StrucTab: A Structured Optimization Framework for Table Parsing

Gengluo Li[1,4,⋆], Shangpin Peng[2,5,6,⋆], Chengquan Zhang[2,†], Binghong Wu[2], Hao Feng[2], Weinong Wang[2], Pengyuan Lyu[2], Huawen Shen[1,4], Xingyu Wan[2], Zhuotao Tian[5], Han Hu[2], Can Ma[1,4], and Yu Zhou[3,✉]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] LLM Department, Tencent, China
[3] Nankai University, China
[4] School of Cyber Security, University of Chinese Academy of Sciences, China
[5] Shenzhen Loop Area Institute, China
[6] Hong Kong University of Science and Technology, China

ligengluo@iie.ac.cn pspdada0808@gmail.com yzhou@nankai.edu.cn

Abstract. Table parsing aims to convert table images into structured, machine-readable representations, a task requiring the joint perception of complex spatial layouts and textual content. While recent vision-language models (VLMs) enable end-to-end parsing, they typically rely on direct supervision of the final output, thereby bypassing the explicit intermediate reasoning that is crucial for understanding complex table structures. Furthermore, attempts to optimize these models using reinforcement learning (RL) are often hindered by unstable or ambiguous reward designs, limiting potential performance gains. To address these limitations, we propose **StrucTab**, a table parsing model learned through intermediate structural supervision and reward decomposition. At the modeling level, by decomposing the parsing process into human-inspired subtasks, such as row-column counting and merged-cell analysis, StrucTab progressively unifies them through a sequential reasoning strategy. At the optimization level, we introduce **Uni-TabRL**, a unified RL framework that leverages decomposed rewards (validity, structure, and content) to provide stable and informative optimization signals. Finally, at the evaluation level, we present **TableVerse-5K**, a large-scale, challenging benchmark encompassing diverse, real-world table scenarios. Extensive experiments demonstrate the state-of-the-art performance of StrucTab across all evaluated public benchmarks and significant improvements on TableVerse-5K, validating the effectiveness of explicit structural modeling and decomposed reward optimization. Code and benchmark are publicly available at `https://github.com/VirtualLUOUCAS/StrucTab`.

Keywords: Table Parsing · VLMs · Structured Reasoning · RL

⋆ Equal contribution. † Project leader. ✉ Corresponding author.

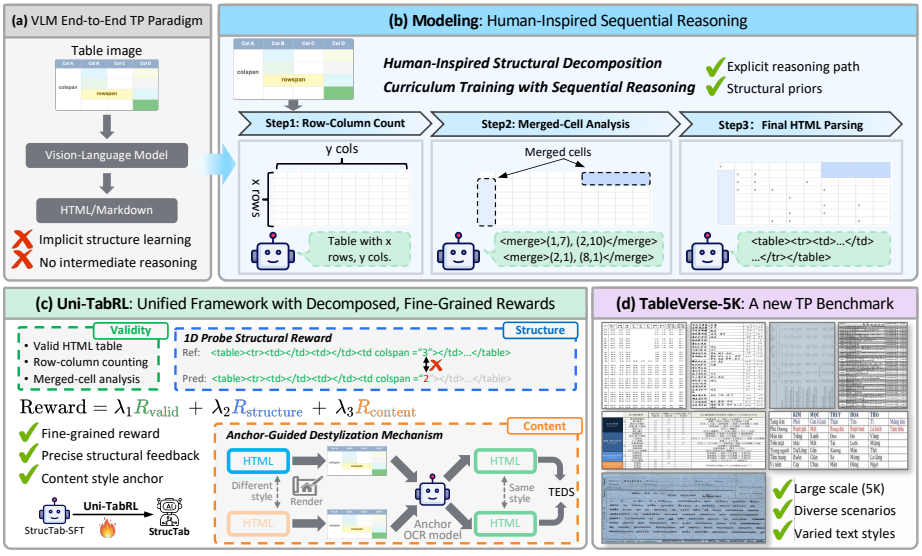


Fig. 1: StrucTab: A table parsing framework that overcomes the limitations of (a) current VLM-based end-to-end paradigms through three key innovations: **(b) Modeling** via human-inspired sequential reasoning; **(c) Optimization** with **Uni-TabRL** using decomposed, fine-grained rewards; and **(d) Evaluation** on **TableVerse-5K**, a large-scale, challenging benchmark of diverse real-world tables.

1 Introduction

Table parsing is a fundamental problem in document understanding, which aims to convert table images into structured, machine-readable representations such as HTML or Markdown [23,38,41,54,62]. Unlike plain text, tables exhibit complex spatial layouts and logical relationships, requiring models to jointly perceive both structural organization and textual content [14,47,65,68]. This intrinsic complexity makes robust table parsing particularly challenging in real-world scenarios.

The methodology for table parsing has undergone a significant paradigm shift from cascaded pipelines to end-to-end generation. Traditionally, this task relied on cascaded pipelines that decoupled structure recognition from optical character recognition (OCR) [3, 11, 20, 26, 48, 50]. While effective in constrained settings, such cascaded designs inevitably suffer from error propagation and limited robustness in diverse, in-the-wild layouts [25]. To address these limitations, VLMs have revolutionized the field by directly parsing images to structured codes [4, 17, 19, 36, 43]. However, as shown in Fig. 1 (a), although this unified modeling simplifies the pipeline, it primarily relies on the autoregressive generation of a linearized sequence, treating table parsing as a flat, one-dimensional image-to-text task. Since tables are inherently structured entities where logical relationships among rows, columns, and merged cells are paramount, this reliance on end-to-end generation raises a fundamental question: *Can we truly*

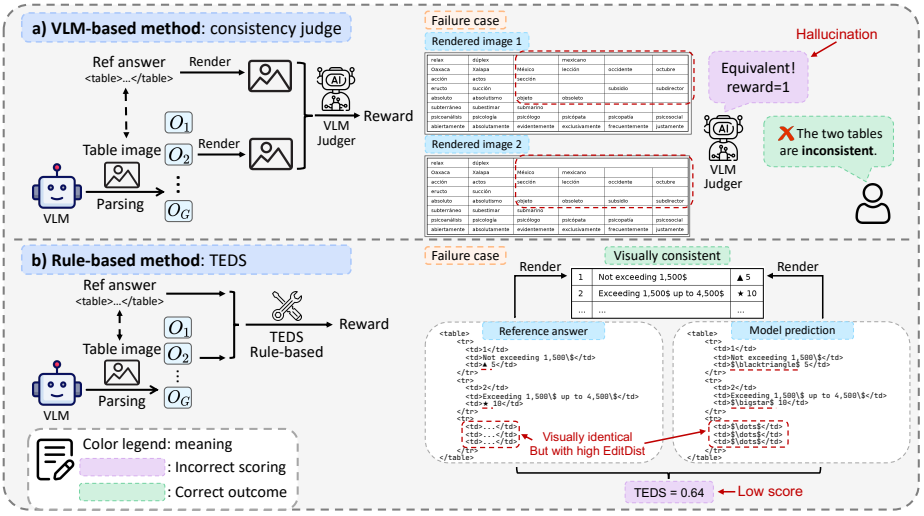


Fig. 2: Limitations of existing RL reward designs for table parsing. Top: VLM-based methods rely on VLM judges to compare rendered predictions with references, but may introduce noisy supervision due to hallucinations or misalignment. Bottom: Rule-based methods measure markup similarity and can be misleading when visually identical tables admit multiple valid expressions, especially in formula-intensive cases. These issues lead to unstable and ambiguous optimization signals.

expect a model to internalize the rigorous topological logic of complex tables solely by mimicking the final serialized output, without explicitly modeling the intermediate reasoning process?

