Adapting Vision-Language Models for Information Extraction from Bilingual Medical Invoices

Anh-Dung DO[1] and Thanh-Ha DO[2]
[1] VNU University of Science, [2] Posts and Telecommunications Institute of Technology
E-mails: doanhdung_t66@hus.edu.vn, dothanhha@ptit.edu.vn

***Abstract*—Recent advances in Vision-Language Models (VLMs) have significantly improved document understanding tasks such as invoice processing and form parsing. In this research, we present a survey of recent VLM-based approaches for information extraction from complex document images, with a focus on pharmacy invoices containing bilingual (English-Vietnamese) content. We analyze and compare state-of-the-art models in terms of architecture, dataset usage, and performance. Furthermore, we propose an adaptation workflow utilizing synthetic datasets and the Vintern-3B model, which addresses key challenges arising from bilingual document structures. Our experiments highlight the effectiveness of background removal and augmentation strategies in enhancing extraction performance.**

I. Introduction

Medical invoices are important documents in healthcare and insurance systems, containing information such as treatment costs, prescribed medications, and services provided [1]. Extracting data automatically from these invoices helps to enhance efficiency, reduce manual effort, and improve data accuracy. However, medical invoices have challenges compared to traditional commercial invoices because they often lack standardized formats, with each healthcare facility using different layouts, terminologies, and table structures that incorporate domain-specific abbreviations and contextual details [2]. As a result, traditional approaches based on Optical Character Recognition (OCR) and rule-based methods achieve accurate data extraction but are limited in their ability to interpret complex semantic relationships. Smartphone-captured images further increase these challenges as they often suffer from quality issues such as skewness, blurriness, poor lighting, and the presence of extraneous elements like hands or shadows.

Deep learning models have achieved significant advancements in image document analysis. However, they are effective because they have a lot of data. While there are published datasets for document information extraction, such as CORD [3] and SROIE [4], none are specifically to the medical domain. These datasets primarily consist of general-purpose receipts or invoices and lack the specialized structure, domain-specific terminology, and complexity commonly found in healthcare-related documents.

According to the MIT Sloan Management Review (2025) [5], unstructured data is re-emerging as a critical focus for AI applications, driven by advancements in Generative AI. The report highlights that 97% of data in large organizations exists in unstructured forms (images, free-text, videos), yet processing this data remains labor-intensive, requiring manual curation, tagging, and normalization. In medical invoice processing, these challenges include critical information (e.g., physician names, drug dosages) scattered across dynamic layouts, often obscured by stamps or handwritten notes.

Given these limitations, a more intelligent and adaptive approach is necessary to extract structured information from medical invoices accurately. Recent advancements in Vision-Language Models (VLMs) have demonstrated strong capabilities by integrating layout awareness, semantic understanding, and multimodal processing. These models offer a promising solution to overcome the shortcomings of traditional OCR and rule-based methods.

This paper has three contributions to Vision-Language-based information extraction. First, we independently construct a synthetic medicine invoice dataset, designing diverse invoice layouts and embedding healthcare-specific terminology to simulate real-world documents better. To the best of our knowledge, this is the first dataset focusing on medicine invoices with bilingual (English-Vietnamese) content, addressing a previously unexplored gap in multilingual document understanding. Second, we enrich the dataset by integrating synthetic invoices with real-world document images captured via smartphones, explicitly addressing complex scenarios such as overlapping documents and cluttered backgrounds, thereby improving content separation in noisy visual contexts. Third, we present a pipeline for Vision-Language Models, utilizing the constructed dataset for structured information extraction.

The paper is organized into four main sections. Section I highlights the challenges of extracting structured information from Vietnamese medical invoices. Section II reviews recent advancements in VLMs for document understanding, comparing OCR-based and OCR-free approaches in terms of architecture, benchmark performance, and applicability to structured documents. Section III presents the paper's core contribution, a unified framework for structured information extraction. This includes the creation of a synthetic dataset to be enriched with realistic features such as stamps, signatures, QR codes, and diverse layouts, along with a robust data augmentation pipeline. Background removal is handled using DeepLabV3 with a MobileNetV3 backbone, and structured fields are extracted using the Vintern-3B model to generate machine-readable JSON. The experiments and results section IV evaluates both synthetic and authentic invoice images using standard metrics, demonstrating significant improvements due to the proposed preprocessing and augmentation strategies. Finally, Section V

979-8-3315-7206-8/25/$31.00

TABLE I: Summary of Vision-Language Models with Document Understanding Benchmarks

