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

Ahmad Mohammadshirazi [1] **Pinaki Prasad Guha Neogi** [1] **Ser-Nam Lim** [2] **Rajiv Ramnath** [1]

Abstract

Document Visual Question Answering (VQA) demands robust integration of text detection, recognition, and spatial reasoning to interpret complex document layouts. In this work, we introduce DLaVA, a novel, training-free pipeline that leverages Multimodal Large Language Models (MLLMs) for zero-shot answer localization in order to improve trustworthiness, interpretability, and explainability. By leveraging an innovative OCR-free approach that organizes text regions with unique bounding box IDs, the proposed method preserves spatial contexts without relying on iterative OCR or chain-of-thought reasoning, thus substantially reducing the computational complexity. We further enhance the evaluation protocol by integrating Intersection over Union (IoU) metrics alongside Average Normalized Levenshtein Similarity (ANLS), thereby ensuring that not only textual accuracy is considered, but spatial accuracy is taken into account, ultimately reducing the risks of AI hallucinations and improving trustworthiness. Experiments on benchmark datasets demonstrate competitive performance compared to state-of-the-art techniques, with significantly lower computational complexity and enhanced accuracies and reliability for high-stakes applications. The code and datasets utilized in this study for DLaVA are accessible at: https://github.com/ahmad-shirazi/AnnotMLLM.

1. Introduction

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 contextual understanding to bridge the gap between visual data and linguistic queries (Ishmam et al., 2024). Figure 1 presents some examples of visual information extraction, showcasing document annotations from the CORD dataset (see Appendix A and B for more details).

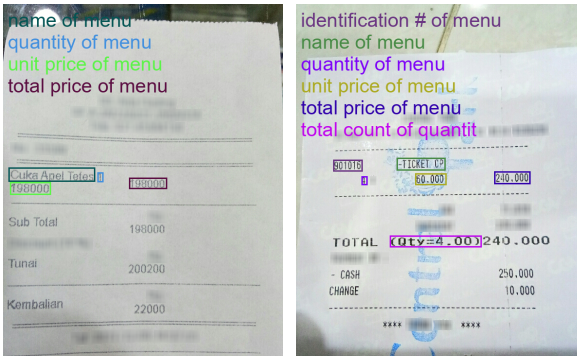


Figure 1. Examples of visual information extraction on images from the CORD dataset (Park et al., 2019): questions are displayed at the top in colored fonts, with the corresponding answers highlighted by matching colored boundary boxes.

Existing approaches, such as LayoutLMv3 (Huang et al., 2022), LayoutLLM (Luo et al., 2024), LayTextLLM (Lu et al., 2024), and DocLayLLM (Liao et al., 2024), have made significant progress in visual question answering and layout analysis. However, these methods come with several limitations. They often rely on chain-of-thought (CoT) reasoning or iterative OCR processes for spatial grounding, which incur high computational costs and require extensive fine-tuning. Furthermore, these methods are evaluated on metrics like Average Normalized Levenshtein Similarity (ANLS) (Yujian & Bo, 2007) that focus primarily on textual accuracy while overlooking the spatial correctness of the predicted answers. As a result, these approaches typically lack precise answer localization, thereby limiting interpretability and explainability—challenges that are particularly critical in high-stakes applications such as legal, medical, and

*Equal contribution [1]Department of Computer Science and Engineering, Ohio State University, Ohio, US [2]Department of Computer Science, University of Central Florida, Florida, US. Correspondence to: Ahmad Mohammadshirazi <mohammadshirazi.2@osu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

