LayoutPointer: A Spatial-Context Adaptive Pointer Network for Visual Information Extraction

Siyuan Huang[1], Yongping Xiong[1†], Guibin Wu[2]
[1]Beijing University of Posts and Telecommunications, [2]Chizhou University
{siyuanhuang, ypxiong}@bupt.edu.cn, wuguibin@czu.edu.cn

Abstract

Visual Information Extraction (VIE), as a crucial task of Document Intelligence, involves two primary sub-tasks: Semantic Entity Recognition (SER) and Relation Extraction (RE). However, VIE faces two significant challenges. Firstly, most existing models inadequately utilize spatial information of entities, often failing to predict connections or incorrectly linking spatially distant entities. Secondly, the improper input order of tokens challenges in extracting complete entity pairs from documents with multi-line entities when text is extracted via PDF parser or OCR. To address these challenges, we propose **LayoutPointer**, a Spatial-Context Adaptive Pointer Network. LayoutPointer explicitly enhances spatial-context relationships by incorporating 2D relative position information and adaptive spatial constraints within self-attention. Furthermore, we recast the RE task as a specialized cycle detection problem, employing a unique tail-to-head pointer to restore the semantic order among multi-line entities. To better evaluate the effectiveness of our proposed method, we reconstruct a multi-line dataset named MLFUD, which more accurately reflects real-world scenarios. Fine-tuning experimental results on FUNSD, XFUND, and MLFUD datasets demonstrate that LayoutPointer significantly outperforms existing state-of-the-art methods in F1 scores for RE tasks (e.g., 5.71% improvement on XFUND using LayoutPointer$_{\text{BASE-X}}$ over LayoutLMv3)[1].

1 Introduction

The Visually-Rich Document Understanding (VRDU) task involves leveraging artificial intelligence to extract and interpret information from digitized or scanned documents (Xu et al., 2020; Cui et al., 2021; Kastanas et al., 2023). Visual Information Extraction (VIE), as a critical component within VRDU, involves two primary sub-tasks: Semantic Entity Recognition (SER) and Relation Extraction (RE). SER focuses on extracting specific entities from documents, while RE aims to identify relations between entities. In recent years, multimodal pre-trained models (Xu et al., 2020; Xu et al., 2021a; Huang et al., 2022; Li et al., 2021b; Appalaraju et al., 2023; Zhang et al., 2020; Yu et al., 2023; Luo et al., 2023; Hong et al., 2022; Wang et al., 2022; Lee et al., 2023), which integrate image, text, and layout information, have demonstrated strong capabilities in the VIE task.

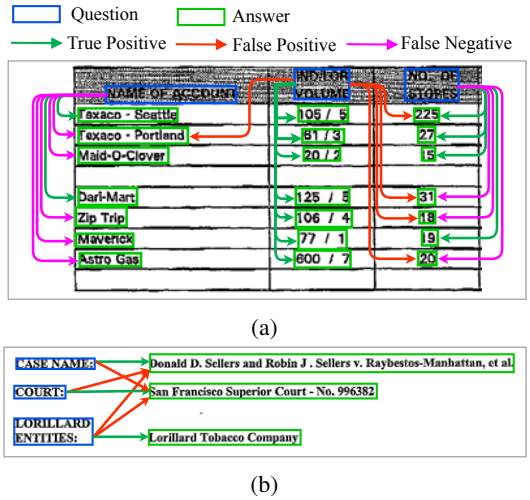
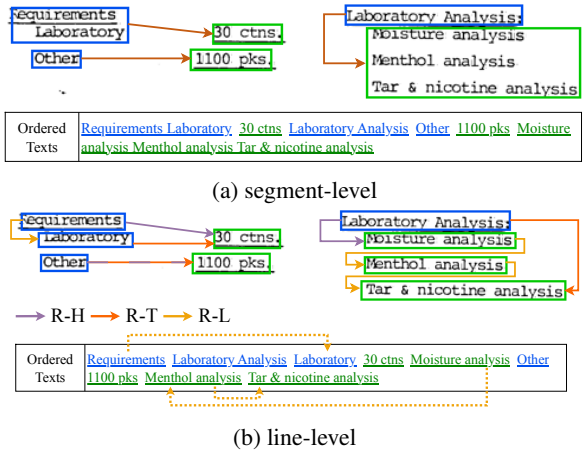


Figure 1: The RE results of LayoutLMv3$_{\text{BASE}}$.

However, the VIE task currently faces two primary challenges. Firstly, most existing models inadequately handle the explicit spatial relationships between entities. We test the capabilities of the state-of-the-art base model LayoutLMv3 (Huang et al., 2022) in the RE task. As shown in Fig. 1a, LayoutLMv3 fails to identify five entity links in the *"NAME OF ACCOUNT"* column of a table. Moreover, Fig. 1b demonstrates that, when processing forms with a left-right structural layout, LayoutLMv3 incorrectly connects entities that are

†Corresponding author.
[1]Available codes and the MLFUD dataset: https://github.com/ThinkSYR/LayoutPointer.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3737–3748
June 16-21, 2024



(a) segment-level

(b) line-level

Figure 2: Examples of document images: (a) Annotation in segment coordinates, (b) Line-annotation after OCR processing.

spatially distant and belong to different rows, such as *"LORILLARD"* and *"Donald"*. These findings highlight the limitations of LayoutLMv3 in perceiving spatial distance and layout patterns between entities. One potential solution involves applying relative distance thresholds along the x and y axes to refine relationship predictions. However, the effectiveness of this solution strongly relies on the specific threshold settings, making it difficult to achieve precise control across different scenarios.

Secondly, the improper input order of tokens makes it difficult for existing methods to extract complete entity pairs from documents with multi-line texts. When extracting text from real-world documents using a PDF parser or OCR, it is common for an entity to be divided across multiple lines. A case in point is illustrated in Fig. 2, where *"Requirements Laboratory"* is split into *"Requirements"* and *"Laboratory"* on separate rows after OCR processing. Tokens within the same entity will be scrambled when sorted by 2D coordinates as coordinates cannot be shared between them. Existing VIE schemes (Xu et al., 2022; Gu et al., 2022) primarily depend on head features of contiguous entities to identify relations. However, the method falls short in reconstructing the correct semantic order of entities, leading to incomplete results.

