A Bounding Box is Worth One Token - Interleaving Layout and Text in a Large Language Model for Document Understanding

Jinghui Lu[*1] **Haiyang Yu**[*2] **Yanjie Wang**[*1] **Yongjie Ye**[1] **Jingqun Tang**[1]
Ziwei Yang[1] **Binghong Wu**[1] **Qi Liu**[1] **Hao Feng**[1] **Han Wang**[1] **Hao Liu**[1] **Can Huang**[†1]
[1]ByteDance Inc. [2]Fudan University
lujinghui@bytedance.com, hyyu20@fudan.edu.cn
{wangyanjie.prince, yeyongjie.ilz, tangjingqun}@bytedance.com
{yangziwei.1221, wubinghong, liuqi.nero}@bytedance.com
{fenghao.2019, wanghan.99, haoliu.0128, can.huang}@bytedance.com

Abstract

Recently, many studies have demonstrated that exclusively incorporating OCR-derived text and spatial layouts with large language models (LLMs) can be highly effective for document understanding tasks. However, existing methods that integrate spatial layouts with text have limitations, such as producing overly long text sequences or failing to fully leverage the autoregressive traits of LLMs. In this work, we introduce *Interleaving* ***Lay****out and* ***Text*** *in a* ***L****arge* ***L****anguage* ***M****odel (LayTextLLM)* for document understanding. LayTextLLM projects each bounding box to a single embedding and interleaves it with text, efficiently avoiding long sequence issues while leveraging autoregressive traits of LLMs. LayTextLLM not only streamlines the interaction of layout and textual data but also shows enhanced performance in KIE and VQA. Comprehensive benchmark evaluations reveal significant improvements of LayTextLLM, with a 15.2% increase on KIE tasks and 10.7% on VQA tasks compared to previous SOTA OCR-based LLMs. All resources are available at https://github.com/LayTextLLM/LayTextLLM.

1 Introduction

Recent research has increasingly explored the use of Large Language Models (LLMs) or MultiModal Large Language Models (MLLMs) (Achiam et al., 2023; Team et al., 2023; Anthropic, 2024; Reid et al., 2024; Feng et al., 2023a,b; Liu et al., 2024c; Lu et al., 2024; Nourbakhsh et al., 2024; Gao et al., 2024; Li et al., 2024a; Zhou et al., 2024; Zhu et al., 2024; Zhao et al., 2024) for document-oriented Visual Question Answering (VQA) and Key Information Extraction (KIE).

A line of research utilizes off-the-shelf OCR tools to extract text and spatial layouts, which are then combined with LLMs to address Visually Rich Document Understanding (VRDU) tasks. These approaches assume that *most valuable information for document comprehension can be derived from the text and its spatial layouts, viewing spatial layouts as "lightweight visual information" (Wang et al., 2024a)*. Following this premise, several studies (Liu et al., 2024c; Perot et al., 2023; Luo et al., 2024; Chen et al., 2023a; He et al., 2023) have explored various approaches that integrate spatial layouts with text for LLMs and achieves results that are competitive with those of MLLMs.

The most natural method to incorporate layout information is by treating spatial layouts as tokens, which allows for the seamless interleaving of text and layout into a unified text sequence (Perot et al., 2023; Chen et al., 2023a; He et al., 2023). For example, Perot et al. (2023) employ format such as *"HARRISBURG 78|09"* to represent OCR text and corresponding layout, where *"HARRISBURG"* is OCR text and *"78|09"* indicates the mean of the horizontal and vertical coordinates, respectively. Similarly, He et al. (2023) use *"[x_min, y_min, x_max, y_max]"* to represent layout information. These approaches can effectively take advantage of autoregressive characteristics of LLMs and is known as the *"coordinate-as-tokens"* scheme (Perot et al., 2023). In contrast, DocLLM (Wang et al., 2024a) explores interacting spatial layouts with text through a disentangled spatial attention mechanism that captures cross-alignment between text and layout modalities.

However, we believe that both of the previous approaches have limitations. As shown in Figure 1, coordinate-as-tokens significantly increases the number of tokens. Additionally, to accurately comprehend coordinates and enhance zero-shot capabilities, this scheme often requires few-shot in-context demonstrations and large-scale language models, such as ChatGPT Davinci-003 (175B) (He et al., 2023), which exacerbates issues related to sequence length and GPU resource demands. Al-

[*] Equal Contribution
[†] Corresponding author

Findings of the Association for Computational Linguistics: ACL 2025, pages 7252–7273
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

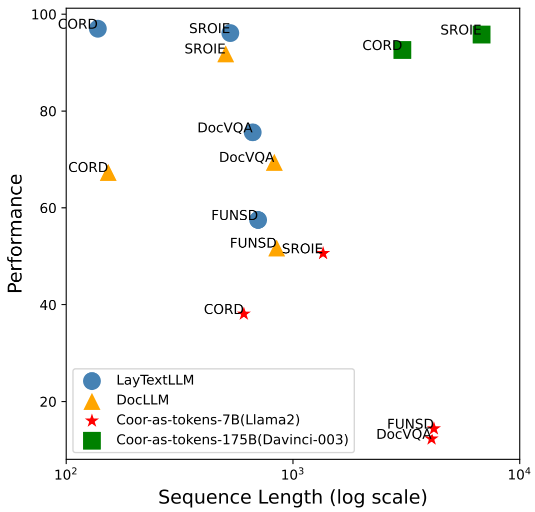


Figure 1: The performance against input sequence length of different datasets across various OCR-based methods where data is from Table 1 and 5.

though DocLLM does not increase sequence length, its performance may be improved by more effectively leveraging the autoregressive traits of LLMs.

To address these problems, this paper explores a simple yet effective approach to enhance the interaction between spatial layouts and text — *Interleaving **Layout** and **Text** in a **Large Language Model** (LayTextLLM)* for document understanding. Adhering to the common practice of interleaving any modality with text (Huang et al., 2023; Peng et al., 2023; Dong et al., 2024), we specifically apply this principle to spatial layouts. In particular, we map each bounding box to a single embedding, which is then interleaved with its corresponding text. As shown in Figure 1, LayTextLLM significantly outperforms the 175B models, while only slightly increasing or even reducing the sequence length compared to DocLLM. Our contributions can be listed as follows:

- We propose LayTextLLM for document understanding. To the best of the authors' knowledge, this is the first work to employ a unified embedding approach (Section 3.1) that interleaves spatial layouts directly with textual data within a LLM. By representing each bounding box with one token, LayTextLLM efficiently addresses sequence length issues brought by coordiante-as-tokens while fully leveraging autoregressive traits for VRDU tasks.

- We propose three tailored pre-training tasks (Section 3.2.1) to improve the model's understanding of the interaction between layout and text, and its ability to generate precise coordinates for regions of interest. These tasks include Line-level Layout Decoding, Text-to-Layout Prediction, and Layout-to-Text Prediction. Besides, we introduce Spatially-Grounded KIE (Section 3.2.2) to further enhance the model's performance on KIE task.

- Extensive experimental results quantitatively demonstrate that LayTextLLM significantly surpasses previous state-of-the-art (SOTA) OCR-based methods. Notably, it outperforms DocLLM by 10.7% on VQA tasks and 15.2% on KIE tasks (Section 4). Furthermore, it achieves superior performance on SOTA OCR-free MLLMs, such as Qwen2-VL among most KIE datasets. Ablations and visualizations demonstrate the utility of the proposed component, with analysis showing that LayTextLLM not only improves performance but also reduces input sequence length compared to current OCR-based models.

2 Related Work

2.1 OCR-based LLMs for VRDU

Early document understanding methods (Hwang et al., 2020; Xu et al., 2020, 2021; Hong et al., 2022; Tang et al., 2022) tend to solve the task in a two-stage manner, *i.e.*, first reading texts from input document images using off-the-shelf OCR engines and then understanding the extracted texts. Considering the advantages of LLMs (*e.g.*, high generalizability), some recent methods endeavor to combine LLMs with OCR-derived results to solve document understanding. Inspired by the coordinate-as-tokens" approach in ICL-D3IE (Perot et al., 2023), He et al. (2023) use numerical tokens to integrate layout information, combining layout and text into a unified sequence that maximizes the autoregressive benefits of LLMs. To reinforce the layout information while avoiding increasing the number of tokens, DocLLM (Wang et al., 2024a) designs a disentangled spatial attention mechanism to capture cross-alignment between text and layout modalities. Recently, LayoutLLM (Luo et al., 2024) utilizes the pre-trained layout-aware model (Huang et al., 2022), to insert the visual information, layout information and text information. However, these methods struggle to leverage autoregressive properties of LLMs while avoiding the computational overhead of increasing token counts. Finding a way to integrate layout information remains a challenge.

2.2 OCR-free MLLMs for VRDU

With the increasing popularity of MLLMs (Feng et al., 2023b; Hu et al., 2024; Liu et al., 2024c; Tang et al., 2024; Chen et al., 2024a; Dong et al., 2024; Li et al., 2024b; Liu et al., 2024a; Lu et al., 2025; Feng et al., 2025; Fei et al., 2025; Wang et al., 2025), various methods are proposed to solve VRDU through explicitly training models on visual text understanding datasets and perform end-to-end inference without using OCR engines. LLaVAR (Zhang et al., 2023) and UniDoc (Feng et al., 2023b) are notable examples that expand upon the document-oriented VQA capabilities of LLaVA (Liu et al., 2024b) by incorporating document-based tasks. These models pioneer the use of MLLMs for predicting texts and coordinates from document images, enabling the development of OCR-free document understanding methods. Additionally, DocPedia (Feng et al., 2023a) operates document images in the frequency domain, allowing for higher input resolution without increasing the input sequence length. Recent advancements in this field, including mPLUG-DocOwl (Ye et al., 2023), Qwen-VL (Bai et al., 2023), Qwen2-VL (Wang et al., 2024b), and TextMonkey (Liu et al., 2024c), leverage publicly available document-related VQA datasets to further enhance the document understanding capability. Although these OCR-free methods have exhibited their advantages, they still struggle with the high-resolution input to reserve more text-related details.

3 Method

In this section, we introduce LayTextLLM. We begin by detailing the model architecture, which features an innovative Spatial Layout Projector (Section 3.1) that transforms four-dimensional layout coordinates into a single-token embedding. Next, we present three layout-text alignment pretraining tasks: line-level layout decoding, text-to-layout prediction, and layout-to-text prediction (Section 3.2.1) to ensure a seamless integration of layout and text understanding. Finally, we describe the incorporation of spatially-grounded key information extraction as a auxiliary task during supervised fine-tuning (SFT) (Section 3.2.2), to enhance the performance in KIE tasks.

3.1 Model Architecture

The overall architecture of LayTextLLM is shown in Figure 2. LayTextLLM is built on the Llama2-7B-chat model (Gao et al., 2023).

Spatial Layout Projector To enable the model to seamlessly integrate spatial layouts with text, we propose a novel **S**patial **L**ayout **P**rojector (SLP). This projector employs a two-layer MLP to transform layout coordinates into bounding box tokens, facilitating the interleaving of spatial and textual information. Concretely, each OCR-derived spatial layout is represented by a bounding box defined by four-dimensional coordinates $[x_1, y_1, x_2, y_2]$, where these coordinates denote the normalized minimum and maximum horizontal (x) and vertical (y) extents of the box, respectively. The SLP maps these coordinates into a high-dimensional embedding space, enabling the LLM to process them as a single token. This is computed as:

$$z = W_2 \cdot (\text{GeLU}(W_1 \cdot c + b_1)) + b_2 \qquad (1)$$

where $c \in \mathbb{R}^4$ is the vector of bounding box coordinates, $W_1 \in \mathbb{R}^{h\times 4}$ and $W_2 \in \mathbb{R}^{d\times h}$ are weight matrices, $b_1 \in \mathbb{R}^{h\times 1}$ and $b_2 \in \mathbb{R}^{d\times 1}$ are bias vectors, h is the hidden dimension of the MLP, and d is the dimension of the final embedding. In this study, we set $h = d$. The resulting bounding box token $z \in \mathbb{R}^d$ is a high-dimensional representation of the spatial layout. Importantly, the SLP is shared across all bounding box tokens, which introduces a minimal number of parameters to the model.