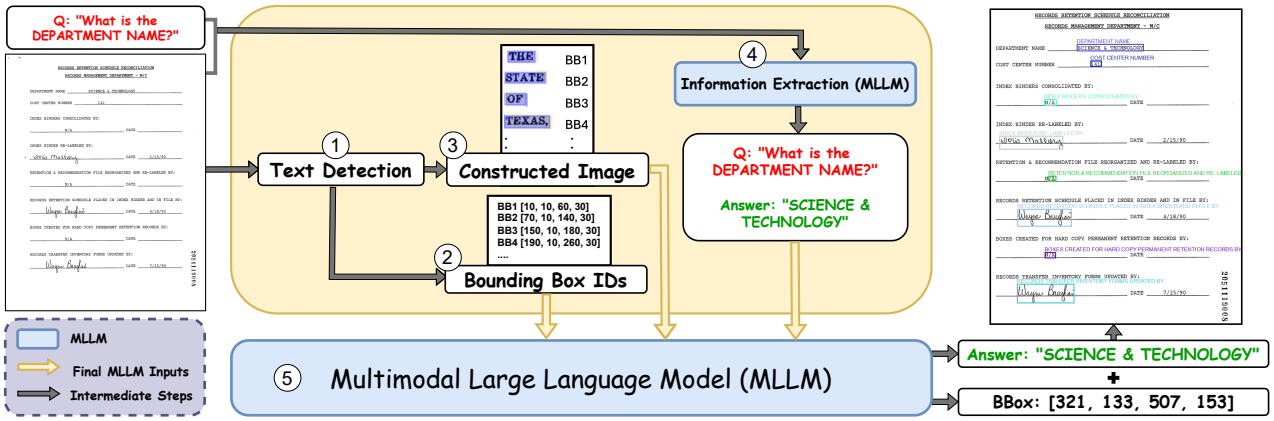


Figure 2. DLaVA Model Architecture. This diagram illustrates our final single-pipeline design. In the text detection step, detected text regions generate two outputs: a series of cropped images that are reorganized into a "constructed image" with unique bounding box identifiers (e.g., BB1, BB2, BB3, etc.) and their corresponding bounding box coordinates (e.g., BB1 [10, 10, 60, 30], BB2 [70, 10, 140, 30], etc.). The approach then leverages a two-stage MLLM pipeline. In Stage 1, the original image and the user's question are provided to the MLLM to derive an initial textual answer. In Stage 2, the constructed image—comprising all cropped images with their BB IDs—along with the recorded bounding box coordinates and the initial QA pair are fed back into the MLLM to refine spatial localization. This integrated design eliminates the need for iterative OCR and reduces computational overhead, culminating in a final annotation module that delivers the final answer along with precise bounding box annotations. Numbered circles denote sequential processing steps (see Section 3 for more details).

financial document analysis (Huang et al., 2024).

Motivated by these challenges, we propose a novel, zero-shot (training-free) OCR-free pipeline that harnesses the inherent visual understanding of MLLMs to directly extract and localize answers from document images. Unlike conventional OCR-dependent methods ((Mohammadshirazi et al., 2024))—which often suffer from cascading errors and high computational complexity—our approach bypasses the need for iterative OCR by constructing a *single* image that comprises detected text regions with unique bounding box identifiers, thereby preserving essential spatial relationships while significantly reducing computational overhead.

The proposed design is driven by several key motivations: firstly, by consolidating text information into a single constructed image rather than sending all recognized text as prompts, typical in OCR-dependent methods, we reduce the token count, which is crucial for avoiding context window overflow (e.g., the 128k token limit for Pixtral) and ensuring that the MLLM can process the input effectively; secondly, the constructed image approach bypasses the iterative OCR processing required for each cropped image, thereby streamlining the pipeline and reducing computational overhead; and finally, instead of processing multiple separate cropped images—which may exceed the MLLM's input limitations—we combine them into a single constructed image, making the model more efficient and suitable for spatial reasoning. We demonstrate the effectiveness of our model by comparing it with an OCR-dependent baseline, and empirical evaluations confirm that our model not only attains state-of-the-art (SoTA) textual accuracy but also achieves robust spatial grounding, establishing its potential as a viable alternative to CoT or OCR-dependent solutions. Building on this foundation, our contributions are threefold:

1. **Zero-shot spatial grounding for MLLMs.** A training-free pipeline that equips off-the-shelf MLLMs with answer localization in document images, reducing complexity versus CoT or fine-tuning approaches.
2. **Constructed-image architecture.** A novel design that integrates text detection into a compact "constructed image," eliminating external OCR and preserving layout context for superior efficiency and accuracy.
3. **Unified ANLS + IoU evaluation.** A rigorous framework combining ANLS and IoU (Rezatofighi et al., 2019) to measure both textual and spatial accuracy, enhancing interpretability and mitigating hallucinations.

2. Related Work

Information extraction systems initially relied on statistical and topic-based classifiers 2020 before transitioning to layout- and vision-aware architectures. Recent multimodal document processing research spans four key areas. Text detection has advanced through differentiable binarization (DBNet (Liao et al., 2020)), irregular-shape handling (FAST (Chen et al., 2021)) and receptive-field fusion (MixNet (Zeng et al., 2023)). Recognition evolved from sequence-and-attention models (CRNN (Shi et al., 2016), SAR (Li et al., 2019), MASTER (Lu et al., 2021)) to transformer-based approaches (ViTSTR (Atienza, 2021), PARSeq (Bautista & Atienza, 2022), MaskOCR (Lyu et al.,

2022), TrOCR (Li et al., 2023), DTrOCR (Fujitake, 2024)). Information extraction leverages both OCR-free architectures (Donut (Kim et al., 2022), UDOP (Tang et al., 2023), OmniParser (Wan et al., 2024)) and OCR-dependent models that incorporate positional cues (ICL-D3IE (He et al., 2023), LATIN-Prompt (Wang et al., 2023), Cream (Kim et al., 2023), InstructDoc (Tanaka et al., 2024)). Layout-aware methods (LayoutLLM (Luo et al., 2024), DocLayLLM (Liao et al., 2024), LayTextLLM (Lu et al., 2024)) further integrate spatial structure, but typically rely on separate OCR steps or extra encoders, increasing complexity and inference time. In contrast, our unified MLLM merges text recognition and spatial reasoning in a single end-to-end model, improving both efficiency and localization precision.

