

# DLaVA: Document Language and Vision Assistant for Answer Localization with Enhanced Interpretability and Trustworthiness

Ahmad Mohammadshirazi  
The Ohio State University  
Columbus, OH  
mohammadshirazi.2@osu.edu

Pinaki Prasad Guha Neogi  
The Ohio State University  
Columbus, OH  
guhaneogi.2@osu.edu

Ser-Nam Lim  
University of Central Florida  
Orlando, FL  
sernam@ucf.edu

Rajiv Ramnath  
The Ohio State University  
Columbus, OH  
ramnath.6@osu.edu

## Abstract

Document Visual Question Answering (VQA) requires models to interpret textual information within complex visual layouts and comprehend spatial relationships to answer questions based on document images. Existing approaches often lack interpretability and fail to precisely localize answers within the document, hindering users' ability to verify responses and understand the reasoning process. Moreover, standard metrics like Average Normalized Levenshtein Similarity (ANLS) focus on text accuracy but overlook spatial correctness. We introduce **DLaVA**, a novel method that enhances Multimodal Large Language Models (MLLMs) with answer localization capabilities for Document VQA. Our approach integrates image annotation directly into the MLLM pipeline, improving interpretability by enabling users to trace the model's reasoning. We present both OCR-dependent and OCR-free architectures, with the OCR-free approach eliminating the need for separate text recognition components, thus reducing complexity. To the best of our knowledge, DLaVA is the first approach to introduce answer localization within multimodal QA, marking a significant step forward in enhancing user trust and reducing the risk of AI hallucinations. Our contributions include enhancing interpretability and reliability by grounding responses in spatially annotated visual content, introducing answer localization in MLLMs, proposing a streamlined pipeline that combines an MLLM with a text detection module, and conducting comprehensive evaluations using both textual and spatial accuracy metrics, including Intersection over Union (IoU). Experimental results on standard datasets demonstrate that DLaVA achieves SOTA performance, significantly enhancing model transparency and re-

liability. Our approach sets a new benchmark for Document VQA, highlighting the critical importance of precise answer localization and model interpretability. The code and datasets utilized in this study for DLaVA are accessible at: <https://github.com/ahmad-shirazi/AnnotMLLM>

## 1. Introduction

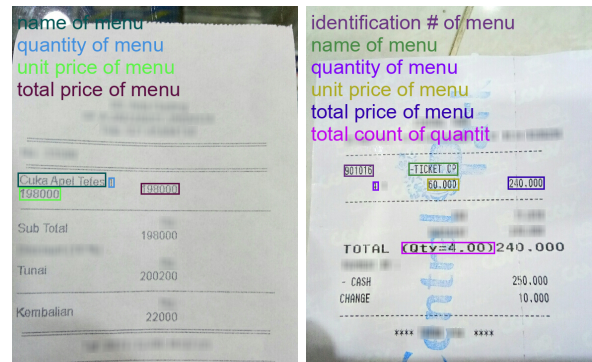


Figure 1. Examples of Answer Annotation in Documents from the CORD Dataset [33]

Document Visual Question Answering (VQA) stands at the intersection of computer vision and natural language processing, aiming to answer questions based on the content of a document image. This task is inherently challenging due to the need for a model to not only accurately recognize and interpret textual information within complex visual layouts but also to reason about the spatial relationships and semantics of the content. Effective solutions require a harmonious integration of text detection, recognition, and

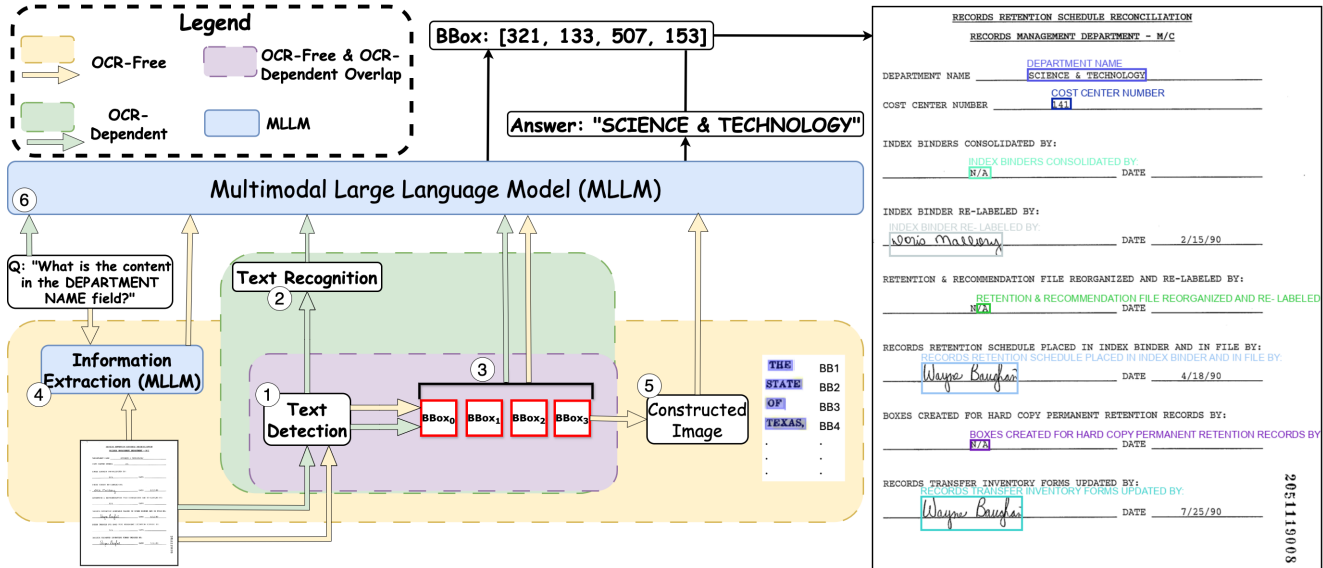


Figure 2. Figure 2. DLaVA Model Architecture for OCR-Dependent and OCR-Free Approaches. The OCR-Dependent approach integrates OCR outputs with positional data to enhance layout comprehension, while the OCR-Free approach utilizes end-to-end MLLM processing to interpret and extract data directly from the image. Both converge in the final annotation module for precise answer localization. The numbers within the circles represent the steps (Refer Section 3 for more details)

contextual understanding to bridge the gap between visual data and linguistic queries [18], as illustrated in Figure 1, which presents document annotations on the CORD dataset (see Appendix A for additional details).

