TDATR: Improving End-to-End Table Recognition via Table Detail-Aware Learning and Cell-Level Visual Alignment

Supplementary Material

A. Document Data Processing

For data from different sources, we employed distinct processing workflows due to their varying formats [3, 28].

Chinese and English webpages: We render HTML file to image using khtmltopdf[1]. Then we utilized a commercial OCR[2] engine to recognize text lines on the webpages, extracting both the textual content and their corresponding coordinates.

Chinese and English papers: For papers with LaTeX source code, we first compile the LaTeX code into a PDF, and then use the PyMuPDF[3] parser to extract text lines and their coordinates from the compiled PDF. For papers available only in PDF format, we utilize a commercial engine to extract text lines and coordinates. Specifically, for mathematical formulas in papers, we employ LatexOCR[4] tool.

README files: We downloaded README files and their referenced content from various GitHub projects. First, we filter out invisible elements from the README files, such as web links, jump markers, and comments, to ensure consistency between the text and rendered images. We then used Pandoc[5] to convert the filtered README files into HTML. Finally, we utilized wkhtmltopdf to convert the HTML content into images. To limit the image size, we segmented the images and extracted the corresponding markdown content as labels.

WuKong dataset [12] and in-house data: We utilized a commercial OCR engine to recognize text lines on images.

B. Table Data Processing

Real-world table refers to images captured through photographing or scanning. Such images often contain geometric distortions, background noise, and low resolution, making recognition considerably more challenging. Digitally-born table refers to images rendered directly from code or digital documents. These images have clean characters and well-aligned layouts.

B.1. Unified Multi-source Table Data Processing

To obtain labels for the table auxiliary tasks from various dataets, we designed a unified processing pipeline.

In the first step, we unify the table label from different sources into a consistent format. In document images, we represent a table using table box, table cells, and table grids. The table box indicates the position of the table area within the document image. Table cells contain cell coordinates, logical coordinates, the text content within each cell and cell ID. Table grids represent the fine-grained structure of a table, showing the results after splitting merged cells. Each table grid includes the ID of the corresponding cell and its coordinates.

In the second step, we conduct data cleaning to eliminate inconsistently labeled table data, ensuring high data quality. First, we remove table data with overlapping logical coordinates for cells. Next, we exclude entries with incomplete table grids, specifically those where grids have not been assigned to their corresponding cells. Finally, we eliminate redundant table grids, which occur when adjacent grid rows and grid columns are identical.

The last step is training data generation. We extract table images from document images by cropping based on the table box. Table cell information is used for label generation in the table HTML parsing task, table cell detection task, and table cell spotting task. Table grid information is utilized for label generation in the table span cell detection task and the table row and column detection task.

B.2. Table Data Augmentation

High-quality labeled table data for photographic scenes is limited [57, 64]. We expanded the iFLYTAB [64] dataset inspired by an identity matrix-based augmentation method [6], resulting in the iFLYTAB-Aug dataset with 82.5k samples.

We made the following modifications to the identity matrix-based augmentation to ensure the generation of complex and realistic table data.

- We restrict the selected table regions to have more than 4 rows and columns.
- We ensure that the selected table sub-region always contains at least one span cell, and all rows and columns containing the span cell are retained. This ensures that the table has a complex structure.
- For wireless tables, the selected region always starts from the first row and the first column. Because the row and column headers provide essential information for distinguishing between rows and columns.

C. Implementation Details

In this section, we provide the detailed input–output designs of the table detail-aware learning tasks, as illustrated in the

[1] https://wkhtmltopdf.org/
[2] https://www.xfyun.cn/services/common-ocr
[3] https://github.com/pymupdf/PyMuPDF
[4] https://github.com/lukas-blecher/LaTeX-OCR
[5] https://pandoc.org/

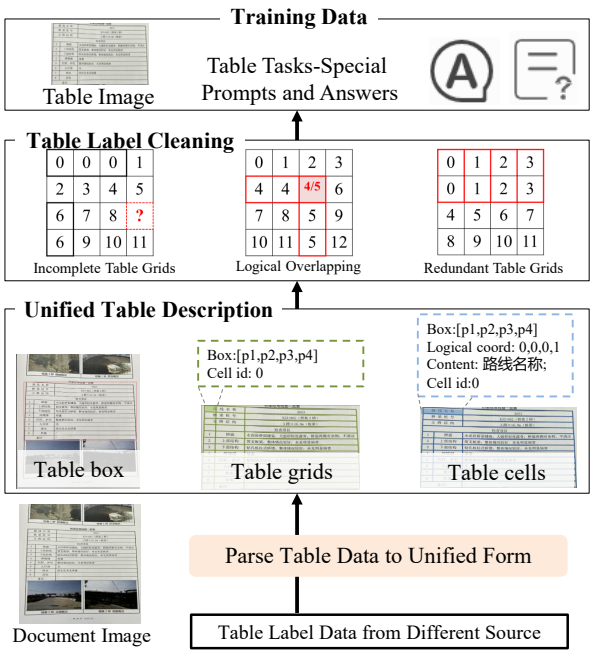


Figure 4. The pipeline of unified multi-source table data processing. The pipeline normalizes heterogeneous table annotations from various sources into a unified representation for model training.

