.2. Form-NLU overall annotation workflow.

document and the sections, respectively. The *Form/Table Key* and *Form/Table Value* refer to the intended questions of the form designer (e.g. Company Name) and the corresponding user input fields. The difference between the *Key/Value* of form and table is that, by our definition, the *Form Key-Value* pairs are horizontally aligned, whereas the *Table Key-Value* pairs are vertically aligned. Any components that do not fall into these four types are defined as *Others*, such as a customised paragraph and cells added and modified by some companies on their own purposes.

While the Layout Analysis (as Task A) focuses on understanding the overall form structure via those general semantic components, the Key Information Extraction (as Task B[4]) aims to comprehend different intentions of the form designers. Thus, in Task B, the *Key* and *Value* components from Task A are further distinguished based on the twelve Key intentions, which include: (1) Company Name (*com_nm*), (2) Company ID (*com_id*), (3) Holder Name (*hold_nm*), (4) Holder ID (*hold_id*), (5) Change Date (*chg_date*), (6) Previous Notice Given Date (*gvn_date*), (7) Previous Notice Date (*ntc_date*), (8) Class of Securities (*class*), (9) Previous Share (*pre_shr*), (10) Previous Voting Percentage (*pre_pct*), (11) Share (*new_shr*), (12) New Voting Percentage (*new_pct*). *Key* (1)-(2) and *Key* (3)-(4) ask for the identity of the listed company and the substantial holder, respectively, whereas *Key* (5)-(12) query the value of the shares being changed. The identifiers *Key* and *Value* were simply appended separately to the twelve (12) *Key* names, and twenty-four (24) labels for the key-value pairs were created accordingly. For example, *"Company Name Key"* and *"Company Name Value"*. Based on these different key intentions, the value patterns are inspected to explore the potential noise caused by the diversified user understanding and uncertainty. The statistical analysis can be found in Section 3.4 (Figure 3.3).

[4]See Section 3.5.2 Task B - Key Information Extraction

3.3.3 Annotation Procedure

Three human annotators and two human reviewers are recruited[5] and performed an iterative annotation for each form component using the DataTorch[6]. Specifically, as shown in Figure 3.2, the annotation procedures are split into three specialised sub-tasks: (1) *Key annotation* for the aforementioned 12 types of *Keys*, (2) *Value annotation* for the paired *Values*, and (3) *Non-Key-Value annotation* for the rest of the non-Key and non-Value components, including *Title*, *Section* and *Others*. Each annotator was assigned one sub-tasks respectively. They identified all possible targeted components and annotated the corresponding rectangular bounding boxes and labels based on the annotation guideline. During the annotation, any uncertain cases would be marked by the annotator and further decided by the discussion and majority voting in the regular meeting. The annotation guidelines were updated whenever needed to handle similar cases later. For instance, one typical type of uncertain case was the annotation of *Form Key* due to its various formats made by different form users, such as *"The previous notice was dated 12 Feb 2003"* as an example for the Key (7) Previous Notice Date, where the final agreed Key *"The previous notice was dated"* is expressed in a complete sentence with the Value *"12 Feb 2023"*. This iterative annotation process produced 757 annotated Digital form documents. In addition, 50 printed and 50 handwritten forms were randomly selected, and the annotation process was conducted like the digital forms. These printed and handwritten forms are used as additional test sets to explore the possibility of handling different forms in real-world scenarios.

To ensure the final quality of the human annotation, the annotated bounding boxes with labels are visualised of all forms and assigned them to the two human reviewers for parallel manual checks. They checked each form one by one with reference to the up-to-date annotation guideline and annotated the correctness. They reviewed any form annotated with the incorrect label by both reviewers or received disagreement between them (i.e. one of them annotated as correct while the other annotated as incorrect). Then, those are modified based on the two reviewers' final agreed-upon decision. To measure the quality of the annotations, the

[5]The three annotators have a background in Computer Science or Financial at the University of Sydney while the two human reviewers are financial domain experts.

[6]https://datatorch.io/features/annotator

annotation agreement rate is calculated using both Cohen's Kappa [21] and the Hamming Loss, which derived the overall scores of **0.998** and **0.003**, respectively, indicating a high annotation quality.

3.3.4 Annotation Format

The final annotation of our proposed **Form-NLU** is provided in *.json* format, separately for Task A and B. The annotations for each form segment are stored as a *dictionary object* containing multiple key-value pairs with indicative key names in the *.json* file. The main attributes for each form segment shared by the two tasks include ***bbox*** for bounding box coordinates, ***text*** for textual tokens of this specific segment (e.g. from OCR or pdfminer[7]), and the ***label*** for this segment based on the specific task, e.g., *"Section"* for a segment in Task A or *"Company ID (com_id)"* for a segment in Task B. Besides, some auxiliary attributes are included derived from the experiments for potential development in the future, such as the ***segmentation*** coordinates for Task A as well as the ***visual_feature*** and ***bert_cls*** for Task B (See Section 3.6.4, 3.6.5).

3.4 Dataset Analysis

3.4.1 Component Distribution

The final version of our annotated **Form-NLU** consists of 857 forms, including 757 Digital, 50 Printed and 50 Handwritten forms. Table 3.2 shows the overall statistics of the data splits and the breakdown distribution of form components. I randomly split the Digital forms using the ratio of 70/10/20, resulting in 535, 76 and 146 forms for training, validation and testing. Besides, the 50 Printed and 50 Handwritten forms are included as our additional test splits. The breakdown distribution shows that the *Key* and *Value* persistently dominate since they are the main contents of the form. *Values* can be less than *Keys* due to the cases of empty values. The *Title* and *Section* are the least and demonstrate similar occurrences because each

[7]PDFMiner is used to extract the text of digital-born forms $\boldsymbol{D}$ and Google Cloud Vision to extract the text of printed and handwritten sets ($\mathcal{P}, \mathcal{H}$)

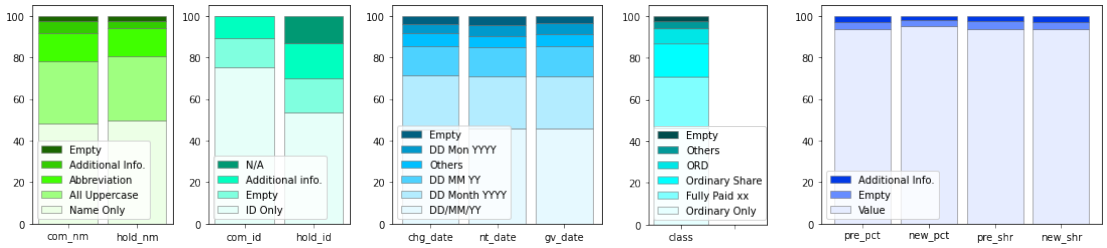
Type	Usage	Form Images	Title	Section	Form		Table		Others
					Key	Value	Key	Value	
Train	**Digital**	535	1068	1070	3708	3568	2669	2669	1691
Val		76	152	152	524	510	380	379	246
Test	**Digital**	146	292	292	1009	978	730	730	458
	Printed	50	98	100	346	332	250	249	152
	Handwritten	50	100	100	348	315	249	226	149
Total Number		857	1710	1714	5935	5703	4278	4253	2696

TABLE 3.2. Number of form components for each data split.

Type	Title	Section	Form		Table	
			Key	Value	Key	Value
Bounding Box Average Width	186.15	136.82	95.46	85.98	57.29	41.92
Bounding Box Average Height	30.44	12.47	12.47	12.47	10.50	10.45
Bounding Box Average PX	4275.94	1720.62	1204.96	1183.10	609.75	456.94
Average Number of Tokens	7.35	7.10	5.16	4.14	4.29	1.74

TABLE 3.3. Average bounding box width, height, number of pixel(PX), and number of tokens for each form component

form document normally contains one main title with optional one or two subtitles while always having the two essential sections. The *Others* are slightly more frequent than the *Title* or *Section*, and the occurrence depends on the different form fillers. This consistent overall distribution across forms is attributed to the use of the Form 604 template. It makes our **Form-NLU** a promising benchmark dataset for exploring solutions to the real-world financial form layout understanding problem when the standard form template is available (our Task A). In addition, I further differentiate the keys and values for comprehending the specific form intentions (our Task B). These key-value pairs are contained in either *Table* (mostly vertical key-value alignment) or *Form* (mostly horizontal key-value alignment), which indicates the variety of spatial relationships that require the model to learn in order to identify the values and align with each key accurately. In Table 3.3, the average bounding box size and textual tokens of each form component are provided, which reflects both their visual and linguistic features that can be potentially helpful for specific key and value identification. *Title* and *Section* tend to have longer textual content and bigger component sizes with larger font sizes. In comparison, *Key* and *Value* are much shorter and are contained in smaller bounding boxes. Especially the *Value* in the *Table* has the shortest length as they are mostly numeric values.

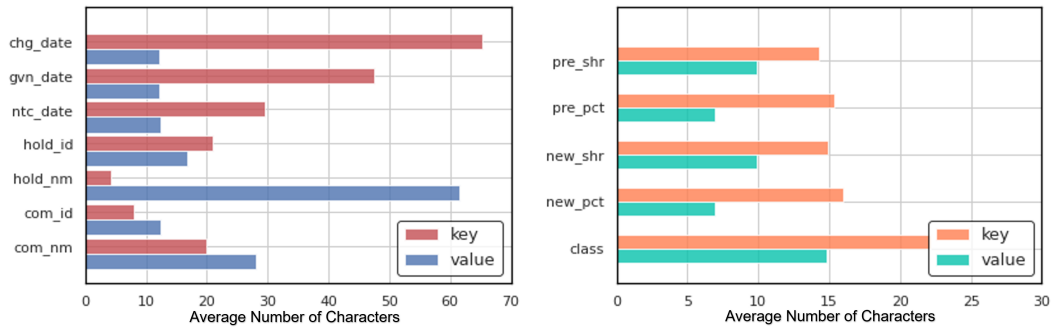


(A) name (B) id (C) dates (D) class (E) percent/share

FIGURE 3.3. Value patterns distributions for each key group.

3.4.2 Value Pattern Proportion

As mentioned before, the filled-in content varies and involves potential noises due to the diversified user understanding and uncertainty. To illustrate the variety of the filled-in content, I summarise the main value patterns for each key intention and provide the proportions in Figure 3.3. Throughout all keys, the largest proportion of value follows the common value pattern that the form users tend to provide minimal information to fulfil the intention of the keys. For example, they put simply the company name only to the key *com_nm* in Figure 3.3a or ID only for 3.3b. However, there are several other value patterns available, including additional information shown in Figure 3.3a, 3.3b and 3.3e, or having different data format or writing style, such as the various date formats in Figure 3.3c. These human-caused variants well reflect the real-world form understanding scenarios and imply potential challenges for achieving precise information retrieval in our Task B. Overall, the Share Class (Figure 3.3d) and the date type values (Figure 3.3c) show comparatively more patterns while the numerical type values in Figure 3.3e tend to be more consistent with the common pattern. This pattern similarity among the keys with the same data type values indicates that solely relying on the linguistic cues may not be enough, as understanding the relative spatial location of the values in the form is also required in order to distinguish these similar values of different keys.

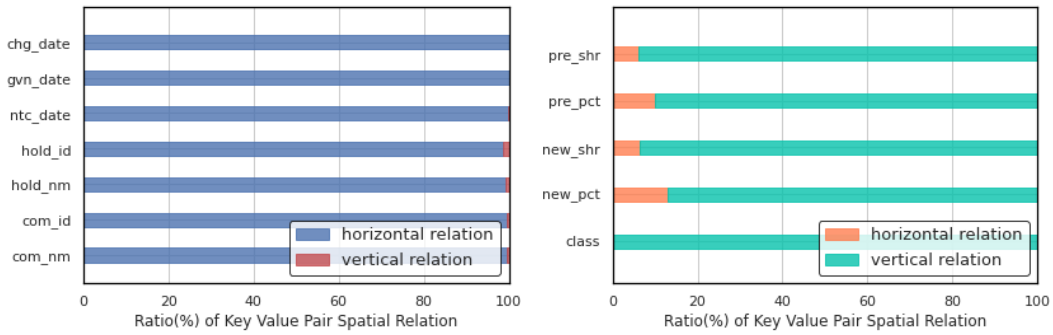


(A) Form Key Value Pairs (B) Table Key Value Pairs

FIGURE 3.4. Average number (#) of characters in each key value pair.

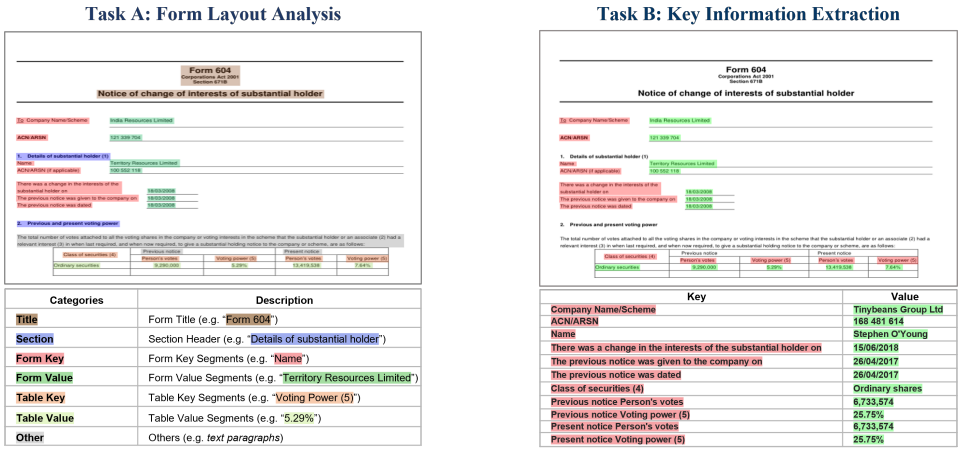
3.4.3 Key-value Pair Comparison Analysis

To provide an in-depth analysis of the nature of key-value pairs, I further summarise the average character number and the ratio of spatial relation for the twelve key-value pairs in Figure 3.4 and 3.5, respectively. It can be observed that typically there are seven fixed key-value pairs in Form (i.e., Figure 3.4a/3.5a) and another fixed five in Table (i.e., Figure 3.4b/3.5b) formatted by the standard Form 604 template. Two groups of key-value pairs demonstrate their own features. As shown in Figure 3.4a for the form-based key-value pairs, the three date type pairs tend to have much longer value than the key where as the other four pairs may have either longer or shorter values than keys. The *hold_nm* key-value shows the largest character length gap among all. In comparison, the five Table-based pairs in Figure 3.4b all have shorter values than keys, and the length gaps are similar. Overall, most keys have short values (e.g., < 20 characters) from which the semantic context could be scarce. On the other hand, as seen from Figure 3.5, the seven Form-based pairs are uniformly aligned horizontally, whereas the other five pairs in Table are mostly vertical. However, some special cases with the opposite alignment are also observed in both groups, such as the vertical key-value pairs for the *hold_id*/*hold_nm* and *com_id*/*com_nm* in the Form-based group (Figure 3.5a), as well as the horizontal pairs for the four percentage type keys and values in the Table-based group (Figure 3.5b). Thus, simply memorizing the spatial alignment for each key cannot lead to optimised value retrieval in Task B.



(A) Form Key Value Pairs (B) Table Key Value Pairs

FIGURE 3.5. Ratio (%) of horizontal and vertical relation for each key value pair.



(A) Task A Sample (B) Task B Sample

FIGURE 3.6. Examples for Task A and B. For Task A, the users need to detect each layout component's bounding box and recognise the associated category. Task B asks the user to feed the fixed key text (red area) into the model to predict the corresponding value of RoI's index (green area).

3.5 Tasks Overview

3.5.1 Task A - Form Layout Analysing

The purpose of Task A is to detect the semantic entities (*Title*, *Section*, *Form_key*, *Form_value*, *Table_key*, *Table_value*, *Others*) of forms. This is a prior task to understand the form layout by detecting the position of each semantic entity, including key, value and other objects. Given a form image I, an object detection model is used to detect a set of RoIs $(r_1, r_2, ..., r_n)$.

Each r_i contains bounding box b_i coordinates (x_i, y_i, w_i, h_i) with semantic category c_i. As shown in Figure 3.6a, the bounding box of each semantic entity in an input form image is detected and coloured based on the recognised categories. The model should effectively detect layout components such as form *Title* (like " Form 604 ") and *Section* headers (such as " Details of substantial holder "). Additionally, I also expect the models can differentiate keys/values located in *Form* or *Table*; for example, " Name " should be detected as a *Form_key* instance, while " Voting Power (5) " is a *Table_key* instance.

3.5.2 Task B - Key Information Extraction

Task B aims to evaluate whether the proposed models could comprehensively understand designers' intentions and user uncertainties to extract valuable information from input forms. It allows using ground truth RoIs' set $\mathcal{R}_{gt} = (R_1, R_2, ..., R_n)$ during the training and inference stage. Given key text information t, a document image I and a set of ground truth RoIs $(r_1, r_2, ..., r_n)$, a model H can output the RoI's index number aligned with input t. As Figure 3.6b shown, each highlighted entity with a red background colour is the key t needs to feed into the proposed model, and the green paired RoI's index is the desired output from the model based on the current input key. For example, inputting the required key text content " Company Name/Scheme ", the desired output from the model is the RoI's index of paired values where the text content can be easily acquired (" Tinybeans Group Ltd ") from it.

3.6 Evaluation Setup

3.6.1 Implementation Detail

For Task A, Faster-RCNN and Mask-RCNN models are fine-tuned with two backbones[8] respectively on our dataset based on Detectron2 platform. I set 5000, 128, and 0.02 as the maximum iteration times, batch size and base learning rate, and other setups are the same

[8]Faster-RCNN backbones are faster_rcnn_R_50_FPN_3x, faster_rcnn_R_101_FPN_3x and Mask-RCNNs are mask_rcnn_R_50_FPN_3x, mask_rcnn_R_101_FPN_3x

as Detectron2 official tutorial [9]. Regarding Task B, various approaches are employed to encode the vision and language features. Firstly, all Task B adopted baselines use pretrained BERT to encode key textual content. Moreover, for the visual aspect, VisualBERT, LXMERT, and M4C models utilise 2048-d features extracted from the Res5 layer of ResNet101. The maximum number of input key text tokens and the number of segments on each page are all defined as 50 and 41, respectively. Task A and B experiments are conducted on 51 GB Tesla V100-SXM2 with CUDA11.2.

3.6.2 Task A Baselines and Metric

Two popular object detection models are adopted, **FasterRCNN** [103] and **MaskRCNN** [39] with different ImageNet pretrained backbones testing on our dataset. To evaluate the performance of the object detection model, a mAP (mean Average Precision) is applied, which is commonly used in object detection tasks. The mAP score is calculated by averaging APs over all categories' overall pre-defined IoU thresholds[10].

3.6.3 Task B Baselines and Metric

Task B mainly focuses on predicting the corresponding value of the RoI index based on input text content (as shown in Figure 3.6b). Thus, several multimodal transformer frameworks are adopted as baselines on **Form-NLU**, of which inputs are multi-aspect RoI features, including large pre-trained **VisualBERT** [66], **LXMERT** [114] and non-pre-trained **M4C** [43] models [11]. Regarding evaluation, the weighted F1-score is used as the primary evaluation metric for representing overall and breakdown performance. Note that some visual language pre-trained models, such as ViLT [55], LayoutLM [140], were excluded since those are mainly based on the image patches or pieces and do not fit into the task demands.

[9] https://colab.research.google.com/drive/16jcaJoc6bCFAQ96jDe2HwtXj7BMD_-m5

[10] We refer [157] to adopt mAP as metrics and follow their thresholds for Task A.

[11] Detailed baseline setup can be found in https://github.com/adlnlp/form_nlu#baseline-model-description

3.6.4 Task A Model

For Task A - Form Layout Analysing task (Section 3.7.1), Faster-RCNN and Mask-RCNN with various depth backbones are used (ResNet-50 and ResNet-101) as baselines and contact experiments to check the effects of diverse model architecture and model size. Additionally, for the real-world case analysis (Section 3.7.3), some methods using open-source PDF parsers are adopted (such as PDFMiner) to analyse the form layout without training the deep learning models. The PDF parser outputs can also be used as inputs for Task B.

3.6.5 Task B Model

For Task B, a new document key information extraction model is proposed (as Figure 3.7 shown) to predict the corresponding values from input key content. To achieve this, we investigate positional and logical relations and their hierarchical structure with two components: multiple aspect features extraction/integration and entity-token dual-level Model.

(1) Multi-Aspect Features

Five aspect features are investigated, including Visual (V), Textual (T), Positional (P), Density (D), and Gap Distance (G) features, with the corresponding encoding approaches to comprehensively explore which features benefit understanding designer intentions and argument form-like document segment representations. Many traditional methods for document understanding tasks have shown the effectiveness and significance of visual, textual, and positional features only [139, 140, 151]. Except for those three aspect features, exploring the effectiveness of text density for form understanding is also significant, which has been demonstrated by [80] for document layout analysis and using the multimodal integration [10]. Moreover, based on the dataset analysis results, the gap distance between entities is crucial for understanding the form structure. Hence, two layout-related features are introduced, including normalised positional features (bounding box coordinates) and the gap distance between an entity and its neighbours. **(2) Entity-Token Dual Level Model**

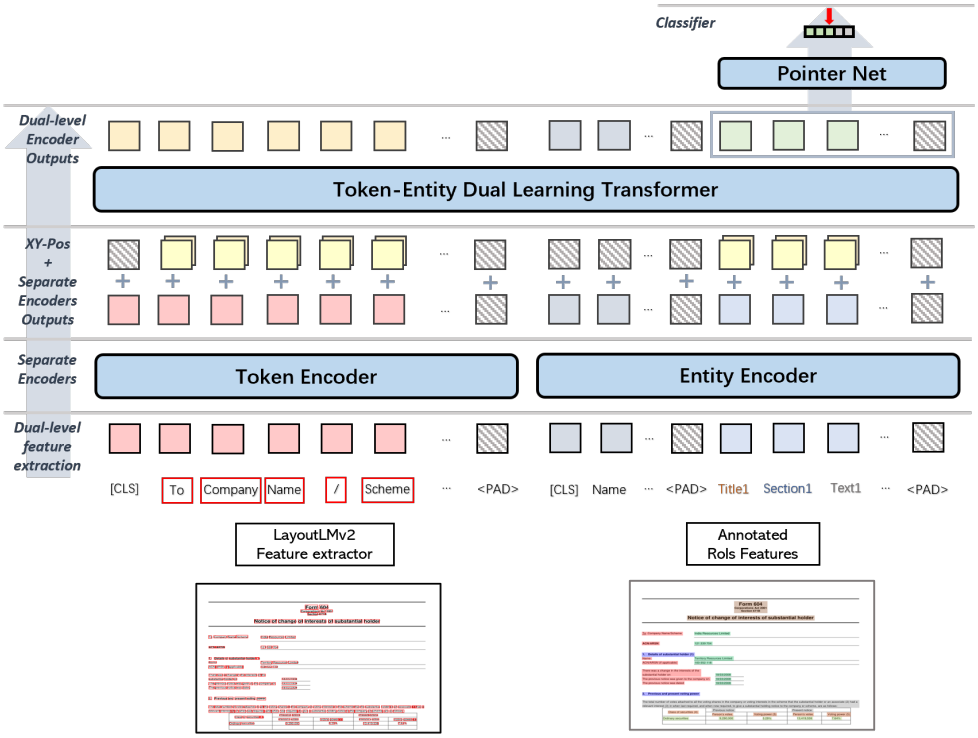


FIGURE 3.7. The proposed key information extraction model architecture for task B.