Existing approaches, such as LayoutLMv3 [15], LayoutLLM [30], LayTextLLM [28], and DocLayLLM [26], have made significant strides in addressing visual question answering and layout analysis. While these models demonstrate proficiency in extracting textual information and providing coordinate predictions, they often fall short in terms of interpretability and explainability. Specifically, they lack the ability to precisely localize answers within the document image, making it difficult for users to verify responses and understand the reasoning behind them. Besides, metrics like Average Normalized Levenshtein Similarity (ANLS) [43] focus on text accuracy but overlook spatial correctness, and Intersection over Union (IoU) [34] evaluations are typically limited to layout datasets without assessing answer localization accuracy. In our work, we demonstrate why ANLS alone is insufficient for evaluating answer correctness and emphasize the importance of incorporating metrics like IoU. By using both metrics together, we address limitations found in previous approaches [30], such as not being able to handle false-positive cases (e.g., when an answer does not exist, and the model hallucinates and gives wrong answers). Besides, the lack of answer localization limits model transparency and reliability, as accurate spatial grounding is essential in applications like legal, medical, and financial document analysis to ensure that answers are

derived from the correct visual context [16].

In this paper, we introduce a novel approach that enhances Multimodal Large Language Models (MLLMs) with answer localization capabilities for Document VQA. Our method integrates answer annotation within images directly into the MLLM pipeline, and this addition not only improves interpretability by allowing users to trace the model’s reasoning but also facilitates the identification and analysis of errors, thereby contributing to a deeper understanding of the model’s decision-making process. Our contributions can be summarized as follows:

- 1. Advancing Interpretability and Reliability:** By grounding responses in spatially annotated visual content, our approach enhances user trust and reduces the risk of AI hallucinations. This advancement sets a new standard for reliability in Document VQA and demonstrates competitive results while improving model interpretability.
- 2. Introducing Answer Localization in MLLMs:** We present a novel pipeline that augments MLLMs with the ability to localize answers within document images, which addresses a significant gap in current Document VQA methodologies and enhances model interpretability.
- 3. Innovative Pipeline Design and Model Analysis:** We propose a streamlined pipeline that combines an MLLM with a text detection module, eliminating the need for a separate text recognition component. This integration reduces complexity and improves cohesiveness.

4. **Comprehensive Evaluation with IOU Metrics:** We conduct an in-depth evaluation of our model’s performance using both spatial and textual accuracy metrics, including IoU for the first time in the context of VQA with MLLMs. This dual assessment provides a more complete picture of the model’s capabilities compared to prior works.

The remainder of this paper is organized as follows: In Section 2, we review related work and present a literature survey. Section 3 details our proposed approach, DLVA, describing both the OCR-Dependent and OCR-Free architectures. Section 4 presents the experimental setup, including dataset descriptions, baseline models, and ablation study architectures. In Section 5, we discuss the results, highlighting the interpretability, trustworthiness, and explainability of the models. Section 6 addresses the limitations and future work. Finally, Section 7 concludes the paper and outlines future work.

## 2. Related Work

Recent advancements in multimodal document processing have significantly enhanced the capabilities of models in text detection, recognition, and information extraction. In this section, we review the relevant literature, focusing on the methods most pertinent to our work.

### 2.1. Text Detection

Accurate text detection is a foundational step for structured data extraction from unstructured documents. Recent methods have focused on improving accuracy and efficiency across various text orientations, sizes, and backgrounds. DBNet [25] introduced a real-time differentiable binarization method that improved boundary localization while maintaining computational efficiency. FAST [4] further improved detection speed and accuracy for irregular text shapes, while MixNet [44] utilized receptive fields and feature fusion to tackle complex scenes, marking significant strides in robust text detection.

### 2.2. Text Recognition

In text recognition, the evolution from sequence models to Transformer-based architectures has yielded models resilient to diverse fonts, distortions, and complex layouts. Early models such as CRNN [35], SAR [23], and MASTER [29] established the groundwork for sequence and attention-based recognition. More recent Transformer-based models, such as ViTSTR [2] and PARSeq [3], further enhanced accuracy by capturing long-range dependencies. Innovations like MaskOCR [31], TrOCR [24], and DTrOCR [10] have integrated masked pretraining with encoder-decoder frameworks, achieving SOTA recognition accuracy across challenging scenarios.

### 2.3. Information Extraction

Recent advancements in MLLMs have utilized both OCR-free and OCR-dependent architectures. OCR-free models, such as Donut [20], UDOP [37], and OmniParser [39], bypass traditional OCR steps, reducing pipeline complexity and mitigating error propagation. Advanced OCR-free MLLMs, including LLaVAR [45], Pixtral-12B [1], Llama 3.2-11B [9], InternVL v2 [5, 6], Qwen-VL [41], and LLaVA-OneVision [22], extend multimodal comprehension, offering efficient extraction of structured data without dependency on external OCR processes.

In contrast, OCR-dependent models integrate OCR data to enhance document layout and positional comprehension. ICL-D3IE [13] and LATIN-Prompt [42] incorporate positional data, though this can lead to increased input sequence length and slower inference. Recent approaches such as Cream [21] and InstructDoc [36] streamline these processes by employing additional encoders to integrate OCR information, improving inference efficiency without compromising comprehension.

