DIANJIN-DOCBENCH: A FULL-SPECTRUM BENCHMARK FOR FINANCIAL DOCUMENT OCR

Anonymous authors
Paper under double-blind review

ABSTRACT

Financial document parsing must preserve both the content of individual pages and the continuity of information fragmented by layout and pagination. We introduce Dianjin-DocBench, a benchmark covering 712 documents and 1,033 pages across 16 financial and finance-related categories in Chinese and English. Its manual annotation workflow separates layout analysis, content recognition, and truncated-content linking, producing page-level ground truth with unmerged regions and document-level ground truth with explicit relations and correctly consolidated text and tables. The evaluation protocol covers text, formula, table, and image parsing, alongside separate same-page and cross-page continuation tasks. We propose a new image evaluation metric that measures overlap between unions of predicted and annotated regions, accommodating differences in region granularity while penalizing both missed visual content and excess predicted area. We evaluate pipeline-based specialized VLMs, end-to-end specialized VLMs, and general VLMs under this unified protocol. By separating recognition, visual preservation, and continuation, Dianjin-DocBench provides a focused evaluation of the information retained in parsed financial documents.

1 INTRODUCTION

Financial documents are the primary record through which institutions communicate economic facts and preserve the evidence used for analysis and decision-making. Most are distributed as visually structured PDFs designed for human reading, leaving their content difficult to search, aggregate, or analyze automatically. Optical character recognition (OCR) provides the bridge from these page images to machine-readable content (Smith, 2007), while modern document parsing further recovers layout and structure in representations such as Markdown or HTML (Blecher et al., 2024; Wei et al., 2024; Poznanski et al., 2025). These representations support financial search, retrieval-augmented generation (Lewis et al., 2020), question answering, and automated analysis. Because financial information is conveyed through both content and structure, a useful OCR output must preserve not only the recognized characters but also the structural relationships needed to interpret them.

Existing benchmarks assess this capability at different levels. Many concentrate on a single task, such as layout detection (Zhong et al., 2019; Pfitzmann et al., 2022) or table recognition (Zhong et al., 2020). More recent benchmarks evaluate several elements jointly, including text, formulas, tables, reading order (Ouyang et al., 2025), and multi-page structure (Li et al., 2025b; Zhou et al., 2026). These benchmarks have established useful protocols for general-domain document parsing, but their data and evaluation objectives do not fully characterize financial documents.

Financial documents follow a markedly different content distribution. Complex numerical tables encode relationships that cannot be recovered from plain text alone, and seals frequently overlap printed content. Visual regions are also integral to financial documents: charts, images, and seals convey information that text transcription alone cannot fully preserve. A parser that transcribes the text while omitting or poorly localizing these regions produces an incomplete representation. Although some benchmarks evaluate images, common box-matching or coverage-based measures can be sensitive to whether the same visual content is annotated as one region or several, and may not jointly reflect missed content and excessive prediction area. Consequently, their scores do not always describe visual preservation faithfully.

Truncated content introduces a second difficulty. Financial documents frequently continue sentences, clauses, and tables across page boundaries, so the final Markdown must reconstruct content that is fragmented in the page images (Qian et al., 2026). Evaluating financial document parsing therefore requires assessing not only the recognition of individual fragments, but also the reconstruction of continuous text and complete tables.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

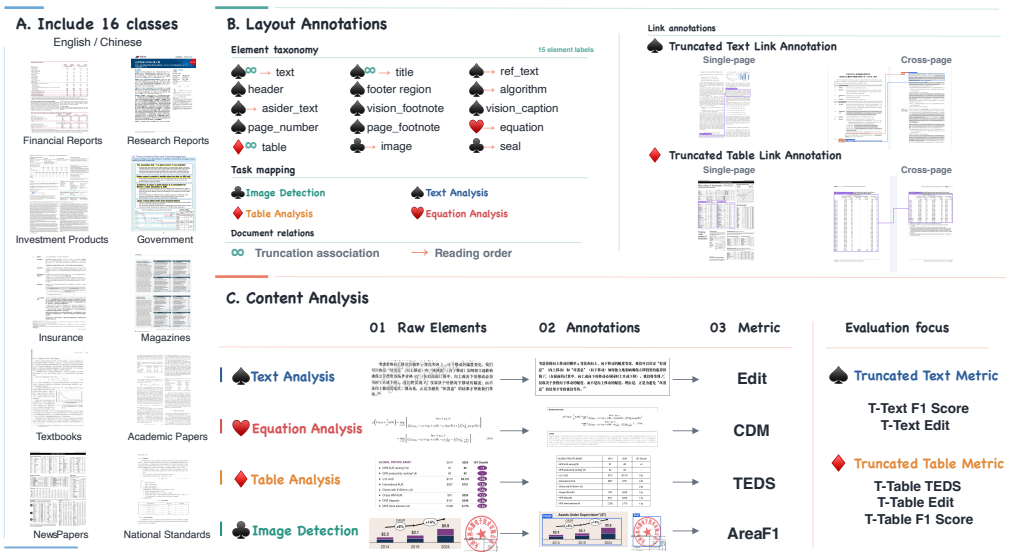

Figure 1: Overview of Dianjin-DocBench. The benchmark covers 16 financial document categories and defines 15 bounding-box classes, which are mapped to four principal page-level scoring tasks: image, text, table, and equation. Reading order and explicit same-page and cross-page links further support document-level evaluation of truncated text and tables.

We introduce **Dianjin-DocBench**, shown in Figure 1, a full-spectrum benchmark designed for financial document OCR. It contains 712 documents and 1,033 pages from 16 financial and finance-related categories in Chinese and English. Dianjin-DocBench provides detailed page annotations together with page- and document-level Markdown, reading order, visual regions, and precise relations for truncated text and tables. Its protocol evaluates local recognition and final-document reconstruction in a unified setting, with granularity-tolerant visual evaluation and document-level measurement of truncated-content reconstruction.

The benchmark makes three contributions:

- **A broad financial document benchmark.** Dianjin-DocBench provides substantial bilingual coverage across 16 financial and finance-related document types, with detailed layout and content annotations as well as precise truncated-text and truncated-table relations. It supports evaluation at both page and document levels.
- **Granularity-tolerant evaluation of financial visual evidence.** We introduce an image evaluation method that measures recovered visual area while accommodating many-to-many region decompositions. The benchmark also reports document-level reconstruction of truncated text and tables from final Markdown.
- **A comprehensive evaluation of current systems.** We benchmark 22 representative pipeline-based specialized VLMs, end-to-end specialized VLMs, and general VLMs under the same data and scoring protocol.

