SRFUND: A Multi-Granularity Hierarchical Structure Reconstruction Benchmark in Form Understanding

Jiefeng Ma[1]   Yan Wang[1]   Chenyu Liu[2]   Jun Du[1*]   Yu Hu[1]
Zhenrong Zhang[1]   Pengfei Hu[1]   Qing Wang[1]   Jianshu Zhang[2]

[1]University of Science and Technology of China, Hefei, China
[2]iFLYTEK, Hefei, China

{jfma, yanwangsa, zzr666, hudeyouxiang}@mail.ustc.edu.cn,
{jundu, yuhu2, qingwang2}@ustc.edu.cn, {cyliu7, jszhang6}@iflytek.com

Abstract

Accurate identification and organizing of textual content is crucial for the automation of document processing in the field of form understanding. Existing datasets, such as FUNSD and XFUND, support entity classification and relationship prediction tasks but are typically limited to local and entity-level annotations. This limitation overlooks the hierarchically structured representation of documents, constraining a comprehensive understanding of complex forms. To address this issue, we present the SRFUND, a hierarchically structured multi-task form understanding benchmark. SRFUND provides refined annotations on top of the original FUNSD and XFUND datasets, encompassing five tasks: (1) **word to text-line merging**, (2) **text-line to entity merging**, (3) **entity category classification**, (4) **item table localization**, and (5) **entity-based full-document hierarchical structure recovery**. We meticulously supplemented the original dataset with missing annotations at various levels of granularity and added detailed annotations for multi-item table regions within the forms. Additionally, we introduce global hierarchical structure dependencies for entity relation prediction tasks, surpassing traditional local key-value associations. The SRFUND dataset includes eight languages including *English, Chinese, Japanese, German, French, Spanish, Italian, and Portuguese*, making it a powerful tool for understanding cross-lingual forms. Extensive experimental results demonstrate that the SRFUND dataset presents new challenges and significant opportunities in handling diverse layouts and global hierarchical structures of forms, thus providing deep insights into the field of form understanding. The original data set and implementations of the baseline methods are available at https://sprateam-ustc.github.io/SRFUND.

1 Introduction

In the United States, billions of individuals and businesses submit tax returns annually,[2] and globally, hundreds of billions of parcels are distributed each year,[3] most of which are accompanied by invoices and delivery notes. Although these documents vary in format, they are all considered forms, which serve as crucial information mediums widely used in global information and merchandise exchange. Compared to storage formats like camera-captured images or scanned documents, digitizing original forms into structured text aids in reducing storage space and facilitates information dissemination

---

*Corresponding author.

[2]https://www.irs.gov/pub/irs-pdf/p55b.pdf

[3]https://www.pitneybowes.com/content/dam/pitneybowes/us/en/shipping-index/
23-mktc-03596-2023_global_parcel_shipping_index_ebook-web.pdf

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

(a) Word level	(b) Text-line level	(c) Entity level	(d) Item table level

(e) Overall form structure based on entities.

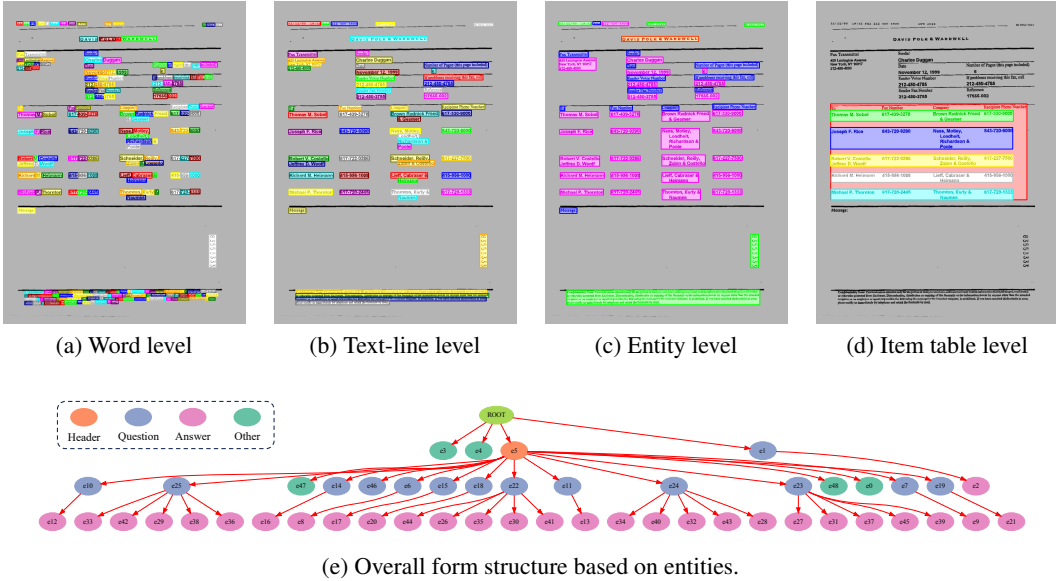



Figure 1: Multiple granularity of annotations and supported tasks on SRFUND. Different colors are used to indicate various element annotations, with subfigure c showing yellow, blue, pink, and green representing four different types of entities: Header, Question, Answer, and Other, respectively. Subfigure e displays the document's entity-level structural relationships based on the entities in subfigure c, where e_i denotes the i-th entity in the document. The colors of different nodes indicate different entity types, with the color-to-type correspondence matching that in subfigure c.

[18]. Consequently, there has been a growing practical demand in recent years for understanding information within forms, including both textual content and document structures across various layouts and languages. With the rapid development of document processing technologies, significant progress has been made in the field of form understanding [22, 40, 45], along with the establishment of a series of benchmark datasets [16, 17, 33, 37, 41, 44]. However, none of these existing datasets have established the global and hierarchical structural dependencies considering all elements at different granularity, including words, text lines, and entities within the forms.

To enhance the applicability of form understanding tasks in hierarchical structure recovery, we introduce the SRFUND, a multilingual form structure reconstruction dataset. The SRFUND dataset comprises 1,592 form images across eight languages, with each language contributing 199 images. As illustrated in Figure 1, each form image is manually annotated with the locations and text contents of each word, text line, and entity. After identifying each independent entity, we categorize these entities into four classes including *Header, Question, Answer*, and *Other*, which is consistent with the definitions of the FUNSD data set [17]. Moreover, all entities in the form are annotated with their hierarchical dependencies, allowing us to reconstruct the global form structure. For the multi-item table regions frequently found in forms, we have specifically annotated the positions of these tables, including their table headers, and grouped each line item within these tables individually. The refined annotations of SRFUND support the evaluation of form structure reconstruction tasks at different granularities. We conducted benchmark tests on several tasks using representative methods from three categories: vision-only, language-only, and multi-modal approaches. These tasks include (1) word to text-line merging, (2) text-line to entity merging, (3) entity category classification, (4) item table localization, and (5) entity-based full-document hierarchical structure recovery. Detailed experimental settings and results are presented in Sec. 4.

2 Related Work

Prior research has divided document structure tasks into two main categories: physical layout analysis and logical structure analysis [13, 30]. The former refers to the physical locations of various regions within a document image, while the latter aims to understand their functional roles and relationships.

2

In this chapter, we will first review the work related to physical and logical structure analysis. Additionally, we will introduce common benchmarks widely used in form understanding tasks.

2.1 Document physical layout analysis

Earlier document layout analysis methods can be classified into two main categories. Algorithms employing a bottom-up strategy start from the finest elements of the document and iteratively merge these elements based on rules or clustering algorithms to create larger and more unified regions [20, 31]. Conversely, top-down strategies begin from the entire document image and use histogram analysis or whitespace refinement methods to segment it into increasingly smaller regions [12, 46]. With the advancement of deep learning technology, several approaches have been proposed to address the document layout analysis challenges in more complex scenarios. These approaches can generally be divided into two categories. Detection-based approaches follow the route of object detection in computer vision, treating different elements within a document as distinct detection targets [24, 32, 35]. Models such as Cascade-RCNN [2], YOLOX [10], and DETR [3] are used to directly predict the positions of various elements in the document images. On the other hand, methods based on instance segmentation employ frameworks used for instance segmentation in natural scene images to segment areas within documents [23, 42, 43]. For example, FCN [28] or Mask R-CNN [14] is utilized to segment text-line regions or other types of areas from complex document images.