Despite these improvements, spatial precision and explainability remain challenging for document VQA applications. Our work addresses these challenges by introducing an integrated MLLM approach that merges text recognition and spatial understanding within a unified model, bypassing the need for separate OCR components and advancing spatial localization in document analysis.

### 2.4. Layout-Aware Document Understanding

Incorporating layout-specific information has proven effective in enhancing spatial comprehension in document understanding. LayoutLLM [30] employs a layout instruction tuning strategy to improve the model’s ability to interpret document layouts. DocLayLLM [26] encodes OCRed textual, visual, and positional information directly within the model, removing the need for additional document encoders and refining comprehension through a Chain-of-Thought (CoT) annealing process. LayTextLLM [28] introduces a Spatial Layout Projector to convert OCR-derived coordinates into bounding box tokens, allowing seamless integration of spatial layouts with textual data. While these models enhance layout awareness, they often require complex adaptations or additional components that may affect model generality and increase computational overhead.

In summary, recent developments in multimodal document processing and layout-aware models have significantly advanced Document VQA capabilities, yet challenges in spatial precision, interpretability, trustworthiness and computational efficiency remain. These research gaps motivated our work, leading us to develop an innovative approach that addresses the challenges.

Table 1. Model OCR-Dependent Comparison using ANLS metric on Document VQA and QA for VIE

Model Category	Models	Document VQA		QA for VIE	
		DocVQA	FUNSD	CORD	SROIE
<b>Text</b>	Llama2-7B-Chat [38]	64.99	48.20	47.70	68.97
	Llama3-8B-Instruct [9]	51.79	68.57	52.31	61.24
<b>Text + BBox</b>	LayTextLLM ( <b>Llama2-7B</b> ) [28]	72.83	78.65	70.81	83.27
<b>Text + BBox + Image</b>	LayoutLLM-7B <sub>CoT</sub> ( <b>Llama2-7B</b> ) [30]	74.25	78.65	62.21	70.97
	LayoutLLM-7B <sub>CoT</sub> ( <b>Vicuna-1.5-7B</b> ) [30]	74.27	79.98	63.10	72.12
	DocLayLLM ( <b>Llama2-7B</b> ) [26]	72.83	78.65	70.81	83.27
	DocLayLLM ( <b>Llama3-7B</b> ) [26]	<b>78.40</b>	<b>84.12</b>	71.34	84.36
	DLaVA <sub>OCR-Dependent</sub> ( <b>Pixtral-12B</b> )	74.02	79.57	<b>84.41</b>	<b>90.45</b>

Table 2. Model OCR-Free Comparison using ANLS metric on Document VQA and QA for VIE

Model Category	Models	Document VQA		QA for VIE	
		DocVQA	FUNSD	CORD	SROIE
<b>Image</b>	Llama3.2-11B [9]	78.4	65.02	42.96	61.42
	Pixtral-12B [1]	80.71	78.26	79.08	82.24
	LLaVA-OneVision-7B [22]	47.59	22.82	32.43	12.10
	Qwen2-VL-7B [40]	64.15	47.44	15.98	45.17
	InternVL2-8B [7]	71.26	57.58	55.88	81.55
<b>Image + BBox</b>	DLaVA <sub>OCR-Free</sub> ( <b>Pixtral-12B</b> )	<b>85.91</b>	<b>87.57</b>	<b>82.08</b>	<b>91.42</b>

### 3. DLaVA

This section describes the two approaches used in DLaVA for information extraction from documents, as illustrated in Figure 2. As discussed in Section 2.3, both the OCR-dependent and OCR-free methods utilize an MLLM to accurately extract and locate information, but they differ in their reliance on OCR, where one approach is computationally efficient, and the other is structurally more accurate. By incorporating both approaches, we aim to achieve an optimal balance of structural accuracy and computational efficiency.

#### 3.1. DLaVA (OCR-dependent)

In the OCR-dependent approach, as shown in Figure 2:

1. **Text Detection Module:** The original document image  $I$  is processed using a text detection model, specifically DB-ResNet-50 [25], as shown in step 1 in Figure 2. This model outputs bounding boxes for each text segment in the image. The detected bounding boxes are represented as:

$$B = \{B_0, B_1, \dots, B_n\}$$

where each  $B_i$  is a bounding box coordinate  $[x_{i1}, y_{i1}, x_{i2}, y_{i2}]$ . Each bounding box  $B_i$  is used to crop a segment of the image  $I$ , isolating individual

words or phrases. The cropped image for  $B_i$  is denoted by:

$$C_i = I[B_i]$$

2. **Text Recognition Module:** Each cropped image  $C_i$  is passed to a text recognition model, PARSeq [3], as illustrated in step 2 of the diagram. This model applies OCR to convert the visual text into strings. The OCR output for each cropped image is:

$$T_i = \text{OCR}(C_i)$$

where  $T_i$  represents the recognized text associated with the bounding box  $B_i$ . This text recognition step yields a set of paired text and bounding box data:

$$\{(T_0, B_0), (T_1, B_1), \dots, (T_n, B_n)\}$$

3. **MLLM Processing:** Finally, the Pixtral-12B model [1] in step 6 takes the output of step 2 (recognized texts), the outputs of step 3 (the boundary box coordinates), and the question  $Q$  as inputs; and it generates the answer  $A$ , and the boundary box of the answer  $B_A$  as outputs. For a question  $Q$  such as ‘‘What is the content in the DEPARTMENT NAME field?’’, the model identifies the bounding box and answer text  $A$  as:

$$A, B_A = \text{MLLM}(Q, \{(T_0, B_0), (T_1, B_1), \dots, (T_n, B_n)\})$$

Here,  $A$  denotes the answer (e.g., “SCIENCE & TECHNOLOGY”), and  $B_A$  is the bounding box where this answer is located.

