SynDoc: A Hybrid Discriminative-Generative Framework for Enhancing Synthetic Domain-Specific Visually-Rich Document Understanding

Anonymous ACL submission

Abstract

Domain-specific Visually Rich Document Understanding (VRDU) presents significant challenges due to the complexity and sensitivity of documents in fields such as medicine, finance, and material science. Existing Large (Multimodal) Language Models (LLMs/MLLMs) achieve promising results but face limitations such as hallucinations, inadequate domain adaptation, and reliance on extensive fine-tuning datasets. This paper introduces SynDoc, a novel framework that combines discriminative and generative models to address these challenges. SynDoc employs a robust synthetic data generation workflow, using structural information extraction and domain-specific query generation to produce high-quality annotations. Through adaptive instruction tuning, SynDoc improves the discriminative model's ability to extract domain-specific knowledge. At the same time, a recursive inferencing mechanism iteratively refines the output of both models for stable and accurate predictions. This framework demonstrates scalable, efficient, and precise document understanding and bridges the gap between domain-specific adaptation and general world knowledge[1].

1 Introduction

Visually Rich Documents combine visual elements and text to convey information in an engaging and thorough way (Ding et al., 2024b). With the increasing demand for domain-specific Visually Rich Document Understanding (VRDU), significant opportunities are emerging in areas such as medicine (Ding et al., 2023b, 2024c), finance (Zhu et al., 2022; Ding et al., 2023a), material science (Khalighinejad et al., 2024), and politics (Wang et al., 2023). These areas often rely on documents that contain extensive domain-specific knowledge and sensitive information, which pose unique challenges to automated understanding systems. As

[1]The code will be released after acceptance

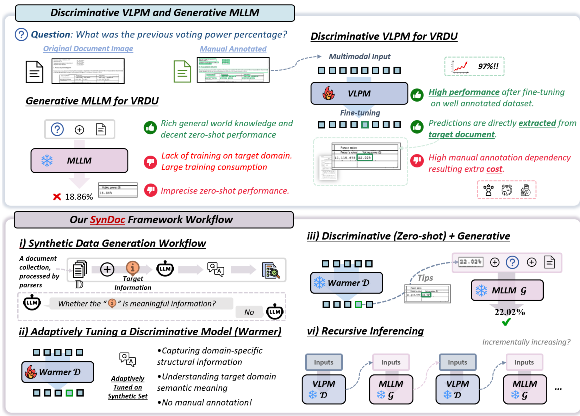

Figure 1: Comparing SynDoc with discriminative and generative VRDU frameworks.

industries increasingly turn to AI-powered solutions for document analysis, the need for robust and adaptable frameworks capable of navigating these intricacies has reached an unprecedented level.

Vision-Language Pretrained Models (VLPMs) (Huang et al., 2022; Gu et al., 2021; Lyu et al., 2024) have demonstrated significant advances in VRDU, normally in a **discriminatory** manner by directly mapping multimodal inputs to structured outputs through classification and sequence labeling. Yet, they encounter several challenges. First, they are heavily dependent on extensive fine-tuning datasets (Ding et al., 2024a). Second, their practical use, particularly in zero-shot scenarios, is limited by hallucinations and inconsistent domain adaptation. Multimodal Large Language Models (MLLMs) have been applied to VRDU in a **generative** manner (Hu et al., 2024b; Feng et al., 2024), achieving remarkable progress due to their rich general knowledge; however, they suffer from a lack of target domain knowledge, leading to unreliable and imprecise outputs in VRDU applications. For instance, as shown in Figure 1, an MLLM extracts the *present voting power* "18.86%" instead of the requested *previous voting power* ("22.02%"), highlighting its limitations in understanding the structure of tables.

Recent existing research has explored various strategies to address these challenges in VRDU, with synthetic data generation increasingly emerging as a crucial approach, driving advances in both discriminative and generative models. The discriminative framework uses domain-adaptive techniques in the VLPM backbone, achieving promising results through fine-tuning on curated annotated datasets (Ding et al., 2024a). However, this approach remains constrained by high annotation costs and limited zero-shot performance. However, generative models leverage synthetic data for self-supervised pretraining (Hu et al., 2024b; Feng et al., 2024) and instructive tuning (Hu et al., 2024a; Tang et al., 2024; Zhang et al., 2024) to enhance multimodal VRD comprehension. However, the massive computational demands and suboptimal performance in zero-shot scenarios in a new domain are challenges. The synthetic generation method powered by MLLMs (Ding et al., 2024c) often faces issues generating meaningful or inconsistent question-answer pairs. Therefore, the field still sees a gap in research on how to improve the quality of these generated QA pairs.

In this study, we propose **SynDoc**, a new hybrid framework that leverages discriminative and generative models to enhance VRDU through a multifaceted approach. Compared to previous studies, SynDoc offers several advantages. First, SynDoc employs a robust synthetic data generation workflow that blends structural information extraction techniques, such as OCR (Optical Character Recognition) and PDF parsing, with multi-task inquiry generation and quality verification modules. This workflow ensures the creation of high-quality synthetic annotations that accurately reflect both document structure and content, enabling a nuanced understanding of complex domain-specific documents. Second, SynDoc integrates a discriminative model, referred to as the *warmer*, with a generative MLLM to combine their complementary strengths. The discriminative model leverages pre-trained backbones, adaptively fine-tuned on synthetic datasets, to effectively extract domain-specific knowledge. Simultaneously, the generative model utilizes state-of-the-art MLLM to generate abstractive answers through zero-shot prompting. Third, SynDoc employs adaptive instruction tuning incorporating multimodal features- including text, visuals, layouts, and structural elements- with predictions from MLLMs. This approach enables the discriminative warmer to provide detailed, context-aware information, thus enhancing the outputs of the generative model. Finally, a key innovation in SynDoc is its recursive inferencing mechanism, where outputs from both the discriminative and generative models undergo iterative refinement through cross-feeding. This iterative process contributes to more stable and accurate responses in zero-shot settings. By integrating these components, we hypothesize that SynDoc offers a scalable and robust framework for domain-specific document understanding; we demonstrate its effectiveness on three domain-specific datasets and assess its generalizability using a cross-domain dataset.

2 Related Work

Curated and synthetic data for VRDU. Heuristic (Watanabe et al., 1995; Seki et al., 2007) and statistical learning methods (Oliveira and Viana, 2017) perform well in domain-specific document understanding but rely on expert efforts, limiting cross-domain adaptability. (Huang et al., 2022; Tang et al., 2023; Lyu et al., 2024; Xu et al., 2021a; Wang et al., 2022a; Hong et al., 2022) address this limitation by employing self-supervised learning on large-scale, unannotated, and multi-source document collections such as RVL-CDIP (Harley et al., 2015), thereby improving generalizability and multimodal comprehension in broader VRDU tasks. Fine-tuning these frameworks with curated datasets achieves state-of-the-art performance in specific VRDU tasks. However, the creation of high-quality curated datasets (Jaume et al., 2019; Park et al., 2019; Ding et al., 2023b) is resource-intensive, posing challenges for scalability and applicability to novel document collections. Recent research (Ding et al., 2024c) has explored using LLMs/MLLMs to generate synthetic datasets with well-designed prompts and human verification. Some VRDU MLLMs also create large-scale synthetic datasets to conduct self-supervised pretraining (Hu et al., 2024b; Feng et al., 2024) or instruct-tuning (Hu et al., 2024a; Tang et al., 2024; Zhang et al., 2024) to enhance multimodal document understanding. A recent work DAViD (Ding et al., 2024a) pretrains VRDU models with synthetic QA pairs, followed by semi-supervised refinement, achieving performance comparable to full supervision. However, there remains a limited exploration into optimizing synthetic dataset generation and integrating SoTA MLLMs for real-world applications.