2.2 Document logical structure analysis

Logical structure analysis of documents focuses on analyzing the types and relationships of document elements at a logical level, which is often built upon the results of the physical layout analysis [30].

In document element classification tasks, early approaches are often based on deterministic grammar rules [8, 21]. Such approaches typically require detailed rule specifications for a particular document layout and struggle to generalize to different document scenarios. Deep learning methods have boosted the performance on this task. Vision-only approaches for multi-class detection or segmentation do not rely on extracting the text content and position of document elements but instead employ visual models to directly locate different categories of document elements [24, 32, 43]. Additionally, approaches based on natural language models [27] or multi-modal language models [40, 45] are used to determine the types of document elements when the text content and positions of document elements are provided.

In document structure analysis tasks, early solutions employ formal grammar or logical trees to represent and arrange hierarchical relationships among elements in documents. This often requires manually designing rules tailored to the current layout. To automatically learn relationships among document elements from diverse layout data, some deep learning-based approaches have been utilized for relationship prediction tasks. DocStruct [39] and StrucText [25] utilize a single learnable asymmetric parameter matrix for predicting asymmetric relationships between any two document elements. GraphDCM [38] introduces a *Merger* module containing a set of asymmetric parameter matrices, further aggregating fine-grained elements with the same category in the document into coarse-grained elements. In LayoutXLM [41], all possible head-tail pairs are first collected, and a bi-affine classifier [9] is used to determine if a relationship exists between them. GeoLayoutLM [29] proposes a relationship classification head composed of bi-linear layers and lightweight transformers, further enhancing the performance of element relationship classification tasks.

2.3 Form understanding benchmarks

The development process of form understanding tasks is closely related to relevant benchmarks. SROIE [16] comprises 973 scanned English receipts, each annotated with line-level texts, corresponding bounding boxes, and a structured extraction target with four predefined field types. CORD [33] is another receipt dataset collected from various sources, including shops and restaurants, containing 1,000 receipt images from Indonesia. Compared to SROIE, the CORD dataset includes annotations at the word and entity levels with richer extraction field types. It also provides classification attributes for key-value pairs and group information within the same item, enabling the recovery of local relationships between different entities. EPHOIE [37] consists of 1,494 examination paper headers collected from Chinese school exams, annotated with ten types of entities. It offers annotations for text-line-level positions and contents, as well as classification attributes for key-value pairs. However,

3

all the aforementioned datasets are provided under specific form categories, lacking diversity in form types. The FUNSD [17] dataset contains 199 noisy scanned English documents, along with annotations at the word and entity levels. It categorizes all entities within forms into four classes: *Header*, *Question*, *Answer*, and *Other*, and provides local relationships between entities, supporting entity labeling and entity linking tasks. XFUND [41] is an extension of FUNSD in multiple languages, collecting additional forms in seven languages and providing similar annotations as FUNSD. It is worth noting that XFUND suffers from some entity definition confusion, where different text lines that should belong to the same entity are split and labeled as distinct entities. The SIBR dataset [44] is a publicly available dataset designed for visual information extraction, comprising 1,000 form images, including 600 Chinese invoices, 300 English bills of entry, and 100 bilingual receipts. It offers text-line-level position and content information, along with two types of links to represent document element relationships: one for linking different text lines within the same entity and another for indicating relationships between different entities. However, these datasets lack uniform granularity in comprehensive annotations for words, text lines, and entities. In addition, they focus on local key-value relationships and ignore the nested relationships between elements of different hierarchies in the document, resulting in incomplete representation of form information.

3 SRFUND benchmark

3.1 Data collection and annotation

The objective of SRFUND is to advance the development of form understanding and structured reconstruction tasks. Among existing form datasets, FUNSD and XFUND are two prominent works that have made outstanding contributions to the establishment of typical form understanding tasks and the extension to multilingual data scenarios, respectively. Our dataset, SRFUND, is built upon these two representative datasets, encompassing all document images from both datasets. SRFUND comprises 1,592 form images across eight languages, with each language contributing 199 images.

Upon careful analysis of the existing annotations in FUNSD and XFUND, we found inconsistencies in the granularity of annotations between the two datasets. The FUNSD dataset encapsulates complete semantic information of entity elements, regardless of whether this information is distributed across single or multiple lines, while the XFUND dataset includes cases where consecutive semantic multi-line text is independently counted as separate entities without connectivity between these text lines. Fortunately, both FUNSD and XFUND datasets cover word-level textual contents and positions. Leveraging the word-level annotation information from the original datasets, we meticulously followed several procedures to ensure a rigorous construction process for the dataset: (1) Adjust inaccurate word-level bounding boxes and supplement missing textual information. (2) Aggregate consecutive words with continuous semantics into one text-line and annotate the corresponding rectangular bounding box. It's noteworthy that separate values with bullet points, keys and values in key-value pairs are considered different text lines, even when they are visually connected to each other. (3) For entities composed of consecutive semantic multi-line text, we annotate the polygonal bounding box of the entity. (4) Based on the roles of entities in the current form, modify or assign correct categories to different entities. When the IOU between the annotation box of an entity and the existing entity box from the original datasets (i.e. FUNSD or XFUND) is greater than 0.8, the original label of the entity is adopted as the initial label of the current entity. (5) Determine the location of item tables and annotate the headers and each individual row item within the tables. (6) For separate entities with linkage relationships, annotate the relationships between these entities (unidirectional or bidirectional).

To ensure the accuracy of the annotations, all annotated results underwent at least three rounds of cross-checking, with any disputed annotations resolved by domain experts in the field of document processing. Given the multilingual coverage within the SRFUND dataset, a commercial translation engine was employed during the annotation process to translate form images into the native languages of annotators, providing semantic context for reference. The collection, annotation, and refinement processes of the dataset collectively consumed approximately 6,000 person-hours.

3.2 Dataset analysis

Supported tasks: As illustrated in Figure 1 and Table 1, SRFUND covers annotations at different levels, enabling the dataset to support a wider range of tasks than all existing form understanding

4

Table 1: Comparison with existing form understanding datasets.

Dataset	Supported Tasks					Statistics		
	Word to Text-line	Text-line to Entity	Entity Labeling	Item Table	Structure Recovery	Language	Images	Avg. Form Tree Depth
SROIE [16]	✗	✗	✓	✗	-	EN	1,000	-
CORD [33]	✓	✓	✓	✗	Local	IND	1,000	1.173
EPHOIE [37]	✗	✗	✓	✗	Local	ZH	1,494	1.115
SIBR [44]	✗	✓	✓	✗	Local	ZH, EN	1,000	1.515
FUNSD [17]	✗	✗	✓	✗	Local	EN	199	1.570
XFUND [41]	✗	✗	✓	✗	Local	ZH, JA, ES, FR, IT, DE, PT	1,393	1.699
SRFUND (Ours)	✓	✓	✓	✓	Global	EN, ZH, JA, ES, FR, IT, DE, PT	1,592	3.049

datasets. It is noteworthy that previous datasets focused solely on the structural relationships between local document entities, whereas we meticulously annotated the logical relationships among all entities. This makes SRFUND the first dataset supporting the task of structure recovery for each entity at a global level. Furthermore, with the complement of item tables and their constituent items, SRFUND is also the first dataset supporting the localization of item tables within forms. The recovery of global structures and the localization of item tables require models to possess a strong understanding of entities within forms, posing significant challenges to form understanding tasks.

Table 2: Statistics of different granularity categories of SRFUND. *Total*, *Same*, *New* refer to the total amount of a class in SRFUND, the same amount as in the original dataset (FUNSD/XFUND), and the amount added or modified, respectively.

	Words	Lines	Entities	Tables	Table Items	Links
Total	529,711	112,662	96,824	591	1,954	122,594
Same	529,372	0	81,462	0	0	47,113
New	339	112,662	15,362	591	1,954	75,481

Statistical metrics: (a) As shown in Table 1, the SRFUND dataset includes annotations in eight different languages, making it more diverse than existing datasets and addressing form understanding needs across various languages. In terms of document hierarchy, we constructed global entity relationships, resulting in an average tree depth of 3.049, which surpasses previous datasets significantly. (b) As depicted in Table 2, the SRFUND dataset introduces a substantial number of annotations across different hierarchical levels in addition to the original version. We modified 339 word-level annotations and provided annotations for 112,662 text lines. Our dataset contains a total of 96,824 entities, out of which 15,362 are newly added or modified. Additionally, we provided detailed annotations for item tables in documents, totaling 591 item tables and 1,954 item contents within them. Furthermore, we added 75,481 entity links, resulting in a total of 122,594 entity links.