### 3.2. DLaVA(OCR-free)

The OCR-free approach involves the following steps:

1. **Text Detection Module:** Similar to the OCR-dependent approach, we do text-detection (step 1) resulting in a set of bounding boxes (step 3) and the corresponding cropped images.
2. **Constructed Image Creation:** Instead of performing OCR on each cropped image (corresponding to the boundary boxes), the bounding box images are arranged to form a “constructed image,” illustrated in step 5. Each bounding box  $B_i$  is assigned a unique ID for easy reference. The constructed image,  $I_C$ , is an assembly where each line contains a cropped image, followed by their boundary box ID  $B_i$ :

$$I_C = \{(C_0, B_0), (C_1, B_1), \dots, (C_n, B_n)\}$$

For example, if the document contains sentences like “THE STATE OF TEXAS...”, after text detection, we obtain cropped images of each individual word, such as “THE” ( $C_0$ ), “STATE” ( $C_1$ ), “OF” ( $C_2$ ), and “TEXAS” ( $C_3$ ). In the constructed image  $I_C$ , each line would display the words with their bounding box IDs in sequence: the first line would contain “THE ( $B_0$ )”, the second line “STATE ( $B_1$ )”, and so on.

3. **Information Extraction Model:** In parallel, the Pixtral-12B model [1] (step 4) receives the input image  $I$  and the query  $Q$  to generate the answer text  $A$ . These generated answers along with their corresponding questions together ( $Q+A$ ) go to step 6.
4. **MLLM Processing:** Finally, the Pixtral-12B model [1] in step 6 takes the output of step 3 (boundary box coordinates), step 4 ( $Q+A$ ), and step 5 (constructed image  $I_C$ ) as inputs; and generates the boundary boxes for the answers ( $B_A$ ) and return them together with the answers ( $A$ ) generated in step 4.

In both approaches, the goal is to produce the answer  $A$  to the query  $Q$  along with the bounding box  $B_A$ , enabling precise extraction and localization of information from the document image.

## 4. Experiments

### 4.1. Datasets and Experimental Setup

We evaluated our proposed model on several well-established, text-rich document datasets commonly used for Visual Information Extraction (VIE) and Document Visual Question Answering (VQA) tasks. For VIE-related question answering, we utilized the FUNSD [19], CORD [33],

and SROIE [17] datasets. In the domain of Document VQA, we assessed performance using the DocVQA [32] dataset. All models, including our proposed approach and baseline comparisons, were trained and evaluated on a single NVIDIA A100 GPU with 80 GB of memory. This consistent computational environment ensures fair and reliable comparisons across different experimental settings.

We evaluated our model using two metrics to assess textual accuracy and spatial alignment, following established protocols. For textual accuracy, we used ANLS [43], which measures normalized Levenshtein distance between predicted and ground truth answers, with values from 0 to 1 (1 indicating a perfect match). For spatial alignment, we employed IoU [34], which assesses overlap between predicted and ground truth bounding boxes. Performance was evaluated using  $\text{mAP@IoU}[0.50:0.95]$ , where mean Average Precision (mAP) is computed across IoU thresholds from 0.50 to 0.95 in increments of 0.05. This metric captures the model’s ability to localize answer regions accurately across varying levels of spatial precision, providing a comprehensive measure of answer correctness and localization.

### 4.2. Baseline Models

To evaluate the effectiveness of our proposed approach, we compare it against several baseline models, categorized into OCR-free and OCR-dependent MLLMs. For OCR-free MLLMs, we selected state-of-the-art models as appropriate baselines for document-oriented VQA and VIE tasks. These include PixTral-12B [1], InternVL v2-8B [5, 6], Qwen-VL 7B [40], LLaVA-OneVision (OV) 7B [22], and LLaMA 3.2-11B [9]. For OCR-dependent models, we selected LLaMA 2-7B-Chat [38], LLaMA 3-8B-Instruct [9], LayoutLLM-7B [30], DocLayLLM [26], and LayTextLLM [28] as appropriate baselines due to their strong performance in document-oriented VQA and VIE tasks, along with their effective integration of OCR-derived information.

### 4.3. Ablation study

In our ablation study, we focus only on evaluating the OCR-Free approach as the OCR-Dependent model relies on interdependent components, and removing any of these components would prevent it from functioning effectively. For the OCR-Free model, we conduct two specific ablation experiments. In **Ablation 1**, we feed the original input image  $I$  as an additional input to the final MLLM model (step 6 in Figure 2) besides rest of the input components. In **Ablation 2**, we remove the information extraction step (step 4) entirely and rely solely on the final MLLM (step 6) for both question-answering and providing the corresponding bounding boxes for the answers. These ablations allow us to assess the significance of each component in the OCR-Free pipeline and understand their contributions to overall performance.

Table 3. Model Comparison on Document QA for VIE (mAP@IOU[0.50:0.95])

Model Category	DocVQA	FUNSD	CORD
OCR-Dependent	44.03	38.69	52.21
OCR-Free	46.22	45.52	57.86

## 5. Results and Discussion

In this section, we present a comprehensive analysis of our proposed models’ performance compared to SOTA baseline methods on Document VQA and VIE tasks.

### 5.1. Performance Analysis of OCR-Dependent Models

Here, we analyze the performance of our OCR-dependent model, DLaVA<sub>OCR-Dependent</sub>, in comparison with existing baseline models. The results are summarized in Table 1, which presents the ANLS scores on Document VQA datasets (DocVQA) and VIE datasets (FUNSD, CORD, and SROIE).