To answer this, we revisit task modeling through a cognitive lens. From a human perspective, table parsing is rarely a single-step inference. Instead, as illustrated in Fig. 1 (b), it follows an explicit reasoning trajectory: first perceiving the global layout, then resolving local dependencies such as merged cells, and finally reconstructing the content. Conversely, current VLM-based methods *bypass these intermediate steps*, thereby forcing the model to implicitly infer complex structural logic from the final output. This black-box approach lacks structural scaffolding, making optimization inefficient and leading to suboptimal generalization.

Beyond task modeling, effectively optimizing table parsing models requires advanced learning paradigms. Reinforcement learning (RL) has emerged as a promising avenue to refine parsing quality. Unlike supervised fine-tuning (SFT), RL encourages the learning of robust, generalizable representations rather than superficial token-level pattern matching [4]. Existing reward designs generally fall into two categories: VLM-based approaches that employ VLMs as judges to compare rendered predictions against ground truths [59, 60], and rule-based approaches built on structure-aware metrics like TEDS [43, 44, 69]. However, as shown in Fig. 2, both exhibit fundamental flaws. VLM-based judges struggle

to reliably distinguish fine-grained differences between rendered images, while rule-based metrics may yield misleading rewards when visually identical content admits multiple valid representations (e.g., in formula-intensive tables). Consequently, current reward designs fail to provide stable and reliable optimization signals, leaving the potential of RL for table parsing largely untapped.

Furthermore, progress in table parsing is severely constrained by the limitations of existing benchmarks. Recent document understanding benchmarks [12, 30, 58] typically treat table parsing as an auxiliary task, resulting in small evaluation sets with restricted scenario diversity. Consequently, they may fail to adequately capture the complexity of real-world table parsing, where variations in acquisition conditions, complex layouts, and diverse writing styles pose significant challenges to model robustness and generalization [8, 46].

To address these critical limitations across task modeling, optimization, and evaluation, we propose a comprehensive framework that rethinks table parsing from a full-stack perspective. At the core of our framework is **StrucTab**, a 1B-parameter VLM. At the modeling level (Fig. 1 (b)), we decompose the parsing process into three complementary subtasks: row–column counting, merged-cell analysis, and final HTML generation. These tasks are initially introduced independently during early pretraining to build atomic skills, and are subsequently unified via a sequential reasoning strategy. This step-by-step formulation progressively equips the model with robust table-understanding capabilities. At the optimization level (Fig. 1 (c)), we introduce **Uni-TabRL**, a novel RL framework tailored for table parsing. Building upon GRPO [40], it incorporates fine-grained, decomposed rewards, thereby providing more stable and reliable optimization signals than existing VLM- or rule-based approaches, leading to more pronounced performance gains. Finally, at the evaluation level (Fig. 1 (d)), we present **TableVerse-5K**, a large-scale, challenging benchmark spanning diverse real-world scenarios, including photographed, printed, and handwritten tables. Extensive experiments demonstrate that StrucTab achieves state-of-the-art performance across all evaluated public benchmarks and significantly outperforms all the baselines on TableVerse-5K.

We summarize our main contributions as follows:

- **Modeling.** We present **StrucTab**, a structure-aware VLM that achieves state-of-the-art performance across multiple benchmarks. Through explicitly incorporating human-inspired structural reasoning, it substantially improves robustness in complex tables over standard black-box approaches.
- **Optimization.** We introduce **Uni-TabRL**, a unified RL framework for table parsing that leverages fine-grained, decomposed rewards to ensure stable and reliable optimization signals, yielding further improvements.
- **Evaluation.** We curate **TableVerse-5K**, a large-scale, challenging table parsing benchmark encompassing diverse real-world table scenarios to facilitate comprehensive evaluation of model generalization.

2 Background

2.1 Preliminaries

TEDS. Tree-Edit-Distance-based Similarity (TEDS) [68] is widely used in table parsing tasks to measure differences between prediction and ground truth, jointly evaluating structural alignment and cell content correctness [12, 30, 58]. As in Eq. (1), T_{pred} and T_{gt} denote the predicted and ground-truth HTML trees, with $|T|$ representing the node count of T. The edit distance $\text{EditDist}(T_{\text{pred}}, T_{\text{gt}})$ is the minimum number of node operations required to transform T_{pred} into T_{gt}:

$$\text{TEDS}(T_{\text{pred}}, T_{\text{gt}}) = 1 - \frac{\text{EditDist}(T_{\text{pred}}, T_{\text{gt}})}{\max(|T_{\text{pred}}|, |T_{\text{gt}}|)}. \tag{1}$$

Instead, TEDS-S (structure-only TEDS) applies the same formulation but ignores textual content, evaluating the structural accuracy of prediction alone.

GRPO. Group Relative Policy Optimization (GRPO) [40] has been widely adopted in RL as a critic-free algorithm that directly optimizes policies using group-normalized rewards from sampled responses. For each input x, multiple outputs $\{o_1, o_2, \ldots, o_G\}$ are sampled from the current policy θ_{old}, and rewards R_i are computed. The group-normalized advantage for the i-th response is:

$$A_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}. \tag{2}$$

The policy model π_θ is then optimized by maximizing:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{(x,y)\sim D,\, \{o_i\}_{i=1}^{G}\sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\frac{\pi_\theta(o_i|x)}{\pi_{\theta_{\text{old}}}(o_i|x)} A_i, \text{clip}\left(\frac{\pi_\theta(o_i|x)}{\pi_{\theta_{\text{old}}}(o_i|x)}, 1-\epsilon, 1+\epsilon \right) A_i \right) \right], \tag{3}$$

where ϵ is a hyperparameter for clipping the policy ratio.

2.2 Related Work

Table Parsing. Traditional table parsing methods primarily adopted multi-stage pipelines, decomposing the task into structure recognition and OCR [7, 9, 10, 22, 31, 48, 50, 64]. While effective on clean documents, these approaches suffer from error propagation and limited robustness in real-world scenarios [25]. To mitigate these issues, image-to-markup approaches reformulate parsing as an end-to-end sequence generation task, directly converting table images into structured formats such as HTML [13, 27, 35]. Modern VLMs have further advanced this paradigm through unified modeling and large-scale pretraining [4, 19, 36, 51]. However, most current end-to-end methods optimize for the final HTML sequence directly, overlooking explicit intermediate structural reasoning. This holistic optimization may fail to capture intricate table structures robustly.

