Rethinking Reading Order: Toward Generalizable Document Understanding with LLM-based Relation Modeling

Weishi Wang*, **Hengchang Hu, Daniel Dahlmeier**
SAP
{weishi.wang,hengchang.hu,d.dahlmeier}@sap.com

Abstract

Document understanding requires modeling both structural and semantic relationships between the layout elements within the document, with human-perceived reading order (RO) playing a crucial yet often neglected role compared to heuristic OCR sequences used by most existing models. Previous approaches depend on costly, inconsistent human annotations, limiting scalability and generalization. To bridge the gap, we propose a cost-effective paradigm that leverages large language models (LLMs) to infer global RO and inter-element layout relations without human supervision. By explicitly incorporating RO as structural guidance, our method captures hierarchical, document-level dependencies beyond local adjacency. Experiments on Semantic Entity Recognition, Entity Linking, and Document Question Answering show consistent improvements over baseline methods. Notably, LLM-inferred RO, even when differing from ground-truth adjacency, provides richer global structural priors and yields superior downstream performance. These results and findings demonstrate the scalability and significance of RO-aware modeling, advancing both LLMs and lightweight layout-aware models for robust document understanding. Code, data, and more details will be made publicly available after corporate review, in accordance with SAP's corporate open-source policy.

1 Introduction

Understanding document layouts is essential for modern document intelligence systems, powering applications such as *Semantic Entity Recognition* (Perot et al., 2024; Lamott et al., 2024; Mao et al., 2024), *Entity Linking* (Fan et al., 2024; Huang et al., 2021; Yao et al., 2019; Zhou et al., 2021), and *Document Question Answering* (Wang et al., 2023b; Mathew et al., 2021). While recent efforts have been dedicated to incorporating document layout structures, including explicitly encoding spatial information through bounding box coordinates (Lu et al., 2024; Perot et al., 2024; Kim et al., 2023) and implicitly modeling positions of layout elements (Wang et al., 2024; Luo et al., 2024; Liao et al., 2024), most models still rely on OCR-based token sequences (Gu et al., 2022). These sequences often reflect engine-specific or heuristic ordering rather than the natural reading flow perceived by humans.

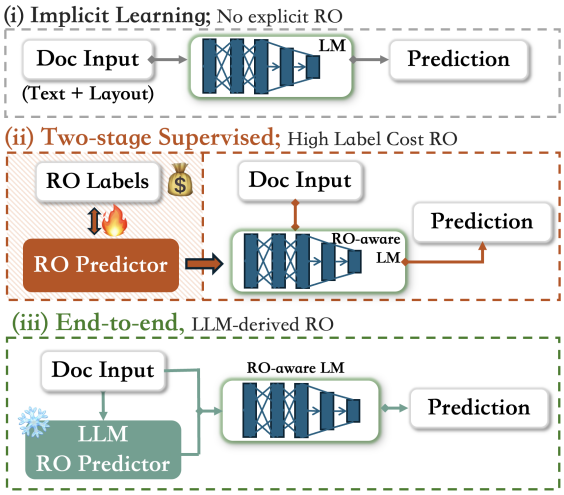


Figure 1: Comparison of three paradigms for modeling reading order. Our end-to-end, unsupervised approach (iii) eliminates costly human supervision while delivering stronger, structure-aware performance.

Despite advances in layout-aware modeling, a key aspect of human document comprehension remains underexplored: the *reading order* (RO). RO reflects the natural sequence in which textual elements are processed, serving as a crucial bridge between spatial layout and semantic understanding. By encoding document flow and structural hierarchy, RO plays a vital role in capturing inter-element semantic relationships (Zhang et al., 2023; Wang et al., 2025). However, existing methods typically

*Corresponding author: weishi.wang@sap.com

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4110–4130

treat the OCR-based token sequence as fixed input and rarely reason about global RO relations, limiting their ability to model document-level dependencies that humans naturally perceive.

Recent efforts have attempted to explicitly model RO by introducing human-annotated supervision (Zhang et al., 2024a). As shown in Figure 1 (ii), these annotations are commonly used to construct or prune document RO relations for developing downstream RO-aware language models. While this two-stage paradigm has demonstrated strong in-domain performance, it suffers from practical limitations. Human annotations are costly and time-consuming to produce, often exhibit inconsistency, and fail to generalize across diverse layouts and tasks (Sylolypavan et al., 2023; Wu et al., 2024). Our analysis in Section 5 further reveals that human-annotated RO relations are often biased toward local adjacency, neglecting broader document-level semantics. These limitations highlight the need for a more scalable and generalizable solution to RO modeling.

To address these limitations, we propose a novel paradigm that integrates RO as a natural structural prior directly into layout-aware document models (Figure 1, iii). Rather than relying on manual annotations or positional heuristics, we leverage large language models (LLMs) to infer global RO and inter-element layout relations. These inferred RO signals are incorporated into strong document understanding backbones, such as LayoutLMv3 (Huang et al., 2022) and GeoLayoutLM (Luo et al., 2023), via a lightweight RO-aware attention—requiring no architectural modifications. This approach enables a model-agnostic, scalable, and cost-effective solution: inferring RO relations for 6,000 pages requires only 0.7 GPU-hours (replacing 25 hours of human annotation) and adds less than 0.5 ms per instance after caching, with under 1% parameter overhead.

Comprehensive evaluation across Semantic Entity Recognition (SER), Entity Linking (EL), and generative document Question Answering (QA) tasks demonstrates the effectiveness of our approach. LLM-inferred RO relations significantly enhance EL performance and yield consistent gains on SER tasks across diverse types of documents. Notably, our method outperforms models trained on ground-truth RO signals, which are constrained by adjacency-based annotations. Although LLM-inferred RO aligns less closely with these ground-truth labels, it captures richer, document-level structural relationships that lead to superior downstream results. This counterintuitive finding exposes a key limitation of current evaluation metrics: optimizing for local adjacency does not necessarily improve document understanding. Our results underscore the importance of modeling global dependencies for robust and generalizable document analysis. For generative QA tasks, integration of RO relations into prompts enables GPT-4o to achieve improved results on DocVQA and InfoVQA benchmarks, further highlighting the value of global RO-awareness. Our main contributions are:

- We identify fundamental limitations in existing document structure modeling approaches and introduce reading order as an explicit signal to better capture document organization.
- We leverage LLMs to infer global reading order without supervision, developing a simple yet effective RO-aware attention module that integrates this information into both LLM-based or lightweight layout-aware models.
- We conduct extensive experiments across diverse document understanding tasks, demonstrating significant improvements over baselines for both LLMs and lightweight models.

2 Related Work

Layout-aware Document Understanding has increasingly focused on incorporating layout information to capture the spatial structure. Early methods (Xu et al., 2021; Katti et al., 2018) embed 2D positional coordinates alongside text to model spatial relationships. More recent methods (Huang et al., 2022; Yu et al., 2023) further enable OCR-free document understanding (Tang et al., 2023) by incorporating visual features using masked language modeling across image and text. To model inter-entity structure, relation-aware methods construct layout graphs with tokens or fields as nodes. BROs (Hong et al., 2022) encodes pairwise box relations with relation heads, while DHFormer (Xing et al., 2024) introduces hierarchical decoding to fuse local and global layout dependencies. Zhang et al. (2023) extends this by using supervised reading order graphs as generation priors. However, these methods rely on manually labeled layout relations, which can be expensive and domain-specific, thereby limiting scalability.

LLM-based Document Understanding has increasingly been framed as a generative or

instruction-following problem. DocLLM (Wang et al., 2024), UniDoc (Feng et al., 2023), and LMDX (Perot et al., 2024) demonstrate that LLMs can extract entities, classify documents, and answer questions via prompt-based supervision. These approaches typically serialize documents as flat text using coordinate-sorted reading order (Wang et al., 2023b), often losing spatial nuances. To remedy this, layout-aware LLM prompting techniques have emerged, including CoT-style reasoning with spatial indicators (Lamott et al., 2024; Liao et al., 2024), and box-token augmentation (Lu et al., 2024). Nonetheless, these prompt-only methods suffer from long sequence lengths and require significant prompt engineering. Unified models like InstructDoc (Tanaka et al., 2024) and LayTextLLM (Lu et al., 2024) embed spatial-textual signals into hidden states, bridging layout and semantics more effectively. In this work, we exploit LLMs for inferring relational RO priors and seamlessly integrate them into diverse architectures, enabling comprehensive document comprehension.

3 Approach

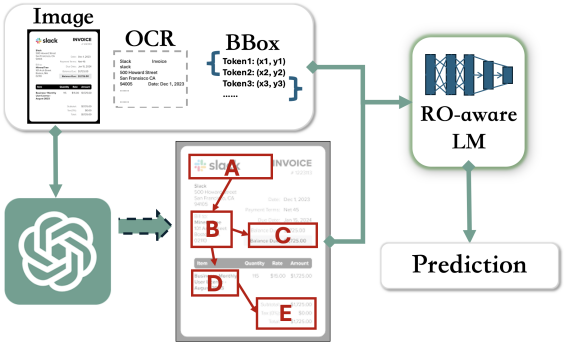


Figure 2: Illustration of our end-to-end proposed paradigm. Given a document image, OCR text, and their bounding boxes, it then uses an LLM to infer global RO relations. The inferred RO is integrated into a layout-aware model to enhance document understanding.

We propose a unified framework that explicitly models document reading order as a structural prior to enhance general document understanding. As illustrated in Figure 2, our approach first employs an LLM-based reading order prediction (ROP) model to extract global-aware RO relations. Then, an RO-aware model integrates these RO relations for enhanced document comprehension, supporting both LLM-based and lightweight backbone models.

We detail our approach in the following section, beginning with the formal ROP task definition in Section 3.1, followed by our ROP models in Section 3.2, and discussing our RO-aware document understanding framework in Section 3.3.