2 Related Work

Vision-language models. Vision-language models (VLMs) combine visual encoding with language generation, enabling models to respond to image–text inputs through multimodal instruction tuning (Liu et al., 2023). With advances in visual recognition and structured generation, general-purpose VLMs can extract content from document images and organize it according to user instructions (Kimi Team, 2026; Qwen Team, 2026b). This provides a prompt-based interface for document parsing without additional task-specific fine-tuning, allowing the same model to perform different content extraction and organization tasks.

Document OCR models. Document OCR has evolved from text recognition toward the direct generation of structured document representations (Kim et al., 2022; Blecher et al., 2024). Current specialized parsers primarily follow pipeline-based and end-to-end designs. Pipeline-based methods perform layout localization, region-level content recognition, and document assembly in stages, allowing different regions to be processed with appropriate recognition and post-processing procedures (Niu et al., 2026; Cui et al., 2025). End-to-end methods instead learn document perception and structured generation within a unified model, directly converting page images into

parsed outputs (Lu et al., 2026; Dong et al., 2026). These differences in region localization, content generation, and output assembly motivate evaluation beyond recognition accuracy to assess how faithfully the final outputs preserve document information.

Table 1: Comparison of representative document parsing benchmarks by document-type coverage, supported evaluation targets, and evaluation paradigm.

Benchmark	Types	Evaluation coverage								Evaluation paradigm	
		Layout	Text	Table	Formula	Reading order	Figure	T-Text	T-Table	Single-task	End-to-end
Task-focused benchmarks											
PubLayNet (Zhong et al., 2019)	1	✓	–	–	–	–	–	–	–	✓	–
DocLayNet (Pfitzmann et al., 2022)	6	✓	–	–	–	–	–	–	–	✓	–
Robust Reading (Karatzas et al., 2015)	1	–	✓	–	–	–	–	–	–	✓	–
PubTabNet (Zhong et al., 2020)	1	–	–	✓	–	–	–	–	–	✓	–
SciTSR (Chi et al., 2019)	1	–	–	✓	–	–	–	–	–	✓	–
Im2LaTeX-100K (Deng et al., 2017)	1	–	–	–	✓	–	–	–	–	✓	–
UniMER-Test (Wang et al., 2024a)	1	–	–	–	✓	–	–	–	–	✓	–
ReadingBank (Wang et al., 2021)	1	–	–	–	–	✓	–	–	–	✓	–
HRDoc (Ma et al., 2023)	1	–	–	–	–	–	–	–	–	✓	–
DocHieNet (Xing et al., 2024)	5	–	–	–	–	–	–	–	–	✓	–
Page-level end-to-end benchmarks											
Fox (Liu et al., 2024)	2	–	✓	–	–	–	–	–	–	–	✓
Nougat (Blecher et al., 2024)	1	–	✓	✓	✓	–	–	–	–	–	✓
GOT OCR 2.0 (Wei et al., 2024)	2	–	✓	✓	✓	–	–	–	–	–	✓
olmOCR-Bench (Poznanski et al., 2025)	7	–	✓	✓	✓	✓	–	–	–	–	✓
OmniDocBench (Ouyang et al., 2025)	9	✓	✓	✓	✓	✓	–	–	–	✓	✓
Document-level end-to-end benchmarks											
READoc (Li et al., 2025b)	3	–	✓	✓	✓	✓	–	–	–	–	✓
MPDocBench-Parse (Zhou et al., 2026)	15	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
FinDocBench (Qian et al., 2026)	6	✓	✓	✓	✓	✓	–	–	✓	–	✓
Dianjin-DocBench	**16**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Document parsing benchmarks. Early document parsing benchmarks primarily assess individual capabilities, such as layout analysis, table recognition, and formula recognition (Zhong et al., 2019; 2020; Wang et al., 2024a). While useful for evaluating specific components, these isolated tasks do not fully characterize the quality of complete document outputs. Subsequent benchmarks extend evaluation to full-page and document-level structured extraction, covering multiple content elements, reading order, and document structure (Ouyang et al., 2025; Li et al., 2025b). Photographed-document and multilingual benchmarks broaden the input conditions considered in evaluation (Du et al., 2025; Li et al., 2026b), while multi-page benchmarks explicitly assess content continuity and reconstruction across layout regions and page boundaries (Zhou et al., 2026). Financial-document benchmarks further bring domain-specific content and document-level parsing into focus (Qian et al., 2026). Building on this progression, Dianjin-DocBench focuses on the combination of complex numerical tables, charts and seals, and within-page and cross-page content fragmentation in financial documents. It jointly evaluates page-level content recognition, visual preservation under different region granularities, and document-level reconstruction of truncated content.

3 DIANJIN-DOCBENCH DATASET

The benchmark is organized around two complementary requirements: preserving the content visible on each financial document page and recovering the text and tables that form continuous units across layout or page boundaries. Both are annotated from the same source documents, so page-level recognition and document-level reconstruction can be evaluated against consistent ground truth.

3.1 DOCUMENT COLLECTION AND COVERAGE

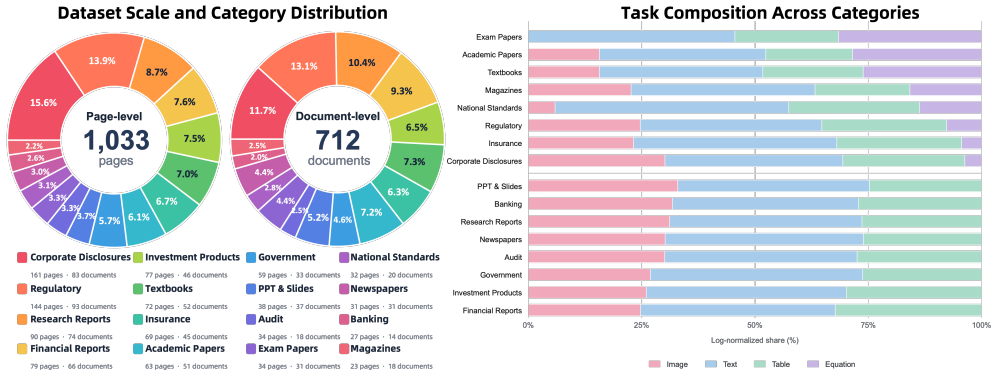

Figure 2: Dataset scale and category distribution (left), together with the relative composition of image, text, table, and equation regions across categories (right).