The proposed new entity-token dual-level form information extraction model contains four components, including Entity Encoder $Encoder_{entity}$, Token Encoder $Encoder_{token}$, Dual Encoder $Encoder_{dual}$ and Pointer Net-based classifier ϕ.

1) Entity Encoder aims to learn the semantic relations between entities for enhancing the vanilla entity representations. The pretrained LXMERT [114] is selected as the entity encoder based on the preliminary results. The inputs of the LXMERT-based entity encoder include key text T_{key}, Entities' visual features V_{entity}, and the bounding box coordinates B_{entity}. After feeding those features into pretrained LXMERT, the enhanced Entity RoIs' representations $E \in \mathbb{R}^{768}$ can be extracted.

2) Token Encoder aims to acquire cross-modal token representations. The SoTA layout-aware pre-trained visual language transformer, LayoutLMv2 [139], is adopted as the token encoder. *LayoutLMV2FeatureExtractor* is firstly employed to extract and encode the token level features, including token text T_{token}, token RoIs' bounding box B_{token} with pixel features I_{pixel}. Then, I follow the original LayoutLMV2 setup to process the extracted token-level

features and feed them into our token-level encoder $Encoder_{token}$, pretrained LayoutLMv2 to get the token-level representation $T \in \mathbb{R}^{768}$.

3) XY-Pos Dual Encoder is designed to learn geometrically sensitive multi-grained and multi-modality feature representations. After getting entity and token level feature representations from pretrained models, are treated as the sequence inputs of a dual-level mutual learning encoder $Encoder_{dual}$, where a 6-layer transformer with 8-*heads* self-attention is used as the basic framework. Unlike the original transformer that adopts sinusoidal positional encoding [123], I proposed a new geometric-sensitive positional encoding *XY-Pos*. It flattens the stacked sets of normalised bounding box coordinates along with the X or Y axis to get X_i^{pos} and Y_i^{pos}. Supposing $b_i = (x_i, y_i, w_i, h_i)$ is the bounding box of r_i of $Encoder_{dual}$ in document page D_j, the size of D_j is (W_j, H_j).

$$pos_i^x = [\frac{x_i + d_{i_1}}{W_j}, \frac{x_i + d_{i_2}}{W_j}, ..., \frac{x_i + d_{i_m}}{W_j}] \tag{3.1}$$

where the l-th step $d_{i_l} = \frac{w_i \times l}{m}$

$$X_i^{pos} = flatten([pos_i^x] \times n) \tag{3.2}$$

I set m and n as 32 and 24 to get a 768-d vector. Finally X_i^{pos} is flattened into a $m \times n$ dimensional vector, to have $X_i^{pos} \in \mathbb{R}^{768}$. Similar procedures are used to generate positional encoding Y_i^{pos} along *Y-axis* of r_i. The final input representations before feeding into $Encoder_{dual}$ can be represented as:

$$E_{pos_{xy}} = E + X_{entity}^{pos} + Y_{entity}^{pos} \tag{3.3}$$

$$T_{pos_{xy}} = T + X_{token}^{pos} + Y_{token}^{pos} \tag{3.4}$$

Unlike most VLPMs [114, 139] adopted positional encoding methods through linear projecting 4-d bounding box coordinates into high dimensional vectors. *XY-pos* could capture more geometric features between entities and tokens. Then, they are fed into $Encoder_{dual}$ and get the updated token and entity representations T_{dual}, E_{dual}.

Test set	Model	mAP	Breakdowns ($Precision$)						
			Title	**Section**	**Form_key**	**Form_value**	**Table_key**	**Table_value**	**Others**
D	**F-50**	68.98	73.54	69.36	67.93	70.37	**69.18**	66.76	65.73
	F-101	69.99	71.32	68.30	72.37	71.70	68.74	72.84	64.65
	M-50	71.74	77.29	69.06	**73.01**	70.30	68.40	73.25	**70.90**
	M-101	**71.93**	**79.54**	**69.64**	71.25	**72.38**	67.02	**74.38**	69.33
P	**F-50**	56.46	57.44	53.98	54.49	51.46	59.31	62.86	55.67
	F-101	59.54	55.72	52.81	63.13	61.32	59.42	68.11	56.26
	M-50	63.06	59.83	60.98	63.74	**61.91**	70.75	65.93	58.29
	M-101	**65.47**	**63.05**	**62.49**	**66.84**	61.80	**71.82**	**70.19**	**62.10**
H	**F-50**	49.87	58.73	43.85	63.14	25.15	63.48	40.14	54.56
	F-101	50.39	53.92	33.36	65.28	36.68	66.97	41.06	55.43
	M-50	57.99	61.45	48.46	68.16	43.14	71.76	**56.95**	56.05
	M-101	**60.22**	**64.81**	**54.46**	**69.97**	**46.82**	**73.34**	52.81	**59.32**

TABLE 3.4. Overall performance and breakdown results for layout analysing task (Task A). F-50 and F-100 represent the Faster-RCNN with ResNet50 and ResNet101 as backbones. The same patterns occur for Mask-RCNN (M-50 or M-101).

4) Pointer Net-based Classifier The multi-aspect features are concatenated with E_{dual} to get E_{multi}. E_{multi} is fed into the pointer net [124] based classifier ϕ to get a score vector which following a $softmax$ to retrieve the final prediction results y_{pre}.

3.7 Results

3.7.1 Task A: Form Layout Analysing

(1) Overall and Breakdown Performance

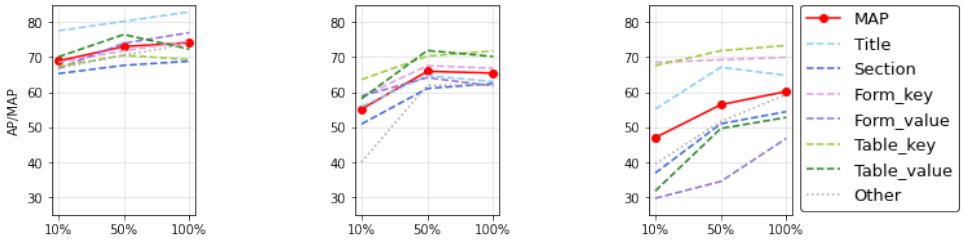
This section shows the test performance of layout analysing (Task A) models on digital (***D***), printed (***P***), and handwritten (***H***) sets. From Table 3.4, it can be observed that the mAP of Mask-RCNNs can achieve around 2%, 6% and 10% higher than Faster-RCNNs with identical backbones on ***D***, ***P*** and ***H***, respectively. This may result from the finer spatial localisation of auxiliary instance segmentation tasks adopted by Mask-RCNNs. Then, the effects of various backbones are explored, like Faster-RCNN with ResNet-101 (F-101) or Mask-RCNN with ResNet-50 (M-50). From Table 3.4, it could be observed that F-101 and M-101 consistently achieve around 1% to 2% higher than F-50 and M-50. It demonstrates that the deeper or

large-scale pretrained backbones may generate more comprehensive visual representations. Notably, the overall performance of ***D*** is around 5% higher than ***P*** and even about 11% more than ***H***. The main reason may result from the apparent difference between scanned (both printed ***P*** and handwritten ***H***) and digital (***D***) forms, especially the ***H*** involving more user-uncertainties such as writing mistakes or scanning rotation.

As our Form-NLU provides fine-grained layout component types such as *Value* subdivided into *Form_value* and *Table_value*, it enables us to observe and analyse the performance of subdivided components affected by distinct layout position distribution. For example, certain layout components show stable performance among the three test sets, such as *Form_key* and *Table_key*. It may result from those components designed by form designers with more shared layout and visual patterns like font and layout arrangement. However, *Form_value* and *Table_value* components may involve more uncertainties and noise from users, electric devices or propagation process, leading to mAP on ***H*** (46.82% and 52.81%) being lower than ***P*** (61.80% and 70.19%) and much lower than ***D*** (72.38% and 74.38%). In addition, for *Title* and *Others* components, the performance of M-101 on ***D*** (79.54% and 69.33%) is apparently higher than ***P*** (63.05% and 62.10%) and ***H*** (64.81% and 59.32), while ***P*** and ***H*** almost have similar performance. The reason may come from the difference between scanned (***P*** and ***H***) and digital forms, such as the rotation or lower resolution of scanned forms.

(2) Stepped Training Set Ratios

I set stepped ratios of training set size ($\mathcal{T}$) (10%, 50% and 100%) to train M-101 and evaluate it on ***D***, ***P***, and ***H*** sets to explore the effects of training size on Task A. It can be seen from Figure 3.8 as $\mathcal{T}$ size increases, overall and breakdown performance improve, whereas this trend becomes inapparent when $\mathcal{T}$ reaches 50% on ***D*** and ***P***. Regarding performance on ***H***, increasing the training size could boost the performance of all types of layout components, significantly for *Table_value*, *Section* and *Others*. Especially, as one significant layout component type for downstream tasks, *Table_value* hold around 30% on ***H***. In general, the trends shown in Figure 3.8 demonstrate that our training set ($\mathcal{T}$) contains various form types to enable handling distinct application scenarios, even with the small ratio of $\mathcal{T}$. Moreover, it also represents increasing the size of digital training sets ($\mathcal{T}$) that could effectively improve



(A) Digital(***D***) (B) Printed(***P***) (C) Handwritten(***H***)

FIGURE 3.8. Different training ratio mAP on digital (***D***), printed (***P***), and handwritten (***H***) test sets for task A.

printed and handwritten performance. More training samples may enable catching more shared feature patterns, such as spatial distributions, typographical similarity, font type and size of specific components.

3.7.2 Task B: Key Information Extraction

(1) Overall and Multi-aspect Feature Performance

This section compares and analyses the performance between our model and three widely used baselines under different configurations on digital (***D***), printed (***P***) and handwritten (***H***) test sets. Firstly, it focuses on comparing the performance among vanilla models (the first row of each model group in Table 3.5) in which input features and model architectures are the same as their paper described. Our model achieves better performance than all vanilla baselines on ***D***, ***P*** and ***H***. It indicates that iterative learning with token-level features could enhance entity-level representations to boost downstream performance. In addition, for three vanilla baselines, LXMERT can get 2.28%, 6.38% and 15.36% higher than VisualBERT on ***D***, ***P*** and ***H***, respectively, which may result from inputs of LXMERT containing positional information, but VisualBERT is pre-trained on the visual feature of input RoIs only. Subsequently, compared with VisualBERT, the non-pretrained M4C can increase by around 1.5%, 3% and 9% on ***D***, ***P*** and ***H***, respectively. It may illustrate the significance of textual and positional features. However, due to the input feature differences between vanilla models, external experiments are conducted to explore the effects of multi-aspect features on all adopted models for a fair comparison.

Model	Input Features					Overall (F1-score)		
	V	T	P	D	G	Digital(***D***)	Printed(***P***)	Handwritten(***H***)
M4C	◯	◯	◯	×	×	96.91	88.62	74.06
	◯	◯	◯	◯	×	96.32	89.52	74.98
	◯	◯	◯	×	◯	97.00	88.81	75.04
	◯	◯	◯	◯	◯	97.22	89.78	74.89
VisualBERT	◯	×	×	×	×	95.55	85.91	65.25
	◯	×	◯	×	×	95.84	85.93	67.25
	◯	◯	◯	×	×	96.07	85.90	70.14
	◯	◯	◯	×	◯	96.61	87.43	70.79
	◯	◯	◯	◯	×	96.70	87.00	71.28
	◯	◯	◯	◯	◯	96.73	87.18	72.65
LXMERT	◯	×	◯	×	×	97.83	92.29	80.51
	◯	◯	◯	×	×	97.67	94.15	82.80
	◯	◯	◯	◯	×	97.70	94.59	82.60
	◯	◯	◯	×	◯	97.74	94.49	83.71
	◯	◯	◯	◯	◯	97.74	95.07	84.43
Our Model	◯	◯	◯	×	×	99.06	95.07	84.56
	◯	◯	◯	◯	×	**99.30**	95.32	84.75
	◯	◯	◯	×	◯	99.12	95.50	85.77
	◯	◯	◯	◯	◯	99.09	**95.50**	**86.98**

TABLE 3.5. Results of vanilla model (first row of each model group), model with all aspect features (last row of each) on ***D***, ***P***, and ***H*** test sets. RoI inputs contain visual (V), textual (T), positional (P), text density (D), gap distance (G) features, where ◯ and × represent using or not using the specific feature in that column.

M4C initially contains V, T, and P of input RoIs, and is a non-pretrained model with random initial weights. Thus, it may cause only slight improvements can be found in Table 3.5 after adding D and G into the model on three test sets. For VisualBERT, the vanilla model only contains RoIs visual features; after stacking more aspect features, the performance is gradually improved where the F1-score on ***H*** increases from 65.25% (vanilla) to 72.65% (with V, T, S, D, G). There is an apparent increase after adding D and G into VisualBERT, which may contribute to the additional features making the input representations more comprehensive. A similar trend also can be found in LXMERTs where T may contribute to a more positive effect (about 2% increasing) than VisualBERT on ***P*** and ***H***. Regarding our model, although there is no noticeable improvement observed on ***D***, the positive effects on ***P*** and ***H*** can be found, especially for ***H*** (from 84.56 % to 86.98%). Generally, the proposed additional features could help the model better understand form structures to improve the

PE Approach	Digital(***D***)	Printed(***P***)	Handwritten(***H***)
Without Positional Encoding	98.83	94.34	85.98
Linear Projection	99.09	95.50	86.98
XY-Positional Encoding	**99.22**	**96.14**	**89.13**

TABLE 3.6. Performance of various positional encodings (PE).

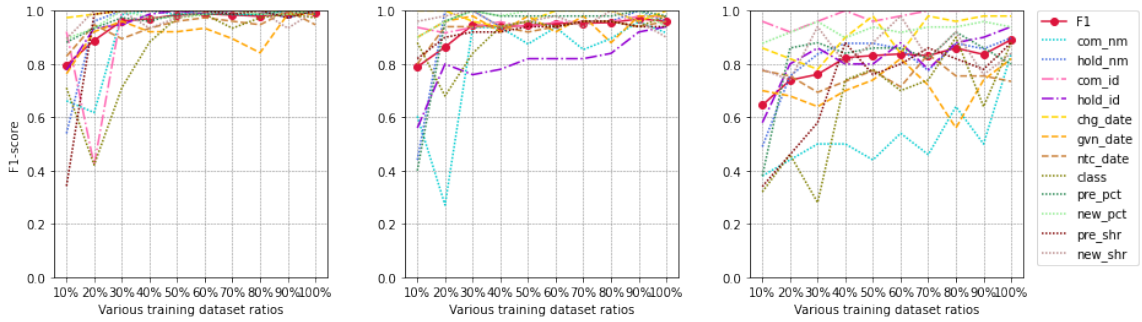
model generalisation ability, notably when the appearance of input forms differs during the training and inference stage.

(2)Positional Encoding Validation

To demonstrate the effectiveness of XY-pos, external experiments are performed to compare the performance of the model with and without linear projection PE methods. Compared to the model without any PE in Table 3.6, linear projection and XY-pos can have an apparent increase in all three test sets. It illustrates that positional information is significant in understanding the layout structure for form understanding. Furthermore, from Table 6, we can find XY-pos reach a better performance than linear projection on ***D***, ***P*** and ***H***. Significantly, the F1 increased from 86.98% to 89.13% on ***H***. It demonstrates that XY-Pos may capture more positional information to better understand various layout structures of input forms.

(3) Fine-grained Training Set Ratio

For exploring the training size influences for form understanding, fine-grained training set ratios are defined (from 10% to 100%) to train the model and represent the evaluation performance on ***D***, ***P*** and ***H*** in Figure 3.9. With increasing training size, fluctuating increases can be observed for overall and breakdown performance on three test sets. The rapid increases can be observed on ***D*** and ***P*** before reaching 50% $\mathcal{T}$, following the stable trends after reaching this critical ratio. However, for the handwritten set ***H***, an apparent overall performance increase can be found during the entire training size interval. It may demonstrate the wide variety of our training dataset ***H*** can result in a more generic trained model to extract specific key information more effectively. Additionally, unlike the datasets providing general key-value annotations, Form-NLU can show the training-size sensitivity of specific key-value pairs to analyse the robustness of the trained model with limited training data. For example,



(A) Digital (***D***) (B) Printed (***P***) (C) Handwritten (***H***)

FIGURE 3.9. Performance of our model with fine-grained training set ratios on three test sets.

Model	Acc	Breakdowns (Accuracy)											
		nm(name)		id		date			class	pct(percent)		shr(share)	
		com	hold	com	hold	chg	ntc	gvn		pre	new	pre	new
LXMERT	43.78	86.30	80.14	69.18	52.74	72.60	65.07	54.79	4.79	6.85	24.66	5.48	2.74
VisualBERT	57.48	88.36	78.08	63.01	59.59	39.04	70.55	37.67	47.95	50.00	28.77	53.42	73.29
M4C	69.35	83.56	80.82	63.70	73.29	65.75	57.53	82.88	61.64	65.75	58.90	59.59	78.77
Ours	72.72	91.78	81.51	71.92	71.23	80.82	80.82	84.93	71.92	70.55	80.14	50.00	57.53

TABLE 3.7. Overall performance and breakdown results of key information extraction task with PDFminer.

com_nm, *com_id*, and *hold_nm* show more sensitive to training size, while date-related keys such as *ntc_date* are less sensitive.

3.7.3 Task B with PDF Parser

Real-world users, especially non-deep learning users, are challenged to get input RoIs through well pretrained layout analysing models because of lacking ground truth annotations and relevant background knowledge. Thus, using textlines extracted by specific PDF parsing tools is an alternative way to obtain the input of adopted well-trained models. PDFminer (a widely used PDF parser [12]) is used to extract RoIs for replacing the ground truth RoIs in a digital set (***D***) and feed them into three trained baseline models and our models. I set IoU = 0.5 as a threshold to calculate accuracy, as Table 3.7 shows.

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

Different from the trend of feeding ground truth RoIs during the testing stage, the non-pretrained model M4C can achieve much better accuracy (69.35%) compared with pre-trained LXMERT (35.24%) and VisualBERT (53.34%). It might be due to the noisy RoIs detected by PDFminer being fed into the large-scale pretrained models, which decreases those heavy models' RoIs feature representation ability. Our model can achieve the highest overall performance among all tested models. It demonstrates that the proposed XY-pos enhanced dual-level model can improve generality to understand the input contents better. Regarding breakdown results, *com_nm* and *hold_nm* could always perform better. Most of the key information extracted by our model could hold higher accuracy than other baselines, such as *com_id* and *chg_date*. Notably, LXMERT shows high sensitivity to the bounding box precision of input RoIs. However, because PDFminer extracted inaccurate bounding boxes, especially in the table area, the keys located in the table are much lower than other models, such as *class*, *pre_pct*, *pre_shr*, etc.

3.8 Case Study: Transfer Learning

In order to check the feasibility of our proposed data and model, a case study is conducted with transfer learning. Note that transfer learning aims to store knowledge gained while solving our dataset and apply pre-trained models to a different but related problem. In this case study, we applied the best (our model) and second-best models (LXMERT), trained by the Form-NLU dataset, to the publicly available benchmark FUNSD. We use the selected key (question)-value (answer) pairs from the FUNSD [49] dataset[13]. Based on the testing results, even if the nature of FUNSD is entirely different from our Form-NLU dataset, the models trained on our dataset can still achieve sound performance. The result shows that LXMERT (the second best of in Task B) correctly predicted 80 out of the 150 (53.33%) FUNSD key(question)-value(answer) pairs while our model achieved **94 (62.67%)** correct predictions. It demonstrates that our Form-NLU dataset can learn the general form layout

[13]Please refer to https://github.com/adlnlp/form_nlu/blob/main/README.md#case-study-setup to check the setup detail and samples

and be applied in other benchmarks; in addition to this, our proposed model is efficient in extracting a feature representation for form understanding.

3.9 Conclusion

This paper proposes Form-NLU, a new form structure understanding and key information extraction dataset. The proposed dataset covers the important point of view, enabling the interpretation of the form designer's specific intention and the alignment of user-written value on it. The dataset includes three form types: digital, printed, and handwritten, which cover diverse form appearances/layouts and deal with their noises. Moreover, a novel model is proposed for form structure understanding and key-value information extraction, which applies robust positional and logical relations. Our model outperformed all state-of-the-art models in key-value information extraction tasks. I do hope that our proposed dataset and model can provide great insight into form structure and information analysis. Hence, I adopted an off-the-shelf PDF layout extraction tool and also provided its feasibility by conducting transfer learning.

CHAPTER 4

Multi-page Visually Rich Document Question Answering

This Chapter is the work ***MMVQA: A Comprehensive Dataset for Investigating Multipage Multimodal Information Retrieval in PDF-based Visual Question Answering*** [28] accepted at **IJCAI 2024**. I formulated the research aim, designed the dataset structure and construction, analysed and visualised the dataset, designed the methodology and implementation, and wrote the whole paper.