3.1 Task Formulation

Let $\mathbb{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$ be a set of documents that includes the modalities of image and Optical Character Recognition (OCR). For each document $\mathcal{D}_i$, it associates with a collection of layout elements $\mathcal{D}_i = \{(d_j, bbox_j)\}_{j=1}^{|\mathcal{D}_i|}$, where d_j denotes the j-th layout element and $bbox_j$ represents its corresponding bounding box. Specifically, $bbox_j = (x_j^0, y_j^0, x_j^1, y_j^1)$, where the coordinates (x_j^0, y_j^0) and (x_j^1, y_j^1) indicate the bottom-left and top-right vertices of d_j's bounding box, respectively. And the j-th layout element $d_j = \{ocr_j, p_j\}$, where ocr_j indicates the textual information extracted by an OCR engine and $p_j = (d_j, d_k)$ denotes the direct relationship link from d_j to d_k $\forall\ j \leq k \leq |\mathcal{D}_i|$. Formally, layout modeling is defined as a document reading order prediction to recognize all the reading order relation pairs within the document.

3.2 Reading Order Prediction (ROP)

Baseline ROP Models. We adapt a neural global pointer (Su et al., 2022) as a baseline method, which is widely used in relation extraction tasks in NLP (Wang et al., 2022). For a document $\mathcal{D}$ with N layout elements $\mathcal{D} = \{(d_i, bbox_i)\}_{i=1}^{N}$, let $\mathcal{L}$ be the ground truth label of reading order relation, where $(d_i, d_j) \in \mathcal{L}$ indicates that the j-th layout element is an immediate successor of the i-th layout element. The tokenized OCR within layout elements $\mathbb{T}_\mathcal{D} = (t_1^1, \ldots, t_1^{n_1}, \ldots, t_N^1, \ldots, t_N^{n_N})$ and corresponding bounding box $\mathbb{B}_\mathcal{D} = (bbox_1^1, \ldots, bbox_1^{n_1}, \ldots, bbox_N^1, \ldots, bbox_N^{n_N})$ are then fed to a transformer-based (Vaswani et al., 2017) encoder f_θ to obtain the layout-aware text embeddings with average pooling:

$$\begin{aligned} h_1^1, \ldots, h_N^{n_N} &= f_\theta(\mathbb{T}_\mathcal{D}; \mathbb{B}_\mathcal{D}) \\ h_i &= \text{AveragePool}(h_i^1, \ldots, h_i^{n_i}) \end{aligned} \tag{1}$$

where h_i is the representation of the i-th layout element. Then the layout embedding h_i are fed into the global pointer network for relation extraction (Zhang et al., 2024a):

$$\begin{aligned} q_i &= \mathbf{W}_q h_i + \mathbf{b}_q \\ k_j &= \mathbf{W}_k h_j + \mathbf{b}_k \\ s_{ij} &= q_i^T k_j \end{aligned} \tag{2}$$

where $\mathbf{W}_{q,k}$ and $\mathbf{b}_{q,k}$ are learnable parameters; s_{ij} is the predicted score of $(i, j) \in \mathcal{L}$. Model weights

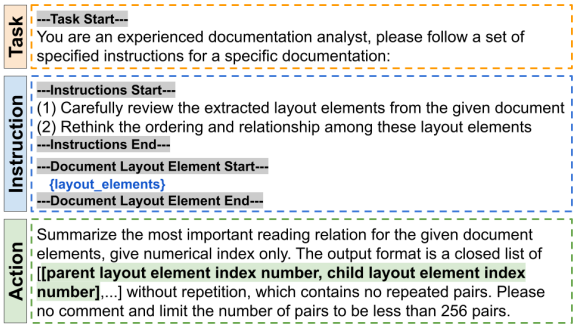


Figure 3: Example of our text-only prompt that is fed into LLMs to generate reading relations of a given document, denoted as *LLM-text* prompt. While the document image is attached to further improve the contextual representation, namely *LLM-multimodal* prompt.

are optimized by a class-imbalance loss (Su et al., 2022). During inference, the model identifies reading order relation pairs by filtering the predictions with a threshold:

$$\hat{\mathcal{L}} = \{(i, j) | s_{ij} > 0\} \tag{3}$$

where $\hat{\mathcal{L}}$ is the set of predicted reading order relation pairs. The baseline approach demonstrates its efficacy in recognizing the reading order of in-domain documents. However, documents always comprise rich and diverse reading patterns, which significantly impair the performance of the ROP models. Furthermore, existing ROP models overrely on human-annotated features of reading orders rather than identifying semantic correlations between layout elements (Wang et al., 2023a; Xing et al., 2024), leading to suboptimal generalization capabilities to out-of-domain examples.

LLM-based ROP Model. Since recent advanced LLMs are often pre-trained with multimodal knowledge, and fine-tuned to follow natural instructions, these models can understand unseen document layouts (Hu et al., 2024; Hegde et al., 2023). Therefore, to bridge the limitations of conventional ROP methods, we propose to condition an LLM g_θ to extract a set of RO relation pairs $\mathcal{L}_{g_\theta}$ on the given document $\mathcal{D}$:

$$\mathcal{L}_{g_\theta} = g_\theta(\mathcal{D}, \mathcal{P}) \tag{4}$$

where $\mathcal{P}$ is the instruction prompt. As illustrated in Figure 3, we first index the layout element within the document. Each prompt specifies the textual content of the layout elements with their corresponding boxes. For multimodal LLM, we explicitly introduce the visual information from the document image. Compared to conventional ROP models that depend on annotated RO labels, our method leverages LLMs' instruction-following ability to infer RO relations conditioned on layout content and spatial context. This enables zero-shot generalization to unseen layouts while reducing dependence on handcrafted features.

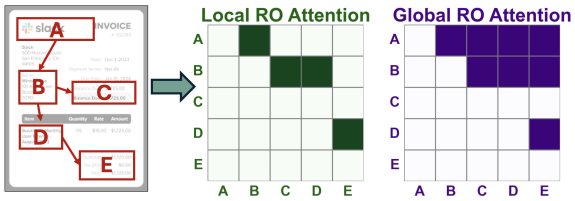


Figure 4: Illustration of RO matrices. The local RO matrix (green) captures direct succession between adjacent layout elements, while the global RO matrix (purple) models document-level RO awareness. We integrate both into RO-aware attention to enhance lightweight models for document understanding (see Section 3.3).

3.3 RO-aware Document Understanding

While conventional ROP models excel at capturing local, adjacent layout relationships, they often overlook semantic connections that span across distant document layout elements. These global-aware RO relationships, such as the logical connections between questions and their corresponding answers across multiple document regions, or the semantic alignment between headers and their associated content blocks distributed throughout the document, are fundamental to comprehensive document understanding yet remain largely unaddressed by current approaches. Our detailed case studies presented in Section A.4 highlight that an exclusive focus on local RO relations often leads to suboptimal document comprehension capabilities.

To address these limitations, we propose an end-to-end RO-aware framework that integrates our LLM-based ROP model to generate enhanced RO relations, enabling the capture of both local spatial layout patterns and broader document-wide semantic relations, thereby providing a robust foundation for accurate document comprehension.

RO-aware Attention for Lightweight Models. As illustrated in Figure 4, we represent local and global RO signals as binary matrices α and β, respectively. We introduce a lightweight RO-aware attention mechanism that incorporates these matrices directly into the attention computation. Specifi-

cally, we modify the attention weights as follows:

$$attn = softmax\left(\frac{\left(\mathbf{Q}\mathbf{K}^T + \lambda\alpha + (1-\lambda)\beta\right)}{\sqrt{d_k}}\right)\mathbf{V} \tag{5}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are the query, key, and value vectors, respectively; d_k is the dimension of the query and key vectors; λ is a learnable scalar as the weight of RO relation matrix α.

RO-aware Instruction for LLM. To enable LLMs to effectively utilize document structure, we explicitly inject RO relations into instruction prompts. Concretely, our method begins by processing a set of indexed layout elements within the document through our LLM-based ROP model, which generates directional reading links $\mathcal{L}_{g_\theta}$ between layout elements. These RO relations are then serialized into the prompt as concise natural language tuples (e.g., "Element 3 is followed by Element 7"), and strategically positioned before the main task instruction in the prompt. For multimodal LLMs, we maintain spatial context by incorporating both the document image and precise bounding box coordinates.

Extensive experimental results demonstrate that our paradigm improves downstream document understanding tasks by a huge margin, showing remarkable performance gains as well as robustness in both lightweight and LLM-based RO-aware document understanding.

4 Experimental Setup

4.1 Dataset

In this paper, we mainly focus on document understanding tasks, including Semantic Entity Recognition (SER), Entity Linking (EL), and generative document Question Answering (QA). We choose ROOR (Zhang et al., 2024a) to train a baseline ROP model, containing 10,967 expert-level annotated RO linking pairs. For SER, we choose ROOR, EC-FUNSD (Zhang et al., 2024b) and CORD (Park et al., 2019) for evaluation, which are widely used SER benchmarks. For EL task, we have employed ROOR and EC-FUNSD for evaluation. For QA task, we leverage the test sets of DocVQA (Mathew et al., 2021) and InfoVQA (Mathew et al., 2022) for evaluation, which are two challenging and popular document QA benchmarks. The details of all datasets are shown in Section A.1.

4.2 Baselines and Implementation Details

We evaluate our method on two categories with different model sizes and capabilities. For *foundational LLMs*, as discussed in Section 3.2, we leverage open-weights and closed-sources proprietary LLMs, including Gemma-2-9B-Instruct and Gemma-2-27B-Instruct (Rivière et al., 2024), Mistral-Nemo-Instruct[1], Mistral-Large-Instruct[2], Gemini-1.0 (Anil et al., 2023), Gemini-1.5: Google GCP VertexAI gemini-1.5-pro (v1) (Anil et al., 2023), Claude-3.5-Sonnet: AWS Bedrock anthropic–claude-3.5-sonnet (v1)[3], GPT-4o: OpenAI gpt-4o-2024-05-13 (Hurst et al., 2024) and GPT-4V (Zhao et al., 2024), where they are prompted to extract RO information from a given document. Specifically, we use GPT-4o for evaluating QA tasks. To ensure a fair comparison, all LLMs were configured with identical decoding parameters, including a maximum token limit of 4,096. We maintained deterministic outputs by setting the temperature to 0 and utilizing a consistent seed across both the LLMs and the lightweight models.