4 Experiments

To comprehensively assess the SRFUND dataset, we conducted benchmark tests using models across three different modalities: language models based on pure text input, detection models based on purely visual inputs, and document pre-trained language models that utilize multi-modal inputs. We performed experimental analyses on five tasks, detailed as follows: (1) word to text-line merging, (2) text-line to entity merging, (3) entity category classification, (4) item table localization, and (5) entity-based full-document hierarchical structure recovery.

4.1 Setting

As illustrated in Figure 2, the vision-only models rely on the document image as the input, and tasks 1 to 4 can be regarded as the text-line detection task, the entity detection task, the multi-class entity detection task and the line item table area detection task respectively. We selected three distinct

5

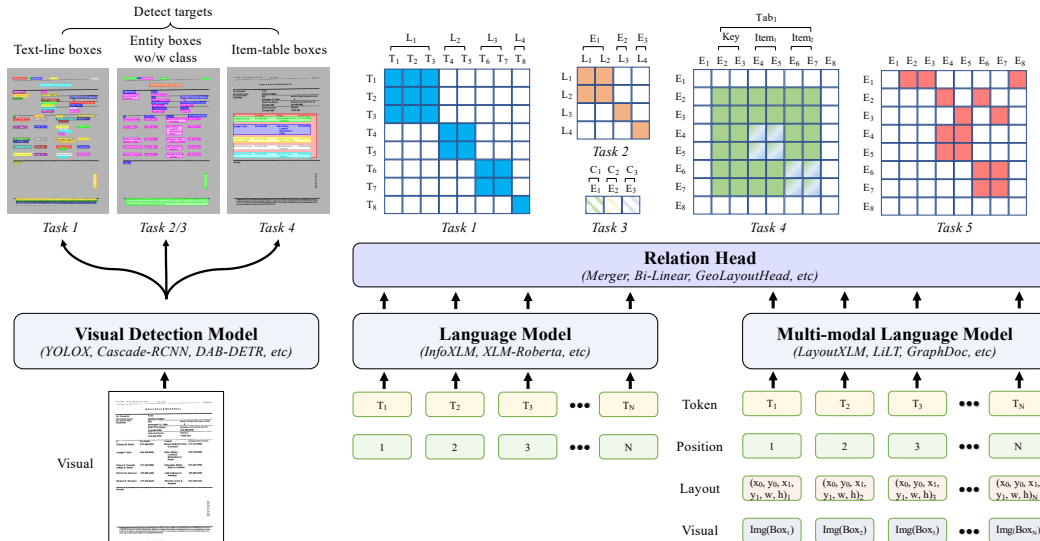



Figure 2: Models with varied modalities used for evaluating on the SRFUND benchmark. The *left* side shows the input (original document image) and predicted targets for the visual model in tasks 1∼4, treated as detection tasks. Tasks 1, 2, and 4 involve single-category detection, while Task 3 involves four-category detection. The *right* side presents the input and output of the pure language model and multimodal pretraining model. The abbreviations T/L/E/C/Tab/Key/Item stand for Token/Text-Line/Entity/Category/Item-Table/Header/Table-Item, with □ indicating the merging relationship from fine-grained to coarse-grained elements.

types of visual detection models for comparison: YOLOX [10], a single-stage detector based on the YOLO architecture; Cascade-RCNN [2], a multi-stage detector based on the RCNN framework; and DAB-DETR [26], an improved version of the DETR model with faster convergence speed and detection accuracy. All three detection models utilize a ResNet-50 [15] backbone pre-trained on ImageNet [7]. In the training stage, we follow the original configuration adopted in the mmdetection [4]. During testing, we only preserve predicted boxes with a score threshold larger than 0.3 and adopt the non-maximum suppression algorithm for further filtering. Since the task is framed as a coarse-to-fine merging task, we use the standard F1 score as the main evaluation metric. Unlike the common practice in object detection, where true positives are determined by thresholding the Intersection-over-Union, we use a different criterion tailored to evaluate the usefulness of detections for text read-out. Inspired by the CLEval metric [1] used in text detection, we measure whether the predicted area contains nothing but the related word-level box centers as visualized in Figure 3.

The text-only models rely on the word texts as the input, and the multi-modal models rely on the texts, two-dimensional coordinates, and form images as the input. Given the multilingual nature of the SRFUND dataset, we employed several multilingual language models and document pre-trained language models that support multilingual inputs. The language models include InfoXLM [5] and

(a) Correct detection examples   (b) Incorrect detection examples

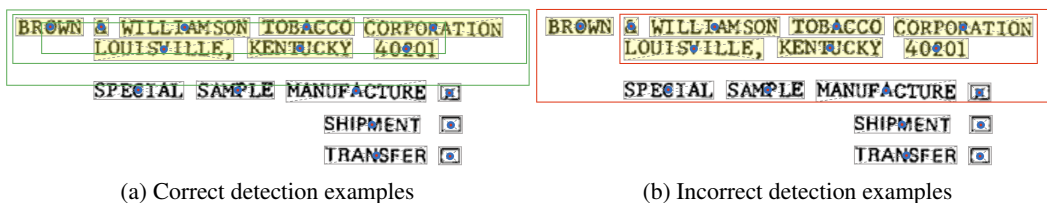



Figure 3: Visualization of correct (the green boxes) and incorrect (the red boxes) bounding box predictions to capture the *Header* entity (texts with yellow background). Bounding box must include exactly the word-level centers that lie within the ground truth annotation. Note: in Figure 3a, only one of the predictions would be considered correct if all three boxes were predicted.

6

XLM-Roberta [6], while the document pre-trained language models include LayoutXLM [41], LiLT [36], and GraphDoc [45], while the latter two models incorporate InfoXLM-base as the language model. As illustrated in Figure 2, for tasks 1, 2, and 4, we use a symmetric attention relation matrix as the learning target, where each aggregation target may contain multiple fine-grained element sets. Each aggregation target is considered an independent prediction sample, and a prediction is deemed correct only if all elements within the target set are completely aggregated together. For task 5, we utilize an asymmetric attention relation matrix to learn the relationships between different entities , while a pair prediction is only considered correct if the directional relationship between two entities is accurately predicted. We employ the F1 score as the final evaluation metric for all tasks, and *Merger* [38] is adopted as the default relation classification head in the following sections.

All models were run on servers equipped with eight 48 GB A40 graphics cards with a batch size of 8. To ensure that different models achieved their optimal performance, we followed the training strategies for vision-only models as the original version in mmdetection [4], with specific settings for learning rate, optimizer, and the number of epochs detailed in the appendix. The text-only and multi-modal models uniformly utilized the Adam [19] optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$ and underwent 200 training epochs, with an initial learning rate set at 5e-5. The training began with a linear warm-up during the first 10% epochs, followed by a continuation under a linear decay strategy.

4.2 Results and analysis

4.2.1 Word to text-line merging

Table 3: Results of the word to text-line merging task, using F1-score as the metric.

