VRD-IU: Lessons from Visually Rich Document Intelligence and Understanding

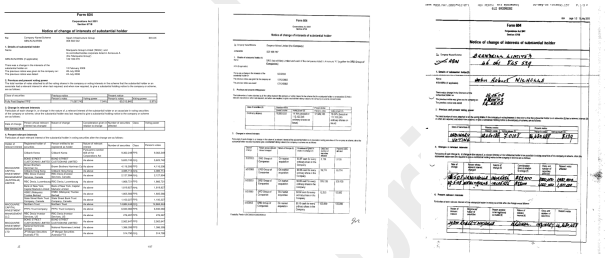
Yihao Ding[1,2], **Soyeon Caren Han**[1,2], **Yan Li**[2] and **Josiah Poon**[2]
[1]The University of Melbourne
[2]The University of Sydney
{yihao.ding, josiah.poon}@sydney.edu.au, yali3816@uni.sydney.edu.au caren.han@unimelb.edu.au

Abstract

Visually Rich Document Understanding (VRDU) has emerged as a critical field in document intelligence, enabling automated extraction of key information from complex documents across domains such as medical, financial, and educational applications. However, form-like documents pose unique challenges due to their complex layouts, multi-stakeholder involvement, and high structural variability. Addressing these issues, the VRD-IU Competition was introduced, focusing on extracting and localizing key information from multi-format forms within the Form-NLU dataset, which includes digital, printed, and handwritten documents. This paper presents insights from the competition, which featured two tracks: Track A, emphasizing entity-based key information retrieval, and Track B, targeting end-to-end key information localization from raw document images. With over 20 participating teams, the competition showcased various state-of-the-art methodologies, including hierarchical decomposition, transformer-based retrieval, multimodal feature fusion, and advanced object detection techniques. The top-performing models set new benchmarks in VRDU, providing valuable insights into document intelligence.

1 Introduction

Visually Rich Document Understanding (VRDU) has garnered significant attention across various domains, including medical [Ding *et al.*, 2024b], receipts [Park *et al.*, 2019; Huang *et al.*, 2019], financial [Stanisławek *et al.*, 2021], and educational [Wang *et al.*, 2021] fields, becoming a pivotal approach for recording, preserving, and sharing information. Recent advancements in deep learning, particularly self-supervised pretrained transformers [Huang *et al.*, 2022; Tang *et al.*, 2023; Xu *et al.*, 2021] and large language models [Xue *et al.*, 2024; OpenAI, 2024; Liu *et al.*, 2024], have demonstrated notable success in this domain. However, form-like documents present unique challenges due to their involvement of multiple parties (designers, deliverers, and fillers), complex layouts, and higher uncertainties.



(a) Digital (b) Printed (c) Handwritten

Figure 1: Digital, Printed and Handwritten Form Samples from Form-NLU dataset.

These factors create significant barriers to directly applying existing methods to practical tasks [Ding *et al.*, 2024c; Ding *et al.*, 2024a].

To address the challenges caused by form-like documents, we introduce the VRD-IU Competition, which focuses on extracting key information from multi-format forms within the Form-NLU dataset [Ding *et al.*, 2023], encompassing digital, printed, and handwritten documents, as illustrated in Figure 1. The competition aims to tackle the challenges associated with the diverse and complex nature of such documents, which often involve multiple stakeholders and contain essential information that is difficult to extract. By offering two distinct tracks tailored to different skill levels, it accommodates participants ranging from newcomers to seasoned professionals in the field of deep learning. This initiative not only fosters innovation in document understanding and accelerates advancements in information extraction techniques but also aims to engage a broader community, encouraging innovators to contribute to the evolution of efficient information analysis methodologies.

More than 20 teams participated in the VRD-IU competition, with the top 3 teams in Track A and the top 2 teams in Track B submitting their code and abstract papers to share their proposed solutions. Various methods were proposed, focusing on achieving more comprehensive document understanding. In Track A, the methods mainly focused on acquiring more representative entity features, while in Track B, end-to-end frameworks were proposed to enable the framework to understand both the semantic and structural aspects of form images to extract and locate the relevant information.