Dianjin-DocBench contains 712 documents and 1,033 pages spanning 16 financial and finance-related categories in Chinese and English. The collection includes corporate disclosures, regulatory documents, research reports, financial reports, investment products, textbooks, insurance documents, academic papers, government documents, PPT and slides, audit documents, exam papers, national standards, newspapers, banking documents, and magazines. As shown in Figure 2, the proportions of text, tables, formulas, and images vary across document categories, resulting in different parsing demands. Financial content motivates evaluation beyond transcription. A correctly transcribed paragraph may still lose the chart or seal that accompanies it; a table may be locally accurate but incomplete when its continuation is left on the next page. We therefore provide page-level annotations with fragments kept separate and document-level annotations with fragments correctly joined, to evaluate content recognition and truncated-content linking. We also report evaluation results separately by document category and language.

3.2 THREE-STAGE MANUAL ANNOTATION

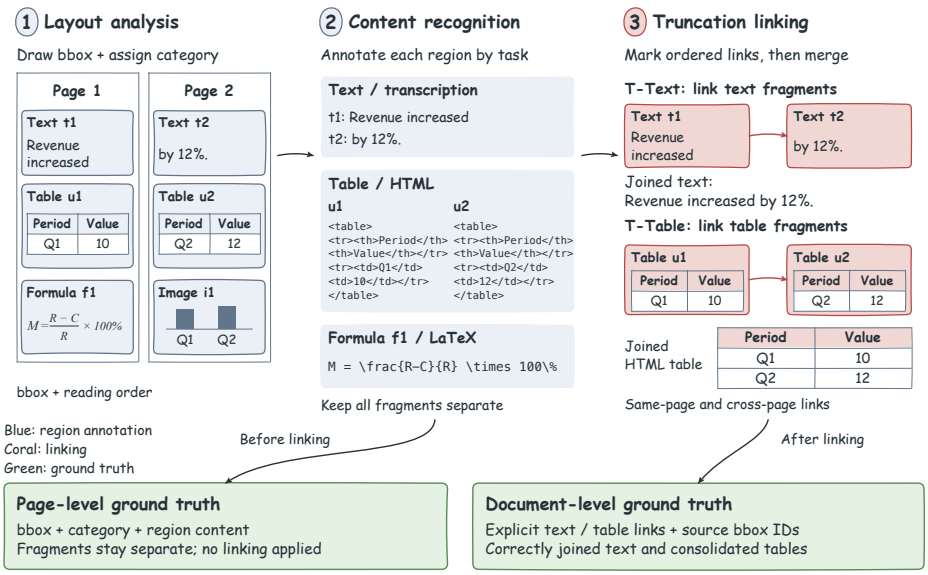

Figure 3: Three-stage manual annotation workflow. Region boxes and categories are first annotated, followed by task-specific text, table, and formula content. Continuation relations are then marked to construct correctly joined text and tables. Stage 2 supplies page-level annotations with fragments kept separate; Stage 3 supplies document-level annotations with explicit links and consolidated content.

The annotation process consists of layout analysis, content recognition, and truncated-content linking, with human annotation at each stage. The process produces both an unmerged page-level representation and a linked document-level representation.

Stage 1: Layout analysis. Bounding boxes are manually drawn on document pages and assigned semantic categories. The schema contains 15 region classes: text, title, table, image, header, footer, algorithm, side text, seal, visual caption, visual footnote, page number, reference text, equation, and page footnote. A region retains its box coordinates, allowing later content and relation annotations to refer to the same visual evidence. Reading-order annotations describe the sequence in which textual regions should be read.

Stage 2: Content recognition. The content within each box is manually annotated according to its task. Textual regions receive text transcriptions, tables receive structured HTML annotations, and formulas receive LaTeX annotations. The region categories connect these task-specific targets to the layout layer. Image and seal regions retain their spatial annotations for visual-region evaluation. Image annotations also cover images embedded within table cells, enabling evaluation of parsers' ability to identify and preserve such visual content in tables. At this stage, fragments remain separate even when they belong to one continuous paragraph or table. The resulting page-level ground truth records what appears on each page without applying truncation links or merging the fragments.

Stage 3: Truncated-content linking. Text and table boxes requiring continuation are manually inspected, and the relevant boxes and their connection order are annotated. Relations may connect separated regions within a page or extend across pages. Correctly joined text or consolidated

table content is then constructed to form the linked document representation. A continuation label therefore identifies both the source fragments and the complete content they should form; it is not inferred solely from spatial adjacency. The page-level representation remains available alongside this document-level ground truth.

3.3 Two annotation levels for two evaluation targets

The page-level target preserves region boundaries and their original content. It supports text, table, and formula recognition, reading order, and visual localization. The document-level target additionally encodes the relations and consolidated content needed for truncated-text and truncated-table evaluation. A pair of fragments can consequently be recognized correctly at page level while still being joined incorrectly at document level.

For text, a continuation group is an ordered sequence of fragments; adjacent members define the required joining edges. For tables, the group identifies the component table regions and their consolidated HTML target. Same-page and cross-page relations are retained as separate evaluation scopes. This separation tests continuity without allowing strong transcription scores to stand in for successful joining.

4 Evaluation Protocol

4.1 Visual-region metric design

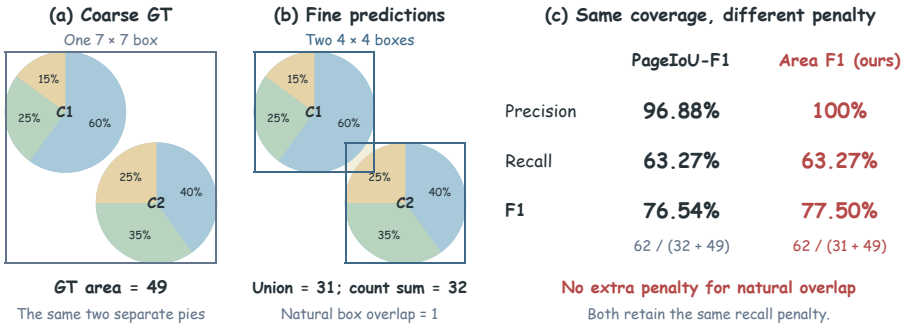

Figure 4: Natural overlap under different image-region granularities. Tight predicted boxes preserve both pie charts, but their overlap introduces an additional precision penalty under coverage counting. Area F1 avoids this penalty. The constructed comparison evaluates PageIoU-style count F1 and Area F1 over full bounding-box areas.