Large Language Model As shown in Figure 2, the bounding box token z is interleaved with its corresponding textual embeddings and fed into the LLM. To introduce additional trainable parameters for layout information, we integrate a Partial Low-Rank Adaptation (P-LoRA) module proposed in InternLM-XComposer2 (Dong et al., 2024) detailed in Appendix A. Additionally, to improve the efficiency of coordinate decoding, we introduce 1,000 special tokens, *i.e., "<B0>"* through *"<B999>"* to represent output coordinates.

3.2 Training Tasks

LayTextLLM is pre-trained using three innovative tasks designed to align layout and text. During the SFT phase, we introduce a novel Spatially-Grounded Key Information Extraction task as a auxiliary task, which significantly enhances the model's performance on KIE-related tasks. Figures 3 and 4 illustrate the above tasks.

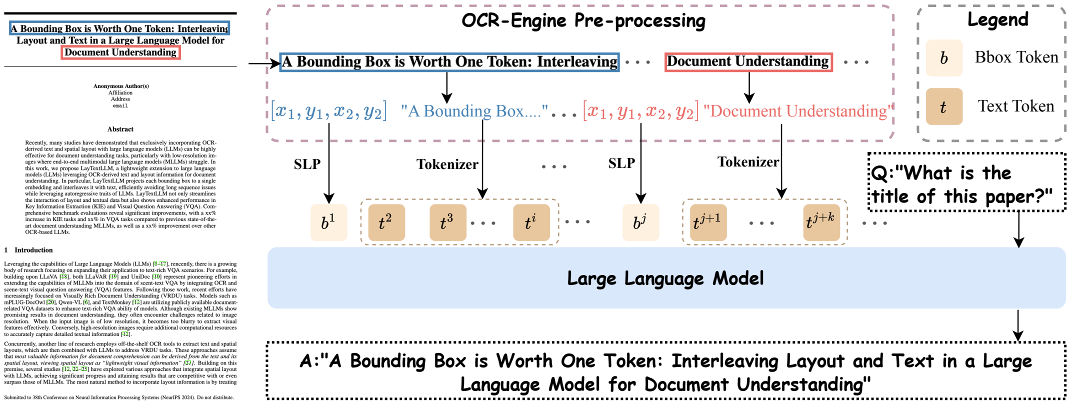
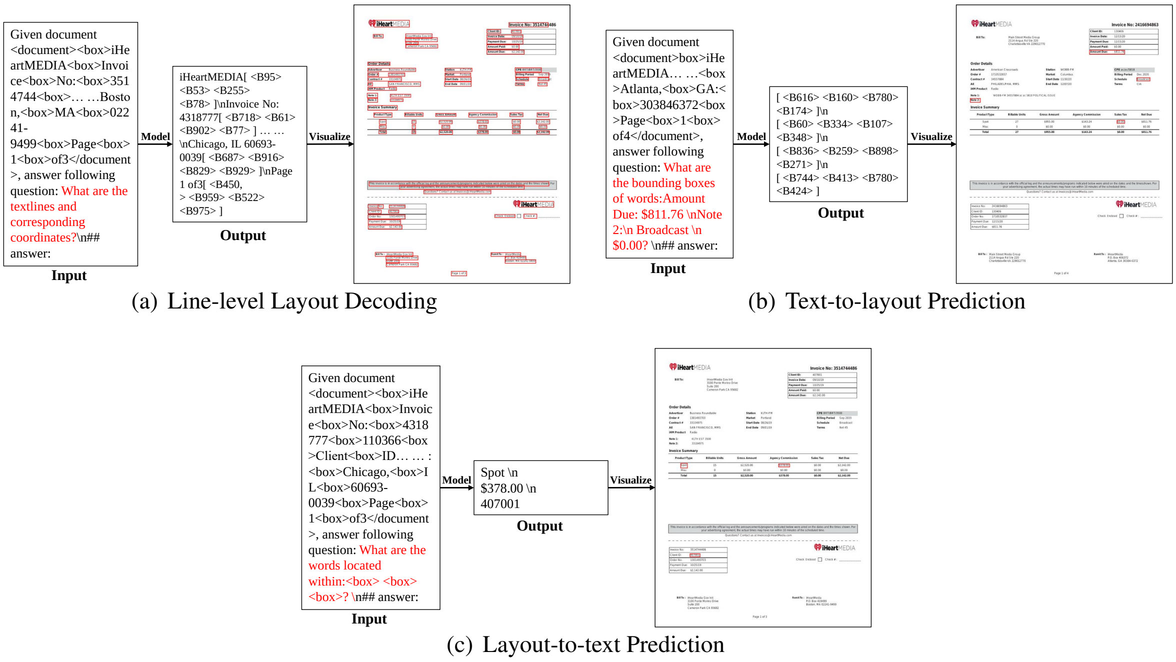

OCR-Engine Pre-processing
A Bounding Box is Worth One Token: Interleaving · · ·
Document Understanding · · ·
$[x_1, y_1, x_2, y_2]$ "A Bounding Box...." · · · $[x_1, y_1, x_2, y_2]$ "Document Understanding"
Legend
b Bbox Token
t Text Token
SLP
Tokenizer
· · ·
SLP
Tokenizer
Q:"What is the title of this paper?"
b^1 t^2 t^3 · · · t^i · · · b^j t^{j+1} · · · t^{j+k} · · ·
Large Language Model
A:"A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding"

Figure 2: An overview of LayTextLLM incorporates interleaving bounding box tokens (b^i) with text tokens (t^i), where the superscripts represent the sequence positions of the tokens.



Given document <document><box>iHeartMEDIA<box>Invoice<box>No:<box>3514744<box>… …Boston,<box>MA<box>02241-9499<box>Page<box>1<box>of3</document>, answer following question: What are the textlines and corresponding coordinates?\n## answer:

iHeartMEDIA[<B95> <B53> <B255> <B78>]\nInvoice No: 4318777[<B718> <B61> <B902> <B77>] … … \nChicago, IL 60693-0039[<B687> <B916> <B829> <B929>]\nPage 1 of3[<B450, > <B959> <B522> <B975>]

(a) Line-level Layout Decoding

Given document <document>box>iHeartMEDIA… …<box>Atlanta,<box>GA:<box>303846372<box>Page<box>1<box>of4</document>, answer following question: What are the bounding boxes of words:Amount Due: $811.76 \nNote 2:\n Broadcast \n $0.00? \n## answer:

[<B616> <B160> <B780> <B174>]\n
[<B60> <B334> <B107> <B348>]\n
[<B836> <B259> <B898> <B271>]\n
[<B744> <B413> <B780> <B424>]

(b) Text-to-layout Prediction

Given document <document><box>iHeartMEDIA<box>Invoice<box>No:<box>4318777<box>110366<box>Client<box>ID… … :<box>Chicago,<box>IL<box>60693-0039<box>Page<box>1<box>of3</document>, answer following question: What are the words located within:<box> <box> <box>? \n## answer:

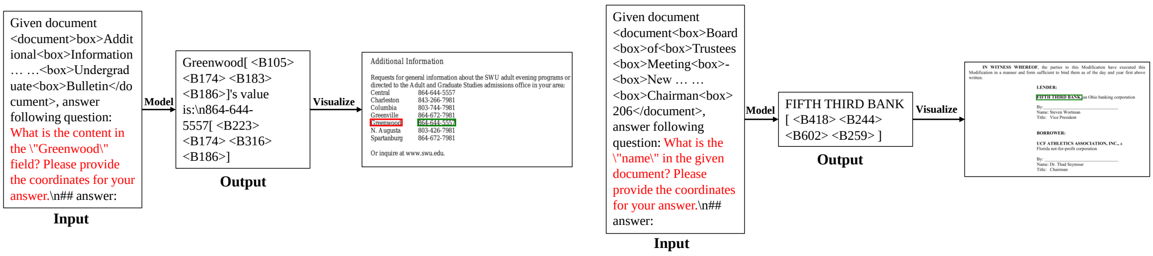
Spot \n
$378.00 \n
407001

(c) Layout-to-text Prediction

Figure 3: Illustration of layout-text alignment pre-training tasks. <box> is the placeholder for bounding box tokens.

3.2.1 Layout-text Alignment Pre-training

Line-level Layout Decoding To enhance the model's ability to interpret and reconstruct layout information, we introduce the Line-level Layout Decoding task. This task leverages the bounding box embeddings, which encode spatial layout details, and challenges the model to decode these embeddings back into precise coordinates. Specifically, the model is provided with word-level OCR texts and their corresponding layout coordinates as input. It is then prompted with the question: *"What are the textlines and corresponding coordinates?"* The model is expected to intelligently merge word-level OCR texts into coherent line-level texts while simultaneously generating the coordinates that represent the layout of these line-level texts. The output consists of two components: (1) the reconstructed line-level texts and (2) the corresponding combined coordinates, which are derived by aggregating the word-level bounding boxes to reflect the spatial arrangement of the line-level OCR. Through this task, the model is expected to demonstrate two key abilities: (1) the ability to logically group word-level texts into line-level texts using layout information, and (2) the ability to accurately decode bounding box embeddings back into spatial coordinates. By doing so, the model demonstrates a deeper understanding of both textual content and its spatial organization within a document.



Given document <document>box>Additional<box>Information … …<box>Undergraduate<box>Bulletin</document>, answer following question: What is the content in the \"Greenwood\" field? Please provide the coordinates for your answer.\n## answer:

Greenwood[<B105> <B174> <B183> <B186>]'s value is:\n864-644-5557[<B223> <B174> <B316> <B186>]

Additional Information

Requests for general information about the SWU adult evening programs or directed to the Adult and Graduate Studies admissions office in your area:

Central	864-644-5557
Charleston	843-266-7981
Columbia	803-744-7981
Greenville	864-672-7981
Greenwood	864-644-5557
N. Augusta	803-426-7981
Spartanburg	864-672-7981

Or inquire at www.swu.edu.

(a) SG-KIE for Entity Linking

Given document <document<box>Board<box>of<box>Trustees<box>Meeting<box>-<box>New … …<box>Chairman<box>206</document>, answer following question: What is the \"name\" in the given document? Please provide the coordinates for your answer.\n## answer:

FIFTH THIRD BANK [<B418> <B244> <B602> <B259>]

IN WITNESS WHEREOF, the parties to this Modification have executed this Modification in a manner and form sufficient to bind them as of the day and year first above written.

LENDER:

FIFTH THIRD BANK, an Ohio banking corporation

By:____________________
Name: Steven Wortman
Title: Vice President

BORROWER:

UCF ATHLETICS ASSOCIATION, INC., a Florida not-for-profit corporation

By:____________________
Name: Dr. Thad Seymour
Title: Chairman

(b) SG-KIE for Semantic Entity Recognition

Figure 4: Illustration of Spatially-Grounded KIE task. <box> is the placeholder for bounding box tokens.