Fig. 5 and 6.

D. Baseline Protocol

Thanks for pointing out the ambiguity in Tables. 3 and 4. Our compared baselines can be grouped into: (1) Dataset-specific setting: methods fine-tuned on each target dataset. (2) Unified setting (marked with "†"): a single checkpoint evaluated across multiple datasets. Table 9 presents the training data configurations of all baseline methods used in this paper.

E. Structure-guided Cell Localization Module

We leverage logical relationships between cells to perform bidirectional structure-guided enhancement. We take the row-based enhancement as an example to illustrate the computation process. We first project the cell representation $\mathbf{C}'$ into a row feature space using a linear layer to obtain row-level similarity features $\mathbf{C}^{row}$, as shown in Eq. 2. We then compute pairwise similarity scores via inner product to estimate whether two cells belong to the same row. After thresholding, we obtain the row similarity matrix, i.e., a binary relationship matrix indicating which cell pairs share the same row, as defined in Eq. 3. As illustrated in Fig. 7, Cell 2 and Cell 3 are in the same row, thus $M^{row}_{2,3} = 1$. This matrix is subsequently used as a mask in self-attention to reinforce feature interactions among cells within the same row.

F. Evaluation Benchmarks

iFLYTAB-full obtains 5,419 test samples. The samples come from diverse sources—including screenshots, scans, and camera-captured images—covering a wide range of image qualities that allow evaluation of model robustness. The dataset exhibits large variations in image resolution, testing the model's capability to handle multi-resolution inputs. It also contains grid tables, bordered three-line tables, and borderless tables. The absence of visible cell boundaries in borderless tables introduces significant challenges for TR.

TabRecSet contains 7,548 validation samples, all captured in real-world scenarios with strong perspective distortion and low image quality. Borderless and three-line tables are generated by erasing the ruling lines of grid tables, creating a domain gap between these synthetic styles and real-world data. The dataset includes both Chinese and English tables.

PubTabNet consists of 9,015 validation samples and 9,064 test samples, with the validation set commonly used for benchmarking. Its annotations are produced by an automated pipeline, resulting in low-resolution images and inconsistent visual-HTML alignment (e.g., cell over-segmentation). Such inconsistencies lead to contradictory training signals and may underestimate performance during evaluation. Models often require dataset-specific fine-tuning to adapt to these inconsistencies.

PubTables-1M contains 93,834 test samples and is sourced from the same corpus as PubTabNet. It applies automated consistency checks to correct the annotation inconsistencies present in PubTabNet, resulting in significantly improved label reliability.

OmniDocBench v1.5. Following PaddleOCR-VL, we crop 512 table samples from the benchmark. The dataset covers a wide spectrum of table types, including challenging note-style tables where continuous content and background ruling lines visually disrupt cell boundaries, often causing over-segmentation. Successful recognition requires semantic understanding of cell content beyond visual boundary cues.

CC-OCR. The 300 table test samples in CC-OCR cover both Chinese and English, spanning real-world and digital-document scenarios. The dataset includes long tables, dense tables, and heavily rotated cases, posing significant challenges for structure parsing and spatial reasoning.

OCRBench includes 700 table-related samples in both Chinese and English. Using the provided table boxes and our internal table detector, we crop table regions for recognition. Many samples come from financial reports, whose formatting introduces unique difficulties, e.g., large spacing between "$" and numbers is easily mistaken for column separators.

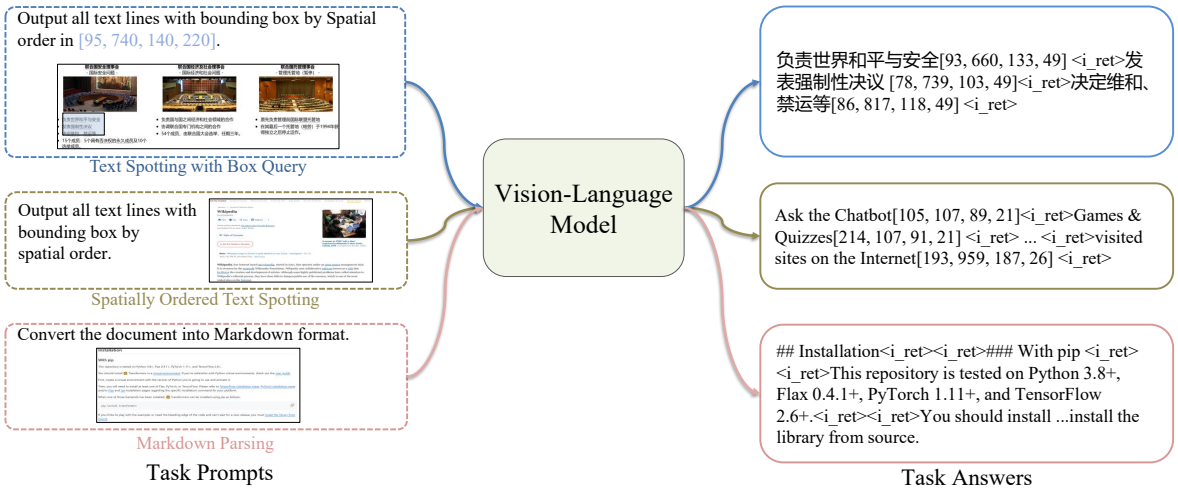


Figure 5. Illustration of table content recognition tasks. These tasks leverage diverse document data to enable text recognition, text localization, and reading-order understanding.