VRDU frameworks. Self-supervised frameworks (Wang et al., 2022b; Appalaraju et al., 2023; Kim et al., 2022) employ diverse pretraining tasks to enhance multimodal learning, achieving strong performance on downstream tasks when fine-tuned with curated datasets. However, most discriminative models rely heavily on off-the-shelf OCR tools such as LayoutLM-series (Xu et al., 2020, 2021a; Huang et al., 2022; Xu et al., 2021b), making extractive predictions vulnerable to cumulative errors from both the models and OCR systems. To mitigate this, end-to-end OCR-free frameworks (Kim et al., 2022; Abramovich et al., 2024; Lyu et al., 2024) bypass OCR dependency. Despite these advances, their smaller model sizes and limited training resources constrain world knowledge, reducing generalization without substantial annotations. LLMs/MLLMs (OpenAI, 2024; Team et al., 2024; Bai et al., 2023; Laurençon et al., 2024; OpenAI, 2023), benefiting from scaling laws, leverage extensive training to capture broad knowledge, supporting zero-shot and few-shot learning in VRD tasks (He et al., 2023). However, issues like hallucination and lack of domain-specific knowledge limit their reliability. Our SynDoc aims to bridge this gap by introducing an adaptively tuned discriminative warmer that provides domain-specific knowledge, which is then integrated into a generative MLLM. This approach enables the model to refine the inference process recursively, leveraging both domain-aware information and broad world knowledge to enhance accuracy and reliability.

3 Methods

3.1 Overview of SynDoc

Let $\mathbb{D}$ be a document collection within a *specific domain*. We propose a framework to predict the answer to a user-provided natural language query Q concerning a specific document $d \in \mathbb{D}$. This framework integrates a discriminative model $\mathcal{D}$ and a generative model $\mathcal{G}$ to address Q in extractive and abstractive manners, respectively. $\mathcal{D}$ employs pretrained backbones to capture target-domain knowledge named as a **warmer**, while $\mathcal{G}$ employs state-of-the-art LLMs/MLLMs and applies specific prompts P to predict answer of in zero-shot scenarios.

To ensure the workflow is functional, we first generate the synthetic dataset (Figure 2). This process begins with structural information extraction using off-the-shelf tools (e.g., OCR or PDF parsers[2]). Next, synthetic domain-specific queries are generated using MLLMs. Therefore, $\mathcal{D}$ incorporates multimodal representations, including textual, visual, layout, and structural features, along with predictions from MLLM. During inference, the outputs from $\mathcal{D}$ and $\mathcal{G}$ undergo iterative refinement through cross-feeding until they achieve convergence (e.g., stable predictions). The following subsections describe four key modules in SynDoc: Synthetic Data Generator, Discriminative Warmer Architecture, Adaptive Instruction Tuning, and Recursive Inference.

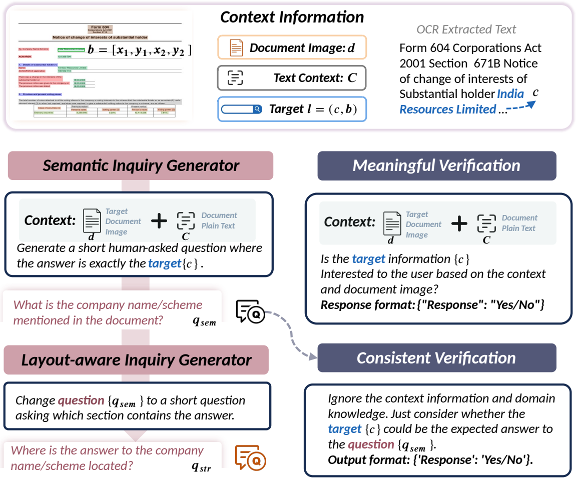

Figure 2: Workflow of the Synthetic Data Generator.

3.2 Synthetic Data Generator

VRD Structure Parsing We use off-the-shelf tools to extract the text content and layout structure of a target document collection (Figure 2). For document images, we employ vision-based OCR tools to get text line entities L. Each $l = (b, c) \in L$ contains the bounding boxes b with corresponding textual content c. We use $(x_{min}, y_{min}, x_{max}, y_{max})$ to represent coordinates of each box. For text-embedded PDF files, we employ the PDF parsing tools to acquire text line or document semantic entity sets L (e.g., paragraph, list, section) along with more accurate structural information.

MLLM-driven Inquiry Generation For $\mathcal{D}$ to capture knowledge from the target domain, we propose a MLLM-driven workflow with two modules (Figure 2). *i) Multi-Task Inquiry Generation* produces diverse inquiries to instruct-tune $\mathcal{D}$ to enhance its structural and semantic understanding of the domain. Specifically, a set of text lines is

[2] https://github.com/PaddlePaddle/PaddleOCR or https://pypi.org/project/pdfminer/

randomly selected and fed to an LLM to generate two types of QA pairs. First, *Semantic* QA pairs guide $\mathcal{D}$ to extract target information from a document. By inputting the target entity content along with its document and context information into an MLLM, we generate pairs (q_{sem}, c), where c is the answer to the generated question q_{sem}. Second, *Spatial-aware* QA pairs facilitate $\mathcal{D}$ in capturing both semantic and spatial correlations. Here, we transform q_{sem} into q_{spt} by identifying the document region (e.g., top-left, top-middle, top-right) where the target information c is located. *ii) Multi-Aspect Quality Verification* is implemented to filter out low-quality questions by assessing factors including meaningfulness and question-answer consistency. It first determines whether c is relevant to the end user (e.g., "Is the target information interesting to the end user?"). It then verifies that c adequately answers q_{sem} (e.g., "Whether the target information c could be expected answer of a question q_{sem}?").

3.3 Warmer Architecture

Warmer ($\mathcal{D}$) utilizes a vision-language pre-trained model (VLPM) as its backbone, optimized for discriminative answer extraction through adaptively tuning on synthetic datasets. The adopted VLPM is pre-trained on layout-aware tasks and fine-tuned on well-annotated datasets, exhibiting decent performance in targeted VRDU tasks. To address zero-shot scenarios, we design the warmer architecture based on the VLPM backbone, enabling $\mathcal{D}$ to learn multi-aspect domain-aware knowledge from synthetic datasets. We will first introduce the initial feature representation of $\mathcal{D}$ and then describe the detailed architecture.

Initial Feature Representation For a synthetically acquired entity set L of document I_d, a pre-trained vision model extracts visual representation v from b and a text model extracts sentence representation s from c (Ding et al., 2024c). b's coordinates are linearly projected to match s (Tan and Bansal, 2019). A textual sequence $C = \{\tau_i\}_{i=1}^{n}$ encodes context, summed with projected coordinates $B = \{b_i\}_{i=1}^{n}$ and, if relevant, concatenated with document image patches P. For each semantic query q_{sem}, the MLLM-generated answer a can aid localization. Grid embeddings $G = \{g_i\}_{i=1}^{j \times k}$ result from resizing and flattening pixel data over a $j \times k$ grid of the document image.

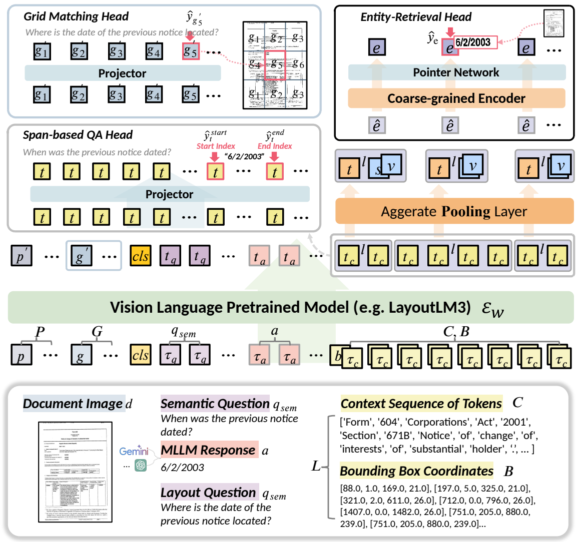

Figure 3: Architecture of the discriminative Warmer.