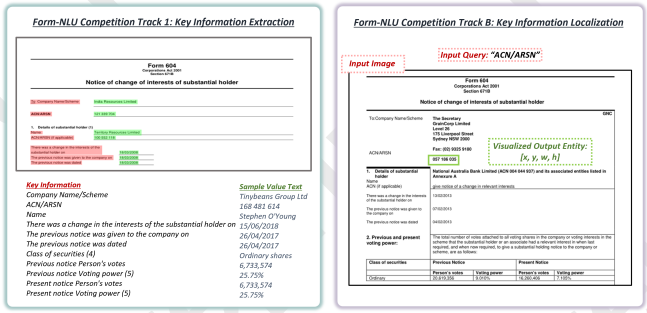


Figure 2: VRD-IU Competition Track A and Track B Samples.

The proposed methods achieved new state-of-the-art performance and paved the way for future work in complex form understanding direction.

The contributions of this paper are as follows:

1) To provide a redefined task definition for the Form-NLU dataset, focusing on Key Information Localization (Track-B), which involves both extracting and localizing target information from input form images.

2) To introduce the state-of-the-art approaches proposed by competition participants, along with a qualitative analysis of these methods.

3) To offer key insights into future directions for VRDU understanding, highlighting potential areas for improvement and innovation.

2 VRD-IU Competition Task Definition

The adopted dataset for the VRD-IU competition is derived from the Form-NLU dataset. Unlike the original tasks, which emphasize both form structure and content understanding, this competition specifically targets the extraction of key information from complex form documents. To address varying levels of task complexity, two distinct tracks are introduced. Track A aligns with Form-NLU Task B, focusing on retrieving key information from provided semantic entities within the document. In contrast, Track B presents a more challenging scenario, offering only the form image and query without any prior information about the document structure. The detailed definitions of these tracks are provided to support future research in this area.

Track A: Form Key Information Extraction [1] Track A involves developing deep learning-based retrieval models to extract specific form components in response to a given key query. Human-annotated bounding box coordinates representing the semantic entities of input form documents are provided. The task requires accurate identification and localization of the target entity based on the query. Performance is evaluated using the **F1-Score**, following the evaluation methodology defined in Form-NLU Task B.

Track B: Form Key Information Localization [2] Track B focuses on designing an end-to-end framework to predict bounding box coordinates of form components directly from input document images based on a given key query. Unlike Track A, ground truth bounding boxes for semantic entities are not provided; the inputs consist solely of form images and query keys. Evaluation is based on the **Mean Average Precision (MAP)** of the predicted bounding boxes, reflecting both accuracy and localization effectiveness.

[1]The dataset used for Track A can be found via https://www.kaggle.com/competitions/vrd-iu2024-tracka

[2]The dataset used for Track B can be found via https://www.kaggle.com/competitions/vrd-iu2024-trackb

3 Methods

3.1 Methods for Track-A

To effectively retrieve key information from domain-specific semantic entities based on user queries, several solutions leverage enhanced document understanding through pretrained transformers. These approaches incorporate advanced components and strategies hierarchical decomposition, token classification, and multimodal feature fusion, to better comprehend the structure and semantics of forms, enabling more accurate and efficient extraction of target information.

Key Information Extraction with Hierarchical Layout Decomposition, proposed by Choi et al. (**Team rb-ai**, *Top-1 Team in Track-A*), utilizes prior template knowledge from the Form-NLU dataset to segment form images into top, middle, and bottom sections. Then, section-specific KIE models are then applied for key information extraction. To address misclassification issues within each section, heuristic post-processing techniques are employed. The approach uses a pretrained transformer-based YOLOv8 as an object detector for section segmentation, while LayoutLMv3 [Huang *et al.*, 2022] is fine-tuned as the KIE backbone for each segmented section.

Redefining Information Extraction from Visually Rich Documents as Token Classification is introduced by Song et.al (**Team vipski**, *Top-2 Team in Track A*) to convert the entity retrieving task as a sequence labelling task with maintaining the original resolution and aspect ratios of document images, the approach reduces visual distortion, enhancing model accuracy. After fine-tuning LayoutLM-v3 with a token classifier, the proposed method achieve cutting-edge performance.