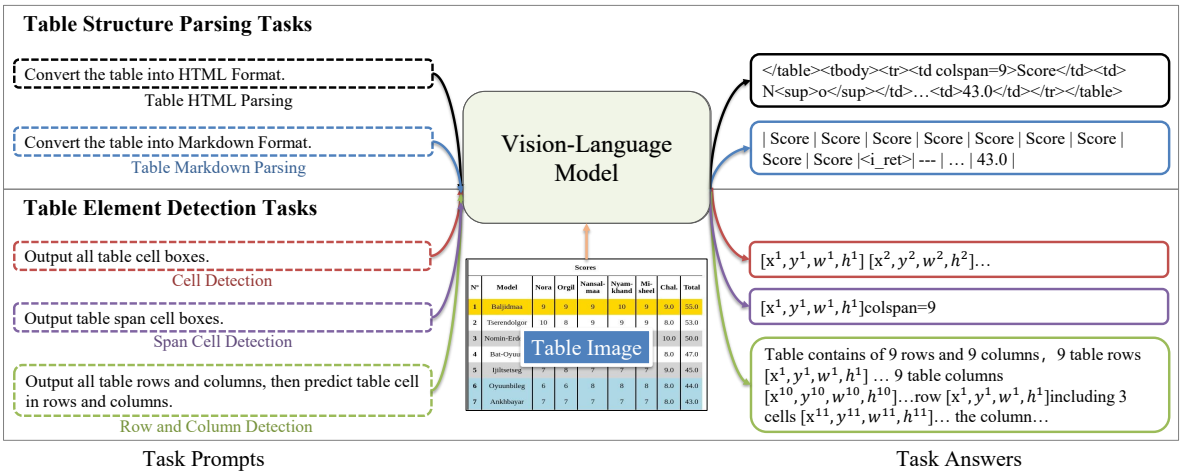


Figure 6. Illustration of table structure understanding tasks. These tasks equip the model with structure-awareness from both the cell level and the row/column level.

G. Single-dataset Training Variant

We conduct a single-dataset comparison by performing only-PubTabNet table detail fusion fine-tuning starting from our table detail-aware pretrained model. On PubTabNet-val, we achieve TEDS-S 96.78 / TEDS 96.10, outperforming the second-best TR dataset-specific baseline, TableFormer (96.75 / 93.60). This supports the effectiveness of our "perceive-then-fuse" paradigm in a single-dataset setting.

H. SGCL Inference Efficiency

TDATR leverages SGCL to localize cells in parallel, conditioned on the generated cell tokens. Since TR baselines such as Dolphin or EDD do not output cell boxes, a direct efficiency comparison is not applicable. For a fair comparison, we implement a matched baseline, "ED Loc Gen" in Table. 6, that autoregressively generates discretized coordinates after the cell tokens. We evaluate 40 randomly sampled PubTabNet images (max side length 1024), with an average of 26.75 cells and 190.38 TR tokens. Measured on an NPU with batch size 1, TDATR achieves 9.7s end-to-end latency, compared to 15.7s for the baseline (1.6× faster). Importantly, SGCL contributes only 0.28s to the end-to-end latency, confirming that parallel refinement keeps localization overhead low. TDATR and the baselines have comparable max reserved memory (15.77 GiB vs. 15.36 GiB).

Table 9. Summary of the training data configurations of the baseline methods. For each method, we report the paradigm, table training data, auxiliary data, whether table-specific fine-tuning is applied, and additional notes.