Detailed Architecture. $\mathcal{D}$ processes the input word sequence (q, a, C, P and B). These inputs are passed through a VLPM backbone, $\mathcal{E}_w$, to derive embedded feature representations:

$$(P', G', T_q, T_a, T_c) = \mathcal{E}_w(P, G, q, a, C + B) \quad (1)$$

where T represented corresponding encoded textual features, while P' and G' represent the encoded patch and grid features, respectively.

For each $l \in L$ extracted using parsing tools, a pooling layer aggregates the token features to obtain the entity-level representation e.

$$\hat{e} = \text{Pooling}\left(\{\mathcal{E}_w(c_i), c_i \in c\}\right) \quad (2)$$

$$e = \hat{e} \oplus v \oplus s \quad (3)$$

The enhanced entity features, $E = \{e_l \mid l \in L\}$, are processed by an **Entity-Retrieval Head**, which includes a coarse-grained transformer encoder for improving entity-level contextual understanding and a pointer network (Ding et al., 2024c) to predict the final entity index. Additionally, a fine-grained **Span-based QA Head** is employed to predict the start and end indices of the answer span based on the input query q. A **Grid Matching Head** is introduced to enhance structural understanding within the target domain. This matching head predicts the grid index of the input set G' by leveraging specially aware queries. A different head is trained on diverse stages to enable warmer capture of adequate domain-specific knowledge.

3.4 Adaptively Warmer Tuning

Step-by-step training enables the warmer $\mathcal{D}$ to effectively adapt to the target domain, starting with

structural adaptation to enhance the domain-specific structural understanding, followed by the task-oriented **semantic adaptation** for locating target information based on the input query.

Structural Adaptation enhances both semantic and layout understanding by guiding $\mathcal{D}$ to identify the most relevant document grid $g' \in G'$ for a given structural query q_{str}. For example, given the query "*Where is the date of the previous notice located?*", $\mathcal{D}$ predicts the grid g_5 that contains the answer (Figure 3). A pointer network computes the logit for each candidate grid (Ding et al., 2024c), and the probability over grids is obtained using the softmax function. The model optimization employs the cross-entropy loss function to compute the structure adaptation loss $\mathcal{L}_{str}$:

$$\mathcal{L}_{str} = -\sum\nolimits_{g' \in G'} y_{g'} \log \hat{y}_{g'} \tag{4}$$

where $y_{g'}$ represents the ground truth indicator of each grid. This adaptation process ensures that the model effectively learns to associate structural queries with relevant document regions, improving both retrieval accuracy and layout-aware reasoning.

Semantic Adaptation enables $\mathcal{D}$ to pretrain on a synthetic semantic QA set P, allowing it to better understand document image I_d and q_{sem} for zero-shot extractive QA in real-world scenarios. The model employs two extractive QA heads: a fine-grained, span-based QA head and a coarse-grained entity-retrieving head. The fine-grained head predicts the start and end token indices using a linear projector, with the cross-entropy loss defined as:

$$\mathcal{L}_{fg} = -\sum\nolimits_{t \in \mathcal{E}_w(c)} y_t^{\text{start}} \log \hat{y}_t^{\text{start}} + y_t^{\text{end}} \log \hat{y}_t^{\text{end}} \tag{5}$$

where y_t^{start} and y_t^{end} denote the ground truth indices, while $\hat{y}_t^{\text{start}}$ and $\hat{y}_t^{\text{end}}$ represent the predicted probabilities after Softmax.

The coarse-grained entity retrieving head retrieves entities based on entity logits and is optimized with a cross-entropy loss function:

$$\mathcal{L}_{cg} = -\sum\nolimits_{e \in E} y_e \log \hat{y}_e \tag{6}$$

where y_e represents the ground truth probability distribution over the entity set E, and $\hat{y}_e$ is the predicted Softmax normalised probability. The final optimization objective combines both losses as:

$$\mathcal{L} = \lambda_{fg}\mathcal{L}_{fg} + \lambda_{cg}\mathcal{L}_{cg} \tag{7}$$

where λ_{fg} and λ_{cg} control the balance between the fine-grained and coarse-grained QA losses. During the semantic adaptation process, different synthetic subsets may be selected based on *Multi-Aspect Quality Verification* results, possibly leading to varying performance, as described in Section 5.2.

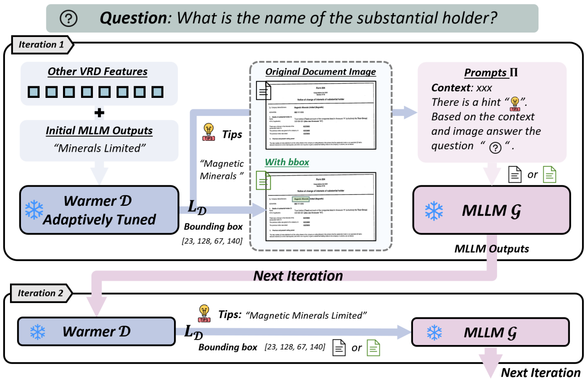

Figure 4: An illustration of the recursively inferencing framework for zero-shot question answering on VRDs. Given a question, *"What is the name of the substantial holder?"*, the initial MLLM output is enhanced using retrieved entity hints ($L_\mathcal{D}$) from the adaptively tuned warmer. Bounding box hints and other VRD features guide MLLM toward more precise answers in subsequent iterations.

3.5 Recursively Inferencing

We propose a recursively inferencing framework to harness $\mathcal{D}$ and $\mathcal{G}$ for zero-shot question answering on VRDs (Figure 4). The retrieved top-k entities $L_\mathcal{D}$ serve as domain-specific guidance to enhance MLLM responses. Originally, given the prompt $(I_d, C, q_{sem}) \rightarrow \Pi$, $\mathcal{G}$ generates an answer $A_\mathcal{G}$. In the t-th recursive process, $\mathcal{D}$ refines its retrieval based on the previous $A_\mathcal{G}^{(t)}$, leading to an updated prompt that integrates the extracted entity information:

$$L_\mathcal{D}^{(t+1)} = \mathcal{D}(A_\mathcal{G}^{(t)}) \tag{8}$$

$$\Pi^{(t+1)} = \text{UpdatePrompt}(\Pi^{(t)}, L_\mathcal{D}^{(t+1)}) \tag{9}$$

$$A_\mathcal{G}^{(t+1)} = \mathcal{G}(\Pi^{(t+1)}) \tag{10}$$

This allows $\mathcal{G}$ to acquire more domain-specific knowledge, improving its ability to comprehend and locate question-relevant information within the context with greater accuracy and reliability. The iterative refinement process enhances both extractive and generative responses over time.

4 Experimental Settings

4.1 Datasets

We used four datasets from different domains to evaluate SynDoc: FormNLU (financial forms)

Model	F-P	F-H	CORD	Ephoie	FUNSD
Idefics2	57.54	33.31	54.45	15.22	62.11
InternVL2	66.56	45.47	66.84	68.92	74.95
Qwen2-VL	78.05	43.65	77.86	70.36	79.12
GPT-4o	76.16	56.49	79.05	79.40	80.05
Gemini	76.09	66.86	84.35	81.82	83.56
SynDoc (Gemini)					
Top-1	80.29	67.73	85.19	81.80	82.77
Top-K	81.60	66.90	83.57	81.33	82.12
Top-1 R	80.29	67.73	85.19	**82.15**	83.02
Top-K R	**81.91**	68.09	84.57	81.58	82.40
w/bbox	80.93	**68.13**	**85.40**	82.08	**83.87**

Table 1: Results using Zero-shot MLLM. The last row shows the best configuration with bounding boxes.