Text-to-layout Prediction To enhance the model's ability to comprehend and predict document layouts, we introduce the Text-to-Layout Prediction task. In this task, the model predicts spatial coordinates for text segments based on word-level OCR inputs and their corresponding layout information. Specifically, given a prompt such as *"What are the bounding boxes of the words: {word1}\n {word2}\n {word3}...?"*, where {word} represents line-level text randomly selected from the input (number of selected words limited to 5), the model is required to generate precise spatial coordinates for each of the specified words.

Layout-to-text Prediction We also propose the Layout-to-Text Prediction task. In this task, the model predicts textual content based on spatial layout information and bounding box coordinates. Given a prompt such as *"What are the words located within: {bbox1}\n {bbox2}\n {bbox3}...?"*, where {bbox} is the placeholder of bounding box embedding representing the spatial coordinates of text regions (with the number of bounding boxes limited to 5), the model generates the corresponding textual content for each specified region. The Text-to-Layout Prediction and Layout-to-Text Prediction tasks offer complementary advantages to advance document layout understanding. All word-level and line-level OCR results can be easily obtained using off-the-shelf OCR tools, making it easy to scale up for large-scale pre-training.

3.2.2 Supervised Fine-tuning

During the SFT phase, we fine-tuned the pre-trained model with the Document Dense Description (DDD) and Layout-aware SFT datasets from Luo et al. (2024). Additionally, we introduce **S**patially-**G**rounded **K**ey **I**nformation **E**xtraction (SG-KIE) task, which requires the model to not only answer questions (*i.e.,* extract specific values) but also provide the coordinates of these answers by responding to the prompt *"Please provide the coordinates for your answer."* as a auxiliary task to further improve the model performance on KIE tasks.

In the literature, KIE tasks are classified into two types: Entity Linking (EL) and Semantic Entity Recognition (SER). EL is an open-set KIE task in which both the key and its corresponding value are present in the input. In contrast, SER is a closed-set KIE task where the key has a predefined meaning, and the value must be extracted from the document.

For the EL task, SG-KIE requires the model to output the answer in the following format: *"{key}{key_bbox}'s value is {value}{value_bbox}"*, where {key} and {value} represent the respective key and value, and {key_bbox} and {value_bbox} denotes the spatial layout information of the corresponding textual content. For the SER task, the answer format is: *"{value}{value_bbox}"*, where {value} refers to the extracted value, and {value_bbox} represents the spatial layout of the extracted text in the document. The illustrations of SG-KIE for these tasks are presented in Figure 4.

4 Experiments

4.1 Datasets

Layout-text Alignment Pre-training Data In training process, we exclusively used open-source data to facilitate replication. We subsampled data from two datasets for layout-text alignment pre-training: (1) **DocILE** (Šimsa et al., 2023) and (2) **RVL_CDIP** (Harley et al., 2015).

SFT data We selected **KVP10k** (Naparstek et al., 2024) and **SIBR** (Yang et al., 2023) datasets to create training examples of SG-KIE tasks. For document-oriented VQA, we selected **Document Dense Description (DDD)** and **Layout-aware SFT** data used in Luo et al. (2024), which are two synthetic datasets generated by GPT-4. Besides, **DocVQA** (Mathew et al., 2021), **InfoVQA** (Mathew et al., 2022), **ChartQA** (Masry et al., 2022), **VisualMRC** (Tanaka et al., 2021) is included following (Liu et al., 2024c). For KIE task, we selected **SROIE** (Huang et al., 2019), **CORD** (Park et al., 2019), **FUNSD** (Jaume et al., 2019) datasets following Wang et al. (2024a); Luo et al. (2024); Liu et al. (2024c). The dataset statistics are provided in Appendix C.

4.2 Implementation Detail

The LLM component of LayTextLLM is initialized from the Llama2-7B-chat (Touvron et al., 2023), consistent with previous OCR-based methods like DocLLM (Wang et al., 2024a), which also use Llama2-7B. We also replicated the results of the coor-as-tokens scheme using Llama2-7B for consistency. Noting the LayoutLLM (Luo et al., 2024) utilizes Llama2-7B and Vicuna 1.5 7B, which is fine-tuned from Llama2-7B. Thus, for the majority of our comparisons, the models are based on the same or similar LLM backbones, allowing for a fair comparison between approaches. Other MLLM baselines use backbones like Qwen-VL (Bai et al., 2023), Qwen2-VL (Wang et al., 2024b), InternVL (Chen et al., 2024b), and Vicuna (Chen et al., 2024a), all with at least 7B parameters, excluding the visual encoder. This also makes the comparison fair.

In this study, we developed two versions of LayTextLLM to facilitate a comparative analysis under different training configurations. Following the terminology established by Luo et al. (2024), the term "zero-shot" denotes models that are trained without exposure to data from downstream test datasets. For the first version, **LayTextLLM**$_{zero}$, we utilized DDD, Layout-aware SFT data, KVP10k, and SIBR for training. The second version, **LayTextLLM**$_{all}$, extends this training regimen by incorporating a broader array of VQA and KIE datasets, including DocVQA, InfoVQA, VisualMRC, ChartQA, FUNSD, CORD, and SROIE. Both versions are initialized with the same pretrained LayTextLLM weights, with the key difference being that LayTextLLM$_{all}$ benefits from the inclusion of additional downstream training datasets. We used word-level and line-level OCR provided by the respective datasets for a fair comparison, with the exception of the ChartQA dataset, which does not provide OCR. Detailed setup can be found in Appendix D.

4.3 Baselines

OCR-based baselines For OCR-based baseline models, we implemented a basic approach using only OCR-derived text as input. This was done using two versions: **Llama2-7B-base** and **Llama2-7B-chat**. We also adapted the coordinate-as-tokens scheme from He et al. (2023) for these models, resulting in two new variants: **Llama2-7B-base**$_{coor}$ and **Llama2-7B-chat**$_{coor}$. Additionally, we included results from a stronger baseline using the ChatGPT Davinci-003 (175B) model (He et al., 2023), termed **Davinci-003-175B**$_{coor}$. One other recent SOTA OCR-based approach, **DocLLM** (Wang et al., 2024a) is also included.

OCR-free baselines These baselines include **UniDoc** (Feng et al., 2023b), **DocPedia** (Feng et al., 2023a), **Monkey** (Li et al., 2023), **InternVL** (Chen et al., 2023b), **InternLM-XComposer2** (Dong et al., 2024), **TextMonkey**, **TextMonkey**$_{+}$ (Liu et al., 2024c), **Qwen2-VL** (Wang et al., 2024b). We selected the above models as baselines due to their superior performance in both document-oriented VQA and KIE tasks.

Visual+OCR baselines We selected **LayoutLLM**$_{Llama2^{CoT}}$ (Luo et al., 2024) and the most recent SOTA method **DocLayLLM**$_{Llama2^{CoT}}$ (Liao et al., 2024), which integrates visual cues, text and layout, as stronger baselines.

4.4 Evaluation Metrics

To ensure a fair comparison with other OCR-based methods, we conducted additional evaluations using original metrics specific to certain datasets, such as F1 score (Wang et al., 2024a; He et al., 2023), ANLS (Gao et al., 2019; Wang et al., 2024a; Luo et al., 2024) and CIDEr (Vedantam et al., 2015; Wang et al., 2024a). To ensure a fair comparison with OCR-free methods, we adopted the accuracy metric (Liu et al., 2024c; Feng et al., 2023b), where a response from the model is considered correct if it fully captures the ground truth.

4.5 Quantitative Results

Comparison with SOTA OCR-based Methods For the primary comparison in our work, we evaluate against other SOTA pure OCR-based methods. The experimental results, as presented in Table 1, demonstrate significant performance improvements achieved by the LayTextLLM models compared to DocLLM (Wang et al., 2024a). Specifically,

	Document-Oriented VQA			**KIE**			
	DocVQA	VisualMRC	Avg	FUNSD	CORD	SROIE	Avg
Metric	*ANLS % / CIDEr*			*F-score %*			
Text							
Llama2-7B-base	34.0	182.7	108.3	25.6	51.9	43.4	40.3
Llama2-7B-chat	20.5	6.3	13.4	23.4	51.8	58.6	44.6
Text + Coordinates							
Llama2-7B-base$_{coor}$ (He et al., 2023)	8.4	3.8	6.1	6.0	46.4	34.7	29.0
Llama2-7B-chat$_{coor}$ (He et al., 2023)	12.3	28.0	20.1	14.4	38.1	50.6	34.3
Davinci-003-175B$_{coor}$ (He et al., 2023)	-	-	-	-	92.6	95.8	-
DocLLM (Wang et al., 2024a)	69.5*	264.1*	166.8	51.8*	67.4*	91.9*	70.4
LayTextLLM$_{zero}$ (Ours)	66.6	229.1	147.9	57.6	87.3	89.4	78.1
LayTextLLM$_{all}$ (Ours)	**75.6***	**279.4***	**177.5**	**63.3***	**97.3***	**96.0***	**85.6**

Table 1: Comparison with SOTA OCR-based methods. The asterisk(*) indicates that the model was trained using the training set associated with the evaluation set.

	Document-Oriented VQA			**KIE**				
	DocVQA	InfoVQA	Avg	FUNSD	SROIE	POIE	CORD	Avg
Metric	*Accuracy %*							
OCR-free								
UniDoc (Feng et al., 2023b)	7.7	14.7	11.2	1.0	2.9	5.1	-	-
DocPedia (Feng et al., 2023a)	47.1*	15.2*	31.2	29.9	21.4	39.9	-	-
Monkey (Li et al., 2023)	50.1*	25.8*	38.0	24.1	41.9	19.9	-	-
InternVL (Chen et al., 2023b)	28.7*	23.6*	26.2	6.5	26.4	25.9	-	-
InternLM-XComposer2 (Dong et al., 2024)	39.7	28.6	34.2	15.3	34.2	49.3	-	-
TextMonkey (Liu et al., 2024c)	64.3*	28.2*	46.3	32.3	47.0	27.9	-	-
TextMonkey$_{+}$ (Liu et al., 2024c)	66.7*	28.6*	47.7	42.9	46.2	32.0	-	-
Qwen2-VL (Wang et al., 2024b)	**81.4***	**45.2***	**63.3**	53.2	71.3	**85.7**	78.8	72.2
Text + Coordinates								
LayTextLLM$_{zero}$ (Ours)	70.4	29.8	50.1	54.9	88.3	65.1	86.9	73.8
LayTextLLM$_{all}$ (Ours)	77.7*	40.1*	59.0	**60.1***	**95.5***	68.1	**96.7***	**80.1**

Table 2: Comparison with SOTA OCR-free MLLMs.

LayTextLLM$_{zero}$ exhibits notably superior performance, with its zero-shot capabilities even rivaling SFT approaches. For instance, in the KIE task, LayTextLLM$_{zero}$ achieves an overall performance of 78.1%, significantly outperforming DocLLM's score of 70.4%. Furthermore, under the same training conditions, LayTextLLM$_{all}$ surpasses the previous OCR-based SOTA by a substantial margin, achieving an overall improvement of 10.7% in the VQA task and 15.2% in the KIE tasks. Besides, we found that the spatial information can be decoded back into coordinates even without visual information, as discussed in Appendix I, which is not exhibited in DocLLM. Similarly, when contrasting with coordinate-as-tokens employed in Llama2-7B, LayTextLLM$_{zero}$ again outperforms significantly. More qualitative results are shown in Appendix B. More discussion of subperformance of DocLLM and coordinate-as-tokens can be seen Appendix F.

Comparison with SOTA OCR-free Methods We also compare LayTextLLM with other OCR-free methods, and the results in Table 2 highlight its exceptional performance across various tasks. Due to fairness concerns, results for ChartQA are reported separately in Appendix G, as the dataset lacks OCR-derived outputs, and we employed in-house OCR tools instead.

