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

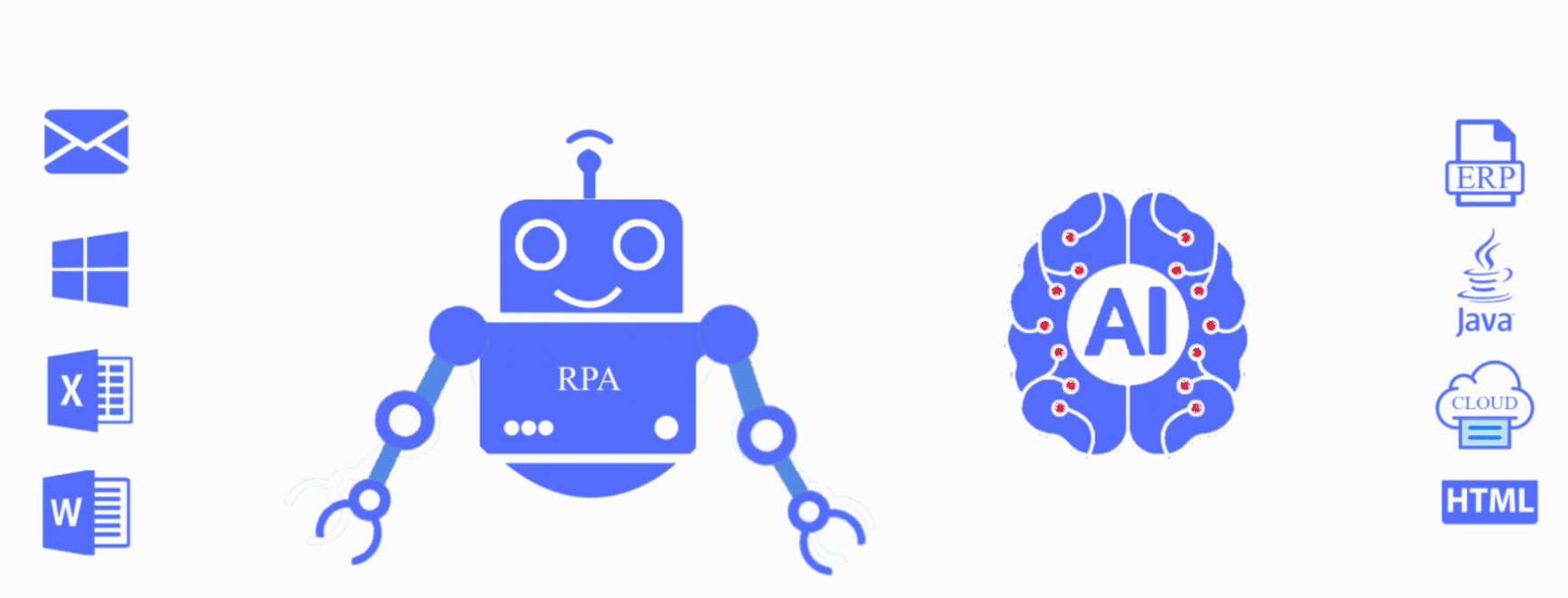
Presenter：Jiefeng Ma

Group Name： SPRAT Lab of NERC-SLIP, USTC

Project website: https://sprateam-ustc.github.io/SRFUND/

Intro Exsiting tasks on form understanding

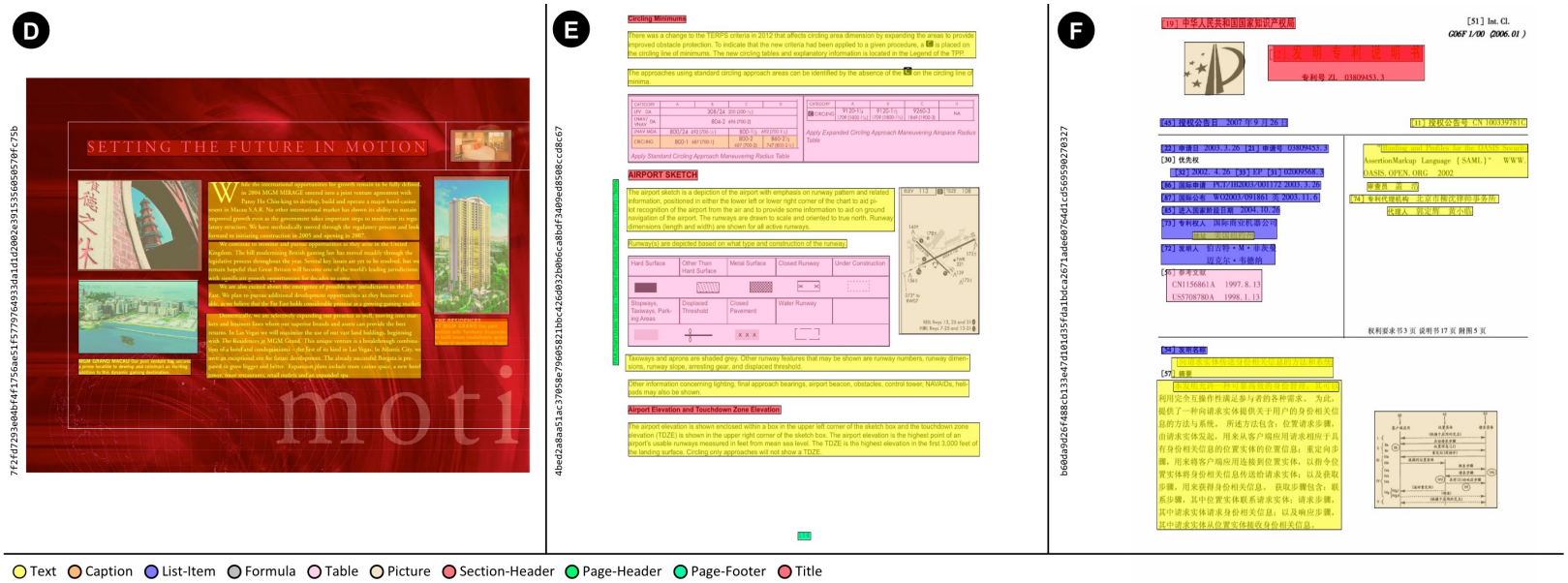
Exsiting tasks