Method	Paradigm	Table training data	Extra data	Dataset specific	Notes
TableMaster	TSR	PubTabNet	–	Yes	–
LORE	TSR	PubTabNet, TabRecSet, and iFLYTAB	–	Yes	20k images were randomly sampled from PubTabNet for training. TabRecSet and iFLYTAB were reproduced by us based on the released code.
BGTR (PT)	TSR	TabRecSet, iFLYTAB, PubTabNet, FinTabNet and SynthTabNet	–	Yes	–
UniTabNet	TSR	iFLYTAB, PubTables-1M, and PubTabNet.	Pre-training: a synthetic dataset comprising 1.4 million Chinese and English samples from SynthDog, and PubTables-1M	Yes	–
EDD	E2E-TR	PubTabNet	–	Yes	–
SEMv3 + PPOCR	M-TR	PubTabNet and iFLYTAB	PPOCR relies on general text recognition data	Yes	"+PPOCR" indicates that the cell content is obtained from the PPOCR model.
GTE	TSR	PubTabNet and FinTabNet	–	Yes	The model is pre-trained on PubTabNet and fine-tuned on multiple datasets.
Davar-Lab	TSR	PubTabNet	–	Yes	–
LGPMA + R2AM	M-TR	PubTabNet	Additional data required by R2AM	Yes	"+R2AM" indicates that the cell content is obtained from the R2AM model.
TableFormer + GT	M-TR	PubTabNet, FinTabNet, and SynthTabNet	–	Yes	"+GT" indicates that the ground-truth cell content.
RapidTable	M-TR	–	–	–	–
OmniParser	OCR-VLM	PubTabNet and FinTabNet	Large-scale document parsing data	Yes	–
DocOwl1.5	OCR-VLM	TURL and PubTabNet	Unified structure-learning data from documents, webpages, charts, and natural images	No	–
Dolphin	OCR-VLM	PubTabNet and PubTab1M	Large-scale document parsing data	Yes	–
MinerU2.5	OCR-VLM	In-house	Large-scale document parsing data	No	–
DeepSeek-OCR	OCR-VLM	In-house	Large-scale document parsing data	No	–
PaddleOCR-VL	OCR-VLM	In-house	Large-scale document parsing data	No	–
dots.ocr	OCR-VLM	In-house	Large-scale document parsing data	No	–
GOT	OCR-VLM	In-house	Large-scale document parsing data	No	–
TabPedia	TSR	PubTabNet and PubTab1M	–	No	–
DETR + PDF	M-TR	PubTables-1M	–	Yes	"+PDF" indicates that the cell content is obtained from the source PDF files.

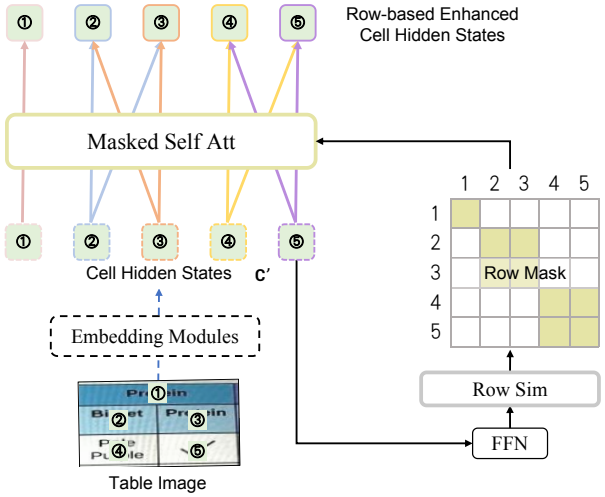


Figure 7. The architecture of the structure-guided cell localization module, illustrated with the row-based cell enhancement example.

I. Visualization of Table Recognition

We visualize several challenging table samples. Real-world tables (Fig. 8) contain background noise, perspective distortion, and uneven illumination. Long tables (Fig. 9) feature lengthy sequences, numerous cells, and long text contents. Complex-structure tables (Fig. 10) include extensive row or column spanning. Our method performs robustly across all these cases, demonstrating strong generalization and effectiveness.

配件	品牌型号	价格
CPU	intel 酷睿i9 12900KF 盒装	¥4699
散热器	华硕ROG 龙神360 水冷	¥1799
主板	华硕ROG STRIX Z690-A GAMING WIFI DDR4 吹雪	¥2699
内存	金士顿骇客16G*2 3200 DDR4	¥798
显卡	华硕TUF-RTX3070-08G-GAMING	¥6099
硬盘	西数SN850 1T NVME PCIe4.0	¥1099
电源	鑫谷昆仑KL850 金牌全模组	¥1089
机箱	安钛克复仇者X	¥429
键鼠装	自选	---
显示器	自选	---
报价：	18711元	