Type	Method	English	Chinese	Japanese	German	French	Spanish	Italian	Portuguese	Avg.
Vision-only	YOLOX [10]	0.8222	0.8053	0.6959	0.8587	0.7310	0.8301	0.7470	0.7900	0.7850
	Cascade-RCNN [2]	0.8520	0.8842	0.7569	0.8683	0.8191	0.8404	0.7590	0.7710	0.8189
	DAB-DETR [26]	0.8437	0.8500	0.7394	0.8795	0.8082	0.8468	0.7926	0.7954	0.8194
Text-only	XLM-RoBerta [6]	0.6290	0.6272	0.6093	0.6982	0.6921	0.6470	0.6285	0.6780	0.6509
	InfoXLM [5]	0.6426	0.6482	0.6298	0.7011	0.6974	0.6551	0.6253	0.6921	0.6611
Multi-modal	LayoutXLM [40]	**0.9081**	0.9360	**0.9118**	**0.9255**	**0.9282**	**0.9372**	**0.9157**	**0.9387**	**0.9260**
	LiLT [36]	0.8887	**0.9387**	0.8803	0.9193	0.9223	0.9202	0.8962	0.9054	0.9094
	GraphDoc [45]	0.8755	0.9100	0.8005	0.9167	0.8954	0.8993	0.8471	0.8708	0.8758

Aggregating words into text lines presents significant challenges for single-modal approaches. As demonstrated in Table 3, text-only models faced difficulties due to the absence of two-dimensional spatial coordinates and visual cues. This lack of information impedes their ability to accurately identify breakpoints in semantically continuous texts that span multiple lines. Conversely, vision-only models depend exclusively on the visual boundary features of text lines to assess word aggregation. This approach often fails to correctly segment words that are visually close yet semantically distinct, such as *Question* and *Answer* appearing on the same line. Additionally, the intricate form layouts of languages like Japanese and Italian pose further challenges in precise text-line segmentation. In contrast, document pre-trained language models that integrate multiple modalities significantly improve performance by leveraging a broader range of data, thereby overcoming the limitations of single-modal systems.

4.2.2 Text-line to entity merging

The task of merging text lines into entities relies more on semantic information. Compared to task 1, we can observe from Table 4 that the performance of text-only models surpasses that of vision-only models. This indicates that pure vision models have a weaker capability in understanding multi-line entities, while linguistic information significantly aids in capturing the semantic continuity between different text lines.

The pre-training process greatly impacts the effectiveness of document pre-trained language models. The GraphDoc model, which leverages sentence-level semantic information and is only pre-trained on English documents, performs well in the task of merging text lines in English forms. In contrast, LayoutXLM is pre-trained using documents in all languages included in the SRFUND dataset, which allows it to demonstrate superior performance on forms in other languages.

7

Table 4: Results of the text-line to entity merging task, using F1-score as the metric.

Type	Method	English	Chinese	Japanese	German	French	Spanish	Italian	Portuguese	Avg.
Vision-only	YOLOX [10]	0.7415	0.7243	0.5891	0.7309	0.6504	0.7449	0.6238	0.6594	0.6830
	Cascade-RCNN [2]	0.7918	0.8336	0.6873	0.7997	0.8060	0.8138	0.7153	0.7560	0.7754
	DAB-DETR [26]	0.7681	0.7794	0.6332	0.7893	0.7344	0.7663	0.7075	0.7346	0.7391
Text-only	XLM-RoBerta [6]	0.8767	0.9354	0.8974	0.8850	0.9014	0.9044	0.8836	0.9226	0.9029
	InfoXLM [5]	0.8773	0.9411	0.8921	0.8729	0.9010	0.9026	0.8847	0.9188	0.9012
Multi-modal	LayoutXLM [40]	0.9151	**0.9681**	**0.9387**	**0.9157**	**0.9408**	**0.9463**	**0.9280**	**0.9594**	**0.9412**
	LiLT [36]	0.9047	0.9542	0.9117	0.9140	0.9368	0.9351	0.9134	0.9430	0.9283
	GraphDoc [45]	**0.9229**	0.9343	0.8770	0.9113	0.9260	0.9326	0.9060	0.9314	0.9181

4.2.3 Entity category classification

Table 5: Results of the entity category classification task, using F1-score as the metric.

Type	Method	English	Chinese	Japanese	German	French	Spanish	Italian	Portuguese	Avg.
Vision-only	YOLOX [10]	0.5284	0.6040	0.4619	0.4976	0.4743	0.5385	0.4466	0.4244	0.4969
	Cascade-RCNN [2]	0.6739	0.7482	0.6124	0.7123	0.7749	0.7318	0.6662	0.6707	0.6988
	DAB-DETR [26]	0.6531	0.6631	0.5286	0.6735	0.6863	0.6574	0.6067	0.6152	0.6355
Text-only	XLM-RoBerta [6]	0.8558	0.9666	0.8847	0.8912	0.9067	0.9161	0.8955	0.8884	0.9028
	InfoXLM [5]	0.8589	0.9570	0.8782	0.8953	0.9107	0.9221	0.8995	0.8840	0.9025
Multi-modal	LayoutXLM [40]	**0.9045**	**0.9718**	**0.8957**	0.9216	**0.9299**	**0.9320**	**0.9269**	**0.9086**	**0.9248**
	LiLT [36]	0.8678	0.9631	0.8876	0.9006	0.9217	0.9270	0.9135	0.8967	0.9118
	GraphDoc [45]	0.8930	0.9619	0.8620	**0.9261**	0.9129	0.9250	0.9169	0.8897	0.9113

Visual modalities can also address the task of entity classification by learning layout information. Apart from *Header* entities, which typically exhibit characteristics such as boldface and larger font sizes, the variations in font styles among other entity categories are minimal. However, as shown in Table 5, detection models are also capable of effectively handling the detection tasks for various entity categories. This indicates that the visual modality can acquire layout information, such as *Question* typically being located to the left of *Answer*, and entities with multi-line texts usually being categorized as *Answer* or *Other*.

4.2.4 Item table localization

Table 6: Results of the item table localization task, using F1-score as the metric.

Type	Method	English	Chinese	Japanese	German	French	Spanish	Italian	Portuguese	Avg.
Vision-only	YOLOX [10]	0.1100	0.1721	0.0100	0.0467	0.1000	0.0710	0.0600	0.0911	0.0826
	Cascade-RCNN [2]	0.0839	0.2081	0.0433	0.0800	0.1327	**0.1427**	0.0817	**0.1486**	0.1151
	DAB-DETR [26]	0.1399	0.2670	0.0000	0.0667	0.1333	0.0903	0.0767	0.1100	0.1105
Text-only	XLM-RoBerta [6]	0.0526	0.2090	0.0800	**0.5714**	0.2222	0.0000	0.1333	0.0526	0.1514
	InfoXLM [5]	0.0513	0.1846	0.0000	0.4545	0.2143	0.0000	0.0000	0.0000	0.1124
Multi-modal	LayoutXLM [40]	**0.7273**	**0.3333**	**0.1053**	0.4348	0.1053	0.0588	**0.3158**	0.1250	**0.3022**
	LiLT [36]	0.2273	0.1867	0.0000	0.5263	0.0769	0.0000	0.0000	0.0417	0.1306
	GraphDoc [45]	0.0000	0.0556	0.0000	0.3333	**0.3333**	0.0606	0.0000	0.1224	0.0945

The item table localization task presents significant challenges. Since this task requires that all entities within the item table be included, any missing entity predictions result in the outcome being deemed incomplete. This requirement makes it difficult for models of any modality to accurately locate item tables within documents, as shown in Table 6.

There is significant variability in performance across forms in different languages. Item table localization in forms of different languages requires optimization through models of various modalities. For some languages, the localization of item tables relies more heavily on the capabilities of vision models, such as in Spanish and Portuguese forms.

8

4.2.5 Hierarchical structure recovery

Table 7: Results of the hierarchical structure recovery task, using F1-score as the metric.

Type	Method	English	Chinese	Japanese	German	French	Spanish	Italian	Portuguese	Avg.
Text-only	XLM-RoBerta [6]	0.5270	0.6514	0.5388	0.6637	0.6054	0.6121	0.5839	0.5081	0.5830
	InfoXLM [5]	0.5305	0.6436	0.5227	0.6695	0.6071	0.5941	0.5736	0.4872	0.5732
Multi-modal	LayoutXLM [40]	0.7135	0.7601	0.6626	0.7734	0.7415	0.7009	0.6710	0.6310	0.7013
	LiLT [36]	0.7050	0.7578	0.6538	0.7499	0.7153	0.6940	0.6702	0.5747	0.6821
	GraphDoc [45]	**0.7938**	**0.7881**	**0.6714**	**0.7976**	**0.7754**	**0.7416**	**0.6969**	**0.6648**	**0.7349**

The granularity of semantic information must align with the task requirements. Among the three document pre-trained language models, only the GraphDoc model maintains sentence-level or entity-level linguistic input during both the pre-training stage and the training/testing phases of the current task. This alignment enables the GraphDoc model to achieve the best performance in the task of entity-based document structure recovery.