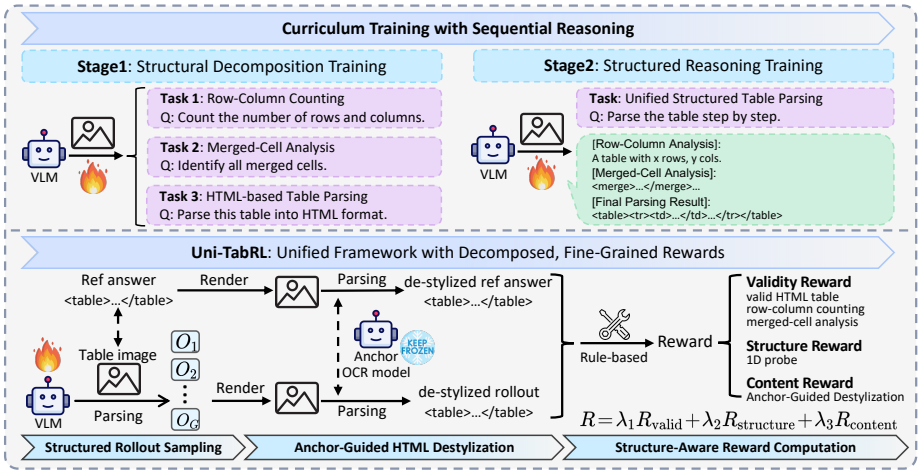


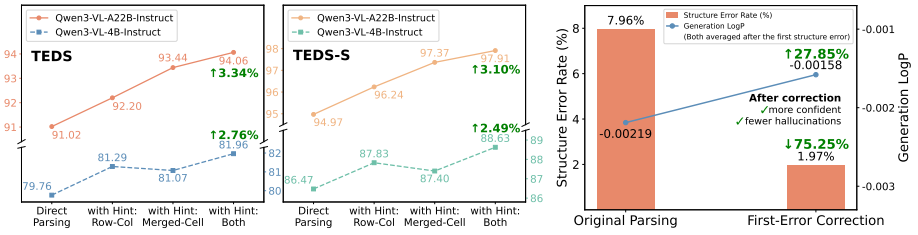
Fig. 3: Structured Optimization Framework for StrucTab. Top: Curriculum training progressively integrates atomic structural skills into unified reasoning. Bottom: Uni-TabRL enables stable optimization via fine-grained, decomposed rewards.

RL for Table Parsing. Recent studies explore RL for table parsing via task-specific rewards. FD-RL [69] adopts rule-based TEDS, which may be misleading when semantically equivalent content (e.g., formulas) admits multiple syntactic forms. MonkeyOCR-v1.5 [59] uses a VLM-based judge to assess visual consistency between rendered predictions and ground truth, but limited VLM sensitivity to fine-grained details introduces noisy feedback. TRivia [60] employs auxiliary QA tasks as reinforcement signals, focusing on local regions rather than global structural correctness. Collectively, these works highlight the potential of RL for table parsing, while revealing the challenge of designing stable, fine-grained rewards for holistic table understanding [16, 33, 56, 57].

Table Parsing Benchmark. Table parsing is typically evaluated on small subsets of document understanding benchmarks, such as OmniDocBench [30], CC-OCR [58], and OCRBench v2 [12]. As parsing serves only as an auxiliary subtask in these suites, the samples are limited in scale and structural diversity. Consequently, they may fail to fully reflect real-world complexities such as intricate merged cells, diverse acquisition conditions, and handwritten content, highlighting the need for more dedicated and comprehensive benchmarks.

3 StrucTab: Framework and Training Strategy

In this section, we present the structured reasoning and optimization framework of StrucTab. We first introduce the human-inspired structural decomposition for table parsing in Sec. 3.1, followed by a curriculum training strategy with sequential reasoning in Sec. 3.2. Finally, we describe the Uni-TabRL optimization framework in Sec. 3.3. The overall optimization pipeline is illustrated in Fig. 3.



(a) Impact of explicit structural cues **(b)** Impact of early error correction

Fig. 4: **(a)** Providing structural cues (e.g., row–column counts) explicitly improves parsing performance on OmniDocBench. **(b)** Correcting the first structural error reduces subsequent structural errors while improving downstream generation confidence.

3.1 Human-Inspired Structural Decomposition

Inspired by the explicit reasoning process of humans, we hypothesize that intermediate structural signals play an important role in table understanding but are not explicitly modeled in current end-to-end training. To verify this, we conduct a training-free preliminary study by injecting structural cues during parsing. As shown in Fig. 4a, providing row–column counts or merged-cell information as hints consistently improves table parsing accuracy. This result suggests that intermediate structural signals are indeed underutilized in existing models.

Motivated by this observation, we reformulate table parsing as structured reasoning and introduce intermediate objectives into training. Specifically, we decompose the task into two complementary subtasks: (a) row–column counting and (b) merged-cell analysis. These components inject explicit structural priors before final HTML decoding, guiding the model toward more robust and stable table understanding. Both subtasks are lightweight and directly derived from standard HTML annotations, enabling scalable supervision without additional manual labeling. Details are provided in the Appendix.

Row-Column Counting. This subtask targets global structural perception by estimating the number of rows and columns in a table. Given a table image, the model is prompted: *"Count the number of rows and columns in this table."* The target is formatted as a concise description (e.g., *"x rows and y columns"*). By predicting global dimensions before detailed parsing, the model develops a holistic understanding of table layout.

Merged-Cell Analysis. This subtask models local structural dependencies by identifying merged cells. Using the `rowspan` and `colspan` attributes in HTML annotations, we derive merged regions and represent them in a structured tokenized format: `<merge>(r1,c1),(r2,c2)</merge>`, indicating a cell spanning from (r_1, c_1) to (r_2, c_2). Merged regions are extracted row by row and concatenated in a fixed order to form the target sequence. The prompt is: *"Identify all merged cells in this table."* Explicit supervision of merged relationships equips the model with localized structural awareness.

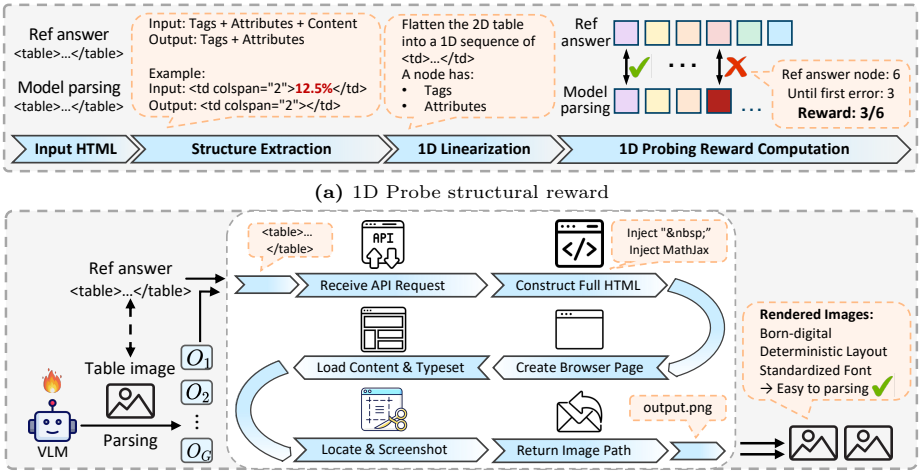


Fig. 5: Uni-TabRL Implementation Pipeline. Details are in the Appendix.

3.2 Curriculum Training with Sequential Reasoning

To bridge structural decomposition and end-to-end parsing, we design a curriculum training strategy with sequential reasoning. In the first stage of pretraining, row-column counting, merged-cell analysis, and HTML parsing are optimized as three independent question–answer tasks. This stage focuses on learning atomic structural perception and parsing skills in isolation.