(Ding et al., 2023a), CORD (receipts) (Park et al., 2019), Ephoie (exam papers) (Wang et al., 2021), and FUNSD (Jaume et al., 2019) (multi-domains). (Appendix A.1 for more details). Form-NLU was further divided into Printed (F-P) and Handwritten (F-H) subsets. The document images in each **test set** were processed using the *Synthetic Data Generation* module to produce synthetic structure annotations and QA pairs with verification results. During inference, QA pairs or key-value/question pairs from the original dataset are utilized.

For the FUNSD and CORD datasets, we utilized the processed test sets from (Luo et al., 2024). For Form-NLU and Ephoie, we converted the key-value pairs into QA pairs for inference. Consistent with (Mathew et al., 2021; Luo et al., 2024), we used the Averaged Normalized Levenshtein Similarity (ANLS) as our primary **evaluation metric**.

4.2 Baselines and Implementation Details

We compared SynDoc with state-of-the-art baselines (Appendix B). These include both open source (i.e., Qwen2-VL (Wang et al., 2024), Idefics2 (Laurençon et al., 2024), and InternVL2 (Chen et al., 2024)) and proprietary models (i.e., GPT-4o (OpenAI, 2024) and Gemini 1.5 (Team et al., 2024)). We selected these models due to their remarkable performance on various document-related benchmarks.

All MLLMs were tested using their default settings in the Huggingface environment[3] with access to up to 2× A100 80G GPUs.

5 Results and Discussion

5.1 Main Results

Table 1 shows that proprietary models generally outperform their open-source counterparts. This advantage is particularly evident in complex scenarios (e.g., F-H and Ephoie). Among similarly sized open-source MLLMs, Qwen2-VL achieves the highest performance, benefiting from its extensive multimodal training data and advanced OCR capabilities. Intern-VL2 also demonstrates strong performance across all datasets, whereas Idefics2 encounters challenges, particularly with structurally complex documents in Ephoie.

Since Gemini shows better performance across most benchmark datasets compared to GPT-4o, we present the results of the Gemini-based SynDoc framework. Overall, incorporating adaptively tuned warmer knowledge into MLLMs enhances performance on domain-specific datasets; however, it may introduce noise in cross-domain benchmarks such as FUNSD. The results also suggest that employing top-K candidate hints or recursive inference (top-K R) substantially improves MLLM performance in zero-shot scenarios.

[3] https://huggingface.co/

Adapt	St	Prior	F-P	F-H	CORD	Ephoie	FUNSD
1	✗	✗	31.39	18.18	41.48	19.23	44.37
2	✗	✗	42.56	16.41	46.71	20.64	48.66
3	✗	✗	33.87	14.61	41.16	22.74	42.77
4	✗	✗	44.23	12.23	50.44	23.78	44.67
1	✗	✓	59.26	30.67	65.6	22.94	56.83
2	✗	✓	65.67	31.63	66.37	22.06	57.77
3	✗	✓	64.68	27.85	65.9	25.48	57.43
4	✗	✓	65.75	29.31	65.08	24.76	59.86
1	✓	✓	62.67	30.25	66.21	24.12	58.08
2	✓	✓	66.03	**31.64**	**67.26**	24.13	58.05
3	✓	✓	65.2	28.83	63.94	25.29	61.01
4	✓	✓	**66.19**	28.29	66.25	**27.16**	**61.24**

Table 2: Results under various Warmer Adaptive Tuning Configurations. Adapt - Four types of adaptive tuning sets: (1) full synthetic set, (2) meaningful verification filtered set, (3) consistency verification filtered set, and (4) dual verification filtered set. St - structure adaptation. Prior - prior MLLM outputs.

5.2 Warmer Performance Analysis

Here, we evaluated the effectiveness of the *Synthetic Data Generation* workflow and *Warmer*'s capability to capture domain-specific knowledge.

Adaptive Tuning Strategies. We first evaluated the adaptive tuning methods in three settings.

i) Effects of adaptive tuning sets. Table 2 shows that both verification methods improve performance and enhance domain adaptation. However, meaningfulness verification consistently provides performance gains, while consistency verification can sometimes negatively affect tuning. This negative impact may be attributed to OCR errors, which

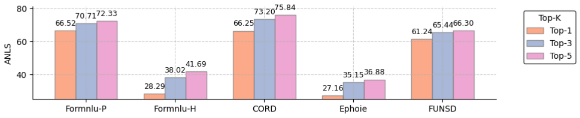

Figure 5: Top-K retrieved entity performance using LayoutLMv3 as the backbone.

can lead to inaccurate MLLM justifications.

ii) Impact of prior MLLM outputs. Table 2 also shows that incorporating MLLM outputs as Warmer input helps Warmer efficiently locate relevant information with improved accuracy.

iii) Structural Adaption Tuning (St) is introduced to enhance the Warmer model by improving its comprehension of layout and semantic correlations within a specific domain. Table 2 consistently demonstrates its efficacy across all datasets. The result indicates that the proposed self-supervised structural adaptation effectively warms up the Warmer, enabling it to capture richer structural and semantic correlations while enhancing subsequent semantic adaptation.

Top-K Retrieved Entity Performance. Here, we compared the Top-1, Top-3, and Top-5 retrieved entities, selecting the entity with the highest ANLS when multiple entities are given. Figure 5 shows that the Top-3 predictions significantly improve the retrieval of relevant information compared to Top-1. However, the performance gain between Top-3 and Top-5 is marginal. Notably, for datasets with lower OCR accuracy, the improvement from Top-1 to Top-3 is more pronounced, indicating the benefit of broader retrieval in error-prone scenarios.

Various Warmer Backbones. We selected three commonly used models to assess the effectiveness of various backbones: the text-only RoBERTa (Liu, 2019), the text and layout-aware LiLT (Wang et al., 2022a), and the text, layout, and vision-aware LayoutLMv3 (Huang et al., 2022). Table 3 shows that multimodal frameworks tend to outperform the monomodal RoBERTa, particularly when OCR errors impact the input text sequence. However, LayoutLMv3-Chinese exhibits weaker feature representation, significantly underperforming compared to LiLT and RoBERTa, despite all three using the same xlm-RoBERTa-base checkpoints. Interestingly, there are instances where the monomodal RoBERTa outperforms multimodal backbones, indicating that multimodal architectures do not always guarantee superior performance or enhanced domain-specific knowledge extraction.

Model	F-P	F-H	CORD	Ephoie	FUNSD
Roberta	64.18	23.85	**70.40**	31.57	59.44
LiLT	63.82	30.89	67.87	**31.97**	**60.94**
LayoutLMv3	**65.75**	**31.63**	66.37	25.48	59.86

Table 3: Results under different Warmer backbones.

	F-P		F-H		CORD		Ephoie	
Model	Vani.	Ours	Vani.	Ours	Vani.	Ours	Vani.	Ours
InternVL	66.56	↑ 68.09	45.47	↑ 46.81	66.84	↑ 68.8	68.92	↑ 70.29
QWenVL	78.05	↓ 77.27	43.65	↑ 44.43	77.86	↑ 78.44	70.36	↑ 75.03
Gemini	76.09	↑ 81.91	66.86	↑ 68.02	84.35	↑ 85.19	81.82	↑ 82.15

Table 4: Comparison of Warmer to Generative Models.

5.3 Recursive Inference Results

Here, we assessed how effectively the zero-shot trained Warmer enhances MLLM inference and explored the impact of the recursive inferencing mechanism across various MLLMs.

Performance on Various MLLMs. Table 4 presents the results of two high-performing open-source models (InternVL and QWenVL) and the best-performing proprietary model (Gemini). The result shows that the inclusion of Warmer outputs consistently improves performance across all models and datasets.