LayTextLLM$_{zero}$ significantly outperforms most OCR-free methods except for Qwen2-VL. Notably, even without exposure to the dataset's training set, LayTextLLM$_{zero}$ achieves competitive VQA performance, rivaling models like TextMonkey+, which were trained on corresponding datasets. When fine-tuned with relevant data, LayTextLLM$_{all}$ exhibits even greater performance improvements. Compared to the SOTA MLLM Qwen2-VL, LayTextLLM sub-performs on VQA tasks which is further discussed in Limitation (Section 5). However, it outperforms Qwen2-VL in terms of KIE tasks. Notably, LayTextLLM$_{zero}$ exceeds Qwen2-VL on three out of four KIE benchmarks, with significant improvements of 1.7% on FUNSD, 17% on SROIE, and 8.1% on CORD.

Comparison with SOTA Visual+OCR Methods As shown in Table 3, in zero-shot scenarios, our approach outperforms LayoutLLM and DocLayLLM on most KIE datasets, with improvements of 12.4% and 5.4%, respectively. This is noteworthy given that both LayoutLLM and DocLayLLM utilize visual, OCR text, and layout information as inputs and inference with Chain-of-thought, highlighting our ability to effectively leverage OCR-based results. However, similar to the comparison results with MLLMs, LayTextLLM exhibits limitations in

	Document-Oriented VQA			KIE			
	DocVQA	VisualMRC	Avg	FUNSD$^-$	CORD$^-$	SROIE$^-$	Avg
Metric	*ANLS %*						
Visual + Text + Coordinates							
LayoutLLM$_{Llama2CoT}$ (Luo et al., 2024)	74.2	**55.7**	**64.9**	78.6	62.2	70.9	70.6
DocLayLLM$_{Llama2CoT}$ (Liao et al., 2024)	72.8	55.0	63.9	78.7	70.8	83.2	77.6
Text + Coordinates							
LayTextLLM$_{zero}$ (Ours)	66.6	37.9	52.3	79.0	79.8	90.2	83.0
LayTextLLM$_{all}$ (Ours)	**75.6***	42.3*	59.0	**83.4***	**83.1***	**95.6***	**87.4**

Table 3: Comparison with LayoutLLM. The superscript minus($^-$) indicates that the cleaned test set used in Luo et al. (2024).

				Document-Oriented VQA				KIE			
SLP	L-T PT	SG-KIE	P-LoRA	DocVQA	InfoVQA	VisualMRC	Avg	FUNSD	CORD	SROIE	Avg
×	✓	✓	✓	65.8	25.3	28.7	39.9	49.3	65.8	61.9	59.0
✓	×	✓	✓	**78.2**	**39.7**	28.3	**48.7**	52.1	76.5	86.8	71.8
✓	✓	×	✓	69.1	28.7	29.3	42.3	52.3	82.4	84.0	72.9
✓	✓	✓	×	74.6	36.6	**32.6**	47.9	54.8	86.0	**91.3**	**77.4**
✓	✓	✓	✓	70.4	29.8	31.7	44.0	**54.9**	**86.9**	88.3	76.7

Table 4: Ablations on each component of LayTextLLM (Accuracy).

document-oriented VQA tasks, particularly when addressing questions that heavily depend on visual information. A more detailed analysis of these challenges is provided in Limitations (Section 5).

4.6 Analysis

Ablations To better assess the utility of each component in LayTextLLM, an ablation study was conducted, the results of which are presented in Table 4. Detailed information on the training setup for all variants is provided in Appendix D. The results clearly show that incorporating interleaved spatial layouts and texts significantly enhances the performance, evidenced by a 4.1% improvement in VQA and a 17.7% increase in KIE (the first row vs. the fourth row), indicating that SLP is a critical component. Interestingly, using next-token-prediction as the pre-training task (*i.e.,* the second row) generally outperforms layout-text alignment pre-training across almost all VQA tasks. However, for KIE tasks, layout-text alignment pre-training remains more effective. We hypothesize that layout-text alignment pre-training helps the model learn the relationship between layout and text, which is particularly useful for layout-aware tasks like KIE. In contrast, next-token-prediction focuses on reconstructing the entire document, which is more beneficial for semantic-rich tasks like VQA. Furthermore, including SG-KIE results in a modest performance increase of 1.7% in VQA (the third row vs. the fourth row) but a significant improvement in KIE tasks (*i.e.,* 3.8%), which is as expected. Intriguingly, excluding P-LoRA improves performance on VQA and KIE tasks, suggesting it adds unnecessary complexity or interference, which further highlights the benefits of interleaving texts and layouts.

Sequence Length Table 5 presents statistics on the average input sequence length across different datasets. Intriguingly, despite interleaving bounding box tokens, LayTextLLM consistently exhibits the shortest sequence length in three out of four datasets, even surpassing DocLLM, which is counterintuitive. We attribute this to the tokenizer mechanism. For example, using `tokenizer.encode()`, a single word from the OCR engine, like *"International"* is encoded into a single ID [4623]. Conversely, when the entire OCR output is processed as one sequence, such as *"... CPC,International,Inc..."*, the word *"International"* is split into two IDs [17579, 1288], corresponding to *"Intern"* and *"ational"* respectively. This type of case occurs frequently, we provide further discussion in Appendix E.

Dataset	LayTextLLM	DocLLM	Coor-as-tokens
DocVQA	**664.3**	827.5	4085.7
CORD	**137.9**	153.2	607.3
FUNSD	**701.9**	847.5	4183.4
SROIE	529.2	**505.1**	1357.7

Table 5: Average sequence length.

5 Conclusion

We propose LayTextLLM, interleaving spatial layouts and text to improve predictions through an innovative SLP, the Layout-text Alignment pre-training and the SG-KIE tasks. Extensive experiments show the effectiveness of LayTextLLM.

Limitations

Although LayTextLLM has shown significant capabilities in text-rich VQA and KIE tasks, this alone does not suffice for all real-world applications. There are some instances where reasoning must be based solely on visual cues (*e.g.* size, color, objects)—a challenge that remains unmet. Questions such as *"What is the difference between the highest and the lowest green bar?"* and *"What is written on the card on the palm?"* illustrate this gap. Two bad cases, detailed in Figures 6 and 7, also underscore these limitations. Addressing these challenges underscores the need for future advancements that incorporate visual cues into the capabilities of LayTextLLM. Since the integration with MLLMs is not the primary focus of this work, the preliminary experiments exploring this approach are discussed in Appendix J.

References

OpenAI:Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, FlorenciaLeoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, HyungWon Chung, Dave Cummings, and Jeremiah Currier. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023a. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. 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.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Xiang Fei, Jinghui Lu, Qi Sun, Hao Feng, Yanjie Wang, Wei Shi, An-Lan Wang, Jingqun Tang, and Can Huang. 2025. Advancing sequential numerical prediction in autoregressive models. *arXiv preprint arXiv:2505.13077*.

Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2023a. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*.

Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. 2023b. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*.

Hao Feng, Shu Wei, Xiang Fei, Wei Shi, Yingdong Han, Lei Liao, Jinghui Lu, Binghong Wu, Qi Liu, Chunhui Lin, et al. 2025. Dolphin: Document image parsing via heterogeneous anchor prompting. *arXiv preprint arXiv:2505.14059*.

Chufan Gao, Xuan Wang, and Jimeng Sun. 2024. TTM-RE: Memory-augmented document-level relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 443–458, Bangkok, Thailand. Association for Computational Linguistics.

Liangcai Gao, Yilun Huang, Herve Dejean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. 2019. Icdar 2019 competition on table detection and recognition (ctdar). In *International Conference on Document Analysis and Recognition*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 10767–10775.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. 2024. mPLUG-DocOwl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv:2302.14045*.

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.

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

Wonseok Hwang, Jinyeong Yim, Seung-Hyun Park, Sohee Yang, and Minjoon Seo. 2020. Spatial dependency parsing for semi-structured document information extraction. *Cornell University - arXiv,Cornell University - arXiv*.

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.

Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. 2023. Visual information extraction in the wild: practical dataset and end-to-end solution. In *International Conference on Document Analysis and Recognition*, pages 36–53. Springer.

Qiwei Li, Zuchao Li, Ping Wang, Haojun Ai, and Hai Zhao. 2024a. Hypergraph based understanding for document semantic entity recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2950–2960, Bangkok, Thailand. Association for Computational Linguistics.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024b. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*.

Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. 2024. Doclayllm: An efficient and effective multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024c. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

Jinghui Lu, Yanjie Wang, Ziwei Yang, Xuejing Liu, Brian Mac Namee, and Can Huang. 2024. PadeLLM-NER: Parallel decoding in large language models for named entity recognition. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jinghui Lu, Haiyang Yu, Siliang Xu, Shiwei Ran, Guozhi Tang, Siqi Wang, Bin Shan, Teng Fu, Hao Feng, Jingqun Tang, et al. 2025. Prolonged reasoning is not all you need: Certainty-based adaptive routing for efficient llm/mllm reasoning. *arXiv preprint arXiv:2505.15154*.

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. *CVPR 2024*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

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.

Oshri Naparstek, Ophir Azulai, Inbar Shapira, Elad Amrani, Yevgeny Yaroker, Yevgeny Burshtein, Roi Pony, Nadav Rubinstein, Foad Abo Dahood, Orit Prince, et al. 2024. Kvp10k: A comprehensive dataset for key-value pair extraction in business documents. In *International Conference on Document Analysis and Recognition*, pages 97–116. Springer.

Armineh Nourbakhsh, Sameena Shah, and Carolyn Rose. 2024. Towards a new research agenda for multimodal enterprise document understanding: What are we missing? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14610–14622, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Jiaqi Mu, Hao Zhang, and Nan Hua. 2023. Lmdx: Language model-based document information extraction and localization. *arXiv preprint arXiv:2309.10952*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, et al. 2023. Docile benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition*, pages 147–166. Springer.

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

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

Zineng Tang, Zhenfeng Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Zhu C, Michael Zeng, Zhang Cha, and Mohit Bansal. 2022. Unifying vision, text, and layout for universal document processing. *Cornell University - arXiv,Cornell University - arXiv*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024a. DocLLM: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8529–8548, Bangkok, Thailand. Association for Computational Linguistics.

Han Wang, Yongjie Ye, Bingru Li, Yuxiang Nie, Jinghui Lu, Jingqun Tang, Yanjie Wang, and Can Huang. 2025. Vision as lora. *arXiv preprint arXiv:2503.20680*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. 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)*.

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*.

Zhibo Yang, Rujiao Long, Pengfei Wang, Sibo Song, Humen Zhong, Wenqing Cheng, Xiang Bai, and Cong Yao. 2023. Modeling entities as semantic points for visual information extraction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15358–15367.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023. mPLUG-DocOwl: Modularized multimodal large language model for document understanding. *arXiv:2307.02499*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand. Association for Computational Linguistics.

Hanzhang Zhou, Junlang Qian, Zijian Feng, Lu Hui, Zixiao Zhu, and Kezhi Mao. 2024. LLMs learn task heuristics from demonstrations: A heuristic-driven prompting strategy for document-level event argument extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11972–11990, Bangkok, Thailand. Association for Computational Linguistics.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. FanOutQA: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37, Bangkok, Thailand. Association for Computational Linguistics.