4.2.6 Multimodal Large Language Models (MLLMs) evaluation

Table 8: Results of the two leading performance Multimodal Large Language Models (MLLMs) on SRFUND, using F1-score as the metric.

Model	Tasks	English	Chinese	Japanese	German	French	Spanish	Italian	Portuguese	Avg.
GPT4o	Word to text-line merging	0.4607	0.57	0.3941	0.5676	0.4248	0.5614	0.4098	0.4944	0.4866
GPT4o	Text-line to entity merging	0.2705	0.2035	0.426	0.2415	0.313	0.4644	0.196	0.2397	0.2936
GPT4o	Entity category classification	0.5608	0.3352	0.5298	0.4879	0.4735	0.4478	0.4559	0.4625	0.469
GPT4o	Item table localization	0.0667	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.047
GPT4o	Hierarchical structure recovery	0.1312	0.122	0.1098	0.1713	0.1205	0.0851	0.0864	0.0562	0.1103
GPT4o-mini	Word to text-line merging	0.1866	0.0779	0.0644	0.2509	0.118	0.164	0.1778	0.2488	0.1611
GPT4o-mini	Text-line to entity merging	0.666	0.768	0.8241	0.7999	0.7572	0.8487	0.7971	0.7828	0.7805
GPT4o-mini	Entity category classification	0.4643	0.2364	0.196	0.3498	0.3491	0.3172	0.2201	0.2398	0.2966
GPT4o-mini	Item table localization	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GPT4o-mini	Hierarchical structure recovery	0.371	0.1905	0.183	0.162	0.1865	0.1724	0.1376	0.1892	0.1985

MLLMs like GPT-4o and GPT-4o-mini performed reasonably well on document tasks but were outperformed by smaller, domain-specific models. We conducted experiments using GPT-4o and GPT-4o-mini, two leading general-domain multimodal models provided by OpenAI via API as *gpt-4o-2024-08-06* and *gpt-4o-mini* [4]. Document images served as the input, following a prompt structure comprising a role definition, task description, and the positions and textual content of words, text lines, or entities. The results are presented in Table 8. The findings demonstrate that, even in zero-shot scenarios, these MLLMs achieved reasonable performance, though they were surpassed by smaller, domain-specific models. Notably, GPT-4o outperformed GPT-4o-mini in most tasks, except in the Text-line to entity merging task, where GPT-4o-mini showed superior performance. More details could be found in the Appendix section.

4.2.7 Out-of-domain (OOD) evaluation

Table 9: Cross-validation results between models trained on other datasets and SRFUND, evaluated using Precision/Recall/F1-score. $A \rightarrow B$ denotes training on dataset A and reporting results on the test set of dataset B. The results on SRFUND are averaged across all languages.

Tasks	Word to text-line merging	Text-line to entity merging	Hierarchical structure recovery
CORD → SRFUND	0.2078 / 0.2192 / 0.2133	0.0771 / 0.2107 / 0.1128	0.1565 / 0.0476 / 0.0730
SIBR → SRFUND	- / - / -	0.0813 / 0.2859 / 0.1266	0.4322 / 0.1279 / 0.1974
SRFUND → CORD	0.8660 / 0.7821 / **0.8219**	0.9474 / 0.9309 / **0.9390**	0.1342 / 0.8169 / **0.2305**
SRFUND → SIBR	- / - / -	0.8780 / 0.7925 / **0.8331**	0.2984 / 0.6453 / **0.4081**

Models fine-tuned on SRFUND demonstrate superior out-of-domain (OOD) performance. To assess the out-of-domain generalization capabilities of models fine-tuned on SRFUND, we conducted

---

[4]Experiments taking place on August 16, 2024

9

cross-validation experiments on Tasks 1, 2, and 5 using the CORD [33] and SIBR [44] datasets. Unfortunately, due to the differences in entity categories and the absence of item table information in these external datasets, cross-validation was not feasible for Tasks 3 and 4. We used the LayoutXLM model for these experiments, maintaining the same training configurations as described in the original paper. The results are presented in Table 9. The results indicate that models trained on SRFUND exhibit superior out-of-domain generalization compared to those trained on other datasets. This improvement is likely due to SRFUND's finer-grained annotations and its diverse distribution of form types and layouts, which may better capture the variability present in real-world documents.

4.2.8   Overall analysis

The results from the aforementioned five tasks demonstrate that uni-modal models exhibit relatively poor performance, while document pre-trained models that incorporate multiple input modalities perform significantly better. Concurrently, no single model has consistently outperformed others across all tasks and languages, indicating that in practical applications, we cannot simply rely on a single model or approach to handle all types of form structuring tasks. Instead, we need to select appropriate models and strategies based on the specific requirements of the task and the characteristics of the language involved. This finding underscores the importance of adopting a nuanced and tailored approach when tackling form structuring challenges, rather than employing a one-size-fits-all solution.

5   Limitations

SRFUND is built upon the FUNSD and XFUND datasets, primarily comprising several types of forms such as application forms, registration forms, and dealing slips, all stored as grayscale images. Although these forms exhibit relatively rich layouts and designs, they lack diverse background images and colors. Furthermore, due to the limited size of the original datasets and the need for meticulous manual intervention during the creation of annotated data to ensure annotation quality, the current dataset remains relatively small. These limitations can be partly addressed by using pre-trained models on datasets like DocLayNet [34] (a layout analysis dataset containing 80.9K document images, mainly in English) and ADOPD [11] (a multi-task dataset containing 120K document images, focusing mainly on English and East Asian languages such as Chinese, Japanese, and Korean) for initial annotations, followed by repeating the current annotation process. However, establishing a scalable data annotation process for multilingual documents remains a challenge and is a focus of future work.

6   Conclusion

In summary, this paper presents two main contributions. Firstly, we propose a multilingual, multitask document hierarchical structuring benchmark named SRFUND, encompassing 1,592 forms from eight languages. To the best of our knowledge, this is the first benchmark in form understanding that integrates multi-level structure reconstruction, spanning from words to the global structure of forms. Secondly, we conducted baseline experiments on five tasks using various representative approaches from different modalities, demonstrating that the SRFUND benchmark introduces new challenges to the field of form understanding. We believe that the SRFUND benchmark holds significant potential for future academic research, contributing continuously to the in-depth study of global form structures.

References

[1] Youngmin Baek, Daehyun Nam, Sungrae Park, Junyeop Lee, Seung Shin, Jeonghun Baek, Chae Young Lee, and Hwalsuk Lee. Cleval: Character-level evaluation for text detection and recognition tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 564–565, 2020.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

10

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020.

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[5] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, 2021.

[6] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 2019.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] D Derrien-Peden. Frame-based system for macrotypographical structure analysis in scientific papers. In *Proceedings of 1st International Conference on Document Analysis and Recognition*, pages 311–319, 1991.

[9] Timothy Dozat and Christopher D Manning. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*, 2016.

[10] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[11] Jiuxiang Gu, Xiangxi Shi, Jason Kuen, Lu Qi, Ruiyi Zhang, Anqi Liu, Ani Nenkova, and Tong Sun. Adopd: A large-scale document page decomposition dataset. In *The Twelfth International Conference on Learning Representations*, 2024.

[12] Jaekyu Ha, Robert M Haralick, and Ihsin T Phillips. Recursive xy cut using bounding boxes of connected components. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pages 952–955, 1995.

[13] Haralick. Document image understanding: Geometric and logical layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 385–390, 1994.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[16] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *Proceedings of 15th International Conference on Document Analysis and Recognition*, pages 1516–1520, 2019.

[17] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops*, pages 1–6, 2019.

[18] Saima Khan, Shazia Khan, and Mohsina Aftab. Digitization and its impact on economy. *International Journal of Digital Library Services*, 5(2):138–149, 2015.

11

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[20] Koichi Kise, Akinori Sato, and Motoi Iwata. Segmentation of page images using the area voronoi diagram. *Computer Vision and Image Understanding*, 70(3):370–382, 1998.

[21] Mukkai Krishnamoorthy, George Nagy, Sharad Seth, and Mahesh Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(7):737–747, 1993.

[22] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 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*, pages 6309–6318, 2021.

[23] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8610–8617, 2019.