For *lightweight RO-aware Models*, as discussed in Section 3.3, we consider LayoutLMv3 (Huang et al., 2022) and GeoLayoutLM (Luo et al., 2023) for SER and EL tasks via direct fine-tuning. The implementation details can be found in Section A.2.

4.3 Evaluation Metrics

To evaluate the performance of various models for the task of SER, ROP, and EL, we apply F1 score as our primary evaluation metric, following the methodology of ROOR (Zhang et al., 2024a). Following the common practice, we use the Average Normalized Levenshtein Similarity (ANLS) and Exact Match (EM) accuracy for QA tasks (Biten et al., 2019; Mathew et al., 2021). ANLS is a looser metric to evaluate the closeness of a generated answer to the expected answer, while EM is a stricter metric requiring the prediction to be identical to the ground truth.

5 Experimental Results and Analysis

5.1 Results on Information Extraction

We first evaluate the accuracy of various ROP models in capturing reading order structures. As shown in Table 1, the baseline LayoutLMv3 shows

[1]https://mistral.ai/news/mistral-nem
[2]https://mistral.ai/news/mistral-large
[3]https://www.anthropic.com/news/claude-3-5-sonnet

Method	Word-level	Segment-level
LR (Wang et al., 2021)	-	9.44
TPP (Zhang et al., 2023)	-	42.96
LayoutLMv3-base†	88.77	67.97
LayoutLMv3-large†	94.14	80.49
GPT-4o-text	80.85	30.25
GPT-4o-multimodal	79.23	27.47
Human (Zhang et al., 2024a)	-	99.28

Table 1: F1 results on RO relation prediction. Human performance indicates the annotation consistency between two annotators. † denotes our reproduced results.

Method	Avg. RO-D	# Avg. RO-L	
	Document	Document	Element
GT	11.84	55.11	1.09
ROP-prediction (RORE)	11.90	51.51	1.11
GPT-4o-text	14.66	42.55	6.14
GPT-4o-multimodal	**15.14**	**45.29**	**4.45**
Gemini-1.5-pro-multimodal	13.88	35.75	5.65
Calude-3.5-sonnet-multimodal	13.90	36.07	2.60

Table 2: Statistics of RO distances (RO-D) and link density (RO-L). The average distance between connected layout elements (Avg. RO-D) was calculated at the document level. For RO links (# Avg. RO-L), we evaluate two scales: the document-level, representing the total RO links per document; and the element-level, representing the average number of links originating from an individual layout element. Full RO statics are provided in Table 9.

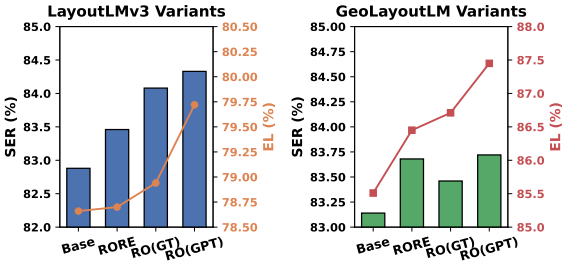


Figure 5: Performance comparison on the ROOR dataset. SER is shown as bars, and EL as lines. Each subplot corresponds to a model backbone (LayoutLMv3 or GeoLayoutLM), with enhanced variants demonstrating consistent improvements across both tasks.

Method	EC-FUNSD		CORD
	SER	EL	SER
LayoutLMv3-large	82.88	78.66	96.82
+ RORE	83.46	78.70	96.79
+ RO(GT)	84.08	78.94	–
+ RO(GPT)	**84.33**	**79.72**	**97.04**
GeoLayoutLM	83.14	85.51	96.45
+ RORE	83.68	86.45	96.89
+ RO(GT)	83.46	86.71	–
+ RO(GPT)	**83.72**	**87.45**	**97.34**

Table 3: Document understanding results on EC-FUNSD and CORD. RO supervision improves both SER and EL performance, with GPT-4o-inferred RO achieving the best overall F1 results across models.

strong performance in modeling local, immediate-succession RO relations. GPT-4o achieves an impressive F1 score of 80.85% at the word level using only textual layout information, validating its capability in capturing localized RO links. However, at the segment level, its F1 score drops to 30.25%, suggesting a distinct global reading preference. While deviating from human annotations, it reflects a more comprehensive capture of document-wide semantic flow rather than simple adjacency-based ordering, which provides meaningful implications for downstream document understanding tasks. Table 2 indicates that LLM-inferred ROs prioritize global structural relationships.

To thoroughly assess the downstream impact of these RO representations, we conduct extensive experiments comparing four configurations: 1) **Base**, with no RO supervision; 2) **RORE**, which uses RO relations predicted by a fine-tuned LayoutLMv3 model; 3) **RO(GT)**, which leverages ground-truth RO annotations; and 4) **RO(GPT)**, which uses RO relations inferred by GPT-4o. Our results, illustrated in Figure 5, demonstrate that RO(GPT) significantly outperforms other configurations with the highest F1 scores, particularly in the EL task, where it surpasses RO(GT) by a substantial margin—highlighting the effectiveness of LLM-inferred RO priors. While RO(GT) shows occasional performance degradation in SER tasks, likely due to annotation inconsistencies or overfitting to adjacency-based RO relations. In contrast, RO(GPT) implies consistent improvements across both LayoutLMv3 and GeoLayoutLM backbones.

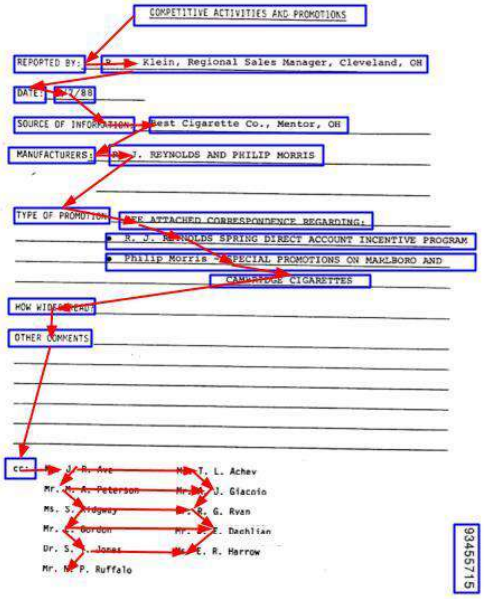
Further evaluations on EC-FUNSD and CORD are presented in Table 3, revealing robust generalization capabilities of our approach in diverse document domains. Incorporating GPT-4o-inferred RO relations into lightweight models via RO-aware attention consistently improves document understanding performance, achieving their best results. These results confirm that modeling global RO relations significantly enhances document understanding capabilities across diverse domains.

5.2 Results on Generative Document QA

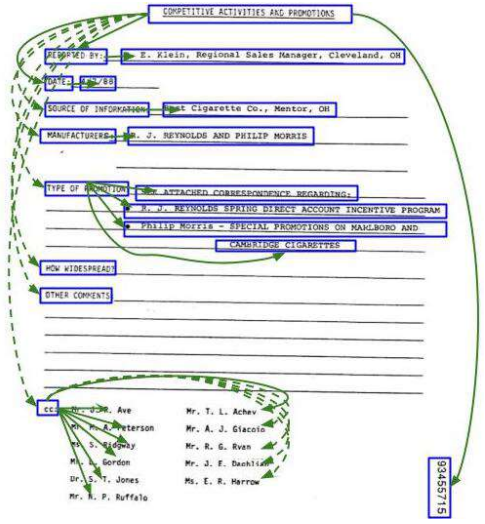
To comprehensively evaluate the effectiveness of our LLM-based RO-aware document understand-

Method	DocVQA	InfoVQA
LayoutLMv3-base† (Zhang et al., 2024a)	69.13	22.62
RORE-LayoutLMv3-base (Zhang et al., 2024a)	73.53	29.20
LayoutLMv2-base (Xu et al., 2021)	78.08	-
LayoutLMv3-base (Huang et al., 2022)	78.76	-
LayoutLMv3-large (Huang et al., 2022)	83.37	45.10
LayoutLMv2-large (Xu et al., 2021)	83.48	-
TILT-large* (Powalski et al., 2021)	87.05	61.20
UDOP* (Tang et al., 2023)	87.80	63.00
DocFormerv2-large* (Appalaraju et al., 2024)	87.84	48.80
PaLI-3* (Chen et al., 2023)	88.60	62.40
GPT-4o-text	80.35	50.83
+ RORE	81.36	52.70
+ RO(GPT-4o)	81.68	53.27
GPT-4o-multimodal	91.89	76.30
+ RORE	92.53	76.61
+ RO(GPT-4o)	**92.77**	**77.15**
Human (Mathew et al., 2021)	98.10	-

Table 4: Evaluation ANLS results on the test set of DocVQA and InfoVQA, * indicates training with additional VQA data. Baseline results are obtained from their original paper, where LayoutLMv3-base† is the reproduced result by Zhang et al. (2024a).



(a)



(b)

Figure 6: (a): human annotated RO (ground-truth label), capturing local RO dependency between adjacent layout elements. (b): LLM-inferred RO, capturing global RO relations across layout elements.

ing paradigm, we extend our experiments to encompass a broader range of generative document QA tasks. The results are presented in Table 4, demonstrating the versatility and robustness of our approach. Specifically, our methodology incorporates generated RO relations directly into the prompt (refer Section A.3 for more details), instructing GPT-4o to synthesize answers for both DocVQA and InfoVQA tasks in an end-to-end manner. From the table, we can conclude that our proposed paradigm demonstrates exceptional effectiveness in enhancing generative document QA tasks, achieving enhanced generative document QA performance against baseline approaches. The integration of RO relations serves as an informative auxiliary feature, enabling our LLM-based backbones to assess global-aware reading patterns, better capture document-level context and relationships, and generate more accurate and contextually relevant answers. The results suggest a novel and promising direction towards achieving human-level document understanding capabilities, bridging the gap between machine and human comprehension of complex documents.