Effectiveness of Top-K Candidates. Figure 6 shows that providing top-K candidates from the warmer can enhance the likelihood of integrating relevant extracted information into MLLMs and improve performance. For instance, in FormNLU, retrieving additional information from the warmer can guide Gemini to focus on relevant context, thereby enhancing its performance. However, this approach also introduces the risk of incorporating noise into the prompt, which may negatively impact the generative model's performance. This effect is particularly notable in InternVL2 and QWenVL2, when applied to datasets with OCR-challenging like F-H and Ephoie.

Effectiveness of Iterative Tuning. Table 5 shows that models exhibit improved performance when more than one iteration is conducted. This demonstrates that Warmer and the LLM generator can mutually reinforce each other, enabling the model to generate more accurate final predictions. Additionally, we observed that open-source models (InternVL, QWenVL) typically require more iterations to reach peak performance, while the closed-source Gemini often achieves its best results with fewer iterations. Moreover, datasets that present

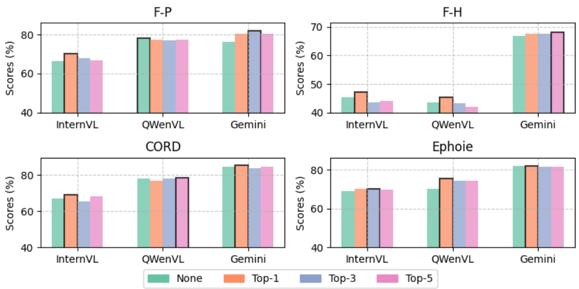

Figure 6: Result comparison by feeding Top-K Warmer-Retrieved Candidates into MLLM.

Iter.	F-P			F-H			CORD			EPHOIE		
	Int	Qw	Gemi	Int	QW	Gemi	Int	QW	Gemi	Int	QW	Gemi
Vani.	66.56	78.05	76.09	45.47	43.65	66.86	66.84	77.86	84.35	68.92	70.36	81.82
Iter 1	68.09	76.53	**80.29**	46.81	44.43	67.73	**68.80**	76.93	**85.19**	68.54	75.03	81.80
Iter 2	**70.12**	77.22	80.17	46.17	**45.27**	67.60	67.89	76.70	84.67	69.49	**75.55**	**81.91**
Iter 3	68.54	76.75	80.15	**47.23**	44.50	67.32	67.29	76.93	84.65	**70.24**	75.44	81.71
Iter 4	68.28	**77.27**	79.88	45.54	45.26	**67.63**	66.84	76.70	84.39	68.99	75.55	82.15
Iter 5	70.21	76.75	80.06	44.86	44.51	67.63	67.28	**76.93**	84.40	70.07	75.44	81.86

Table 5: Performance trends of iterative tuning. Int: InternVL2; QW: QWenVL2; Gemi: Gemini.

OCR challenges (F-H and Ephoie) benefit from additional iterations, with all models requiring at least two iterations for optimal performance.

Recursive Warmer Performance. Table 6 shows that recursive inference enhances both discriminative Warmer and generative MLLM performance. Notably, the FormNLU dataset exhibits significant improvement, with scores rising from 66.19 to 73.76 on the printed set and from 31.64 to 39.15 on the handwritten set. An interesting finding is that the performance peaks for Warmer and MLLM do not always coincide at the same iteration. This may suggest that while Warmer improves retrieval, Gemini might not immediately capitalize on these improvements due to its integration and reasoning process.

6 Case Study

To further illustrate the effectiveness of SynDoc, Figure 7 visualizes several examples where initial MLLM predictions are refined using SynDoc[4]. In ***Q1***, a question regarding the present voting count initially yields an incorrect answer of 15,41, which is subsequently corrected to 27,210 with the aid of the warmer. This example highlights how the warmer effectively introduces domain-specific knowledge, mitigating hallucinations and reducing the imprecision of MLLM predictions.

Additionally, relying solely on the Top-1 retrieved answer from the warmer may not always capture the most relevant information needed for accurate answering. As demonstrated in ***Q2***, providing Top-3 entities enhances performance by leveraging both the warmer's domain knowledge and the MLLM's general world knowledge, thereby refining the final prediction.

The last example ***Q3*** highlights the effectiveness of the iterative inference mechanism. Here, the warmer and MLLM incrementally improve each other's performance, leading to an almost correct prediction. Notably, even when the warmer provides the perfect hints in the final iteration, OCR errors may still be present. However, the MLLM compensates by leveraging its large-scale general world knowledge to generate the correct prediction.

	F-P		F-H		CORD		Ephoie	
	Warmer	Gemini	Warmer	Gemini	Warmer	Gemini	Warmer	Gemini
Vanilla	66.19	76.09	31.64	66.86	**67.26**	84.35	27.16	81.82
1	73.57	**80.29**	38.11	**67.73**	63.37	**85.19**	**27.98**	81.80
2	**73.76**	80.17	38.79	67.60	64.15	84.67	25.94	81.91
3	73.72	80.15	**39.15**	67.32	64.32	84.65	26.03	81.71
4	73.76	79.88	38.84	67.63	64.32	84.39	25.94	**82.15**
5	73.60	80.06	38.92	67.63	64.04	84.40	26.12	81.86

Table 6: Impact of iterations on Warmer and Gemini.

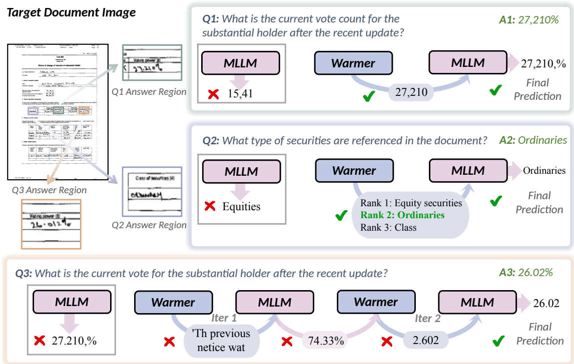

Figure 7: Qualitative Case Studies.

7 Conclusion

In this paper, we introduced a novel VRDU framework, SynDoc, which effectively integrates discriminative VLPMs and generative MLLMs to advance domain-specific VRDU performance, particularly in zero-shot settings. Our extensive experiments show that the proposed *Synthetic Data Generator* and *Adaptive Warmer Tuning* enable the discriminative warmer to efficiently acquire domain knowledge and, together with recursive inference, drive continual performance gains for both the warmer and the MLLM. While the framework exhibits robust results on multiple domain-specific datasets, however, further enhancements may be required to maximize generalizability and robustness in cross-domain applications.

[4] Please refer to Appendix for more case studies.

Limitations

While SynDoc achieves strong results in domain-specific VRDU tasks, it has several limitations. The framework's performance is sensitive to the quality of synthetic data and the accuracy of external tools like OCR and PDF parsers, making it vulnerable to errors from noisy or complex documents. Its domain adaptation strategy, though effective within target domains, often struggles to generalize across diverse document types, as shown by performance drops in cross-domain settings such as FUNSD dataset. Additionally, the iterative inference process increases computational cost, and the current evaluation is limited to a handful of public datasets, leaving broader real-world applicability for future exploration.

References

Ofir Abramovich, Niv Nayman, Sharon Fogel, Inbal Lavi, Ron Litman, Shahar Tsiper, Royee Tichauer, Srikar Appalaraju, Shai Mazor, and R Manmatha. 2024. Visfocus: Prompt-guided vision encoders for ocr-free dense document understanding. In *European Conference on Computer Vision*, pages 241–259. Springer.

Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. 2023. Docformerv2: Local features for document understanding. *arXiv preprint arXiv:2306.01733*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

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

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

Yihao Ding, Siqu Long, Jiabin Huang, Kaixuan Ren, Xingxiang Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023a. 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.

Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023b. Pdf-vqa: A new dataset for real-world vqa on pdf documents. In *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*, pages 585–601. Springer Nature Switzerland.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024c. Mmvqa: A comprehensive dataset for investigating multipage multimodal information retrieval in pdf-based visual question answering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI*, pages 3–9.

Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. 2024. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences*, 67(12):1–14.

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

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19485–19494.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10767–10775.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024a. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024b. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *CoRR*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. 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.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Ghazal Khalighinejad, Sharon Scott, Ollie Liu, Kelly L Anderson, Rickard Stureborg, Aman Tyagi, and Bhuwan Dhingra. 2024. Matvix: Multimodal information extraction from visually rich articles. *arXiv preprint arXiv:2410.20494*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 498–517. Springer.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. *arXiv preprint arXiv:2404.05225*.

Pengyuan Lyu, Yulin Li, Hao Zhou, Weihong Ma, Xingyu Wan, Qunyi Xie, Liang Wu, Chengquan Zhang, Kun Yao, Errui Ding, and 1 others. 2024. Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond. *arXiv preprint arXiv:2405.21013*.

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

Dario Augusto Borges Oliveira and Matheus Palhares Viana. 2017. Fast cnn-based document layout analysis. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1173–1180. IEEE.

OpenAI. 2023. Chatgpt: A conversational agent.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

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

Minenobu Seki, Masakazu Fujio, Takeshi Nagasaki, Hiroshi Shinjo, and Katsumi Marukawa. 2007. Information management system using structure analysis of paper/electronic documents and its applications. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 689–693. IEEE.

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

Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Hao Feng, Yang Li, Siqi Wang, Lei Liao, and 1 others. 2024. Textsquare: Scaling up text-centric visual instruction tuning. *arXiv preprint arXiv:2404.12803*.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. 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.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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

Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021. 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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wenjin Wang, Zhengjie Huang, Bin Luo, Qianglong Chen, Qiming Peng, Yinxu Pan, Weichong Yin, Shikun Feng, Yu Sun, Dianhai Yu, and 1 others. 2022b. Ernie-mmlayout: Multi-grained multimodal transformer for document understanding. *arXiv preprint arXiv:2209.08569*.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023. Vrdu: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5184–5193.

Toyohide Watanabe, Qin Luo, and Noboru Sugie. 1995. Layout recognition of multi-kinds of table-form documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(4):432–445.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, and 1 others. 2021a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

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

Ruiyi Zhang, Yufan Zhou, Jian Chen, Jiuxiang Gu, Changyou Chen, and Tong Sun. 2024. Llava-read: Enhancing reading ability of multimodal language models. *arXiv preprint arXiv:2407.19185*.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866.

A Detailed Dataset Information

A.1 Dataset Description

Form-NLU (Ding et al., 2023a) is introduced for financial-domain form layout and content understanding, focusing on single-template, multi-format forms, including digital, printed, and handwritten variations. This dataset specifically addresses KIE tasks, which involve extracting 12 types of key information from more challenging printed and handwritten documents. Examples of these key information fields include "*Substantial Holder Name*", "*Previous Persons' Votes*", and others.

CORD (Park et al., 2019) is proposed for receipt understanding with diverse receipt templates. This dataset focuses on the sub-task of KIE to extract fine-grained key information from scanned receipts, such as "*store name*" and "*item quantity*".

Ephoie (Wang et al., 2021) is a dataset proposed for understanding scanned Chinese exam paper headers. The collected exam papers have diverse templates and handwritten information. This dataset focuses on the KIE sub-task to extract information from these exam papers, such as "*Score*," "*School*," and "*Student Name*."

FUNSD (Jaume et al., 2019) is a dataset for form understanding, comprising scanned form images from diverse sources with varying templates. Each form contains predefined key-value pairs categorized as "*Question*" and "*Answer*" in the metadata. This dataset is utilized to assess the capability of the proposed framework in handling cross-domain scenarios.

Domain	Category	# Doc	# QA	Set 1	Set 2	Set 3	Set 4
FormNLU-P	Financial Form	50	596	1937	1137	1073	676
FormNLU-H	Financial Form	50	597	1998	621	815	302
CORD	Receipt	100	156	1644	1535	988	968
EPHOIE	Exam Paper	311	928	2488	1746	1553	1159
FUNSD	Cross-domain	50	467	2036	1905	1088	1022

Table 7: Dataset statistics across different dataset including the size of original test and the synthetic dataset.

B Detailed Model Information

B.1 Warmer Variants Details

RoBERTa (Liu, 2019): RoBERTa is a self-supervised text-only language model trained on a large corpus, including BookCorpus, English Wikipedia, CommonCrawl News, OpenWebText, and Stories datasets. RoBERTa removes the next-sentence prediction (NSP) objective and uses dynamic masking, larger batch sizes, and longer sequences.

LiLT (Wang et al., 2022a): LiLT (Language-independent Layout Transformer) extends pretrained text encoders with a lightweight layout encoder. It is pretrained on the IIT-CDIP scanned document corpus. LiLT features a dual-stream architecture to separately encode text and layout (bounding box) information, with Bi-directional Attention Complementation (BiACM) to enhance cross-modal alignment.

LayoutLMv3 (Huang et al., 2022): LayoutLMv3 is a multimodal Transformer that jointly encodes text, layout, and image information. It is pretrained on the IIT-CDIP corpus and synthetic document data, using masked language modeling (MLM), masked image modeling (MIM), and word-patch alignment (WPA) tasks.

B.2 Large Vision-Language Models details

B.2.1 Close Source Models

GPT-4o (OpenAI, 2024): GPT-4o is a multimodal model capable of processing text, images, and audio, with an estimated size in the hundreds of billions to 1 trillion parameters. Trained on web-scale text, images, and audio, GPT-4o features native multimodal reasoning, multilingual support, and high-speed inference.

Gemini 1.5 (Team et al., 2024): Gemini 1.5 Pro is a mid-size multimodal model with a Mixture-of-Experts (MoE) architecture, trained on a vast multimodal corpus with a focus on long-context tasks up to 1 million tokens.

B.2.2 Open Source Models

InternVL2 (Chen et al., 2024): InternVL2 combines a vision Transformer and a language model. It is pretrained on 5M curated multimodal samples, including documents, forms, scientific charts, and medical images. InternVL2 ranges from 1B to 108B parameters, pretrained on curated multimodal data including documents, forms, scientific charts, and medical images. It achieves competitive results on specific document-centric tasks, such as DocVQA.

QwenVL2 (Wang et al., 2024): QwenVL2 is trained on 1.4T tokens, including image-text pairs, OCR data, video, and interleaved documents. With innovations like Naive Dynamic Resolution and Multimodal RoPE, QwenVL2 achieves competitive performance on multimodal benchmarks, establishing itself as a leading open-source option.

Model	Params	Modality	Training Data	Status
RoBERTa	125M	Text	Web, Books	Open
LiLT	131M	Text+Layout	IIT-CDIP	Open
LayoutLMv3	133M	Text+Layout+Vision	IIT-CDIP	Open
GPT-4o	~200B	Text+Vision+Audio	Web+Images+Audio	Closed
Gemini 1.5	175B	Text+Vision+Audio	Web+Multimodal	Closed
InternVL2	8B	Text+Vision	Documents, Medical	Open
QwenVL2	72B	Text+Vision+Video	Web, OCR, Video	Open
Idefics2	8B	Text+Vision	Web, Documents	Open

Table 8: Baseline Models for Visual-rich Document Understanding (Appendix)

Idefics2 (Laurençon et al., 2024): Idefics2 combines a Mistral-7B language model with a SigLIP vision encoder. Trained on interleaved web documents, captions, OCR data, and diagram-text mappings, it supports arbitrary sequences of text and images. Despite its smaller size, it achieves comparable performance to 30B+ models.

C Detailed Prompts

We list all the prompts used in this paper for synthetic data generation in Table 9 and MLLM zero-shot testing in Table 10.