This work progresses from constructing a multimodal, multi-page, retrieval-based document visual question-answering dataset to introducing a series of frameworks designed to address challenges in understanding multi-page documents.

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. The new VRD-QA frameworks are proposed to grasp textual contents and relations among document layouts simultaneously, extending page-level understanding to the entire multi-page document. The aim is to enhance the capabilities of existing vision-and-language models in handling challenges posed by text-dominant documents in VRD-QA.

4.1 Introduction

4.1.1 Background

The growing demands for visually rich document (VRD) question-answering (QA) areas are becoming increasingly evident, especially in specialised fields such as finance, medicine, and industry. VRDs, which include forms [26], academic papers [27], and industrial reports [84], typically comprise text-dense and visually rich components such as *titles*, *paragraphs*, *tables*, and *charts*. These components, referred to as ***document semantic entities***, are not only knowledge-intensive but are also organised in a pre-defined 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. Therefore, in VRDs, it is essential to comprehend the layout and logical structure, especially in multi-page documents, to accurately locate and use these document entities as reliable evidence for answering knowledge-intensive questions.

Generative models [72, 89, 120] have made impressive progress in providing interactive and human-like responses to user queries by memorising vast knowledge [152]. The trend is inevitable in the domain of VRD understanding. These large generative models rely on plain text to learn textual content [120] and use image patches to encode visual cues [143]. This approach makes understanding document entities' layout and logical relationships in VRDs difficult. Additionally, generative models are suffered from hallucinations [144], high costs [41], and updating knowledge difficulties[44]. Retrieval-based QA [74] is introduced to address 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 with less cost. While recent retrieval-based applications mainly focus on web-crowded domains like Wikipedia and web images [44], VRD-QA requires a deep understanding of domain-specific multimodal knowledge.

A few VRD-QA datasets [84, 116] 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 [119], potentially causing recently proposed models [45, 147] to excel at retrieving annotated in-line text but disregarding the logical and layout connections among document entities. Moreover, they may be limited in handling the entire lengthy document. To address these limitations, entity-level document understanding tasks have been introduced by [26] and [27]. However, a common issue with these datasets is their emphasis on text-dense mono-modal information extraction, overlooking visually rich entities such as *tables*, *figures*, and *charts*. This oversight is particularly crucial in fields such as minerals and finance, garnering increased attention.

4.1.2 Research Aims

This paper proposes a new multi-page, multimodal document entity retrieval dataset, MMVQA, that is pivotal for advancing information retrieval systems in knowledge-intensive applications. MMVQA addresses the limitations of generative models in answering knowledge-intensive questions. It 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 development of innovative models capable of navigating and interpreting real-world documents at a multi-page or entire document level, leveraging Joint-grained and multimodal information. The paper proposes a set of frameworks demonstrating how to effectively use existing VLPMs and pretrained language models with long sequence support to locate target entities from MMVQA.

4.1.3 Contributions

Contributions are summarised as follows: This paper introduces 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, leveraging the implicit knowledge from VLPMs and fine-grained level information to enhance model effectiveness and robustness. A series of experiments combining quantitative and qualitative analyses are performed to provide deeper insights

into MMVQA and demonstrate the effectiveness of our proposed techniques for multimodal multi-page document entity retrieval.

4.2 Related Work

4.2.1 Document Visual Question Answering Datasets

Dataset Name	Multi-Page	Source	# Docs	# Images	QA num	Answer Types	Answer Categories	Question Generation
DocVQA	✗	Industry Documents	N/A	12,767	50,000	Text	Ext.	Human
DocCVQA	✓	Financial Reports	N/A	14,362	20	Text	Ext.	Human
RDVQA	✗	Promo Exceptions	N/A	8,362	8,514	Text	Ext.	Online Dialog
CS-DVQA	✗	Industry Documents	N/A	600	1,000	Text	Abs.	Human
SlideVQA	✓	Slides	2,619	52,480	14,484	Text	Ext. + Abs.	Human
VisualMRC	✗	Webpage	N/A	10,197	30,562	Text	Abs.	Human
TAT-DQA	✓	Financial Reports	2,758	3,067	16,558	Text	Ext. + Abs.	Human
PDFVQA	✓	Journal Article	1,147	12,000	84K/54K/5.7K	Text/Text Dense Entity	fixed answer + Ext.	Template
MP-DOCVQA	✓	Industry Documents	6,000	48,000	46,000	Text	Ext.	Human
OD-DocVQA	✗	Webpage	N/A	158,000	15,000	Text	Ext.	User-log
MMVQA	✓	Journal Article	3,146	30,239	262,928	**Multimodal Entities** ***(E.g. Table, Paragraph, Figure)***	Ext.	Human & ChatGPT

TABLE 4.1. Visually rich document (VRD) or similar domain question answering datasets.

The first proposed question was answered over document images [84] as the DocVQA dataset. The scanned documents in the dataset are industry 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 [31] 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 [137], 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 [118] and SlideVQA [115] 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 to answer 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 [116] 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 multimodal VRD-QA dataset that retrieves target document entities across multiple pages.

4.2.2 Document Understanding in VRD-QA

Recently, there has been a surge in the use of VLPMs pretrained on the general domain for document understanding [26, 27]. Models like [17, 96, 114] have shown impressive performance on single-page document understanding tasks by leveraging implicit knowledge from the general domain. Recently, various document understanding models have emerged, employing diverse pretraining techniques to capture correlations between different modalities like text, visuals, and layouts. Unfortunately, most of these models [45, 139, 147] rely solely on token sequences obtained through OCR, ignoring the structural relationships between document entities and facing computational challenges. Some models, such as UDoc [33], highlight the importance of entity information for document understanding, but they tend to focus on single-page scenarios, not reflecting real-world complexities. Furthermore, current cross-page understanding models have limitations in tasks like document parsing [81, 101], as they do not extend to VRD-QA. In our paper, we introduce multimodal cross-page document understanding frameworks. These frameworks not only leverage implicit knowledge from various VLPMs but also incorporate mechanisms to utilise fine-grained (textual tokens) and coarse-grained (document semantic entities) information. This approach enhances entity representations and improves correlation learning for comprehensive document understanding.

4.3 MMVQA

4.3.1 Dataset Collection

The documents are collected from PubMed Central[1], a biomedical and life science journal literature archive. Its Open Access Subset contains millions of full-text 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. Figure 4.1 shows the annual-basis distribution of collected document types, where 87.3% of collected documents are research articles. Most of these research articles were published in the last five years.

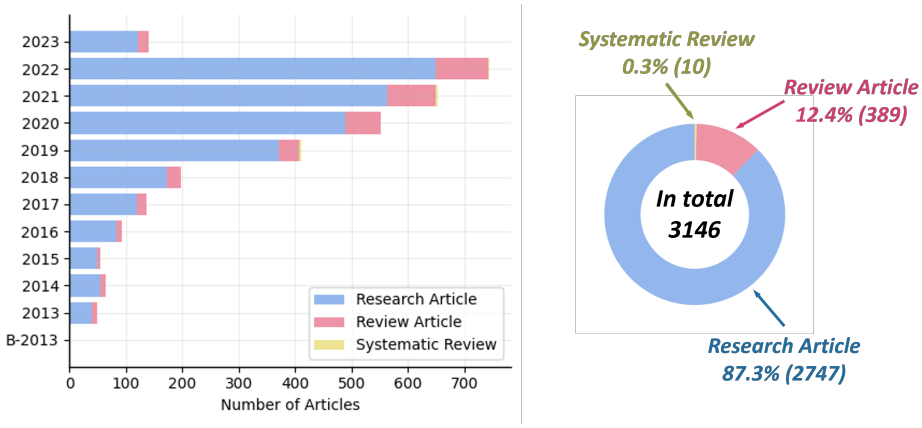


FIGURE 4.1. Document type distribution analysis of collected documents.

4.3.2 Dataset Preprocessing

The dataset includes both PDF images and segmented document components, which are categorised into predefined semantic categories such as *Title*, *Section*, *Paragraph*, *List*, *Figure*, *Table*, *Figure Caption*, and *Table Caption*. The segmented document components are defined as ***document semantic entities***, which contain associated text within its bounding box. PDFminer [2] is used, following [157], to extract the bounding box coordinates and text contents

[1] https://www.ncbi.nlm.nih.gov/pmc/
[2] https://pypi.org/project/pdfminer/

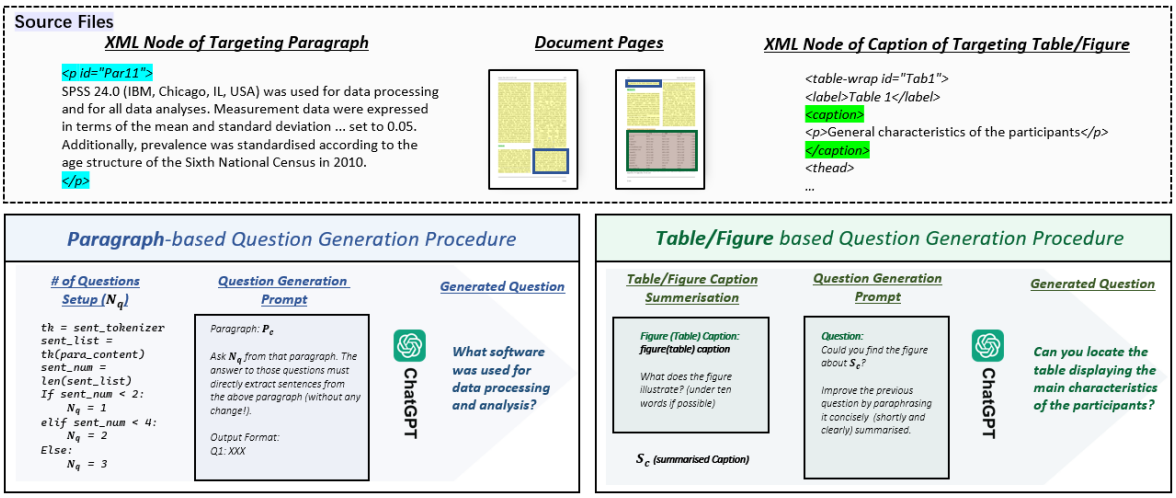


FIGURE 4.2. The sample question generation progress.

of each document page's textbox, textline, image, and geometric shapes. Then, 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.

4.3.3 Question Generation

A large number of diverse types of content-related questions are generated, associated with different multimodal document entities of journal articles. To do so, ChatGPT[3] is applied to automatically generate 1-3 questions based on the contents of each paragraph of these main sections. As shown in Figure 4.2, the number of sentences in the paragraph determines the number of questions (n_q) to be generated for paragraph-based questions. 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.

4.3.4 Dataset Format

The MMVQA dataset is divided into three sets: training, validation, and testing, with the statistics provided in Table 4.4. The randomly selected dataset samples are provided in

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

Table 4.2. Each set comprises a DataFrame (CSV file) with attributes such as "*question*", "*answer*", and "*document_id*". Additionally, extra annotations for "*context*" and "*page_range*" are included, presenting the text content and the covered page range (in the first-level/top-level section) of the answer for the question. We provide a detailed description of each column's attribute:

- *Question*: the processed LLM generative natural language question which is related to the content details about a paragraph (paragraph-based) or table/figure (table/figure-based).
- *Document_ID*: the document_id to link the question to the corresponding document for acquiring semantic entity information (Please refer to the document pickle file for more information).
- *Answer_objt_id*: the target document entity ID ordered by reading order function defined by [25].
- *super_section*: the Super-section type of the target answer entity located in the first-level section for conducting Super-Section-based breakdown analysis.
- *ID*: a unique ID for each question sample.
- *Page_range*: the target entity in the first-level section covered pages.
- *Context*: the first level section content only applies to paragraph-based questions.

For each set, the metadata information is provided (in an additional JSON file). It contains annotated features, including document entity *bounding box*, *text content*, *category*, etc, which are essential for model implementation.

- *bbox*: bounding box coordinates following COCO format (x1,y1,w,h), where x1 and y1 are the left corner locations; w and h are the width and height of the box.
- *text*: the text content of each document semantic entity.
- *object_id*: unique object ID to locate an object.
- *category*: semantic type of document entities such as "*Paragraph*", "*Table*", "*Figure*", etc.

Question	Document_ID	Answer_objt_id	Super-Section	ID	Page_range	Context
What mutant was constructed with an amino acid substitution inactivating EndoU during MERS-CoV infection?	PMC9173776	[16]	results and discussion	9022	(1, 2)	In order to study the effects of EndoU activity on the dsRNA-induced antiviral innate...
What was the role of pyridine-2-aldoxime in the activation process?	PMC8697031	[32]	conclusions	20635	(3, 4)	In summary, we have first noticed the ortho C–H bond activation in a coordinated 7,8-benzoquinoline in...
What is the survivorship rate of Total Hip Arthroplasty at 25-year follow-up?	PMC8987314	[7, 8]	introduction	21045	(0, 1)	Total hip arthroplasty (THA) is a highly successful operation with greater than 85% ...
Can you locate the table comparing CMV characteristics in patients treated with ICI drugs?	PMC9399572	[64]	table	36776	(2, 6)	N/A
Can you locate the graphic that depicts the connection between GERD and atrophic gastritis?	PMC8900577	[68]	figure	37893	(4, 9)	N/A

TABLE 4.2. Randomly selected samples from MMVQA dataset. For table/Figure-based questions, no context information is provided.

4.3.5 Human Evaluation

The questions are filtered by pre-defined rules to ensure quality and evaluated by raters. Three raters are invited to evaluate 20 randomly selected questions from three aspects, including Relevance, Meaningfulness and Correctness. A human evaluation survey example can be found in Figure 4.3. **Relevance**, which examines if the question aligns with and is applicable to the provided PDF file context; **Meaningfulness**, assessing if the question is practical and likely to be asked in a real-world scenario; and **Correctness**, determining whether the provided answer accurately and comprehensively addresses the question based on the context. Each criterion is further broken down into three levels: (1) **Relevance**: *Valid & Relevant*, *Moderately Valid & Relevant*,and *Invalid or Irrelevant*; (2) **Meaningfulness**: *Highly Meaningful*, *Moderately Meaningful*, and *Lacks Meaningfulness*; and (3) **Correctness**: *Comprehensive & Accurate*, *Partial or Incomplete*, and *Inadequate or Incorrect*.

The percentage agreement measures the percentage of all raters agreeing on a specific option. Percentage Agreement for Option $P_k = \left(\frac{\text{Number of items where all raters chose option } k}{\text{Total number of items}}\right) \times 100$. As shown by Table 4.3, the high scores in *Relevance*, *Meaningfulness*, and *Correctness* indicate a consistent recognition of quality across different evaluation criteria by the raters. The strong positive rates, particularly in 'Correctness' at 72.73%, suggest that the majority of the raters not only found the content acceptable but highly appropriate and accurate according to the standards set for evaluation.

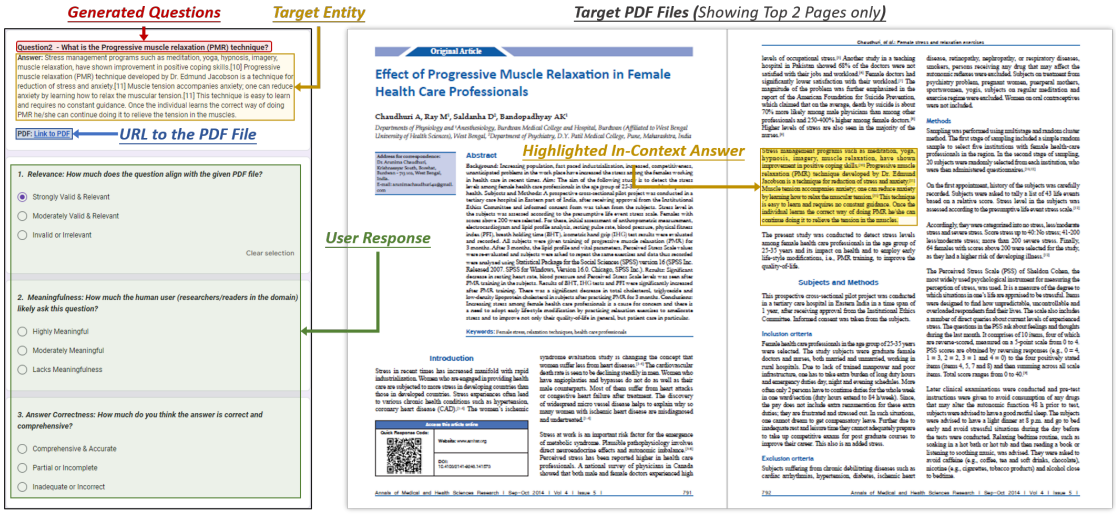


FIGURE 4.3. A sample of human evaluation survey.

	Positive Rate	Strong Positive Rate	Negative Rate
Relevance	90.91	59.9	9.09
Meaningfulness	100	38.1	0
Correctness	100	72.73	0

TABLE 4.3. Rater agreement on evaluation metrics for generated questions.

4.4 Dataset Analysis

4.4.1 Document Analysis

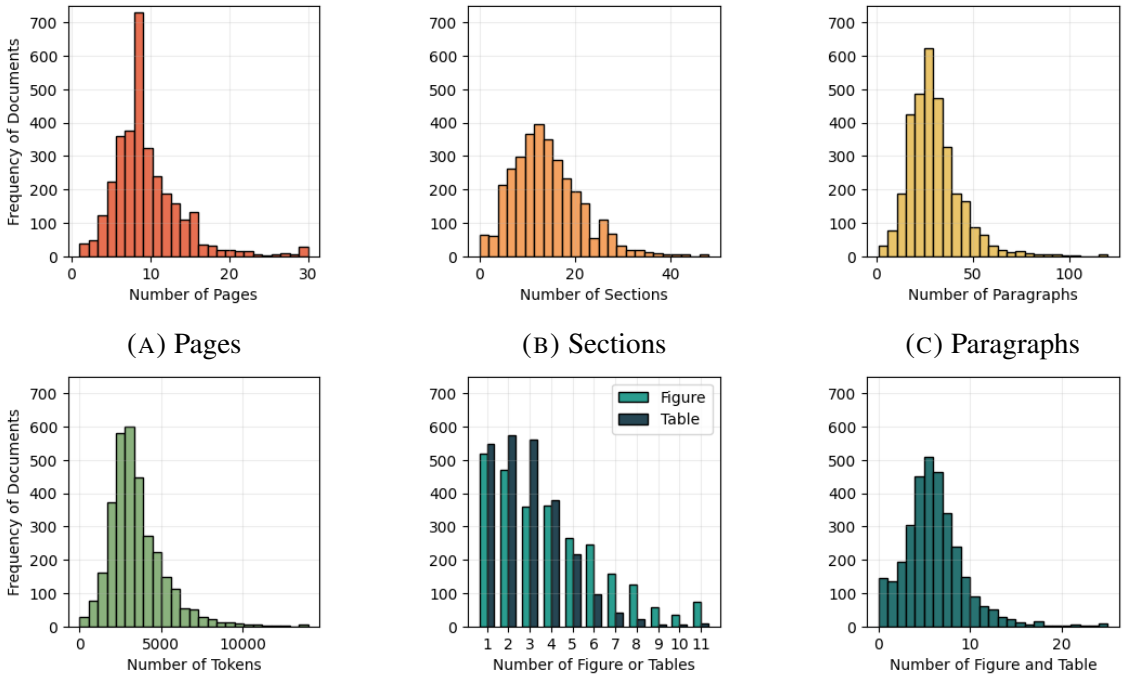
Document Collection Topic Analysis To analyse the key topics of the collected documents in our proposed dataset, the frequent words are represented from the *Title*, *Keywords* and *Abstract*. As expected, Figure 4.4 shows the word clouds analysis result that these documents are highly related to medical topics with keywords, such as '*covid-19*' and '*treatment*', repeatedly appearing. Note our datasets are collected via PubMed Central.

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 represented by Figure 4.5, it can be found the number of document components is quite consistent. Most documents contain

(A) Article Title (B) Keywords (C) Abstract

FIGURE 4.4. Word cloud of key phrases and topics from document *Titles*, *Keywords*, and *Abstracts*.

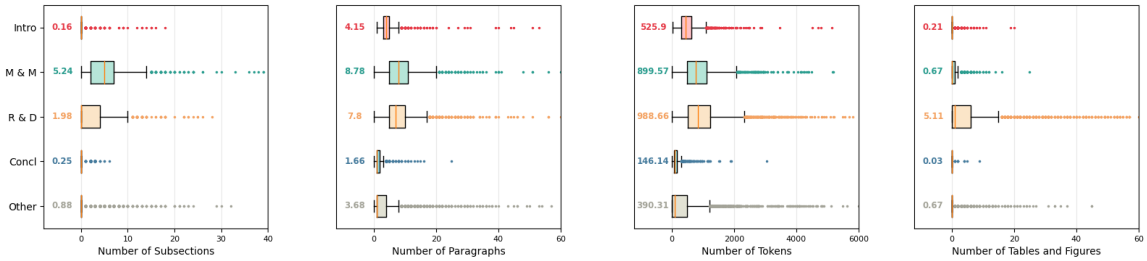
around ten pages and have 10-20 different sections with around 20-40 paragraphs of 2000-4000 tokens. Hence, the analysis can ensure that the collected documents are mostly lengthy and have a complex structure enough to evaluate the model's feasibility to contextualise understanding over multiple consecutive 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.



(A) Pages (B) Sections (C) Paragraphs

(D) Tokens (E) Tables/Figures (F) Tables & Figures

FIGURE 4.5. Distribution of various document component types



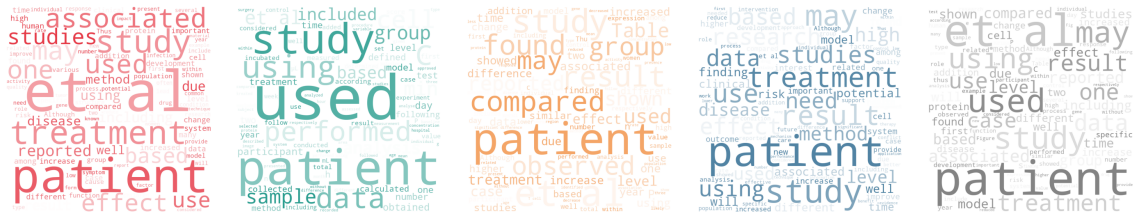
(a) Subsection (b) Paragraph (c) Token (d) Table&Figure

FIGURE 4.6. Distribution of various document components and semantic entities of each Super-Section type.