3. DLaVA

This section describes our proposed DLaVA approach for zero-shot, OCR-free information extraction from documents, as illustrated in Figure 2. By harnessing the power of MLLM, our method directly extracts and localizes information from document images without relying on iterative OCR processing, thereby achieving robust structural accuracy while balancing computational efficiency with precise spatial grounding. The DLaVA approach is comprised of the following steps: **(1) Text Detection Module**: The original document image I is processed using a text detection model—DB-ResNet-50 (Liao et al., 2020), as shown in step 1 in Figure 2 as its real-time differentiable binarization method delivers superior boundary localization with high computational efficiency—a critical balance for structured data extraction in document images that is not as effectively achieved by FAST's emphasis on irregular text shapes or MixNet's complexity in handling intricate scenes. This model outputs bounding boxes for each text segment in the image. The detected bounding boxes are represented as:

$$BB = \{BB_1, BB_2, \ldots, BB_n\}$$

where each BB_i is a bounding box coordinate $[x_{i1}, y_{i1}, x_{i2}, y_{i2}]$, labeled as step 2 in Figure 2. Each bounding box BB_i is used to crop a segment of the image I, isolating individual words or phrases. The cropped image for BB_i is denoted by: $C_i = I[BB_i]$

(2) Constructed Image Creation: Instead of performing OCR on each cropped image, the bounding box images are arranged to form a "constructed image," illustrated in step 3 of Figure 2. Each bounding box BB_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 its corresponding bounding box ID:

$$I_C = \{(C_1, BB_1), (C_2, BB_2), \ldots, (C_n, B_n)\}$$

For example, if the document contains sentences like "THE STATE OF TEXAS...", after text detection, we obtain cropped images of individual words such as "THE" (C_1), "STATE" (C_2), "OF" (C_3), and "TEXAS" (C_4). In the constructed image I_C, each line would display the words with their bounding box IDs in sequence (e.g., the first line shows "THE (BB_1)", the second line "STATE (BB_2)", etc.).

(3) Information Extraction Model: In parallel, the MLLM - Pixtral-12B model (Agrawal et al., 2024) receives the input image I and the query Q (step 4) to generate the answer text A. The generated answers, together with their corresponding questions (Q+A), are passed as an input to the final MLLM.

(4) Final MLLM Processing: In the final step (step 5), the Pixtral-12B model utilizes the bounding box coordinates from step 2, the constructed image I_C from step 3, and the question-answer pair from step 4 to generate the answer's bounding box B_A and return it along with the answer A. Subsequently, post-processing scripts are applied to annotate the returned answer based on the coordinates of B_A.

Handling Cascading Errors: Our approach avoids cascading errors by eliminating the explicit text recognition (OCR) step entirely. In traditional OCR-based systems, any misrecognition of text in the initial OCR stage propagates through subsequent stages, leading to errors in answer extraction and localization. In contrast, our method leverages an MLLM to directly extract and localize information from document images. We first detect text regions and then create a "constructed image" that consolidates these regions along with their unique bounding box identifiers and corresponding coordinates. This unified representation is processed in one go—first to generate an initial answer and later to refine spatial localization—thereby bypassing the need for iterative OCR and preventing errors from accumulating. Furthermore, any inaccuracies introduced during the text detection phase are mitigated by the final MLLM (step 5), which leverages the overall contextual information to correct inconsistencies (Liu et al., 2024). This streamlined pipeline not only enhances accuracy but also improves computational efficiency and robustness in answer localization.

4. Experiments

4.1. Datasets and Experimental Setup

We evaluated our proposed model on several well-established, text-rich document datasets commonly used for VIE and Document VQA tasks. For VIE-related question answering, we utilized the **FUNSD** (Jaume et al., 2019), **CORD** (Park et al., 2019), and **SROIE** (Huang et al., 2019) datasets. In the domain of Document VQA, we assessed performance using the **DocVQA** (Mathew et al., 2021), **RICO** (Deka et al., 2017) datasets, and Scene Text+Evidence Visual Question Answering **(STE-VQA)** (Wang et al., 2020). All experiments run on a single NVIDIA A100 GPU (80 GB) for fair comparison.

Table 1. Comparison of DLaVA with SoTA models on benchmark datasets using ANLS evaluation metric

Model Category	Models	Document VQA			QA for VIE		
		DocVQA	STE-VQA	RICO	FUNSD	CORD	SROIE
Text	Llama2-7B-Chat (Touvron et al., 2023)	64.99	52.14	59.49	48.20	47.70	68.97
	Llama3-8B-Instruct (Dubey et al., 2024)	51.79	54.65	58.81	68.57	52.31	61.24
Text + BBox	LayTextLLM **(Llama2-7B)** (Lu et al., 2024)	72.83	-	-	78.65	70.81	83.27
Text + BBox + Image	LayoutLLM-7B $_{\text{CoT}}$ **(Llama2-7B)** (Luo et al., 2024)	74.25	-	-	78.65	62.21	70.97
	LayoutLLM-7B $_{\text{CoT}}$ **(Vicuna-1.5-7B)** (Luo et al., 2024)	74.27	-	-	79.98	63.10	72.12
	DocLayLLM **(Llama2-7B)** (Liao et al., 2024)	72.83	-	-	78.65	70.81	83.27
	DocLayLLM **(Llama3-7B)** (Liao et al., 2024)	78.40	-	-	84.12	71.34	84.36
Image	Phi4-14B (Abdin et al., 2024)	79.84	60.22	68.49	77.64	77.03	80.12
	Llama3.2-11B (Dubey et al., 2024)	78.4	48.14	53.47	65.02	42.96	61.42
	Pixtral-12B (Agrawal et al., 2024)	80.71	61.67	70.31	78.26	79.08	82.24
	LLaVA-NeXT-13B (Liu et al., 2023)	51.01	13.77	25.12	19.71	33.5	13.41
	LLaVA-OneVision-7B (Li et al., 2024)	47.59	22.39	19.54	22.82	32.43	12.10
	Qwen2.5-VL-7B (Bai et al., 2025)	68.54	61.41	56.42	58.44	39.01	56.37
	InternVL2-8B (Chen et al., 2024b)	71.26	59.74	44.81	57.58	55.88	81.55
Image + BBox	DLaVA **(Pixtral-12B)**	**85.91**	**66.96**	**76.34**	**87.57**	**82.08**	**91.42**

We report the ANLS (Yujian & Bo, 2007), and the mean Average Precision at IoU thresholds 0.50–0.95 (mAP@IoU[0.50:0.95]) (Rezatofighi et al., 2019) for evaluating our model against baselines. Hyperparameters and prompt formats are detailed in Appendices C and D.