D Computational Cost

Table 11 presents the training and inference resource consumption across five benchmark datasets with a consistent batch size of 16. The GPU memory usage remains within a reasonable range (approximately 25.5GB–28GB), demonstrating the framework's efficiency and scalability on standard hardware. The structural and semantic training times per epoch are well-balanced, typically ranging from 2 to 8 minutes, depending on dataset complexity. Notably, the inference time remains minimal—under 2.5 minutes for all datasets—highlighting the framework's practical deployment potential. These results indicate that the proposed framework achieves a favorable trade-off between training cost and performance, making it suitable for both research and real-world applications.

E Additional Evaluation Results

E.1 Various Prompt Method Performance

We present the results obtained using various prompting methods for baseline MLLMs and the Gemini-based SynDoc framework. The findings indicate that multimodal prompting, which integrates OCR-extracted textual context with document images, generally enhances performance. However, the OCR Challenging dataset exhibits difficulties

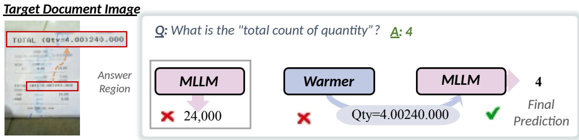

Figure 8: Qualitative case studies about CORD dataset for demonstrating the effectiveness of Warmer retrieved the content and the MLLM self-correction ability for OCR-error.

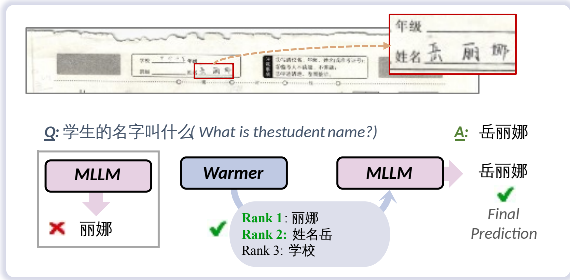

Figure 9: Qualitative case studies about Ephoie dataset for demonstrating the effectiveness of Top-K.

in certain cases. For image-only prompting, some open-source models demonstrate relatively lower performance. Consequently, our SynDoc framework adopts the Image + Text context prompt as the primary approach for overall evaluation and ablation studies.

E.2 More Detailed Experimental Results

We provide the detailed experimental results of different configurations for the MLLM inferencing, from Table 13 to Table 25.

F Additionaly Case Studies

Figures 8 and 9 present qualitative case studies from the CORD and Ephoie datasets, respectively, highlighting the complementary strengths of MLLM-based self-correction pipeline and the Top-K retrieval. In Figure 8, the MLLM initially predicts "24,000", and the Warmer module retrieves a noisy string "Qty=4.00240.000". Despite the noise, the final MLLM module successfully interprets the correct answer as "4", demonstrating its robustness to OCR errors and its ability to reason over imperfect retrieved content. In Figure 9, a query about a student's name is given, where the initial MLLM output is incorrect. However, the Warmer module retrieves relevant entities, ranking the correct answer within the Top-3, which enables the final MLLM stage to recover the accurate result. These examples collectively demonstrate the pipeline's

Module	Prompt Description	Prompt Template
User-Input Verification	Checks whether the target information was entered by the user or is part of the form template.	Based on the provided Context `{}` from the target form and the form image itself, check if the target information itself (do not consider the context) "`{}`" was entered by the form user (not part of the form template). Only output "Yes" if the `{}` is exactly provided by user not from the form template, do not consider context information. The response should follow the format below: "Response": "Yes/No"
Semantic Question Generation	Generates a short human-asked question where the answer exactly matches the target.	Based on the above context `{}` and target document image, generate a human-asked SHORT question (output question only) of which answer is exactly same as "`{}`"
Answer Verification	Verifies whether the given target could be the expected answer to the given question.	Ignore the context information and domain knowledge (e.g. FAX NUMBER). Just consider whether '`{}`' could be the expected answer to the question '`{}`'. Output format: `{'Response': 'Yes/No', 'Explanation': 'xxx'}`.
Layout-Aware Question Reformulation	Reformulates a question into a short question about the location of the answer in the document.	Change the question `{}` to a very short question about finding the position of the answer from input document image. For example, where is the answer of xx located?

Table 9: Synthetic Data Generator Prompt Example

effectiveness in overcoming early-stage retrieval errors and OCR-related noise in complex document QA tasks.

Module	Prompt Description	Prompt Template
Text-Image QA without Tips	Generates a response to a question based on an image and text context, without any additional Tips.	Above is the context { } of the target { }. Please answer the question '{ }' based on the context and image. The output format must strictly follow: Answer: xxx
Text-Image QA with One Tip	Generates a response to a question based on an image and text context, with a single Tip.	The above is the context { } of the target { }. This is a Tip: '{ }' (which may not be correct). Please answer the question '{ }' based on the context and image. The output format must strictly follow: Answer: xxx
Text-Image QA with Multiple Tips	Generates a response to a question based on an image and text context, with multiple ranked Tips.	The above is the context { } of the target { }. These are the Tips (which may not be correct): Please answer the question '{ }' based on the context and image. The output format must strictly follow: Answer: xxx
Text-Image QA with Bounding Boxes (No Tips)	Generates a response to a question based on an image, text context, and bounding box overlays, without any additional Tips.	Above is the context { } of the target { } document, Please answer the question { }, Based on the context and image, The output format strictly follows: Answer: xxx
Text-Image QA with Bounding Boxes (One Tip)	Generates a response to a question based on an image, text context, and bounding box overlays, with a single Tip.	The above is the context { } of the target { } document. This is a Tip: '{ }' (which may not be correct). Please answer the question { }, Based on the context and image, The output format strictly follows: Answer: xxx
Text-Image QA with Bounding Boxes (Multiple Tips)	Generates a response to a question based on an image, text context, and bounding box overlays, with multiple ranked Tips.	The above is the context { } of the target { } document. These are Tips: '{ }', (which may not be correct.) Please answer the question { }, Based on the context and images, The output format strictly follows: Answer: xxx

Table 10: Summary of Inference Prompt Functions and Their Templates

Dataset	Batch Size	GPU Consumption	Structural Time (1 Epoch)	Semantic Time (1 Epoch)	Inference Time
FormNLU-P	16	27983.4M	00:03:46	00:03:08	00:01:10
FormNLU-H	16	25736.0M	00:03:58	00:03:01	00:01:02
CORD	16	26174.5M	00:04:30	00:04:02	00:02:01
EPHOIE	16	27993.1M	00:06:01	00:03:12	00:01:14
FUNSD	16	25566.2M	00:08:10	00:02:01	00:00:59

Table 11: Per-epoch GPU consumption and time cost across different datasets with a fixed batch size of 16. The reported times correspond to the most effective training configurations: 2 epochs for structural adaptation and 10 epochs for semantic adaptation.