Image evaluation measures whether a parser preserves the spatial extent of visual evidence in financial documents, including charts, other images, and seals. A key challenge is the difference in region granularity between predictions and annotations: the same multi-panel figure may be represented by one enclosing box or several boxes around individual subfigures. Evaluation should accommodate reasonable differences in this grouping while accounting for omitted regions and excess predicted area. MinerU2.5 introduces PageIoU to address granularity differences through per-pixel coverage counts (Niu et al., 2026). However, counting overlapping coverage can introduce an additional penalty in the corresponding count-based F1, even though the overlap does not change the union of the predicted regions. Figure 4 illustrates this issue: the tight boxes of two separate pie charts naturally overlap, but this overlap does not indicate incorrect extraction. To avoid penalizing such valid decompositions solely for repeated coverage, we propose **Area F1**, which compares the unions of predicted and annotated visual regions rather than their coverage multiplicities.

For page p, let G_p be the union of ground-truth image and seal boxes, and P_p the union of predicted visual boxes. All boxes are expressed in a common page coordinate system, and areas are computed over the full rectangles. With $I_p = |G_p \cap P_p|$, area precision, recall, and F1 are

$$\text{Precision}_p = \frac{I_p}{|P_p|}, \qquad \text{Recall}_p = \frac{I_p}{|G_p|}, \qquad \text{AreaF1}_p = \frac{2I_p}{|G_p| + |P_p|}. \tag{1}$$

Unioning boxes counts each covered location only once, making Area F1 invariant to changes in box decomposition or overlapping coverage that leave the region unions unchanged. Precision penalizes predicted area outside the annotated union, while recall penalizes annotated area left uncovered. This removes the additional penalty for overlap without ignoring spatial discrepancies between coarse and fine boxes. When only one of G_p and P_p is empty, the score is zero; pages with both sets empty are

excluded. Image F1 is the mean Area F1 over eligible pages. The metric evaluates visual localization and preservation, not interpretation of the information within a chart.

4.2 PAGE-LEVEL RECOGNITION AND READING ORDER

Predicted Markdown is separated into text, tables, and formulas and aligned with the corresponding page annotations. Text uses normalized edit distance (Edit), tables use tree-edit-distance-based similarity (TEDS) (Zhong et al., 2020) and its structure-only variant (TEDS-S), and formulas use character detection matching (CDM) (Wang et al., 2024b). Images are evaluated using Area F1, defined in Section 4.1, to measure visual-region localization and preservation. Reading order is evaluated using edit distance on the aligned textual sequence. Lower Edit indicates fewer recognition or ordering errors; higher TEDS, TEDS-S, CDM, and Area F1 indicate better performance on their respective tasks.

Headers, footers, page numbers, and page footnotes are excluded from the main text and reading-order scores, following the treatment of inconsistently emitted peripheral content in page-level parsing evaluation. They remain annotated in the dataset.

The page-level Overall score equally weights text, formula, table, and image performance. With text Edit on a 0–1 scale and formula CDM, table TEDS, and Image F1 on a 0–100 scale,

$$\text{Overall} = \frac{100(1 - E_{\text{text}}) + C_{\text{formula}} + T_{\text{table}} + F_{\text{image}}}{4}. \tag{2}$$

Here E_{text}, C_{formula}, T_{table}, and F_{image} denote text Edit, formula CDM, table TEDS, and Image F1, respectively. Systems without usable visual localization receive zero image credit. Reading order and truncated-content reconstruction are reported separately and excluded from Overall.

4.3 TRUNCATED-CONTENT RECONSTRUCTION

We evaluate truncated text (T-Text) and tables (T-Table) in the final document Markdown against the linked document-level annotations, reporting same-page and cross-page results separately. The evaluator aligns and scores the output without adding missing links or merging fragments.

For T-Text, annotated fragments are located through recognition-tolerant matching. A link is recovered only when adjacent fragments appear in order within one reconstructed text block. We report edge precision, recall, and F1; group recall requires all links in a group to be recovered. Edit measures reconstructed text fidelity separately.

For T-Table, predicted tables are aligned with the consolidated HTML targets. TEDS, TEDS-S, and Edit assess content and structure and are averaged over annotated continuation groups. F1 evaluates

Table 2: Unified page-level results on Dianjin-DocBench, grouped by inference design and ordered by Overall within each group. Overall equally weights text, formula, table, and figure scores; N/A layout contributes zero. Overall, CDM, Area F1, TEDS, and TEDS-S are reported on a 0–100 scale; Edit remains on a 0–1 scale. Lower is better for Edit metrics; higher is better otherwise.

Model	Overall ↑	Text Edit↓	Formula CDM↑	Figure Area F1↑	Table TEDS↑	Table TEDS-S↑	Reading order Edit↓
Pipeline-based Specialized VLMs							
TeleOCR (Cai et al., 2026)	**90.29**	0.0430	93.32	**83.56**	88.57	90.14	0.3116
MinerU2.5-Pro (Wang et al., 2026)	89.68	0.0430	96.03	75.69	**91.30**	**92.65**	0.3102
PaddleOCR-VL-1.6 (Zhang et al., 2026)	88.68	0.0630	94.85	79.51	86.65	88.63	0.3311
PaddleOCR-VL-1.5 (Cui et al., 2026)	88.29	0.0623	94.21	79.51	85.69	87.69	0.3309
GLM-OCR (Duan et al., 2026)	87.27	0.0903	92.81	78.91	86.38	88.32	0.3571
PaddleOCR-VL (Cui et al., 2025)	87.22	0.0595	95.08	74.44	85.33	87.20	0.3326
MinerU2.5 (Niu et al., 2026)	87.07	0.0425	96.12	69.24	87.15	88.73	0.3129
Youtu-Parsing (Cao et al., 2026)	84.47	0.0282	89.48	61.86	89.35	91.34	0.3091
MonkeyOCR-Pro-3B (Li et al., 2025a)	79.73	0.0762	86.35	60.98	79.20	82.18	0.3590
Dolphin-v2 (Feng et al., 2026)	78.23	0.0674	86.99	46.81	85.87	87.69	0.3324
End-to-End Specialized VLMs							
OvisOCR2 (Lu et al., 2026)	86.94	**0.0218**	94.35	65.84	89.76	90.99	**0.2919**
dots.mocr (Zheng et al., 2026)	85.42	0.0675	89.34	70.50	88.59	90.42	0.3368
Logics-Parsing-v2 (Chen et al., 2025)	83.04	0.0382	95.55	51.88	88.56	90.49	0.3107
Qianfan-OCR (Dong et al., 2026)	81.65	0.0837	93.02	57.39	84.58	86.95	0.3320
FireRed-OCR (Wu et al., 2026)	69.13	0.0469	**96.27**	N/A	84.96	86.85	0.3145
DeepSeek-OCR-2 (Wei et al., 2026)	67.68	0.0497	93.70	N/A	81.99	84.72	0.3256
HunyuanOCR-1.5 (Li et al., 2026a)	66.51	0.0584	91.54	N/A	80.33	82.70	0.3331
General VLMs							
Kimi-K3 (Kimi Team, 2026)	85.99	0.0560	95.05	66.73	87.77	89.81	0.3243
Kimi-K2.6	84.79	0.0547	93.49	63.26	87.87	89.81	0.3330
Qwen3.5-Plus (Qwen Team, 2026a)	80.34	0.0821	90.91	59.02	79.62	82.37	0.3495
Ovis2.6-30B-A3B (Lu et al., 2025)	74.65	0.0413	95.24	17.69	89.78	91.46	0.3089
Qwen3.8-Max (Qwen Team, 2026b)	73.35	0.0979	94.46	23.51	85.22	86.89	0.3223

recovery of links between table fragments and is aggregated at document level, distinguishing content similarity from successful joining.