Model	CORD (F1)	FUNSD (F1)	DocVQA (ANLS)	OpenViVQA (Score)	InfoVQA (ANLS)	TextVQA (Acc)
LayoutLMv3 [6]	96.42	85.94	–	–	–	–
Donut [7]	96.6	–	–	–	–	–
Pix2Struct [8]	–	–	76.6	–	40.0	–
PaLI-3 [9]	–	–	87.6	–	76.7	84.1
Vintern-1B [10]	–	–	–	7.7	–	–
Gemini 1.5 [11]	–	–	93.1	–	81.0	78.7
DeepSeek-VL2 [12]	–	–	93.3	–	78.1	–
Qwen2.5-VL [13]	–	–	96.4	–	87.3	–
InternVL3 [14]	–	–	95.4	–	86.5	–

discusses the practical implications and outlines future work.

II. Survey of Vision-Language Models for Document Information Extraction

Document Information Extraction (Doc-IE) is an essential task in the field of document AI, aiming to extract structured content, such as key-value pairs, tables, and hierarchical fields from visually rich documents. Compared to general Visual Question Answering (VQA), Doc-IE structured outputs (e.g., JSON, XML) directly from document images without requiring natural language queries. This task is crucial for real-world applications, including automated form processing, financial receipt parsing, and healthcare invoice analysis. A comprehensive comparison of recent vision-language models and their performance on Doc-IE-related benchmarks is presented in Table I.

Traditional approaches are OCR-based approaches, where textual content is extracted via OCR engines and passed to layout-aware models, such as LayoutLMv3 [6], for downstream field classification. While effective on datasets such as FUNSD [15], CORD [3], and SROIE [4], these methods are limited by OCR quality and are brittle under noisy inputs or handwritten content. Additionally, they often require hand-crafted annotations and domain-specific post-processing rules. lassification. While effective on datasets such as FUNSD [15], CORD [3], and SROIE [4], these methods are limited by OCR quality and are brittle under noisy inputs or handwritten content. Additionally, they often require hand-crafted annotations and domain-specific post-processing rules.

Recent advancements toward end-to-end, OCR-free architectures that directly process raw document images. Notable models include Donut [7], which frames document parsing as a sequence generation task using visual prompts, and Pix2Struct [8], which maps screenshots to structured outputs via transformer decoders. These models eliminate the need for OCR and demonstrate robustness across diverse layouts and low-quality scans.

Large Vision-Language Models (VLMs) such as PaLI-3 [9], InternVL3 [14], and DeepSeek-VL2 [12] have further pushed the boundaries of document understanding. Pretrained on massive multimodal datasets, they combine visual reasoning with strong language modeling capabilities, enabling them to generalize across diverse document types. Several models adopt structured prompting strategies and long-context transformers to support dense extraction over full-page documents, including tables and nested forms.

Of particular interest are lightweight and domain-specific models. Vintern-1B [10] and Vintern-3B [16] are notable examples designed for document understanding in low-resource languages such as Vietnamese. Trained on synthetic and real-world documents from medical, financial, and administrative domains, they have strong performance on structured parsing tasks while maintaining practical efficiency for real-world deployment.

Despite these advances, current benchmarks primarily focus on general-purpose documents. Datasets like DocVQA [17] and InfoVQA [18] include question-answering and visual reasoning but do not reflect the structure and terminology found in domain-specific documents like medical invoices. Similarly, existing IE datasets such as FUNSD, CORD, and SROIE cover limited domains (e.g., receipts or forms) and lack healthcare-specific annotations.

To the best of our knowledge, there is no publicly available dataset or comprehensive evaluation dedicated to structured information extraction from medicine invoices. This work addresses these gaps by (1) constructing a medical invoice dataset, including synthetic augmentation to handle format variability, and (2) applying and evaluating Vintern-3B for structured information extraction, producing machine-readable outputs such as JSON for healthcare and insurance systems.

III. Proposed Approach: Structured Information Extraction from medicine Invoices

This section presents a proposed approach for extracting structured information from Vietnamese medical invoices. Our model requires directly parsing raw document images and outputting structured information by a predefined JSON schema, including critical medical and financial entities such as patient name, hospital, invoice date, prescribing physician, and a detailed list of prescribed medications with associated quantities and pricing.

We also propose a multi-step process to create the synthetic dataset [1] (see Figure 1). The process begins with structured information in JSON format, containing key fields merged with a LaTeX invoice template to render a clean invoice image, complete with essential components such as barcodes, signatures, and stamps.