Ablation Study. To rigorously validate the contribution of each component in our system, we conducted a comprehensive ablation study, shown in Table 5. The study systematically examines three key components: 1) RO information integration; 2) textual document layout information incorporation; 3) visual document information utilization. Our ablation analysis reveals several insights. The optimal configuration is consistently achieved when combining all three components. The results demonstrate that these components work synergistically, with each component contributing uniquely to the

Method	DocVQA		InfoVQA	
	ANLS	EM	ANLS	EM
GPT-4o+RO	92.77	79.64	77.15	43.63
w/o RO	92.53	78.44	76.61	42.56
w/o RO, text	91.89	73.95	76.30	39.98
w/o RO, image	80.35	62.59	50.83	23.89

Table 5: Ablation results of document generative VQA.

overall performance. All experimental results and findings collectively validate that our proposed unsupervised end-to-end LLM-based paradigm is a simple yet effective method for enhancing document understanding capabilities, particularly in a challenging context of generative QA tasks. The results demonstrate the practical applicability of our proposed paradigm in real-world document processing scenarios, with self-generated RO relations.

Case Study. We provide a detailed case study to demonstrate the advantage of our proposed LLM-inferred RO relations compared to the human-annotated RO signals in Figure 6. It demonstrates the practical benefits and real-world implications of our proposed approach, revealing improved capability in recognizing document RO relations in global layout comprehension. For detailed examples and additional analysis, please refer to Section A.4, which provides comprehensive evidence of our model's enhanced capabilities in handling challenging generative document VQA tasks through improved global layout awareness.

5.3 Further Analysis

Effectiveness of RO-aware Attention. To examine how RO information contributes to document modeling at the architectural level, we investigate two attention variants for integrating local and global RO relations into lightweight models. As shown in Table 6, the combined use of local and global RO signals yields the best performance, confirming their complementary roles in enhancing layout-sensitive attention mechanisms.

LLM-based RO Generation Analysis. We evaluate the quality of the RO predictions in various LLMs, ranging from open-weights to closed-sources proprietary LLMs, under both text-only and multimodal configurations. As evidenced in Table 7, the predicted RO relationships of GPT-4o consistently demonstrate superior performance in downstream SER and EL tasks across all three backbone models (LayoutLMv3-base, LayoutLMv3-large, GeoLayoutLM-large), validating its robust capability to infer semantically correlated and layout-aware reading order structures without task-specific fine-tuning. Notably, we observe that GPT-4o's text-only variant achieves remarkable results, suggesting that LLMs can effectively generalize beyond human heuristics to provide more consistent and semantically aligned reading paths than traditional ground truth annotations. While GPT-4o's multimodal variant further strengthens these findings, demonstrating that the integration of visual and textual cues significantly enhances reading order prediction quality, particularly in documents with complex structural layouts.

RO-aware Document Representation. To evaluate the impact of RO information on document representations, we reformulate DocVQA and InfoVQA as retrieval tasks, measuring the relevance between document-question pairs. We leverage GPT-4o to generate RO pairs for both tasks, with a LayoutLMv3-base trained ROP model serving as a baseline for comparison. The evaluation methodology involves reordering document tokens according to the predicted RO and processing them through OpenAI text-embedding-3-small [4] to obtain unified document embeddings, while question embeddings are generated through the same model. This approach allows us to quantitatively assess the effectiveness of different RO strategies through average cosine similarity scores between document and question embeddings, providing a direct measure of semantic alignment.

The experimental results, presented in Table 8, demonstrate the superior performance of our GPT-4o-based RO approach in enhancing document-question alignment. Specifically, our method achieves significant improvements in average cosine similarity scores, showing increases of 0.60% on DocVQA and 2.73% on InfoVQA compared to baseline RO relations. These improvements are particularly pronounced in text-extraction-focused tasks like DocVQA, where accurate RO is crucial for comprehension. In contrast, InfoVQA shows a slight decrease in average cosine similarity (47.74% $\rightarrow$ 46.53%), this observation aligns with the task's inherent characteristics, as InfoVQA heavily emphasizes reasoning (22%) and counting capabilities

[4] https://platform.openai.com/docs/models/text-embedding-3-small

Relation Order Generation Method	LayoutLMv3-base		LayoutLMv3-large		GeoLayoutLM-large	
	Local Attn	Local+Global Attn	Local Attn	Local+Global Attn	Local Attn	Local+Global Attn
Human Annotation	64.34	63.54▼	78.94	77.49▼	83.71	86.71▲
ROP-prediction (RORE)	63.78	63.96▲	77.97	78.70▲	85.45	86.45▲
Gemma-2-27b-text	62.86	63.32▲	78.09	76.12▼	87.12	85.99▼
Mistral-large-instruct-text	63.15	65.22▲	78.45	78.44▼	86.11	84.87▼
GPT-4o-text	63.61	65.33▲	76.08	77.89▲	85.44	86.61▲
Gemini-1.5-pro-text	63.28	64.21▲	76.53	78.31▲	84.86	86.35▲
GPT-4o-multimodal	63.74	65.21▲	76.62	79.72▲	86.77	87.45▲
Gemini-1.5-pro-multimodal	62.12	63.16▲	76.16	78.11▲	86.57	86.46▼

Table 6: Attention ablation studies on EC-FUNSD EL task. Blue ▲ indicates Global Attn improves over Local Attn; red ▼ indicates a drop of F1 score. Full results available in Table 10.

Relation Order Generation Method	LayoutLMv3-base		LayoutLMv3-large		GeoLayoutLM-large	
	SER	EL	SER	EL	SER	EL
No-layout	81.47	61.11	82.88	78.66	83.14	85.51
Human Annotation	81.55	64.34	84.08	78.94	83.46	86.71
Gemma-2-9b-text†	81.13	64.61	84.13	79.77	83.61	86.23
Gemma-2-27b-text†	81.38	63.32	83.62	78.09	83.24	87.12
Mistral-nemo-instruct-text†	81.27	62.04	83.17	78.44	83.33	85.94
Mistral-large-instruct-text†	81.47	65.22	83.29	78.45	83.66	86.11
GPT-4o-text*	81.34	**65.33**	**84.33**	77.89	83.64	86.61
GPT-4v-text*	80.92	61.90	82.65	78.30	82.79	86.89
Gemini-1.0-pro-text*	80.87	63.86	83.96	**80.04**	83.19	86.38
Gemini-1.5-pro-text*	81.22	64.21	83.08	78.31	82.63	86.35
Claude-3.5-sonnet-text*	**81.82**	64.08	84.08	78.44	83.04	86.63
GPT-4o-multimodal*	81.56	65.21	83.84	79.72	**83.72**	**87.45**
GPT-4v-multimodal*	81.20	63.93	83.50	77.68	83.49	86.26
Gemini-1.5-flash-multimodal*	81.32	63.48	83.72	77.34	83.61	86.26
Gemini-1.5-pro-multimodal*	81.55	63.16	83.31	78.11	82.96	86.57
Claude-3.5-sonnet-multimodal*	81.56	62.93	83.04	78.72	83.22	87.21

Table 7: EC-FUNSD – SER↑ / EL↑. Best F1 scores per column are bolded. † indicates open-weights models, * indicates closed-sources proprietary models. Multimodal models are shaded.

Method	Cosine Similarity	
	DocVQA	InfoVQA
Original	35.84	**47.74**
Baseline ROP	35.66	43.80
GPT-4o	**36.26**	46.53

Table 8: Results on zero-shot question-to-document cosine similarity. Baseline ROP represents the augmented document by following the reading order relations generated by the baseline ROP model. GPT-4o indicates the reordered document by following the reading order relations generated by the GPT-4o model.

(33%) (Zhang, 2024), depending less on RO. This nuanced performance across different task types validates the effectiveness of our approach and provides valuable insights into the role of RO in various document understanding scenarios.

6 Conclusion

In this paper, we propose a novel paradigm for document understanding that revises how structural information is incorporated by explicitly leveraging reading order (RO) as a structural prior. By utilizing LLMs to infer document-wide RO relations without manual annotation, our dual-pathway framework—combining lightweight RO-aware attention for compact backbones and RO-prompting for LLMs—enables end-to-end, scalable RO modeling. Extensive experiments on semantic entity recognition, entity linking, and document question answering demonstrate consistent and significant improvements over strong baselines, surpassing even models trained with human-annotated RO supervision. Our comprehensive analysis shows that LLM-inferred RO substantially enhances structural alignment and generalization, especially for complex or out-of-domain documents.

Limitations

Although we evaluate our approach across three representative document understanding tasks (Semantic Entity Recognition, Entity Linking, and Question Answering), the effectiveness of our method can be impacted by underlying data quality issues, particularly OCR errors in the original datasets that disproportionately affect text-only LLM document understanding performance. Furthermore, while we pioneer the investigation of LLM-based reading order relations, the computational resources required by our approach currently constrains its application to documents with extensive long-range context or multi-page scenarios—a limitation that becomes particularly relevant when scaling to more complex document understanding tasks. These limitations not only highlight opportunities for future optimization but also underscore the need for more efficient approaches to handle increasingly complex document structures.

Ethics Statement

Our work complies with the ACL Ethics Policy. In this work, we use LLMs to generate reading order relations using publicly available document understanding benchmarks, which are widely used to evaluate the document understanding performance. We provide detailed procedures to generate reading order relations and provide proper citations to their source benchmarks. We will publicly release our generated reading order information.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R. Manmatha. 2024. Docformerv2: Local features for document understanding. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 709–718. AAAI Press.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez, Marçal Rusiñol, Minesh Mathew, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. ICDAR 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1563–1570. IEEE.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. 2023. Pali-3 vision language models: Smaller, faster, stronger. *CoRR*, abs/2310.09199.

Shengda Fan, Yanting Wang, Shasha Mo, and Jianwei Niu. 2024. LogicST: A logical self-training framework for document-level relation extraction with incomplete annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5496–5510. Association for Computational Linguistics.

