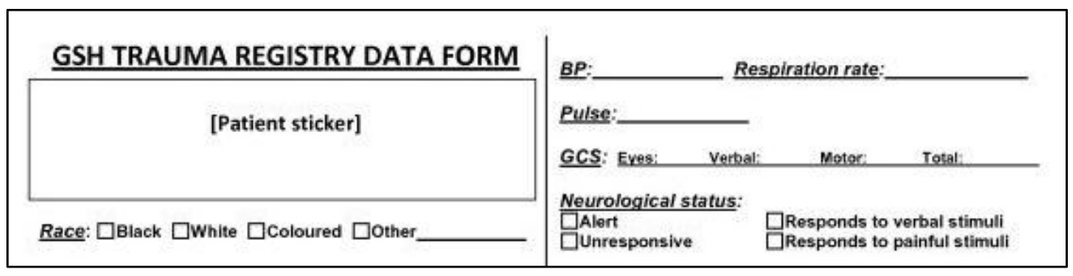
# Enhancing Visually-Rich Document Understanding via Layout Structure Modeling

Qiwei Li†, Zuchao Li†*, Xiantao Cai*, Bo Du and Hai Zhao

## Limiations of Previus Work

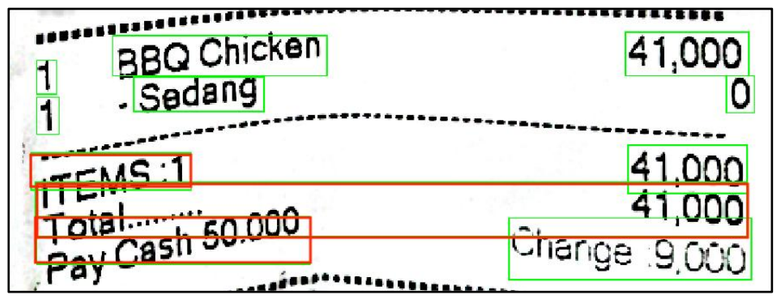
**1. Mismatch between the raw order and proper understanding order.**



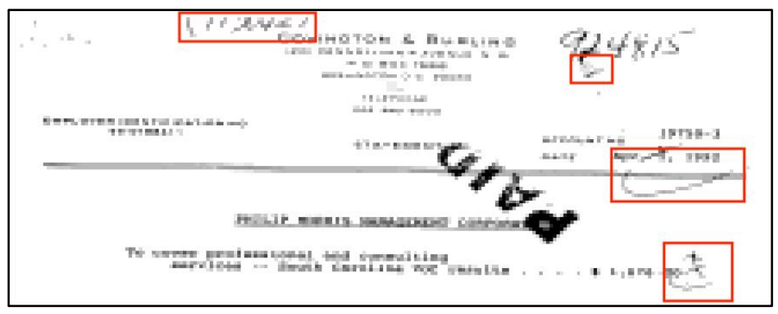

**(a) Raw Order:** *GSH TRAUMA REGISTERY DATA FORM BP: Respiration rate: [Patient sticker] Pulse: GCS: Eyes: Verbal: Motor: Total: Neurological status: Race: Black White Coloured Other Alert Responds to verbal stimuli Unresponsive Responds to painful stimuli*

**(b) Proper Order:** *GSH TRAUMA REGISTERY DATA FORM [Patient sticker] Race: Black White Coloured Other BP: Respiration rate: Pulse: GCS: Eyes: Verbal: Motor: Total: Neurological status: Alert Responds to verbal stimuli Unresponsive Responds to painful stimuli*

**2. Inadequate disclosure of document structure by visual information.**




**(a) Faulty Text Box**



**(b) Interference Information in Resized Image**

## GraphLayoutLM: Layout Language Model Enhanced by Layout Struture Graph

**1. Layout graph modeling.**

- **Hierarchical relationship**

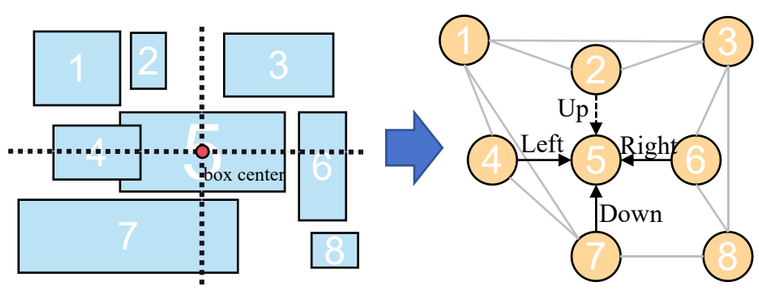
Parent node identification rule of region $P$:

$$L(n_i) = \begin{cases} 1, & box_i = Top(P)\ and\ box_i = Left(P) \\ 0, & otherwise \end{cases}$$