Models	**Prompt**	**Formnlu-P**	**Formnlu-H**	**CORD**	**Ephoie**	**Funsd**
InternVL2	Context-only	59.65	7.16	44.00	54.39	53.48
Qwen2-VL		72.12	10.04	65.20	61.59	68.87
Idefics2		28.52	3.33	4.33	8.90	21.98
GPT-4o		71.64	1.45	69.88	59.78	68.71
Gemini		70.88	5.91	71.53	59.94	68.21
InternVL2	Image-only	68.28	48.85	62.86	63.92	74.85
Qwen2-VL		79.17	55.35	75.85	83.79	83.06
Idefics2		46.97	35.64	51.54	2.97	58.48
GPT-4o		74.81	56.51	77.63	62.23	80.32
Gemini		79.78	66.29	81.48	76.07	83.79
InternVL2	Context + Image	66.56	45.47	66.84	68.92	74.95
Qwen2-VL		79.71	55.33	79.12	83.35	82.77
Idefics2		57.54	33.31	54.45	15.22	62.11
GPT-4o		76.16	56.49	79.05	79.40	80.05
Gemini		76.09	66.86	84.35	81.82	83.56
SynDoc	Context + Image	81.91	68.02	85.19	82.15	83.02
SynDoc	Context + Image + bbox	80.93	68.13	85.40	82.08	83.87

Table 12: Performance comparison of various models on different datasets.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	67.26	66.84	63.38	68.80	57.74	67.89	58.50	67.29	59.70	66.84	57.82	67.28
QWenVL (2B)	67.26	12.17	63.38	16.36	59.75	16.75	59.86	16.43	59.75	16.75	59.86	16.43
QWenVL (7B)	67.26	77.86	63.38	76.93	59.89	76.70	59.64	76.93	59.89	76.70	59.64	76.93
QWenVL (72B)	67.26	79.12	63.38	78.02	59.98	77.96	60.30	77.81	59.98	77.96	60.30	77.81
Gemini	67.26	84.35	63.37	85.19	64.15	84.67	64.32	84.65	64.32	84.39	64.04	84.40

Table 13: Performance comparison across iterations for different models on the CORD dataset with Top-1 warmer retrieved entity.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	66.19	66.56	73.57	68.09	68.65	70.12	70.32	68.54	69.20	68.28	69.19	70.21
QWenVL (2B)	66.19	44.85	73.57	50.34	61.63	50.45	61.82	50.52	61.76	50.48	61.84	50.54
QWenVL (7B)	66.19	78.05	73.57	76.53	72.52	77.22	73.18	76.75	72.61	77.27	73.18	76.75
QWenVL (72B)	66.19	79.71	73.57	81.21	74.41	81.42	74.54	81.20	74.58	81.42	74.54	81.20
Gemini	66.19	76.09	73.57	80.29	73.76	80.17	73.72	80.15	73.76	79.88	73.60	80.06

Table 14: Performance comparison across iterations for different models on the Printed dataset with Top-1 warmer retrieved entity.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	31.64	45.47	38.11	46.81	32.29	46.17	32.70	47.23	32.06	45.54	32.76	44.86
QWenVL (2B)	31.64	14.56	38.11	19.21	24.95	19.33	25.45	19.19	25.02	19.36	25.44	19.20
QWenVL (7B)	31.64	43.65	38.11	44.43	34.51	45.27	35.25	44.50	34.83	45.26	35.26	44.51
QWenVL (72B)	31.64	55.33	38.11	58.33	38.37	58.40	38.33	58.37	38.48	58.58	38.36	58.37
Gemini	31.64	66.86	38.11	67.73	38.79	67.60	39.15	67.32	38.84	67.63	38.92	67.63

Table 15: Performance comparison across iterations for different models on the Handwritten dataset with Top-1 warmer retrieved entity.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	27.16	68.92	27.98	68.54	25.78	69.49	25.94	70.24	26.00	68.99	25.96	70.07
QWenVL (2B)	27.16	46.13	27.98	36.51	27.10	37.00	26.78	36.39	27.10	36.97	26.78	36.39
QWenVL (7B)	27.16	70.36	27.98	75.03	26.79	75.55	26.76	75.44	26.79	75.55	26.76	75.44
QWenVL (72B)	27.16	83.35	27.98	81.95	26.38	82.08	26.51	82.06	26.38	82.08	26.51	82.06
Gemini	27.16	81.82	27.98	81.80	25.94	81.91	26.03	81.71	25.94	82.15	26.12	81.86

Table 16: Performance comparison across iterations for different models on the Ephoie dataset with Top-1 warmer retrieved entity.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	61.24	74.95	59.64	73.18	58.30	72.13	58.44	73.41	58.97	73.57	58.98	73.12
QWenVL	61.24	79.12	61.94	74.84	60.03	75.73	60.92	74.57	59.93	75.73	60.92	74.57
Gemini	61.24	83.56	59.17	82.77	59.77	83.02	60.06	82.38	59.54	82.91	60.11	82.36

Table 17: Performance comparison across iterations for different models on the FUNSD dataset with Top-1 warmer retrieved entity.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	67.26	66.84	63.38	61.61	60.19	65.31	53.73	64.75	53.52	61.70	54.06	62.22
QWenVL	67.26	77.86	63.38	78.16	59.65	77.96	59.34	78.12	59.65	77.96	59.34	78.12
Gemini	67.26	84.35	63.38	83.46	63.79	82.34	63.42	83.07	63.42	83.07	63.69	83.00

Table 18: Top-3 Performance comparison across iterations for different models on the CORD dataset.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	66.19	66.56	73.56	65.91	67.70	67.85	67.55	67.12	68.53	66.21	66.92	66.38
QWenVL	66.19	78.05	73.57	77.08	72.93	76.60	72.81	76.63	72.53	76.72	72.80	76.67
Gemini	66.19	76.09	73.99	81.60	74.12	81.91	74.30	81.63	74.01	81.58	74.28	81.46

Table 19: Top-3 Performance comparison across iterations for different models on the Printed dataset.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	31.64	45.47	38.11	43.48	32.64	43.24	30.92	42.02	31.75	43.15	31.93	43.52
QWenVL	31.64	43.65	38.11	42.03	33.65	43.37	33.68	41.66	32.60	42.62	33.28	41.55
Gemini	31.64	66.86	38.11	66.82	39.35	67.68	39.48	67.12	39.15	66.80	38.79	67.49

Table 20: Top-3 Performance comparison across iterations for different models on the Handwritten dataset.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	27.16	68.92	27.98	70.29	26.00	69.04	26.08	69.35	26.17	68.15	26.30	69.41
QWenVL	27.16	70.36	27.98	73.91	26.36	74.29	26.68	74.18	26.28	74.29	26.68	74.18
Gemini	27.16	81.82	27.98	81.18	26.23	81.13	26.25	81.16	26.27	81.43	26.10	81.32

Table 21: Top-3 Warmer Retrieved Entity Performance comparison across iterations for different models on the Ephoie dataset.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	67.26	66.84	63.37	64.25	57.02	63.21	54.63	68.18	55.93	66.76	57.87	65.13
QWenVL	67.26	77.86	63.38	78.20	59.54	77.49	58.91	78.44	60.08	77.53	58.91	78.16
Gemini	67.26	84.35	63.38	84.57	63.79	82.85	63.99	83.37	63.79	82.77	63.79	83.39

Table 22: Top-5 Performance comparison across iterations for different models on the CORD dataset.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	66.19	66.56	73.57	66.88	69.19	66.17	67.46	65.23	65.67	65.67	66.27	66.27
QWenVL	66.19	78.05	73.57	76.35	72.22	77.01	72.66	76.67	72.34	77.27	72.70	76.21
Gemini	66.19	76.09	73.58	80.10	73.35	80.35	73.70	80.20	73.40	80.36	73.54	80.08

Table 23: Top-5 Performance comparison across iterations for different models on the Printed dataset.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	31.64	45.47	38.11	43.82	32.79	44.13	33.66	43.78	31.87	41.55	32.11	43.22
QWenVL	31.64	43.65	38.11	40.12	32.51	41.97	33.13	40.18	32.26	41.75	32.99	39.78
Gemini	31.64	66.86	38.11	66.90	39.05	67.33	39.06	67.51	38.99	67.01	39.11	68.02

Table 24: Top-5 Performance comparison across iterations for different models on the Handwritten dataset.

Model	Baseline		Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM	Warmer	LLM
InternVL	27.16	68.92	27.98	68.88	26.18	68.66	25.93	67.90	26.06	69.61	25.87	69.77
QWenVL	27.16	70.36	27.98	74.32	26.32	74.32	26.56	74.35	26.32	74.34	26.67	74.24
Gemini	27.16	81.82	27.98	81.33	26.35	81.18	26.31	81.58	26.37	81.23	26.28	81.45

Table 25: Top-5 Performance comparison across iterations for different models on the Ephoie dataset.