- Document VQA Performance:** DLaVA<sub>OCR-Dependent</sub> achieves strong performance on the DocVQA benchmark, scoring 74.02% in ANLS, closely aligning with the results of top-performing baselines such as LayoutLLM-7B<sub>CoT</sub> (Vicuna-1.5-7B) at 74.27% and DocLayLLM (Llama3-7B) at 78.40%. Unlike DocLayLLM and LayoutLLM-7B, which requires computationally expensive CoT pretraining and annealing, DLaVA is out-of-the-box, operating in a zero-shot paradigm, thereby offering efficient performance with reduced computational overhead.
- VIE Task Performance:** In VIE tasks, DLaVA<sub>OCR-Dependent</sub> demonstrates notable advantages, particularly on the CORD and SROIE datasets, achieving ANLS scores of 84.41% and 90.45%, respectively, outperforming other OCR-dependent models, including DocLayLLM (Llama3-7B) at 71.34% for CORD and 84.36 for SROIE. On the FUNSD dataset, DLaVA scores 79.57%, slightly below DocLayLLM’s 84.12%, but without the need for extensive pretraining. These results underscore DLaVA’s ability to deliver competitive accuracy in document understanding tasks with significantly lower computational demands.

In addition, we evaluated the IoU performance of DLaVA<sub>OCR-Dependent</sub>, as presented in Table 3, for the DocVQA, FUNSD, and CORD datasets. The IoU scores obtained were 38.69% for FUNSD, 52.21% for CORD, and 44.03% for DocVQA. While these IoU scores are lower than the corresponding ANLS scores, they provide valuable insights into the model’s spatial alignment capabilities. The lower IoU scores can be attributed to several fac-

tors inherent in document processing tasks [Appendix B], e.g., IoU is sensitive to even slight misalignments in bounding box placement, and complex document layouts with small fonts, stylized text, or overlapping elements make precise spatial localization challenging. The combination of ANLS and IoU allows us to capture both the textual accuracy and spatial precision of the model’s predictions, offering a more holistic assessment. The ANLS scores reflect strong text recognition and content accuracy, while the IoU scores highlight areas where fine-grained spatial alignment can further enhance answer localization. Using both metrics, we gain a nuanced understanding of the model’s strengths and areas for refinement, demonstrating that DLaVA<sub>OCR-Dependent</sub> is adept at recognizing textual content while offering targeted insights into the precision of answer localization within document images.

The enhanced performance of DLaVA<sub>OCR-Dependent</sub> can be attributed to two primary factors. First, by leveraging both textual and visual features alongside bounding box information, our model effectively captures the complex relationships within documents. Second, the use of the PixTral-12B [1] backbone provides a larger parameter space, enhancing the model’s capacity to understand and generate accurate responses.

Table 4. Ablation Study on Model OCR-Free Comparison using ANLS metric on Document VQA and QA for VIE

Models	DocVQA	FUNSD	CORD	SROIE
DLaVA <sub>OCR-Free</sub>	<b>85.91</b>	<b>87.57</b>	<b>82.08</b>	<b>91.42</b>
Ablation 1	83.55	83.28	79.08	85.36
Ablation 2	82.26	84.35	82.91	86.02

Table 5. Ablation Study on Model OCR-Free Comparison using IoU (mAP@IOU[0.50:0.95]) metric on Document VQA and QA for VIE

Models	DocVQA	FUNSD	CORD
DLaVA <sub>OCR-Free</sub>	<b>46.22</b>	<b>45.52</b>	<b>57.86</b>
Ablation 1	44.01	32.71	45.45
Ablation 2	39.41	37.12	46.69

### 5.2. Performance Analysis of OCR-Free Models

Similarly, we examine the performance of our OCR-free model, DLaVA<sub>OCR-Free</sub>, in comparison with existing OCR-free baseline models. The results are summarized in Table 2, which presents the ANLS scores on Document VQA datasets (DocVQA) and VIE datasets (FUNSD, CORD, and SROIE).

- Document VQA Performance:** Our OCR-free model, DLaVA<sub>OCR-Free</sub> (Pixtral-12B), achieves the highest ANLS scores on the DocVQA dataset, with scores of 85.91%. This represents a significant improvement over the previous best OCR-free model, Pixtral-12B, which scored 80.71% on DocVQA.
- VIE Task Performance:** In VIE tasks, DLaVA<sub>OCR-Free</sub> demonstrates exceptional performance across all datasets. In the FUNSD dataset, it achieves an ANLS score of 87.57%, outperforming Pixtral-12B’s 78.26% by a substantial margin. In the CORD dataset, it scored 82.08, surpassing the next-best OCR-free model, Pixtral-12B, which scored 79.08%. In the SROIE dataset, it attains an ANLS score of 91.42%, significantly higher than Pixtral-12B’s 82.24%.

Additionally, we have evaluated the IoU performance of the OCR-Free model, as presented in Table 3, for the DocVQA, FUNSD, and CORD datasets. The IoU scores obtained are 57.86% for CORD, 45.52% for FUNSD, and 46.22% for DocVQA. The comparatively lower values of IoU can be explained based on the same logic as presented in Section 5.1.

The remarkable performance of DLaVA<sub>OCR-Free</sub> can be attributed to a number of factors. First, by operating without reliance on OCR, our model eliminates error propagation from text recognition inaccuracies by utilizing the visual language model’s inherent text recognition capabilities and employing the constructed image with bounding box identifiers. This approach leverages the MLLM’s strength in interpreting visual content directly, resulting in higher overall accuracy, as evidenced by higher ANLS scores. Second, incorporating bounding box information directly into the model enhances spatial reasoning, allowing for more precise answer localization within documents. Although the IoU scores indicate there is room for improvement in spatial alignment, the integration of bounding boxes still significantly contributes to the model’s understanding of document layouts.

Furthermore, the OCR-Free approach proves advantageous over OCR-dependent methods due to the reduced context window requirements. By sending all identified text regions as a single constructed image, we avoid the need to input each word or text separately, minimizing context length and optimizing model performance. This efficiency, combined with DLaVA’s ability to integrate visual and textual information effectively, enables it to handle diverse document layouts and content without additional preprocessing steps. Operating in a zero-shot learning paradigm, DLaVA adapts efficiently to various document types, demonstrating strong generalization capabilities across different datasets. The synergy of these factors leads to a robust model that excels in text recognition and, to a substantial extent, spatial localization, thereby advancing

the field of document understanding.