Layout Hierarchical subtree construction of region $P$:

$$T_p = < n_p, \mathrm{C}_p n_p, Parent - Child >$$

- **Positional relationship**




- **Layout structure graph construction**

**Hierarchical relationships** are used to build the layout structure tree $T_g$.

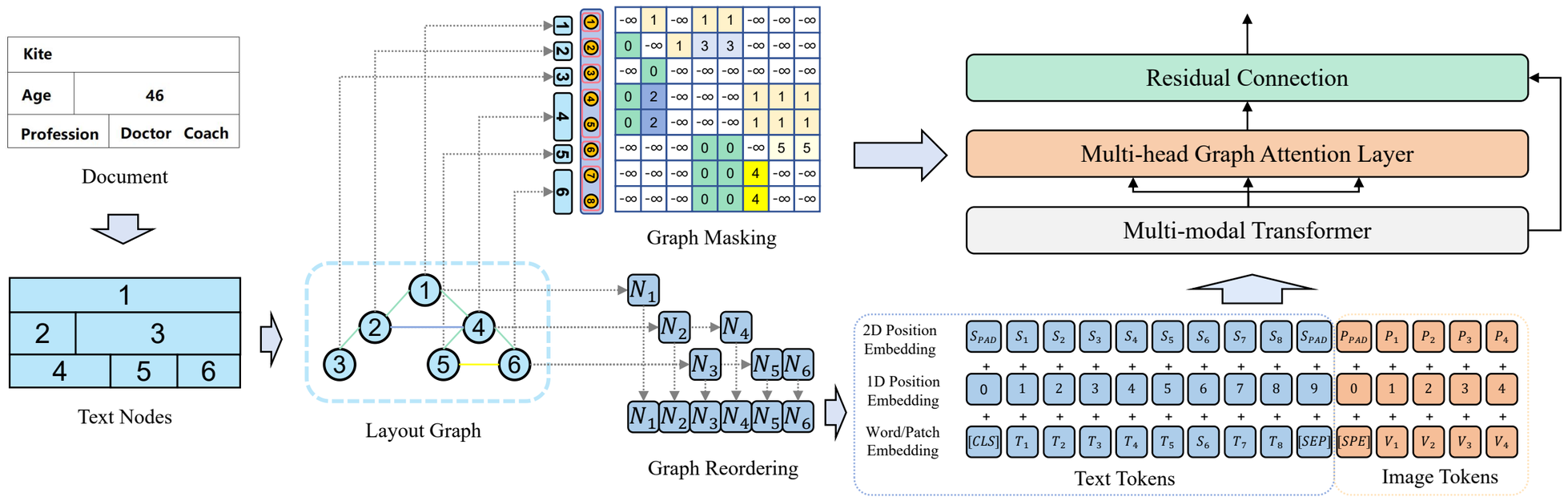
**Positional relationships** are used to enrich the tree into a layout structure graph $G$.

Representation of $G$:

$$G = T_g \cup \{< n_i, n_j, rel(n_i, n_j) > | n_i, n_j \in T_g.sibling\}$$

**2. The architecture of GraphLayoutLM.**

- **GraphLayoutLM utilizes document layout structures graph through Graph Reordering and Graph Masking strategies.**




