Method Type	Methods	Text$^{Edit}\downarrow$		Formula$^{Edit}\downarrow$		Formula$^{CDM}\uparrow$		Table$^{TEDS}\uparrow$		Table$^{Edit}\downarrow$		Read Order$^{Edit}\downarrow$		Overall$^{Edit}\downarrow$	
		EN	*ZH*	*EN*	*ZH*	*EN*	*ZH*	*EN*	*ZH*	*EN*	*ZH*	*EN*	*ZH*	*EN*	*ZH*
Pipeline Tools	MinerU [42]	**0.061**	**0.215**	**0.278**	0.577	57.3	42.9	**78.6**	62.1	**0.18**	0.344	**0.079**	0.292	**0.15**	0.357
	Marker [34]	0.08	0.315	0.53	0.883	17.6	11.7	67.6	49.2	0.619	0.685	0.114	0.34	0.336	0.556
	Mathpix [4]	0.105	0.384	0.306	**0.454**	62.7	**62.1**	77.0	**67.1**	0.243	**0.32**	0.108	0.304	0.191	0.365
Expert VLMs	GOT-OCR [45]	0.189	0.315	0.360	0.528	74.3	45.3	53.2	47.2	0.459	0.52	0.141	0.28	0.287	0.411
	Nougat [7]	0.365	0.998	0.488	0.941	15.1	16.8	39.9	0.0	0.572	1.000	0.382	0.954	0.452	0.973
General VLMs	GPT4o [2]	0.144	0.409	0.425	0.606	72.8	42.8	72.0	62.9	0.234	0.329	0.128	0.251	0.233	0.399
	Qwen2-VL-72B [44]	0.096	0.218	0.404	0.487	**82.2**	61.2	76.8	76.4	0.387	0.408	0.119	**0.193**	0.252	**0.327**
	InternVL2-76B [8]	0.353	0.290	0.543	0.701	67.4	44.1	63.0	60.2	0.547	0.555	0.317	0.228	0.44	0.443

Table 2. Comprehensive evaluation of document parsing algorithms on OmniDocBench: performance metrics for text, formula, table, and reading order extraction, with overall scores derived from ground truth comparisons.

Model Type	Models	Book	Slides	Financial Report	Textbook	Exam Paper	Magazine	Academic Papers	Notes	Newspaper	Overall
Pipeline Tools	MinerU [42]	**0.055**	0.124	**0.033**	**0.102**	**0.159**	0.072	**0.025**	0.984	**0.171**	0.206
	Marker [34]	0.074	0.34	0.089	0.319	0.452	0.153	0.059	0.651	0.192	0.274
	Mathpix [4]	0.131	0.22	0.202	0.216	0.278	0.147	0.091	0.634	0.69	0.3
Expert VLMs	GOT-OCR [45]	0.111	0.222	0.067	0.132	0.204	0.198	0.179	0.388	0.771	0.267
	Nougat [7]	0.734	0.958	1.000	0.820	0.930	0.83	0.214	0.991	0.871	0.806
General VLMs	GPT4o [2]	0.157	0.163	0.348	0.187	0.281	0.173	0.146	0.607	0.751	0.316
	Qwen2-VL-72B [44]	0.096	**0.061**	0.047	0.149	0.195	**0.071**	0.085	**0.168**	0.676	**0.179**
	InternVL2-76B [8]	0.216	0.098	0.162	0.184	0.247	0.150	0.419	0.226	0.903	0.3

Table 3. End-to-end text recognition performance on OmniDocBench: evaluation using edit distance **across 9 PDF page types**.

Models	Fuzzy	Water	Color	None
MinerU [42]	0.15/0.048	**0.151/0.031**	0.107/0.052	0.079/**0.035**
Marker [34]	0.333/0.092	0.484/0.126	0.319/0.127	**0.062**/0.125
Mathpix [4]	0.294/0.064	0.290/0.059	0.216/0.09	0.135/0.043
GOT-OCR [45]	0.175/0.05	0.190/0.056	0.186/0.097	0.177/0.081
Nougat [7]	0.934/0.051	0.915/0.071	0.873/0.096	0.615/0.208
GPT4o [2]	0.263/0.078	0.195/0.057	0.184/0.078	0.186/0.072
Qwen2-VL-72B [44]	**0.082 /0.01**	0.172/ 0.078	**0.104/0.05**	0.084/0.042
InternVL2-76B [8]	0.120/0.013	0.197/0.042	0.155/0.059	0.261/0.082

Table 4. End-to-end text recognition on OmniDocBench: evaluation **under various page attributes** using the edit distance metric. The value is **Mean/Variance** of scores in the attribute group. Columns represent: *Fuzzy* (Fuzzy scan), *Water* (Watermark), *Color* (Colorful background). *None* (No special issue)

Models	Single	Double	Three	Complex
MinerU [42]	0.311/0.187	**0.101/0.013**	**0.117**/0.046	0.385/**0.057**
Marker [34]	0.299/0.143	0.299/0.299	0.149/0.063	**0.363**/0.086
Mathpix [4]	0.207/0.123	0.188/0.07	0.225/**0.029**	0.452/0.177
GOT-OCR [45]	0.163/0.106	0.145/0.059	0.257/0.072	0.468/0.185
Nougat [7]	0.852/0.084	0.601/0.224	0.662/0.093	0.873/0.09
GPT4o [2]	0.109/0.112	0.204/0.076	0.254/0.046	0.426/0.188
Qwen2-VL-72B [44]	**0.066/0.048**	0.145/0.049	0.204/0.055	0.394/0.203
InternVL2-76B [8]	0.082/0.052	0.312/0.069	0.682/0.098	0.444/0.174

Table 5. End-to-end reading order evaluation on OmniDocBench: results **across different column layout types** using Normalized Edit Distance. The value is **Mean/Variance** of scores in the attribute group.

Ignore Handling. We implement an ignore logic for certain components in PDF page content, meaning they participate in matching but are excluded from metric calculations. This is mainly because of inconsistent output standards among models, which should not affect the validation results. For fairness, we ignore: (1) Headers, footers, page numbers, and page footnotes, which are handled inconsistently by different models. (2) Captions for figures, tables, and footnotes often have uncertain placements, thus complicating the reading order. Additionally, some models embed table captions in HTML or LaTeX tables, while others treat them as plain text.

4.3. Metric Calculation

Pure Text. We calculate Normalized Edit Distance [21], averaging these metrics at the sample level to obtain the final scores.

Tables. All tables are converted to HTML format before calculating the Tree-Edit-Distance-based Similarity (TEDS) [54] metric and Normalized Edit Distance.

Formulas. Formulas are currently evaluated using the Character Detection Matching (CDM) metric [41], Normalized Edit Distance, and BLEU [33].

Reading Order. Reading order is evaluated using the Normalized Edit Distance as metric. It only involves text com-

[5] https://github.com/tesseract-ocr/tesseract
[6] https://github.com/VikParuchuri/surya
[7] https://github.com/lukas-blecher/LaTeX-OCR