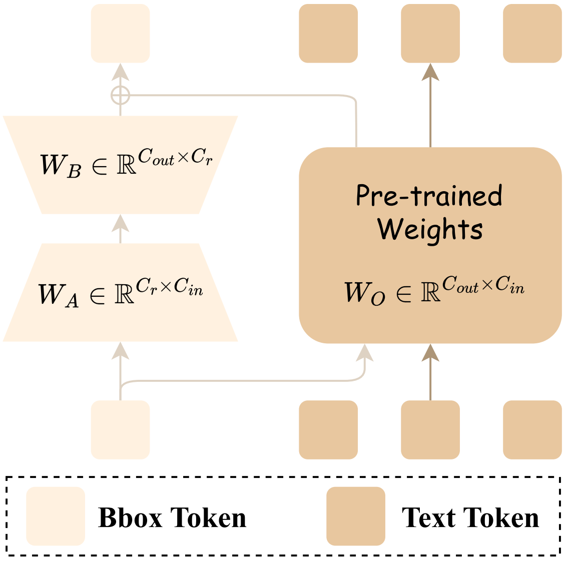

$W_B \in \mathbb{R}^{C_{out}\times C_r}$
$W_A \in \mathbb{R}^{C_r\times C_{in}}$
Pre-trained Weights
$W_O \in \mathbb{R}^{C_{out}\times C_{in}}$
Bbox Token
Text Token

Figure 5: The illustration of P-LoRA, adapted from Dong et al. (2024).

A Layout Partial Low-Rank Adaptation

After using the SLP to generate bounding box tokens and a tokenizer to produce text tokens, these two modalities are then interacted using a Layout Partial Low-Rank Adaptation (P-LoRA) module in LLMs. P-LoRA, introduced in InternLM-XComposer2 (Dong et al., 2024), is originally used to adapt LLMs to the visual modality. It applies plug-in low-rank modules specified to the visual tokens, which adds minimal parameters while preserving the LLMs inherent knowledge.

Formally, for a linear layer in the LLM, the original weights $W_O \in \mathbb{R}^{C_{out}\times C_{in}}$ and bias $B_O \in \mathbb{R}^{C_{out}}$ are specified for input and output dimensions C_{in} and C_{out}. P-LoRA modifies this setup by incorporating two additional matrices, $W_A \in \mathbb{R}^{C_r\times C_{in}}$ and $W_B \in \mathbb{R}^{C_{out}\times C_r}$. These matrices are lower-rank, with C_r being considerably smaller than both C_{in} and C_{out}, and are specifically designed to interact with new modality tokens, which in our case are bounding box tokens. For example, given an input $x = [x_b, x_t]$ comprising of bounding box tokens (x_b) and textual tokens (x_t) is fed into the system, the forward process is as follows, where $\hat{x}_t, \hat{x}_b$ and $\hat{x}$ are outputs:

$$\begin{aligned} \hat{x}_t &= W_0 x_t + B_0 \\ \hat{x}_b &= W_0 x_b + W_B W_A x_b + B_0 \\ \hat{x} &= [\hat{x}_b, \hat{x}_t] \end{aligned} \quad (2)$$

B Qualitative Examples

Qualitative examples of document-oriented VQA (upper row) and KIE (bottom row) are shown in Figure 8. The results indicate that LayTextLLM is highly effective in utilizing spatial layout information to make more accurate predictions for these challenging examples. For example, in the upper right figure, many numeric texts in the receipt act as noise for the baseline method. In contrast, LayTextLLM integrates layout information to accurately predict the total price, as demonstrated by the other examples, underscoring the utility of LayTextLLM.

C Dataset Statistics

Table 6 and 7 show the statistics of datasets used in layout-text alignment pre-training and SFT, respectively. In layout-text alignment pre-training, for training efficiency, we randomly selected around 50,000 documents from each of the DocILE and RVL_CDIP datasets. For every document, we generated two tasks: line-level layout decoding and either a text-to-layout or layout-to-text prediction task, which yields a total of around 200,000 pre-training examples. We also tested the model on a KIE dataset **POIE** (Kuang et al., 2023).

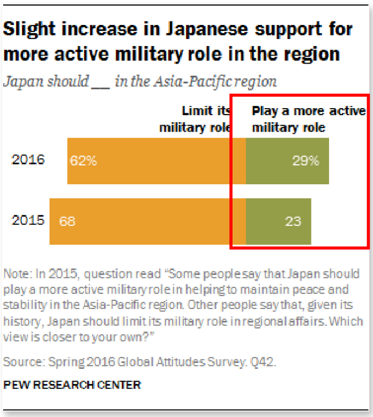
Dataset	DocILE	RVL_CDIP
Num Documents	55,719	59444
Num Examples	111,438	118,888
Num Tokens	75,952,078	67,340,246

Table 6: Dataset statistics for layout-text alignment pre-training (using Llama-2 Tokenizer).

D Implementation Detail

All training and inference procedures are conducted on eight NVIDIA A100 GPUs.

Training LayTextLLM is initialized with Llama2-7b-chat model, the pre-training, SFT, and other model hyper-parameters can be seen in Table 8. Additional parameters including SLP and P-LoRA are randomly initialized. During pre-training and SFT, all parameters are trainable. Please note that all variants of LayTextLLM, including those utilized in ablation studies, are trained in accordance with the same settings. Specifically, for all variants in ablation study, we train with the same setting and dataset in accordance with LayTextLLM$_{zero}$. For the variant

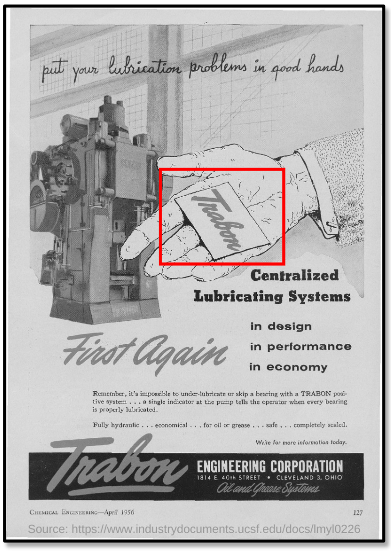


Question: What is the **difference** between the highest and the lowest **green bar**?

GroundTruth: **6**

Our Prediction: **40**

Figure 6: A failure case of LayTextLLM on ChartQA.


put your lubrication problems in good hands
Trabon
Centralized Lubricating Systems
First Again
in design
in performance
in economy
Trabon
ENGINEERING CORPORATION
1814 E. 40th STREET • CLEVELAND 3, OHIO
Oil and Grease Systems
Source: https://www.industrydocuments.ucsf.edu/docs/lmyl0226

What is written on the card on the palm?

GroundTruth: Trabon

Our Prediction: put your lubrication problems in good hands

Figure 7: A failure case of LayTextLLM on DocVQA.

without SLP, we replace the bounding box token placeholder "*<box>*" with "*\n*". For the variant without layout-text alignment pre-training, we pre-train the model on the same dataset using a conventional next-token prediction task, excluding the loss computation for the bounding box token. After pre-training, we fine-tune the model on the SFT datasets. For the variant without SG-KIE tasks, we remove the SG-KIE data from the SFT datasets while retaining the original SER and EL tasks in KVP10k and SIBR to ensure the total number of training examples remains unchanged. For the variant without P-LoRA, we replace all P-LoRA modules with linear layers, as was previously done.

All baseline results are sourced from Liu et al. (2024c) or respective original papers, with the exception of the Llama2-7B series, the Llama2-7B$_{\text{coor}}$ series, and Qwen2-VL, these results were re-implemented by authors.

Inference For the document-oriented VQA test set, we use the original question-answer pairs as the prompt and ground truth, respectively. For KIE tasks, we reformat the key-value pairs into a question-answer format, as described in Wang et al. (2024a); Luo et al. (2024); Liu et al. (2024c). Additionally, for the FUNSD dataset, we focus our testing on the entity linking annotations as described in Luo et al. (2024). Note that for KIE tasks, we report the result of directly generating the answer texts, instead of generating the answer with the coordinates (SG-KIE). The discussion regarding inference with SG-KIE can be found in

Dataset	DDD	Layout-aware SFT	KVP10k	SIBR	DocVQA	InfoVQA	ChartQA	VisualMRC	FUNSD	CORD	SROIE
Num Documents	115,955	50,409	4,249	600	10,192	4,405	3,699	7,012	147	794	626
Num Examples	115,955	280,073	50,661	12,978	39,459	23,945	7,398	7,013	2,375	8,932	2,503
Num Tokens	71,067,212	101,209,393	27,018,563	8,045,694	17,621,621	1,024,236	1,052,752	1,622,387	11,543,711	1,140,437	1,066,930

Table 7: Dataset statistics for SFT (using Llama-2 Tokenizer).

	Backbone	**Plora rank**	**Batch size**	**Max length**	**Precision**	**Train params**	**Fix params**
Lay-Text Pretrain	Llama2-7B-base	256	128	4096	bf16	7.4 B	0B
SFT	Llama2-7B-base	256	128	4096	bf16	7.4 B	0B
	Learning rate	**Weight decay**	**Scheduler**	**Adam betas**	**Adam epsilon**	**Warm up**	**Epoch**
Lay-Text Pretrain	5.0e-05	0.01	cosine	[0.9, 0.999]	1.0e-08	0.005	4
SFT	1.0e-05	0.01	cosine	[0.9, 0.999]	1.0e-08	0.005	4

Table 8: LayTextLLM trainng Hyper-parameters.

Appendix H.

To eliminate the impact of randomness on evaluation, no sampling methods are employed during testing for any of the models. Instead, beam search with a beam size of 1 is used for generation across all models. Additionally, the maximum number of new tokens is set to 512, while the maximum number of input tokens is set to 4096.

E Discussion of Input Sequence Length

As mentioned in Section 4.6, it is intriguing that LayTextLLM has fewer input sequences than DocLLM, which is counterintuitive given that LayTextLLM interleaves bounding box tokens, typically resulting in longer sequence lengths. We attribute this to the Byte Pair Encoding (BPE) tokenizers (Sennrich et al., 2016) prevalently used in modern LLMs such as Llama2.

BPE operates by building a vocabulary of commonly occurring subwords (or token pieces) derived from the training data. Initially, it tokenizes the text at the character level and then progressively merges the most frequent adjacent pairs of characters or sequences. The objective is to strike a balance between minimizing vocabulary size and maximizing encoding efficiency.

Thus, when tokenizing a single word like *"International"* on its own, the tokenizer might identify it as a common sequence in the training data and encode it as a single token. This is especially likely if *"International"* frequently appears as a standalone word in the training contexts. However, when the word *"International"* is part of a larger sequence of words such as including in a long sequence of OCR-derived texts like *"...335 CPC,International,Inc..."*, the context changes. The tokenizer might split *"International"* into sub-tokens like *"Intern"* and *"ational"* because, in various contexts within the training data, these subwords might appear more frequently in different combinations or are more useful for the model to understand variations in meaning or syntax.

When using LayTextLLM, we input word-level OCR results into the tokenizer, typically resulting in the former situation, where words are encoded as single tokens. Conversely, with DocLLM, the entire OCR output is processed as one large sequence, leading to the latter situation and a longer sequence length than in LayTextLLM. This difference underscores the utility of LayTextLLM in achieving both accuracy and inference efficiency due to its shorter sequence length.