To address the above two challenges, we propose a Spatial-Context Adaptive Pointer Network, named **LayoutPointer**. LayoutPointer consists of three main components: **(1) Feature Extractor**, leveraging LayoutLMv3 for multimodal feature extraction; **(2) Spatial-Context Adaptive Enhancer**, building on multi-layer Encoders that explicitly introduces 2D relative position information and adaptive spatial constraints through **RC-RoPE** and **Soft Layout Mask** mechanisms in self-attention; **(3) Score Pointer**, generating global score maps for pointers which mark special links between token pairs. As depicted in Fig. 2b, to effectively process the VIE task with multi-line entities, we employ three pointers **{R-H, R-T, R-L}** to represent entity relations, with R-L specifically indicating the semantic order among row entities. This method converts complete entity pairs into specialized cyclic graphs, decoding relations by identifying cycle paths in document graphs.

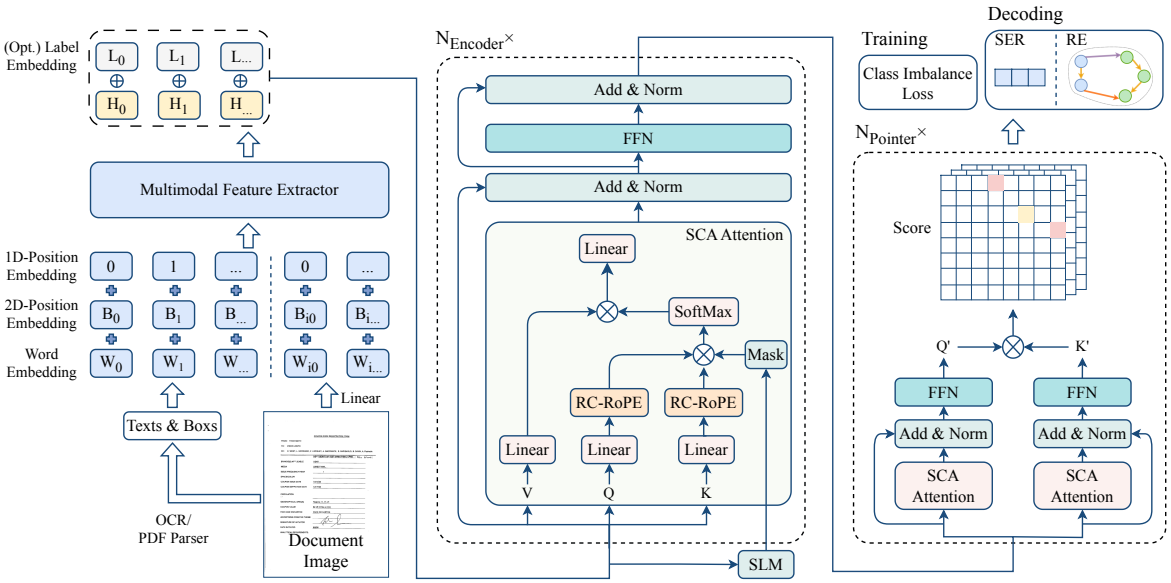
To evaluate the effectiveness of our proposed method, we extract samples from public datasets FUNSD and XFUND and reconstruct a multi-line dataset, named MLFUD. Fine-tuning experimental results reveal that LayoutPointer outperforms current state-of-the-art methods, improving F1 scores by 5.68% over GeoLayoutLM and 5.71% over LayoutLMv3 for the RE task on FUNSD and XFUND, respectively, and significantly improving by 30.29% over LayoutLMv3 on MLFUD.

2 Related Works

Graph-based methods convert text and coordinate information into nodes in a graph structure, with edges representing spatial relationships (Liu et al., 2019; Sun et al., 2021). Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017; Yu et al., 2021; Zhang et al., 2021) and Graph Attention Networks (GANs) (Velickovic et al., 2018; Zhang et al., 2023) are then used to enhance text-visual feature fusion.

End-to-end methods using single-modal image input employ CNNs or Transformers to extract regional image features and text in a unified framework (Zhang et al., 2020; Wang et al., 2021). To further improve the model, pre-training tasks such as text region-level image masking (Yu et al., 2023) are integrated into the system.

Pre-trained multimodal Transformer-based methods combine text, coordinates, and image data, utilizing Transformer-based pre-trained language models for feature extraction (Huang et al., 2022; Luo et al., 2023; Hong et al., 2022; Wang et al., 2022; Appalaraju et al., 2023; Li et al., 2021a; Lee et al., 2023; Gu et al., 2022; Peng et al., 2022). These models are typically fine-tuned via feed-forward networks (FFNs) for SER and bi-affine classifiers for RE (Xu et al., 2022). Differently,



(a) Feature Extractor (b) Spatial-Context Adaptive Enhancer (c) Score Pointer

Figure 3: An overview of LayoutPointer

BROS (Hong et al., 2022) introduces the SPADE Decoder for RE. KVPFormer (Hu et al., 2023) employs a coarse-to-fine answer prediction approach, which identifies each question first and then selects the most likely answer. GeoLayoutLM (Luo et al., 2023) design a Transformer-based relation head with geometric pre-training.

Most models utilize predefined spatial relative position information or 2D position embeddings to handle 2D coordinates. This paper investigates the integration of explicit adaptive spatial information to improve the performance of RE through fine-tuning only. Additionally, we explore methods for extracting relations between multi-line entities.

3 Methodology

3.1 Problem Formulation

Given a visually-rich document D accompanied by an image, comprising n tokens, denoted as $D = \{d_0, d_1, \ldots, d_n\}$, each token d_i consists of a word$_i$ and its corresponding bounding box$_i$. The SER task aims to identify entities within these tokens and classify them into m_{SER} categories, represented as $C = \{c_1, c_2, \ldots, c_{m_{\text{SER}}}\}$. The RE task aims to extract a set of entity pairs $\{(\text{subject}_j, \text{object}_j, r_i), r_i \in \mathcal{R}\}$, where subject$_j$ and object$_j$ represent two related semantic entities and $\mathcal{R}$ represent relation labels. Following LayoutXLM (Xu et al., 2022), this paper primarily focuses on key-value relation extraction.

For relation extraction involving multi-line entities, the target entity pair can be expanded to the relation between multiple joint sub-entities. Each subject$_j$ can be considered as composed of multiple sub-entities (subject$_{j1}, \ldots,$ subject$_{jk}$). Two adjacent sub-entities subject$_{jk-1}$ and subject$_{jk}$ are semantically connected in sequence.