Super-Section Component Analysis The first-level sections of each document are defined 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*. 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 detailed in Figure 4.6. 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 and shorter content. The *Other* Super-Section, encompassing diverse contents like *Supplementary* or *Fundings*, exhibits a larger interquartile range and more outliers, reflecting its varied nature.

Supersection-based Topic Analysis The frequent words based on the content of each Super-Section are represented, as shown in Figure 4.7. *Introduction* normally focuses on the purpose of the study, *M&M* focuses on the analysis and method of the study, *R&D* focuses on the effect and difference of the study, and *Conclusion* focuses on the risk and effect. These frequent words in each super-section also imply the possible answers to the questions normally asked about each super-section in the dataset.

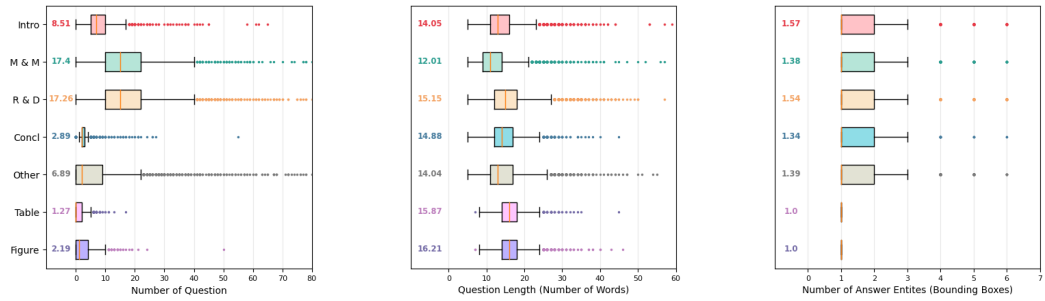


(A) Intro (B) M&M (C) R&D (D) Concl (E) Other

FIGURE 4.7. Topic keyword analysis for each Super-Section types.

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 4.4. Dataset distribution across different splits with question count by Super-Section category.



(a) Question Number (b) Question Length (c) Entity Number

FIGURE 4.8. Question pattern analysis of each Super-Section type.

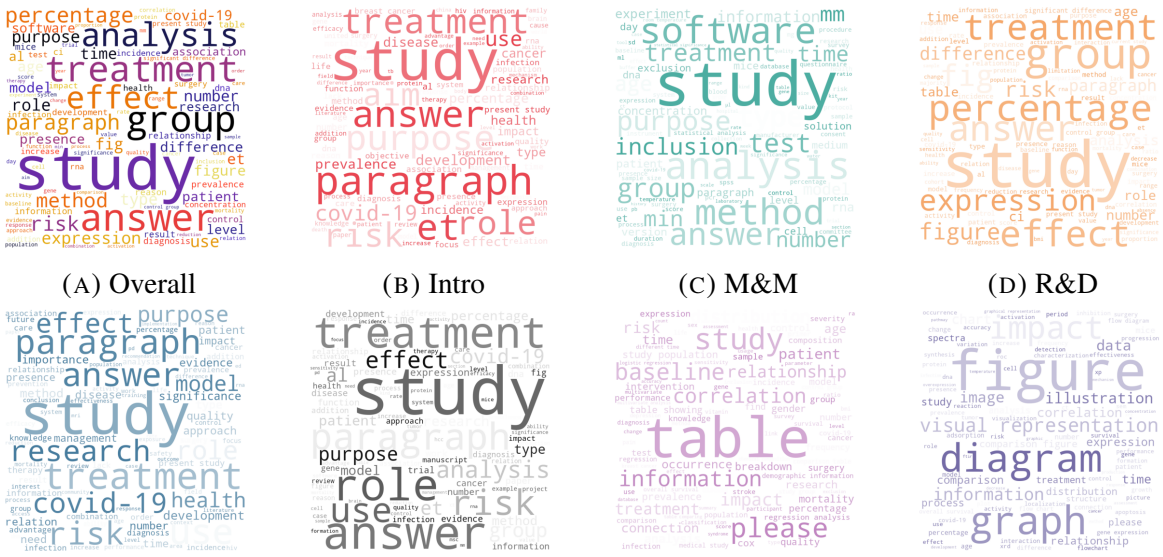
4.4.2 Question-based Analysis

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 4.4.

Super-Section-oriented Question-Answer Distribution The distribution of questions over each Super-Section in documents is shown in Figure 4.8a. Most of the questions are asked over *M&M* and *R&D* sections, each having an average of around 17 questions. The average

question length is shown in Figure 4.8b. *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). In contrast, for other Super-Section questions, answers may located in more than one document entity.

Super-Section-based Question Content Analysis The Wordcloud of the questions are visualised regarding the whole dataset and each Super-Section, respectively, in Figure 4.9. Most question words of each Super-Section are correspondingly related to that Super-Section's contents. For example, questions for the *M&M* section frequently contain the word '*method*', and table/figure-related questions have question words such as "*table*", "*figure*", "*diagram*", "*graph*".



(A) Overall (B) Intro (C) M&M (D) R&D

(E) Concl (F) Other (G) Table (H) Figure

FIGURE 4.9. Visualised wordcloud of the entire dataset and section-specific question analysis.

4.5 Task Definitions and Metrics

The main task of MMVQA is ***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 [27, 33], 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 4.10, 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})[4]. Additionally, to address diverse application scenarios and effectively meet specific requirements, a set of distinct evaluation metrics are introduced for more adaptive performance assessment, including ***Exact Matching*** (*EM*), ***Partial Matching*** (*PM*), and ***Multi-Label Recall*** (*MR*). More details can be articulated in Section 4.5.2.

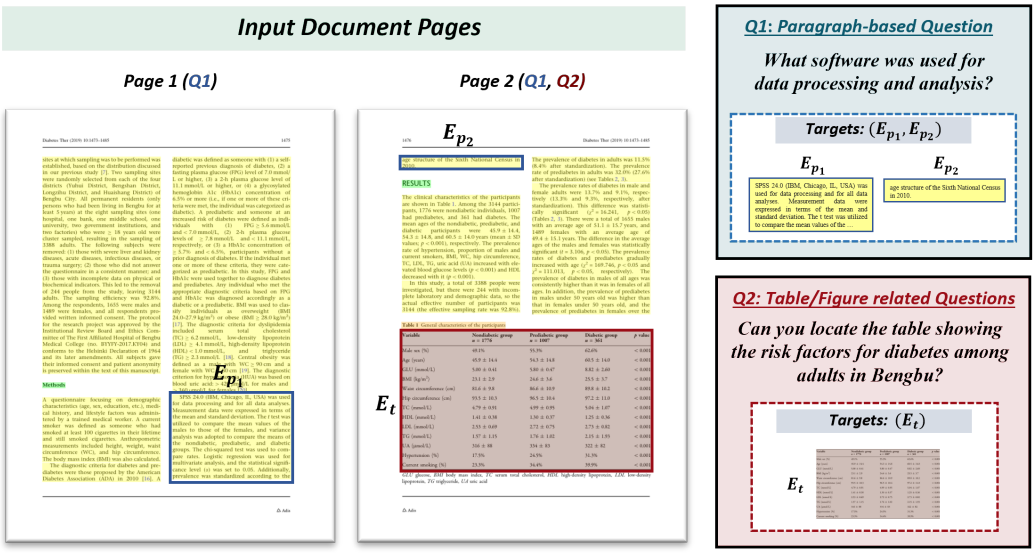


FIGURE 4.10. Defining tasks of multimodal cross-page information retrieval with illustrative examples.

4.5.1 Task Definition

This section outlines how our multimodal DIR task is 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

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

more than one entity (as in Figure 4.10). 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 4.10, 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 4.10, the ground truth set only contains the table entity E_t.

4.5.2 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. **Partial Matching Accuracy***(PM)* is introduced 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. The detailed definitions of each metric are defined as follows.

- ***Exactly Matching Accuracy (EM)***: denoted as EM, is defined as $\frac{N^{EM}_{\Phi pos}}{N_\Phi}$, where N_Φ is the total number of instances in a data split Φ, which could be the entire test set or a subset (e.g., QA pairs in the *Introduction*). $N^{EM}_{\Phi pos}$ represents the number of positive samples in Φ. An instance is considered positive only if the predicted entity set $S_{E_{pre}}$ is the same as the ground truth $S_{E_{gt}}$; otherwise, it is treated as negative. For instance, the predicted entity set $S_{E_{Q1_{pre}}} = E_{p_1}$ is treated as a negative sample for Q_1 in Figure 4.10. *EM* is a stringent metric suitable for scenarios requiring precise, unambiguous information retrieval, particularly when used as supporting evidence or reliable references.
- ***Partial Matching Accuracy (PM)***: is similar to exact matching, defined as $PM = \frac{N^{PM}_{\Phi pos}}{N_\Phi}$. The key distinction lies in its tolerance for partial matches. A prediction is

considered positive if it meets $S_{E_{pre}} \in S_{E_{gt}}$ and $S_{E_{Q1_{pre}}} \neq \emptyset$. Thus, $S_{E_{Q1_{pre}}} = \{E_{p_1}\}$ is a negative case for *EM* but positive for *PM* for $Q1$ in Figure 4.10. This metric 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)***: denoted as R, serves as a crucial metric in multi-label classification, defined as $R_\Phi = \frac{TP_\Phi}{TP_\Phi + FN_\Phi}$, where TP_Φ and FN_Φ represent the number of true positive and false negative instances of set Φ. For example, the *MR* of the predicted $S^{pre}_{EQ1} = \{E_{p_1}\}$ of $Q1$ in Figure 4.10 should be $R_{Q1} = 0.5$, where $TP_{Q1} = 1, FN_{Q1} = 1$. This metric aims to assess the proportion of correctly identified actual positives in situations where identifying all positive instances is critical.

4.6 Baseline Framework Setup

4.6.1 RoI-based Model

A detailed description of how to set up a RoI-based framework for multipage document information retrieval is provided.

- ***Vanilla Transformer***: directly uses the initial visual embedding $\mathbb{V}$ with textual embedding $\mathbb{T}$ to generate the entity embedding $\mathbb{E}$, where $\mathbb{E} = \mathcal{L}_{vl}(\mathbb{V} \oplus \mathbb{T})$. The sequence of question tokens Q are encoded by *bert-base-cased*. The combined embedding $[\mathbb{E}+\mathbb{P}+\mathbb{B}+\mathbb{L}, Q]$ is fed into the retriever $\mathcal{R}$. This model is a foundational benchmark for evaluating the impact of various pretrained techniques in comparison studies.
- ***VisualBERT*** [66]: is a pretrained transformer-based VLPM to learn the contextual relationship between RoI visual cues and plain text on the general domain. For applying VisualBERT in this dataset, the question token sequence and entity visual embedding $\mathbb{V}$ are fed into a pretrained VisualBERT backbone to get Q and $\mathbb{V}'$. The

entity representation is then derived as $\mathbb{E} = \mathcal{L}_{vl}(\mathbb{V}' \oplus \mathbb{T})$ for downstream procedures. VisualBERT will be applied to demonstrate whether the pretrained general domain backbone can generate a more robust question and entity visual aspect representation.

- ***LXMERT*** [114]: layout information aware VLPM utilising the bounding box co-ordinates of RoI to enhance the cross-modality representations. The only difference from applying VisualBERT is the inputs of the LXMERT backbone contain linear projected bounding box coordinates. LXMERT can help determine the impact of layout-aware processing in visually rich document understanding tasks.

4.6.2 Patch-based Model

The detailed model setup for the Patch-based model is also provided as follows:

- ***CLIP*** [96]: jointly trains the embedding acquired from vision ($\mathcal{E}_{clip_v}$) and textual ($\mathcal{E}_{clip_t}$) encoders by predicting the correct text-image pairs. Question text Q and the sequence of image patches are fed into $\mathcal{E}_{clip_t}$ and $\mathcal{E}_{clip_v}$, respectively. The jointly learnt question and visual embedding acquired from $\mathcal{E}_{clip_{joint}}$ are fed into $\mathcal{R}$ with entity embedding $\mathbb{E}$.
- ***ViLT*** [55]: is the first pertained vision-language model using a shared convolutional-free multimodal encoder to learn cross-modal interaction with improved memory and time efficiency. The token and image patch sequences are fed into a cross-modal encoder $\mathcal{E}_{vilt}$ to get question token embedding Q and patch embedding P.
- ***BridgeTower*** [138]: use multiple bridge layers to learn the relation between the top layers of uni-modality encoders. The procedures for employing BridgeTower are closely with the CLIP framework. However, unlike CLIP, which provides another joint modality encoder, BridgeTower mainly focuses on learning contextual relations on the top layers.

4.7 Implementation Detail

Input Representation: The BERT [CLS] token (768-d) and ResNET R5 Layer (2048-d) are extracted as the initial textual and visual embedding of each RoI. For positional embedding, I follow [26] to linearly project a 4-D bounding box into a 768-D vector of each RoI. Similar procedures are conducted on label embedding by linearly projecting the one-hot embedding into a 768-d vector.

Model Configuration: In Multimodal multi-page Retriever $\mathcal{R}$, the Multimodal Entity Encoder $\mathcal{E}$ and Decoder $\mathcal{D}$ have six encoder layers with eight heads and 768-d hidden states. The number of maximum document entities is 200, and an extra token is placed to deal with documents exceeding 200 entities. Additionally, the maximum input length of questions is set as 100. For the model default text input length of less than 100, I adopted their default setup (e.g. ViLT up to 40 tokens, CLIP up to 76 tokens). A pretrained Bigbird checkpoint is used for the Joint-grained Retriever, and the maximum input length is 2048.

Training Setup: All models use *CrossEntropy* as the loss function, with *Adam* optimiser and 2*e-5 as the learning rate. The maximum training epoch is set to 15, and the best-performed model is saved based on Validation sets performance. The values of batch size are 32, 16 and 8 for RoI-based, Patch-based and Joint-grained models, respectively.

Environmental Setup: All experiments are conducted on a 48 GB *NVIDIA RTX A6000* GPU with *CUDA 11.8*.

4.8 Methodology

4.8.1 Multimodal Multi-page Retriever

Existing document understanding models [45, 54, 127] and datasets [84, 116] 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

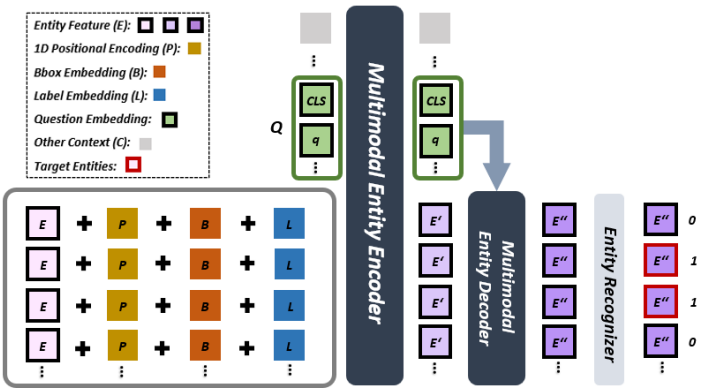


FIGURE 4.11. Multimodal multi-page retriever framework.

in more prevalent multi-page scenarios. Instead of employing sequences of tokens that lead to significant memory consumption, a multimodal entity-level retrieval framework $\mathcal{R}$ is introduced to identify the target entity set S_Q from the cross-page entity sequence in a given question Q, as illustrated in Figure 4.11.

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 4.8.2, are combined with 1D positional encoding $\mathbb{P}$, bounding box embedding $\mathbb{B}$, and label embedding $\mathbb{L}$. 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 $\mathbb{E}$.

4.8.2 VLPM Augmented Retriever

Existing Vision Language Pretraining 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

[77]. 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 [55, 66, 114] in tasks related to understanding documents. Thus, the methods are proposed to harness the implicit information embedded in pretrained VLPMs to obtain more comprehensive and robust representations of multimodal entities.

(1) RoI-based Frameworks

RoI-based VLPMs focus on learning the contextual entity relationships and correlation between textual content and associated visual cues of each RoI, which in our scenario are document-semantic entities (e.g. *section*, *paragraph*, *table*, etc.). As shown in Figure 4.12, $\mathcal{F}_{roi}$ 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 $\mathcal{R}$. Vanilla **Transformer** is used as a foundational benchmark for evaluating the impact of various pretrained techniques in comparative studies [26]. Additionally, **VisualBERT** [66] and **LXMERT** [114] are adopted to enhance the initial visual embedding of each document entity. The improved visual embeddings are concatenated with $\mathbb{T}$ to obtain $\mathbb{E}$.

(2) 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 [45, 139]. Despite these advancements, the demands of cross-page document understanding remain insufficiently addressed. Consequently, our research investigates the effectiveness of patch-based VLPMs in the general domain in cross-page

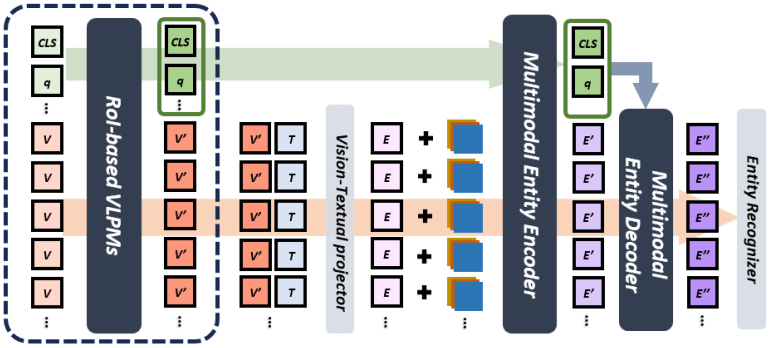


FIGURE 4.12. Multimodal IR on RoI-based VLPM.

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.

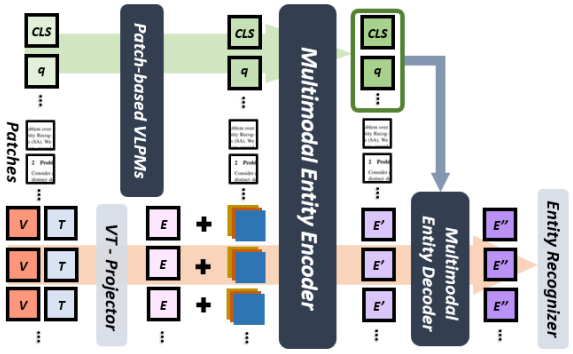


FIGURE 4.13. Multimodal IR on patch-based VLPM.

Figure 4.13 illustrates the procedure of multimodal information retrieval on the proposed Image Patch-based VLPM. To apply a vision-language model for cross-page document understanding, multiple document pages are merged $\mathbb{I} = \{I_1, I_2, ..., I_m\}$ 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, it could have $[Q', P', \mathbb{E}'] = \mathcal{E}([Q, P, \mathbb{E}])$, 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. The patch-based VLPMs are introduced to

obtain contextual patch embedding P, including models such as **CLIP** [96], **ViLT** [55], **BridgeTower** [138].

4.8.3 Joint-grained Retriever

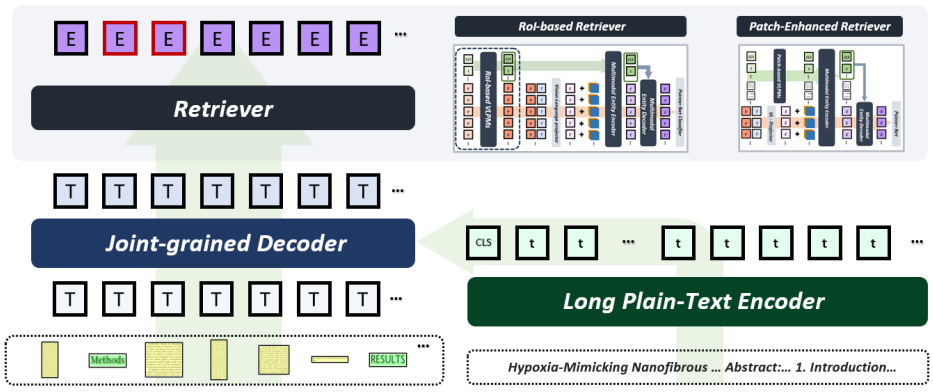


FIGURE 4.14. Joint-grained(coarse-and-fine grained) Retriever.

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.

To address this, a **Joint-grained Retriever** (Jg) architecture is introduced, shown in Figure 4.14, 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 pretrained 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 [148]. 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.

4.9 Experiments and Discussions

4.9.1 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

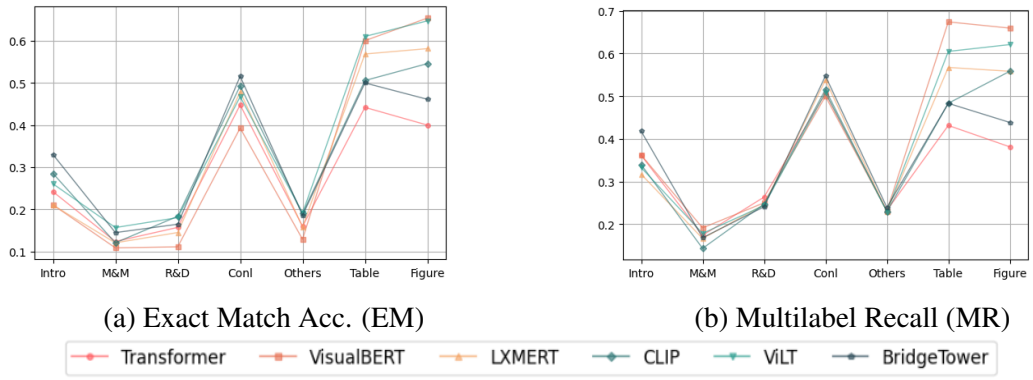
TABLE 4.5. Overall performance under various evaluation metrics.

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 *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 pretrained 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 pretrained 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 4.5 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 4.9.3 and 4.9.4.