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

Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 2022. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4573–4582. IEEE.

Nidhi Hegde, Sujoy Paul, Gagan Madan, and Gaurav Aggarwal. 2023. Analyzing the efficacy of an llm-only approach for image-based document question answering. *CoRR*, abs/2309.14389.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. BROS: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence,*

EAAI 2022 Virtual Event, February 22 - March 1, 2022, pages 10767–10775. AAAI Press.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 3096–3120. Association for Computational Linguistics.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021. Three sentences are all you need: Local path enhanced document relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 998–1004. Association for Computational Linguistics.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document AI with unified text and image masking. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4083–4091. ACM.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Anoop R. Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4459–4469. Association for Computational Linguistics.

Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. 2023. Visually-situated natural language understanding with contrastive reading model and frozen large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11989–12010. Association for Computational Linguistics.

Marcel Lamott, Yves-Noel Weweler, Adrian Ulges, Faisal Shafait, Dirk Krechel, and Darko Obradovic. 2024. LAPDoc: Layout-aware prompting for documents. In *Document Analysis and Recognition - ICDAR 2024 - 18th International Conference, Athens, Greece, August 30 - September 4, 2024, Proceedings, Part IV*, volume 14807 of *Lecture Notes in Computer Science*, pages 142–159. Springer.

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

Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, Hao Liu, and Can Huang. 2024. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. *CoRR*, abs/2407.01976.

Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. 2023. Geolayoutlm: Geometric pre-training for visual information extraction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 7092–7101. IEEE.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 15630–15640. IEEE.

Zhiming Mao, Haoli Bai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Visually guided generative text-layout pre-training for document intelligence. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4713–4730. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: A consolidated receipt dataset for post-ocr parsing.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, Chen-Yu Lee, and Nan Hua. 2024. LMDX: language model-based document information extraction and localization. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15140–15168. Association for Computational Linguistics.

Rafal Powalski, Lukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Palka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 732–747. Springer.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. *CoRR*, abs/2208.03054.

Aneeta Sylolypavan, Derek H. Sleeman, Honghan Wu, and Malcolm Sim. 2023. The impact of inconsistent human annotations on AI driven clinical decision making. *npj Digit. Medicine*, 6.

Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2024. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19071–19079. AAAI Press.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19254–19264. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024. Docllm: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8529–8548. Association for Computational Linguistics.

Renshen Wang, Yasuhisa Fujii, and Alessandro Bissacco. 2023a. Text reading order in uncontrolled conditions by sparse graph segmentation. In *Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part VI*, volume 14192 of *Lecture Notes in Computer Science*, pages 3–21. Springer.

Weishi Wang, Hengchang Hu, Zhijie Zhang, Zhaochen Li, Hongxin Shao, and Daniel Dahlmeier. 2025. Document intelligence in the era of large language models: A survey. *CoRR*, abs/2510.13366.

Wenjin Wang, Yunhao Li, Yixin Ou, and Yin Zhang. 2023b. Layout and task aware instruction prompt for zero-shot document image question answering. *CoRR*, abs/2306.00526.

Yanbo J. Wang, Sheng Chen, Hengxing Cai, Wei Wei, Kuo Yan, Zhe Sun, Hui Qin, Yuming Li, and Xiaochen Cai. 2022. A GlobalPointer based robust approach for information extraction from dialog transcripts. In *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 13–18, Abu Dhabi, Beijing (Hybrid). Association for Computational Linguistics.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. Layoutreader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4735–4744. Association for Computational Linguistics.

Xingjiao Wu, Luwei Xiao, Xiangcheng Du, Yingbin Zheng, Xin Li, Tianlong Ma, Cheng Jin, and Liang He. 2024. Cross-domain document layout analysis using document style guide. *Expert Syst. Appl.*, 245:123039.

Hangdi Xing, Changxu Cheng, Feiyu Gao, Zirui Shao, Zhi Yu, Jiajun Bu, Qi Zheng, and Cong Yao. 2024. Dochienet: A large and diverse dataset for document hierarchy parsing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1129–1142. Association for Computational Linguistics.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2579–2591. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 764–777. Association for Computational Linguistics.

Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2023. StrucTexTv2: Masked visual-textual prediction for document image pre-training. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Chong Zhang, Ya Guo, Yi Tu, Huan Chen, Jinyang Tang, Huijia Zhu, Qi Zhang, and Tao Gui. 2023. Reading order matters: Information extraction from visually-rich documents by token path prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13716–13730. Association for Computational Linguistics.

Chong Zhang, Yi Tu, Yixi Zhao, Chenshu Yuan, Huan Chen, Yue Zhang, Mingxu Chai, Ya Guo, Huijia Zhu, Qi Zhang, and Tao Gui. 2024a. Modeling layout reading order as ordering relations for visually-rich document understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 9658–9678. Association for Computational Linguistics.

Chong Zhang, Yixi Zhao, Chenshu Yuan, Yi Tu, Ya Guo, and Qi Zhang. 2024b. Rethinking the evaluation of pre-trained text-and-layout models from an entity-centric perspective. *arXiv preprint arXiv:2402.02379*.

Jinxu Zhang. 2024. Read and think: An efficient step-wise multimodal language model for document understanding and reasoning. *arXiv preprint arXiv:2403.00816*.

Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. 2024. Lora land: 310 fine-tuned llms that rival gpt-4, A technical report. *CoRR*, abs/2405.00732.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14612–14620. AAAI Press.

A Appendix

A.1 Dataset

We follow ROOR (Zhang et al., 2024a) to manage the document layout elements in metadata, transforming all benchmarks to a unified format.

ROOR contains 199 samples, including 10,662 segments, 31,297 words, and 10,967 expert-level annotated reading order linking pairs. We leverage the reading order linking pairs to develop and evaluate the baseline ROP model.

EC-FUNSD (Zhang et al., 2024b) contains word and segment-level layout annotations with expert-level entity and relation annotations, comprising 10,662 segments, 31,297 words, 8,398 entities, 3,912 linking relations. We evaluate SER and EL on the EC-FUNSD dataset

CORD (Park et al., 2019) is a receipt OCR parsing dataset, including both box-level text and parsing class annotations. It has 1,000 samples with 30 entity classes, supporting our SER evaluation.

DocVQA (Mathew et al., 2021) is a popular large-scale industry document VQA dataset, including 12, 767 documents and 50, 000 questions and answers. Each document contains diverse textual, graphical, and structural elements.

InfoVQA (Mathew et al., 2022) is a famous dataset comprising a diverse collection of infographics, natural language questions, and answers annotations. The collected questions require methods to jointly reason over the document layout, textual content, graphical elements, and data visualizations. It has around 5,400 images and 30,000 QA pairs.

A.2 Implementation Details

All of our experiments that require supervised fine-tuning are conducted on a single NVIDIA A100-80GB GPU. The implementation is based on the deep learning framework PyTorch [5].

Baseline ROP Model. We employ LayoutLMv3-base (Huang et al., 2022) as our backbone ROP model. The maximum sequence length of textual tokens for both of them is 2048. The maximum layout elements is 512 for the word-level ROP and 256 for the segment-level ROP. We fine-tune the ROP model on the ROOR dataset for 500 epochs with a patience of 50 for early stopping. The dimension of query and key vectors are set to 128. We use an AdamW optimizer with 10% linear warming-up steps, a 0.1 dropout rate, and a 1e-5 weight decay, together with a cosine scheduler. The learning rate is set to 2e-5 with a batch size of 16.

Lightweight RO-aware Model. For SER and EL tasks, we follow the implementation of Zhang et al. (2024a), with all experiments using the learning rate of 1e-5, batch size of 16, number of fine-tuning epoches of 500 and early stopping patience of 50. The initial value for λ is 10 in all layers, and the learning rate of the λ are searched between the original learning rate and 1e-2, where the latter is for faster convergence. We use the RO information in the first 4 layers, and their λ values are set to 0.1 initially.

Foundational LLMs. We use their official API to invoke the foundational LLMs. As illustrated in Figure 3, we instruct LLMs to generate RO pairs. We then incorporate the generated RO relations directly into the instruction prompt, which comprises OCR text, bounding box, image, and generated RO information. The details could be found in Section A.3.

A.3 Generative Document QA Prompts

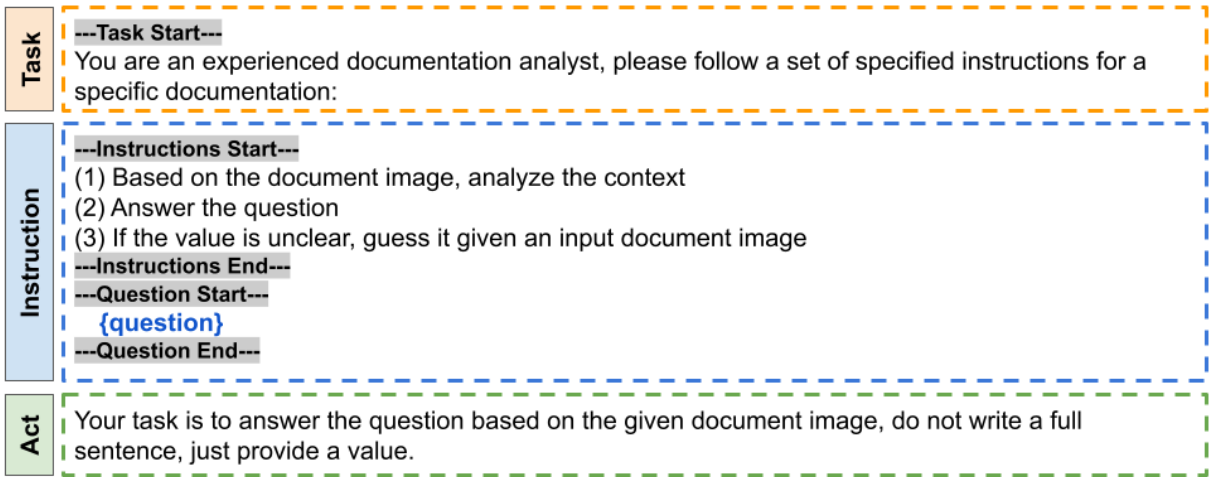
Our prompt design encompasses multiple modality variants tailored to specific document understanding scenarios. The complete prompt templates are illustrated in Figure 7:

Visual-based Document QA. Figure 7a demonstrates our base prompt design for image-based visual document QA tasks. This template is the foundation for handling purely visual document understanding scenarios without integrating RO relations.