### 5.3. Ablation Study on the OCR-Free Model

The results of the ablation study are summarized in Table 4 and Table 5. Our model, DLaVA<sub>OCR-Free</sub>, achieves the highest ANLS and IoU scores across all datasets, confirming the effectiveness of integrating both bounding box annotations and the information extraction step.

In Ablation 1, where we provide the original input image as an additional input, there was a decline in performance across all datasets. For instance, the ANLS score on the DocVQA dataset decreased from 85.91% to 83.55%, and the IoU score dropped from 46.22% to 44.01%. This decline can be attributed to redundant information when including the input image along with other pipeline components, as the required information was already extracted in prior steps. This redundancy likely introduces noise, deteriorating the final model’s performance.

In Ablation 2, where we removed the information extraction step entirely, we observed mixed results. While there was a slight improvement in the ANLS score on the CORD dataset (from 82.08% to 82.91%), the IoU score decreased from 57.86% to 46.69%. This suggests that separating tasks into distinct steps (i.e., finding answers in one step and boundary box annotation in another) enhances performance, as the model is less effective when tasked with multiple objectives simultaneously. Overall, the full pipeline benefits from explicit extraction and integration of textual and spatial information, particularly for precise answer localization.

### 5.4. Interpretability of the Proposed DLaVA Model

Interpretability refers to understanding the internal workings of the model, such as pipeline design and architecture. The proposed DLaVA model enhances interpretability through its OCR-free architecture, particularly in its handling of document images and spatial data. Key aspects that improve interpretability include:

- Visual Representation of Text Regions:** DLaVA’s OCR-free approach utilizes a constructed image  $I_C$ , where detected text regions are organized with unique bounding box IDs. This arrangement preserves spatial relationships, allowing easy inspection of text areas directly within the document layout.
- Direct Mapping Between Inputs and Outputs:** DLaVA generates answers associated with specific bounding box IDs  $B_A$ , establishing a transparent link between the input text regions and output answers, which aids in understanding the model’s decision-making process.
- Simplified Pipeline without OCR Complexity:** By bypassing OCR and focusing on visual and spatial patterns with the Pixtral-12B MLLM model [1], DLaVA avoids

OCR-related complexities, offering a clearer interpretive pathway through the document’s visual content.

- **Transparent and Modular Processing Steps:** The OCR-free pipeline is composed of distinct stages—from text detection with DB-ResNet-50 [25] to constructed image creation—each of which can be independently inspected and analyzed, adding to the model’s interpretability.

Through these design choices, the DLaVA model provides an interpretable framework for Document Visual Question Answering, offering users a more transparent and trustworthy system for document analysis.

## 5.5. Explainability and Trustworthiness

Trustworthiness in Document VQA is crucial, and the proposed DLaVA model enhances it by delivering precise answer localization, allowing users to verify answers directly within the document images.

The assignment of unique bounding box IDs to text regions in  $I_C$  strengthens spatial reasoning and answer localization. By referencing these bounding boxes during response generation, the model improves both accuracy and traceability, enabling users to pinpoint exact answer locations within the document. This spatial grounding provides a verifiable link between the model’s outputs and the visual content, bolstering user trust in the model’s responses.

In terms of explainability, our model provides insights into its decision-making through the relationship between ANLS and IoU scores. While high ANLS scores confirm textual accuracy, IoU evaluates the precision of answer localization, offering a multi-dimensional view of model performance. However, despite these contributions to explainability, achieving complete clarity remains challenging due to the inherent complexities of MLLMs. These models’ probabilistic nature and intricate internal workings can sometimes obscure the exact rationale behind certain outputs. Overall, these design elements contribute significantly to the model’s trustworthiness and enhance interpretability, providing users with confidence in its outputs while acknowledging the limitations in fully transparent explainability.

## 6. Limitations and Future Work

**Limitations:** While our model demonstrates superior performance on benchmark datasets, certain limitations remain. First, though we have tested our model’s performance on the benchmark datasets and achieved better results compared to the SOTA baselines, it still does not suffice for all real-world applications because there could be complex situations where the model needs to have the ability to comprehend visual charts and images, etc. Additionally, the model’s reliance on standard MLLM outputs introduces occasional unpredictability due to their probabilistic nature,

such as formatting inconsistencies in JSON responses with nested structures, which may require post-processing adjustments.

**Future Work:** Our future work focuses on addressing challenges associated with lower IoU scores by refining bounding box annotations through fine-tuning techniques such as LoRA[14], LoRA+[12], QLoRA[8], and DoRA[27]. Additionally, we plan to utilize Retrieval-Augmented Generation (RAG) [11] to enhance the model’s performance and adaptability to diverse document types. Though our current model mainly focuses on enhancing the interpretability and trustworthiness, we aim to further improve the model’s explainability as well, making it more suitable for a broader range of real-world applications.

## 7. Conclusion

This paper introduces DLaVA, a document language model equipped not only to answer questions based on information in document images but also to localize it via bounding boxes around textual answers within the images. By directly integrating image annotation capabilities into the MLLM pipeline, DLaVA eliminates the need for supplementary encoders or extensive techniques like CoT. Operating in an out-of-the-box learning paradigm, it generalizes across diverse document types without additional training, ensuring both adaptability and high accuracy.

Our approach addresses critical limitations of existing models by removing the separate text recognition component and enhancing spatial accuracy. The integration of bounding box annotations enhances spatial reasoning, leading to higher accuracy. This advancement not only streamlines the processing pipeline but also significantly improves the explainability and precision of Document VQA tasks.