3.2 Model Architecture

Figure 3 presents the overall architecture of LayoutPointer, which consists of three primary components: a Feature Extractor employing LayoutLMv3 for multimodal feature representation, a Spatial-Context Adaptive Enhancer designed to reinforce spatial relationships by integrating RC-RoPE and Soft Layout Mask mechanisms into self-attention, and a Score Pointer for generating global score maps according to feature sequences.

3.2.1 Feature Extractor

We use LayoutLMv3 (Huang et al., 2022) to extract feature representations. This is a pre-trained multimodal Transformer architecture that uses three embedding layers to process inputs, including text, 2D position, and images. Assuming the feature sequence are represented as $X = \{x_i, i = 1, 2, \ldots, n\}$. Following LayoutXLM (Xu et al., 2021b) and SERA (Zhang et al., 2021), in the RE task, the label embedding is concatenated with x_i as prior knowledge:

$$x_i^* = x_i \oplus \text{LabelEmbedding}(c_i), \quad (1)$$

where c_i represents the entity label for the i-th token, and $\oplus$ represents the concatenation operator.

3.2.2 Spatial-Context Adaptive Enhancer

As illustrated in Fig. 3b, the SCA Enhancer consists of multi-layer Transformer Encoders with SCA Attention. Initially, given that SCA Attention consists of H attention heads, the input sequence is passed through linear transformations to yield $Q^h, K^h, V^h \in \mathbb{R}^{n\times d}$:

$$q_i^h = x_i W_Q^h, \quad k_i^h = x_i W_K^h, \quad v_i^h = x_i W_V^h, \tag{2}$$

where $q_i^h, k_i^h \in \mathbb{R}^d$, and $h \in [1, 2, \ldots, H]$. Here, W_Q^h, W_K^h, W_V^h represent three parameter matrices corresponding to the h-th attention head.

RC-RoPE Inspired by Rotatory Position Encoding (RoPE) (Su et al., 2023) and its 2D variant [2], we introduce RC-RoPE specifically for visually-rich documents. The rotation matrix is calculated as follows:

$$\boldsymbol{R}_{\mathrm{x,y}}^{\theta} = \begin{bmatrix} \cos x\theta & -\sin x\theta & 0 & 0 \\ \sin x\theta & \cos x\theta & 0 & 0 \\ 0 & 0 & \cos y\theta & -\sin y\theta \\ 0 & 0 & \sin y\theta & \cos y\theta \end{bmatrix} \tag{3}$$

$$\boldsymbol{R}_{\mathrm{x,y}}^{\Theta} = \operatorname{diag}\left(\boldsymbol{R}_{\mathrm{x,y}}^{\theta_1}, \boldsymbol{R}_{\mathrm{x,y}}^{\theta_2}, \ldots, \boldsymbol{R}_{\mathrm{x,y}}^{\frac{\theta_d}{4}}\right), \tag{4}$$

where the initialization of angles is given by $\Theta = \{\theta_i = 10000^{-4(i-1)/d}, i \in [1, 2, \ldots, d/4]\}$. As shown in Eq. (5), similar to properties of RoPE, the multiplication of 2D rotation matrix naturally integrates 2D relative position.

$$\begin{aligned} F(x, y, q_i^h, k_i^h) &= \frac{1}{\sqrt{d}} \left(R_{x,y}^{\Theta} q_i^h\right)^{\top} \left(R_{x,y}^{\Theta} k_i^h\right) \\ &= \frac{1}{\sqrt{d}} (q_i^h)^{\top} R_{x_k - x_q, y_k - y_q}^{\Theta} (k_i^h). \end{aligned} \tag{5}$$

According to the text box $(x_0^{(i)}, y_0^{(i)}, x_1^{(i)}, y_1^{(i)})$, we apply 2D RoPE to Q^h and K^h to obtain the attention score $A^h \in \mathbb{R}^{n\times n}$. The element $\alpha_{i,j}^h$ is calculated as follows:

$$\alpha_{i,j}^h = \begin{cases} F(x_0, y_0, q_i^h, k_i^h), & \text{if } \mathbb{1}_{\text{odd}}(h) = 1 \\ F(x_1, y_1, q_i^h, k_i^h), & \text{if } \mathbb{1}_{\text{even}}(h) = 1 \end{cases} \tag{6}$$

where $\mathbb{1}_{\text{odd}}(h)$ equals 1 for odd h and $\mathbb{1}_{\text{even}}(h)$ equals 1 for even h. Eq. (6) indicates that RC-RoPE alternately applies 2D RoPE on coordinates of the top-left point (x_0, y_0) and the bottom-right point (x_1, y_1) to Q and K across different attention heads. Aggregating information from various attention heads, the model explicitly encodes 2D spatial positions within tokens.

Soft Layout Mask (SLM) Building upon the 1D relative position decay mechanism (Press et al., 2022; Chi et al., 2022; Chi et al., 2023), we propose Soft Layout Mask (SLM). Initially, we compute the 2D intervals between entities:

$$\Delta_{i,j}^h = \begin{cases} \Delta x_{i,j}, & \text{for } h \in [1, 2, \ldots, \lfloor H/2 \rfloor] \\ \Delta y_{i,j}, & \text{for } h \in [\lfloor H/2 \rfloor + 1, \ldots, H] \end{cases} \tag{7}$$

$\Delta_{i,j}^h$ is defined as a head-specific 2D relative distance. Considering that the width of a text box is usually much larger than the height, we define the x-axis geometric distance as the minimum of distances between the left and right endpoints and midpoints, i.e., $\Delta x_{i,j} = \min(\Delta x_0^{ij}, \Delta x_1^{ij}, \Delta x_{\text{mid}}^{ij})$. The y-axis geometric distance is defined as $\Delta y_{i,j} = \Delta y_{\text{mid}}^{ij}$. Subsequently, we apply Soft Layout Mask to the attention scores:

$$\varphi_{i,j}^h = \alpha_{i,j}^h - \gamma_i^h \log(1 + \beta_i^h \Delta_{i,j}^h), \tag{8}$$

where γ_i^h and β_i^h are parameters controlling the magnitude and rate of distance decay respectively, and are computed based on the input sequence x_i:

$$\gamma_i = \sigma(x_i W_\gamma + b_\gamma), \tag{9}$$

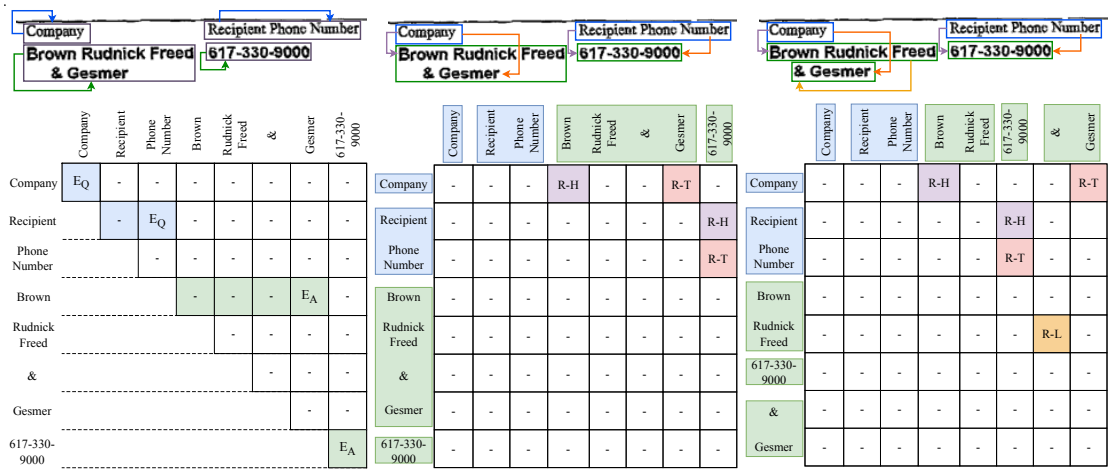
$$\beta_i = \sigma(x_i W_\beta + b_\beta), \tag{10}$$

where $W_\gamma, W_\beta \in \mathbb{R}^{d\times H}$ and $b_\gamma, b_\beta \in \mathbb{R}^H$ represent learnable weights and biases, respectively. The activation function is set as $\sigma(\cdot) = \text{ELU}(\cdot) + 1$ to ensure that $\gamma_i, \beta_i \geq 0$. Eq. (7) to (10) demonstrate how each token's attention awareness is constrained by horizontal or vertical relative distances. The adaptive adjustment of γ and β enhances local attention focus within a controlled range and imposes a higher score penalty on irrelevant information that is layout-distant. Then, we utilize the SoftMax function to calculate the output vector:

$$o_i^h = \sum_j \frac{\exp\left(\varphi_{ij}'\right)}{\sum_k \exp\left(\varphi_{ik}'\right)} v_j^h. \tag{11}$$

Finally, the output of SCA Enhancer is derived through residual connections, normalization, and a Feed-Forward Network (FFN). In this paper, diverging from the traditional Transformer Encoder, we employ RMSNorm (Zhang and Sennrich, 2019) instead of LayerNorm and employ GELU (Hendrycks and Gimpel, 2023) within FFN.

[2] Inroduced at https://spaces.ac.cn/archives/8397.



(a) SER (b) RE (w/o multi-line) (c) RE (with multi-line)

Figure 4: The illustration of pointer schema on SER and RE tasks.

3.2.3 Score Pointer

We employ the pointer to indicate a particular link between tokens and use SCA Attention-based module to compute the global score map for each pointer. Specifically, with the feature sequence $\{p_0, p_1, \ldots, p_n\}$, the Score is calculated based on the token pair (p_i, p_j) as follows:

$$q'_{i,c} = \text{Mapper}^{(Q)}(p_i), \tag{12}$$

$$k'_{j,c} = \text{Mapper}^{(K)}(p_j), \tag{13}$$

$$\text{Score}_u(i, j|c) = q'^T_{i,c} k'_{j,c}, \tag{14}$$

where $\text{Mapper}^{(Q)}$ and $\text{Mapper}^{(K)}$ represent two structurally identical but independent subnetworks. Each network comprises a single layer of SCA Attention with FFN. We define a threshold s_t, such that if $\text{Score}_u(i, j|c) > s_t$, the link **from-i-to-j** is considered as the **pointer** u of class c.

3.3 Pointer Schema with Decoding Algorithm

In the VIE task, we design four special pointers {E, R-H, R-T, R-L} to decode entities with their relations. Figure 4 schematically illustrates these pointers, and their specific meanings and decoding algorithms are outlined below:

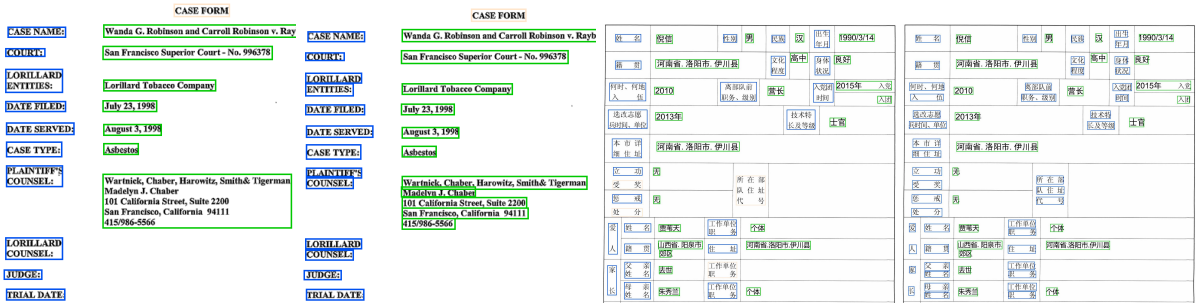
SER: We use the pointer **E** to identify entities, as shown in Fig. 4a, where it represents the link between the start and end tokens of an entity **(head-to-tail)**.

During decoding, we traverse Score to find position pairs (i, j) that satisfy $\text{Score}_E(i, j|c_k) > 0$. The span of tokens from the i-th to the j-th token is then concatenated in order, forming an entity of category c_k.

RE: To identify relations between entities, two types of pointers are employed: R-H and R-T. **R-H** identifies the connection between the start tokens of the subject and object entities **(head-to-head)**, while **R-T** establishes the link between their end tokens **(tail-to-tail)**. A relationship between two entities is confirmed if and only if both links are established. For instance, in Fig. 4b, the term *"company"* is linked with *"Brown"* at the start and with *"Gesmer"* at the end, indicating a key-value relationship between the two entities.

RE with multi-line: We utilize the pointer R-L for restoring the semantic sequence between row entities. **R-L** is designated to identify a unidirectional matching relationship from the end token of one entity to the start of another **(tail-to-head)**. For instance, in Fig. 4c, the entity *"Brown Rudnick Freed & Gesmer"* is split into two lines in multi-line text parsing. In such cases, R-L identifies that *"Brown Rudnick Freed"* and *"& Gesmer"* are semantically connected and constitute parts of the same entity.