In the second stage, we integrate these subtasks into a unified reasoning sequence that follows the human-inspired perception order. Specifically, we construct single-round dialogues requiring the model to sequentially generate (a) row-column counts, (b) merged-cell analysis results, and (c) the final HTML representation within one coherent response. This sequential formulation imposes an explicit reasoning path, encouraging the model to condition its final parsing decisions on intermediate structural predictions.

Through curriculum-based integration and sequential supervision, StrucTab learns step-by-step structured reasoning, leading to more robust and interpretable table parsing behavior compared to conventional direct end-to-end training.

3.3 Uni-TabRL

RL is effective in improving document understanding models by designing task-specific rewards [4, 47], but existing reward signals are often noisy or ambiguous for structural reasoning ability learning, thus remaining suboptimal for robust table parsing. To supply fine-grained and reliable supervision for RL, we decompose the overall reward into three complementary components that separately supervise *generation validity*, *structural correctness*, and *content fidelity*:

$$R = \lambda_1 R_{\text{valid}} + \lambda_2 R_{\text{structure}} + \lambda_3 R_{\text{content}} , \tag{4}$$

where λ_1, λ_2, and λ_3 are hyperparameters balancing the contributions of each component. Details are provided in the Appendix.

Validity Reward. To guarantee the feasibility of generated outputs, we introduce this component as a hard gating mechanism. Given the strict syntax requirements of HTML, any failure to produce a complete and valid table or correct prerequisite outputs (incorrect row-column counting or merged-cell analysis) results in a zero validity reward; otherwise, $R_{\text{valid}} = 1$. This binary constraint forces the model to master basic structural validity and task decomposition as a foundation for subsequent fine-grained optimization.

Structure Reward. To provide precise feedback on the structural correctness, we move beyond holistic metrics like TEDS-S as a reward, which yield a single aggregated similarity score without identifying where structural errors occur or how they propagate during decoding. Such non-decomposable signals fail to provide fine-grained, step-level guidance for autoregressive generation. Our preliminary study in Fig. 4b reveals that early correction of structural deviations is crucial for stabilizing long-form decoding and leads to more reliable table generation. Thus, we propose a **1D Probe structural reward** tailored for autoregressive models. As shown in Fig. 5a, by linearizing both the predicted and reference tables into cell sequences, we compute the reward as the ratio of correctly matched cells before the first mismatch. This mechanism explicitly penalizes early structural divergence and encourages the model to maintain a correct structural trajectory, aligning the optimization signal with the sequential generation process.

Content Reward. While standard TEDS is widely used to evaluate content accuracy, it often penalizes valid outputs that differ stylistically from the reference (e.g., visually equivalent formulas), introducing noise into reward signals. Modern OCR models are trained extensively on synthetic and rendered document images, enabling high-precision and deterministic parsing on standardized inputs. Leveraging this property, we propose **Anchor-Guided Destylization** to resolve this ambiguity. Specifically, both the reference and the model's output are rendered into standardized images (illustrated in Fig. 5b) and then re-parsed by a frozen OCR model acting as a style anchor. This process normalizes stylistic variations into a unified canonical format. By computing TEDS between these de-stylized representations, we obtain a robust, style-invariant reward that focuses purely on semantic content fidelity.

By combining validity constraints, localized structural feedback, and style-robust content supervision, our decomposed reward design yields stable and informative signals, effectively guiding the model toward robust table parsing.

4 TableVerse-5K Benchmark

For more comprehensive real-world evaluation, we construct **TableVerse-5K**, a large-scale benchmark featuring diverse table sources and high-quality annotations. Comparison with existing benchmarks is presented in Tab. 1, and the construction pipeline is shown in Fig. 6. Details are provided in the Appendix.

Table 1: Comparison of Table Parsing Benchmarks. Our TableVerse-5K is the largest, most diverse, and structurally complex benchmark for real-world table parsing.

Benchmark	Release Date	Size	Merged Cells Avg. Count	Avg. Dim.		Scenario		
				Rows	Cols	Scan	Photo (Printed)	Photo (Handwritten)
OmniDocBench [30]	2024.12	512	1.34	10.04	5.40	✓	✗	✗
CC-OCR [58]	2024.12	300	4.63	**18.99**	**7.37**	✓	✓	✓
OCRBench v2 [12]	2025.01	700	1.37	8.19	4.55	✓	✓	✗
TableVerse-5K (Ours)	2026.06	**5,000**	**5.10**	14.39	7.13	✓	✓	✓

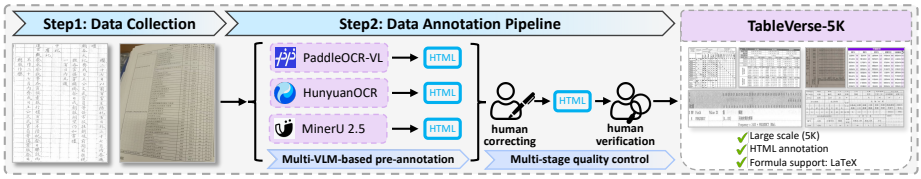


Fig. 6: TableVerse-5K Construction Pipeline.

Data Collection. All table images in TableVerse-5K are manually collected from diverse real-world sources, including academic papers, posters, reports, and technical documents [32]. The tables are manually cropped to ensure accurate localization. The dataset covers both Chinese and English, spanning printed tables, photographed tables, and handwritten content, thereby reflecting the diversity and complexity of practical scenarios.

Annotation Pipeline. To ensure both efficiency and annotation quality, we adopt a hybrid annotation strategy that combines multi-VLM-based annotation with rigorous human verification. Specifically, we first obtain candidate parsing results using three strong models: PaddleOCR-VL [6], MinerU2.5 [28], and HunyuanOCR [44]. Human annotators then select the most complete prediction as the initial draft and refine it by carefully correcting structural elements (e.g., layout and merged cells) as well as cell contents (including text and formulas). The refined annotations are subsequently reviewed by a second annotator for further verification and correction, forming a multi-stage quality control process [18,32].

Dataset Summary. Following this pipeline, we construct TableVerse-5K, a benchmark comprising 5K high-quality annotated tables. Compared with existing benchmarks, it offers greater scenario diversity, richer structural variations, and stricter annotation standards, making it better suited for evaluating the robustness and generalization of table parsing models in real-world settings. A detailed comparison and additional visualizations are provided in the Appendix.

5 Experiments

In this section, we conduct comprehensive experiments to demonstrate the superior performance of StrucTab. We detail the experimental settings in Sec. 5.1, report the results in Sec. 5.2, and provide ablation studies in Sec. 5.3. Finally, we present qualitative visualizations in Sec. 5.4.

5.1 Experimental Settings

Training Pipeline. We adopt HunyuanOCR [44] as the vision-language backbone, as it currently represents one of the strongest lightweight OCR models. Full-parameter fine-tuning is performed throughout all training stages. The training pipeline consists of three stages: pretraining, supervised fine-tuning (SFT), and RL. All experiments are conducted on 16 NVIDIA H20 GPUs. Additional implementation details are provided in the Appendix.