5 Experiments and Analysis

We organize the evaluated systems into pipeline-based specialized VLMs, end-to-end specialized VLMs, and general VLMs. Pipeline systems expose native layout, recognition, and assembly outputs; end-to-end parsers produce structured content directly; general VLMs are evaluated through page-level calls and ordered multi-image document calls. For truncated-content experiments, we compare only models that support the corresponding tasks.

5.1 Page-level completeness and visual evidence

As shown in Table 2, pipeline-based systems lead overall page-level performance, while end-to-end specialized models achieve the best results on text, formula, and reading-order tasks. Strong recognition performance is therefore not restricted to a single inference design. General VLMs are also competitive on formula and table recognition, but their image-preservation results vary much more widely. The main distinction between these approaches lies not in whether they can recognize document content, but in how consistently they handle its different forms.

Pipeline systems preserve visual evidence well. The leading pipeline systems retain image regions more effectively than both end-to-end specialized models and general VLMs. Their explicit layout-detection and region-extraction stages provide a dedicated path for preserving visual content, which may help explain this advantage. Nevertheless, substantial variation within the pipeline group shows that adopting a staged design alone is insufficient. Reliable image preservation remains an important area for improvement across parsing approaches.

5.2 Same-page continuity

Table 3: Same-page T-Text results. P, R, and F1 measure the recovery of links between text fragments; Group R measures the recall of completely connected groups. Edit evaluates the reconstructed text.

Model	Edit↓	P↑	R↑	F1↑	Group R↑
Pipeline-based Specialized VLMs					
MinerU2.5-Pro	0.2209	80.00	32.62	46.34	34.25
PaddleOCR-VL-1.6	0.1448	90.71	54.51	68.10	55.71
PaddleOCR-VL-1.5	0.1356	90.85	55.36	68.80	56.62
End-to-End Specialized VLMs					
OvisOCR2	0.1205	**96.67**	62.23	75.72	63.47
FireRed-OCR	0.1207	95.21	59.66	73.35	60.73
DeepSeek-OCR-2	0.2419	85.54	60.94	71.18	62.56
HunyuanOCR-1.5	0.1727	88.77	71.24	79.05	73.52
General VLMs					
Kimi-K3	**0.0910**	90.21	75.11	**81.97**	76.71
Kimi-K2.6	0.1331	89.67	70.82	79.14	73.52
Ovis2.6-30B-A3B	0.1912	93.63	63.09	75.38	63.93
Qwen3.8-Max	0.2039	82.52	36.48	50.60	36.53
Qwen3.5-Plus	0.1288	80.45	**75.97**	78.15	**77.63**

Truncated text. As shown in Table 3, end-to-end specialized models are effective at recovering same-page text continuity, outperforming the compared pipeline systems on connection recovery. They generate continuous text directly without an additional joining tool, suggesting that page context provides useful cues for joining separated fragments. General VLMs can also perform strongly in this setting. However, recall remains lower than precision, and complete groups are still missed. The remaining difficulty is therefore not simply recognizing individual fragments, but recovering all of the connections needed to reconstruct a passage.

Table 4: Same-page T-Table results. F1 measures the recovery of links between table fragments; TEDS, TEDS-S, and Edit evaluate the reconstructed tables.

Model	TEDS↑	TEDS-S↑	Edit↓	F1↑
General VLMs				
Kimi-K3	**67.32**	**72.83**	**0.2984**	**12.92**
Kimi-K2.6	42.47	60.28	0.4438	2.94
Ovis2.6-30B-A3B	48.33	53.48	0.5021	0.00
Qwen3.8-Max	55.59	65.51	0.3885	2.94
Qwen3.5-Plus	45.78	64.55	0.4468	3.57

Truncated tables. As shown in Table 4, same-page table joining remains largely unresolved. Since current specialized OCR models do not support this task, we evaluate only general VLMs. Even the strongest model, Kimi-K3, achieves an F1 of only 12.92. Partial recovery of table content and structure is much more successful than recovery of the connections between fragments. Unlike continuous prose, these fragments must be integrated into a shared row-and-column structure. The results suggest that this structural reconstruction remains a major obstacle beyond recognizing the individual fragments.

5.3 CROSS-PAGE CONTINUITY

Table 5: Cross-page T-Text results. P, R, and F1 measure the recovery of links between text fragments; Group R measures the recall of completely connected groups. Edit evaluates the reconstructed text.

Model	Edit↓	P↑	R↑	F1↑	Group R↑
General VLMs					
Kimi-K2.6	**0.1233**	92.86	61.90	74.29	61.90
Kimi K3	0.1563	96.36	63.10	76.26	63.10
Qwen3.5-Plus	0.1648	95.59	**77.38**	**85.53**	**77.38**
Qwen3.8-Max	0.3046	85.71	14.29	24.49	14.29
Ovis2.6-30B-A3B	0.2287	**96.77**	35.71	52.17	35.71

Truncated text. As shown in Table 5, general VLMs can recover cross-page text connections from ordered page images, but performance is uneven. Precision remains relatively high across models, whereas recall varies substantially. This pattern indicates that the models often make valid connections when they attempt to join fragments, yet leave many true continuations unresolved. Missed connections are thus a persistent limitation: providing adjacent pages makes cross-page recovery possible, but does not ensure that the available context is used completely.

Table 6: Cross-page T-Table results. F1 measures the recovery of links between table fragments; TEDS, TEDS-S, and Edit evaluate the reconstructed tables.