[24] Kai Li, Curtis Wigington, Chris Tensmeyer, Handong Zhao, Nikolaos Barmpalios, Vlad I Morariu, Varun Manjunatha, Tong Sun, and Yun Fu. Cross-domain document object detection: Benchmark suite and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12915–12924, 2020.

[25] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1912–1920, 2021.

[26] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *International Conference on Learning Representations*, 2021.

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[29] Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. Geolayoutlm: Geometric pre-training for visual information extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7092–7101, 2023.

[30] Anoop M Namboodiri and Anil K Jain. Document structure and layout analysis. In *Digital Document Processing: Major Directions and Recent Advances*, pages 29–48. 2007.

[31] Lawrence O'Gorman. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1162–1173, 1993.

[32] Dario Augusto Borges Oliveira and Matheus Palhares Viana. Fast cnn-based document layout analysis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1173–1180, 2017.

[33] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.

[34] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3743–3751, 2022.

12

[35] Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 572–573, 2020.

[36] Jiapeng Wang, Lianwen Jin, and Kai Ding. 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*, pages 7747–7757, 2022.

[37] Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. Towards robust visual information extraction in real world: new dataset and novel solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2738–2745, 2021.

[38] Yan Wang, Jun Du, Jiefeng Ma, Pengfei Hu, Zhenrong Zhang, and Jianshu Zhang. Ustc-iflytek at docile: a multi-modal approach using domain-specific graphdoc. *Working Notes of CLEF*, 2023.

[39] Zilong Wang, Mingjie Zhan, Xuebo Liu, and Ding Liang. Docstruct: A multimodal method to extract hierarchy structure in document for general form understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 898–908, 2020.

[40] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 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, 2020.

[41] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. Xfund: a benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3214–3224, 2022.

[42] Yue Xu, Fei Yin, Zhaoxiang Zhang, and Cheng-Lin Liu. Multi-task layout analysis for historical handwritten documents using fully convolutional networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1057–1063, 2018.

[43] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324, 2017.

[44] Zhibo Yang, Rujiao Long, Pengfei Wang, Sibo Song, Humen Zhong, Wenqing Cheng, Xiang Bai, and Cong Yao. Modeling entities as semantic points for visual information extraction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15358–15367, 2023.

[45] Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. Multimodal pre-training based on graph attention network for document understanding. *IEEE Transactions on Multimedia*, 25:6743–6755, 2023.

[46] Yefeng Zheng, Huiping Li, and David Doermann. A model-based line detection algorithm in documents. In *Proceedings of 7th International Conference on Document Analysis and Recognition*, pages 44–48, 2003.

13

A Datasheets for SRFUND

A.1 Motivation

For what purpose was the dataset created?

The purpose of creating SRFUND dataset is to advance the development of form understanding and structured reconstruction tasks by covering forms of various layouts and languages. Although some benchmarks datasets [16, 17, 33, 37, 41, 44] have been established, none of them have established the global and hierarchical structural dependencies that consider all elements at different granularity, including words, text lines, and entities within the forms. To enhance the applicability of form understanding tasks in hierarchical structure recovery, we introduce the SRFUND, a multilingual document structure reconstruction dataset. To the best of our knowledge, this is the first benchmark in form understanding that integrates multi-level structure reconstruction, spanning from words to the global structure of forms, and we believe that the SRFUND dataset will significantly promote the development of form understanding and structured reconstruction.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The SRFUND dataset was created by the NERC-SLIP of University of Science and Technology of China.

Who funded the creation of the dataset?

The iFLYTEK Research.

A.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The SRFUND dataset comprises 1,592 form images, which across eight languages with each language contributing 199 images, accompanied by their respective annotation files. These images represent scanned or photographed forms, and images in English are stored in the Portable Network Graphics (PNG) format, while images in other languages are stored in the Joint Photographic Experts Group (JPEG) format. The annotations are stored in JSON format, capturing the locations and text content of every word, text-line, and entity, including their hierarchical dependencies. Furthermore, the entities are categorized into four classifications including Header, Question, Answer, and Other. The multi-item table regions which are frequently found in forms are also specifically annotated.

How many instances are there in total (of each type, if appropriate)?

The SRFUND dataset comprises a collection of 1,592 images, with 96,824 entities, 112,662 text lines, 529,711 words, and 122,594 linkings.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The SRFUND dataset contains all possible instances.

What data does each instance consist of?

Each instance in the SRFUND consists of an image along with corresponding annotations. These annotations include bounding boxes and text content of every word, text-line, entity, and item table, including their hierarchical dependencies. Moreover, every entity is assigned a categorical label, namely Header, Question, Answer, or Other.

Is there a label or target associated with each instance?

Yes. The label contains bounding boxes and text content of every word, text-line, entity, and item table, including their hierarchical dependencies, as well as a categorical label for every entity.

Is any information missing from individual instances?

No. There is no missing information from individual instances in the SRFUND dataset.

14

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Yes. Images belonging to the same language are contained in the same folder.

Are there recommended data splits (e.g., training, development/validation, testing)?

The SRFUND dataset is divided into training and validation sets in a ratio of approximately 3:1.

Are there any errors, sources of noise, or redundancies in the dataset?

Despite the SRFUND dataset undergoing rigorous multiple checks and expert verification, there may still be instances of minor errors, such as in sections of handwritten text. Should any annotation mistakes be identified, or if users report such errors, we will promptly address these in the maintenance process to ensure the accuracy of the data.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The SRFUND dataset is self-contained.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

No. The SRFUND dataset provides refined annotations on top of the original FUNSD and XFUND datasets. The XFUND dataset collected the documents publicly available on the internet and removed the content within the documents while only keeping the templates to manually fill in synthetic information. The FUNSD dataset was annotated with a subset of the Truth Tobacco Industry Document (TTID), an archive collection of scientific research, marketing, and advertising documents of the largest US tobacco firms, which aims to advance information retrieval research.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

No.

A.3 Collection Process

How was the data associated with each instance acquired?

The SRFUND dataset provides refined annotations on top of the original FUNSD and XFUND datasets. The XFUND dataset collected the documents publicly available on the internet and removed the content within the documents while only keeping the templates to manually fill in synthetic information. The FUNSD dataset was annotated with a subset of the Truth Tobacco Industry Document (TTID). For more information about data collection, please refer to the FUNSD and XFUND datasets. The annotation process is described in Sec. 3.1 of the main paper.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?

The SRFUND dataset provides refined annotations on top of the original FUNSD and XFUND datasets. We did not collect any data ourselves, but used X-anylabeling for finer annotations.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Ten crowdworkers and four graduate students participated in the data collection process. The crowdworkers were responsible for providing initial annotations and cross-checking them. For each form annotated or checked, crowdworkers received a compensation of $1 or $0.2, respectively. The graduate students were tasked with resolving conflicts in the annotations provided by the crowdworkers. They performed these corrections based on a detailed pre-established annotation

15

guideline and their specialized knowledge in the field of form understanding. The graduate students were compensated through research grants.

Over what timeframe was the data collected?

For more information about data collection, please refer to the FUNSD and XFUND datasets. The annotation process is described in Sec. 3.1 of the main paper. The collection, annotation, and refinement processes of the dataset collectively consumed approximately 6,000 person-hours, spanning approximately 5 months.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

Does the dataset relate to people?

Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We collected the data from other sources, including the FUNSD and XFUND datasets.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

N/A.

A.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

There was no preprocessing/cleaning of the data done. The annotation process is described in Sec. 3.1 of the main paper.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

No, but all of the source data products are freely available online.

Is the software that was used to preprocess/clean/label the data available?

We used X-anylabeling which was available at https://github.com/CVHub520/X-AnyLabeling for finer annotations.

A.5 Uses

Has the dataset been used for any tasks already?

No.

Is there a repository that links to any or all papers or systems that use the dataset?

The current paper and the code used for experiments are available at `https://sprateam-ustc.github.io/SRFUND`.

What (other) tasks could the dataset be used for?

The SRFUND dataset can also be utilized for tasks such as hierarchical text recognition and the generation of document-based question answering data, relying on global structural analysis.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

16

No.

Are there tasks for which the dataset should not be used?

No.