F Discussion on Advantage of Interleaving Layout and Text

Discussion on DocLLM We visualize the attention patterns between input and output tokens in Figure 9. The attention pattern is insightful with the specific question, *"What is the quantity of -TICKET CP?<0x0A>"*

As shown in Figure 9(a), when the model begins predicting the answer *"Final"*, *"<0x0A>"(newline symbol)* is heavily focusing on layout information, as seen by the significant attention on the bounding box embedding *"<unk>"* token before *"(Qty"*. This highlights the model's effort to orient itself spatially and understand the structural context of the tokens. At this stage, the model is developing a cognitive understanding of how the elements are laid out on the page. We extract and visualize the attention scores that *"<0x0A>"* assigns to each bounding box in Figure 9(c). The visualization shows that the model focuses most on *"Qty"*, followed by *"-TICKET"* and *"2.00"*, which reflects the layout information essential for making the prediction. In the final layer (Figure 9(b)), the model's attention shifts dramatically towards the *"Qty"* token, which holds the semantic meaning necessary to answer the question. This shift from layout-based cognition to content-based reasoning illustrates how the bounding box tokens act as spatial anchors that help the model pinpoint and organize the relevant information (such as *"Qty"*) to make the correct prediction.

The attention of LayTextLLM exhibits a distinct pattern compared to models like DocLLM, which uses block infilling to predict missing blocks from both preceding and succeeding context. In contrast, LayTextLLM adheres to an auto-regressive approach, focusing its attention solely on preceding information. Furthermore, interleaving bounding box and text embeddings creates strong attention connections between textual and spatial representations, as shown in Figure 9. In contrast, DocLLM integrates spatial information into the calculation of attention score which is implicitly. As shown in Table 1, LayTextLLM significantly outperforms DocLLM, again underscoring the advantage of interleaving bounding box and text embeddings. Also, we found that the spatial information can be decoded back into coordinates even without inputing visual information, as discussed in Appendix I, which is not exhibited in DocLLM.

We also conduct a fairer experiment by re-implementing DocLLM using the identical training settings as LayTextLLM$_{zero}$. In order to ensure a more intuitive and fair comparison between the two layout adaptation methods (*i.e.,* SLP versus disentangled spatial attention), we exclude the use of P-LoRA in LayTextLLM$_{zero}$. Table 9 demonstrates that SLP is a more effective method for incorporating layout information, as evidenced by a 6.7% improvement in VQA and an 8.4% improvement in KIE. Additionally, while DocLLM introduces a suite of attention weights for layout information, it significantly increases the number of parameters in LLaMA-2 from 6.73B to 8.37B. In contrast, LayTextLLM introduces a much smaller increase in parameters.

Discussion on coordinate-as-tokens The sub-performance of coordinate-as-tokens methods can be attributed to the following three reasons: (1) The coordinate-as-tokens approach tends to introduce an excessive number of tokens, often exceeding the pre-defined maximum length of Llama2-7B (*i.e.*, 4096). Consequently, this leads to a lack of crucial OCR information, resulting in hallucination and subpar performance. (2) When re-implementing the coordinate-as-tokens method with Llama2-7B, we did not introduce the ICL strategy, as it would contribute additional length to the input sequence. (3) The coordinate-as-tokens approach necessitates a considerably larger-sized LLM to comprehend the numerical tokens effectively.

G Results of ChartQA

As shown in Figure 6, the question-answer pairs in ChartQA (Masry et al., 2022) tend to involve the visual cues for reasoning. However, with only text and layout information as input, the proposed LayTextLLM inevitably have difficulties in reasoning visual-related information. Thus, on the ChartQA dataset, LayTextLLM can hardly achieve better performance than previous methods that include visual inputs. Although the visual information is not used in LayTextLLM, it can still exhibit better zero-shot ability than UniDoc (Feng et al., 2023b). After incorporating the training set of ChartQA, the performance of LayTextLLM can be boosted to 42.2%. Considering the importance of visual cues in ChartQA-like tasks, we will try to involve the visual information into LayTextLLM in future work. A preliminary discussion can be seen in Appendix J.

H Inference with SG-KIE

As discussed in Section 4.6, incorporating SG-KIE as an auxiliary task in SFT has been shown to enhance the performance of KIE tasks. In this section, we investigate the effectiveness of using SG-KIE as a direct inference task for KIE. The results are shown in Table 11. We can observe that, for the FUNSD$^-$ and CORD$^-$ datasets, SG-KIE inference demonstrates improved performance. However, for the SROIE$^-$ dataset, there is a slight decrease in performance. We manually reviewed the problematic cases of SG-KIE and identified two main reasons for the performance drop: (1) incorrect format, which leads to parsing errors such as *"432.60[SR @ 6%[<B-1013><B453> <B><B478>]"*, and (2) ambiguous key types in the SROIE$^-$ dataset. For instance, the key "total" can refer to "grand total" and if the model has not been trained with the dataset, SG-KIE may mistakenly localize it to the wrong value. A notable instance of this issue is shown in Figure 10. These types of errors occur

	Document-Oriented VQA				KIE				Num Params
Methods	DocVQA	InfoVQA	VisualMRC	Avg	FUNSD	CORD	SROIE	Avg	
DocLLM	66.6	28.3	28.6	41.2	51.3	71.8	83.9	69.0	8.37B
LayTextLLM	**74.6**	**36.6**	**32.6**	**47.9**	**54.8**	**86.0**	**91.3**	**77.4**	**6.76B**

Table 9: Comparison of two layout adaptation methods, *i.e.,* SLP in LayTextLLM and Disentangled Spatial Attention in DocLLM.

	ChartQA
OCR-free	
UniDoc (Feng et al., 2023b)	10.9
DocPedia (Feng et al., 2023a)	46.9*
Monkey (Li et al., 2023)	54.0*
InternVL (Chen et al., 2023b)	45.6*
InternLM-XComposer2 (Dong et al., 2024)	51.6*
TextMonkey (Liu et al., 2024c)	58.2*
TextMonkey$_+$ (Liu et al., 2024c)	**59.9***
Qwen2-VL (Wang et al., 2024b)	**61.9***
Text + Coordinates	
LayTextLLM$_{zero}$ (Ours)	30.2
LayTextLLM$_{all}$ (Ours)	42.6*

Table 10: Comparison with SOTA OCR-free MLLMs on ChartQA (accuracy). * denotes the use of the dataset's training set.

frequently in the dataset.

For improvement, we observed that SG-KIE performs better when processing complex answers that require the aggregation of multiple consecutive word-level OCR results, leading to more accurate and complete outputs, as illustrated in Figure 11.

Dataset	FUNSD$^-$	CORD$^-$	SROIE$^-$
LayTextLLM$_{zero}$	79.6	81.3	**87.0**
LayTextLLM$_{zero-sg}$	**80.0**	**81.9**	86.0

Table 11: Inference with SG-KIE vs. without SG-KIE (accuracy).

I Decoding Bounding Box Coordinates

We also evaluate the model's ability to decode bounding box embeddings into coordinates. Since the SG-KIE task requires the model to generate precise coordinates for answers, this task can be used to assess the performance in accurately predicting bounding boxes. Specifically, we select the examples with correct predictions for textual answer and compute the Intersection over Union (IoU) score (Rezatofighi et al., 2019) between the predicted and ground truth coordinates. We tested the on three datasets: FUNSD, which is not used to train LayTextLLM$_{zero}$. If the IoU exceeds 0.5, we consider the bounding box prediction to be correct. Accuracy is used as the metric to evaluate this capability, we compute accuracy for the coordinates for both key and value. Results show that about 77.5% bounding box is correctly predicted, cases are visualized in Figure 12. Also, we visualize the coordinates prediction for the pre-training task—line-level layout decoding—in Figure 13. Moreover, SG-KIE produces coordinates, which is obviously interpretable, and providing coordinates seems to be more valuable for certain downstream tasks.

FUNSD	LayTextLLM$_{zero}$
Accuracy	77.5

Table 12: Coordinate prediction accuracy.

J Combination with MLLMs

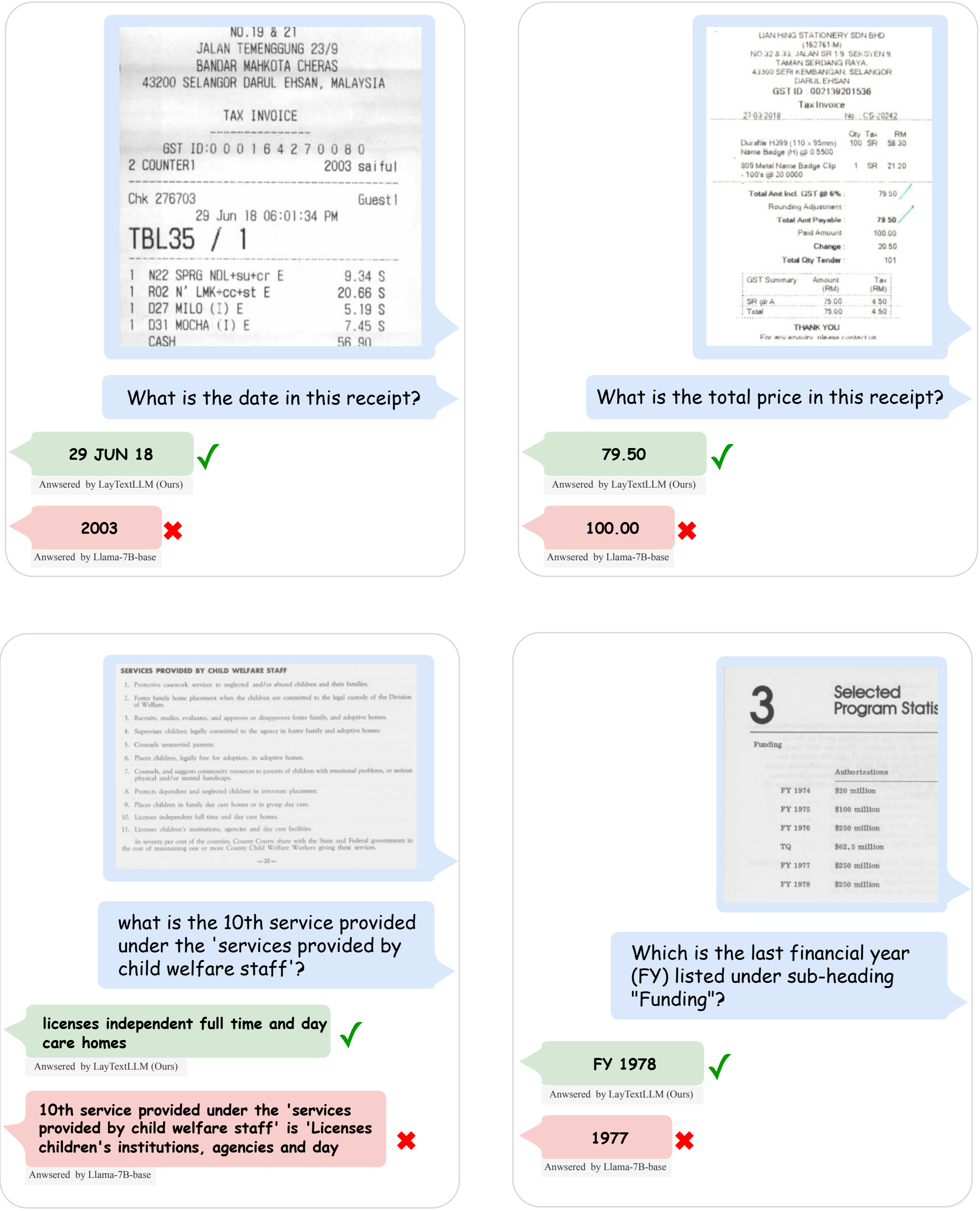
As discussed in Limitation (Section 5), LayTextLLM faces challenges with VQA tasks that require the comprehension of visual elements such as font, size, shape, objects, color, and other visual attributes. To address this limitation, we conducted a preliminary experiment combining LayTextLLM with a MLLM to explore the potential of leveraging visual information while preserving the strengths of LayTextLLM.