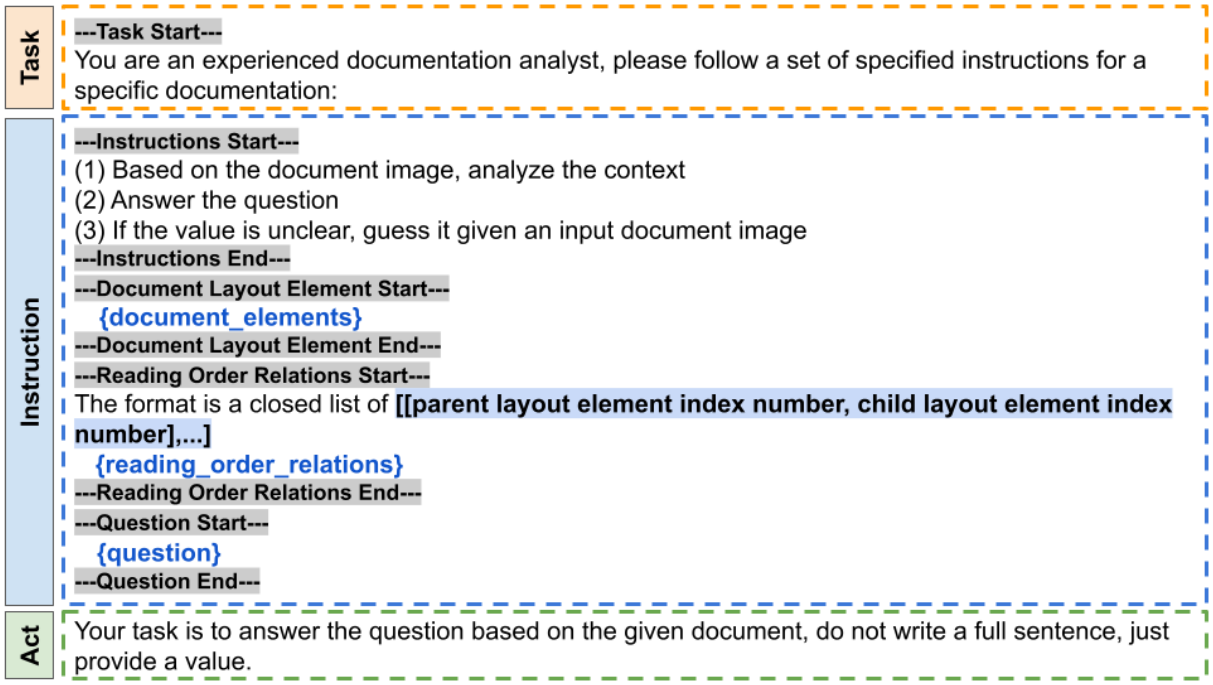
Text-Based Document QA. Figure 7b presents our specialized RO-aware prompt designed for text-based document QA settings. This template is optimized for scenarios where document OCR textual content and structure are the primary focus.

Multimodal Document QA. We consider both the document image and OCR textual content, attaching the document image to Figure 7b to form the multimodal RO-aware model for improved document comprehension and question answering performance.

[5] https://pytorch.org/



(a) Image QA prompt.



(b) Text QA prompt.

Figure 7: Generative VQA prompts illustration

A.4 Case Study

We systematically analyzed the performance of our proposed paradigm across different document types and question categories to provide a comprehensive understanding of its capabilities. Our approach shows robust performance across diverse real-world document types, as shown in Figure 9. Our method demonstrates advanced strength in handling industry text-intensive documents with complex layout structures. Meanwhile, Figure 10 shows our proposed method could effectively handle visual-intensive documents, capturing complicated spatial relationships.

This case study implies that our approach achieves robust performance across diverse real-world document types and complexities. And the improvements in global layout awareness suggest better scalability to address more complex document understanding tasks, where the consistent performance across different document types indicates strong generalization capabilities

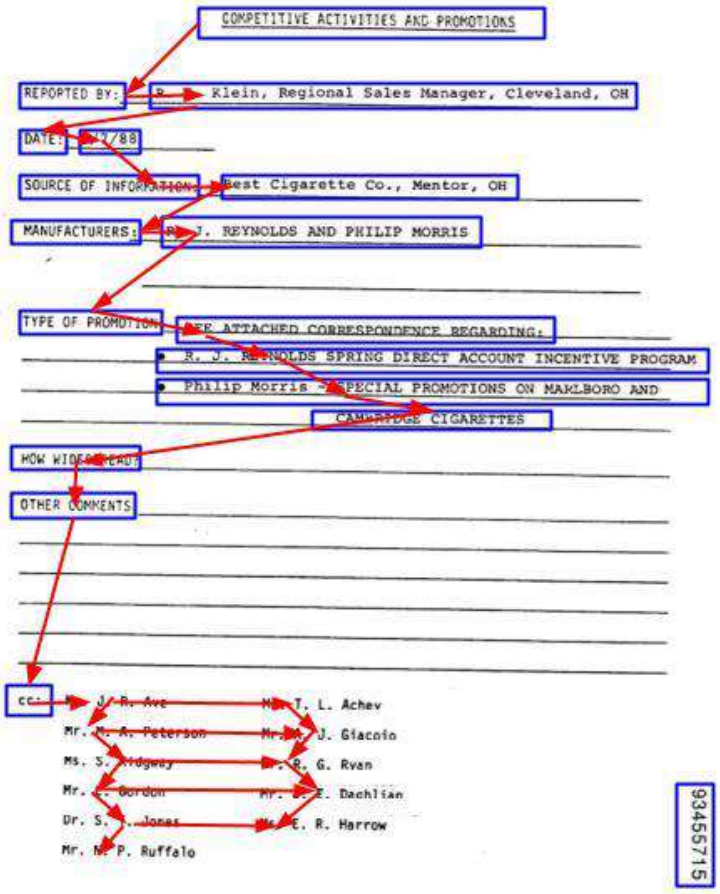
A.5 More Experimental Results

RO statistics.

Effectiveness of RO-aware Attention. To examine how RO information contributes to docu-

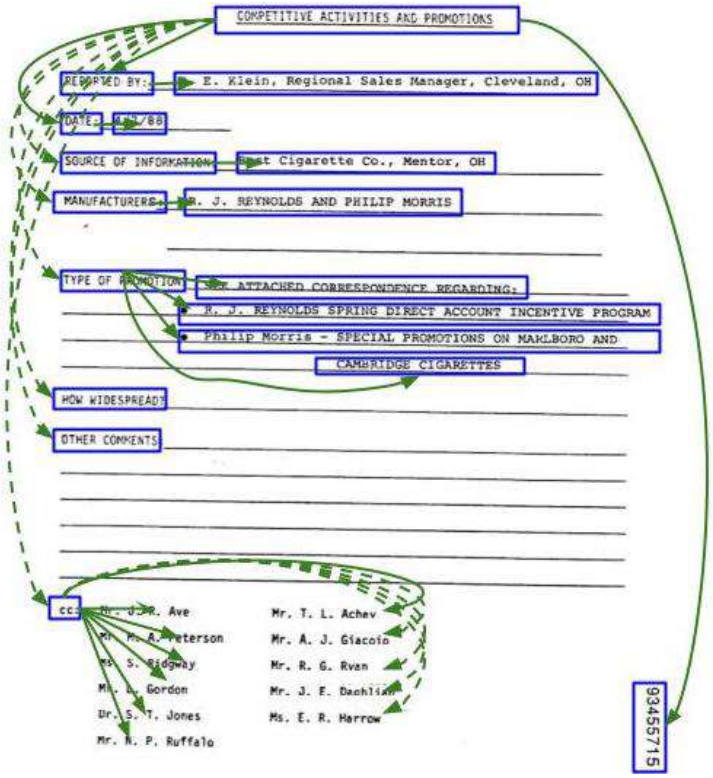
---Document Layout Element Start---
0: TYPE OF PROMOTION:<bbox>[58, 358, 196, 376]
1: cc:<bbox>[58, 720, 85, 733]
2: OTHER COMMENTS:<bbox>[60, 533, 173, 552]
3: HOW WIDESPREAD?<bbox>[61, 485, 172, 507]
4: MANUFACTURERS:<bbox>[63, 271, 174, 287]
5: REPORTED BY:<bbox>[66, 136, 154, 160]
6: DATE:<bbox>[67, 178, 104, 202]
7: SOURCE OF INFORMATION:<bbox>[67, 223, 229, 245]
8: Mr. M. A. Peterson<bbox>[100, 750, 237, 766]
. . .
14: 4/7/88<bbox>[118, 184, 173, 200]
15: R. E. Klein, Regional Sales Manager, Cleveland, OH<bbox>[184, 136, 627, 158]
16: - R. J. REYNOLDS SPRING DIRECT ACCOUNT INCENTIVE PROGRAM<bbox>[187, 391, 687, 409]
17: - Philip Morris- SPECIAL PROMOTIONS ON MARLBORO AND<bbox>[187, 418, 654, 442]
18: CAMBRIDGE CIGARETTES<bbox>[352, 451, 529, 466]
19: R. J. REYNOLDS AND PHILIP MORRIS<bbox>[193, 271, 476, 289]
20: SEE ATTACHED CORRESPONDENCE REGARDING:<bbox>[207, 364, 547, 380]
21: COMPETITIVE ACTIVITIES AND PROMOTIONS<bbox>[247, 64, 523, 88]
22: Best Cigarette Co., Mentor, OH<bbox>[247, 225, 517, 246]
23: Mr. A. J. Giacoio<bbox>[281, 751, 413, 769]
. . .
28: 93455715<bbox>[681, 810, 699, 903]
---Document Layout Element End---

(a) Layout element example of EC-FUNSD.