4.9.2 Breakdown Analysis

(1) Breakdown Performance: Super-Section



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

FIGURE 4.15. Analysing baseline performance: visualisation of breakdowns across various Super-Sections

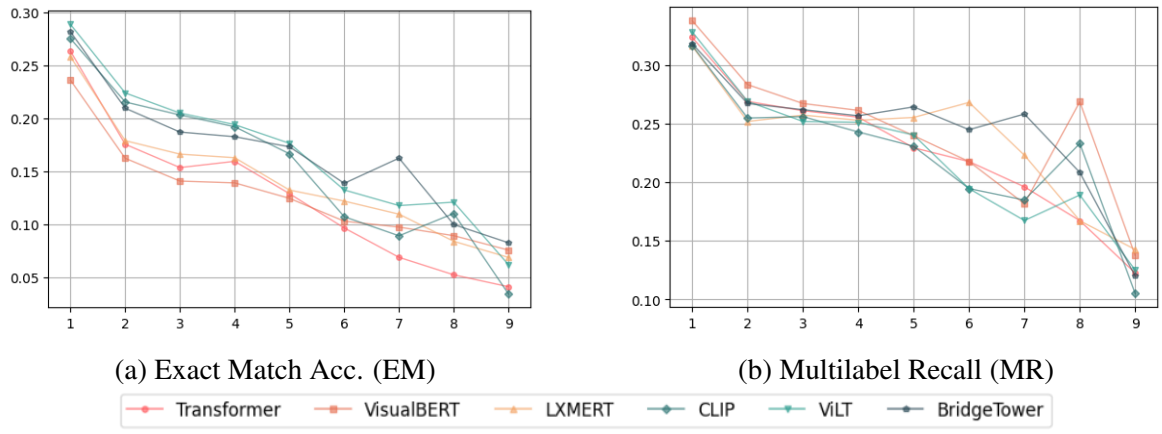
To thoroughly examine the benefits and limitations of various frameworks, addition breakdown analyses are performed, focusing on different question types, including ***paragraph***-based (*Intro*, *M&M*, *R&D*, *Conl*, *Other*) and ***table/figure***-based. As indicated in Figure 4.15, the *table/figure-based* questions tend to outperform *paragraph-based* ones regarding both *EM* and *MR* because table/figure questions typically have a single target entity (as Figure 4.8c shown). Additionally, VLPMs only rely solely on visual aspects of documents, resulting in more effective embeddings for visually rich entities. Moreover, as demonstrated in Figure 4.15a,

patch-augmented models generally yield better results for paragraph-based questions compared to RoI-based models, indicating that patch-augmented approaches may be more adept at capturing contextual information for text-dense entities, e.g. *Paragraph*, *List*.

Paragraph-based questions: It could be further categorised by their structural complexity. Introduction and Conclusion sections, characterised by simpler structures with fewer subsections and shorter contexts (refer to Figures 4.6a and 4.6b), demonstrate better performance in *EM* and *MR* compared to the *Material and Method* (*M&M*) and *Result and Discussion* (*R&D*) sections. This indicates that more complex structures with longer contexts might require advanced techniques for improved robustness.

Table/figure-based questions: distinct trends emerge compared to overall and paragraph-based performance. Specifically, RoI pretrained frameworks like VisualBERT and LXMERT reveal superior *EM* scores, as shown in Figure 4.15a. This improvement likely results from the enhanced capacity to represent visually-rich entities. Interestingly, some models, such as VisualBERT, may display lower *EM* but satisfactory *MR* performance, suggesting they can correctly identify target entities despite some answer inaccuracies. This characteristic is particularly beneficial in practical scenarios, such as aiding LLMs or Multimodal Generative Models in summarisation or QA tasks.

(2) Breakdown Performance: Page-range



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

FIGURE 4.16. Comparative performance analysis: visualising baseline performance across varied input page ranges in the test set.

As different question-answer pairs will have distinct numbers of document images, page-range-oriented breakdown analysis is conducted to examine the robustness of proposed frameworks on a changed number of input pages. Overall, the apparent decrease trend of EM can be observed in Figure 4.16a from near 30% on a single page to less than 5% on nine pages. This demonstrates that with the number of pages increasing, the difficulty of understanding the document dramatically improves. Notably, the Patch-based models retain better robustness under long page scenarios, which may benefit from the document-patch embedding, enabling more comprehensive and robust entity representations. However, different trends can be observed in *MR*. For RoI-based frameworks, frameworks with pretrained backbones (VisualBERT and LXMERT) exhibit greater robustness compared to vanilla transformers. Additionally, for Patch-based, CLIP and BridgeTower show strong robustness in *MR* but show weakness in ViLT. This may reveal the dual-encoder configurations can effectively recognise the target entities but may be sensitive to noise.

4.9.3 Joint-grained Framework Results

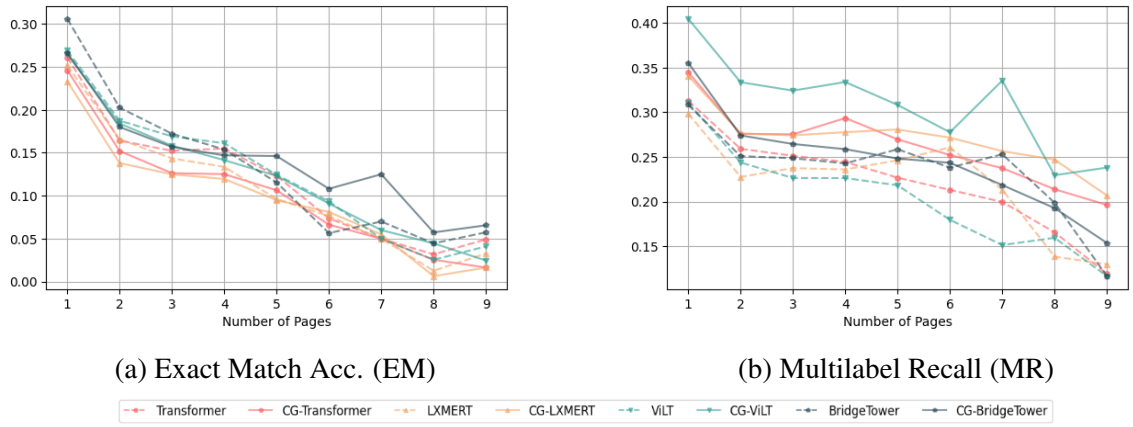
(1) Overall and Super-Section Breakdown Performance

To illustrate the effectiveness of the proposed Joint-grained framework (Figure 4.14), a performance comparison between the top two vanilla frameworks is conducted 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.

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

(2) Breakdown Performance: Page-range



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

FIGURE 4.17. Visualised breakdown performance of each model across different input page ranges.

To assess the Joint-grained framework's robustness across different input page numbers, a comparative analysis is conducted, shown in Figure 4.17. Figure 4.17a 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 4.17b, 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.

4.9.4 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 4.7. Comprehensive breakdown performance: BridgeTower Joint-grained frameworks based on various sourced textual token sequences, overall and super-Section based breakdown.

The real-world efficacy of the proposed Joint-grained framework is demonstrated by evaluating its performance using text extracted from off-the-shelf tools. Significant improvements are highlighted for BridgeTower, as shown in Table 4.6. The performance of BridgeTower-based Joint-grained frameworks on various text token sequences is presented, including those from the MMVQA dataset (Jg-BridgeTower), PDF parser (Jg-BridgeTower-PDFMiner), and OCR tools (Jg-BridgeTower-OCR). As shown in Table 4.7, incorporating fine-grained textual 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.

4.9.5 Category-oriented Entity Representation

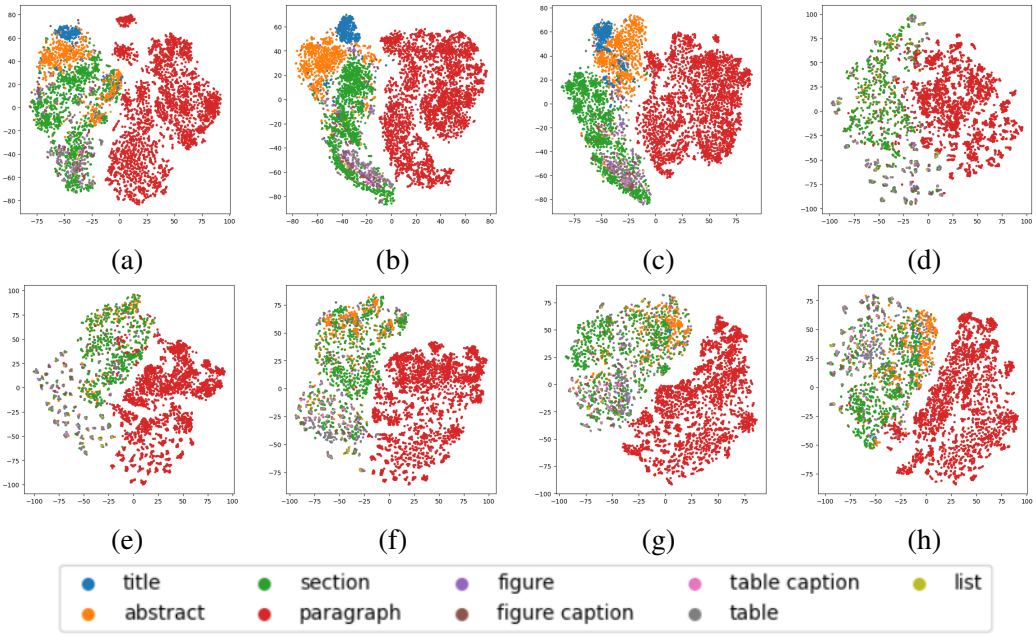


FIGURE 4.18. Category-oriented entity representation T-SNE analysis of various frameworks including (a) Transformer, (b) VisualBERT, (c) LXMERT, (d) CLIP, (e) ViLT, (f) BridgeTower, (g) Jg-BridgeTower-PDFMiner, (h) Jg-BridgeTower-OCR.

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 4.18. 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 4.5. The possible reason is although they benefit from pretrained 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.

For RoI-based frameworks, Transformer underperforms VisualBERT and LXERMT in *Table* and *Figure* question types. This performance gap can be attributed to the distinctiveness of

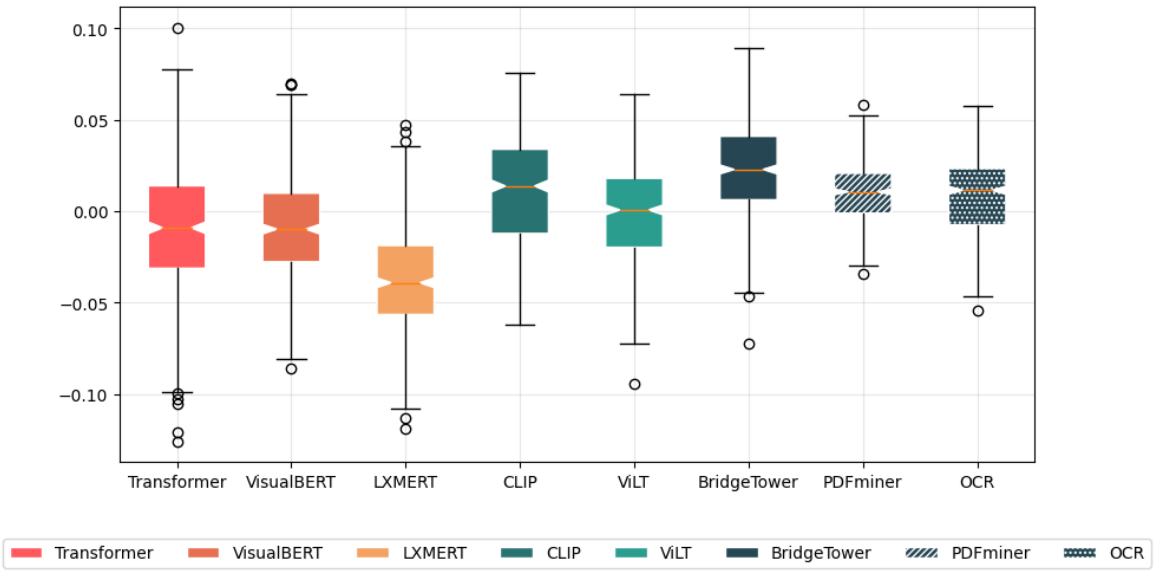


FIGURE 4.19. Question-Answering embedding-based cosine similarity/correlation distribution.

entity embeddings for *Figure* and *Table*, as shown in Figures 4.18b and 4.18c for VisualBERT and LXMERT, respectively, compared to Transformer (Figure 4.18a). Additionally, for Patch-based models, BridgeTower outperforms other counterparts on paragraph-based questions. This may be linked to BridgeTower's focused pretraining on textual content and clearer clustering of text-dense entities as illustrated in Figure 4.18f. Moreover, compared to the vanilla Bridgetower framework (Figure 4.18f), Joint-grained information-augmented models (Figure 4.18g, 4.18h) tend to have more representative entity representations, especially for text-dense document entities, e.g. *Abstract*, *Section*.

4.9.6 Question-Answering Correlation

The correlation between question and target entity representations is assessed by calculating the average cosine similarity between the *[CLS]* embedding of the question and the averaged target entity embedding. Our findings reveal that models enhanced with image patches exhibit a higher question-answer correlation, potentially explaining their superior performance over RoI-based models despite the latter's more category-specific representational features. However, Joint-grained models like BridgeTower-PDFminer and BridgeTower-OCR show

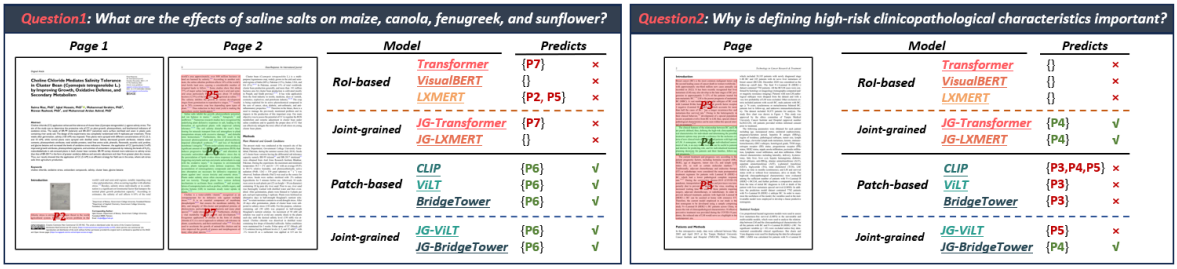


FIGURE 4.20. Qualitative analysis of various model performance on two sample questions.

lower cosine similarity scores compared to the original BridgeTower, even though they have better performance (as Table 4.7 shown). This prompts further qualitative case studies to delve into the efficacy of these models in the MMVQA context.

4.10 Case Study

To demonstrate the effectiveness of the proposed frameworks, the predictions of various architectures are represented and analysed qualitatively. As shown in Figure 4.20, all RoI-based frameworks, including Joint-grained models, failed to identify the correct answer paragraph, even with errors including predicting entities on entirely different pages (LXMERT). Conversely, patch-based models successfully located paragraph $P6$, suggesting that integrating patch embeddings with entity representations ($\mathbb{E}$) enhances understanding of document layout. However, as displayed with Question 2 in Figure 4.20, patch embeddings alone were insufficient for accurate paragraph location. Instead, Joint-grained frameworks that incorporate fine-grained information achieved correct predictions, underlining the effectiveness of fine-grained data in improving entity representation robustness.

4.11 Conclusion

This paper presents a significant contribution by introducing the MMVQA dataset and a novel joint-grained architecture. The MMVQA, sourced from PubMed Central, showcases diverse document types, complex structures, and extensive content-related questions in multi-page

documents. A robust benchmark, the Joint-grained retrieval architecture, is also introduced, consistently enhancing model performance, particularly in complex document sections. This research aims to advance the understanding of multi-page document comprehension and establish a foundation for future exploration and refinement of models in this domain, marking a significant step forward in document understanding research.

CHAPTER 5

Knowledge Transfer for VRD Understanding

This chapter is based on the work ***3MVRD: Multimodal Multi-task Multi-teacher Visually-Rich Form Document Understanding***, accepted at **ACL 2024** [29]. I was responsible for formulating the research aim, designing the methodology, implementing the experiments, and writing the entire paper.

This work begins with the introduction of a joint-grained framework to address the limitations of both fine-grained and coarse-grained Visual Reasoning and Document Understanding (VRDU) models. It then steps into utilizing multi-teacher knowledge distillation methods to leverage the implicit knowledge from pretrained models, enhancing the performance of lightweight student models.

This paper presents a groundbreaking multimodal, multi-task, multi-teacher joint-grained knowledge distillation model for visually-rich form document understanding. The model is designed to leverage insights from both fine-grained and coarse-grained levels by facilitating a nuanced correlation between token and entity representations, addressing the complexities inherent in form documents. Additionally, the new intra-grained and cross-grained loss functions are introduced to further refine diverse multi-teacher knowledge distillation transfer process, presenting distribution gaps and a harmonised understanding of form documents. Through a comprehensive evaluation across publicly available form document understanding datasets, the proposed model consistently outperforms existing baselines, showcasing its efficacy in handling the intricate structures and content of visually complex form documents.

5.1 Introduction

5.1.1 Background

Understanding and extracting structural information from Visually-Rich Documents (VRDs), such as academic papers [27, 157], receipts [91], and forms [26, 49], holds immense value for Natural Language Processing (NLP) tasks, particularly in information extraction and retrieval. While significant progress has been made in solving various VRD benchmark challenges, including layout analysis and table structure recognition, the task of form document understanding remains notably challenging.

This complexity of the form document understanding arises from two main factors: 1) the involvement of two distinct authors in a form and 2) the integration of diverse visual cues. Firstly, forms mainly involve two primary authors: form designers and users. Form designers create a structured form to collect necessary information as a user interface. Unfortunately, the form layouts, designed to collect varied information, often lead to complex logical relationships, causing confusion for form users and heightening the challenges in form document understanding. Secondly, diverse authors in forms may encounter a combination of different document natures, such as digital, printed, or handwritten forms. Users may submit forms in various formats, introducing noise such as low resolution, uneven scanning, and unclear handwriting. Traditional document understanding models do not account for the diverse carriers of document versions and their associated noises, exacerbating challenges in understanding form structures and their components. Most VRD understanding models inherently hold implicit multimodal document structure analysis (Vision and Text understanding) knowledge either at fine-grained [45, 127] or coarse-grained [66, 114] levels. The fine-grained models mainly focus on learning detailed logical layout arrangement, which cannot handle complex relationships of multimodal components, while the coarse-grained models tend to omit significant words or phrases.

5.1.2 Research Aims

Hence, a novel joint-grained document understanding approach is introduced with multimodal multi-teacher knowledge distillation. This method leverages knowledge from various task-based teachers (pretrained models) throughout the training process, intending to create more inclusive and representative multi- and joint-grained document representations.

5.1.3 Contributions

The contributions can be summarised as follows: 1) We present a groundbreaking multimodal, multi-task, multi-teacher joint-grained knowledge distillation model designed explicitly to understand visually rich form documents. 2) The proposed model outperforms across publicly available form document datasets. 3) To the best of my knowledge, this research marks the first in adopting multi-task knowledge distillation, focusing on incorporating multimodal form document components.

Model	Modalities	Pretraining Datasets	Pretraining Tasks	Downstream Tasks	Granularity
Donut [54]	V	IIT-CDIP	NTP	DC, VQA, KIE	Token
Pix2struct [61]	V	C4 corpus	NTP	VQA	Token
LiLT [127]	T, S	IIT-CDIP	MVLM, KPL, CAI	DC, KIE	Token
BROS [42]	T, S	IIT-CDIP	MLM, A-MLM	KIE	Token
LayoutLMv3 [45]	T, S, V	IIT-CDIP	MLM, MIM, WPA	DC, VQA, KIE	Token
DocFormerv2 [2]	T, S, V	IDL	TTL, TTG, MLM	DC, VQA, KIE	Token
Fast-StrucText [149]	T, S, V	IIT-CDIP	MVLM, GTR, SOP, TIA	KIE	Token
FormetNV2 [60]	T, S, V	IIT-CDIP	MLM, GCL	KIE	Token
M^3-DVQA (3MVRD)	T, S, V	FUNSD, Form-NLU	Multi-teacher Knowledge Distillation	KIE	Token, Entity

TABLE 5.1. Comparison with state-of-the-art models for receipt and form understanding. In the *Modalities* column, *T* represents textual information, *V* represents visual information, and *S* represents spatial information.

5.2 Related Works

Visually Rich Document (VRD) understanding entails comprehending the structure and content of documents by capturing the underlying relations between textual and visual

modalities. Several downstream tasks, such as Key Information Extraction (KIE), Document Classification (DC), and Visual Question Answering (VQA), have contributed to raising the attention of the multimodal learning community. In this work, form documents are addressed, whose structure and content are particularly challenging to understand [111]. Form documents possess intricate structures involving collaboration between form designers, who craft clear structures for data collection, and form users, who interact with the forms based on their comprehension, with clarity and ease of understanding being variable. Table 5.1 enumerates the main characteristics of the latest models proposed for document understanding.

Vision-only approaches: They exclusively rely on the visual representation (denoted by *V* modality in Table 5.1) of the document components thus circumventing the limitations of state-of-the-art text recognition tools (e.g., Donut [54] and Pix2struct [61]). Their document representations are commonly pretrained using a Next Token Prediction (NTP) strategy, offering alternative solutions to traditional techniques based on Natural Language Processing.

Multimodal approaches: They leverage both the recognised text and the spatial relations (denoted by *T* and *S*) between document components (see, for instance, LiLT [127] and BROS [42]). The main goal is to complement the understanding of raw content with layout information. Expanding upon this multimodal framework, models such as LayoutLMv3 [45], DocFormerv2 [2], Fast-StrucText [149], and, FormNetV2 [60] integrate the visual modality with text and layout information. These approaches are capable of capturing nuances in the document content hidden in prior works. To leverage multimodal relations, these models are typically pretrained in a multi-task fashion, exploiting a curated set of token- or word-based pretraining tasks, such as masking or alignment.