We adopt a specialized cycle detection algorithm to decode the RE task. As illustrated in Fig. 4c and 2b, each independent continuous entity can be treated as a node in the graph, and the R-H, R-T, and R-L pointers act as undirected edges. Within this graph, a complete entity can be represented as a path, while the relation is represented by cycles incorporating special edges. During decoding, cycles that pass through R-H and R-T edges are detected by using the depth-first search algorithm. A special cycle is confirmed as a complete entity pair if and only if it contains exactly one R-H and one R-T edge.



(a) (b) (c) (d)

Figure 5: Comparative example images between public datasets and MLFUD. (a) from FUNSD, (c) from XFUND, (b) and (d) from MLFUD.

3.4 Loss Function

During training, we employ the Class Imbalance Loss (Su et al., 2022):

$$\mathcal{L} = \log\left(e^{s_t} + \sum_{(q,k)\in\Omega_{\text{neg}}} e^{s(q,k)}\right) + \log\left(e^{-s_t} + \sum_{(q,k)\in\Omega_{\text{pos}}} e^{-s(q,k)}\right), \tag{15}$$

where q and k represent the indices corresponding to the start or end tokens of entities, and s denotes the score map. The threshold s_t is used to divide target and non-target classes. Ω_{neg} and Ω_{pos} respectively denote the sets of (q, k) where $s(q, k) \leq s_t$ and $s(q, k) > s_t$. Following GlobalPointer, we set $s_t = 0$.

4 Experiments

4.1 Datasets

This paper primarily focuses on SER and RE tasks. We conduct fine-tuning experiments on two public datasets, FUNSD and XFUND. FUNSD, primarily used for understanding the form content in English, comprises 199 scanned document images. XFUND is a multilingual dataset consisting of 1,393 form images across seven languages (Chinese, Japanese, Spanish, French, Italian, German, and Portuguese).

As illustrated in Fig. 5a and 5c, existing public datasets, which typically employ segment-level annotations, overlook semantic order issues between multi-line entities. In light of this, we have chosen the FUNSD and the Chinese dataset XFUND-ZH for re-annotation and reconstruction. Our reconstruction process begins with the utilization of the word positions provided in these datasets. We calculate the midpoint distances between the vertical coordinates of adjacent words. When the midpoint distance between two words is found to be less than the average height of a single character of the document, we merge these words into a single line. Following the initial merging, we employ manual annotation to meticulously correct relationships between entities that span multiple lines.

The newly constructed dataset is named MLFUD. MLFUD contains 398 samples, including 199 in Chinese and 199 in English, divided into 298 training and 100 test samples. It encompasses a total of 26,025 line-level continuous entities with 9,679 key-value pairs, and 1,210 multi-line relations. Fig. 5b and 5d compare the re-constructed annotation samples with their originals, demonstrating that line-level annotations more closely resemble real-world applications.

4.2 Experiment Implementation

4.2.1 Implementation Details

We initialize the parameters of feature extractor using LayoutLMv3$_{\text{BASE}}$ (133M), LayoutLMv3$_{\text{BASE-CHINESE}}$ (133M), and LayoutLMv3$_{\text{LARGE}}$ (368M), resulting in three corresponding versions: LayoutPointer$_{\text{BASE}}$, LayoutPointer$_{\text{LARGE}}$, and LayoutPointer$_{\text{BASE-X}}$. The SCA Enhancer in each model is uniformly composed of 2-layer encoders, where the basic model is set to 12 heads and 768 dimensions, and the large model is set to 16 heads and 1024 dimensions. We use 144-dim label embedding in the RE task. The training details can be found in the appendix A.

Inspired by KVPFormer (Hu et al., 2023) and SERA (Zhang et al., 2021), we adopt a Greedy Key-Value Match (GKVM) strategy, which aims to find the key with the highest positive score for

Methods	#Params	SER			RE		
		Precision	Recall	F1	Precision	Recall	F1
$\text{BROS}_{\text{BASE}}$	110M	81.16	85.02	83.05	-	-	71.46
SERA	-	-	-	-	74.02	77.77	75.83
$\text{LayoutLMv3}_{\text{BASE}}$	133M	-	-	90.29	63.69	85.56	73.02†
$\text{LayoutLMv3}_{\text{BASE}}$+GP†	134M	90.92	**92.19**	91.55	85.52	87.53	86.52
$\text{LayoutPointer}_{\text{BASE}}$	165M (SER) 176M (RE)	**92.85**	91.59	**92.21**	**92.20**	**91.98**	**92.09**
$\text{BROS}_{\text{LARGE}}$	340M	82.81	86.31	84.52	-	-	77.01
GeoLayoutLM	399M	-	-	**92.86**	88.94	89.96	89.45
$\text{LayoutLMv3}_{\text{LARGE}}$	368M	-	-	92.08	75.82	85.45	80.35‡
$\text{LayoutLMv3}_{\text{LARGE}}$+GP†	369M	**92.71**	92.89	92.80	92.87	93.33	93.10
$\text{LayoutPointer}_{\text{LARGE}}$	419M (SER) 439M (RE)	92.44	**93.04**	92.74	**96.21**	**94.07**	**95.13**
KVPFormer	-	-	-	-	94.06	87.88	90.86
$\text{LayoutPointer}_{\text{BASE}}$ +GKVM	176M (RE)	-	-	-	**96.89**	**92.35**	**94.56**

Table 1: Experimental results of both SER and RE tasks on FUNSD dataset. "†" means the results are re-implemented by us and "‡" means the results are from the GeoLayoutLM paper (Luo et al., 2023).

each value as the result:

$$\text{Score}(i,j) = \begin{cases} 1, & \text{if } i = \underset{i}{\text{argmax}}\ Score(i,j) \\ -1, & \text{otherwise} \end{cases} \tag{16}$$

We additionally discuss the actual influence of GKVM on LayoutPointer in 4.3.

4.2.2 Baselines

We compare our proposed method with various multimodal frameworks, including BROS (Hong et al., 2022), SERA (Zhang et al., 2021), GeoLayoutLM (Luo et al., 2023), KVPFormer (Hu et al., 2023), LayoutXLM (Xu et al., 2022), LiLT (Wang et al., 2022), and ECN (Déjean et al., 2022). To more clearly demonstrate the improvements brought by LayoutPointer, we implement two baselines for comparison:

- **LayoutLMv3**. LayoutXLM employs FFNs for SER and a bi-affine classifier for RE. Therefore, we continue to use these two approaches on Layoutlmv3 for comparison.
- **LayoutLMv3+GP (GlobalPointer)**. As a baseline comparison model for pointer networks, we leverage LayoutLMv3 as the feature extractor and use GlobalPointer (Su et al., 2022) to implement SER and RE with label embedding.