Specifically, we upgrade the multimodal version of LayTextLLM by building upon Qwen2-VL and incorporating a SLP. For simplicity, neither P-LoRA nor special tokens are introduced. we layout-text alignment pre-trained and SFT the modified Qwen2-VL on the same datasets used for LayTextLLM$_{zero}$, resulting a *Qwen2-VL-LayText* model. We also trained a counterpart of Qwen2-VL-LayText by incorporating only OCR text, excluding layout information. This model, which is identical in training settings to Qwen2-VL-LayText, was named *Qwen2-VL-Text* and serves as a baseline. The model performance can be seen in Table 13. Although it shows a slight drop in performance on VQA tasks, Qwen2-VL-LayText achieves significant improvements in KIE tasks, with an overall accuracy of 76.4% compared to

67.7%. This further demonstrates the effectiveness of interleaving layouts and text. Interestingly, simply adding OCR text (*i.e.,* Qwen2-VL-Text) also results in a notable improvement in KIE tasks when paired with Qwen2-VL. We believe this is because datasets with poor performance, such as CORD and SROIE, primarily consist of text with small or blurred fonts. In these cases, off-the-shelf OCR engines still outperform MLLMs in text recognition.

	Document-Oriented VQA			KIE			
	DocVQA	InfoVQA	Avg	FUNSD	CORD	SROIE	Avg
Metric	*ANLS %*						
Visual + Text + Coordinates							
Qwen2-VL (Wang et al., 2024b)	**81.4**	**45.2**	**63.3**	53.2	71.3	78.8	67.7
Qwen2-VL$_{text}$	77.0	43.5	60.2	46.0	90.2	83.5	73.2
Qwen2-VL$_{LayText}$	**81.4**	42.7	62.1	**54.2**	**91.2**	**83.7**	**76.4**

Table 13: Comparison with Qwen2-VL-LayText with other baselines (accuracy).



NO.19 & 21
JALAN TEMENGGUNG 23/9
BANDAR MAHKOTA CHERAS
43200 SELANGOR DARUL EHSAN, MALAYSIA

TAX INVOICE

GST ID:0 0 0 1 6 4 2 7 0 0 8 0
2 COUNTER1 2003 saiful

Chk 276703 Guest1
29 Jun 18 06:01:34 PM
TBL35 / 1

1	N22 SPRG NDL+su+cr E	9.34 S
1	R02 N' LMK+cc+st E	20.66 S
1	D27 MILO (I) E	5.19 S
1	D31 MOCHA (I) E	7.45 S
	CASH	56.90

What is the date in this receipt?

29 JUN 18 ✓

Anwsered by LayTextLLM (Ours)

2003 ✗

Anwsered by Llama-7B-base

LIAN HING STATIONERY SDN BHD
(162761-M)
NO.32 & 33, JALAN SR 1/9, SEKSYEN 9,
TAMAN SERDANG RAYA,
43300 SERI KEMBANGAN, SELANGOR
DARUL EHSAN
GST ID : 002139201536
Tax Invoice
27/03/2018 No. : CS-20242

	Qty	Tax	RM
Durafile H399 (110 x 95mm) Name Badge (H) @ 0.5500	100	SR	58.30
809 Metal Name Badge Clip - 100's @ 20.0000	1	SR	21.20

Total Amt Incl. GST @ 6% :	79.50
Rounding Adjustment :	
Total Amt Payable :	79.50
Paid Amount :	100.00
Change :	20.50
Total Qty Tender :	101

GST Summary	Amount (RM)	Tax (RM)
SR @ A	75.00	4.50
Total	75.00	4.50

THANK YOU
For any enquiry, please contact us

What is the total price in this receipt?

79.50 ✓

Anwsered by LayTextLLM (Ours)

100.00 ✗

Anwsered by Llama-7B-base

SERVICES PROVIDED BY CHILD WELFARE STAFF

1. Protective casework services to neglected and/or abused children and their families.
2. Foster family home placement when the children are committed to the legal custody of the Division of Welfare.
3. Recruits, studies, evaluates, and approves or disapproves foster family, and adoptive homes.
4. Supervises children legally committed to the agency in foster family and adoptive homes.
5. Counsels unmarried parents.
6. Places children, legally free for adoption, in adoptive homes.
7. Counsels, and suggests community resources to parents of children with emotional problems, or serious physical and/or mental handicaps.
8. Protects dependent and neglected children in interstate placement.
9. Places children in family day care homes or in group day care.
10. Licenses independent full time and day care homes.
11. Licenses children's institutions, agencies and day care facilities.

In seventy per cent of the counties, County Courts share with the State and Federal governments in the cost of maintaining one or more County Child Welfare Workers giving these services.

—20—

what is the 10th service provided under the 'services provided by child welfare staff'?

licenses independent full time and day care homes ✓

Anwsered by LayTextLLM (Ours)

10th service provided under the 'services provided by child welfare staff' is 'Licenses children's institutions, agencies and day ✗

Anwsered by Llama-7B-base

3 Selected Program Statis

Funding

	Authorizations
FY 1974	$20 million
FY 1975	$100 million
FY 1976	$250 million
TQ	$62.5 million
FY 1977	$250 million
FY 1978	$250 million

Which is the last financial year (FY) listed under sub-heading "Funding"?

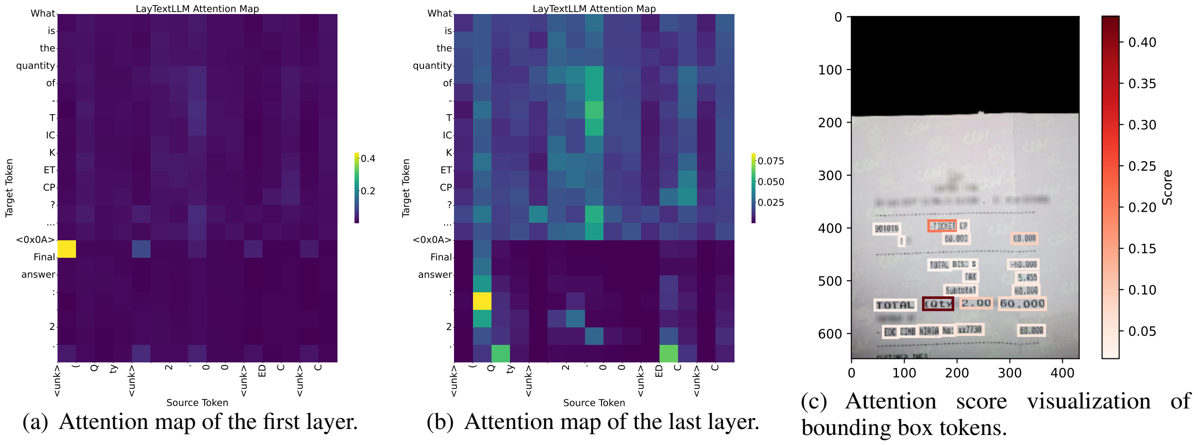
FY 1978 ✓

Anwsered by LayTextLLM (Ours)

1977 ✗

Anwsered by Llama-7B-base

Figure 8: Qualitative comparison with the baseline method.

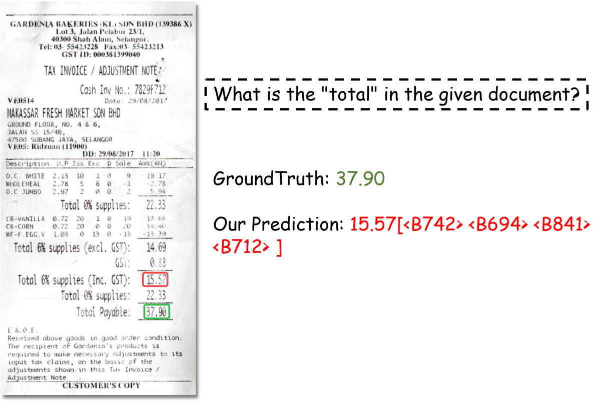


(a) Attention map of the first layer.

(b) Attention map of the last layer.

(c) Attention score visualization of bounding box tokens.

Figure 9: Visualization of attention maps of LayTextLLM. Best viewed in color and with zoom. *"<unk>"* is the placeholder for the bounding box token.



GARDENIA BAKERIES (KL) SDN BHD (139386 X)
Lot 3, Jalan Pelabur 23/1,
40300 Shah Alam, Selangor.
Tel: 03- 55423228 Fax:03- 55423213
GST ID: 000381399040

TAX INVOICE / ADJUSTMENT NOTE

Cash Inv No.: 7829F712
VE0514 Date: 29/08/2017
MAKASSAR FRESH MARKET SDN BHD
GROUND FLOOR, NO. 4 & 6,
JALAN SS 15/4B,
47500 SUBANG JAYA, SELANGOR
VE05: Ridzuan (11900)
DD: 29/08/2017 11:20

Description	U.P	Iss	Exc	D	Sale	Amt(RM)
O.C. WHITE	2.13	10	1	0	9	19.17
WHOLEMEAL	2.78	5	6	0	-1	-2.78
O.C JUMBO	2.97	2	0	0	2	5.94
Total 0% supplies:						22.33
CR-VANILLA	0.72	20	1	0	19	13.68
CR-CORN	0.72	20	0	0	20	14.40
WF-F.EGG.V	1.03	0	13	0	-13	-13.39
Total 6% supplies (excl. GST):						14.69
GST:						0.88
Total 6% supplies (Inc. GST):						15.57
Total 0% supplies:						22.33
Total Payable:						37.90

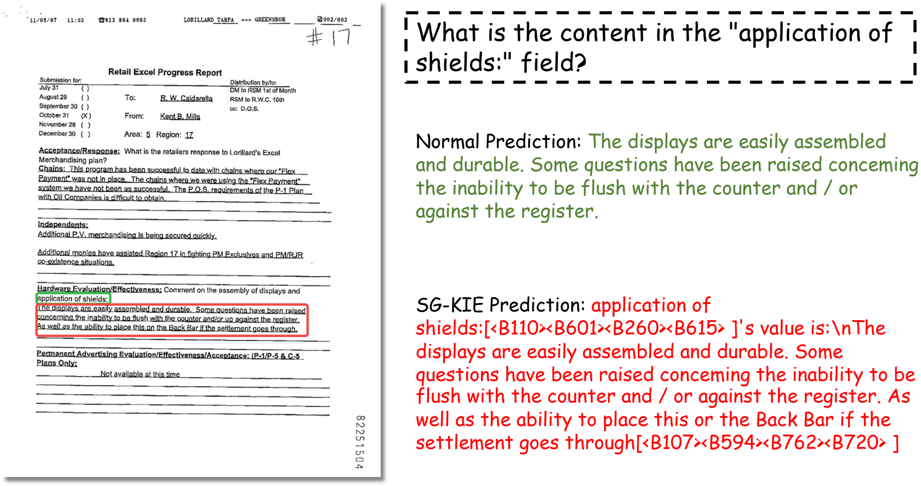
E.&.O.E.
Received above goods in good order condition. The recipient of Gardenia's products is required to make necessary adjustments to its input tax claims, on the basic of the adjustments shown in this Tax Invoice / Adjustment Note
CUSTOMER'S COPY

What is the "total" in the given document?

GroundTruth: 37.90

Our Prediction: 15.57[<B742> <B694> <B841> <B712>]

Figure 10: A failure case of SG-KIE in SROIE⁻. The red box indicates the ground truth and the green box is the prediction.



