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.

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 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?

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.

B Further data analysis

B.1 Annotation distribution

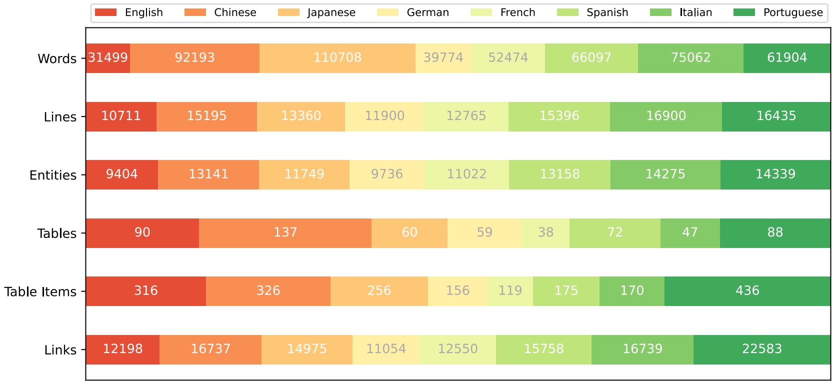


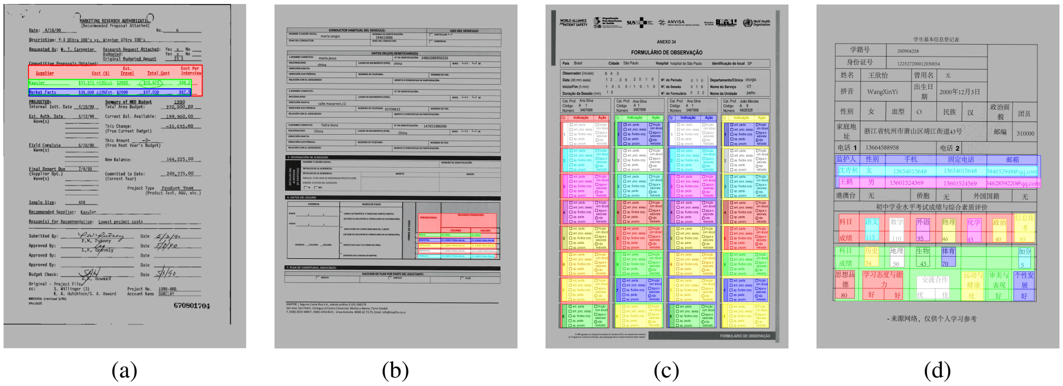
Figure 4: The number of annotations of different granularity in each language in the SRFUND dataset.

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 6i** 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 6j** 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 6k** 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 6l** 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.



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

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

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**

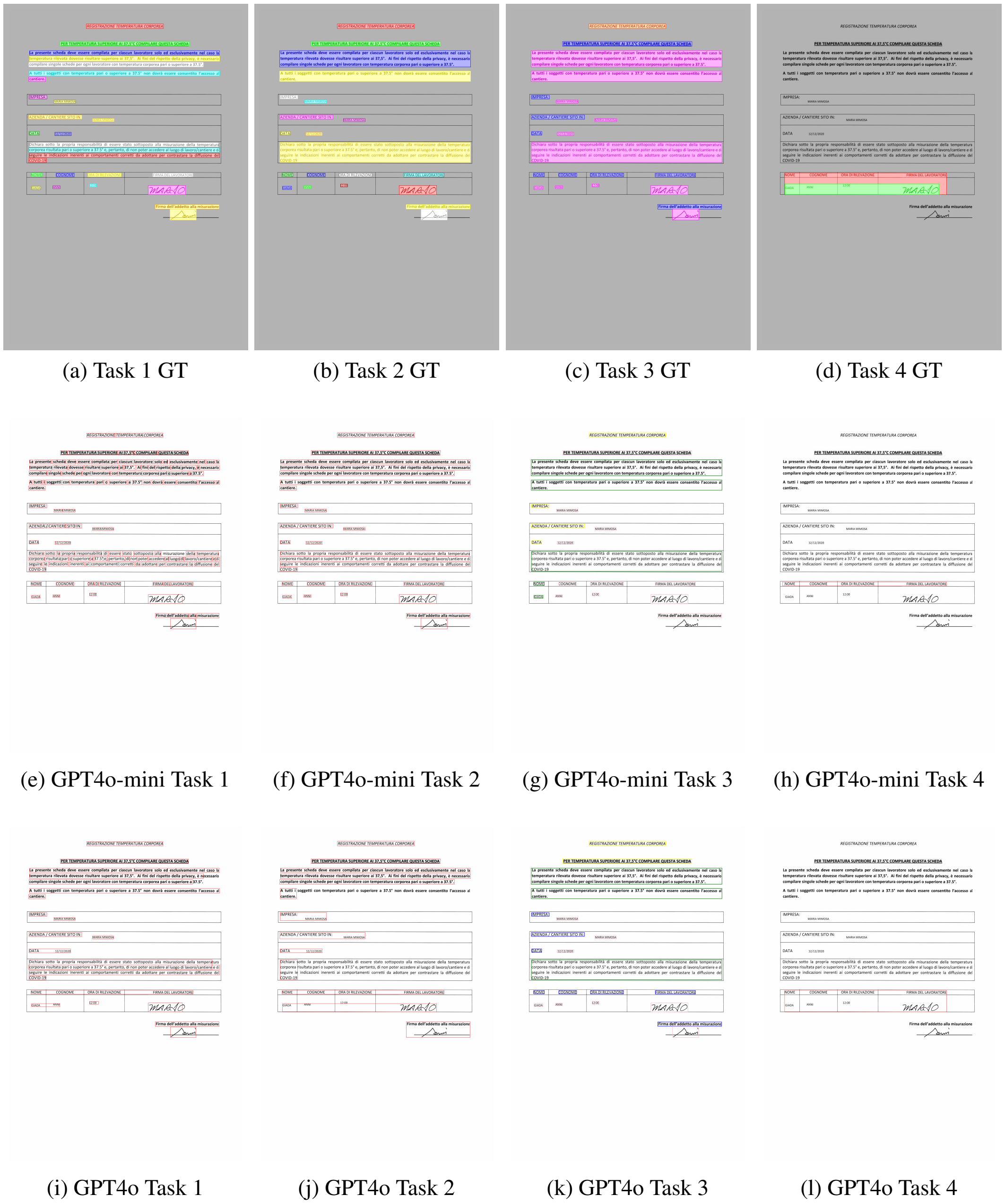
exhibited cross-lingual transferability, typically performing best in their original training languages. Additionally, a certain similarity was observed between forms of languages belonging to the Indo-European family; for instance, models trained on Spanish and Portuguese forms performed very well on 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**.

Test Train	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



(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.

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]