(b) Human annotated reading order ground truth, capturing local reading order dependency between adjacent layout elements, especially for the carbon copy recipients.

ment modeling at the architectural level, we investigate two attention variants for integrating local and global RO relations into lightweight models. We present the full evaluation results in Table 10.



(c) LLM-based ROP model generated reading order relations, capturing global reading order relations across layout elements

Figure 8: An example of EC-FUNSD. The human-annotated ground truth is suboptimal, where GPT-4o can capture the document-level RO relations. LLM-inferred RO relations provide appropriate RO information and dependency, leading to enhanced downstream document understanding performance.

A.6 Usage of AI Assistant

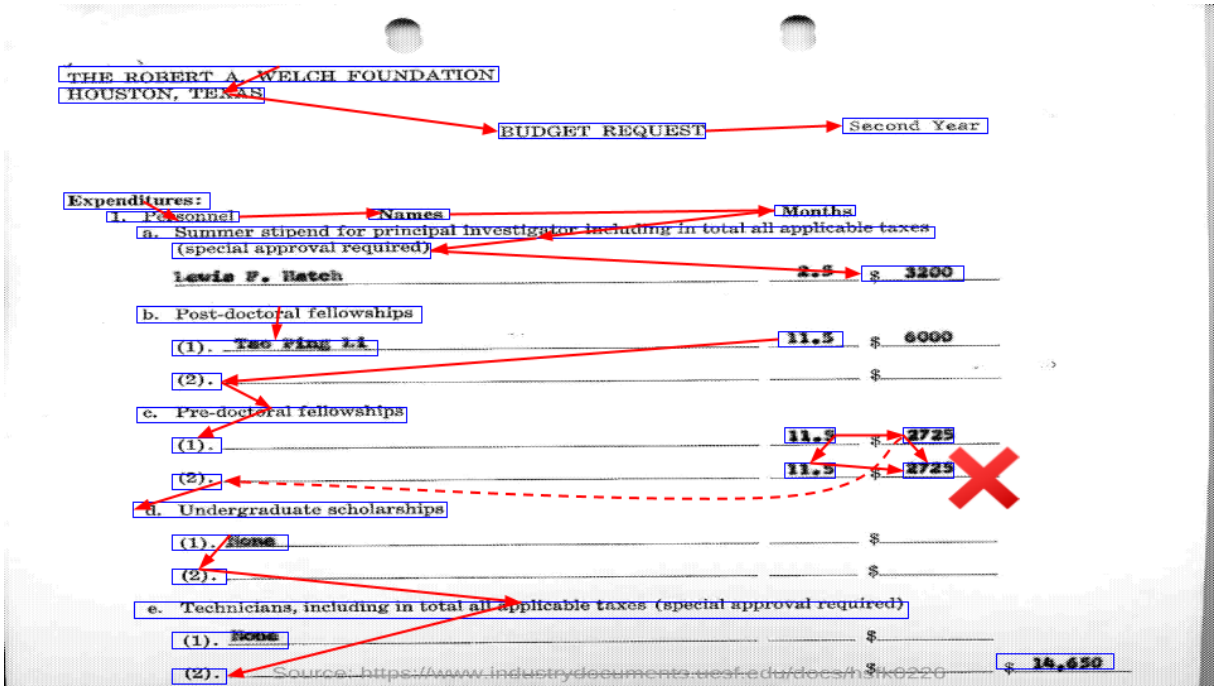
We use AI assistants or tools such as ChatGPT and Grammarly to correct grammar errors and refine the content.

---Document Layout Element Start---

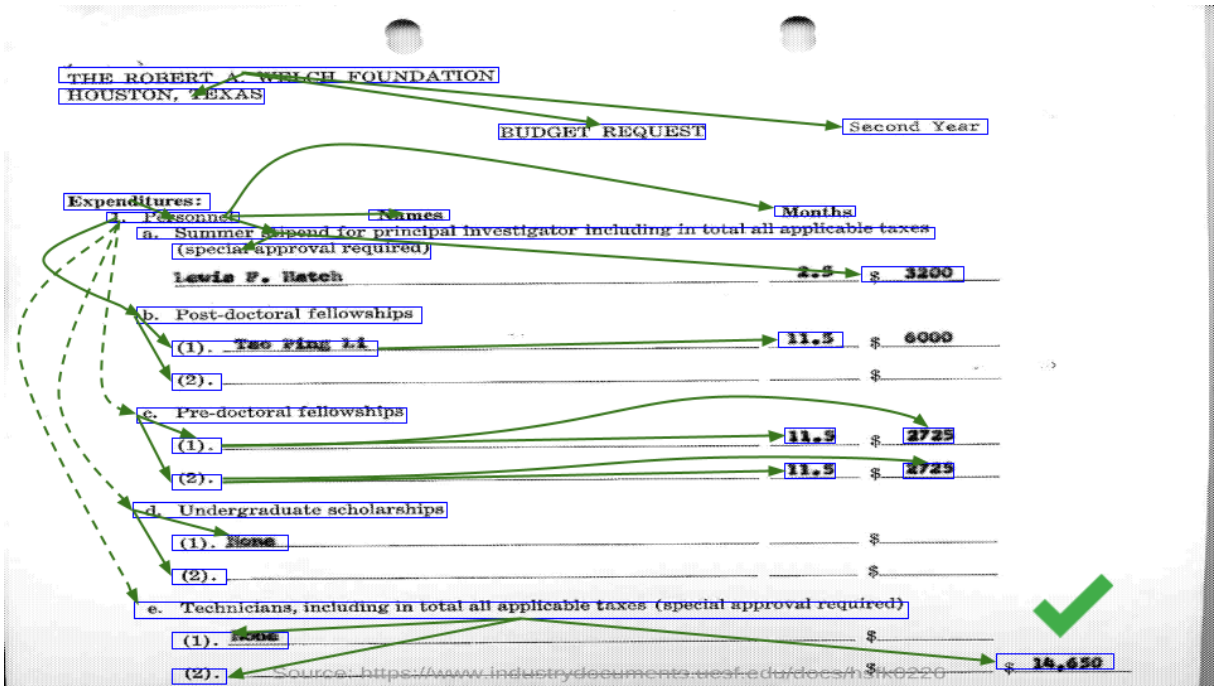
```
0: THE ROBERT A. WELCH FOUNDATION<bbox>[88, 128, 690, 166]
1: HOUSTON, TEXAS<bbox>[84, 165, 365, 200]
2: BUDGET REQUEST<bbox>[692, 243, 985, 275]
3: Second Year<bbox>[1182, 233, 1369, 265]
4: Expenditures:<bbox>[86, 387, 280, 417]
5: 1. Personnel<bbox>[147, 417, 324, 450]
6: Names<bbox>[521, 415, 621, 446]
7: Months<bbox>[1085, 409, 1193, 439]
8: a. Summer stipend for principal investigator including in total all applicable taxes<bbox>[185, 439, 1302, 484]
9: (special approval required)<bbox>[244, 481, 599, 517]
10: $.3800<bbox>[1208, 530, 1331, 571]
11: b. Post-doctoral fellowships<bbox>[187, 615, 579, 651]
12: (1). Too Ping LA<bbox>[239, 675, 513, 717]
13: 11.5<bbox>[1094, 667, 1166, 699]
14: (2) .<bbox>[241, 751, 292, 784]
15: c. Pre-doctoral fellowships<bbox>[194, 816, 564, 851]
16: (1).<bbox>[241, 885, 292, 919]
17: 11.5<bbox>[1094, 866, 1175, 896]
18: 1725<bbox>[1258, 863, 1330, 894]
19: (2).<bbox>[245, 951, 298, 982]
20: 11.5<bbox>[1097, 934, 1170, 966]
21: 2725<bbox>[1259, 932, 1332, 963]
22: d. Undergraduate scholarships<bbox>[194, 1015, 620, 1049]
23: (1). None.<bbox>[243, 1079, 396, 1117]
24: (2).<bbox>[253, 1151, 293, 1182]
25: e. Technicians, including in total all applicable taxes (special approval required)<bbox>[197, 1206, 1260, 1252]
26: (1). Boos<bbox>[251, 1274, 389, 1316]
27: (2).<bbox>[249, 1351, 309, 1380]
28: 314.650<bbox>[1401, 1328, 1539, 1368]
```

---Document Layout Element End---

(a) Layout element example of DocVQA.



(b) Baseline ROP model generated reading order relations, leading to the wrong answer.



(c) LLM-based ROP model generated reading order relations, leading to the correct answer.

Figure 9: An example of DocVQA, with the question of "What is the total amount?". LLM-based reading order relations benefit the model for enhanced document understanding capabilities to generate the correct answer 14,650.