[1]For access to the dataset of synthetic invoice, interested readers are encouraged to contact the author for further information.

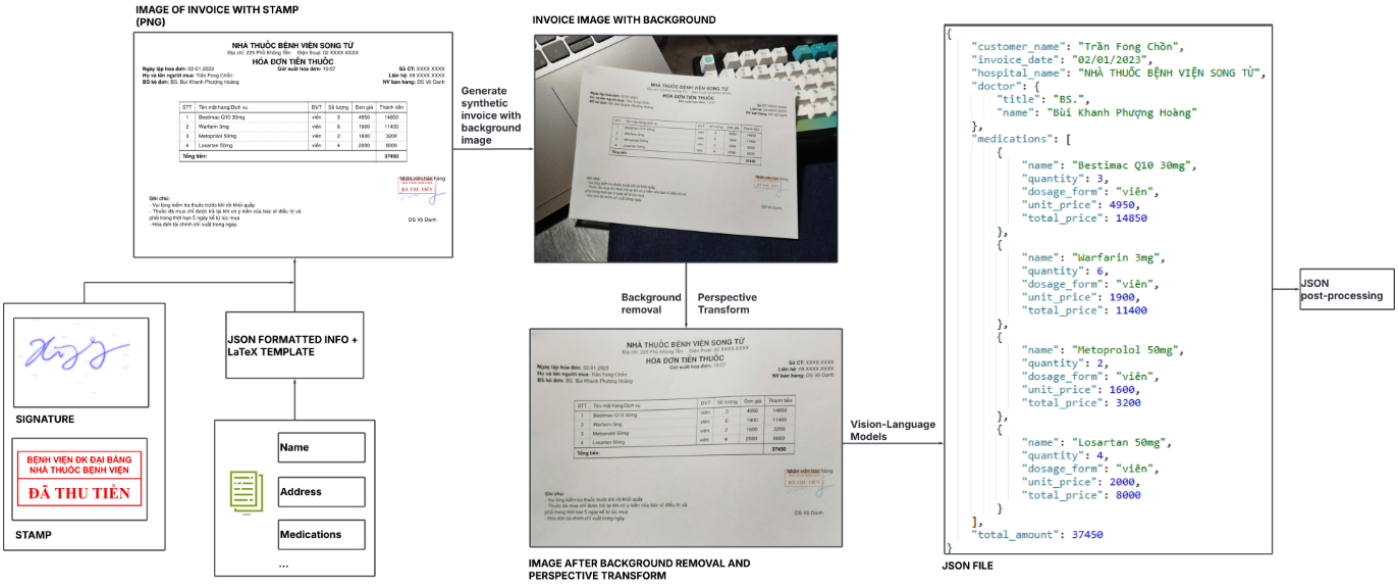


Fig. 1: Overall flow of Vietnamese Information Analysis and Synthesis from Medicine Invoices Image

Stage 1: Create Initial Dataset Components

The first step involves building the core data elements, which include structured content, such as medication details and names, as well as visual elements, like stamps, signatures, and barcodes.

Medication data has both generic and trade names across various medical categories, and drug names are collected by patients, doctors, and hospitals to ensure rich diversity. Visual realism was enhanced using real handwritten signatures from the CEDAR [19] dataset, synthetic barcodes, and various stamps. Table II illustrates the number of each items of each component of dataset, both before and after data augmentation.

TABLE II: Initial Dataset Components. The dash indicate no data modification was applied.

Dataset	Description	Before	After
Medication	Name, quantity, dosage form, unit price	–	98
Name	Doctors, patients, hospitals	–	50 / 14
Stamp	Basic stamps	4	20
Signature	From CEDAR dataset	2,640	225

Stage 2: Create ground-truth JSON file

Each invoice image is paired with a structured JSON file that includes information such as the customer's Name, invoice date, hospital name, doctor's title and Name, list of medications (with details), and the total amount.

Stage 3: Create Synthetic Invoice Images

Based on the initial data and the ground-truth JSON file, medical invoices were generated using LaTeX templates with dynamic values populated using the Jinja2 engine. The created documents simulate a variety of hospital invoice styles with varying fonts, table formats, and field structures. These documents were then exported to PNG images for the next stage.

Some augmentation techniques were also applied at this step to increase layout diversity, as illustrated in Table III. Figure 2b examples some different invoice layouts.