Model	TEDS↑	TEDS-S↑	Edit↓	F1↑
Pipeline-based Specialized VLMs				
MinerU2.5-Pro	86.69	88.73	0.1306	74.24
PaddleOCR-VL-1.6	77.32	80.59	0.2073	69.00
PaddleOCR-VL-1.5	77.50	80.87	0.2042	68.01
PaddleOCR-VL	75.29	78.35	0.2144	60.07
MinerU2.5	82.41	84.93	0.1421	75.90
General VLMs				
Kimi-K2.6	81.02	83.02	0.1833	73.93
Kimi K3	**87.64**	**89.61**	**0.1147**	**83.64**
Qwen3.5-Plus	69.47	75.12	0.2636	51.04
Qwen3.8-Max	71.99	74.94	0.2615	43.25
Ovis2.6-30B-A3B	64.48	66.71	0.3398	17.75

Truncated tables. As shown in Table 6, cross-page table joining is better supported than same-page joining. Pipeline systems with native table-consolidation stages recover connections effectively, showing that explicit assembly remains a useful approach to document-level reconstruction. General VLMs can achieve similarly strong or better results, but their performance varies substantially. Notably, the same general models perform much worse on same-page table connections. Thus, cross-page success should not be interpreted as a general ability to reconnect tables fragmented by different layouts.

5.4 LANGUAGE AND DOCUMENT-CATEGORY VARIATION

As shown in Table 7, similar overall performance across languages can conceal task-specific differences. The leading pipeline systems perform strongly on both English and Chinese subsets, but individual models may remain consistent on text while differing considerably on images or tables. Moreover, the direction of these differences is not uniform across tasks. Cross-language robustness is therefore content-dependent, and an aggregate language score alone can obscure weaknesses in particular types of document content.

Table 7: Page-level results on the English and Chinese subsets of Dianjin-DocBench.

Model	Text Edit↓		Formula CDM↑		Image F1↑		Table TEDS↑		Overall↑	
	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
Pipeline-based Specialized VLMs										
TeleOCR	0.0369	0.0468	95.28	91.46	**85.24**	**82.14**	83.19	91.94	90.00	**90.22**
MinerU2.5-Pro	0.0250	0.0540	96.09	95.97	78.34	73.51	**88.28**	93.19	**90.05**	89.32
PaddleOCR-VL-1.6	0.0639	0.0625	94.47	95.20	79.25	79.73	80.17	90.70	86.87	89.85
PaddleOCR-VL-1.5	0.0634	0.0617	95.66	92.82	79.25	79.73	78.34	90.28	86.73	89.17
GLM-OCR	0.0992	0.0849	95.05	90.69	78.35	79.38	79.36	90.77	85.71	88.09
PaddleOCR-VL	0.0475	0.0668	97.04	93.21	76.80	72.53	78.56	89.57	86.91	87.16
MinerU2.5	0.0308	0.0496	97.09	95.20	73.31	65.86	81.18	90.88	87.12	86.75
Youtu-Parsing	0.0385	0.0220	92.10	86.97	69.89	55.40	84.42	92.43	85.64	83.15
MonkeyOCR-Pro-3B	0.0849	0.0709	86.21	86.48	66.11	56.67	71.82	83.82	78.91	79.97
Dolphin-v2	0.0692	0.0663	87.91	86.10	52.30	42.62	78.48	90.48	77.94	78.15
End-to-End Specialized VLMs										
OvisOCR2	**0.0244**	**0.0203**	93.00	95.63	74.92	58.35	83.91	**93.43**	87.35	86.34
dots.mocr	0.0722	0.0647	93.60	85.30	72.65	68.84	83.37	91.86	85.60	84.88
Logics-Parsing-v2	0.0410	0.0365	94.71	**96.35**	54.63	49.49	84.60	91.04	82.46	83.31
Qianfan-OCR	0.0938	0.0776	91.59	94.38	68.41	48.26	76.77	89.46	81.85	81.09
FireRed-OCR	0.0433	0.0492	98.35	94.30	0.00	0.00	77.91	89.36	67.98	69.69
DeepSeek-OCR-2	0.0605	0.0432	98.44	89.20	0.00	0.00	72.78	87.75	66.29	68.16
HunyuanOCR-1.5	0.0703	0.0512	93.44	89.73	0.00	0.00	72.56	85.19	64.74	67.45
General VLMs										
Kimi-K2.6	0.0774	0.0408	95.21	91.86	61.52	64.71	83.85	90.39	83.21	85.72
Kimi K3	0.0649	0.0506	98.00	92.24	65.31	67.90	83.21	90.61	85.01	86.43
Qwen3.5-Plus	0.0678	0.0908	94.40	87.59	62.22	56.34	79.67	79.59	82.38	78.61
Qwen3.8-Max	0.0667	0.1169	94.79	94.14	22.44	24.38	83.61	86.22	73.54	73.26
Ovis2.6-30B-A3B	0.0332	0.0462	**98.75**	91.91	25.81	11.68	86.52	91.83	76.94	72.70

Table 8: Performance across the 16 document categories in Dianjin-DocBench.