Method	Avg. RO-D	# Avg. RO-L		# Max. RO-L
	Document	Document	Element	Document
GT	11.84	55.11	1.09	1.74
ROP-prediction (RORE)	11.90	51.51	1.11	1.96
Gemma-2-9b-text	16.54	43.88	14.31	24.11
Gemma-2-27b-text	13.82	45.60	10.57	20.67
Mistral-nemo-instruct-text	20.11	55.39	22.83	33.05
Mistral-large-instruct-text	14.38	60.58	11.17	23.23
GPT-4o-text	**14.66**	**42.55**	**6.14**	**12.51**
GPT-4v-text	14.49	42.06	7.71	13.89
Gemini-1.0-pro-text	18.30	51.33	19.41	27.78
Gemini-1.5-pro-text	19.15	38.67	5.26	14.52
Calude-3.5-sonnet-text	13.62	33.58	2.30	8.42
GPT-4o-multimodal	**15.14**	**45.29**	**4.45**	**11.28**
GPT-4v-multimodal	14.69	40.55	2.49	9.19
Gemini-1.5-flash-multimodal	14.81	44.37	5.12	15.28
Gemini-1.5-pro-multimodal	13.88	35.75	5.65	14.43
Calude-3.5-sonnet-multimodal	13.90	36.07	2.60	9.51

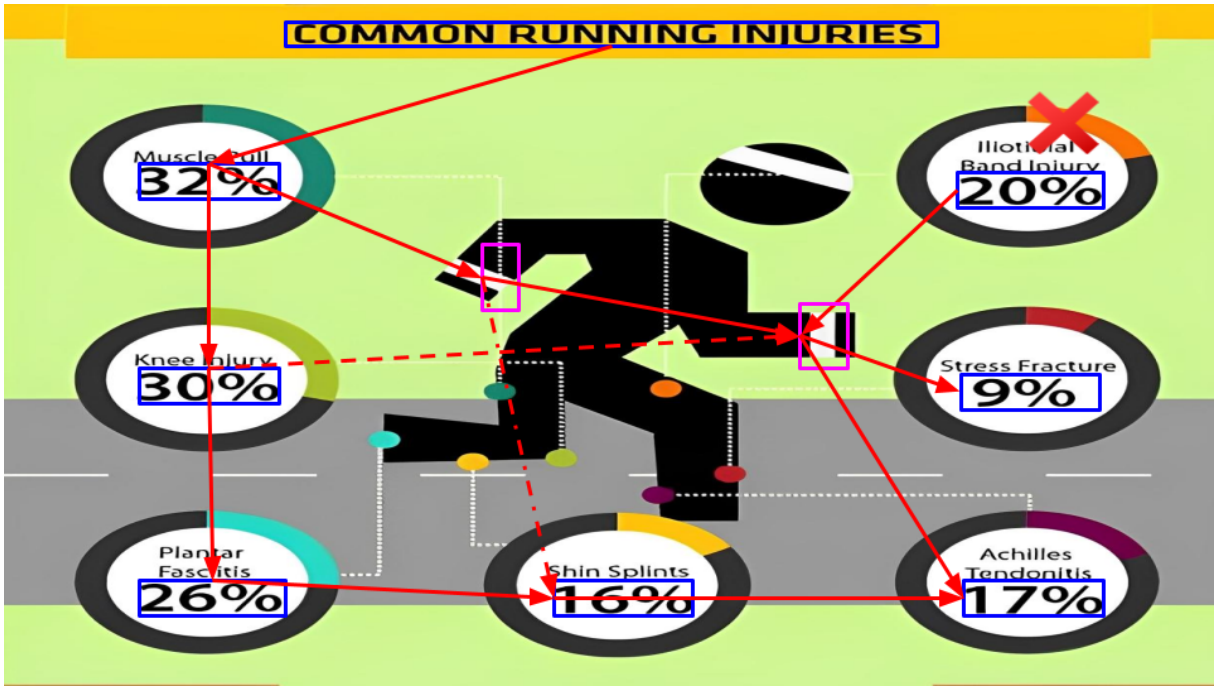
Table 9: Statistics of RO distances (RO-D) and link density (RO-L). The average distance between connected layout elements (Avg. RO-D) was calculated at the document level. For RO links (# Avg. RO-L), we evaluate two scales: the document-level, representing the total RO links per document; and the element-level, representing the average number of links originating from an individual layout element. For RO links (# Max. RO-L), we analyze the maximum number of links originating from an individual layout element.

---Document Layout Element Start---

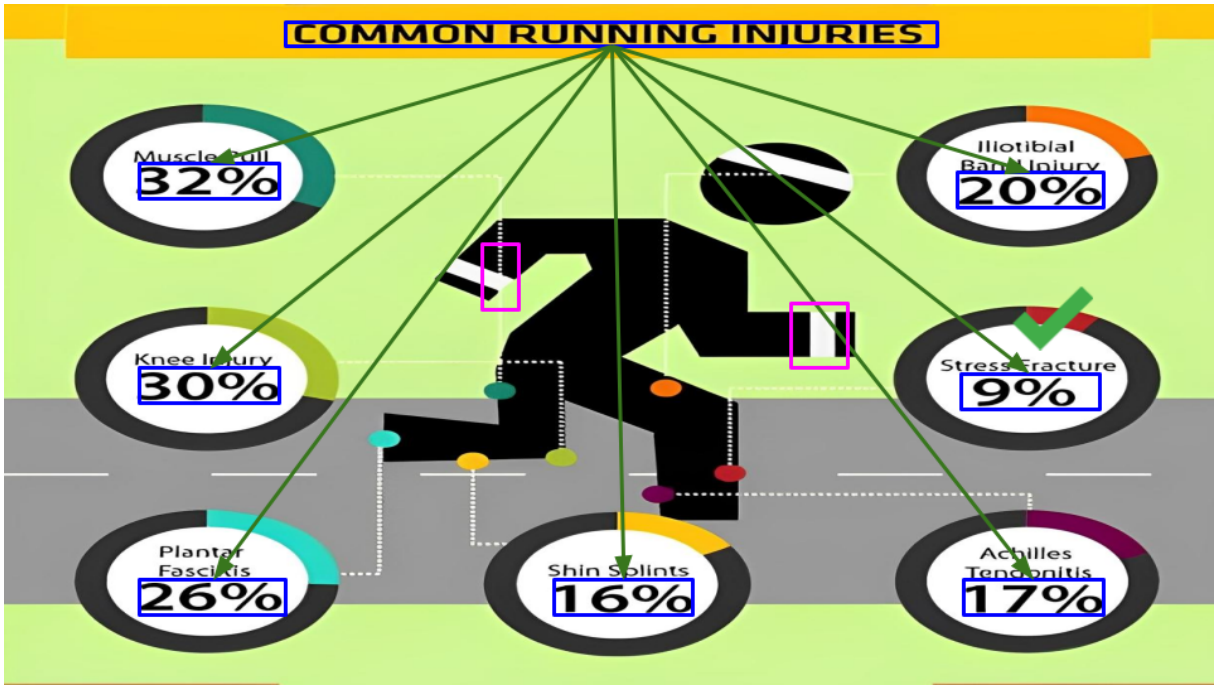
0: STATE ATE OF THE OF THE UNION 2014<bbox>[0.2, 0.03, 0.8, 0.09]
1: KEY TERMS FROM PRESIDENT OBAMA'S SPEECH<bbox>[0.19, 0.12, 0.81, 0.15]
2: better<bbox>[0.65, 0.25, 0.71, 0.28]
3: reform want take<bbox>[0.31, 0.27, 0.49, 0.33]
4: 63<bbox>[0.61, 0.29, 0.62, 0.3]
5: young economy<bbox>[0.12, 0.33, 0.33, 0.37]
6: without just,<bbox>[0.34, 0.33, 0.5, 0.38]
7: he<bbox>[0.64, 0.29, 0.72, 0.36]
8: -<bbox>[0.09, 0.37, 0.11, 0.39]
9: one<bbox>[0.17, 0.37, 0.25, 0.42]
10: businesses<bbox>[0.33, 0.38, 0.53, 0.42]
11: as<bbox>[0.59, 0.37, 0.6, 0.39]
12: home'<bbox>[0.61, 0.37, 0.7, 0.41]
13: next<bbox>[0.69, 0.36, 0.74, 0.39]
14: jobs<bbox>[0.78, 0.3, 0.92, 0.4]
15: a<bbox>[0.09, 0.39, 0.11, 0.41]
16: don I working<bbox>[0.03, 0.4, 0.27, 0.47]
17: together<bbox>[0.05, 0.45, 0.18, 0.5]
18: 63 co 65 like make - co Congress business<bbox>[0.32, 0.39, 0.91, 0.5]
19: ever day<bbox>[0.06, 0.5, 0.15, 0.53]
20: -<bbox>[0.32, 0.5, 0.34, 0.51]
...

---Document Layout Element End---

(a) Layout element example of InfoVQA.



(b) Baseline ROP model generated reading order relations, leading to the wrong answer.



(c) LLM-based ROP model generated reading order relations, leading to the correct answer.

Figure 10: An example of InfoVQA, with the question of "What percentage of common running injuries include stress fracture?". LLM-based reading order relations benefit the model for enhanced document understanding capabilities to generate the correct answer 9%.

Relation Order Generation Method	LayoutLMv3-base		LayoutLMv3-large		GeoLayoutLM-large	
	Local Attn	Global Attn	Local Attn	Global Attn	Local Attn	Global Attn
Human Annotation	64.34	63.54▼	78.94	77.49▼	83.71	86.71▲
ROP-prediction (RORE)	63.78	63.96▲	77.97	78.70▲	85.45	86.45▲
Gemma-2-9b-text	64.30	64.61▲	77.98	79.77▲	85.74	86.23▲
Gemma-2-27b-text	62.86	63.32▲	78.09	76.12▼	87.12	85.99▼
Mistral-nemo-instruct-text	61.55	62.04▲	78.44	77.43▼	85.94	85.89▼
Mistral-large-instruct-text	63.15	65.22▲	78.45	78.44▼	86.11	84.87▼
GPT-4o-text	63.61	65.33▲	76.08	77.89▲	85.44	86.61▲
GPT-4v-text	60.95	61.90▲	78.30	77.44▼	86.89	85.96▼
Gemini-1.0-pro-text	63.86	61.50▼	77.75	80.04▲	85.99	86.38▲
Gemini-1.5-pro-text	63.28	64.21▲	76.53	78.31▲	84.86	86.35▲
Calude-3.5-sonnet-text	64.08	62.27▼	77.60	78.44▲	86.63	86.29▼
GPT-4o-multimodal	63.74	65.21▲	76.62	79.72▲	86.77	87.45▲
GPT-4v-multimodal	63.83	63.93▲	75.66	77.68▲	86.26	85.53▼
Gemini-1.5-flash-multimodal	63.48	61.39▼	76.27	77.34▲	85.41	86.26▲
Gemini-1.5-pro-multimodal	62.12	63.16▲	76.16	78.11▲	86.57	86.46▼
Calude-3.5-sonnet-multimodal	62.93	62.75▼	78.72	78.17▼	87.21	86.64▼

Table 10: Attention ablation studies on EC-FUNSD EL task. Blue ▲ indicates Global Attn improves over Local Attn; red ▼ indicates a drop of F1 score.