3MVRD aligns with the multimodal model paradigm, distinguishing itself by eschewing generic pretraining tasks reliant on masking, alignment, or NTP. Instead, it leverages the direct extraction of knowledge from *multiple teachers*, each trained on downstream datasets, encompassing *both entity and token levels* of analysis with the proposed inter-grained and cross-grained losses. This enriches the depth of understanding in visual documents, capturing intricate relationships and semantic structures beyond individual tokens.

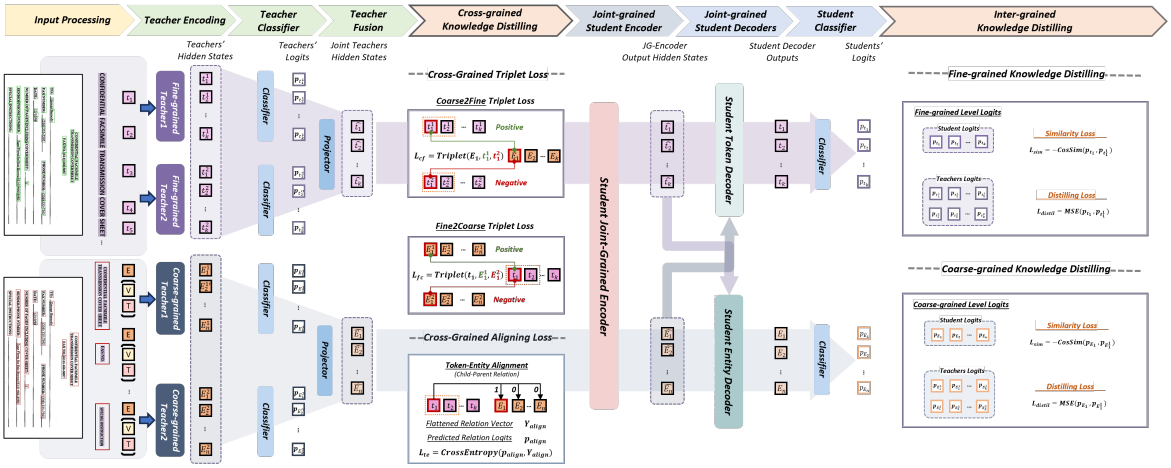


FIGURE 5.1. Multimodal Multi-task Multi-Teacher Joint-Grained Document Understanding Framework. Each section is aligned with the specific colours, Green: Section 5.3.2, Blue: Section 5.3.2, Orange: Section 5.3.3

5.3 Methodology

As previously noted, this paper focuses on interpreting visually rich documents, particularly form documents created and used collaboratively by multiple parties. To accomplish this objective, two tiers of multimodal information are introduced and employed: fine-grained and coarse-grained levels, which play a crucial role in understanding the structure and content of an input form page. Note that existing pretrained visual-language models, whether designed for generic documents, possess implicit knowledge on either fine-grained or coarse-grained aspects. Hence, an approach is proposed that harnesses knowledge from diverse pretrained models throughout training. This strategy aims to generate more comprehensive and representative multi- and joint-grained document representations, ultimately enhancing the effectiveness of downstream tasks related to document understanding.

5.3.1 Preliminary Definitions

Prior to going through the proposed approach in detail, formal definitions will be given for the terminology employed throughout this paper. To establishing clear and precise definitions could contribute to a comprehensive understanding of the concepts and terms integral to this research.

1) Fine-grained Document Understanding [42, 45, 127] is a pivotal aspect of document analysis, involving frameworks that offer detailed insights to comprehend document content, particularly when addressing token-level tasks, such as span-based information extraction and question answering. Regarding ***input features***, existing pretrained models at the fine-grained level harness multimodal features, such as positional information and image-patch embedding, to enhance the fine-grained token representations. The ***pretraining phase*** incorporates several learning techniques, including Masked Visual-Language Modelling, Text-Image Matching, and Multi-label Document Classification, strategically designed to acquire inter or cross-modality correlations and contextual knowledge. However, it is essential to acknowledge the ***limitations*** of fine-grained frameworks, as their primary focus lies in learning the logical and layout arrangement of input documents. These frameworks may encounter challenges in handling complex multimodal components.

2) Coarse-grained Document Understanding [66, 114] is a vital component in document analysis, with frameworks adept at grasping the logical relations and layout structures within input documents. Particularly well-suited for tasks like document component entity parsing, coarse-grained models excel in capturing high-level document understanding. Despite the dominant trend of fine-grained document understanding models, some research recognises [66, 114] that the knowledge from general domain-based Visual-Language Pretrained Models (VLPMs) could be leveraged to form a foundational document understanding. However, the coarse-grained document understanding models have significant ***limitations***, including their tendency to overlook detailed information, leading to the omission of significant words or phrases. Preliminary entity-level annotations are often necessary, and the current backbone models are pretrained on the general domain, highlighting the need for document domain frameworks specifically pretrained at the coarse-grained level.

5.3.2 Multimodal Multi-task Multi-teacher Joint-grained Document Understanding[1]

Therefore, a joint-grained document understanding framework $\mathcal{F}_{jg}$ is introduced, designed to harness pretrained knowledge from both fine-grained and coarse-grained levels. This approach integrates insights from multiple pretrained backbones, facilitating a unified understanding of document content encompassing detailed nuances and high-level structures. It aims to synergise the strengths of fine-grained and coarse-grained models, enhancing the overall effectiveness of form understanding tasks.

(1) Multimodal Multi-task Multi-Teacher

To facilitate this joint-grained framework, **M**ultimodal **M**ulti-teachers from two **M**ulti-tasks, fine-grained and coarse-grained tasks within the framework, are employed. While the fine-grained teacher $\mathcal{F}_{fg}$ is characterised by checkpoints explicitly fine-tuned for the **token classification**, the coarse-grained teacher $\mathcal{F}_{cg}$ utilises fine-tuning checkpoints for the document component **entity classification**. The details of fine-grained and coarse-grained teacher models are articulated in Section 5.4.3. The ablation study of those teacher models is in Section 5.5.3. $\mathcal{F}_{fg}$ and $\mathcal{F}_{cg}$ get the encoded inputs of token and entity level, respectively, to acquire the corresponding last layer hidden states and logits for downstream procedures. For example, after feeding the sequence of tokens $\widetilde{\mathbf{t}} = \{\widetilde{t_1}, \widetilde{t_2}, ..., \widetilde{t_k}\}$ and sequence of multimodal entity embeddings $\widetilde{\mathbf{E}} = \{\widetilde{E_1}, \widetilde{E_2}, ..., \widetilde{E_n}\}$ into $\mathcal{F}_{fg_1}$ and $\mathcal{F}_{cg_1}$, respectively, we acquire the hidden states $\mathbf{t^1} = \{t_1^1, t_2^1, ..., t_k^1\}$ and $\mathbf{E^1} = \{E_1^1, E_2^1, ..., E_n^1\}$, as well as classification logits $\mathbf{p_{t^1}} = \{p_{t_1^1}, p_{t_2^1}, ..., p_{t_k^1}\}$ and $\mathbf{p_{E^1}} = \{p_{E_1^1}, p_{E_2^1}, ..., p_{E_n^1}\}$. Supposing $\mathbb{T} = \{\mathbf{t^1}, \mathbf{t^2}, ...\}$ and $\mathbb{E} = \{\mathbf{E^1}, \mathbf{E^2}, ...\}$ are hidden states from multiple teachers, the combined representations are fed into corresponding projection layers $\mathcal{L}_{fg}$ and $\mathcal{L}_{cg}$ to get the multi-teacher representations $\widehat{\mathbf{t}} = \{\widehat{t_1}, \widehat{t_2}, ..., \widehat{t_k}\}$ and $\widehat{\mathbf{E}} = \{\widehat{E_1}, \widehat{E_2}, ..., \widehat{E_n}\}$ for each grain.

(2) Joint-Grained Learning

The joint-grained learning framework comprises a joint-grained encoder and decoder.

[1]Subsections are aligned with different colour in Figure 5.1, Green: Section 5.3.2, Blue: Section 5.3.2, Orange: Section 5.3.3

The joint-grained encoder $\mathcal{E}$, implemented as a transformer encoder, is designed to learn the contextual correlation between fine-grained $\widehat{\mathbf{t}}$ and coarse-grained $\widehat{\mathbf{E}}$ representations. This enables the model to capture nuanced details at the token level while simultaneously grasping the high-level structures represented by entities within the document.

The joint-grained decoders $\mathcal{D}$ play a crucial role in processing the augmented joint-grained representations. For the fine-grained decoder $\mathcal{D}_{fg}$, the input comprises fine-grained token representations $\widehat{\mathbf{t}}$, with the entity representation serving as memory $\widehat{\mathbf{E}}$. This configuration allows the decoder to focus on refining and generating augmented token representations $\mathbf{t}$ based on the contextual information provided by both token and entity representations. In contrast, for coarse-grained decoder $\mathcal{D}_{cg}$, the input is the entity representation $\widehat{\mathbf{E}}$, while the memory consists of token representations $\widehat{\mathbf{t}}$. This approach enables the coarse-grained decoders to emphasise broader structures and relationships at the entity level, leveraging the memory of fine-grained token information to generate a more comprehensive entity representation $\mathbf{E}$. Overall, the proposed joint-grained architecture facilitates a comprehensive understanding of document content by incorporating fine-grained and coarse-grained perspectives.

The pretraining of different teacher models involves diverse techniques and features, so a simplistic approach of merely concatenating or pooling hidden states may not fully leverage the individual strengths of each model. Traditional self-/cross attention-based transformer encoders or decoders might encounter challenges in integrating knowledge from various grains, potentially introducing noise to specific grained weights. To address this concern, multiple types of losses are proposed to thoroughly explore implicit knowledge within the diverse teachers (pretrained models).

5.3.3 Multimodal Multi-task Multi-Teacher Knowledge Distillation

This section introduces the multi-loss strategy to enhance inter-grained and cross-grained knowledge exchange, ensuring a more nuanced and effective integration of insights from fine-grained and coarse-grained representations. The accompanying multi-loss ablation study

(Section 5.5.4) aims to optimise the synergies between multiple teacher models, thereby contributing to a more robust and comprehensive joint-grained learning process.

(1) Task-oriented Cross Entropy Loss

The Task-oriented Cross Entropy (CE) loss is pivotal in facilitating a task-based knowledge distillation strategy. This is computed by comparing the predictions of the student model with the ground truth for each specific task. Adopting the CE loss provides the student model with direct supervisory signals, thereby aiding and guiding its learning process. Note that two task-oriented CE losses are used within the proposed approach, one from the token classification task and the other from the entity classification task. The output hidden states from $\mathcal{D}_{fg}$ and $\mathcal{D}_{cg}$ are fed into classifiers to get the output logits $\mathbf{p_t} = \{p_{t_1}, p_{t_2}, ..., p_{t_k}\}$ and $\mathbf{p_E} = \{p_{E_1}, p_{E_2}, ..., p_{E_n}\}$. Supposing the label sets for fine-grained and entity-level tasks are $\mathbf{Y_t} = \{y_{t_1}, y_{t_2}, ..., y_{t_k}\}$ and $\mathbf{Y_E} = \{y_{E_1}, y_{E_2}, ..., y_{E_n}\}$, the fine-grained and coarse-grained Task-oriented Cross Entropy losses l_t and l_E are calculated as:

$$l_t = CrossEntropy(\mathbf{p_t}, \mathbf{Y_t}) \tag{5.1}$$

$$l_e = CrossEntropy(\mathbf{p_E}, \mathbf{Y_E}) \tag{5.2}$$

(2) Inter-Grained Loss Functions

Since various pretrained models provide different specific knowledge to understand the form comprehensively, effectively distilling valuable information from selected fine-tuned checkpoints may generate more representative token representations. In addressing this, two target-oriented loss functions are introduced tailored to distil knowledge from teachers at different levels. These aim to project the label-based distribution from fine-grained $\mathbf{p}_\mathbb{T} = \{\mathbf{p_{t^1}}, \mathbf{p_{t^2}}, ...\}$ or coarse-grained teacher logits $\mathbf{p}_\mathbb{E} = \{\mathbf{p_{E^1}}, \mathbf{p_{E^2}}, ...\}$ to corresponding student logits $\mathbf{p_t}$ and $\mathbf{p_E}$, enabling efficient learning of label distributions.

Similarity Loss: This is introduced as an effective method to distil knowledge from the output logits $\mathbf{p_t}$ and $\mathbf{p_E}$ of selected fine-grained or coarse-grained teacher checkpoints from $\mathbf{p}_\mathbb{T}$ and $\mathbf{p}_\mathbb{E}$. It aims to mitigate the logit differences between the student classifier and the

chosen teachers using cosine similarity ($CosSim$), promoting a more aligned understanding of the label-based distribution. Supposing have n_t and n_e teachers for fine-grained and coarse-grained tasks, respectively, the similarity loss of fine-grained l_{sim_t} and coarse-grained l_{sim_e} can be calculated by:

$$l_{sim_t} = -\sum_{i}^{i=n_t}\sum_{j}^{j=k} CosSim(p_{t_j^i}, p_{t_j}) \tag{5.3}$$

$$l_{sim_e} = -\sum_{i}^{i=n_e}\sum_{j}^{j=n} CosSim(p_{E_j^i}, p_{E_j}) \tag{5.4}$$

Distilling Loss: Inspired by [93], an extreme logit learning model is adopted for the distilling loss. This loss implements knowledge distillation using Mean Squared Error (MSE) between the students' logits $\mathbf{p_t}$ and $\mathbf{p_E}$ and the teachers' logit sets $\mathbf{p}_{\mathbb{T}}$ and $\mathbf{p}_{\mathbb{E}}$. This method is employed to refine the knowledge transfer process further, promoting a more accurate alignment between the student and teacher models.

$$l_{distil_t} = \frac{1}{k}\sum_{j}^{j=k} MSE(p_{t_j^i}, p_{t_j}) \tag{5.5}$$

$$l_{distil_e} = \frac{1}{n}\sum_{j}^{j=n} MSE(p_{E_j^i}, p_{E_j}) \tag{5.6}$$

The introduction of these inter-grained loss functions, including the similarity loss and the distilling loss, contributes to mitigating distribution gaps and fostering a synchronised understanding of the form across various levels of granularity.

(3) Cross-Grained Loss Functions In addition, the cross-grained loss functions are incorporated. While fine-grained and coarse-grained information inherently align, the joint-grained framework employs self-attention and cross-attention to approximate the correlation between token and entity representations. $\mathbb{T}$ and $\mathbb{E}$ are teachers hidden states sets, each $\mathbf{t^i} \in \mathbb{T}$ and $\mathbf{E^i} \in \mathbb{E}$ are represented $\mathbf{t^i} = \{t_1^i, t_2^i, ..., t_k^i\}$ and $\mathbf{E^i} = \{E_1^i, E_2^i, ..., E_n^i\}$ and $\mathbf{t}$ and $\mathbf{E}$ are hidden states from student decoder.

Cross-grained Triplet Loss: Inherent in each grained feature are parent-child relations between tokens and aligned semantic form entities. The introduction of triplet loss aids the framework in automatically selecting more representative feature representations by measuring the feature distance from one grain to another-grained aligned representation. This

effectively enhances joint-grained knowledge transfer, optimising the overall understanding of the form. For acquiring the loss $l_{triplet_{fg}}$ to select fine-grained teachers based on coarse-grained distribution adaptively, the anchor is defined as each entity $E_i \in \mathbf{E}$ which has the paired token representations $t_i^1 \in \mathbf{t^1}$ and $t_i^2 \in \mathbf{t^2}$ (if the number of teachers is more significant than 2, randomly select two of them). The L-2 norm distance is used to measure the distance between fine-grained teachers (t_i^1, t_i^2) and anchor E_j, where the more similar entities are treated as positive samples (t_i^{pos}) otherwise negative (t_i^{neg}). For coarse-grained triplet loss $l_{triplet_{cg}}$, the same measurements are adopted for coarse-grained teacher positive (E_j^{pos}) and negative selection (E_j^{neg}) for an anchor t_i. Supposing the j-th, $l_{triplet_{fg}}$ and $l_{triplet_{cg}}$ are defined:

$$l_{triplet_{fg}} = \frac{1}{k} \sum_{i}^{i=k} Triplets(E_j, t_i^{pos}, t_i^{neg}) \tag{5.7}$$

$$l_{triplet_{cg}} = \frac{1}{k} \sum_{i}^{i=k} Triplets(t_i, E_j^{pos}, E_j^{neg}) \tag{5.8}$$

As one entity is typically paired with more than one token, when calculating $l_{triplet_{cg}}$, all k entity-token pairs are considered.

Cross-grained Alignment Loss: In addition to the triplet loss, designed to filter out less representative teachers, another auxiliary task is introduced. This task focuses on predicting the relations between tokens and entities, providing an additional layer of refinement to the joint-grained framework. The cross-grained alignment loss further contributes to the comprehensive learning and alignment of token and entity representations, reinforcing the joint-grained understanding of the form document. For an input form document page containing k tokens and n entities, a targeting tensor $\mathbf{Y_{align}}$ is used where $Dim(\mathbf{Y_{align}}) = \mathbb{R}^{k \times n}$. The acquired alignment logit is employed $\mathbf{p_{align}} = \mathbf{t} \times \mathbf{E}$ to represent the predicted token-entity alignments. The cross-grained alignment loss l_{align} can be calculated by:

$$l_{align} = CrossEntropy(\mathbf{p_{align}}, \mathbf{Y_{align}}) \tag{5.9}$$

5.4 Evaluation Setup

5.4.1 Datasets[2]

FUNSD [49] comprises 199 noisy scanned documents from various domains, including marketing, advertising, and science reports related to US tobacco firms. It is split into train and test sets (149/50 documents), and each document is presented in either printed or handwritten format with low resolutions. The evaluation focuses on the semantic-entity labelling task that identifies four predefined labels (i.e., question, answer, header, and other) based on input text content.

Form-NLU [26] consists of 867 financial form documents collected from Australian Stock Exchange (ASX) filings. It includes three form types: digital ***(D)***, printed ***(P)***, and handwritten ***(H)***, and is split into five sets: train-***D*** (535), val-*D* (76), test-*D* (146), test-***P*** (50), and test-***H*** (50 documents) and supports two tasks: Layout Analysis and Key Information Extraction. The evaluation focuses on the layout analysis that identifies seven labels (i.e., title, section, form key, form value, table key, table value, and others), detecting each document entity, especially for ***P*** and ***H***, the complex multimodal form document.

Table 5.2 and 5.3 demonstrate the number of tokens(length) and the number of document entities. While FUNSD has four types(Question, Answer, Header, Other) of document entities, Form-NLU has seven types(Title, Section, Form Key, Form Value, Table Key, Table Value, and Other). For the Form-NLU, two types of test sets are used, including Printed **P** and Handwritten **H**.

FUNSD (Testing)	**Question**	**Answer**	**Header**	**Other**	**Total**
Entity	1077	821	122	312	**2332**
Token	2654	3294	374	2385	**8707**

TABLE 5.2. FUNSD testing dataset distribution by label.

[2]The statistics of token/entity are shown in Table 5.2 and 5.3.

Form-NLU (Testing)		Title	Section	Form Key	Form Value	Table Key	Table Value	Others	Total
P	**Entity**	98	100	346	332	250	249	152	**1527**
H		100	100	348	315	249	226	149	**1487**
P	**Token**	700	1258	1934	1557	993	389	3321	**10152**
H		742	1031	1805	866	779	366	2918	**8507**

TABLE 5.3. Form-NLU test dataset distribution by categories, where ***P*** and ***H*** are printed and handwritten sets.

5.4.2 Baselines and Metrics

For **token-level information extraction** baselines, three Document Understanding (DU) models are adopted: LayoutLMv3 [45], LiLT [127], and BROS [42]. LayoutLMv3 employs a word-image patch alignment that utilises a document image along with its corresponding text and layout position information. In contrast, LiLT and BROS focus only on text and layout information without incorporating images. LiLT uses a bi-directional attention mechanism across token embedding and layout embedding, whereas BROS uses a relative spatial encoding between text blocks. For **entity-level information extraction** baselines, two vision-language (VL) models are used: LXMERT [114] and VisualBERT [66]. Compared to the two DU models, these VL models use both image and text input without layout information. LXMERT focuses on cross-modality learning between word-level sentence embeddings and object-level image embeddings, while VisualBERT simply inputs image regions and text, relying on implicit alignments within the network. For **evaluation metrics**, inspired by [49] and [26], the F1-score is primarily used to represent both overall and detailed performance breakdowns, aligning with other baselines.

5.4.3 Implementation Details

In token-level experiments, LayoutLMv3-base is fine-tuned using its text tokenizer and image feature extractor. LiLT is also fine-tuned and combined with the RoBERTa base. In entity-level experiments, pretrained BERT (768-d) is employed for encoding textual content, while ResNet101 (2048-d) is used for the region-of-interest(RoI) feature to capture the visual aspect. These extracted features serve as input for fine-tuning LXMERT and VisualBERT. All fine-tuned models serve as teacher models. The hyperparameter testing involves a maximum of 50

Fine-grained	Coarse-Grained	Configure	# Para	# Trainable
LiLT	N/A	Teacher	130,169,799	130,169,799
LayoutLMv3	N/A	Teacher	125,332,359	125,332,359
	LXMERT	JG-Encoder	393,227,514	19,586,415
		JG-Decoder	423,952,890	50,311,791
		JG-$\mathcal{E}\&\mathcal{D}$	440,494,842	66,853,743
	VisualBERT&LXMERT		557,260,798	70,394,991
Layoutlmv3&LiLT	LXMERT		574,205,889	68,034,159
	VisualBERT&LXMERT		688,611,013	71,575,407

TABLE 5.4. Model configurations and number of parameters.

epochs with learning rates set at 1e-5 and 2e-5. All are conducted on a Tesla V100-SXM2 with 16GB graphic memory and 51 GB memory, CUDA 11.2.

5.4.4 Number of Parameters

The table presented in Table 5.4 outlines the number of total parameters and trainable parameters across various model configurations. It is evident that the choice of teacher models primarily determines the total number of parameters. As the number of teachers increases, there is a corresponding enhancement in the total parameter count. Furthermore, the architecture of the student model significantly influences the number of trainable parameters. For instance, encoder-decoder-based student models exhibit a higher count of trainable parameters compared to architectures employing only an encoder or decoder. This discrepancy implies that training encoder-decoder models demands more computational resources. Despite the variation in trainable parameters among different student model architectures, it is noteworthy that the overall number remains substantially smaller than that of single-teacher fine-tuning processes. This observation underscores the efficiency of student model training in comparison to fine-tuning pretrained models.