4.3 Comparison with the SOTAs

FUNSD: The results from FUNSD, as shown in Tab. 1, reveal that both $\text{LayoutLMv3}_{\text{BASE}}$+GP and LayoutPointer surpassed LayoutLMv3 for VIE, underscoring the effectiveness of pointer networks. Specifically, in the SER task, both models produced similar results, which is attributed to the fact that tokens within an entity typically share the same segment coordinates, limiting the role of spatial context information. In contrast, LayoutPointer showed significant superiority in the RE task. According to F1 scores, $\text{LayoutPointer}_{\text{BASE}}$ achieves a 5.57% improvement over $\text{LayoutLMv3}_{\text{BASE}}$+GP. Compared to the previous state-of-the-art method, GeoLayoutLM, $\text{LayoutPointer}_{\text{LARGE}}$ shows a 5.68% improvement, and even $\text{LayoutPointer}_{\text{BASE}}$ demonstrates a 2.64% improvement. The case study is given in appendix B.

Additionally, we implement $\text{LayoutPointer}_{\text{BASE}}$ with GKVM strategy to facilitate comparison with KVPFormer. It can be seen that $\text{LayoutPointer}_{\text{BASE}}$+GKVM improves by 3.7% compared to KVPFormer, and improves by 2.47% compared to before using this strategy.

XFUND We conduct multilingual fine-tuning experiments of the RE task on XFUND, with the results presented in Tab.2. LayoutPointer achieves state-of-the-art performance across almost all lan-

Methods	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
LayoutXLM$_{\text{BASE}}$	66.71	82.41	81.42	8104	82.21	83.10	78.54	70.44	78.23
LiLT[InfoXLM$_{\text{BASE}}$]	74.07	84.71	83.45	93.35	84.66	84.58	78.78	76.43	81.25
ECN	89.27	90.82	86.67	89.66	92.22	86.08	85.72	81.64	87.76
LayoutLMv3$_{\text{BASE-X}}$†	90.94	93.63	89.76	93.11	91.42	90.58	87.05	86.86	90.42
LayoutLMv3$_{\text{BASE-X}}$+GP†	95.00	95.11	92.61	93.48	91.25	93.27	87.49	88.42	92.08
LayoutPointer$_{\text{BASE-X}}$	**97.23**	**97.70**	**97.97**	**96.95**	**96.15**	**95.56**	**94.74**	**93.74**	**96.13**
KVPFormer	95.70	94.27	94.23	95.23	**97.19**	94.11	92.41	**92.19**	94.42
LayoutPointer$_{\text{BASE}}$ +GKVM	**96.91**	**97.59**	**97.91**	**96.42**	96.11	**96.22**	**95.05**	92.10	**96.04**

Table 2: Multilingual fine-tuning F1 score of the RE task on XFUND (fine-tuning on 8 languages all, testing on X).

Methods	entity pairs	R-H	R-T	R-L
LayoutLMv3$_{\text{BASE-X}}$	50.83	80.70	79.83	79.13
LayoutLMv3$_{\text{BASE-X}}$+GP	77.41	88.71	88.27	**88.96**
LayoutPointer$_{\text{BASE-X}}$	**81.12**	**92.39**	**91.81**	88.87

Table 3: F1 scores of the RE task with multi-line entities on MLFUD

guages and the average F1 scores of LayoutPointer show an improvement of 5.71% over LayoutLMv3. In addition, LayoutPointer$_{\text{BASE-X}}$+GKVM achieves almost the same effect as without the strategy, registering a 1.62% improvement in average F1 score compared to KVPFormer.

MLFUD In the RE task involving multi-line entities, we evaluate our proposed method using the self-constructed dataset MLFUD, with results presented in Tab. 3. These results indicate that pointer networks are particularly effective for extracting complete entity pairs and LayoutPointer$_{\text{BASE-X}}$ exhibits a 3.71% improvement over LayoutLMv3$_{\text{BASE-X}}$+GP. Additionally, we test F1 scores on the three pointers {R-H, R-T, R-L}. LayoutPointer$_{\text{BASE-X}}$ demonstrates a more significant advantage in predicting head-to-head and tail-to-tail links between entities than LayoutLMv3$_{\text{BASE-X}}$+GP.

4.4 Ablation Study

Effects of RC-RoPE: The ablation study on position encoding types is detailed in Tab. 4. Comparative analysis of #1a and #1b with the F1 score, shows that 1D RoPE has a marginal impact on the RE task. Conversely, as indicated in #2a, RC-RoPE yields a 1.31% improvement on FUNSD and 1.02% on XFUND. This improvement in both precision and recall demonstrates the significance of explicit spatial information for the RE task.

#	RoPE	SLM	FUNSD			XFUND		
			P	R	F1	P	R	F1
1a	×	×	89.95	82.84	86.25	95.52	90.85	93.11
1b	1D	×	90.86	80.99	85.64	96.60	90.12	93.24
2a	RC	×	92.10	83.46	87.56	96.79	91.64	94.13
2b	×	✓	88.31	87.65	87.98	95.96	93.29	94.60
3a	RC	✓	**92.20**	**91.98**	**92.09**	**96.82**	**95.46**	**96.13**

Table 4: Ablation study on different components of SCA Attention. "1D" means original RoPE while "RC" means "RC-RoPE".

#	Enhancer	Output Pointer	FUNSD	XFUND
1a	×	GP	86.52	92.08
1b	1-Layer	GP	86.88	93.52
1c	2-Layer	GP	89.13	94.01
2a	×	SP w/o SCAA	86.66	92.25
2b	1-Layer	SP w/o SCAA	87.94	94.61
2c	2-Layer	SP w/o SCAA	89.15	95.29
3a	×	SP	87.74	94.77
3b	1-Layer	SP	90.59	95.19
3c	2-Layer	SP	**92.09**	**96.13**

Table 5: Ablation study on the model structure. "GP" means "GlobalPointer". "SP w/o SCAA" means "Score Pointer without SCA Attention"

Effects of Soft Layout Mask (SLM): The ablation study on SLM is detailed in Tab. 4. A comparison of #1a and #2b with the F1 score, indicates that SLM resulted in an improvement of 1.73% on FUNSD and 1.49% on XFUND. Moreover, SLM's introduction of adaptive spatial constraints enhances focused attention between tokens, significantly boosting recall. Further comparisons from #2a, #2b, and #3a reveal that the combination of RC-RoPE and SLM yields improvements both on FUNSD and XFUND. This indicates that RC-RoPE and SLM function synergistically, enhancing each other's effectiveness.

