Model Type	Model	Language			Text background			Text Rotate			
		EN	*ZH*	*Mixed*	*White*	*Single*	*Multi*	*Normal*	*Rotate90*	*Rotate270*	*Horizontal*
Expert Vision Models	PaddleOCR [23]	0.071	**0.055**	**0.118**	**0.060**	**0.038**	**0.085**	**0.060**	**0.015**	0.285	**0.021**
	Tesseract OCR [5]	0.179	0.553	0.553	0.453	0.463	0.394	0.448	0.369	0.979	0.982
	Surya [6]	0.057	0.123	0.164	0.093	0.186	0.235	0.104	0.634	0.767	0.255
	GOT-OCR [45]	0.041	0.112	0.135	0.092	0.052	0.155	0.091	0.562	0.966	0.097
	Mathpix [4]	0.033	0.240	0.261	0.185	0.121	0.166	0.180	0.038	**0.185**	0.638
Vision Language Models	Qwen2-VL-72B [44]	0.072	0.274	0.286	0.234	0.155	0.148	0.223	0.273	0.721	0.067
	InternVL2-76B [8]	0.074	0.155	0.242	0.113	0.352	0.269	0.132	0.610	0.907	0.595
	GPT4o [2]	**0.020**	0.224	0.125	0.167	0.140	0.220	0.168	0.115	0.718	0.132

Table 8. Component-level evaluation on OmniDocBench OCR subset: results grouped by text attributes using the edit distance metric.

Models	CDM	ExpRate@CDM	BLEU	Norm Edit
GOT-OCR [45]	74.1	28.0	55.07	0.290
Mathpix [4]	86.6	2.8	**66.56**	0.322
Pix2Tex [7]	73.9	39.5	46.00	0.337
UniMERNet-B [40]	85.0	60.2	60.84	**0.238**
GPT4o [2]	**86.8**	**65.5**	45.17	0.282
InternVL2-76B [8]	67.4	54.5	47.63	0.308
Qwen2-VL-72B [44]	83.8	55.4	53.71	0.285

Table 9. Component-level formula recognition evaluation on OmniDocBench formula subset.

ever, struggle with high-density documents like newspapers due to limitations in input resolution and token length. In contrast, pipeline tools leverage layout-based segmentation to process components individually, maintaining accuracy in complex layouts. Enhancing VLMs with layout-aware designs and domain-specific fine-tuning offers a promising path forward. OmniDocBench facilitates this by providing detailed annotations for layout, text, formulas, and tables, enabling comprehensive benchmarking and modular tool development for diverse document parsing tasks.

5.2. Single Task Evaluation Results

Layout Detection Results. Layout detection is the first step in document parsing using pipeline tools. A robust layout detection algorithm should perform well across a variety of document types. Table 6 presents an evaluation of leading layout detection models. The DocLayout-YOLO method, which is pre-trained on diverse synthetic document data, significantly outperforms other approaches. This superiority is a key factor in MinerU's integration of DocLayout-YOLO, contributing to its outstanding overall performance. Other methods perform well on books and academic literature but struggle with more diverse formats due to limited training data.

Table Recognition Results. In Table 7, We evaluate table recognition models across three dimensions on our OmniDocBench table subset: language diversity, table frame types, and special situations. Among all models, OCR-based models demonstrate superior overall performance, with RapidTable achieving the highest scores in language diversity and maintaining stable performance across different frame types. Expert VLMs show competitive results in specific scenarios, with StructEqTable [55] excelling in no-frame tables and showing better rotation robustness. General VLMs (Qwen2-VL-7B and InternVL2-8B) exhibit relatively lower but consistent performance, suggesting that while general-purpose VLMs have made progress in table understanding, they still lag behind specialized solutions.

Text Recognition Results. Table 8 compares OCR tools across languages, backgrounds, and rotations using Edit Distance. PaddleOCR outperforms all competitors, followed by GOT-OCR and Mathpix. General VLMs struggle to handle text rotation or mixed-language scenarios.

Formula Recognition Results. Table 9 presents results on formula parsing, using CDM, BLEU, and normalized Edit Distance. GPT-4o, Mathpix, and UniMERNet achieve results of 86.8%, 86.6%, and 85.0%, respectively. Notably, GPT-4o excels with a recall rate of 65.5% under strict conditions requiring perfect character accuracy. Although Mathpix shows high character-level precision, it occasionally omits punctuation, such as commas, leading to a lower overall correctness rate. Nonetheless, all three models are strong candidates for formula recognition tasks.

6. Conclusion

This paper addresses the lack of diverse and realistic benchmarks in document parsing research by introducing OmniDocBench, a dataset featuring a variety of page types with comprehensive annotations, along with a flexible and reliable evaluation framework. OmniDocBench enables systematic and fair assessments of document parsing methods, providing crucial insights for advancing the field. Its task-specific and attribute-level evaluations facilitate targeted model optimization, promoting more robust and effective parsing solutions.

7. Acknowledgments

This project was supported by National Key R&D Program of China (NO.2022ZD0160102) and Shanghai Artificial Intelligence Laboratory.