TABLE III: Types of augmentations applied to increase layout diversity

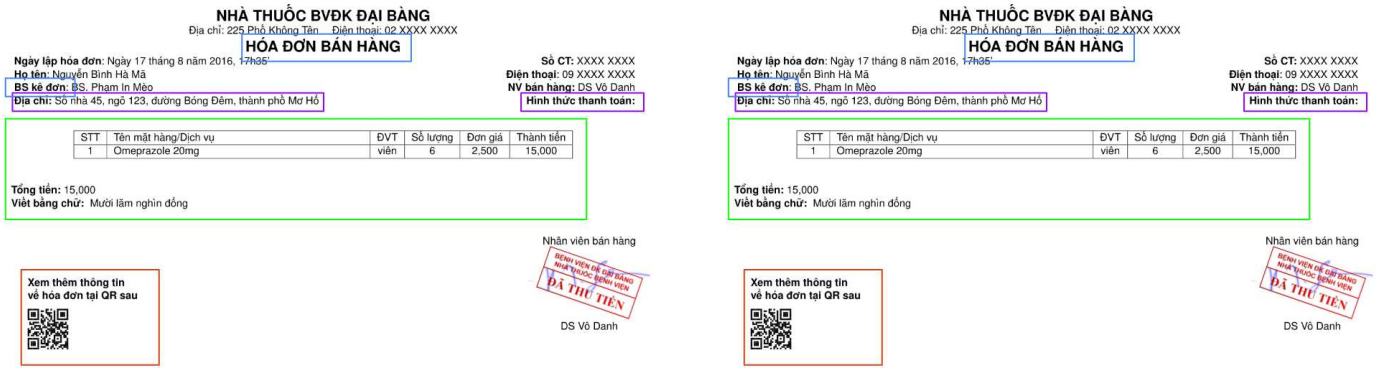
Type of Change	Details
Table styles	Full border, no border, table without horizontal lines
Field label formats	Example: "Phone" → "Tel"
QR code size	1–2 cm
Written amounts	Example: "500,000" → "Five hundred thousand dong"
Random insertion/removal of fields	Adding or removing fields such as address or payment notes

Stage 4.1: Generate Invoice Images with Synthetic Backgrounds

To simulate real-world background conditions, invoice images were composited with synthetic backgrounds using a multi-step image processing pipeline. The initial implementation was adapted from an open-source tutorial [2] on document background removal but was significantly customized to better fit our use case involving synthetic invoice rendering.

We also apply several enhancements to improve realism and dataset quality, including flipping heavily torn or distorted documents, applying perspective distortion, and adjusting brightness and contrast. Additionally, with the introduction

[2]LearnOpenCV. *Deep Learning Based Document Segmentation Using Semantic Segmentation (DeepLabv3) on Custom Dataset*, 2022.



(a) Invoice layout A

(b) Invoice layout B

Fig. 2: Examples of different Vietnamese invoice layouts showing structural diversity in the dataset.

of document-overlapping backgrounds, we simulate cluttered desks and complex real-world layouts. Furthermore, the original shadow overlay method was replaced with a custom shadow rendering mechanism, providing a more realistic integration of shadows without introducing visual artifacts.

Stage 4.2: Capture Images with Phone Camera: To further enhance realism, printed invoices were captured using a mobile phone under various natural conditions, including indoor lighting, shadows, desk clutter, and angular distortion. Each image was manually annotated with a mask to assist in supervised training.

Final Dataset Summary: The final dataset comprises a combination of synthetic images and real-world camera-captured samples. A total of 3550 images were created to ensure both diversity and realism during training.

For the synthetic dataset, 50 invoice templates were rendered onto two background types: clean backgrounds and backgrounds already containing elements similar to those found on invoices. Each invoice was composited with 25 background images in the "clean-background" group and 25 in the "background-with-invoice" group, resulting in 1,250 images per category.

In addition, a total of 250 real-world invoice images were captured using a mobile phone in various environmental settings. These images were divided into two subsets: 200 images were used for training and validation, while 50 images were used for testing. To introduce additional diversity, each of the 200 training and validation photos was augmented five times, resulting in 1000 augmented images. The 50 test images, on the other hand, were kept unchanged to ensure an unbiased evaluation of the model's performance.

Table IV summarizes the composition of the dataset.

IV. Results and Discussion

In this paper, the evaluation process consists of two main components: background removal and OCR-based information extraction, which are divided into training (3,150 images), validation (350 images), and testing (50 images), as summarized in Table V.