4.2. Baseline Models

To evaluate the effectiveness of our proposed approach, we compare it against a diverse set of state-of-the-art baselines spanning both OCR-free and OCR-dependent paradigms. The OCR-free baselines include Phi4-14B (Abdin et al., 2024), PixTral-12B (Agrawal et al., 2024), InternVL v2-8B (Chen et al., 2023; 2024a), Qwen2.5-VL 7B (Bai et al., 2025), LLaVA-OneVision 7B (Li et al., 2024), LLaVA-NeXT-13B (Liu et al., 2023), and LLaMA 3.2-11B (Dubey et al., 2024). For OCR-dependent models, we include LLaMA 2-7B-Chat (Touvron et al., 2023), LLaMA 3-8B-Instruct (Dubey et al., 2024), LayoutLLM-7B (Luo et al., 2024), DocLayLLM (Liao et al., 2024), and LayTextLLM (Lu et al., 2024). These baselines allow for comprehensive comparison in both textual accuracy and spatial localization.

It is to be noted that STE-VQA and RICO entries are omitted for Text+BBox and Text+BBox+Image models due to lack of publicly available implementation and inference support for these datasets at the time of writing. See Appendix E.1 for a detailed justification of baseline selection and categorization.

5. Results and Discussion

We benchmark DLaVA against leading baselines on six datasets using ANLS (Table 1), and it's evident that DLaVA outperforms the strongest competing model by +11.6 pp on DocVQA and +7.1 pp on SROIE, with average gains of +8.5 pp across all six ANLS benchmarks. These results underscore how our zero-shot OCR-free, constructed-image pipeline elevates both multilingual text understanding and layout-aware reasoning.

Appendix E presents three ablations—(1) adding raw image input, (2) removing the information extraction module, and (3) using an OCR-dependent variant—demonstrating that both bounding-box fusion and the two-stage design each contribute 2–5 pp improvements in ANLS and IoU.

By collapsing multiple stages into a single zero-shot MLLM call, DLaVA greatly simplifies the system architecture, cutting inference latency and memory footprint, making it a streamlined, resource-efficient design that minimize compute, time, memory, bandwidth, and energy requirements.

6. Conclusion

In this paper, we introduce DLaVA, a unified, zero-shot OCR-free document VQA model built on a two-stage MLLM pipeline that removes the separate OCR step—reducing token overhead and error cascades—while directly injecting spatial context for precise localization. Our approach achieves SoTA ANLS and IoU on benchmarks like DocVQA and VIE, demonstrating both superior textual accuracy and robust bounding-box alignment. Crucially, explicit spatial annotations enhance interpretability and trustworthiness by enabling users to verify each answer against its source region. DLaVA's streamlined, resource-efficient design thus sets a new standard for reliable and transparent document understanding.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Agrawal, P., Antoniak, S., Hanna, E. B., Chaplot, D., Chudnovsky, J., Garg, S., Gervet, T., Ghosh, S., Héliou, A., Jacob, P., et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Atienza, R. Vision transformer for fast and efficient scene text recognition. In *International conference on document analysis and recognition*, pp. 319–334. Springer, 2021.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Bautista, D. and Atienza, R. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, pp. 178–196. Springer, 2022.

Chen, Z., Wang, J., Wang, W., Chen, G., Xie, E., Luo, P., and Lu, T. Fast: Faster arbitrarily-shaped text detector with minimalist kernel representation. *arXiv preprint arXiv:2111.02394*, 2021.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., and Dai, J. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., 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*, 2024a.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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*, pp. 24185–24198, 2024b.

Deka, B., Huang, Z., Franzen, C., Hibschman, J., Afergan, D., Li, Y., Nichols, J., and Kumar, R. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pp. 845–854, 2017.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Fujitake, M. Dtrocr: Decoder-only transformer for optical character recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8025–8035, 2024.

Guha Neogi, P. P. and Goswami, S. Force of gravity oriented classification technique in machine learning. In Sharma, N., Chakrabarti, A., Balas, V. E., and Martinovic, J. (eds.), *Data Management, Analytics and Innovation*, pp. 299–310, Singapore, 2021. Springer Singapore. ISBN 978-981-15-5616-6.

He, J., Wang, L., Hu, Y., Liu, N., Liu, H., Xu, X., and Shen, H. T. Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19485–19494, 2023.

Huang, Y., Lv, T., Cui, L., Lu, Y., and Wei, F. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4083–4091, 2022.

Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q., Li, Y., Gao, C., Huang, Y., Lyu, W., Zhang, Y., et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., and Jawahar, C. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520. IEEE, 2019.