A.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Yes. The SRFUND dataset is available at `https://sprateam-ustc.github.io/SRFUND`.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

The SRFUND dataset is available through the project website at `https://sprateam-ustc.github.io/SRFUND`.

When will the dataset be distributed?

The dataset is already available.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The dataset will be distributed under the Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No.

A.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset will be maintained by the NERC-SLIP of University of Science and Technology of China.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Contact can be made via email at jfma@mail.ustc.edu.cn

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

If substantial errors are raised by dataset users, we will update the dataset accordingly. The updated version of the dataset will be made available through the dataset release link.

Will older versions of the dataset continue to be supported/hosted/maintained?

Yes, with each update, the older versions will remain accessible through their original links.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

The dataset will be distributed under the Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. This means that researchers are free to extend, augment, build upon, and contribute to the dataset for non-commercial purposes. However, any distribution must be under the same license, and any modifications must be documented.

17

B Further data analysis

B.1 Annotation distribution

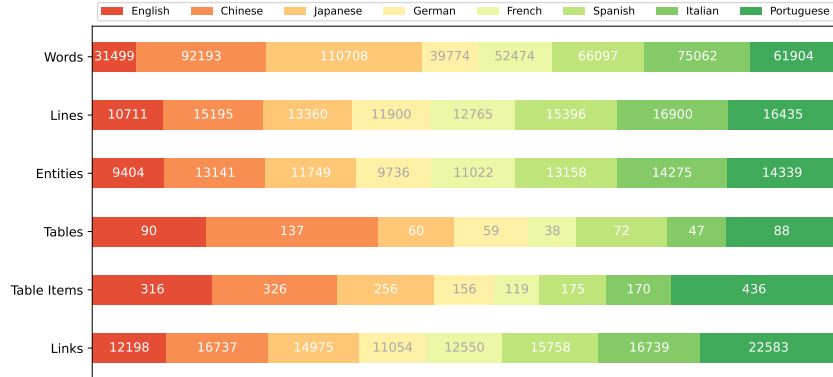



Figure 4: The number of annotations of different granularity in each language in the SRFUND dataset.

In the SRFUND dataset, which encompasses multi-lingual and multi-granularity forms with hierarchical structure annotations, significant variations in annotation types across languages reveal insights into the dataset's composition and potential biases. As depicted in Figure 4, the dataset exhibits a predominance in word-level annotations for forms in Chinese and Japanese, pointing towards extensive textual contents. Italian and Portuguese forms exhibit the highest counts in lines, which could reflect longer or more dispersed document formats. In terms of entities, forms in Portuguese lead, suggesting a denser distribution of entities, which is essential for tasks requiring detailed entity recognition. The Portuguese language also stands out in table items and links between entities, indicating a high degree of structured and relational data integration within forms.

These patterns suggest that the forms in Italian and Portuguese might be rich in structured formats like tables and entity relationships, which are crucial for complex structure analysis tasks. The differences in annotation distribution across languages highlight the diversity in document content, structure, and utility, underscoring the importance of tailored approaches in language-specific data science and natural language processing applications. This comprehensive annotation overview not only aids in understanding the dataset's complexity but also enhances the strategic planning of multilingual structure analysis systems.

B.2 Item table diversity

Figure 5 illustrates the diversity and complexity of item tables in the SRFUND dataset, which contains 591 item tables and 1,954 item group entries across various language forms. The subfigures exemplify the variations in structural and linguistic features characteristic of the dataset: **Subfigure 5a** depicts an item table from an English-language form, embedded directly within the text content without surrounding borders, highlighting the integration of tabular data within texts. **Subfigure 5b** shows an item table from a Spanish-language form, part of a larger bordered table that includes nested item table headings, demonstrating the existence of nested structures within tabular layouts. **Subfigure 5c** presents a Portuguese-language form example, featuring four item tables with identical column headings. These tables incorporate multiple selectable checkbox options within certain cells and are arranged in a vertically elongated format. **Subfigure 5d** from a Chinese-language form features distinct row headings with an item table at the bottom that includes cross-row items, illustrating variations in row-level organization and the challenges of spanning entries. This diversity poses significant challenges for the localization of item tables and the extraction of relationships between different entities within these tables, critical for the automated processing and analysis of form-based data in multilingual contexts.

18

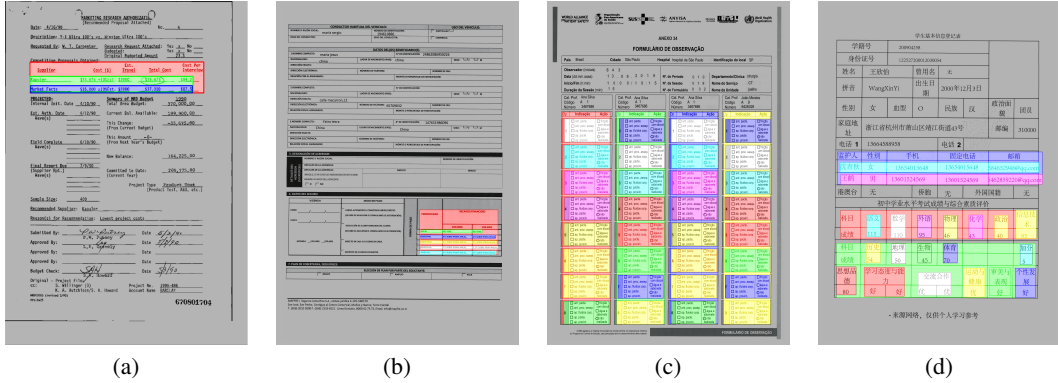



Figure 5: Varied item table annotations that are derived from diverse linguistic sources in the SRFUND dataset. Subfigures (a), (b), (c), and (d) originate from forms in English, Spanish, Portuguese, and Chinese, respectively.

C Hyperparameters

We followed the training strategies for vision-only models as the original version in mmdetection [4]. Detailed configurations are listed in the Table 10.

Table 10: Hyperparameters used in vision-only approaches.

Strategies	YOLOX [10]	Cascade-RCNN [2]	DAB-DETR [26]
Initial learning rate	$1e-2$	$2e-2$	$1e-4$
Optimizer	SGD	SGD	AdamW
Optimizer config	momentum= 0.9, weight_decay= $5e-4$	momentum= 0.9, weight_decay= $1e-4$	weight_decay= $1e-4$
Training epoch	200	48	100

D Extensive experiments

D.1 Relation heads comparison

We conducted experiments on four tasks that involve relationship classification across different levels of granularity. Throughout the experiments, we utilized LayoutXLM-base [41] as the base model, with the results presented in Table 11. It was observed that in some foundational tasks, different structures of relation heads exhibited similarly close performance. This might be due to Tasks 1 and 2 relying more heavily on the base model's capability to understand document layouts, where even relatively simple head designs could achieve satisfactory results. In Task 4, the simplest *Merger* classifier outperformed the other two heads due to the limited training data available for table data. Additionally, it was noted that different models displayed inconsistent performances across various languages in Task 4. This inconsistency might indicate significant divergences in content and layout among table data across languages, as also observed in Figure 4. In Task 5, the relation head in *GeoLayoutLM* demonstrated a clear advantage, exhibiting consistent superiority across different languages, due to its design of a multi-layer classification network ranging from coarse to fine at entity levels.

D.2 Cross language validation

We used LayoutXLM as the base model and *Merger* as the classification head for cross-lingual validation on the hierarchical structure recovery task. For the SRFUND dataset, which includes forms in eight different languages, we trained models separately on each language and tested them across all language forms. As shown in Table 12, the inter-entity relationships trained in each language

19

Table 11: Comparison between different relation heads, using F1-score as the metric. *Task 1* refers to word to text-line merging, *Task 2* refers to text-line to entity merging, *Task 4* refers to item table localization, *Task 5* refers to hierarchical structure recovery. The best average results for each task are shown in **bold**, and the best results for each language are shown in underline.