"Documents are now at the epic-center of these dramatic process transformations and become the new engine of growth!"
--Dr. Tong Sun, July 27th, 2020

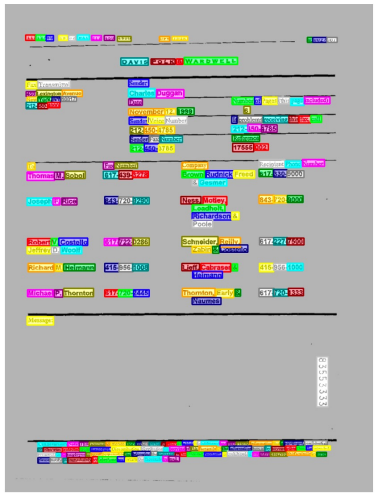
Intro One Key Promblem: localize varied layouts

Documnt Layout Analysis

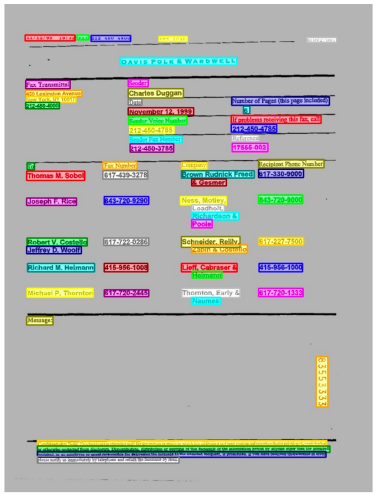


KDD 2022, "DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis"