Ishmam, M. F., Shovon, M. S. H., Mridha, M. F., and Dey, N. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, pp. 102270, 2024.

Jaume, G., Ekenel, H. K., and Thiran, J.-P. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pp. 1–6. IEEE, 2019.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pp. 498–517. Springer, 2022.

Kim, G., Lee, H., Kim, D., Jung, H., Park, S., Kim, Y., Yun, S., Kil, T., Lee, B., and Park, S. Visually-situated natural language understanding with contrastive reading model and frozen large language models. *arXiv preprint arXiv:2305.15080*, 2023.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., and Li, C. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Li, H., Wang, P., Shen, C., and Zhang, G. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8610–8617, 2019.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13094–13102, 2023.

Liao, M., Wan, Z., Yao, C., Chen, K., and Bai, X. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11474–11481, 2020.

Liao, W., Wang, J., Li, H., Wang, C., Huang, J., and Jin, L. Doclayllm: An efficient and effective multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*, 2024.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023.

Liu, Y., Yang, B., Liu, Q., Li, Z., Ma, Z., Zhang, S., and Bai, X. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.

Lu, J., Yu, H., Wang, Y., Ye, Y., Tang, J., Yang, Z., Wu, B., Liu, Q., Feng, H., Wang, H., 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.

Lu, N., Yu, W., Qi, X., Chen, Y., Gong, P., Xiao, R., and Bai, X. Master: Multi-aspect non-local network for scene text recognition. *Pattern Recognition*, 117:107980, 2021.

Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., and Yao, C. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15630–15640, 2024.

Lyu, P., Zhang, C., Liu, S., Qiao, M., Xu, Y., Wu, L., Yao, K., Han, J., Ding, E., and Wang, J. Maskocr: Text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311*, 2022.

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

Mohammadshirazi, A., Firoozsalari, A. N., Zhou, M., Kulshrestha, D., and Ramnath, R. Docparsenet: Advanced semantic segmentation and ocr embeddings for efficient scanned document annotation. *arXiv preprint arXiv:2406.17591*, 2024.

Neogi, P. P. G., Das, A. K., Goswami, S., and Mustafi, J. Topic modeling for text classification. In Mandal, J. K. and Bhattacharya, D. (eds.), *Emerging Technology in Modelling and Graphics*, pp. 395–407, Singapore, 2020. Springer Singapore. ISBN 978-981-13-7403-6.

Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., and Lee, H. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. 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*, pp. 658–666, 2019.

Shi, B., Bai, X., and Yao, C. 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.

Tanaka, R., Iki, T., Nishida, K., Saito, K., and Suzuki, J. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19071–19079, 2024.

Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C., and Bansal, M. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19254–19264, 2023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wan, J., Song, S., Yu, W., Liu, Y., Cheng, W., Huang, F., Bai, X., Yao, C., and Yang, Z. 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*, pp. 15641–15653, 2024.

Wang, W., Li, Y., Ou, Y., and Zhang, Y. Layout and task aware instruction prompt for zero-shot document image question answering. *arXiv preprint arXiv:2306.00526*, 2023.

Wang, X., Liu, Y., Shen, C., Ng, C. C., Luo, C., Jin, L., Chan, C. S., Hengel, A. v. d., and Wang, L. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10126–10135, 2020.

Yujian, L. and Bo, L. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

Zeng, Y.-X., Hsieh, J.-W., Li, X., and Chang, M.-C. Mixnet: toward accurate detection of challenging scene text in the wild. *arXiv preprint arXiv:2308.12817*, 2023.

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.

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 C. 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 location 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.

C. Hyperparameter Details

We set the hyperparameters for each component in our framework to achieve an optimal balance between model efficiency and accuracy. After rigorous experiments with various hyperparameter ranges, we determined the following combinations to be optimal for our model. The configurations for each module are as follows:

- **Pixtral-12B Model Hyperparameters:**
  - We set `max_tokens` to 128k to avoid truncation for large multi-modal prompts; this parameter can be adjusted within the range of 8k to 128k.
  - The `temperature` is fixed at 0.1, which lies within the permissible range of 0.0 to 1.0.
  - We use a `top-p` value of 1.0 to enforce greedy selection under these constraints, with the value allowed to vary between 0.0 and 1.0.
- **DB Resnet-50:**
  - The binarization threshold (`bin_thresh`) is set to 0.3, which is within the acceptable range of 0.1 to 0.9.
  - The box threshold is fixed at 0.1, and it may vary between 0.1 and 0.9.
- **PARSeq:**
  - The maximum sequence length for positional embeddings (`max_length`) is set to 32, and it can be adjusted between 16 and 256.
  - All other hyperparameters for PARSeq remain at their default values.
- **Hyperparameter Optimization:** We employed Optuna for hyperparameter optimization, and the final values were selected based on the best performance on the validation set to ensure a robust balance between computational efficiency and model accuracy.