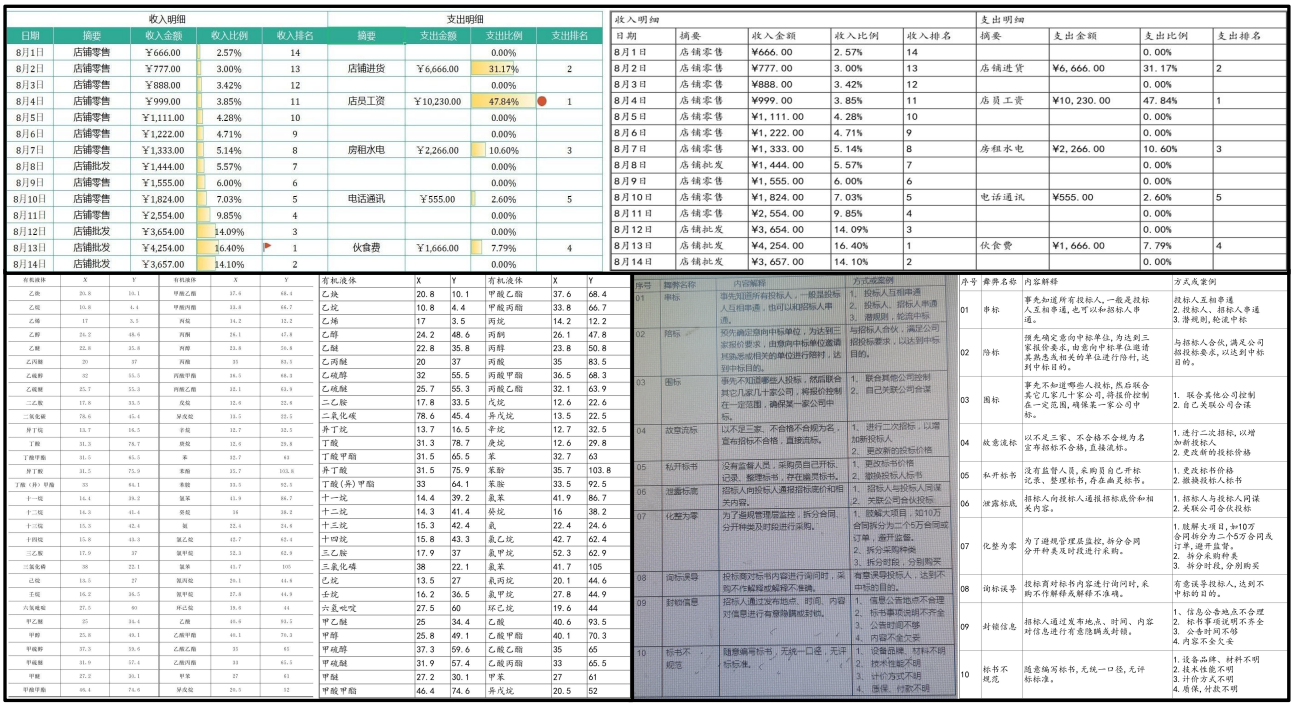
配件	品牌型号	价格
CPU	intel 酷睿i9 12900KF 盒装	¥4699
散热器	华硕ROG龙神360水冷	¥1799
主板	华硕ROGSTRIXZ690-AGAMINGWIFIDDR4吹雪	¥2699
内存	金士顿骇客16G*2320ODDR4	¥798
显卡	华硕TUF-RTX3070-08G-GAMING	¥6099
硬盘	西数SN850ITNVMEPCIe4.0	¥1099
电源	鑫谷昆仑KL850金牌全模组	¥1089
机箱	安钛克复仇者X	¥429
键鼠装	自选	----
显示器	自选	----
报价：	18711元	

项目	每100毫升	营养素参考值%
能量	192千焦	2%
蛋白质	0.7克	15
脂肪	0克	0%
碳水化合物	9.9克	3%
—糖	9.9克	
膳食纤维	1.5克	6%
(以酸葡萄糖计)		
钠	35毫克	2%

每份40g，本包装含有34份		
项目	每份	NRV%
能量	544kJ	6%
蛋白质	0g	0%
脂肪	0g	0%
碳水化合物	33.0g	11%
-糖	26.0g	-
膳食纤维	3.0g	12%

地域	概念
大洲	大陆和它附近的岛屿
大陆	面积广大的陆地
岛屿	面积较小的陆地
洋	海洋的中心部分
海	是洋的一部分，位于大洋边缘，面积较小，靠近大陆
海峡	沟通两个海洋之间的狭窄水道

Figure 8. Visualization of table recognition results on real-world tables. In each subfigure, the left shows the input original table image, and the right presents the HTML-rendered visualization of the corresponding recognition result.



日期	摘要	收入金额	收入比例	收入排名	摘要	支出金额	支出比例	支出排名
	收入明细				支出明细			
8月1日	店铺零售	¥666.00	2.57%	14			0.00%	
8月2日	店铺零售	¥777.00	3.00%	13	店铺进货	¥6,666.00	31.17%	2
8月3日	店铺零售	¥888.00	3.42%	12			0.00%	
8月4日	店铺零售	¥999.00	3.85%	11	店员工资	¥10,230.00	47.84%	1
8月5日	店铺零售	¥1,111.00	4.28%	10			0.00%	
8月6日	店铺零售	¥1,222.00	4.71%	9			0.00%	
8月7日	店铺零售	¥1,333.00	5.14%	8	房租水电	¥2,266.00	10.60%	3
8月8日	店铺批发	¥1,444.00	5.57%	7			0.00%	
8月9日	店铺零售	¥1,555.00	6.00%	6			0.00%	
8月10日	店铺零售	¥1,824.00	7.03%	5	电话通讯	¥555.00	2.60%	5
8月11日	店铺零售	¥2,554.00	9.85%	4			0.00%	
8月12日	店铺批发	¥3,654.00	14.09%	3			0.00%	
8月13日	店铺批发	¥4,254.00	16.40%	1	伙食费	¥1,666.00	7.79%	4
8月14日	店铺批发	¥3,657.00	14.10%	2			0.00%	