Effects of the Structure of SCA Enhancer and Score Pointer: Our evaluation, as detailed in Tab. 5, examines the impact of various configurations of the model structure. The analysis between #1a-2c and #3a-3c indicates that both the addition of single or double layers of SCA Enhancer and the incorporation of SCA Attention in the Score Pointer substantially enhance the F1 score. Notably, the #3c config, which combines a 2-layer SCA Enhancer with SCA Attention in Score Pointer, demonstrates the most significant improvement.

5 Conclusion

In this paper, we propose LayoutPointer, a Spatial-Context Adaptive pointer network. By incorporating RC-RoPE and SLM mechanisms into self-attention, LayoutPointer explicitly enhances spatial-context relations between entities. Moreover, we propose a universal solution to solve RE tasks with multi-line entities, focusing on predicting the semantic order of entities to construct complete entity pairs, with the multi-line dataset MLFUD constructed. Experimental results on XFUND, FUNSD, and MLFUD datasets demonstrate the superior performance of LayoutPointer in RE tasks.

Limitations

In this paper, we leverage the spatial information of entities to assist the model in identifying relations between entities. However, we have not fully considered the utilization of image information. Elements like table lines and special fonts in images can also significantly enhance the recognition of entity relationships. Future work will aim to integrate such image information to enhance the overall effectiveness of the VIE task.

Furthermore, we provide a universal solution for relation extraction among multi-line entities but have not tested its end-to-end effectiveness on real-world digitized images. Practical challenges such as distortions, partial obstructions, and other variabilities in form images, including invoices and identity documents, can impede information extraction performance. In the future, we plan to explore the potential application of our approach to more realistic data. Furthermore, we plan to investigate the applicability of LayoutPointer's core components in more tasks, including document layout analysis and visual question answering.

Acknowledgments

References

Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R. Manmatha. 2023. DocFormerv2: Local Features for Document Understanding. ArXiv:2306.01733 [cs].

Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. 2022. Kerple: Kernelized relative positional embedding for length extrapolation. *Advances in Neural Information Processing Systems*, 35:8386–8399.

Ta-Chung Chi, Ting-Han Fan, Alexander Rudnicky, and Peter Ramadge. 2023. Dissecting Transformer Length Extrapolation via the Lens of Receptive Field Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13522–13537, Toronto, Canada. Association for Computational Linguistics.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document AI: Benchmarks, Models and Applications. ArXiv:2111.08609 [cs].

Hervé Déjean, Stéphane Clinchant, and Jean-Luc Meunier. 2022. LayoutXLM vs. GNN: An Empirical Evaluation of Relation Extraction for Documents. ArXiv:2206.10304 [cs].

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4573–4582, New Orleans, LA, USA. IEEE.

Dan Hendrycks and Kevin Gimpel. 2023. Gaussian Error Linear Units (GELUs). ArXiv:1606.08415 [cs].

Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10767–10775.

Kai Hu, Zhuoyuan Wu, Zhuoyao Zhong, Weihong Lin, Lei Sun, and Qiang Huo. 2023. A Question-Answering Approach to Key Value Pair Extraction from Form-Like Document Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12899–12906.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, Lisboa Portugal. ACM.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 1–6, Sydney, Australia. IEEE.

Sotirios Kastanas, Shaomu Tan, and Yi He. 2023. Document AI: A Comparative Study of Transformer-Based, Graph-Based Models, and Convolutional Neural Networks For Document Layout Analysis. ArXiv:2308.15517 [cs].

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Chen-Yu Lee, Chun-Liang Li, Hao Zhang, Timothy Dozat, Vincent Perot, Guolong Su, Xiang Zhang, Kihyuk Sohn, Nikolay Glushnev, Renshen Wang, Joshua Ainslie, Shangbang Long, Siyang Qin, Yasuhisa Fujii, Nan Hua, and Tomas Pfister. 2023. FormNetV2: Multimodal Graph Contrastive Learning for Form Document Information Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9011–9026, Toronto, Canada. Association for Computational Linguistics.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021a. StructuralLM: Structural Pre-training for Form Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6309–6318, Online. Association for Computational Linguistics.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021b. SelfDoc: Self-Supervised Document Representation Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5648–5656, Nashville, TN, USA. IEEE.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph Convolution for Multimodal Information Extraction from Visually Rich Documents. In *Proceedings of the 2019 Conference of the North*, pages 32–39, Minneapolis - Minnesota. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. 2023. GeoLayoutLM: Geometric Pre-training for Visual Information Extraction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7092–7101, Vancouver, BC, Canada. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Yuhui Cao, Weichong Yin, Yongfeng Chen, Yin Zhang, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. ERNIE-Layout: Layout Knowledge Enhanced Pre-training for Visually-rich Document Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3744–3756, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. ArXiv:2108.12409 [cs].

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, page 127063.

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. ArXiv:2208.03054 [cs].

Hongbin Sun, Zhanghui Kuang, Xiaoyu Yue, Chenhao Lin, and Wayne Zhang. 2021. Spatial Dual-Modality Graph Reasoning for Key Information Extraction. ArXiv:2103.14470 [cs].

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7747–7757, Dublin, Ireland. Association for Computational Linguistics.

Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021. Towards Robust Visual Information Extraction in Real World: New Dataset and Novel Solution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):2738–2745.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021a. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, Virtual Event CA USA. ACM.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021b. LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding. ArXiv:2104.08836 [cs].

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2022. XFUND: A Benchmark Dataset for Multilingual Visually Rich Form Understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3214–3224, Dublin, Ireland. Association for Computational Linguistics.

Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2021. PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4363–4370, Milan, Italy. IEEE.

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. ArXiv:2303.00289 [cs].

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. 2020. TRIE: End-to-End Text Reading and Information Extraction for Document Understanding. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1413–1422, Seattle WA USA. ACM.

Yue Zhang, Zhang Bo, Rui Wang, Junjie Cao, Chen Li, and Zuyi Bao. 2021. Entity Relation Extraction as Dependency Parsing in Visually Rich Documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2759–2768, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. 2023. Multimodal Pre-Training Based on Graph Attention Network for Document Understanding. *IEEE Transactions on Multimedia*, 25:6743–6755.