(a) Pretraining. We conduct pretraining in two progressive stages on a mixture of 6M synthetic table samples and publicly available datasets [27,55,63, 66,68]. In the first stage, row-column counting, merged-cell analysis, and HTML parsing are treated as three independent objectives. The model is trained for one epoch with a learning rate of 4×10^{-5}. In the second stage, the three tasks are unified into single-round dialogues with a fixed reasoning order. This stage is trained for one additional epoch.

(b) SFT. For downstream adaptation, we construct a 130K-sample high-quality dataset comprising 50K samples from public benchmarks and 80K manually annotated HTML tables. Following the second pretraining stage, we use 100K samples for SFT, training for two epochs at 2×10^{-5}.

(c) RL. Finally, we perform RL on the remaining 30K samples using the proposed Uni-TabRL framework to further optimize parsing performance.

Evaluation Benchmarks. We evaluate on the table parsing subsets of three widely used benchmarks: OmniDocBench [30], CC-OCR [58], and OCRBench v2 [12], covering digital, scanned, and photographed tables of varying complexity. We further assess performance on our proposed TableVerse-5K, a larger and more challenging benchmark with substantial real-world diversity.

Baselines. We compare our method against representative baselines across three categories: (a) **Expert table parsing (TP) models**, dedicated models designed explicitly for table structure recognition and parsing; (b) **General-purpose VLMs**, capable of table parsing via unified multimodal modeling; and (c) **Document parsing VLMs**, optimized for holistic document parsing. For each category, we select strong and representative models for comparison to ensure a comprehensive evaluation.

5.2 Experimental results

The quantitative results in Tab. 2 show that StrucTab achieves state-of-the-art performance across all three public benchmarks and TableVerse-5K, consistently outperforming the baselines by a clear margin. Notably, despite its efficient 1B-parameter backbone, StrucTab significantly surpasses massive closed-source VLMs, exceeding Gemini 2.5 Pro and GPT-5 by 7.13% and 19.14% in average TEDS, respectively. This highlights that explicit structural decomposition handles complex layouts much more effectively than the proprietary models.

Furthermore, StrucTab demonstrates clear superiority over existing RL-based table parsers, which often suffer from noisy or ambiguous reward signals. Our

Table 2: Main experimental results. StrucTab achieves the best performance.

Model Type	Model	Release Date	OmniDocBench		CC-OCR		OCRBench V2		TableVerse-5K		Average	
			TEDS	TEDS-S	TEDS	TEDS-S	TEDS	TEDS-S	TEDS	TEDS-S	TEDS	TEDS-S
Expert TP models	UniTable [35]	2024.03	82.76	89.82	57.84	70.47	67.73	78.65	48.55	78.65	53.73	79.15
	TRivia-3B [60]	2025.12	91.60	95.01	84.90	90.17	90.76	94.03	78.15	85.41	80.87	87.31
General Purpose VLMs	GPT-4o [29]	2024.05	78.27	84.56	66.98	79.04	70.51	79.55	63.62	76.41	65.67	77.51
	GPT-5 [42]	2025.08	84.91	89.91	63.25	74.09	79.91	88.69	67.04	78.96	69.65	80.64
	Qwen2.5-VL-72B-Ins. [2]	2025.02	87.85	91.80	81.22	86.48	81.33	86.58	75.23	82.65	77.15	83.97
	InternVL3.5-A28B [49]	2025.08	86.03	90.53	62.87	69.52	79.50	85.81	76.08	84.96	76.62	84.78
	Qwen3-VL-A22B-Ins. [1]	2025.10	91.02	94.97	80.98	86.19	84.12	88.15	78.26	84.23	80.02	85.59
	Seed-1.8 (no-think) [39]	2025.12	89.09	94.37	81.64	85.97	69.91	77.95	79.91	86.03	79.63	85.81
	Kimi K2.5 (no-think) [45]	2026.02	91.44	95.36	82.91	85.79	82.51	89.71	78.75	86.95	80.34	87.85
	Gemini 2.5 Pro [5]	2025.03	90.90	94.32	85.56	90.07	88.94	89.47	79.46	87.13	81.66	88.08
Document Parsing VLMs	MonkeyOCR-pro-1.2B [21]	2025.07	84.24	89.02	68.41	75.62	67.21	72.81	67.98	72.91	69.20	74.29
	MonkeyOCR-pro-3B [21]	2025.07	86.78	90.63	71.42	77.07	76.81	79.69	72.26	77.04	73.85	78.39
	DeepSeek-OCR [52]	2025.10	83.79	87.86	68.95	75.22	82.64	87.33	68.70	76.84	71.39	78.76
	POINTS-Reader [24]	2025.08	77.13	81.66	64.29	73.91	73.96	82.13	72.03	81.13	72.28	80.94
	FD-RL [69]	2025.11	87.35	92.10	72.81	78.29	76.95	82.39	74.31	80.51	75.55	81.52
	dots.ocr [19]	2025.07	88.62	92.86	75.42	81.65	82.04	86.27	73.39	81.84	75.61	83.17
	PaddleOCR-VL [6]	2025.10	91.12	94.62	79.62	85.04	79.29	83.93	77.55	84.08	78.90	84.94
	MinerU 2.5 [28]	2025.09	90.85	94.68	79.76	85.16	87.13	90.62	77.41	84.31	79.62	85.84
	HunyuanOCR [44]	2025.11	95.74	97.71	81.44	86.91	82.34	87.74	77.55	85.22	79.67	86.55
Ours	**StrucTab-SFT**	2026.06	94.62	96.51	84.23	89.91	89.33	91.69	84.59	90.02	85.87	90.70
	StrucTab	2026.06	**95.95**	**97.89**	**86.94**	**91.75**	**91.39**	**94.40**	**87.81**	**92.74**	**88.79**	**93.28**

approach outperforms FD-RL and TRivia-3B by 13.24% and 7.92% in average TEDS, respectively. Crucially, the Uni-TabRL framework alone drives a 2.92% TEDS increase over our supervised baseline (StrucTab-SFT). Overall, these results validate our design: combining intermediate structural reasoning with stable, fine-grained RL optimization significantly enhances parsing robustness.

5.3 Ablation Study

Effectiveness of Human-Inspired Structural Decomposition. We conduct ablation studies during both pretraining and SFT to evaluate the impact of our human-inspired structural decomposition, and summarize the results in Tab. 3. Starting from an end-to-end table-to-HTML parsing baseline (a), we introduce row–column counting and merged-cell analysis as auxiliary supervision signals in settings (b)–(d). Notably, in these settings, tasks are trained independently without sequential reasoning. Results show that incorporating either structural subtask consistently improves performance, and their combination yields further gains. This demonstrates that explicit supervision of intermediate structural properties provides valuable inductive biases, enabling the model to learn robust representations beyond merely fitting the final HTML sequence.

Effectiveness of Curriculum Training and Sequential Reasoning. We further investigate the efficacy of our curriculum training strategy, with results reported in Tab. 3. Comparing setting (d) (independent tasks) with setting (f) (single-round unified sequential reasoning), we observe that integrating subtasks into a coherent reasoning chain significantly boosts performance. This indicates that enforcing an explicit reasoning path allows the model to better condition final table reconstruction on intermediate structural predictions, rather than treating them as isolated auxiliary objectives. Moreover, to validate the necessity of the two-stage curriculum, we conduct a controlled experiment (e) where the first-stage pretraining of independent subtasks is removed, training directly with

Table 3: Impact of Structural Decomposition via Sequential Curriculum. Results on three open-source benchmarks (averaged) and the proposed TableVerse-5K.