Task	Relation Head	English	Chinese	Japanese	German	French	Spanish	Italian	Portuguese	Avg.
Task 1	Merger [38]	0.9081	0.9360	0.9118	0.9255	0.9282	0.9372	0.9157	0.9387	0.9260
	Biaffine [9]	0.9167	0.9493	0.9124	0.9299	0.9309	0.9417	0.9234	0.9500	0.9329
	GeoLayout [29]	0.9175	0.9560	0.9161	0.9365	0.9395	0.9393	0.9240	0.9479	**0.9355**
Task 2	Merger [38]	0.9151	0.9681	0.9387	0.9157	0.9408	0.9463	0.9280	0.9594	0.9412
	Biaffine [9]	0.9286	0.9737	0.9361	0.9277	0.9487	0.9581	0.9334	0.9649	**0.9482**
	GeoLayout [29]	0.9277	0.9753	0.9405	0.9227	0.9433	0.9540	0.9376	0.9619	0.9473
Task 4	Merger [38]	0.7273	0.3333	0.1053	0.4348	0.1053	0.0588	0.3158	0.1250	**0.3022**
	Biaffine [9]	0.3913	0.3200	0.0952	0.6000	0.3478	0.0571	0.2000	0.0392	0.2474
	GeoLayout [29]	0.5000	0.3143	0.1053	0.6250	0.0000	0.1935	0.2000	0.1702	0.2707
Task 5	Merger [38]	0.7135	0.7601	0.6626	0.7734	0.7415	0.7009	0.6710	0.6310	0.7013
	Biaffine [9]	0.7172	0.7737	0.6382	0.7586	0.7452	0.7205	0.6811	0.6097	0.6985
	GeoLayout [29]	0.7623	0.8171	0.6860	0.7999	0.7799	0.7442	0.7086	0.6415	**0.7356**

exhibited cross-lingual transferability, typically performing best in their original training languages. Additionally, a certain similarity was observed between forms of languages belonging to the Indo-European family; for instance, models trained on Spanish and Portuguese forms performed very well on German forms, even surpassing those trained directly on German forms. Furthermore, there was a significant variance in average performance across all languages depending on the training language, suggesting varying degrees of layout complexity and entity relationship complexity among different language forms in the SRFUND dataset. Portuguese forms, due to their complex structure and the highest number of entities and entity relationships as illustrated in Figure 4, achieved results only second to those models trained on the same language data, likely benefiting from their extensive entity count and relational complexity.

D.3 Details for MLLMs evaluation

Our visualized results (see Fig.6) reveal that GPT-4o tends to aggregate fine-grained elements into broader structures, whereas GPT-4o-mini more frequently outputs the input bounding boxes directly. For example, in the Word to text-line merging task, GPT-4o successfully merges words within the same line. However, in the Text-line to entity merging task, GPT-4o encounters difficulties with entity recognition, whereas GPT-4o-mini performs better by directly outputting the text line boxes specified in the prompt.

Table 12: Cross language validation experiment on *Task 5*, i.e. hierarchical structure recovery. We trained on forms in each language and tested across all languages, with the best-performing language results highlighted in **bold**.

Train \ Test	English	Chinese	Japanese	German	French	Spanish	Italian	Portuguese	Avg.
English	**0.5168**	0.3846	0.3249	0.4020	0.3714	0.3555	0.3171	0.3075	0.3634
Chinese	0.3352	**0.6105**	0.4498	0.4742	0.4664	0.4473	0.3899	0.3826	0.4524
Japanese	0.3318	0.4914	**0.5003**	0.4130	0.4108	0.3624	0.3510	0.3094	0.3999
German	0.3488	0.3778	0.2835	**0.5598**	0.4624	0.4227	0.3779	0.3358	0.3926
French	0.3892	0.4127	0.3225	0.5210	**0.5730**	0.4696	0.4633	0.3703	0.4330
Spanish	0.3859	0.4662	0.3681	**0.5485**	0.5431	0.5408	0.4637	0.4385	0.4677
Italian	0.3804	0.4241	0.3585	0.4999	0.5215	0.4759	**0.5560**	0.4272	0.4548
Portuguese	0.4137	0.4922	0.4006	**0.5603**	0.5495	0.5215	0.4807	0.4932	**0.4879**
All (Ref.)	0.7135	0.7601	0.6626	0.7734	0.7415	0.7009	0.6710	0.6310	0.7013

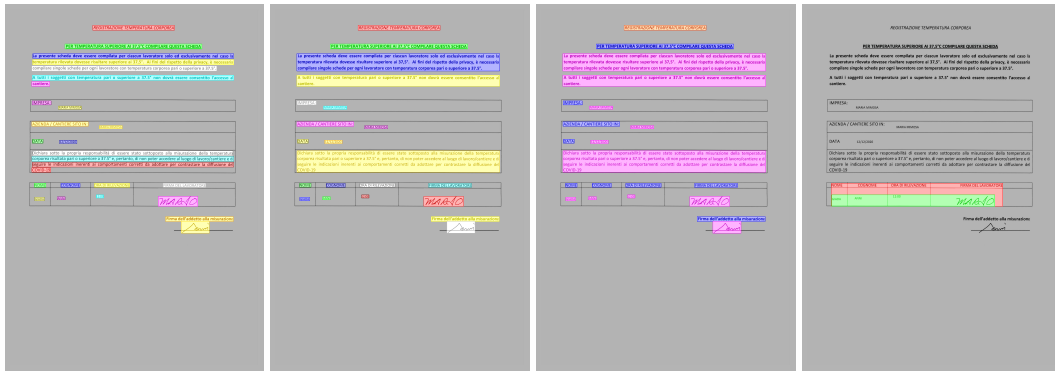
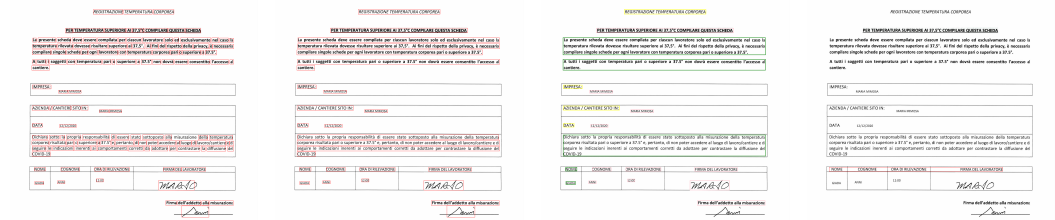
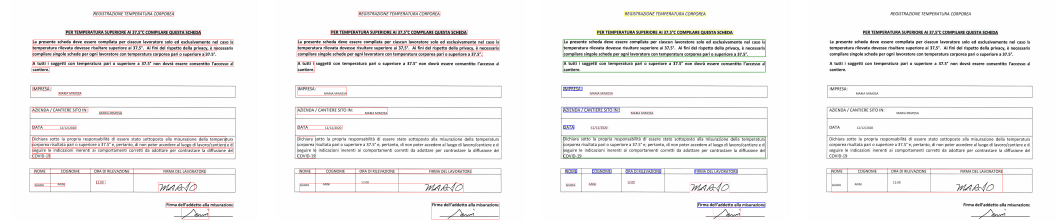
20

(a) Task 1 GT   (b) Task 2 GT   (c) Task 3 GT   (d) Task 4 GT

(e) GPT4o-mini Task 1   (f) GPT4o-mini Task 2   (g) GPT4o-mini Task 3   (h) GPT4o-mini Task 4

(i) GPT4o Task 1   (j) GPT4o Task 2   (k) GPT4o Task 3   (l) GPT4o Task 4

Figure 6: The visualization of the performance of MLLMs on the test set. The first row of images (*a* to *d*) displays the text lines/entities (without categories/with categories) boxes/row item table boxes of the image text. The second row of images (*e* to *h*) shows the predictive results of GPT4o-mini on tasks 1 to 4, and the third row of images (*i* to *l*) shows the predictive results of GPT4o. For task 3, the boxes in yellow, blue, pink, and green represent four different types of entities: Header, Question, Answer, and Other, respectively. Please zoom in for a better view.

21

Checklist

1. For all authors...

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

   (b) Did you describe the limitations of your work? [No]

   (c) Did you discuss any potential negative societal impacts of your work? [No]

   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [N/A]

   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] As a URL, please see the dataset website: `https://sprateam-ustc.github.io/SRFUND/`

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In Sec. 4

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In Sec. 4

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [Yes]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] We provide detailed annotation guidelines to ensure dataset quality. These guidelines, including specific instructions and examples, have been updated on the dataset website and can be accessed via this link: `https://sprateam-ustc.github.io/SRFUND/download/`.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

22