日期	摘要	收入金额	收入比例	收入排名	支出明细			
	收入明细							
8月1日	店铺零售	¥666.00	2.57%	14			0.00%	
8月2日	店铺零售	¥777.00	3.00%	13	店铺进货	¥6,666.00	31.17%	2
8月3日	店铺零售	¥888.00	3.42%	12			0.00%	
8月4日	店铺零售	¥999.00	3.85%	11	店员工资	¥10,230.00	47.84%	1
8月5日	店铺零售	¥1,111.00	4.28%	10			0.00%	
8月6日	店铺零售	¥1,222.00	4.71%	9			0.00%	
8月7日	店铺零售	¥1,333.00	5.14%	8	房租水电	¥2,266.00	10.60%	3
8月8日	店铺批发	¥1,444.00	5.57%	7			0.00%	
8月9日	店铺零售	¥1,555.00	6.00%	6			0.00%	
8月10日	店铺零售	¥1,824.00	7.03%	5	电话通讯	¥555.00	2.60%	5
8月11日	店铺零售	¥2,554.00	9.85%	4			0.00%	
8月12日	店铺批发	¥3,654.00	14.09%	3			0.00%	
8月13日	店铺批发	¥4,254.00	16.40%	1	伙食费	¥1,666.00	7.79%	4
8月14日	店铺批发	¥3,657.00	14.10%	2			0.00%	

有机液体	X	Y	有机液体	X	Y
乙烷	20.8	10.1	甲酸乙酯	37.6	68.4
乙烷	10.8	4.4	甲酸丙酯	33.8	66.7
乙烯	17	3.5	丙烷	14.2	12.2
乙醇	24.2	48.6	丙酮	26.1	47.8
乙醚	22.8	35.8	丙醇	23.8	50.8
乙丙醚	20	37	丙酸	35	83.5
乙硫醇	32	55.5	丙酸甲酯	36.5	68.3
乙硫醚	25.7	55.3	丙酸乙酯	32.1	63.9
二乙胺	17.8	33.5	戊烷	12.6	22.6
二氯化碳	78.6	45.4	异戊烷	13.5	22.5
异丁烷	13.7	16.5	辛烷	12.7	32.5
丁酸	31.3	78.7	庚烷	12.6	29.8
丁酸甲酯	31.5	65.5	苯	32.7	63
异丁酸	31.5	75.9	苯酚	35.7	103.8
丁酸(异)甲酯	33	64.1	苯胺	33.5	92.5
十一烷	14.4	39.2	氨苯	41.9	86.7
十二烷	14.3	41.4	萘烷	16	38.2
十三烷	15.3	42.4	氨	22.4	24.6
十四烷	15.8	43.3	氨乙烷	42.7	62.4
三乙胺	17.9	37	氨甲烷	52.3	62.9
三氯化磷	38	22.1	氨苯	41.7	105
己烷	13.5	27	氨丙烷	20.1	44.6
壬烷	16.2	36.5	氨甲烷	27.8	44.9
六氢吡啶	27.5	60	环己烷	19.6	44
甲乙醚	25	34.4	乙酸	40.6	93.5
甲醇	25.8	49.1	乙酸甲酯	40.1	70.3
甲硫醇	37.3	59.6	乙酸乙酯	35	65
甲硫醚	31.9	57.4	乙酸丙酯	33	65.5
甲醚	27.2	30.1	甲苯	27	61
甲酸甲酯	46.4	74.6	异戊烷	20.5	52

序号	商界名称	内容解释	方式或案例
01	串标	事先知道所有投标人，一般是投标人互相申通，也可以和招标人串通。	1. 投标人互相申通 2. 投标人、招标人串通 3. 潜规则，轮流中标
02	陪标	预先确定意向中标单位，为达到三家报价要求，由意向中标单位邀请其熟悉或相关的单位进行陪标，达到中标目的。	与招标人合伙，满足公司招投标要求，以达到中标目的。
03	围标	事先不知道哪些人投标，然后联合其它几家几十家公司，将报价控制在一定范围，确保某一家公司中标。	1. 联合其他公司控制 2. 自己关联公司合谋
04	故意流标	以不足三家、不合格不合规为名，宣布招标不合格，直接流标。	1. 进行二次招标，以增加新投标人 2. 更改新的投标价格
05	私开标书	没有监督人员，采购员自己开标、记录、整理标书，存在幕黑标书。	1. 更改标书价格 2. 撤换投标人标书
06	泄露标底	招标人向投标人通报招标底价和相关内容。	1. 招标人与投标人同谋 2. 关联公司合伙投标
07	化整为零	为了逃规管理层监控，拆分合同，分开种类及时段进行采购。	1. 就解大项目，如10万合同拆分为二个5万台同或订单，避开监管。 2. 拆分采购种类 3. 拆分时段，分别购买
08	询标误导	投标商对标书内容进行询问时，采购不作解释或解释不准确。	有意误导投标人，达到不中标的目的。
09	封标信息	招标人通过发布地点、时间、内容对信息进行有意隐瞒或封标。	1. 信息公告地点不合理 2. 标书事项说明不完全 3. 公告时间不够 4. 内容不全欠妥
10	标书不规范	随意编写标书，无统一口径，无评标标准。	1. 设备品牌、材料不明 2. 技术性能不明 3. 计价方式不明 4. 服保、付款不明