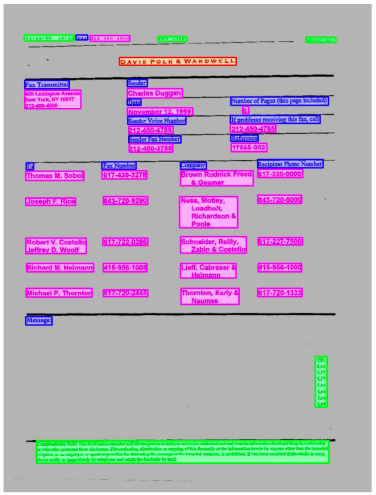
Dataset Our Work on Form Structure Definition



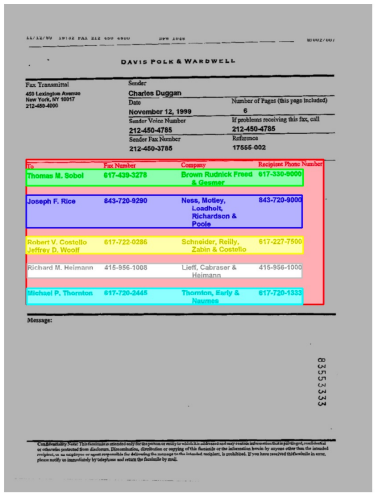
(a) Word level



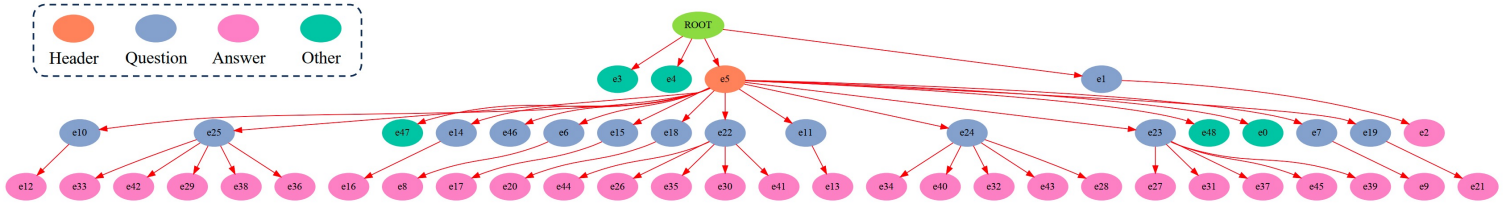
(b) Text-line level



(c) Entity level