## Experiment Results

**1. Experiment results of SER task on English datasets**

| Model | Parameters | Modality | FUNSD Precision | FUNSD Recall | FUNSD F1 | CORD Precision | CORD Recall | CORD F1 |
|---|---|---|---|---|---|---|---|---|
| $\text{BERT}_{\text{BASE}}$ | 110M | T | 54.69 | 67.10 | 60.26 | 88.33 | 91.07 | 89.68 |
| $\text{RoBERTa}_{\text{BASE}}$ | 125M | T | 63.49 | 69.75 | 66.48 | - | - | 93.54 |
| $\text{BROS}_{\text{BASE}}$ | 110M | T+L | 81.16 | 85.02 | 83.05 | 95.58 | 95.14 | 95.36 |
| $\text{LayoutLM}_{\text{BASE}}$ | 160M | T+L+I | 76.77 | 81.95 | 79.27 | 94.37 | 95.08 | 94.72 |
| $\text{XYLayoutLM}_{\text{BASE}}$ | - | T+L+I | - | - | 83.35 | - | - | - |
| $\text{LayoutLMv2}_{\text{BASE}}$ | 200M | T+L+I | 80.29 | 85.37 | 82.76 | 94.53 | 95.39 | 94.95 |
| $\text{DocFormer}_{\text{BASE}}$ | 183M | T+L+I | 80.76 | 86.09 | 83.34 | 96.52 | 96.14 | 96.33 |
| $\text{ERNIE-Layout}_{\text{BASE}}$ | - | T+L+I | - | - | 90.28 | - | - | 96.61 |
| $\text{LayoutLMv3}_{\text{BASE}}$ | 133M | T+L+I | - | - | 90.29 | - | - | 96.56 |
| $\text{LayoutLMv3}_{\text{BASE}}{}^{\dagger}$ | 133M | T+L+I | 90.82 | 91.55 | 91.19 | 96.35 | 96.71 | 96.53 |
| **$\text{GraphLayoutLM}_{\text{BASE}}$ (Ours)** | 135M | T+L+I+G | **92.46** | **93.85** | **93.15** | **97.02** | **97.53** | **97.28** |
| $\text{BERT}_{\text{LARGE}}$ | 340M | T | 61.13 | 70.85 | 65.63 | 88.86 | 91.68 | 90.25 |
| $\text{RoBERTa}_{\text{LARGE}}$ | 355M | T | 67.80 | 73.91 | 70.72 | - | - | 93.80 |
| $\text{LayoutLM}_{\text{LARGE}}$ | 343M | T+L | 75.96 | 82.19 | 78.95 | 94.32 | 95.54 | 94.93 |
| $\text{BROS}_{\text{LARGE}}$ | 340M | T+L | 82.81 | 86.31 | 84.52 | - | - | 97.28 |
| $\text{StructuralLM}_{\text{LARGE}}$ | 355M | T+L | 83.52 | 86.81 | 85.14 | - | - | - |
| $\text{LayoutLMv2}_{\text{LARGE}}$ | 426M | T+L+I | 83.24 | 85.19 | 84.20 | 95.65 | 96.37 | 96.01 |
| $\text{DocFormer}_{\text{LARGE}}$ | 536M | T+L+I | 82.29 | 86.94 | 84.55 | 97.25 | 96.74 | 96.99 |
| $\text{ERNIE-Layout}_{\text{LARGE}}$ | - | T+L+I | - | - | 93.12 | - | - | 97.21 |
| $\text{LayoutLMv3}_{\text{LARGE}}$ | 368M | T+L+I | - | - | 92.08 | - | - | 97.46 |
| $\text{LayoutLMv3}_{\text{LARGE}}{}^{\dagger}$ | 368M | T+L+I | 91.51 | 92.70 | 92.10 | 97.45 | 97.52 | 97.49 |
| **$\text{GraphLayoutLM}_{\text{LARGE}}$ (Ours)** | 372M | T+L+I+G | **94.49** | **94.30** | **94.39** | **97.75** | **97.75** | **97.75** |

**2. Experiment results of SER task on Chinese datasets**

| Model | Modality | XFUND F1 |
|---|---|---|
| $\text{XLM-RoBERTa}_{\text{BASE}}$ | T | 87.74 |
| $\text{XLM-RoBERTa}_{\text{LARGE}}$ | T | 89.25 |
| $\text{LayoutXLM}_{\text{BASE}}$ | T+L+I | 89.24 |
| $\text{LayoutXLM}_{\text{LARGE}}$ | T+L+I | 91.61 |
| $\text{XYLayoutLM}_{\text{BASE}}$ | T+L+I | 91.76 |
| $\text{ERNIE-LayoutX}_{\text{BASE}}{}^{\ddagger}$ | T+L+I | 88.58 |
| $\text{LayoutLMv3-Chinese}_{\text{BASE}}{}^{\ddagger}$ | T+L+I | 92.02 |
| $\text{LayoutLMv3-Chinese}_{\text{BASE}}{}^{\dagger}$ | T+L+I | 91.82 |
| **$\text{GraphLayoutLM-Chinese}_{\text{BASE}}$(Ours)** | T+L+I+G | **93.56** |

**3. Ablation Study**

| Dataset | Graph Reorder | Graph Mask | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| FUNSD | ✗ | ✗ | 84.76 | 90.82 | 91.55 | 91.19 |
| | ✓ | ✗ | 85.70 | 92.36 | 93.15 | 92.75 |
| | ✗ | ✓ | 86.75 | 91.73 | 92.05 | 91.89 |
| | ✓ | ✓ | 88.39 | 92.46 | 93.85 | 93.15 |
| CORD | ✗ | ✗ | 97.11 | 96.35 | 96.71 | 96.53 |
| | ✓ | ✗ | 97.33 | 96.79 | 97.16 | 96.97 |
| | ✗ | ✓ | 97.88 | 96.94 | 97.23 | 97.09 |
| | ✓ | ✓ | 98.01 | 97.02 | 97.53 | 97.28 |
| XFUND | ✗ | ✗ | 85.87 | 89.79 | 93.94 | 91.82 |
| | ✓ | ✗ | 85.61 | 89.98 | 94.37 | 92.12 |
| | ✗ | ✓ | 90.88 | 91.58 | 94.43 | 92.99 |
| | ✓ | ✓ | 91.19 | 91.80 | 95.38 | 93.56 |