Experimental results demonstrate that DLaVA achieves SOTA performance on benchmark datasets while enhancing user trust and reducing the risk of AI hallucinations through spatially grounded responses. By bridging the gap between visual data and linguistic queries with precise answer localization, DLaVA sets a new standard for reliability and transparency in document understanding, thus laying the groundwork for more trustworthy and interpretable AI systems.

## References

- [1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 3, 4, 5, 6, 7
- [2] Rowel Atienza. Vision transformer for fast and efficient scene text recognition. In *International conference on document analysis and recognition*, pages 319–334. Springer, 2021. 3



- [3] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, pages 178–196. Springer, 2022. 3, 4
- [4] Zhe Chen, Jiahao Wang, Wenhai Wang, Guo Chen, Enze Xie, Ping Luo, and Tong Lu. Fast: Faster arbitrarily-shaped text detector with minimalist kernel representation. *arXiv preprint arXiv:2111.02394*, 2021. 3
- [5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 3, 5
- [6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 3, 5
- [7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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, 2024. 4
- [8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024. 8
- [9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 4, 5
- [10] Masato Fujitake. Dtrocr: Decoder-only transformer for optical character recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8025–8035, 2024. 3
- [11] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. 8
- [12] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024. 8
- [13] Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 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, 2023. 3
- [14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 8
- [15] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022. 2
- [16] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024. 2
- [17] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019. 5
- [18] Md Farhan Ishmam, Md Sakib Hossain Shovon, Muhammad Firoz Mridha, and Nilanjan Dey. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, page 102270, 2024. 2
- [19] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 1–6. IEEE, 2019. 5
- [20] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 3
- [21] Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. Visually-situated natural language understanding with contrastive reading model and frozen large language models. *arXiv preprint arXiv:2305.15080*, 2023. 3
- [22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3, 4, 5
- [23] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8610–8617, 2019. 3
- [24] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13094–13102, 2023. 3
- [25] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11474–11481, 2020. 3, 4, 8
- [26] Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient and effective multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*, 2024. 2, 3, 4, 5
- [27] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng,

- and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 8
- [28] Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, et al. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. *arXiv preprint arXiv:2407.01976*, 2024. 2, 3, 4, 5
- [29] Ning Lu, Wenwen Yu, Xianbiao Qi, Yihao Chen, Ping Gong, Rong Xiao, and Xiang Bai. Master: Multi-aspect non-local network for scene text recognition. *Pattern Recognition*, 117: 107980, 2021. 3
- [30] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutlm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640, 2024. 2, 3, 4, 5
- [31] Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Maskocr: Text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311*, 2022. 3
- [32] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 5
- [33] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. 1, 5
- [34] Hamid Rezaatofghi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 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, 2019. 2, 5
- [35] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016. 3
- [36] Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19071–19079, 2024. 3
- [37] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19254–19264, 2023. 3
- [38] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 4, 5
- [39] Jianqiang Wan, Sibao Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. Omniparser: A unified framework for text spotting key information extraction and table recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15641–15653, 2024. 3
- [40] 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. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 4, 5
- [41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3
- [42] Wenjin Wang, Yunhao Li, Yixin Ou, and Yin Zhang. Layout and task aware instruction prompt for zero-shot document image question answering. *arXiv preprint arXiv:2306.00526*, 2023. 3
- [43] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007. 2, 5
- [44] Yu-Xiang Zeng, Jun-Wei Hsieh, Xin Li, and Ming-Ching Chang. Mixnet: toward accurate detection of challenging scene text in the wild. *arXiv preprint arXiv:2308.12817*, 2023. 3
- [45] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Lllavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 3

# DLaVA: Document Language and Vision Assistant for Answer Localization with Enhanced Interpretability and Trustworthiness

## Supplementary Material

### Appendix

#### Appendix A: Examples of Ground Truth Answer Annotations

Appendix A presents some examples of ground truth annotations from the CORD and FUNSD datasets. These examples illustrate how document understanding tasks handle diverse document formats and content types.

Figure 3a depicts a document example from the FUNSD dataset, showcasing the structured layout of annotated key-value pairs in a form-like document. It highlights the ability to capture complex relationships between fields, such as dates, phone numbers, and textual descriptions.

Figure 3b displays a receipt example from the CORD dataset, emphasizing the annotation of essential receipt components like item quantity, unit price, total amount, and item names. This example underscores the importance of annotating critical transactional information typically found in unstructured receipt data.

Figure 3c demonstrates another similar receipt from the CORD dataset.

#### Appendix B: Examples of Predicted Answer Annotations

Appendix B presents the answers and annotations generated by our proposed model, DLaVa (OCR-Free), for the same documents discussed in Appendix A. These examples provide insights into the model’s ability to handle diverse document formats, such as structured forms and unstructured receipts, without relying on OCR. The illustrations highlight how DLaVa identifies key information and maps it to corresponding document regions, showcasing both its strengths and limitations. For example, the model demonstrates high semantic accuracy in extracting answers, as reflected in high ANLS scores, but sometimes struggles with precise spatial alignment, leading to lower IoU scores in some cases. By comparing these predictions with the ground truth annotations in Appendix A, readers can better understand the model’s performance and areas for improvement.

Figure 3c shows a sample document where both the answers and their locations were identified with high precision by our model (as shown in Figure 4c). This resulted in an ANLS score of 100% and an IoU nearly 100%, as the model accurately captured the ground truth information.

Analysis for low IoU score between predicted and ground truth annotations for some cases:

1. First, let us analyze a sample from FUNSD dataset. Figure 3a shows the ground truth answers for this sample along with their annotations, and Figure 4a shows the answers and annotations returned by our model DLaVa (OCR-Free) for the same document.

The IoU score for the “Message” field of this document was observed to be 5.89%, despite achieving a high ANLS score of 70.73%. This discrepancy can be attributed to the differing interpretation of the message’s spatial extent between the ground truth (Figure 3a) and the predicted annotations (Figure 4a).