Setting	Task Design			Opensource Avg.		TableVerse-5K	
	Row-column Counting	Merged-Cell Analysis	Sequential Reason	TEDS	TEDS-S	TEDS	TEDS-S
VLM backbone	-	-	-	86.69	90.95	77.55	85.22
(a) Baseline	-	-	-	86.26	90.37	81.12	85.51
(b) +RC	✓	-	-	87.43	91.20	82.09	86.79
(c) +MC	-	✓	-	87.39	91.28	82.39	87.01
(d) +RC +MC	✓	✓	-	88.93	91.94	83.24	88.61
(e) Seq (w/o Curr)	-	-	✓	89.54	92.41	84.12	89.40
(f) StrucTab-SFT	✓	✓	✓	**90.11**	**92.97**	**84.59**	**90.02**

Table 4: Effectiveness of Uni-TabRL. Both the Anchor-Guided Destylization mechanism and the 1D Probe structural reward contribute to overall performance gains, with their combination achieving the best overall results.

Setting	RL reward design			Opensource Avg.		TableVerse-5K	
	Validity Reward	Structure Reward	Content Reward	TEDS	TEDS-S	TEDS	TEDS-S
StrucTab-SFT	-	-	-	90.11	92.97	84.59	90.02
(a) RL baseline	✓	TEDS-S	TEDS	91.29	94.02	86.45	91.67
(b) +VLM judge	✓	TEDS-S	VLM-based consistency judge	90.51	93.44	85.89	91.57
(c) +Anchor	✓	TEDS-S	Anchor-Guided Destylization	91.76	94.70	87.10	92.16
(d) +1D Probe	✓	1D Probe	TEDS	91.80	94.68	87.07	92.06
(e) **StrucTab**	✓	1D Probe	Anchor-Guided Destylization	**92.05**	**95.06**	**87.81**	**92.74**

sequential reasoning. As shown in (e) vs. (f), bypassing atomic skill learning leads to inferior performance. This highlights that mastering independent structural capabilities before integrating them into a unified reasoning process is crucial for stable optimization and robust table understanding.

Effectiveness of Uni-TabRL. As shown in Tab. 4, variant (a) outperforms the SFT baseline, demonstrating the effectiveness of GRPO-based reinforcement learning for table parsing. Moreover, (c) achieves better performance than both (a) and (b), verifying the advantage of the Anchor-Guided Destylization mechanism, while the inferior results of (b) suggest that VLM-based consistency judges may introduce noisy or unreliable supervision signals that hinder optimization. In addition, (d) improves over (a), confirming the effectiveness of the proposed 1D Probe mechanism as a structure reward. Overall, combining Anchor-Guided Destylization with the 1D Probe structural reward yields the best performance, leading to the final RL model consistently outperforming all other configurations.

1D Probe Structural Reward Degenerate Repetition Behavior. The repetition issue arises when the model fails to generate the closing '</table>' token, leading to an infinite repetition loop during autoregressive decoding. This behavior severely compromises the practicality and deployability of table parsing systems and has therefore drawn significant attention in recent studies [28, 37, 43, 53]. As shown in Tab. 5, both SFT and RL substantially reduce the repetition rate, while the 1D Probe structural reward further mitigates this issue, demonstrating its effectiveness in guiding autoregressive learning.

Table 5: 1D Probe Structural Reward Reduces Repetition. Compared with TEDS-S, the proposed 1D Probe structural reward leads to fewer repetitive generations.

Setting	RL reward design	TableVerse-5K	
		# Repetition	Repetition Rate
VLM backbone	-	94	1.88%
StrucTab-SFT	-	73	1.46%
RL	TEDS + TEDS-S	60	1.20%
RL	TEDS + 1D Probe structural reward	**51**	**1.02%**

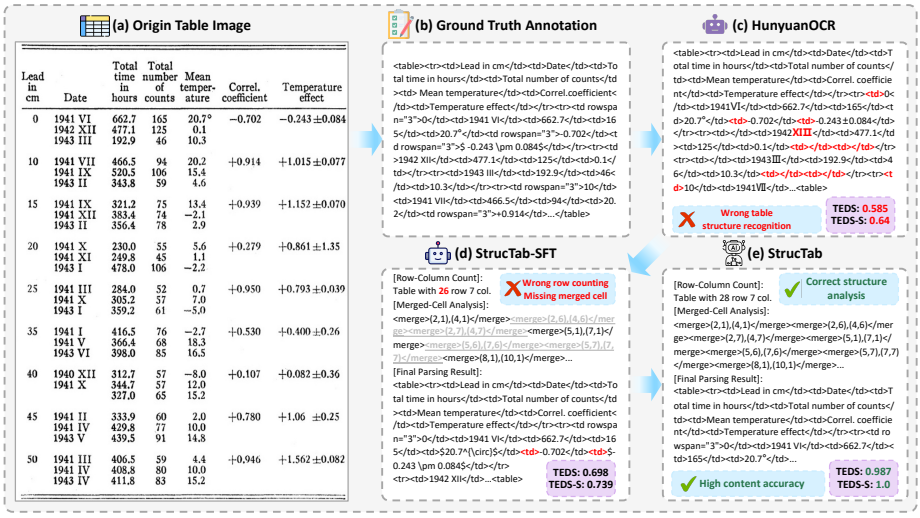


Fig. 7: Qualitative Comparisons on TableVerse-5K. Demonstrating progressive improvement: structural decomposition improves layout consistency in StrucTab-SFT, while Uni-TabRL further enhances structural alignment and content fidelity.

5.4 Qualitative Visualization on TableVerse-5K

Fig. 7 shows representative examples from TableVerse-5K comparing the baseline (c) HunyuanOCR with (d) StrucTab-SFT and (e) StrucTab, where the baseline often produces structural inconsistencies or content errors when directly generating HTML tables. After introducing structural decomposition, StrucTab-SFT improves parsing quality by explicitly reasoning over row–column counts and merged cells before table reconstruction. With further optimization using Uni-TabRL, StrucTab achieves more accurate structural alignment and content consistency, producing predictions that more closely match the ground truth. These visual results are consistent with the experimental improvements observed in our quantitative evaluations, demonstrating the effectiveness of both the structured reasoning paradigm and the proposed Uni-TabRL.

6 Concluding Remarks

Further Discussion. We further analyze why the early-error-focused 1D Probe structural reward is effective. Due to the autoregressive nature of VLMs, early tokens causally influence later predictions, so initial mistakes tend to propagate and amplify downstream errors [61, 67]. Prior work shows that targeting the first failure point is particularly effective: Step-DPO [15] improves mathematical reasoning by optimizing the first incorrect step, and SENTINEL [34] reduces hallucination by intervening at the first erroneous sentence. Consistently, our verification experiment in Fig. 4b shows that correcting the first structural error reduces subsequent structural errors by 75.25% and increases generation confidence by 27.85% (in logps). Motivated by this causal sensitivity, we introduce a 1D Probe structural reward that identifies the earliest structural deviation and provides localized RL feedback, improving structural reliability.