A More Details of Experiments

A.1 More Details of FUNSD and XFUND

FUNSD (Jaume et al., 2019), primarily used for understanding the form content in English, comprises 199 scanned document images encompassing 9,743 entities and 2,399 key-value pairs. It has 149 training examples and 50 testing examples. XFUND (Xu et al., 2022) is a multilingual dataset consisting of 1,393 form images across seven languages (Chinese, Japanese, Spanish, French, Italian, German, Portuguese). Each language of XFUND has 199 examples including 149 training examples and 50 testing examples. Along with FUNSD, it forms an eight-language dataset, containing 106,755 entities with 32,373 key-value pairs.

A.2 Hyperparameters

All experiments are implemented using the PyTorch (Paszke et al., 2019) framework on a single NVIDIA 3090 GPU. We use the AdamW (Loshchilov and Hutter, 2019) optimizer. Detailed hyperparameter settings are reported in Tab. 6. On XFUND, following LayoutXLM for multilingual fine-tuning, $\text{LayoutPointer}_{\text{BASE-X}}$ is trained on the dataset with 8 languages, with separate testing conducted on the test set of each language.

B Case study

To better illustrate the effectiveness of LayoutPointer in enhancing spatial-context information, we conducted a case study on the RE task using the FUNSD dataset. As shown in Fig. 6, when analyzing tables or forms with a top-bottom structure, LayoutLMv3 produces some missed or incorrect links and lacks consistency. In contrast, LayoutPointer exhibits higher sensitivity to the spatial relationships between entities, correctly predicting almost all links. This indicates that LayoutPointer adheres more closely to adaptive spatial constraints, enabling it to accurately predict entity relations.

LayoutPointer version	FE-LayoutLMv3 version	Num of Attn. heads	Hidden size	Training Dataset	Batch size	Initial Learning Rate		Training steps	Warm-up ratio
						FE	Enhancer and SP		
BASE	BASE	12	768	FUNSD	4	3e-5	1e-4	3000	0.1
LARGE	LARGE	16	1024	FUNSD	2	1e-5	1e-4	6000	0.1
BASE-X	BASE-CHINESE	12	768	XFUND	4	3e-5	1e-4	22000	0.1
				MLFUD	4	3e-5	1e-4	8000	0.1

Table 6: Model settings and hyperparameters for different datasets. "FE" means "Feature Extractor". "SP" means "Score Pointer"

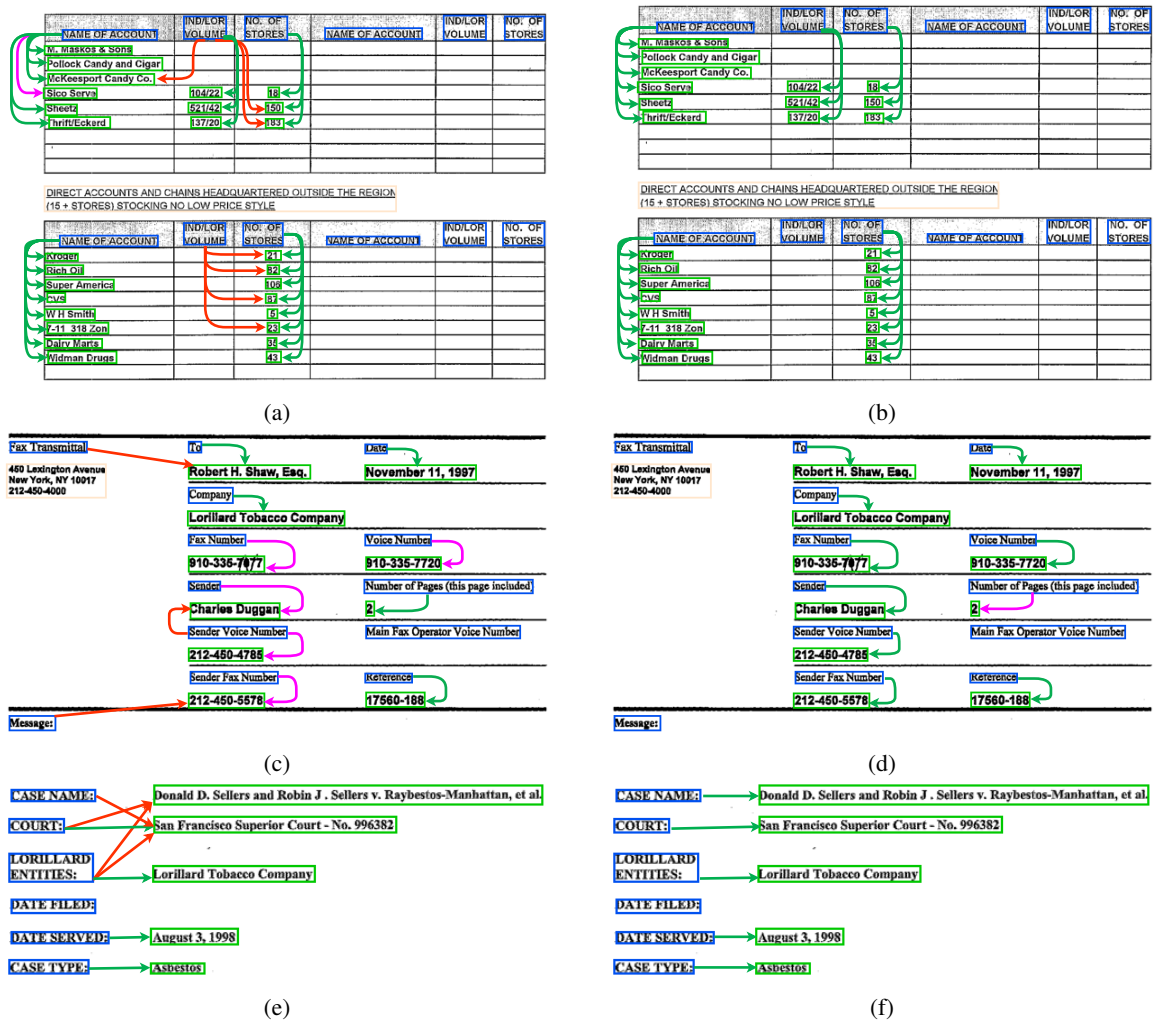


Figure 6: RE Case Study on FUNSD: (a), (c) and (e) from LayoutLMv3, (b), (d) and (f) from LayoutPointer.