序号	商界名称	内容解释	方式或案例
01	串标	事先知道所有投标人，一般是投标人互相申通，也可以和招标人串通。	投标人互相申通 2. 投标人、招标人串通 3. 潜规则，轮流中标
02	陪标	预先确定意向中标单位，为达到三家报价要求，由意向中标单位邀请其熟悉或相关的单位进行陪标，达到中标目的。	与招标人合伙，满足公司招投标要求，以达到中标目的。
03	围标	事先不知道哪些人投标，然后联合其它几家几十家公司，将报价控制在一定范围，确保某一家公司中标。	1. 联合其他公司控制 2. 自己关联公司合谋
04	故意流标	以不足三家、不合格不合规为名，宣布招标不合格，直接流标。	1. 进行二次招标，以增加新投标人 2. 更改新的投标价格
05	私开标书	没有监督人员，采购员自己开标、记录、整理标书，存在幕黑标书。	1. 更改标书价格 2. 撤换投标人标书
06	泄露标底	招标人向投标人通报招标底价和相关内容。	1. 招标人与投标人同谋 2. 关联公司合伙投标
07	化整为零	为了逃规管理层监控，拆分合同，分开种类及时段进行采购。	1. 就解大项目，如10万合同拆分为二个5万台同或订单，避开监管。 2. 拆分采购种类 3. 拆分时段，分别购买
08	询标误导	投标商对标书内容进行询问时，采购不作解释或解释不准确。	有意误导投标人，达到不中标的目的。
09	封标信息	招标人通过发布地点、时间、内容对信息进行有意隐瞒或封标。	1、信息公告地点不合理 2、标书事项说明不完全 3、公告时间不够 4、内容不全欠妥
10	标书不规范	随意编写标书，无统一口径，无评标标准。	1. 设备品牌、材料不明 2. 技术性能不明 3. 计价方式不明 4. 质保、付款不明

Figure 9. Visualization of table recognition results on long tables. In each subfigure, the left shows the input original table image, and the right presents the HTML-rendered visualization of the corresponding recognition result.