11/05/97 11:03 ☎813 884 0663 LORILLARD TAMPA →→→ GREENSBOR 002/003
#17

Retail Excel Progress Report

Submission for:
July 31 ( )
August 29 ( )
September 30 ( )
October 31 (X)
November 28 ( )
December 30 ( )

To: R. W. Caldarella
From: Kent B. Mills
Area: 5 Region: 17

Distribution by/to:
DM to RSM 1st of Month
RSM to R.W.C. 10th
cc: D.O.S.

Acceptance/Response: What is the retailers response to Lorillard's Excel Merchandising plan?
Chains: This program has been successful to date with chains where our "Flex Payment" was not in place. The chains where we were using the "Flex Payment" system we have not been as successful. The P.O.S. requirements of the P-1 Plan with Oil Companies is difficult to obtain.

Independents:
Additional P.V. merchandising is being secured quickly.

Additional monies have assisted Region 17 in fighting PM Exclusives and PM/RJR co-existence situations.

Hardware Evaluation/Effectiveness: Comment on the assembly of displays and application of shields:
The displays are easily assembled and durable. Some questions have been raised concerning the inability to be flush with the counter and/or up against the register. As well as the ability to place this on the Back Bar if the settlement goes through.

Permanent Advertising Evaluation/Effectiveness/Acceptance: (P-1/P-5 & C-5 Plans Only:
Not available at this time

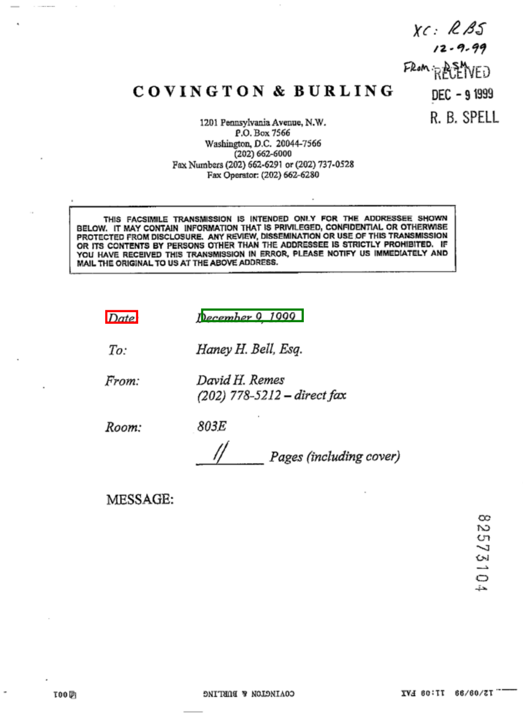
82251504

What is the content in the "application of shields:" field?

Normal Prediction: The displays are easily assembled and durable. Some questions have been raised concerning the inability to be flush with the counter and / or against the register.

SG-KIE Prediction: application of shields:[<B110><B601><B260><B615>]'s value is:\nThe displays are easily assembled and durable. Some questions have been raised concerning the inability to be flush with the counter and / or against the register. As well as the ability to place this or the Back Bar if the settlement goes through[<B107><B594><B762><B720>]

Figure 11: A good case of SG-KIE in FUNSD⁻. The red box indicates the ground truth value and the green box is the key.



XC: RBS
12-9-99
FROM: DSM

RECEIVED
DEC - 9 1999
R. B. SPELL

COVINGTON & BURLING

1201 Pennsylvania Avenue, N.W.
P.O. Box 7566
Washington, D.C. 20044-7566
(202) 662-6000
Fax Numbers (202) 662-6291 or (202) 737-0528
Fax Operator: (202) 662-6280

THIS FACSIMILE TRANSMISSION IS INTENDED ONLY FOR THE ADDRESSEE SHOWN BELOW. IT MAY CONTAIN INFORMATION THAT IS PRIVILEGED, CONFIDENTIAL OR OTHERWISE PROTECTED FROM DISCLOSURE. ANY REVIEW, DISSEMINATION OR USE OF THIS TRANSMISSION OR ITS CONTENTS BY PERSONS OTHER THAN THE ADDRESSEE IS STRICTLY PROHIBITED. IF YOU HAVE RECEIVED THIS TRANSMISSION IN ERROR, PLEASE NOTIFY US IMMEDIATELY AND MAIL THE ORIGINAL TO US AT THE ABOVE ADDRESS.

Date	*December 9, 1999*
To:	*Haney H. Bell, Esq.*
From:	*David H. Remes* *(202) 778-5212 – direct fax*
Room:	*803E*

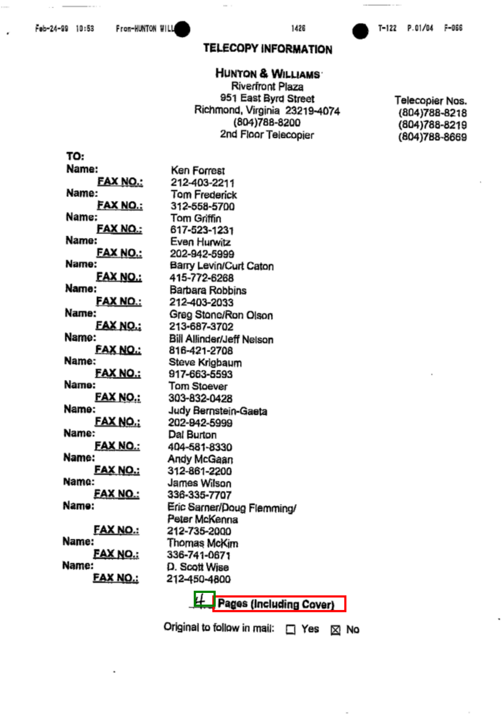
11 *Pages (including cover)*

MESSAGE:

82573104

12/09/99 11:08 FAX COVINGTON & BURLING 001

(a) *Question*: what is the content in the "Date:" field?
Answer: December 9, 1999



Feb-24-99 10:53 From-HUNTON WILL 1426 T-122 P.01/04 F-066

TELECOPY INFORMATION

HUNTON & WILLIAMS
Riverfront Plaza
951 East Byrd Street
Richmond, Virginia 23219-4074
(804)788-8200
2nd Floor Telecopier

Telecopier Nos.
(804)788-8218
(804)788-8219
(804)788-8669

TO:

Name:	Ken Forrest
FAX NO.:	212-403-2211
Name:	Tom Frederick
FAX NO.:	312-558-5700
Name:	Tom Griffin
FAX NO.:	617-523-1231
Name:	Even Hurwitz
FAX NO.:	202-942-5999
Name:	Barry Levin/Curt Caton
FAX NO.:	415-772-6268
Name:	Barbara Robbins
FAX NO.:	212-403-2033
Name:	Greg Stone/Ron Olson
FAX NO.:	213-687-3702
Name:	Bill Allinder/Jeff Nelson
FAX NO.:	816-421-2708
Name:	Steve Krigbaum
FAX NO.:	917-663-5593
Name:	Tom Stoever
FAX NO.:	303-832-0428
Name:	Judy Bernstein-Gaeta
FAX NO.:	202-942-5999
Name:	Dal Burton
FAX NO.:	404-581-8330
Name:	Andy McGaan
FAX NO.:	312-861-2200
Name:	James Wilson
FAX NO.:	336-335-7707
Name:	Eric Sarner/Doug Flemming/ Peter McKenna
FAX NO.:	212-735-2000
Name:	Thomas McKim
FAX NO.:	336-741-0671
Name:	D. Scott Wise
FAX NO.:	212-450-4800

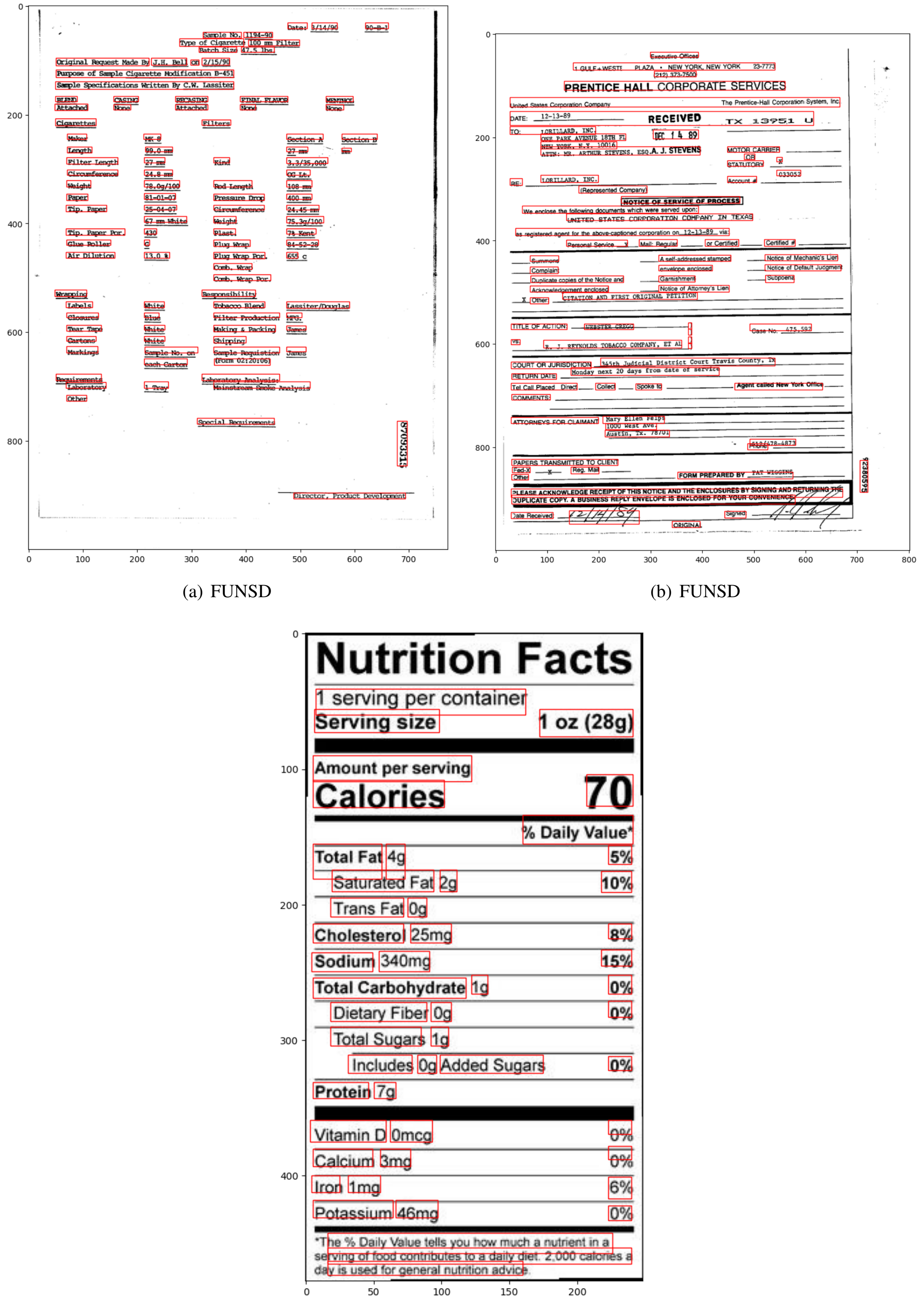
4 **Pages (Including Cover)**

Original to follow in mail: ☐ Yes ☒ No

86328049

(b) *Question*: what is the content in the "Pages (Including Cover)" field?
Answer: 4

Figure 12: Illustration of coordinates prediction for entity linking task. The red box indicates the key text region and the green box indicates the value text region.



(a) FUNSD

(b) FUNSD

(c) POIE

Figure 13: Illustration of coordinates prediction line-level layout decoding. Documents are subsampled from OOD dataset. Red boxes are coordinates for line-level text regions.