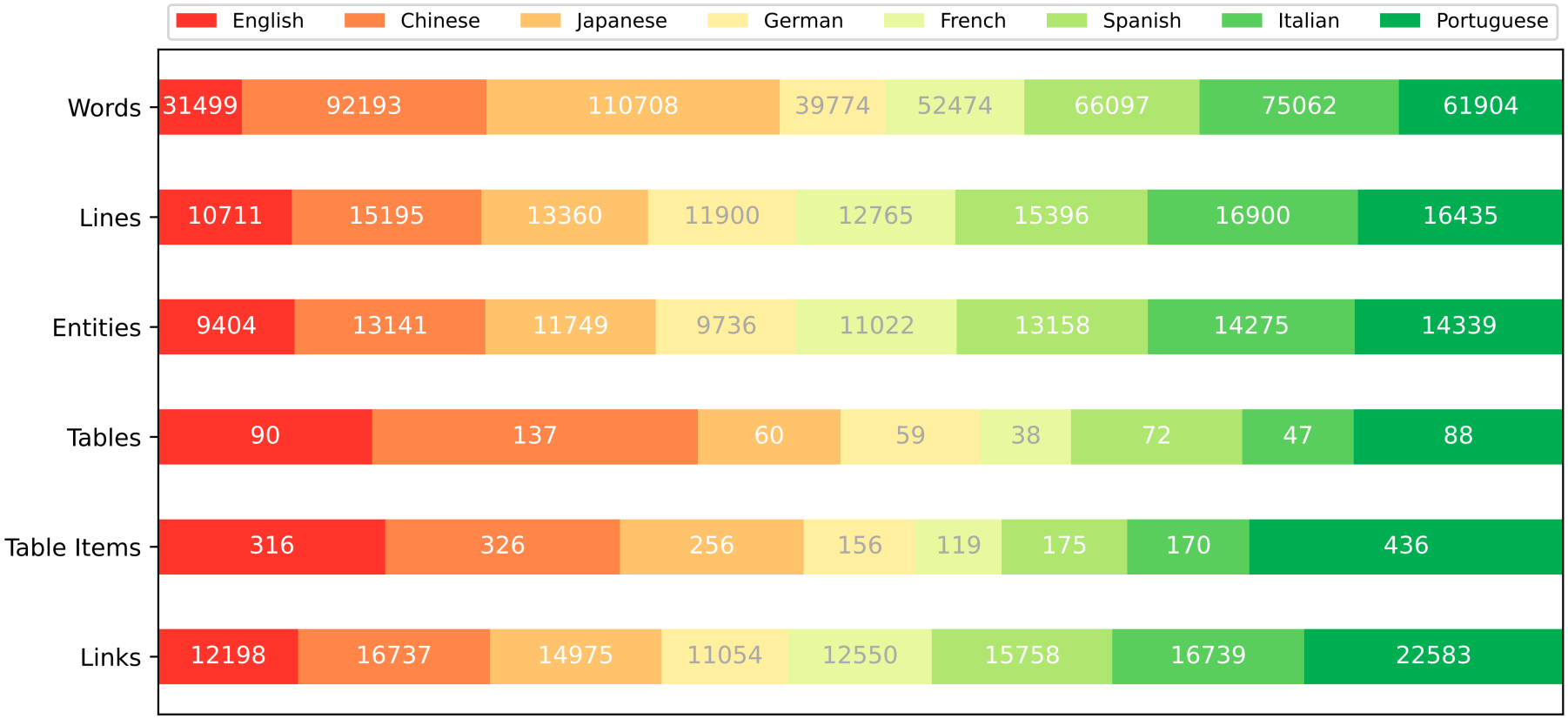
(d) Item table level



(e) Overall form structure based on entities.

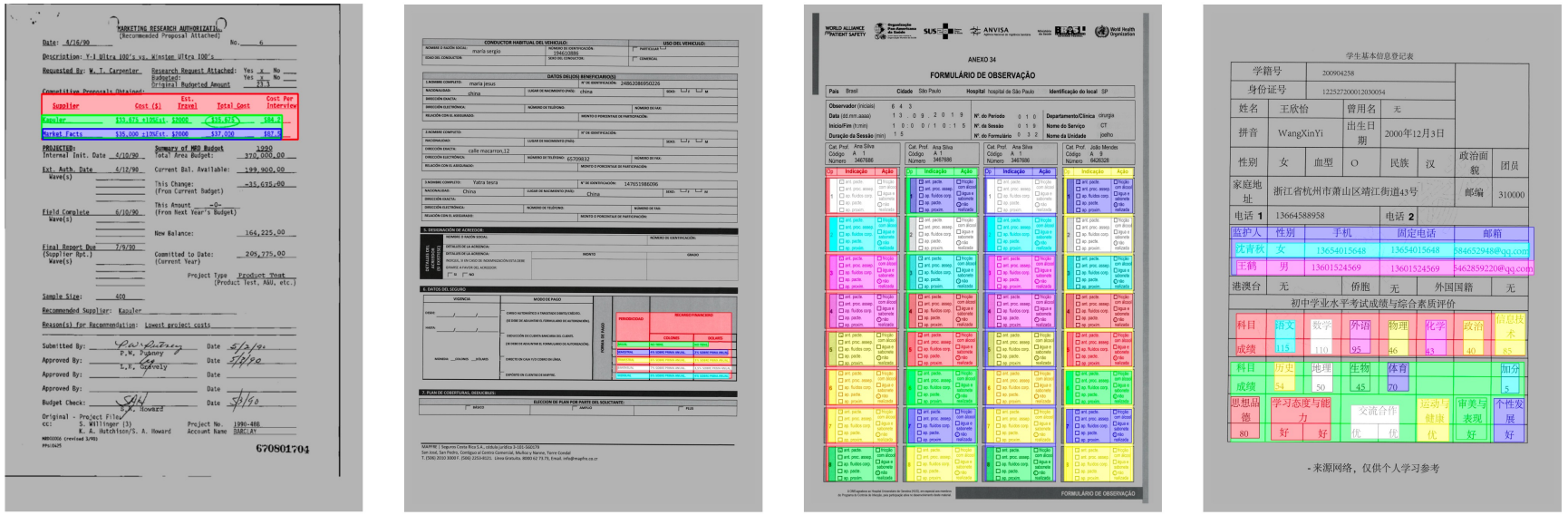
NeurIPS 2024, "SRFUND: A Multi-Granularity Hierarchical Structure Reconstruction Benchmark in Form Understanding"