Model	Config & Loss	FUNSD	Form-NLU	
			P	***H***
BROS	Single Teacher	82.44	92.45	93.68
LiLT	Single Teacher	87.54	<u>96.50</u>	91.35
LayoutLMv3	Single Teacher	<u>90.61</u>	95.99	<u>97.39</u>
JG-$\mathcal{E}$	Joint Cross Entropy	90.45	94.91	96.55
JG-$\mathcal{D}$	Joint Cross Entropy	90.48	95.68	97.62
JG-$\mathcal{E}\&\mathcal{D}$	Joint Cross Entropy	90.57	95.93	97.62
MT-JG-$\mathcal{E}\&\mathcal{D}$ (3MVRD)	Joint Cross Entropy	90.53	97.21	97.75
	+ *Sim*	**91.05**	98.25	98.09
	+ *Distil*	90.90	98.12	97.72
	+ *Triplet*	90.28	97.58	97.28
	+ *Align*	90.55	97.24	97.42
	+ *Sim* + *Distil* + *Triplet* + *Align*	90.92	**98.69**	**98.39**

TABLE 5.5. Overall performance of various model and configurations on Form-NLU printed ***P*** and handwritten ***H***. The full form of acronyms can be found in Section 5.5.1. The best is in **bold**. The best baseline is <u>underlined</u>.

5.5 Results

5.5.1 Overall Performance

Extensive experiments are conducted to highlight the effectiveness of the proposed **Multimodal Multi-task Multi-Teacher** framework, including *joint-grained learning*, *multi-teacher* and *multi-loss* architecture. Table 5.5 shows representative model configurations on various adopted modules. LayoutLMv3 performs notably superior to BROS and LiLT, except for the Form-NLU printed test set. LayoutLMv3 outperforms around 3% and 4%, the second-best baseline on FUNSD and Form-NLU handwritten sets, respectively. This superiority can be attributed to LayoutLMv3's utilisation of patched visual cues and textual and layout features, resulting in more comprehensive multimodal representations. So it can be observed that LayoutLMv3 would be a robust choice for fine-grained baselines in further testing[3]. To find the most suitable **Joint-Grained learning** (JG), the results of single-teacher joint-grained frameworks are compared, including Encoder ($\mathcal{E}$) only, Decoder

[3]I chose LLmv3 and LXMERT for JG and selected LLMv3&LilT and VBERT&LXMERT for MT-JG-$\mathcal{E}\&\mathcal{D}$. More teacher combinations analysis is in Section 5.5.3.

($\mathcal{D}$) only, and Encoder with Decoder ($\mathcal{E}\&\mathcal{D}$). Table 5.5 illustrates $\mathcal{E}\&\mathcal{D}$ achieving the highest performance among three baselines. However, upon integrating multiple teachers from each grain (MT-JG-$\mathcal{E}\&\mathcal{D}$), competitive performance is observed compared to the baselines on both Form-NLU printed (***P***) (from LiLT 96.5% to 97.21%) and handwritten set (***H***) (from LiLT 97.39% to 97.75%). Still, additional techniques may be necessary to distill the cross-grained multi-teacher information better.

To thoroughly distil joint-grained knowledge from multiple teachers, multiple loss functions are introduced encompassing **Multiple auxiliary tasks**. These functions capture teacher knowledge from inter-grained and cross-grained perspectives, generating representative token embeddings. Typically, using either inter-grained or coarse-grained loss individually leads to better performance than the best baselines across various test sets. Inter-grained Similarity (*Sim*) and Distilling (*Distil*) loss consistently achieve higher F1 scores in nearly all test sets. Moreover, cross-grained *Triplet* and alignment (*Align*) losses outperform the best baseline on the Form-NLU (***P***) or (***H***). This highlights the effectiveness of the proposed multi-task learning approach in enhancing token representations by integrating knowledge from joint-grained multi-teachers. Inter-grained loss functions exhibit higher robustness on both datasets, whereas cross-grained loss functions only perform well on Form-NLU. This difference may stem from the FUNSD being sourced from multiple origins, whereas Form-NLU is a single-source dataset. Coarse-grained loss functions may excel on single-source documents by capturing more prevalent knowledge but might introduce noise when applied to multiple sources. Also, the model demonstrates its most competitive performance by integrating all proposed loss functions (+*Sim*+*Distil*+*Triplet*+*Align*). This highlights how the proposed inter-grained and cross-grained loss functions enhance multi-teacher knowledge distillation in form understanding tasks[4].

5.5.2 Breakdown Result Analysis

As shown in Table 5.6, for the FUNSD dataset, it could be found all Joint-Grained(JG-) frameworks can have a delicate performance on recognising *Question* and *Answer*, but

[4]More loss combination analysis is in Section 5.5.4

Model	Config	Overall	Breakdown		
			Header	Question	Answer
LiLT	Teacher	87.54	55.61	90.20	88.34
LayoutLMv3	Teacher	90.61	66.09	91.60	**92.78**
JG-$\mathcal{E}$	Joint CE	90.45	64.94	91.70	92.67
JG-$\mathcal{D}$	Joint CE	90.48	64.07	91.58	92.73
JG-$\mathcal{E}$&$\mathcal{D}$	Joint CE	90.57	64.66	91.48	92.73
MT-JG-$\mathcal{E}$&$\mathcal{D}$	Joint CE	90.53	61.24	92.40	91.75
	Sim	**91.05**	64.81	92.58	92.46
	Distil	90.90	**66.96**	**92.61**	91.97
	Triplet	90.28	62.44	92.00	91.44
	Align	90.55	63.81	91.82	92.29
	+Sim+Distil +Triplet+Align	90.92	64.22	92.54	92.31

TABLE 5.6. Breakdown results of FUNSD dataset.

decreased in Header classification. This might result from the limited number of *Headers* in the FUNSD, leading to inadequate learning of the fine-grained and coarse-grained *Header* information. Multi-task-oriented inter-grained and coarse-grained functions can increase the performance of *Question* recognition by boosting the knowledge distilling from joint-grained multi-teachers. Especially, inter-grained knowledge distillation methods can achieve around 1% higher than LayoutLMv3. The FUNSD dataset cannot illustrate the benefits of cross-grained loss functions well.

Model	Config	Form-NLU Printed Overall and Breakdown							Form-NLU Handwritten Overall and Breakdown						
		Overall	Sec	Title	F_K	F_V	T_K	T_V	Overall	Sec	Title	F_K	F_V	T_K	T_V
LiLT	Teacher	96.50	98.32	96.97	98.84	96.62	96.57	93.60	91.35	95.39	99.50	94.81	90.67	84.19	89.81
LayoutLMV3	Teacher	95.99	98.45	97.96	97.97	96.73	92.37	92.98	97.39	**99.33**	99.01	**99.85**	98.24	93.95	95.95
JG-$\mathcal{E}$	Joint CE	94.91	**99.66**	98.99	98.11	95.73	90.14	90.31	96.55	**99.33**	99.01	99.42	**98.56**	88.37	94.67
JG-$\mathcal{D}$	Joint CE	95.68	**99.66**	**100.00**	98.55	96.45	91.94	91.10	97.62	**99.33**	99.01	**99.85**	**98.56**	93.02	95.98
JG-$\mathcal{E}$&$\mathcal{D}$	Joint CE	95.93	**99.66**	97.96	97.82	97.18	91.97	92.15	97.62	**99.33**	99.01	**99.85**	98.40	93.74	95.75
MT-JG-$\mathcal{E}$&$\mathcal{D}$	Joint CE	97.21	99.32	98.48	99.57	96.58	97.35	95.06	97.75	97.67	99.50	99.13	97.93	95.55	96.41
	Sim	98.25	99.32	99.49	99.28	97.75	**97.96**	**97.12**	98.09	99.00	**100.00**	99.27	98.25	96.45	96.61
	Distil	98.12	99.32	**100.00**	**99.71**	97.90	97.55	96.30	97.72	97.35	**100.00**	99.13	97.62	95.75	97.07
	Triplet	97.58	99.32	99.49	99.28	97.18	97.55	95.87	97.28	98.00	**100.00**	98.83	97.31	93.90	96.83
	Align	97.24	99.32	98.48	**99.71**	96.57	96.13	95.47	97.42	**99.33**	99.50	99.13	96.85	92.86	97.52
	+Sim+Distil +Triplet+Align	**98.69**	99.32	**100.00**	**99.71**	**99.25**	97.35	**97.12**	**98.39**	98.33	**100.00**	99.56	98.09	**96.94**	**97.75**

TABLE 5.7. Overall and breakdown analysis of Form-NLU printed and handwritten set. The categories of Form-NLU dataset Task A include Section (Sec), Title, Form_Key (F_K), Form_Value (F_V), Table_Key (T_K), Table_Value (T_V).

For Form-NLU printed and handwritten sets, the joint-grained framework and proposed loss functions can effectively improve *Section* (*Sec*) and Title recognition. As the *Title*, *Section* and

FG Teacher	CG Teacher	FUNSD	Form-NLU	
			P	**H**
LLmv3	VBERT	90.19	94.72	96.99
	LXMERT	**90.57**	95.93	<u>97.62</u>
	Transformer	90.22	93.65	95.94
LiLT	VBERT	87.66	**97.65**	90.53
	LXMERT	87.34	96.76	91.18
	Transformer	87.91	97.20	90.58
LLmv3	VBERT&LXMERT	90.42	95.05	97.25
LLmv3 & LiLT	LXMERT	90.39	96.73	97.42
LLmv3&LiLT	**VBERT&LXMERT**	<u>90.53</u>	<u>97.21</u>	**97.75**

TABLE 5.8. Comparison of Performance across Teacher Combinations. FG: Fine-Grained, CG: Coarse-Grained, LLmv3: LayoutLMv3, VBERT: VisualBERT. The best is in **bold**. The second best is <u>underlined</u>. This ablation study is based on only Joint Cross Entropy Loss.

Form_key (*F_K*) are normally located at similar positions for single-source forms, this may demonstrate both joint-grained framework and multi-task loss function could distil knowledge. Additionally, baseline models are not good at recognising table keys and values, especially handwritten sets. As layoutLMv3 is used in the joint-grained framework, the performance of recognising table-related tokens is not good for the joint-learning framework. After integrating multiple teachers, the performance has increased from 91.97% to 97.35% on the printed set. The proposed multi-task loss functions may achieve a higher performance of 97.96%. Significant improvements can also be observed across two test sets across all table-related targets. This illustrates that the joint-grained multi-teacher framework can effectively tackle the limitation of one teacher to generate more comprehensive token representations, and the inter-grained and cross-grained loss could boost the effective knowledge exchange to make the generalisation and robustness of the entire framework.

5.5.3 Effect of Multi-Teachers

Various teacher combinations are analysed to ensure they provide sufficient knowledge for improving joint-grained representations, as depicted in Table 5.8. For fine-grained teachers, since BROS underperforms compared to others, this work only includes the performance of its

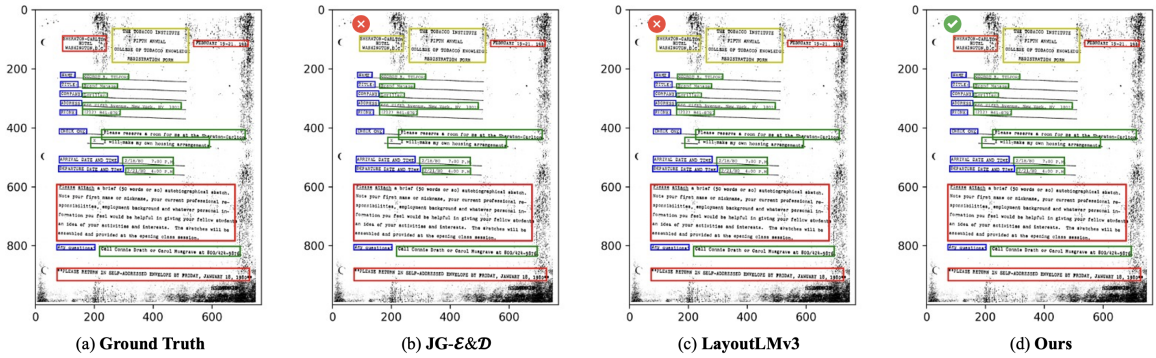
Loss Functions				FUNSD	Form-NLU	
Similarity	Distiling	Triplet	Alignment		***P***	***H***
O	X	X	X	**91.05**	98.25	98.09
X	O	X	X	90.90	98.12	97.72
X	X	O	X	90.28	97.58	97.28
X	X	X	O	90.55	97.24	97.42
O	O	X	X	90.63	98.53	97.22
O	X	O	X	90.51	97.71	97.79
O	X	X	O	90.82	97.80	98.05
X	O	O	X	90.82	98.22	<u>98.35</u>
X	O	X	O	90.83	98.63	97.45
O	O	O	X	90.79	98.56	97.72
O	O	X	O	90.66	**98.72**	97.85
O	O	O	O	<u>90.92</u>	<u>98.69</u>	**98.39**

TABLE 5.9. Performance comparison across loss functions. The best is in **bold**. The second best is <u>underlined</u>.

counterparts. The LayoutLMv3-based joint framework performs better, outperforming LiLT-based by approximately 3% on FUNSD and over 5% on Form-NLU (***H***). This improvement can be attributed to the contextual learning facilitated by visual cues. Notably, LiLT achieves the highest performance on the Form-NLU (***P***), likely due to its well-designed positional-aware pretraining tasks. For coarse-grained teachers, pretrained backbones demonstrate better robustness than randomly initialised Transformers, highlighting the benefits of general domain pretrained knowledge in form understanding tasks. Table 5.8 illustrates multiple teachers cannot always ensure the best performance. However, the robustness of the proposed model is enhanced by capturing more implicit knowledge from cross-grained teachers.

5.5.4 Effect of Loss Functions

To comprehensively investigate the impact of different loss functions and their combinations, Table 5.9 presents various combinations' performance. While employing inter-grained loss individually often proves more effective than using cross-grained loss alone, combining the two losses can enhance knowledge distillation from joint-grained multi-teachers. For instance, concurrently employing distilling(Distil) and Triplet loss improved accuracy from 97.72% to 98.35%. Notably, stacking all proposed loss functions resulted in the best or second-best



(a) **Ground Truth** (b) **JG-$\mathcal{E}\&\mathcal{D}$** (c) **LayoutLMv3** (d) **Ours**

FIGURE 5.2. Visualised prediction results of (a) Ground Truth (b) JG-$\mathcal{E}\&\mathcal{D}$ (c) LayoutLMv3, and (d) 3MVRD on a sample document image of FUNSD dataset. The color code for layout component labels is as follows; Question, Answer, Header, Other. 3MVRD, employing the best loss combination (cross-entropy + similarity) on FUNSD, accurately classified all layout components.

performance across all test sets, showcasing their effectiveness in distilling knowledge from multi-teacher to student models for generating more representative representations. Even though cross-grained Triplet and Alignment losses were ineffective individually, when combined with inter-grained loss, they significantly improved knowledge distillation effectiveness.

5.5.5 Qualitative Analysis: Case Studies

The sample results are visualised for the top 3 - best model with the best configuration, the best baseline *LayoutLMv3* and the second best baseline *JG-$\mathcal{E}\&\mathcal{D}$* of FUNSD in Figure 5.2. It can be noticed that both *LayoutLMv3* and *JG-$\mathcal{E}\&\mathcal{D}$* have wrongly recognised an *Other* (marked by a white cross in red circle), whereas 3MVRD has accurately recognised all document tokens and components.

5.6 Conclusion

A Multimodal Multi-task Multi-Teacher framework is introduced for Visually-Rich form document understanding. The model incorporates *multi-teacher*, *multi-task*, and *multi-loss*, and the results show the robustness in capturing implicit knowledge from multi-teachers for

understanding diverse form document natures, such as scanned, printed, and handwritten. This work may provide valuable insights into leveraging multi-teacher and multi-loss strategies for document understanding research.

CHAPTER 6

Conclusion

This thesis explores the evolving landscape of intelligent understanding of Visually Rich Documents (VRD), highlighting proposing advanced practical solutions and current limitations. To enhance the capabilities and efficacy of Visually Rich Document Understanding (VRDU), this thesis focuses on addressing three critical challenges: neglect of complex VRD structures, lack of multi-page scenarios explorations, and lack of exploration in large model integration.

Chapter 3 introduces Form-NLU, a dataset designed to advance complex form structure understanding and key information extraction, bridging information gaps among stakeholders. Encompassing digital, printed, and handwritten form types, Form-NLU addresses diverse layouts and appearances. A novel model is proposed for efficient key information extraction from forms, supported by a case study on real-world applications and transfer learning evaluations.

Chapter 4 introduces MM-VQA, a pioneering dataset tailored for extracting multimodal information from multi-page VRDs. It proposes a robust model with a joint-grained retrieval architecture to enhance representational capabilities. By leveraging implicit knowledge from VLPMs, this research significantly advances multi-page document understanding, laying a solid foundation for future exploration in this domain.

Chapter 5 introduces the Multimodal Multi-task Multi-Teacher framework in Visually Rich Form Document Understanding (3MVRD). This framework employs different loss functions to distil implicit knowledge from multi-teachers of various grains, enhancing the robustness of document representations. This improvement is particularly vital for handling complex document structures like forms.

Overall, each chapter provides in-depth qualitative and quantitative analyses that illuminate the capabilities of the proposed datasets and the effectiveness of the introduced models. These analyses demonstrate that the proposed approaches effectively address existing research gaps and significantly contribute to the advancement of intelligent document understanding. By tackling real-world demands with innovative solutions, this research marks a significant stride forward in the field of visually rich document analysis.

6.1 Future outlook

With the ongoing advancement of Visually Rich Document Understanding (VRDU) and its growing application across various domains, several potential research directions emerge. These can be summarised into three key areas:

Hierarchical VRD Representation: Current VRDU frameworks primarily concentrate on mono-grained [68, 140], or joint-grained [69] approaches, with a limited exploration into representations that fully acknowledge a document's hierarchical structure and logical correlations. Documents inherently possess a hierarchical structure, where semantic entities are interconnected based on their semantic relationships and structural positioning. For example, an academic paper's hierarchical structure begins with the *Title*, *Abstract*, and *Keywords*, which provide a concise overview of the study. The main body is divided into several first-level sections, such as *Introduction*, *Related Work*, and *Results*, each potentially containing more detailed subsections; for instance, the *Results* section may include both *Qualitative Results* and *Quantitative Results*. While existing pretraining and training methodologies, including self-supervised or supervised approaches, often utilise masked language modelling and multimodal/joint-grained interactive learning, they frequently overlook the importance of hierarchical structures. Consequently, the text representation under the *Qualitative Analysis* subsection fails to understand its semantic role within the broader hierarchical document structure. This oversight can impede the development of hierarchical document representations that are crucial for enhancing retrieval-augmented strategies and supporting various VRDU downstream tasks.

Relational-aware in VRD QA: Exisiting VRD QA has expanded its scope from single-page scenarios [84, 116] to multi-page scenarios [28, 119], encompassing multi-modal document entities. However, existing benchmark datasets and SoTAs primarily focus on retrieving direct answers to questions, often overlooking the contextual information of the target extractive answer and the process of locating the answer. For instance, when a question inquires, "What kinds of backbone models are used in this framework?" within a deep learning model paper, a human would typically need to locate the "Methodology" or "Model" section and check the subsection titles to find the possible answer location. Sometimes, figures depicting the model architecture are also helpful in locating the answer. Therefore, the answer to a question may not be merely an isolated entity or in-line text. The process of answer locating, along with the contextual and correlated content of direct answers, is crucial for understanding the VRD QA process and acquiring more comprehensive answers.

Domain Adaptation in VRDU: Existing Visually Rich Document Understanding (VRDU) frameworks require extensive manual annotation to train or fine-tune models for specific VRDU tasks. In practical applications, obtaining a large volume of well-annotated data is cost-ineffective. Therefore, exploring effective ways to leverage noisy data or domain adaptation methods is crucial. Understanding visually rich documents necessitates comprehending the structure and content of the given domain document collection. Unstructured or semi-structured documents typically require numerous annotations to acquire layout information for feature extraction, as well as to extract key information and question-answer pairs, which demands significant human effort and expert knowledge. The advancement of deep learning has led to the development of off-the-shelf tools and large-scale trained models that can automatically generate structural and content annotations. For example, OCR tools, object detection models, and PDF parsers can generate both fine-grained and coarse-grained information, while LLMs [87] and MLLMs [72] can extract key information and generate question-answer pairs. Leveraging automatically generated noisy training data allows for the capture of more domain-specific knowledge, effectively improving the models' performance in real-world applications with reduced labour, resources, and time consumption.

Bibliography

[1] Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., & Manmatha, R. (2021). Docformer: End-to-end transformer for document understanding. *Proceedings of the IEEE/CVF international conference on computer vision*, 993–1003.

[2] Appalaraju, S., Tang, P., Dong, Q., Sankaran, N., Zhou, Y., & Manmatha, R. (2024). Docformerv2: Local features for document understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(2), 709–718.

[3] Bai, H., Liu, Z., Meng, X., Wentao, L., Liu, S., Luo, Y., Xie, N., Zheng, R., Wang, L., Hou, L., et al. (2023). Wukong-reader: Multi-modal pre-training for fine-grained visual document understanding. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13386–13401.

[4] Bao, H., Dong, L., Piao, S., & Wei, F. (2021). Beit: Bert pre-training of image transformers. *International Conference on Learning Representations*.

[5] Biten, A. F., Litman, R., Xie, Y., Appalaraju, S., & Manmatha, R. (2022). Latr: Layout-aware transformer for scene-text vqa. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16548–16558.

[6] Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Mathew, M., Jawahar, C., Valveny, E., & Karatzas, D. (2019). Icdar 2019 competition on scene text visual question answering. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1563–1570.

[7] Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., & Karatzas, D. (2019). Scene text visual question answering. *Proceedings of the IEEE/CVF international conference on computer vision*, 4291–4301.