All three subsets are used for background removal tasks, whereas the information extraction task employs a purely inference-based approach. To ensure fair evaluation, synthetic images are excluded from the testing phase.

The chosen architecture is DeepLabV3 with MobileNetV3-Large as the backbone for the background removal process. This model is trained on 50 epochs with a batch size of 64. For optimization, the Adam optimizer is used with a learning rate of 0.0001 [20]. To prevent overfitting, a dropout probability of 0.5 is applied.

For the information extraction task, the Vintern-3B [16] vision-language model is employed, with a maximum token length of 1900 and a repetition penalty of 3.5, to mitigate text redundancy.

A. Results

As illustrated in Table VI, the model achieved a high Dice score of 94.86% and an IoU of 91.81% for background removal process. These results indicate that the invoice part was accurately isolated from cluttered backgrounds, including cases that have overlapping documents.

TABLE VI: Performance of model for background removal

Metric	Dice (%)	IoU (%)
Result	94.86	91.81

Figure 3 illustrates typical examples of background removal outcomes. In cases without overlapping content (Figures 3a and 3b), the model segments the invoice with near-perfect precision. When documents overlap (Figures 3c–3f), the effectiveness of background removal is influenced not only by the extent of overlap but also by the visibility of the document's geometric boundaries. In the second case, despite some overlapping content, the four corners and edges of the foreground invoice remain visible. This allows the model to accurately infer the document region and perform successful segmentation, as shown in Figures 3c and 3d. In contrast, the third case indicates a more complex scene where multiple

TABLE IV: Final dataset composition

Dataset Type	Photos	Variations per Photo	Image Count
Synthetic Dataset			
Normal background	50	25 backgrounds	1250
Background with invoice	50	25 backgrounds	1250
Camera-taken Dataset			
Real invoice photos (for train and validation)	200	5 augmentations	1000
Real invoice photos (for test)	50		50
Total			3550

TABLE V: Dataset split for training, validation, and testing. Note that the background removal task uses all three subsets, while the information extraction task only evaluates on the test set of captured images.

Set	Train	Validation	Test	Total
Captured image	900	100	50	1050
Synthetic image	2250	250	-	2500
Total	3150	350	50	3550

documents overlap and several corners of the target invoice are obscured. The document boundaries are unclear and often blend with the surrounding clutter. As a result, the model is complex to distinguish the foreground from the background, leading to lower segmentation accuracy, as illustrated in Figures 3e and 3f.

TABLE VII: OCR Performance Before and After Background Removal

Field	Before		After	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
Customer Name	57.00	55.83	83.67	83.00
Invoice Date*	74.00	74.00	92.00	92.00
Hospital Name	73.17	72.87	90.73	90.07
Doctor Title*	57.00	57.00	63.00	63.00
Doctor Name*	55.07	56.83	78.73	80.83
Total Amount	70.00	70.00	94.00	94.00
Medication Name	71.79	71.11	89.57	89.57
Quantity	71.27	71.27	92.80	92.80
Dosage Form	78.10	78.10	88.90	88.90
Unit Price	64.27	64.27	93.60	93.60
Total Price	63.87	63.87	91.73	91.73

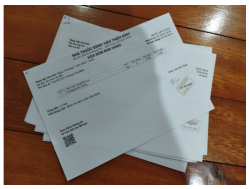
**Fields with post-processing applied.*

The impact of background removal on OCR performance is substantial. As shown in Table VII, all fields benefited from this preprocessing step. On average, precision and recall rates increased by approximately 20–30 percent.

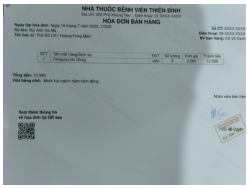
For example, low-performing fields, such as Customer Name and Doctor Name, improved by more than 25 percentage points (from 57.00% to 83.67% and from 55.07% to 78.73%, respectively). Similarly, fields with moderate initial performance, such as Medication Name (from 71.79% to 89.57%), also saw a significant increase of approximately 18 percentage points. Even fields with higher initial performance, such as Invoice Date (from 74.00% to 92.00%) and Total Amount (from 70.00% to 94.00%), showed notable improvements, increasing by approximately 18–24 percentage points. These results demonstrate a consistent upward trend in OCR performance across all fields, with the most significant gains observed in fields that initially had lower recognition rates.