Dataset Multi-granularity multi-lingual distribution



NeurIPS 2024, "SRFUND: A Multi-Granularity Hierarchical Structure Reconstruction Benchmark in Form Understanding"

Dataset What's Item-Table ?



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

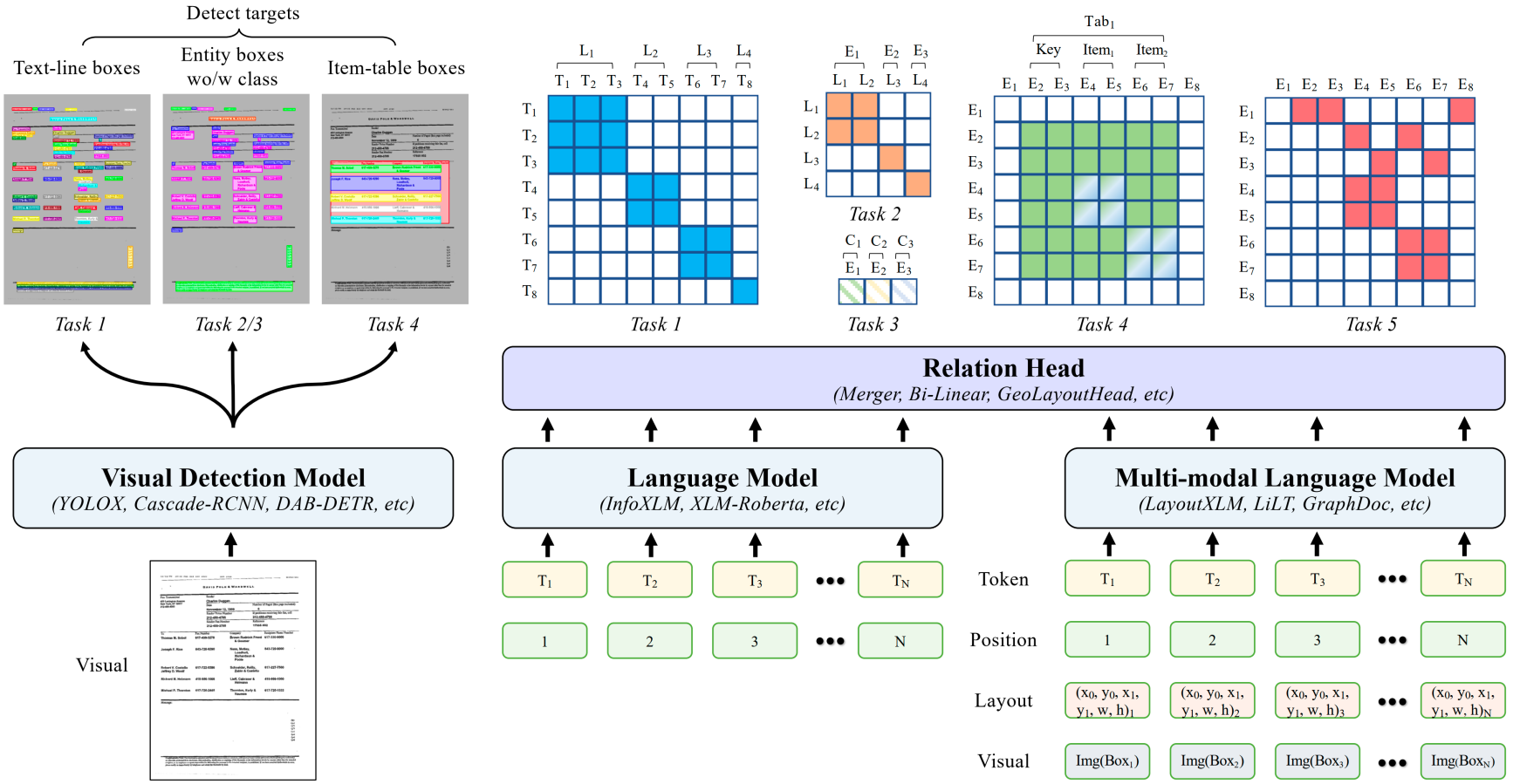
Figure 5: Varied item table annotations that are derived from diverse linguistic sources in the SRFUND dataset. Subfigures (a), (b), (c), and (d) originate from forms in English, Spanish, Portuguese, and Chinese, respectively.