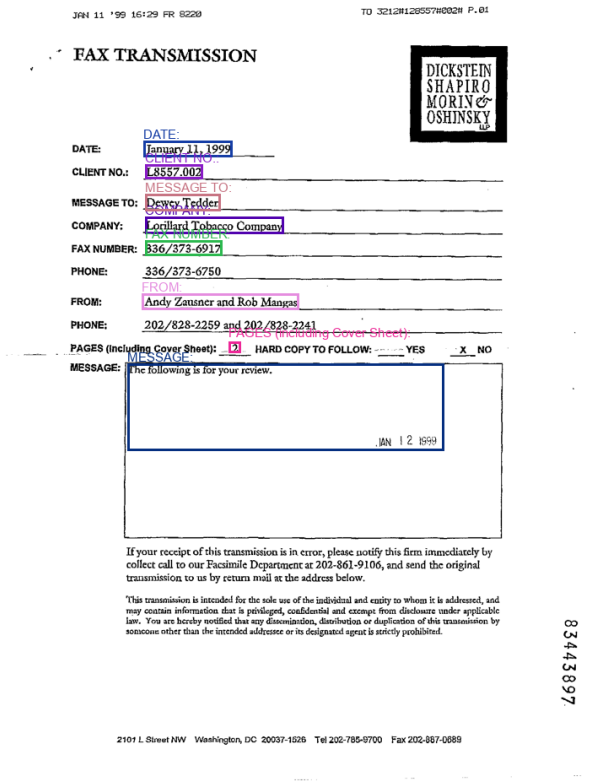
D. Model Prompting Details

The model is prompted using the following inputs:

1. A set of questions.
2. The original input image where the answers to the questions are located.
3. A JSON file containing the Bounding Box IDs (e.g., `BB0`, `BB1`, etc.) along with their corresponding bounding box coordinates for each word in the original input image.
4. A second image displaying all words from the original input image along with their associated Bounding Box IDs.

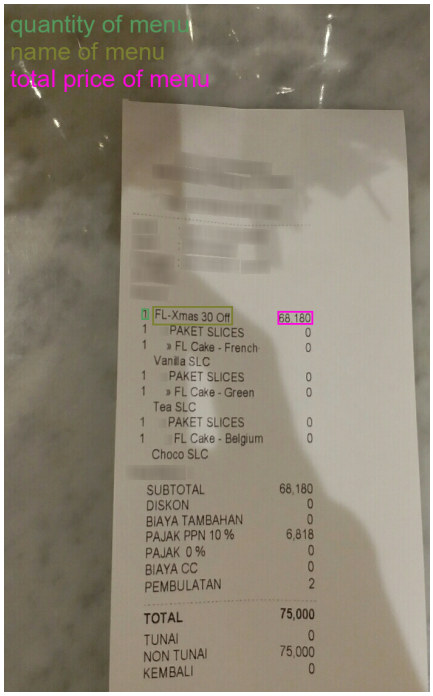
The prompt provided to the model is structured as follows:

```
Questions:
{user_queries}

Bounding Box IDs and Bounding Box
    Coordinates for each word:
{bounding_boxes}

When finding answers to the questions,
    you are STRICTLY allowed to answer
    only using words present in the
    image. So, just return the words
    from the image (AND no description
    of full sentences).
Just match the words that answer the
    question.

Your task is to find the answer to
    these questions from the 1st image,
     and identify the Bounding Box
    Coordinates for each answer.
Return a JSON in the format specified
    below. (NO Additional Information.
    JUST JSON in the following format)

Final Answer: <answer>

where <answer> strictly adheres to the
    following structure:
- <answer> should be in JSON format.
- Each question from the question-
    answer pairs will be a key.
- For each question:
    - "value": The answer text (
        containing only words found in
        the input image; avoid point-
        wise or list-style answers).
    - "bounding_box": [[0.3037,
        0.4863], [0.3257, 0.502]] (The
```