Transformer-Based Visual Feature and Textual Feature Fusion Model for Document Key Information Extraction, proposed by Kim et al. (**Team gcu**, *Top-3 Team in Track A*), aims to enhance document semantic entity representation. RoI visual and textual features are separately processed through a transformer-based vision-language encoder. A cross-encoder then strengthens the correlations between questions and entities. The combined textual and visual representations are fed into a classifier for final information retrieval.

3.2 Methods for Track B

Different from Track A providing the bounding box coordinate of each document semantic entity, Track B requires the model can detect the exact object (semantic entities) from the document image. To realize this, except for the well-designed object detection approaches, some domain-adaptation strategies are adopted.

Key Information LocalizationWith Hierarchical Ensemble is introduced for key information localization by Choi et al. (**Team rb-ai**, *Top-1 Team in Track-A*). The approach uses YOLO [Redmon *et al.*, 2016] and RT-DETR [Zhao *et al.*, 2024] models as baselines and integrates their predictions through a global ensemble and a layout-specific ensemble. Additionally, a layout-based post-editor refines predictions using text recognition and layout-specific detectors to correct mispredictions.

Enhancing Document Key Information Localization Through Data Augmentation is proposed by Dai et al. (**Team chatdy**, *Top-3 Team in Track-A*), which divides the whole workflow to document augmentation phase and an object detection phase. The augmentation phase uses the Augraphy library to mimic the appearance of handwritten documents from digital ones, emplointoying techniques like InkBleed, Letterpress, and JPEG compression to simulate handwritten characteristics. The object detection phase utilizes pretrained VRDU models like DiT [Li *et al.*, 2022] and LayoutLMv3 with Faster R-CNN [Ren *et al.*, 2015] and Mask R-CNN [He *et al.*, 2017] frameworks. The key innovation lies in the augmentation strategy, which enhances the model's ability to generalize to handwritten documents, even when trained only on digital data.

4 Competition Results

The VRD-IU competition attracted significant participation from teams employing diverse methodologies to tackle the key information extraction and localization challenges. Track A and B's evaluation results reveal notable trends in model performance and methodological efficacy.

Track A: Form Key Information Extraction

The results of the top-performing teams in Track A are shown in Table 1. The results demonstrate that pretrained transformer-based approaches, enhanced with hierarchical decomposition and multimodal feature fusion, are highly effective in extracting semantic entities from form documents.

Track B: Form Key Information Localization

Performance in Track B was evaluated using Mean Average Precision (MAP), reflecting both detection accuracy and localization effectiveness, shown in Table 2. Track B posed a greater challenge due to the lack of predefined bounding box annotations. The results indicate a clear performance gap compared to Track A, underscoring the complexity of end-to-end document understanding directly from form images.

Rank	Team Name	Members	Private (%)	Public (%)
1	**rb-ai**	3	**100**	**100**
2	**vipski**	2	97.93	**100**
3	**gcu**	4	96.95	99.67
4	MKAZ	1	93.07	99.57
5	chatdy	1	92.57	99.57
6	Play4fun	1	90.04	99.08
7	Team G.AI	3	80.17	97.78
8	dagmbari	1	78.99	95.55
9	jaehongleee	1	78.87	93.74
10	zuo-zou	1	48.45	80.09
11	DualMonitor	2	42.23	64.13
12	rlawnsxo	1	41.45	54.84

Table 1: Track A Key Information Extraction Results.

Rank	Team Name	Members	Private (%)	Public (%)
1	**rb-ai**	4	**65.79**	**99.70**
2	**oaths11**	3	56.97	80.97
3	**chatdy**	2	55.68	80.49
4	gcu	1	43.93	79.18
5	BiaoXup	1	18.12	13.63
6	Shuo Yang	1	17.06	13.03
7	VRD IU	1	16.80	11.66
8	eunyi lyou	1	0.23	4.88

Table 2: Track B Key Information Localisation Results

5 Key Insights

The competition provided valuable insights into document understanding tasks, particularly within visually rich document (VRD) contexts:

1. Hierarchical Decomposition and Post-Processing Improve Accuracy Track A results highlighted that segmenting form structures into hierarchical components significantly enhances extraction accuracy. Techniques such as section-wise key information extraction (as seen in rb-ai's approach) help mitigate structural misclassification and boost precision.