Summary. In this paper, we propose StrucTab, a table parsing model learned through intermediate structural supervision and reward decomposition. At the modeling level, it decomposes parsing into human-inspired subtasks and progressively integrates them via sequential reasoning. To stabilize reinforcement learning, we introduce Uni-TabRL, which employs fine-grained decomposed rewards, including a 1D Probe for early structural feedback and an Anchor-Guided Destylization strategy for robust, style-invariant content supervision. Extensive experiments show state-of-the-art performance on public benchmarks and our challenging TableVerse-5K benchmark, validating the effectiveness of explicit structural modeling and reward decomposition.

Acknowledgements

Supported by the National Natural Science Foundation of China (Grant NO 62376266 and 62406318) and CAAI-Tencent Rhino-Bird Open Research Fund.

References

1. Bai, S., Cai, Y., et al.: Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923 (2025)
3. Blecher, L., Cucurull Preixens, G., Scialom, T., Stojnic, R.: Nougat: Neural optical understanding for academic documents. In: Proceedings of the International Conference on Learning Representations (2024)
4. Chen, X., Li, S., Zhu, X., Chen, Y., Yang, F., Fang, C., Qu, L., Xu, X., Wei, H., Wu, M.: Logics-Parsing Technical Report. arXiv preprint arXiv:2509.19760 (2025)
5. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
6. Cui, C., Sun, T., Liang, S., Gao, T., Zhang, Z., Liu, J., Wang, X., Zhou, C., Liu, H., Lin, M., et al.: PaddleOCR-VL: Boosting multilingual document parsing via a 0.9B ultra-compact vision-language model. arXiv preprint arXiv:2510.14528 (2025)
7. Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., et al.: PaddleOCR 3.0 Technical Report. arXiv preprint arXiv:2507.05595 (2025)
8. Du, Y., Chen, P., Ying, X., Chen, Z.: DocPTBench: Benchmarking end-to-end photographed document parsing and translation. arXiv preprint arXiv:2511.18434 (2025)
9. Du, Y., Chen, Z., Su, Y., Jia, C., Jiang, Y.G.: Instruction-guided scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 2723–2738 (2025)
10. Du, Y., Chen, Z., Xie, H., Jia, C., Jiang, Y.G.: SVTRv2: CTC beats encoder-decoder models in scene text recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 20147–20156 (2025)
11. Du, Y., Li, C., Guo, R., Cui, C., Liu, W., Zhou, J., Lu, B., Yang, Y., Liu, Q., Hu, X., et al.: PP-OCRv2: Bag of tricks for ultra lightweight OCR system. arXiv preprint arXiv:2109.03144 (2021)
12. Fu, L., Kuang, Z., Song, J., Huang, M., Yang, B., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., et al.: OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. arXiv preprint arXiv:2501.00321 (2024)
13. Huang, Y., Lu, N., Chen, D., Li, Y., Xie, Z., Zhu, S., Gao, L., Peng, W.: Improving table structure recognition with visual-alignment sequential coordinate modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2023)
14. Khang, M., Hong, T.: TFLOP: table structure recognition framework with layout pointer mechanism. arXiv preprint arXiv:2501.11800 (2025)
15. Lai, X., Tian, Z., Chen, Y., Yang, S., Peng, X., Jia, J.: Step-DPO: Step-wise preference optimization for long-chain reasoning of LLMs. arXiv preprint arXiv:2406.18629 (2024)
16. Lei, F., Meng, J., Huang, Y., Chen, T., Zhang, Y., He, S., Zhao, J., Liu, K.: Reasoning-Table: Exploring reinforcement learning for table reasoning. arXiv preprint arXiv:2506.01710 (2025)

17. Li, G., Lyu, P., Zhang, C., Shen, H., Wu, L., Wan, X., Zeng, G., Hu, H., Ma, C., Zhou, Y.: Towards real-world document parsing via realistic scene synthesis and document-aware training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 23709–23719 (2026)
18. Li, G., Peng, S., Wan, X., Zhang, C., Feng, H., Xu, X., Wu, P., Li, B., Ding, Z., Liu, Y., Ye, Y., Yang, Y., Shu, Z., Yan, G., Li, Z., Ma, C., Wang, W., Zhou, Y., Hu, H.: Chronicles-OCR: A cross-temporal perception benchmark for the evolutionary trajectory of chinese characters. arXiv preprint arXiv:2605.11960 (2026)
19. Li, Y., Yang, G., Liu, H., Wang, B., Zhang, C.: dots.ocr: Multilingual document layout parsing in a single vision-language model. arXiv preprint arXiv:2512.02498 (2025)
20. Li, Z., Wei, J., Shen, Z., Ma, C., Wu, Y., Zhou, Y.: PACM: Position-aware cross-modality decoder for handwritten mathematical expression recognition. In: Proceedings of the International Conference on Document Analysis and Recognition. pp. 96–114 (2025)
21. Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Guo, Z., Zhang, J., Wang, X., Bai, X.: MonkeyOCR: Document parsing with a Structure-Recognition-Relation triplet paradigm. arXiv preprint arXiv:2506.05218 (2025)
22. Lin, W., Sun, Z., Ma, C., Li, M., Wang, J., Sun, L., Huo, Q.: TSRFormer: Table structure recognition with transformers. In: Proceedings of the ACM International Conference on Multimedia (2022)
23. Liu, H., Li, X., Liu, B., Jiang, D., Liu, Y., Ren, B.: Neural collaborative graph machines for table structure recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4533–4542 (2022)
24. Liu, Y., Zhao, Z., Tian, L., Wang, H., Ye, X., You, Y., Yu, Z., Wu, C., Xiao, Z., Yu, Y., et al.: POINTS-Reader: Distillation-free adaptation of vision-language models for document conversion. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1576–1601 (2025)
25. Long, R., Wang, W., Xue, N., Gao, F., Yang, Z., Wang, Y., Xia, G.S.: Parsing table structures in the wild. In: Proceedings of the IEEE International Conference on Computer Vision (2021)
26. Lyu, J., Wang, W., Yang, D., Zhong, J., Zhou, Y.: Arbitrary reading order scene text spotter with local semantics guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5919–5927 (2025)
27. Nassar, A., Livathinos, N., Lysak, M., Staar, P.: TableFormer: Table structure understanding with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2022)
28. Niu, J., Liu, Z., Gu, Z., Wang, B., Ouyang, L., Zhao, Z., Chu, T., He, T., Wu, F., Zhang, Q., et al.: MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. arXiv preprint arXiv:2509.22186 (2025)
29. OpenAI: GPT-4o System Card (2024), `https://cdn.openai.com/gpt-4o-system-card.pdf`, accessed: 2026-04-20
30. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., et al.: OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2025)
31. Paruchuri, V.: Marker (2025), `https://github.com/datalab-to/marker`, accessed: 2026-04-20
32. Peng, S., Li, G., Wan, X., Zhang, C., Feng, H., Wu, B., Shen, H., Wang, W., Cai, Z., Tian, Z., Hu, H., Ma, C., Zhou, Y.: ChartArena: Benchmarking chart parsing across languages, scenarios, and formats. arXiv preprint arXiv:2606.01348 (2026)