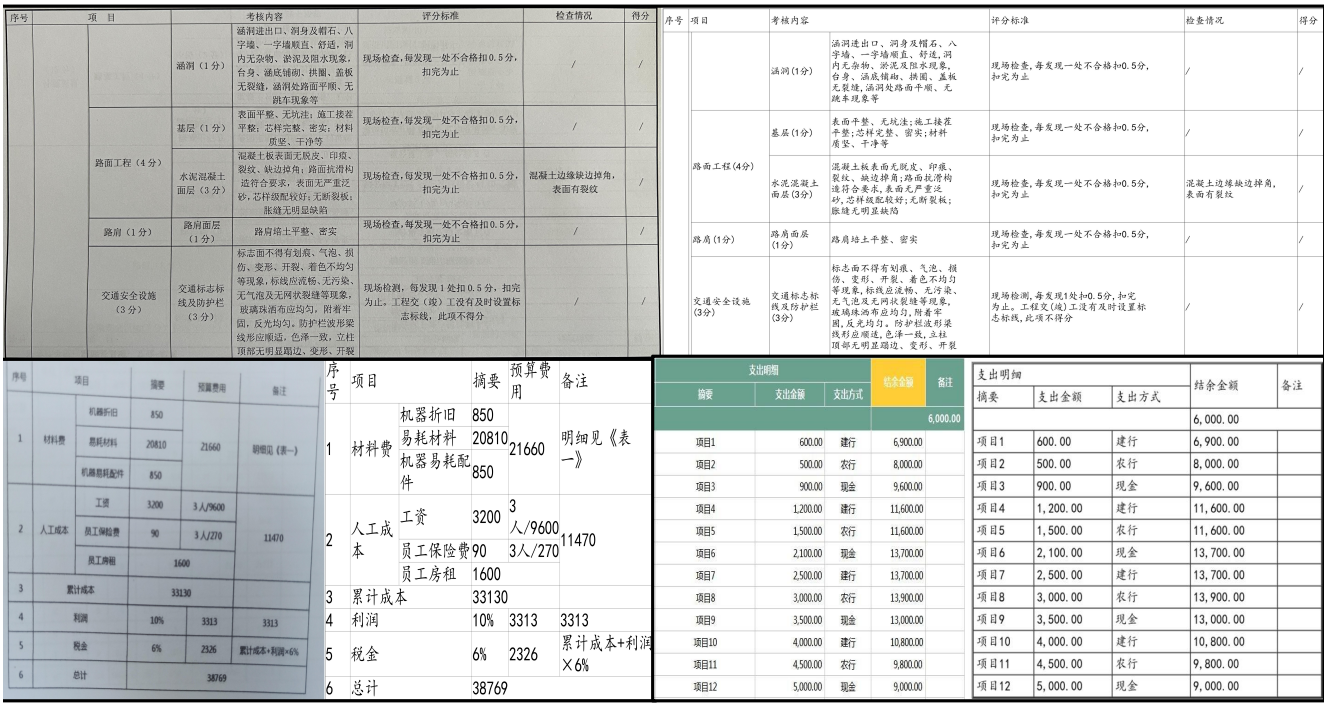


Figure 10. Visualization of table recognition results complex-structure tables. In each subfigure, the left shows the input original table image, and the right presents the HTML-rendered visualization of the corresponding recognition result.

J. Visualization of Cell Localization

We qualitatively compare the cell localization results of several SOTA models,as shown in Fig. 11. SEMv3, which follows a "split-and-merge" strategy by detecting row/column separators to form table grids, is prone to confusing separators with inter-word gaps. LORE employs CornerNet for cell localization and relies solely on visual cues, making it unreliable for empty cells. UniTabNet predicts cell boxes through a single cell token, but compressing spatial information into one token limits its performance on dense tables. "ED Loc Gen" generates cell coordinates sequentially, resulting in excessively long answer sequences that are easily truncated on long tables.

K. Failure Cases Analysis

We observed that these hard cases on iFLYTAB-full are mainly fall into three main error types: (1) Boundary confusion: In borderless tables containing multi-line text the model struggles to distinguish text line spacing from cell delimiters. (2) Span number errors: For cells with large row/column spans (more than 15), the model occasionally predicts the error number. (3) Localization instability: Dense and empty borderless cells lack explicit visual cues causing instability visual-based regression in SGCL. We plan to enhance the decoder's semantic reasoning to resolve these visual ambiguities.

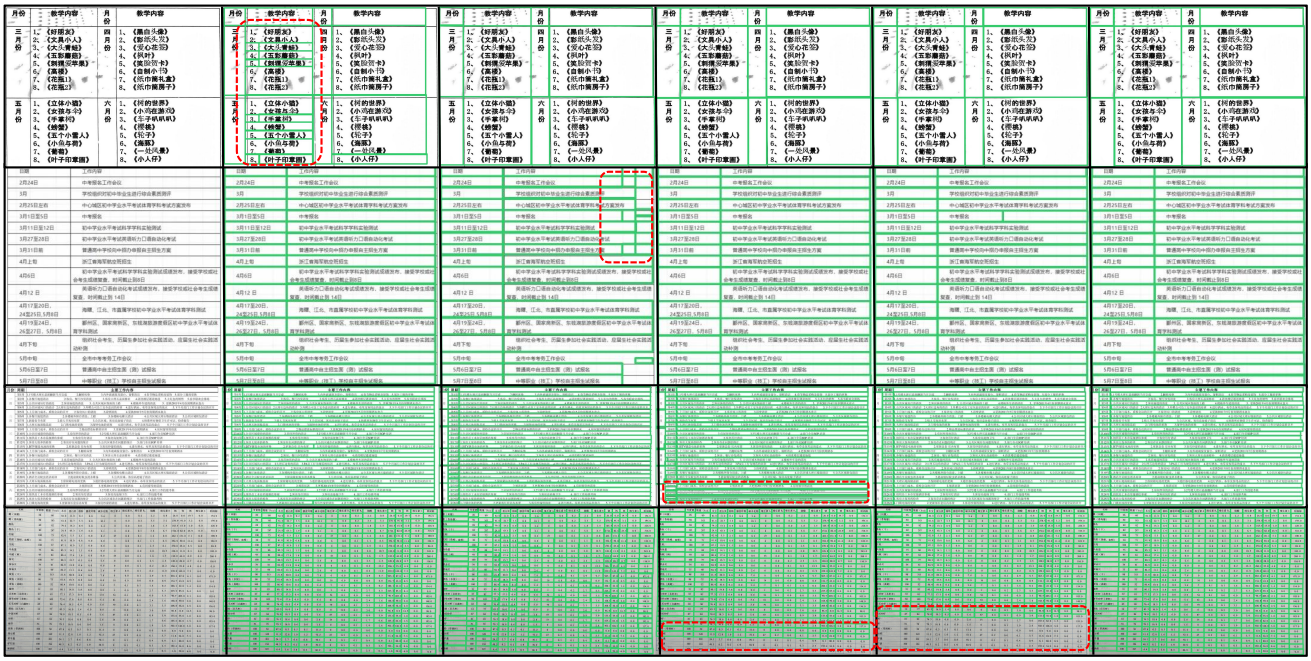


Table Image | SEMv3 | LORE | UniTabNet | ED Loc Gen | TDATR

Figure 11. Qualitatively comparison of the cell localization.