2. Pretrained Transformers Excel in Structured Document Tasks Methods leveraging LayoutLMv3 and other pretrained vision-language transformers outperformed conventional models, reaffirming the effectiveness of multimodal pretraining for VRD tasks. Fine-tuning on domain-specific data further improved entity extraction capabilities.

3. Object Detection Remains a Challenge for Form Localization Track B results highlighted the difficulty of localizing key information without predefined bounding boxes. Despite the effectiveness of YOLO and RT-DETR models, the lower MAP scores indicate that further advances in end-to-end document understanding are needed.

4. Data Augmentation Enhances Generalization Augmenting digital forms with synthetic effects improved model robustness for handwritten document retrieval (e.g., ChatDY's approach). This suggests that document variability must be carefully considered in future dataset preparation.

5. Ensemble Approaches Provide Performance Gains Models integrating multiple object detection techniques (such as rb-ai's hierarchical ensemble strategy) showed superior performance, indicating that fusing diverse model outputs can enhance accuracy in complex layouts.

6 Conclusion

The VRD-IU competition served as an effective benchmark for evaluating state-of-the-art approaches in key information extraction and localization within form documents. Overall, the VRD-IU competition has accelerated research in visually rich document understanding by providing a well-defined benchmark and fostering innovation in key information extraction and localization. The insights gained will be instrumental in guiding the development of more robust, efficient, and generalizable AI-driven document processing systems.

Ethical Statement

There are no ethical issues.

Acknowledgments

Thanks for the support from Google Research, United States.

Contributors

We would like to acknowledge the invaluable contributions of the authors across all competition winners in IJCAI 2024 VRD-IU Competition:

- **Track A - Team rb-ai (Rank 1): Key Information Extraction with Hierarchical Layout Decomposition**: **Hongjun Choi** (*Teamreboott Inc., South Korea*), **Jongho Lee** (*Teamreboott Inc., South Korea*; *Pukyong National University, South Korea*), and **Jaeyoung Kim** (*Teamreboott Inc., South Korea*).
- **Track A - Team vipski (Rank 2): Redefining Information Extraction from Visually Rich Documents as Token Classification**: **Jonghyun Song** and **Eunyi Lyou** (*Graduate School of Data Science, Seoul National University*).
- **Track A - Team gcu (Rank 3): Transformer-Based Visual Feature and Textual Feature Fusion Model for Document Key Information Extraction**: **Wooseok Kim**, **Juhyeong Kim**, **Sangyeon Yu**, **Gyunyeop Kim**, and **Sangwoo Kang** (*School of Computing, Gachon University*).
- **Track B - Team rb-ai (Rank 1):Key Information Localization With Hierarchical Ensemble**: **Hongjun Choi** (*Teamreboott Inc., South Korea*), **Jongho Lee** (*Pukyong National University, South Korea*), and **Jaeyoung Kim** (*Teamreboott Inc., South Korea*).
- **Track B - Team chatdy (Rank 2):Enhancing Document Key Information Localization Through Data Augmentation**: **Yue Dai** (*The University of Western Australia*).

Their dedication and innovative research have significantly advanced the field of key information extraction and visually rich document understanding.

References

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

[Ding *et al.*, 2024a] Yihao Ding, Soyeon Caren Han, Zechuan Li, and Hyunsuk Chung. David: Domain adaptive visually-rich document understanding with synthetic insights. *arXiv preprint arXiv:2410.01609*, 2024.

[Ding *et al.*, 2024b] Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. Mvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*, 2024.

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

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[Huang *et al.*, 2019] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019.

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

[Li *et al.*, 2022] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3530–3539, 2022.

[Liu *et al.*, 2024] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[OpenAI, 2024] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

[Park *et al.*, 2019] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.

[Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[Stanisławek *et al.*, 2021] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I*, pages 564–579. Springer, 2021.

[Tang *et al.*, 2023] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264, 2023.

[Wang *et al.*, 2021] Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. Towards robust visual information extraction in real world: New dataset and novel solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2738–2745, 2021.

[Xu *et al.*, 2021] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*, 2021.

[Xue *et al.*, 2024] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024.

[Zhao *et al.*, 2024] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.