(a) Case 1: Before, without any overlap

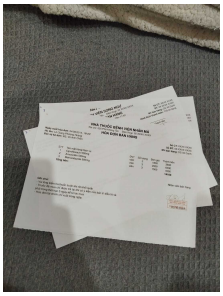
(b) Case 1: After, successful background removal



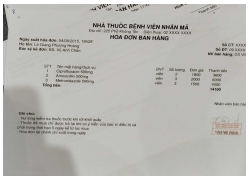
(c) Case 2: Before, with overlapping documents



(d) Case 2: After, successful background removal



(e) Case 3: Before, with overlapping documents



(f) Case 3: After, lower performance

Fig. 3: Compare images under different cases, both before and after background removal. In the first and second case, the result is almost perfect. However, when the top border of the invoice overlap with other invoices, some contents of the underlying invoices still visible.

V. Conclusion and Future Directions

This paper reviews recent developments in VLMs for document understanding and presents a practical application for extracting information from medicine invoices. Experimental results show that VLMs are effective for this task, especially when combined with the background removal process. Prompt-based models, while offering a wide range of flexibility, are highly dependent on the quality of OCR output. Additionally, the use of both synthetic data and image-captured data significantly enhances the overall extraction performance.

In the future, we will enhance the dataset's diversity by incorporating invoice layouts from private pharmacies. Additionally, real-time mobile applications will be developed to improve the usability of the proposed approach.

References

[1] F. Yi, Y.-F. Zhao, G.-Q. Sheng, *et al.*, "Dual model medical invoices recognition," *Sensors*, vol. 19, no. 20, p. 4370, 2019. DOI: 10.3390/s19204370.

[2] T. Saout, F. Lar Deux, and F. Saubion, "An overview of data extraction from invoices," *IEEE Access*, vol. 10, pp. 12 345–12 356, 2024. DOI: 10.1109/ACCESS.2022.3145678.

[3] S. Park, S. Shin, B. Lee, *et al.*, "Cord: A consolidated receipt dataset for post-ocr parsing," in *Document Intelligence Workshop at Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Z. Huang, F. Zhan, L. Jin, *et al.*, "Icdar 2019 competition on scanned receipt ocr and information extraction," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, 2019, pp. 1516–1520. DOI: 10.1109/ICDAR.2019.00244.

[5] T. H. Davenport, R. Bean, and D. Vesset, "Five trends in AI and data science for 2025," *MIT Sloan Management Review*, 2024.

[6] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," in *Proceedings of the 30th ACM International Conference on Multimedia*, Lisboa, Portugal: ACM, 2022, pp. 1–10. DOI: 10.1145/3503161.3548112.

[7] G. Kim, T. Hong, M. Yim, *et al.*, "Ocr-free document understanding transformer," in *Proc. ICLR*, 2022. DOI: 10.48550/arXiv.2304.06549.

[8] K. Lee, M. Joshi, I. Turc, *et al.*, "Pix2struct: Screenshot parsing as pretraining for visual language understanding," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[9] X. Chen, X. Wang, L. Beyer, *et al.*, *Pali-3 vision language models: Smaller, faster, stronger*, Preprint, 2023. arXiv: 2310.09199.

[10] K. T. Doan, B. G. Huynh, D. T. Hoang, *et al.*, *Vintern-1b: An efficient multimodal large language model for vietnamese*, arXiv preprint arXiv:2408.12480, 2024.

[11] G. Team, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*, arXiv preprint arXiv:2403.05530, 2024.

[12] Z. Wu, X. Chen, Z. Pan, *et al.*, "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," *arXiv preprint arXiv:2412.10302*, 2024.

[13] Qwen Team, Alibaba Group, *Qwen2.5-vl technical report*, arXiv preprint arXiv:2502.13923, 2025.

[14] J. Zhu, W. Wang, Z. Chen, *et al.*, *Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models*, arXiv preprint arXiv:2504.10479, 2025.

[15] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "Funsd: A dataset for form understanding in noisy scanned documents," in *2019 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, 2019, pp. 1046–1052. DOI: 10.1109/ICDAR.2019.00180.

[16] K. T. Doan, B. G. Huynh, D. T. Hoang, *et al.*, *Vintern-3b*, 2025.

[17] M. Mathew, D. Karatzas, and C. V. Jawahar, "Docvqa: A dataset for vqa on document images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2200–2209.

[18] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. Jawahar, "Infographicvqa: Visual question answering on infographic images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 401–12 411.

[19] H. Srinivasan and S. N. Srihari, *CEDAR Signature Database*, 2006.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.