33. Peng, S., Wang, W., Tian, Z., Yang, S., Wu, X., Xu, H., Zhang, C., Isobe, T., Hu, B., Zhang, M.: Uni-DPO: A unified paradigm for dynamic preference optimization of LLMs. arXiv preprint arXiv:2506.10054 (2025)
34. Peng, S., Yang, S., Jiang, L., Tian, Z.: Mitigating object hallucinations via sentence-level early intervention. In: Proceedings of the IEEE International Conference on Computer Vision (2025)
35. Peng, S., Chakravarthy, A., Lee, S., Wang, X., Balasubramaniyan, R., Chau, D.H.: UniTable: Towards a unified framework for table recognition via self-supervised pretraining. arXiv preprint arXiv:2403.04822 (2024)
36. Poznanski, J., Rangapur, A., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Wilhelm, C., Lo, K., Soldaini, L.: olmOCR: Unlocking trillions of tokens in PDFs with vision language models. arXiv preprint arXiv:2502.18443 (2025)
37. Poznanski, J., Soldaini, L., Lo, K.: olmOCR 2: Unit test rewards for document OCR. arXiv preprint arXiv:2510.19817 (2025)
38. Raja, S., Mondal, A., Jawahar, C.: Table structure recognition using top-down and bottom-up cues. In: Proceedings of the European Conference on Computer Vision. pp. 70–86 (2020)
39. Seed, B.: Seed1.8 Model Card: Towards generalized real-world agency (2025), `https://github.com/ByteDance-Seed/Seed-1.8/blob/main/Seed-1.8-Modelcard.pdf`, accessed: 2026-04-20
40. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
41. Shen, H., Gao, X., Wei, J., Qiao, L., Zhou, Y., Li, Q., Cheng, Z.: Divide Rows and Conquer Cells: Towards structure recognition for large tables. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. pp. 1369–1377 (2023)
42. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: OpenAI GPT-5 System Card. arXiv preprint arXiv:2601.03267 (2025)
43. Taghadouini, S., Cavaillès, A., Aubertin, B.: LightOnOCR: A 1B end-to-end multilingual vision-language model for state-of-the-art OCR. arXiv preprint arXiv:2601.14251 (2026)
44. Team, H.V., Lyu, P., Wan, X., Li, G., Peng, S., Wang, W., Wu, L., Shen, H., Zhou, Y., Tang, C., et al.: HunyuanOCR Technical Report. arXiv preprint arXiv:2511.19575 (2025)
45. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S., Cao, Y., Charles, Y., Che, H., Chen, C., Chen, G., et al.: Kimi K2.5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276 (2026)
46. Wang, A.L., Tang, J., Liao, L., Feng, H., Liu, Q., Fei, X., Lu, J., Wang, H., Liu, H., Liu, Y., et al.: WildDoc: How far are we from achieving comprehensive and robust document understanding in the wild? In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2025)
47. Wang, B., Wu, B., Li, W., Fang, M., Huang, Z., Huang, J., Wang, H., Liang, Y., Chen, L., Chu, W., et al.: Infinity Parser: Layout aware reinforcement learning for scanned document parsing. arXiv preprint arXiv:2506.03197 (2025)
48. Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., et al.: MinerU: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839 (2024)

49. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
50. Wei, H., Kong, L., Chen, J., Zhao, L., Ge, Z., Yang, J., Sun, J., Han, C., Zhang, X.: Vary: Scaling up the vision vocabulary for large vision-language models. In: Proceedings of the European Conference on Computer Vision (2024)
51. Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al.: General OCR theory: Towards OCR-2.0 via a unified end-to-end model. arXiv preprint arXiv:2409.01704 (2024)
52. Wei, H., Sun, Y., Li, Y.: DeepSeek-OCR: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)
53. Wei, H., Sun, Y., Li, Y.: DeepSeek-OCR 2: Visual causal flow. arXiv preprint arXiv:2601.20552 (2026)
54. Xing, H., Gao, F., Long, R., Bu, J., Zheng, Q., Li, L., Yao, C., Yu, Z.: LORE: Logical location regression network for table structure recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)
55. Yang, F., Hu, L., Liu, X., Huang, S., Gu, Z.: A large-scale dataset for end-to-end table recognition in the wild. Scientific Data **10**(1), 110 (2023)
56. Yang, S., Huang, Q., Yuan, J., Zha, L., Tang, K., Yang, Y., Wang, N., Wei, Y., Li, L., Ye, W., et al.: TableGPT-R1: Advancing tabular reasoning through reinforcement learning. arXiv preprint arXiv:2512.20312 (2025)
57. Yang, Z., Chen, L., Cohan, A., Zhao, Y.: Table-R1: Inference-time scaling for table reasoning. arXiv preprint arXiv:2505.23621 (2025)
58. Yang, Z., Tang, J., Li, Z., Wang, P., Wan, J., Zhong, H., Liu, X., Yang, M., Wang, P., Bai, S., et al.: CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy. In: Proceedings of the IEEE International Conference on Computer Vision (2025)
59. Zhang, J., Liu, Y., Wu, Z., Pang, G., Ye, Z., Zhong, Y., Ma, J., Wei, T., Xu, H., Chen, W., et al.: MonkeyOCR v1.5 Technical Report: Unlocking robust document parsing for complex patterns. arXiv preprint arXiv:2511.10390 (2025)
60. Zhang, J., Wang, B., Zhang, Q., Wu, F., Wen, Z., Lu, J., Shan, J., Zhao, Z., Yang, S., Wang, Z., et al.: TRivia: Self-supervised fine-tuning of vision-language models for table recognition. arXiv preprint arXiv:2512.01248 (2025)
61. Zhang, M., Press, O., Merrill, W., Liu, A., Smith, N.A.: How language model hallucinations can snowball. arXiv preprint arXiv:2305.13534 (2023)
62. Zhang, Q., Wang, B., Huang, V.S.J., Zhang, J., Wang, Z., Liang, H., He, C., Zhang, W.: Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. arXiv preprint arXiv:2410.21169 (2024)
63. Zhang, Z., Hu, P., Ma, J., Du, J., Zhang, J., Yin, B., Yin, B., Liu, C.: SEMv2: Table separation line detection based on instance segmentation. Pattern Recognition **149**, 110279 (2024)
64. Zhang, Z., Zhang, J., Du, J., Wang, F.: Split, embed and merge: An accurate table structure recognizer. Pattern Recognition **126**, 108565 (2022)
65. Zhao, W., Feng, H., Liu, Q., Tang, J., Wei, S., Wu, B., Liao, L., Ye, Y., Liu, H., Zhou, W., et al.: TabPedia: Towards comprehensive visual table understanding with concept synergy. In: Proceedings of Advances in Neural Information Processing Systems. vol. 37, pp. 7185–7212 (2024)
66. Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: Global Table Extractor (GTE): A framework for joint table identification and cell structure recognition using visual context. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (2021)

67. Zhong, W., Feng, X., Zhao, L., Li, Q., Huang, L., Gu, Y., Ma, W., Xu, Y., Qin, B.: Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2024)
68. Zhong, X., ShafieiBavani, E., Jimeno Yepes, A.: Image-based table recognition: data, model, and evaluation. In: Proceedings of the European Conference on Computer Vision (2020)
69. Zhong, Y., Chen, L., Zeng, Z., Zhao, X., Jiang, D., Zheng, L., Huang, J., Qiu, H., Shi, P., Yang, S., et al.: Reading or Reasoning? format decoupled reinforcement learning for document OCR. arXiv preprint arXiv:2601.08834 (2025)