Dataset Comparison with other existing benchmarks

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

Method

Multi-modal systems with vision / language / layout



NeurIPS 2024, "SRFUND: A Multi-Granularity Hierarchical Structure Reconstruction Benchmark in Form Understanding"

Result

Results of five tasks

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

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

NeurIPS 2024, "SRFUND: A Multi-Granularity Hierarchical Structure Reconstruction Benchmark in Form Understanding"

Result

Results of five tasks

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

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

NeurIPS 2024, "SRFUND: A Multi-Granularity Hierarchical Structure Reconstruction Benchmark in Form Understanding"

Result

Results of five tasks

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**

NeurIPS 2024, "SRFUND: A Multi-Granularity Hierarchical Structure Reconstruction Benchmark in Form Understanding"

Result

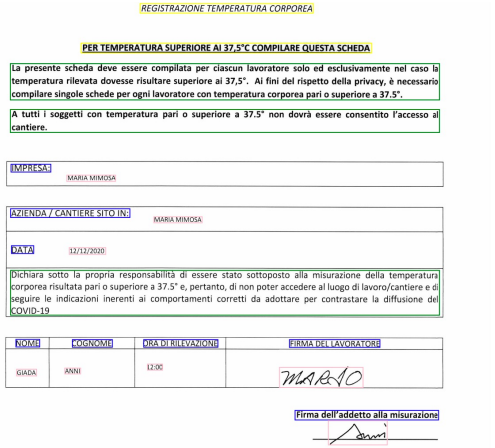
How MLLMs perform ?

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

GPT 4o Task 1

GPT 4o Task 2



GPT 4o Task 3

GPT 4o Task 4

NeurIPS 2024, "SRFUND: A Multi-Granularity Hierarchical Structure Reconstruction Benchmark in Form Understanding"

Result

Relation head matters

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**

NeurIPS 2024, "SRFUND: A Multi-Granularity Hierarchical Structure Reconstruction Benchmark in Form Understanding"

Result Cross-language validation

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

NeurIPS 2024, "SRFUND: A Multi-Granularity Hierarchical Structure Reconstruction Benchmark in Form Understanding"

Result

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 $\rightarrow$ SRFUND	0.2078 / 0.2192 / 0.2133	0.0771 / 0.2107 / 0.1128	0.1565 / 0.0476 / 0.0730
SIBR $\rightarrow$ SRFUND	- / - / -	0.0813 / 0.2859 / 0.1266	0.4322 / 0.1279 / 0.1974
SRFUND $\rightarrow$ CORD	0.8660 / 0.7821 / **0.8219**	0.9474 / 0.9309 / **0.9390**	0.1342 / 0.8169 / **0.2305**
SRFUND $\rightarrow$ SIBR	- / - / -	0.8780 / 0.7925 / **0.8331**	0.2984 / 0.6453 / **0.4081**

Thanks for listening !

Presenter：Jiefeng Ma

Group Name： SPRAT Lab of NERC-SLIP, USTC

Project website: https://sprateam-ustc.github.io/SRFUND/