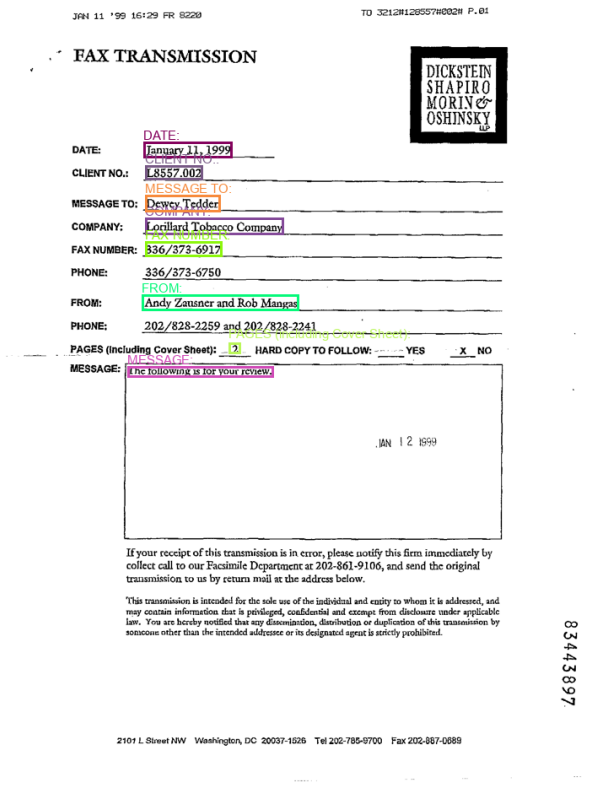
(a) Document Example from FUNSD Dataset

(b) Receipt from CORD Dataset



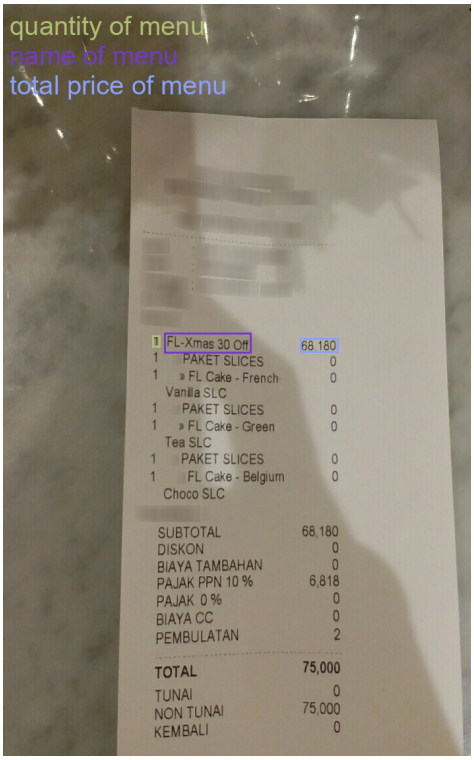
(c) Another receipt from the CORD Dataset

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



(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

Table 2. Selecting best model for Text Recognition based on ANLS (for Ablation 3)

Models	DocVQA	STE-VQA	RICO	FUNSD	CORD	SROIE
PARSeq (Bautista & Atienza, 2022)	**68.22**	58.89	**65.91**	**76.23**	77.21	84.90
MaskOCR (Lyu et al., 2022)	66.83	55.18	59.99	75.42	**77.65**	83.38
TrOCR (Li et al., 2023)	64.86	59.11	63.43	75.01	76.59	81.92
DTrOCR (Fujitake, 2024)	67.93	**60.08**	63.77	76.11	77.19	**85.33**

```
    bounding box coordinates in
    this exact structure).
  (Ensure only numerical digits, no
      NULL or empty values, and
    each coordinate is separated
    by commas).

If the answer consists of multiple
    words:
- Use the following format for "
    bounding_box":
    "value": "1 BLACK SAKURA"
    "bounding_box": [
        [[0.09716796875, 0.458984375],
            [0.22314453125,
            0.4921875]],
        [[0.23486328125, 0.462890625],
            [0.37548828125,
            0.4873046875]],
        [[0.38720703125, 0.4619140625],
             [0.5556640625,
            0.4873046875]]
    ]

Additional Instructions:
- Ensure correct pairing and matching
    of brackets (i.e., (), \{\}, []).
- Each "bounding_box" must contain
    exactly four numerical values
    formatted as two sets of
    coordinates within square brackets.
```

E. Ablation Study

We conduct the following ablation experiments to assess the contributions of different components in our OCR-Free pipeline (Ablation 1 and 2) and also compare it with an OCR-dependent approach (Ablation 3):

- **Ablation 1 - Additional Image Input:** In this experiment, we provide the original input image I as an extra input to the final MLLM model (step 5 in Figure 2) along with the other input components. This helps us evaluate the impact of the full visual context on the model's performance in extracting and localizing answers.

- **Ablation 2 - Removal of Information Extraction:** Here, we remove the information extraction step (step 4) entirely, relying solely on the final MLLM (step 5) for both question-answering and generating the corresponding bounding boxes. This experiment isolates the contribution of the dedicated information extraction module and demonstrates its role in refining spatial localization and answer accuracy.

- **Ablation 3 - OCR-Dependent Approach:** For comparison, we consider an OCR-dependent model that incorporates a text recognition module (PARSeq (Bautista & Atienza, 2022)) to convert cropped images into text. Table 2 compares the text recognition accuracy of several cutting-edge OCR models (PARSeq, MaskOCR, TrOCR, and DTrOCR) across multiple benchmark datasets, and we observe that PARSeq achieves overall higher accuracy, making it the preferred module for our experiments. In this approach, a text detection model (DB-ResNet-50) is first used to obtain the detected cropped images (step 3) along with their corresponding bounding box coordinates (step 2). The cropped images are then passed to the text recognition module (step 3*) to generate textual representations, and the outputs from the text recognition step—together with the bounding box information and the input question—are fed into the final MLLM (step 5) to generate the answer A and its bounding box BB_A. Figure 5 shows the architecture of the OCR-dependent model. This ablation serves as a baseline to highlight the benefits of our unified OCR-free approach over traditional methods that rely on separate text recognition.

Table 3. Ablation Study Results: Comparison of DLaVA and its ablation variants using ANLS metric on Doc. VQA & QA for VIE

Models	DocVQA	EST-VQA	RICO	FUNSD	CORD	SROIE
DLaVA	**85.91**	**66.96**	**76.34**	**87.57**	**84.41**	**91.42**
Ablation 1	83.55	64.01	69.41	83.28	79.08	85.36
Ablation 2	82.26	62.51	73.86	84.35	81.91	86.02
Ablation 3	74.02	62.70	71.99	79.57	82.08	90.45