Model	CD	REG	RR	FR	IP	TB	INS	AP	GOV	PPT	AUD	EXM	STD	NEWS	BANK	MAG
Pipeline-based Specialized VLMs																
TeleOCR	71.24	**92.55**	90.32	**95.27**	**92.49**	**93.57**	61.09	95.90	87.44	80.10	**91.09**	70.46	86.29	75.85	82.43	92.90
MinerU2.5-Pro	70.60	91.52	86.25	93.19	90.43	91.04	54.75	**96.81**	90.71	87.64	82.01	70.87	95.30	**77.35**	**85.03**	96.79
PaddleOCR-VL-1.6	77.63	89.20	91.15	94.00	87.09	85.33	63.56	95.76	88.01	71.71	89.24	69.89	92.80	**77.35**	82.75	89.12
PaddleOCR-VL-1.5	77.70	88.93	**91.41**	92.92	86.38	84.82	55.71	96.64	87.33	71.08	89.67	70.35	92.71	76.96	81.32	84.79
GLM-OCR	76.87	86.50	89.87	92.58	84.74	87.56	63.46	96.56	86.42	69.21	84.65	68.00	90.80	72.61	78.16	79.66
PaddleOCR-VL	87.46	85.09	89.75	88.70	87.83	84.58	53.08	95.79	87.09	79.19	81.29	65.73	**96.25**	75.66	71.31	91.05
MinerU2.5	64.50	88.05	84.28	91.98	91.57	92.39	62.72	95.13	86.88	75.43	70.77	67.99	91.32	73.03	76.86	**96.89**
Youtu-Parsing	83.31	87.70	82.96	92.71	85.91	86.56	**81.17**	93.40	88.12	82.25	73.32	67.89	75.13	70.37	67.06	87.58
MonkeyOCR-Pro-3B	59.84	83.52	78.81	81.18	81.11	83.86	51.73	91.32	79.29	66.98	64.13	61.48	91.24	67.64	58.96	75.04
Dolphin-v2	60.72	77.39	75.06	81.34	80.12	82.81	76.95	89.66	76.45	66.32	63.86	65.34	62.45	64.65	69.64	85.36
End-to-End Specialized VLMs																
OvisOCR2	64.12	87.64	87.87	93.83	87.45	85.39	60.31	96.33	75.37	**87.89**	83.86	68.02	84.29	74.40	77.13	89.89
dots.mocr	86.40	89.02	82.97	87.93	89.99	88.19	54.16	94.32	83.93	80.11	82.11	68.40	69.64	66.15	83.48	88.87
Logics-Parsing-v2	65.12	86.03	84.35	88.24	85.54	78.27	66.66	75.36	83.69	81.24	68.72	69.30	76.35	74.08	71.42	88.83
Qianfan-OCR	61.55	86.86	81.37	91.03	85.00	89.29	61.71	93.68	75.11	71.50	61.01	65.15	93.41	20.90	61.02	89.48
FireRed-OCR	71.84	70.81	60.07	63.68	62.02	69.59	42.85	71.80	61.61	50.95	57.19	70.84	70.45	34.33	52.17	69.47
DeepSeek-OCR-2	71.80	70.01	59.55	62.72	61.84	67.37	42.72	70.53	59.45	48.46	57.32	66.66	62.10	29.96	54.29	65.17
HunyuanOCR-1.5	51.49	69.00	57.91	60.67	61.06	64.26	42.06	71.13	56.88	47.96	56.34	65.60	69.17	36.35	54.94	62.39
General VLMs																
Kimi-K2.6	86.38	80.75	80.56	88.98	86.56	90.05	50.74	92.27	89.42	72.78	82.50	**93.61**	84.33	68.62	75.87	78.51
Kimi K3	**87.53**	85.03	86.92	91.41	85.91	89.40	50.49	94.86	**90.81**	77.50	78.60	92.83	77.11	62.57	84.69	82.57
Qwen3.5-Plus	82.27	82.27	78.93	76.65	80.32	83.37	57.95	85.90	83.85	77.53	76.16	88.54	77.38	63.95	58.60	76.51
Qwen3.8-Max	75.22	74.17	62.71	67.21	62.14	74.68	48.25	83.51	63.10	69.62	67.46	90.98	85.66	48.32	57.30	70.23
Ovis2.6-30B-A3B	76.80	80.05	67.71	66.53	69.97	71.92	49.64	86.91	64.36	66.52	65.72	90.13	68.06	42.71	60.76	75.84

As shown in Table 8, parsing performance also depends strongly on document category. Many specialized systems perform well on financial reports and academic papers, yet struggle with insurance documents and exam papers. General VLMs show a different pattern, with a clear advantage on exam papers despite trailing the leading pipelines overall. These changes in relative performance show that no inference design dominates every category. Broad coverage of financial document types remains challenging even for systems with strong aggregate results.

6 Conclusion

Dianjin-DocBench addresses the need for document parsing evaluation that reflects the requirements of the financial domain. In financial documents, preserving numerical relationships, visual evidence, and content continuity is essential to retaining the meaning of the source material. By making these requirements explicit in its data and evaluation design, the benchmark reveals omissions and reconstruction failures that accurate transcription alone can conceal. Its finance-focused coverage and fine-grained evaluation support more informative comparisons and targeted improvements for financial document OCR. We hope Dianjin-DocBench will help advance document parsing toward more faithful financial information extraction and provide a foundation for research on reliable document understanding in this domain.

AI USE STATEMENT

We used LLM tools to assist with language polishing and translation. The authors reviewed the AI-assisted work and take responsibility for the final manuscript. Generative AI was not used to generate or alter the reported experimental results.

REFERENCES

Lukas Blecher, Guillem Cucurull Preixens, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. In *International Conference on Learning Representations*, volume 2024, pp. 37646–37663, 2024.

Peng Cai, Zhaofan Zou, Shifa Liu, Yikun Wang, Jiawei Tang, Kaicheng Yang, Meng Tong, Zhongjiang He, and Hao Sun. Teleocr: Navigating document parsing across digital and camera-captured documents. *arXiv preprint arXiv:2608.12898*, 2026.

Haoyu Cao, Kun Yin, Yunfei Wu, Bing Liu, Zhongpeng Cai, Xiaotian Li, Huang Chen, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, Yunsheng Wu, Qianyu Li, Antai Guo, Yanzhen Liao, Yanqiu Qu, Haodong Lin, Chengxu He, and Shuangyin Liu. Youtu-parsing: Perception, structuring and recognition via high-parallelism decoding, 2026. URL https://arxiv.org/abs/2601.20430.

Xiangyang Chen, Shuzhao Li, Xiuwen Zhu, Yongfan Chen, Fan Yang, Cheng Fang, Lin Qu, Xiaoxiao Xu, Hu Wei, and Minggang Wu. Logics-parsing technical report, 2025. URL https://arxiv.org/abs/2509.19760.

Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*, 2025.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl-1.5: Towards a multi-task 0.9 b vlm for robust in-the-wild document parsing. *arXiv preprint arXiv:2601.21957*, 2026.

Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M Rush. Image-to-markup generation with coarse-to-fine attention. In *International Conference on Machine Learning*, pp. 980–989. PMLR, 2017.

Daxiang Dong, Mingming Zheng, Dong Xu, Chunhua Luo, Bairong Zhuang, Yuxuan Li, Ruoyun He, Haoran Wang, Wenyu Zhang, Wenbo Wang, Yicheng Wang, Xue Xiong, Ayong Zheng, Xiaoying Zuo, Ziwei Ou, Jingnan Gu, Quanhao Guo, Jianmin Wu, Dawei Yin, and Dou Shen. Qianfan-ocr: A unified end-to-end model for document intelligence, 2026. URL https://arxiv.org/abs/2603.13398.

Yongkun Du, Pinxuan Chen, Xuye Ying, and Zhineng Chen. Docptbench: Benchmarking end-to-end photographed document parsing and translation. *arXiv preprint arXiv:2511.18434*, 2025.

Shuaiqi Duan, Yadong Xue, Weihan Wang, Zhe Su, Huan Liu, Sheng Yang, Guobing Gan, Guo Wang, Zihan Wang, Shengdong Yan, Dexin Jin, Yuxuan Zhang, Guohong Wen, Yanfeng Wang, Yutao Zhang, Xiaohan Zhang, Wenyi Hong, Yukuo Cen, Da Yin, Bin Chen, Wenmeng Yu, Xiaotao Gu, and Jie Tang. Glm-ocr technical report, 2026. URL https://arxiv.org/abs/2603.10910.