In the ground truth annotation, the bounding box includes the specific textual region containing the date component (“Jan 12, 1999”) within the broader message context, towards the end of the box. However, our model’s prediction restricts the bounding box to the “Message” content, omitting the date. This misalignment results in a smaller predicted bounding box compared to the ground truth, thereby reducing the overlap and, consequently, the IoU score.

This outcome highlights a common challenge in document understanding tasks, where predicted annotations may fail to encapsulate all semantically relevant content included in the ground truth. The low IoU score does not necessarily imply poor semantic accuracy but instead reflects a divergence in bounding box definitions.

2. Let us analyze another sample from the CORD dataset. Figure 3b shows the ground truth answers for this sample along with their annotations, and Figure 4b shows the answers and annotations returned by our model DLaVa (OCR-Free) for the same document.

Here, in the task of extracting the “Total Price of Menu” from receipt images, we observed that the IoU score was 0%, despite achieving a perfect ANLS score of 100%. This mismatch highlights an important limitation in the spatial alignment of predicted bounding boxes with the ground truth.

In this instance, the value “11,000” appears multiple times in the document, corresponding to different semantic fields (e.g., item price, subtotal, total price). While the model successfully identified the correct value for the “Total Price of Menu,” it incorrectly annotated a bounding box around the “11,000” value associated with the total price of receipt rather than the ground truth lo-

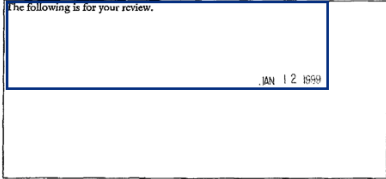
cation of the “11,000” value corresponding to the total price of the menu. This resulted in no overlap between the predicted and ground truth bounding boxes, leading to an IoU score of 0%.

This case illustrates a common challenge in structured document understanding tasks where identical values appear in different semantic contexts. Resolving such issues requires incorporating additional contextual understanding into the model to ensure that annotations are correctly aligned with the intended semantic field. As a part of the future work, we plan to explore incorporating positional priors, cross-field dependencies, or explicit disambiguation mechanisms to improve alignment between predictions and ground truth annotations.

FAX TRANSMISSION



DATE: January 11, 1999  
 CLIENT NO.: 18557 002  
 MESSAGE TO: Dewey Tedder  
 COMPANY: Lorillard Tobacco Company  
 FAX NUMBER: 336/373-6917  
 PHONE: 336/373-6750  
 FROM: Andy Zausner and Rob Manjias  
 PHONE: 202/828-2259 and 202/828-2241  
 PAGES (including Cover Sheet): 2 HARD COPY TO FOLLOW: YES X NO



If your receipt of this transmission is in error, please notify this firm immediately by collect call to our Facsimile Department at 202-861-9106, and send the original transmission to us by return mail at the address below.

This transmission is intended for the sole use of the individual and entity to whom it is addressed, and may contain information that is privileged, confidential and exempt from disclosure under applicable law. You are hereby notified that any dissemination, distribution or duplication of this transmission by someone other than the intended addressee or its designated agent is strictly prohibited.

(a) Document Example from FUNSD Dataset



(b) Receipt from CORD Dataset



(c) Another receipt from the CORD Dataset

Figure 3. Illustrative Examples of Ground Truth Answer Annotations in Documents from the CORD and FUNSD Datasets

FAX TRANSMISSION



DATE: January 11, 1999  
 CLIENT NO.: 18557002  
 MESSAGE TO: Dewey Tedder  
 COMPANY: Lorillard Tobacco Company  
 FAX NUMBER: 336/373-6917  
 PHONE: 336/373-6750  
 FROM: Andy Zausner and Rob Manjas  
 PHONE: 202/828-2259 and 202/828-2241  
 PAGES (including Cover Sheet): 2 HARD COPY TO FOLLOW: YES X NO  
 MESSAGE: The following is for your review.



If your receipt of this transmission is in error, please notify this firm immediately by collect call to our Facsimile Department at 202-861-9106, and send the original transmission to us by return mail at the address below.

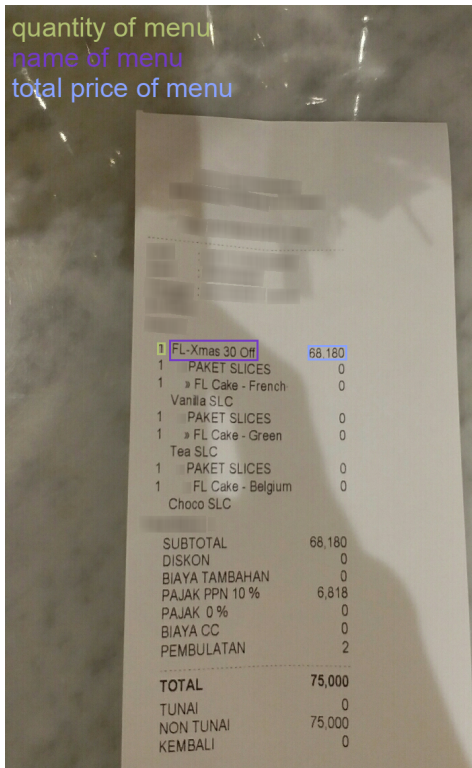
This transmission is intended for the sole use of the individual and entity to whom it is addressed, and may contain information that is privileged, confidential and exempt from disclosure under applicable law. You are hereby notified that any dissemination, distribution or duplication of this transmission by someone other than the intended addressee or its designated agent is strictly prohibited.

83443897

(a) FUNSD-high ANLS, Low IOU



(b) CORD-high ANLS, Low IOU



(c) CORD-high ANLS, High IOU

Figure 4. Examples of Predicted Answer Annotations in Documents from the CORD and FUNSD Datasets