[8] Blau, T., Fogel, S., Ronen, R., Golts, A., Ganz, R., Ben Avraham, E., Aberdam, A., Tsiper, S., & Litman, R. (2024). Gram: Global reasoning for multi-page vqa. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15598–15607.

[9] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

[10] Cao, F., Han, S. C., Long, S., Xu, C., & Poon, J. (2022). Understanding attention for vision-and-language tasks. *Proceedings of the 29th International Conference on Computational Linguistics*, 3438–3453.

[11] Cao, H., Bao, C., Liu, C., Chen, H., Yin, K., Liu, H., Liu, Y., Jiang, D., & Sun, X. (2023). Attention where it matters: Rethinking visual document understanding with selective region concentration. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19517–19527.

[12] Cao, P., Wang, Y., Zhang, Q., & Meng, Z. (2023). Genkie: Robust generative multimodal document key information extraction. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14702–14713.

[13] Chang, M., da Silva Rosa, R., & Ng, W. (2016). The informativeness of substantial shareholder trading in the lead up to a takeover bid. *Asian Finance Association (AsianFA) 2016 Conference*.

[14] Chen, J., Dai, H., Dai, B., Zhang, A., & Wei, W. (2023). On task-personalized multimodal few-shot learning for visually-rich document entity retrieval. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9006–9025.

[15] Chen, J., Lv, T., Cui, L., Zhang, C., & Wei, F. (2022). Xdoc: Unified pre-training for cross-format document understanding. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 1006–1016.

[16] Chen, X., Zhao, Z., Chen, L., Ji, J., Zhang, D., Luo, A., Xiong, Y., & Yu, K. (2021). Websrc: A dataset for web-based structural reading comprehension. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4173–4185.

[17] Chen, Z., Liu, G., Zhang, B.-W., Yang, Q., & Wu, L. (2023). Altclip: Altering the language encoder in clip for extended language capabilities. *Findings of the Association for Computational Linguistics: ACL 2023*, 8666–8682.

[18] Cheng, M., Qiu, M., Shi, X., Huang, J., & Lin, W. (2020). One-shot text field labeling using attention and belief propagation for structure information extraction. *Proceedings of the 28th ACM International Conference on Multimedia*, 340–348.

[19] Chi, Z., Huang, H., Xu, H.-D., Yu, H., Yin, W., & Mao, X.-L. (2019). Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*.

[20] Chu, X., Tian, Z., Zhang, B., Wang, X., & Shen, C. (2022). Conditional positional encodings for vision transformers. *The Eleventh International Conference on Learning Representations*.

[21] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

[22] Davis, B., Morse, B., Price, B., Tensmeyer, C., Wigington, C., & Morariu, V. (2022). End-to-end document recognition and understanding with dessurt. *European Conference on Computer Vision*, 280–296.

[23] Denk, T. I., & Reisswig, C. (2019). Bertgrid: Contextualized embedding for 2d document representation and understanding. *Workshop on Document Intelligence at NeurIPS 2019*.

[24] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *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)*, 4171–4186.

[25] Ding, Y., Huang, Z., Wang, R., Zhang, Y., Chen, X., Ma, Y., Chung, H., & Han, S. C. (2022). V-doc: Visual questions answers with documents. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21492–21498.

[26] Ding, Y., Long, S., Huang, J., Ren, K., Luo, X., Chung, H., & Han, S. C. (2023). Form-nlu: Dataset for the form natural language understanding. *Proceedings of the 46th*

International ACM SIGIR Conference on Research and Development in Information Retrieval, 2807–2816.

[27] Ding, Y., Luo, S., Chung, H., & Han, S. C. (2023). Pdf-vqa: A new dataset for real-world vqa on pdf documents. *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*, 585–601.

[28] Ding, Y., Ren, K., Huang, J., Luo, S., & Han, S. C. (2024). Mmvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*.

[29] Ding, Y., Vaiani, L., Han, C., Lee, J., Garza, P., Poon, J., & Cagliero, L. (2024). 3mvrd: Multimodal multi-task multi-teacher visually-rich form document understanding. *arXiv preprint arXiv:2402.17983*.

[30] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

[31] Du, Q., Wang, Q., Li, K., Tian, J., Xiao, L., & Jin, Y. (2022). Calm: Commen-sense knowledge augmentation for document image understanding. *Proceedings of the 30th ACM International Conference on Multimedia*, 3282–3290.

[32] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in neural information processing systems*, *30*.

[33] Gu, J., Kuen, J., Morariu, V. I., Zhao, H., Jain, R., Barmpalios, N., Nenkova, A., & Sun, T. (2021). Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, *34*, 39–50.

[34] Gu, Z., Meng, C., Wang, K., Lan, J., Wang, W., Gu, M., & Zhang, L. (2022). Xy-layoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4583–4592.

[35] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International conference on machine learning*, 1321–1330.

[36] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. *International conference on machine learning*, 3929–3938.

[37] Harley, A. W., Ufkes, A., & Derpanis, K. G. (2015). Evaluation of deep convolutional nets for document image classification and retrieval. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 991–995.

[38] He, J., Wang, L., Hu, Y., Liu, N., Liu, H., Xu, X., & Shen, H. T. (2023). Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19485–19494.

[39] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

[40] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[41] Hofstätter, S., Chen, J., Raman, K., & Zamani, H. (2023). Fid-light: Efficient and effective retrieval-augmented text generation. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1437–1447.

[42] Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., & Park, S. (2022). Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 10767–10775.

[43] Hu, R., Singh, A., Darrell, T., & Rohrbach, M. (2020). Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9992–10002.

[44] Hu, Z., Iscen, A., Sun, C., Wang, Z., Chang, K.-W., Sun, Y., Schmid, C., Ross, D. A., & Fathi, A. (2023). Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23369–23379.

[45] Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). Layoutlmv3: Pre-training for document ai with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*, 4083–4091.

[46] Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., & Jawahar, C. (2019a). Icdar 2019 robust reading challenge on scanned receipts ocr and information extraction. *International Conference on Document Analysis Recognition*.

[47] Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., & Jawahar, C. (2019b). Icdar2019 competition on scanned receipt ocr and information extraction. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1516–1520.

[48] Jaeger, P. F., Lüth, C. T., Klein, L., & Bungert, T. J. (2022). A call to reflect on evaluation practices for failure detection in image classification. *The Eleventh International Conference on Learning Representations*.

[49] Jaume, G., Ekenel, H. K., & Thiran, J.-P. (2019). Funsd: A dataset for form understanding in noisy scanned documents. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, *2*, 1–6.

[50] Jegou, H., Douze, M., & Schmid, C. (2010). Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, *33*(1), 117–128.

[51] Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, *8*, 64–77.

[52] Kang, L., Tito, R., Valveny, E., & Karatzas, D. (2024). Multi-page document visual question answering using self-attention scoring mechanism. *arXiv preprint arXiv:2404.19024*.

[53] Katti, A. R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., & Faddoul, J. B. (2018). Chargrid: Towards understanding 2d documents. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4459–4469.

[54] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2022). Ocr-free document understanding transformer. *European Conference on Computer Vision*, 498–517.

[55] Kim, W., Son, B., & Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. *International Conference on Machine Learning*, 5583–5594.

[56] Kim, Y. (2014, October). Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1746–1751). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181

[57] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260–270.

[58] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

[59] Lee, C.-Y., Li, C.-L., Dozat, T., Perot, V., Su, G., Hua, N., Ainslie, J., Wang, R., Fujii, Y., & Pfister, T. (2022). Formnet: Structural encoding beyond sequential modeling in form document information extraction. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3735–3754.

[60] Lee, C.-Y., Li, C.-L., Zhang, H., Dozat, T., Perot, V., Su, G., Zhang, X., Sohn, K., Glushnev, N., Wang, R., et al. (2023). Formnetv2: Multimodal graph contrastive learning for form document information extraction. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9011–9026.

[61] Lee, K., Joshi, M., Turc, I. R., Hu, H., Liu, F., Eisenschlos, J. M., Khandelwal, U., Shaw, P., Chang, M.-W., & Toutanova, K. (2023). Pix2struct: Screenshot parsing as pretraining for visual language understanding. *International Conference on Machine Learning*, 18893–18912.

[62] Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., & Heard, J. (2006). Building a test collection for complex document information processing. *Proceedings*

of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 665–666.

[63] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

[64] Li, C., Bi, B., Yan, M., Wang, W., Huang, S., Huang, F., & Si, L. (2021). Structurallm: Structural pre-training for form understanding. *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)*, 6309–6318.

[65] Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., & Wei, F. (2022). Dit: Self-supervised pre-training for document image transformer. *Proceedings of the 30th ACM International Conference on Multimedia*, 3530–3539.

[66] Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

[67] Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., & Zhou, M. (2020). Docbank: A benchmark dataset for document layout analysis. *Proceedings of the 28th International Conference on Computational Linguistics*, 949–960.

[68] Li, P., Gu, J., Kuen, J., Morariu, V. I., Zhao, H., Jain, R., Manjunatha, V., & Liu, H. (2021). Selfdoc: Self-supervised document representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5652–5660.

[69] Li, Y., Qian, Y., Yu, Y., Qin, X., Zhang, C., Liu, Y., Yao, K., Han, J., Liu, J., & Ding, E. (2021). Structext: Structured text understanding with multi-modal transformers. *Proceedings of the 29th ACM International Conference on Multimedia*, 1912–1920.

[70] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

[71] Liu, C., Yin, K., Cao, H., Jiang, X., Li, X., Liu, Y., Jiang, D., Sun, X., & Xu, L. (2024). Hrvda: High-resolution visual document assistant. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15534–15545.

[72] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning.

[73] Liu, X., Gao, F., Zhang, Q., & Zhao, H. (2019). Graph convolution for multimodal information extraction from visually rich documents. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, 32–39.

[74] Liu, X., Tang, W., Ni, X., Lu, J., Zhao, R., Li, Z., & Tan, F. (2023). What large language models bring to text-rich vqa? *arXiv preprint arXiv:2311.07306*.

[75] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[76] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

[77] Long, S., Cao, F., Han, S. C., & Yang, H. (2022, July). Vision-and-language pretrained models: A survey [Survey Track]. In L. D. Raedt (Ed.), *Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI-22* (pp. 5530–5537). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/773

[78] Luo, C., Cheng, C., Zheng, Q., & Yao, C. (2023). Geolayoutlm: Geometric pretraining for visual information extraction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7092–7101.

[79] Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., & Yao, C. (2024). Layoutllm: Layout instruction tuning with large language models for document understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15630–15640.

[80] Luo, S., Ding, Y., Long, S., Poon, J., & Han, S. C. (2022). Doc-gcn: Heterogeneous graph convolutional networks for document layout analysis. *Proceedings of the 29th International Conference on Computational Linguistics*, 2906–2916.

[81] Ma, J., Du, J., Hu, P., Zhang, Z., Zhang, J., Zhu, H., & Liu, C. (2023). Hrdoc: Dataset and baseline method toward hierarchical reconstruction of document structures. *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*.

[82] Majumder, B. P., Potti, N., Tata, S., Wendt, J. B., Zhao, Q., & Najork, M. (2020). Representation learning for information extraction from form-like documents. *proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 6495–6504.

[83] Mao, Z., Bai, H., Hou, L., Shang, L., Jiang, X., Liu, Q., & Wong, K.-F. (2024). Visually guided generative text-layout pre-training for document intelligence. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4713–4730.

[84] Mathew, M., Karatzas, D., & Jawahar, C. (2021). Docvqa: A dataset for vqa on document images. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2200–2209.

[85] Nagy, G., & Seth, S. C. (1984). Hierarchical representation of optically scanned documents.

[86] Oliveira, D. A. B., & Viana, M. P. (2017). Fast cnn-based document layout analysis. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 1173–1180.

[87] OpenAI. (2023). Chatgpt: A conversational agent. https://www.openai.com/chatgpt

[88] Oreshkin, B., Rodríguez López, P., & Lacoste, A. (2018). Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, *31*.

[89] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

[90] Palm, R. B., Laws, F., & Winther, O. (2019). Attend, copy, parse end-to-end information extraction from documents. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 329–336.

[91] Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., & Lee, H. (2019). Cord: A consolidated receipt dataset for post-ocr parsing. *Workshop on Document Intelligence at NeurIPS 2019*.

[92] Perot, V., Kang, K., Luisier, F., Su, G., Sun, X., Boppana, R. S., Wang, Z., Mu, J., Zhang, H., & Hua, N. (2023). Lmdx: Language model-based document information extraction and localization. *arXiv preprint arXiv:2309.10952*.

[93] Phuong, M., & Lampert, C. (2019). Towards understanding knowledge distillation. *International conference on machine learning*, 5142–5151.

[94] Powalski, R., Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka, M., & Pałka, G. (2021). Going full-tilt boogie on document understanding with text-image-layout transformer. *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, 732–747.

[95] Press, O., Smith, N., & Lewis, M. (2021). Train short, test long: Attention with linear biases enables input length extrapolation. *International Conference on Learning Representations*.

[96] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

[97] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020a). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1–67.

[98] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020b). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1–67.

[99] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

[100] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. *International conference on machine learning*, 8821–8831.

[101] Rausch, J., Martinez, O., Bissig, F., Zhang, C., & Feuerriegel, S. (2021). Docparser: Hierarchical document structure parsing from renderings. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(5), 4328–4338.

[102] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

[103] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

[104] Rusinol, M., Benkhelfallah, T., & Poulain dAndecy, V. (2013). Field extraction from administrative documents by incremental structural templates. *2013 12th International Conference on Document Analysis and Recognition*, 1100–1104.

[105] Seki, M., Fujio, M., Nagasaki, T., Shinjo, H., & Marukawa, K. (2007). Information management system using structure analysis of paper/electronic documents and its applications. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, *2*, 689–693.

[106] Shi, D., Liu, S., Du, J., & Zhu, H. (2023). Layoutgcn: A lightweight architecture for visually rich document understanding. *International Conference on Document Analysis and Recognition*, 149–165.

[107] Šimsa, Š., Šulc, M., Uřičář, M., Patel, Y., Hamdi, A., Kocián, M., Skalickỳ, M., Matas, J., Doucet, A., Coustaty, M., et al. (2023). Docile benchmark for document information localization and extraction. *International Conference on Document Analysis and Recognition*, 147–166.

[108] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in neural information processing systems*, *30*.

[109] Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, *29*.

[110] Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI conference on artificial intelligence*, *31*(1).

[111] Srivastava, Y., Murali, V., Dubey, S. R., & Mukherjee, S. (2020). Visual question answering using deep learning: A survey and performance analysis. *Computer Vision and Image Processing - 5th International Conference,CVIP 2020*, *1377*, 75–86.

[112] Stanisławek, T., Graliński, F., Wróblewska, A., Lipiński, D., Kaliska, A., Rosalska, P., Topolski, B., & Biecek, P. (2021a). Kleister: Key information extraction datasets involving long documents with complex layouts. *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I*, 564–579.

[113] Stanisławek, T., Graliński, F., Wróblewska, A., Lipiński, D., Kaliska, A., Rosalska, P., Topolski, B., & Biecek, P. (2021b). Kleister: Key information extraction datasets involving long documents with complex layouts. *International Conference on Document Analysis and Recognition*, 564–579.

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

[115] Tanaka, R., Nishida, K., Nishida, K., Hasegawa, T., Saito, I., & Saito, K. (2023). Slidevqa: A dataset for document visual question answering on multiple images. *arXiv preprint arXiv:2301.04883*.

[116] Tanaka, R., Nishida, K., & Yoshida, S. (2021). Visualmrc: Machine reading comprehension on document images. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(15), 13878–13888.

[117] Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C., & Bansal, M. (2023). Unifying vision, text, and layout for universal document processing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19254–19264.

[118] Tito, R., Karatzas, D., & Valveny, E. (2021). Document collection visual question answering. *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, 778–792.

[119] Tito, R., Karatzas, D., & Valveny, E. (2023). Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, *144*, 109834.

[120] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[121] Tu, Y., Guo, Y., Chen, H., & Tang, J. (2023). Layoutmask: Enhance text-layout interaction in multi-modal pre-training for document understanding. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15200–15212.

[122] Van Landeghem, J., Tito, R., Borchmann, Ł., Pietruszka, M., Joziak, P., Powalski, R., Jurkiewicz, D., Coustaty, M., Anckaert, B., Valveny, E., et al. (2023). Document understanding dataset and evaluation (dude). *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19528–19540.

[123] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[124] Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. *Advances in neural information processing systems*, *28*.

[125] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *International Conference on Learning Representations*.

[126] Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. (2020). Cspnet: A new backbone that can enhance learning capability of cnn. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 390–391.

[127] Wang, J., Jin, L., & Ding, K. (2022). Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7747–7757.

[128] Wang, J., Liu, C., Jin, L., Tang, G., Zhang, J., Zhang, S., Wang, Q., Wu, Y., & Cai, M. (2021). Towards robust visual information extraction in real world: New dataset and novel solution. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(4), 2738–2745.

[129] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., & Yang, H. (2022). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *International Conference on Machine Learning*, 23318–23340.

[130] Wang, Z., Zhang, Z., Devlin, J., Lee, C.-Y., Su, G., Zhang, H., Dy, J., Perot, V., & Pfister, T. (2023). Queryform: A simple zero-shot form entity query framework. *Findings of the Association for Computational Linguistics: ACL 2023*, 4146–4159.

[131] Wang, Z., Gu, J., Tensmeyer, C., Barmpalios, N., Nenkova, A., Sun, T., Shang, J., & Morariu, V. (2022). Mgdoc: Pre-training with multi-granular hierarchy for document image understanding. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3984–3993.

[132] Wang, Z., & Shang, J. (2022). Towards few-shot entity recognition in document images: A label-aware sequence-to-sequence framework. *Findings of the Association for Computational Linguistics: ACL 2022*, 4174–4186.

[133] Wang, Z., Xu, Y., Cui, L., Shang, J., & Wei, F. (2021). Layoutreader: Pre-training of text and layout for reading order detection. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4735–4744.

[134] Wang, Z., Zhou, Y., Wei, W., Lee, C.-Y., & Tata, S. (2023). Vrdu: A benchmark for visually-rich document understanding. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5184–5193.

[135] Watanabe, T., Luo, Q., & Sugie, N. (1995). Layout recognition of multi-kinds of table-form documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(4), 432–445.

[136] Wei, M., He, Y., & Zhang, Q. (2020). Robust layout-aware ie for visually rich documents with pre-trained language models. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2367–2376.

[137] Wu, X., Zheng, D., Wang, R., Sun, J., Hu, M., Feng, F., Wang, X., Jiang, H., & Yang, F. (2022). A region-based document vqa. *Proceedings of the 30th ACM International Conference on Multimedia*, 4909–4920.

[138] Xu, X., Wu, C., Rosenman, S., Lal, V., Che, W., & Duan, N. (2023). Bridgetower: Building bridges between encoders in vision-language representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(9), 10637–10647.

[139] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al. (2021). Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *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)*, 2579–2591.

[140] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). Layoutlm: Pre-training of text and layout for document image understanding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1192–1200.

[141] Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., & Wei, F. (2022). Xfund: A benchmark dataset for multilingual visually rich form understanding. *Findings of the Association for Computational Linguistics: ACL 2022*, 3214–3224.

[142] Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., & Xu, C. (2021). Filip: Fine-grained interactive language-image pre-training. *International Conference on Learning Representations*.

[143] Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., & Yih, W.-T. (2023). Retrieval-augmented multimodal language modeling. *International Conference on Machine Learning*, 39755–39769.

[144] Ye, Y., Hui, B., Yang, M., Li, B., Huang, F., & Li, Y. (2023). Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 174–184.

[145] Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

[146] Yu, W., Lu, N., Qi, X., Gong, P., & Xiao, R. (2021). Pick: Processing key information extraction from documents using improved graph learning-convolutional networks. *2020 25th International Conference on Pattern Recognition (ICPR)*, 4363–4370.

[147] Yu, Y., Li, Y., Zhang, C., Zhang, X., Guo, Z., Qin, X., Yao, K., Han, J., Ding, E., & Wang, J. (2022). Structextv2: Masked visual-textual prediction for document image pre-training. *The Eleventh International Conference on Learning Representations*.

[148] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. (2020). Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, *33*, 17283–17297.

[149] Zhai, M., Li, Y., Qin, X., Yi, C., Xie, Q., Zhang, C., Yao, K., Wu, Y., & Jia, Y. (2023). Fast-structext: An efficient hourglass transformer with modality-guided dynamic token merge for document understanding. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 5269–5277.

[150] Zhang, P., Xu, Y., Cheng, Z., Pu, S., Lu, J., Qiao, L., Niu, Y., & Wu, F. (2020). Trie: End-to-end text reading and information extraction for document understanding. *Proceedings of the 28th ACM International Conference on Multimedia*, 1413–1422.

[151] Zhang, Y., Bo, Z., Wang, R., Cao, J., Li, C., & Bao, Z. (2021). Entity relation extraction as dependency parsing in visually rich documents. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2759–2768.

[152] Zhao, R., Chen, H., Wang, W., Jiao, F., Do, X. L., Qin, C., Ding, B., Guo, X., Li, M., Li, X., et al. (2023). Retrieving multimodal information for augmented generation: A survey. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4736–4756.

[153] Zhao, X., Niu, E., Wu, Z., & Wang, X. (2019). Cutie: Learning to understand documents with convolutional universal text information extractor. *arXiv preprint arXiv:1903.12363*.

[154] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, *36*.

[155] Zheng, X., Burdick, D., Popa, L., Zhong, X., & Wang, N. X. R. (2021). Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 697–706.

[156] Zhong, X., ShafieiBavani, E., & Jimeno Yepes, A. (2020). Image-based table recognition: Data, model, and evaluation. *European Conference on Computer Vision*, 564–580.

[157] Zhong, X., Tang, J., & Yepes, A. J. (2019). Publaynet: Largest dataset ever for document layout analysis. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1015–1022.

[158] Zhu, F., Lei, W., Feng, F., Wang, C., Zhang, H., & Chua, T.-S. (2022). Towards complex document understanding by discrete reasoning. *Proceedings of the 30th ACM International Conference on Multimedia*, 4857–4866.

[159] Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., & Chua, T.-S. (2021). Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *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), 3277–3287.

[160] Zhu, Z., Yu, J., Wang, Y., Sun, Y., Hu, Y., & Wu, Q. (2021). Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 1097–1103.