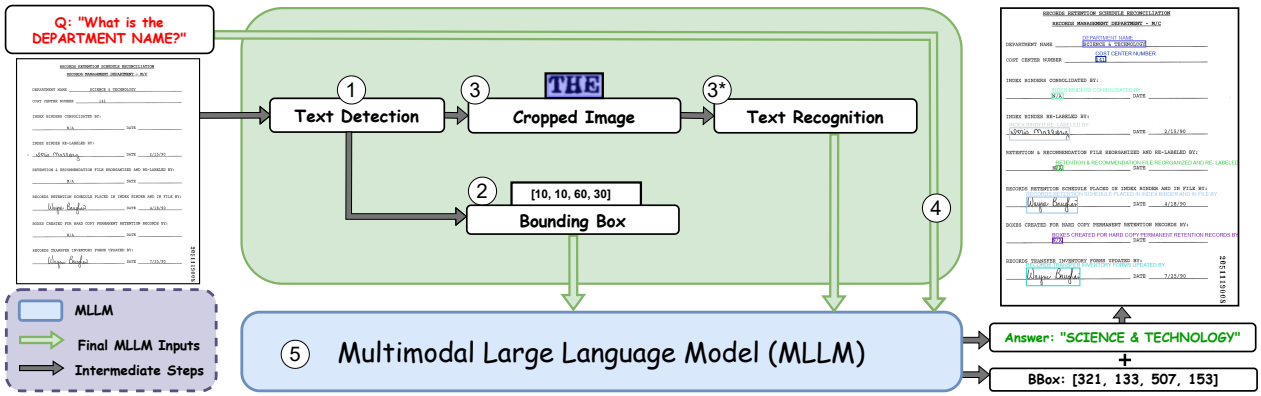


Figure 5. Architecture of OCR-Dependent Model (Ablation 3)

E.1. Ablation Study Results

We evaluate our proposed DLaVA model alongside three ablation variants (Ablation 1, Ablation 2, and Ablation 3 as described in the previous Section) on Document VQA and VIE tasks using the ANLS metric (see Table 3). DLaVA achieves the highest ANLS scores across all evaluated datasets, demonstrating its ability to accurately extract and interpret text information without reliance on iterative OCR. In contrast, Ablation 1 and Ablation 2—where either the original image was added as input or the information extraction step is removed—show reduced ANLS performance, underscoring the importance of both components for boosting textual accuracy and overall model effectiveness. Ablation 3, which incorporates an OCR-dependent process, also exhibits lower ANLS scores, indicating that our zero-shot OCR-free design is more robust to potential errors introduced by text recognition.

Table 4. Ablation Study Results: Comparison of DLaVA and its ablation variants using IoU (mAP@IOU[0.50:0.95]) metric on Document VQA and QA for VIE

Models	DocVQA	EST-VQA	RICO	FUNSD	CORD
DLaVA	**46.22**	**33.65**	**38.13**	**45.52**	**57.86**
Ablation 1	44.01	28.08	29.88	32.71	45.45
Ablation 2	39.41	30.49	33.56	37.12	46.69
Ablation 3	34.93	31.37	32.66	31.98	48.01

Table 4 reports the IoU scores for the same set of ablation experiments, focusing on bounding box localization. DLaVA again outperforms all ablation variants, reflecting its stronger spatial grounding capabilities. In particular, removing the dedicated information extraction step or excluding the original image input leads to noticeably lower IoU scores, highlighting how these design choices facilitate more precise bounding box predictions. Meanwhile, Ablation 3's reliance on an external OCR stage can introduce cascading localization errors, resulting in lower IoU.

Taken together, the ANLS and IoU metrics offer a holistic perspective—capturing both answer quality and localization precision. Compared to rule-driven or handcrafted classifiers (Guha Neogi & Goswami, 2021), DLaVA's unified MLLM pipeline offers better generalization with less tuning, as it eliminates the need for task-specific components like custom matchers or segmenters. This makes DLaVA especially effective in diverse, zero-shot document VQA settings.

Appendix F: Justification for Baseline Selection

To ensure a fair and comprehensive evaluation of DLaVA, we carefully selected baselines that span the full spectrum of document VQA paradigms:

1. OCR-Free Multimodal Baselines: We include state-of-the-art vision-language models such as *Phi4-14B*, *PixTral-12B*, *InternVL2-8B*, *Qwen2.5-VL-7B*, *LLaVA-OneVision*, and *LLaVA-NeXT*. These models directly process images and questions without relying on OCR, allowing us to benchmark against the latest zero-shot visual language models. These models also highlight limitations in spatial reasoning when explicit grounding is not enforced.

2. OCR-Dependent LLM Baselines: We benchmark against OCR-enhanced models like *LayoutLLM*, *DocLayLLM*, *LayTextLLM*, *LLaMA2-7B-Chat*, and *LLaMA3-8B-Instruct*, which leverage textual prompts or positional information extracted through OCR. These represent strong traditional baselines in document understanding that rely heavily on pre-extracted text and bounding boxes, showcasing the trade-off between interpretability and pipeline complexity.

3. Layout-Aware LLMs: To specifically test layout reasoning, we include models that interleave text and spatial signals, such as *DocLayLLM* and *LayoutLLM*. These provide insight into how structured layout-aware processing compares to our constructed image approach in both accuracy and spatial grounding.

This comprehensive mix ensures our evaluation spans across (i) visual-only, (ii) OCR-based, and (iii) layout-enhanced models—allowing us to isolate the benefits of DLaVA's zero-shot, OCR-free, and spatially grounded design.