Hao Feng, Wei Shi, Ke Zhang, Xiang Fei, Lei Liao, Dingkang Yang, Yongkun Du, Xuecheng Wu, Jingqun Tang, Yang Liu, Hong Chen, and Can Huang. Dolphin-v2: Universal document parsing via scalable anchor prompting, 2026. URL https://arxiv.org/abs/2602.05384.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pp. 1156–1160. IEEE, 2015.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pp. 498–517. Springer, 2022.

Kimi Team. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Gengluo Li, Xingyu Wan, Shangpin Peng, Weinong Wang, Hao Feng, Yongkun Du, Binghong Wu, Zheng Ruan, Zhiqiong Lu, Liang Wu, Pengyuan Lyu, Huawen Shen, Zibin Lin, Shijing Hu, Jieneng Yang, Hongbing Wen, Guanghua Yu, Hong Liu, Bochao Wang, Can Ma, Han Hu, Chengquan Zhang, and Yu Zhou. HunyuanOCR-1.5: Making lightweight OCR VLMs faster and better. *arXiv preprint arXiv:2607.04884*, 2026a.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm, 2025a. URL https://arxiv.org/abs/2506.05218.

Zhang Li, Zhibo Lin, Qiang Liu, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiajun Song, Jiarui Zhang, Xiang Bai, and Yuliang Liu. Mdpbench: A benchmark for multilingual document parsing in real-world scenarios. *arXiv preprint arXiv:2603.28130*, 2026b.

Zichao Li, Aizier Abulaiti, Yaojie Lu, Xuanang Chen, Jia Zheng, Hongyu Lin, Xianpei Han, Shanshan Jiang, Bin Dong, and Le Sun. Readoc: A unified benchmark for realistic document structured extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21889–21905, 2025b.

Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, Jiahe Li, Wen Li, Gui Hu, Yiliang Gu, Siran Yang, Jiamang Wang, Hailong Sun, Yibo Wang, Hui Sun, Jinlong Huang, Yuping He, Shengze Shi, Weihong Zhang, Guodong Zheng, Junpeng Jiang, Sensen Gao, Yi-Feng Wu, Sijia Chen, Yuhui Chen, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Ovis2.5 technical report. *arXiv:2508.11737*, 2025.

Shiyin Lu, Yinglun Li, Yu Xia, Yuhui Chen, An-Yang Ji, Jun-Peng Jiang, Qing-Guo Chen, Jianshan Zhao, En Lin, Haijun Li, Cheng Qin, Zhao Xu, and Weihua Luo. Ovisocr2 technical report. *arXiv preprint arXiv:2607.13639*, 2026.

Jiefeng Ma, Jun Du, Pengfei Hu, Zhenrong Zhang, Jianshu Zhang, Huihui Zhu, and Cong Liu. Hrdoc: dataset and baseline method toward hierarchical reconstruction of document structures. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i2.25277. URL https://doi.org/10.1609/aaai.v37i2.25277.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, et al. Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 13–42, 2026.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24838–24848. IEEE, 2025.

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

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *arXiv preprint arXiv:2502.18443*, 2025.

Siyi Qian, Xiongfei Bai, Bingtao Fu, Yichen Lu, Gaoyang Zhang, Xudong Yang, and Peng Zhang. Lingdt-vl-ocr: Structure-aware document-level parsing with fine-grained visual reference, 2026. URL https://arxiv.org/abs/2603.11044.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pp. 629–633. IEEE, 2007.

Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. Unimernet: A universal network for real-world mathematical expression recognition. *arXiv preprint arXiv:2404.15254*, 2024a.

Bin Wang, Fan Wu, Linke Ouyang, Zhuangcheng Gu, Rui Zhang, Renqiu Xia, Bo Zhang, and Conghui He. Cdm: A reliable metric for fair and accurate formula recognition evaluation. *arXiv preprint arXiv:2409.03643*, 5(6), 2024b.

Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, et al. Mineru2. 5-pro: Pushing the limits of data-centric document parsing at scale. *arXiv preprint arXiv:2604.04771*, 2026.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. Layoutreader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 4735–4744, 2021.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024.

Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr 2: Visual causal flow, 2026. URL https://arxiv.org/abs/2601.20552.

Hao Wu, Haoran Lou, Xinyue Li, Zuodong Zhong, Zhaojun Sun, Phellon Chen, Xuanhe Zhou, Kai Zuo, Yibo Chen, Xu Tang, Yao Hu, Boxiang Zhou, Jian Wu, Yongji Wu, Wenxin Yu, Yingmiao Liu, Yuhao Huang, Manjie Xu, Gang Liu, Yidong Ma, Zhichao Sun, and Changhao Qiao. Firered-ocr technical report, 2026. URL https://arxiv.org/abs/2603.01840.

Hangdi Xing, Changxu Cheng, Feiyu Gao, Zirui Shao, Zhi Yu, Jiajun Bu, Qi Zheng, and Cong Yao. Dochienet: A large and diverse dataset for document hierarchy parsing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1129–1142, 2024.

Zelun Zhang, Hongen Liu, Suyin Liang, Yubo Zhang, Yiqing Xiang, Jiaxuan Liu, Ting Sun, Manhui Lin, Yue Zhang, Changda Zhou, et al. Paddleocr-vl-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training. *arXiv preprint arXiv:2606.03264*, 2026.

Handong Zheng, Yumeng Li, Kaile Zhang, Liang Xin, Guangwei Zhao, Hao Liu, Jiayu Chen, Jie Lou, Qi Fu, Rui Yang, Shuo Jiang, Weijian Luo, Weijie Su, Weijun Zhang, Xingyu Zhu, Yabin Li, Yiwei ma, Yu Chen, Yuqiu Ji, Zhaohui Yu, Guang Yang, Colin Zhang, Lei Zhang, Yuliang Liu, and Xiang Bai. Multimodal ocr: Parse anything from documents, 2026. URL https://arxiv.org/abs/2603.13032.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pp. 1015–1022. IEEE, 2019.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pp. 564–580. Springer, 2020.

Bangbang Zhou, Hangdi Xing, Yifan Chen, Jianjun Xu, Qi Zheng, Feiyu Gao, Zhibo Yang, Shuai Bai, Ming Yan, Jieping Ye, et al. Mpdocbench-parse: Benchmarking practical multi-page document parsing. *arXiv preprint arXiv:2605.22100*, 2026.