Benchmark	Document Domain	Annotaion Type					Single-Task Eval				End-to-End Eval			
		BBox	Text	Table	Formula	Attributes	OCR	DLA	TR	MFR	OCR	TR	MFR	ROD
Single-Task Eval Benchmark														
Robust Reading [20]	1	✔	✔				✔							
PubLayNet [49], DocBank [26], DocLayNet [35], M^{6}Doc [9]	1, 1, 5, 6	✔						✔						
PubTabNet [54],TableX [11]	1, 1			✔					✔					
TableBank [25]	1	✔		✔					✔					
Im2Latex-100K [10],UniMER-Test [40]	1				✔					✔				
End-to-end Eval Benchmarks														
Fox [29]	2	✔	✔				✔				✔			
Nougat [7]	1		✔	✔	✔						✔	✔	✔	
GOT OCR 2.0 [45]	2		✔	✔	✔						✔	✔	✔	
OmniDocBench	9	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Table 1. A Comparison between OmniDocBench and existing benchmarks. *BBox*: Bounding boxes. *Text*: Text in Unicode. *Table*: Table in LaTeX/HTML/Markdown. *Formula*: Formula in LaTeX. *Attributes*: Page- and BBox-Level Attributes. *OCR*: Optical Character Recognition; *DLA*: Document Layout Analysis; *TR*: Table Recognition; *MFR*: Math Formula Recognition; *ROD*: Reading Order Detection

this sense, such methods can utilize different expert models to address each specific task. Marker [34] integrates open-source models to parse documents into structured formats such as Markdown, JSON, and HTML. To get higher accuracy, an optional LLM-enabled version can also be integrated to merge tables across pages, handle inline math, and so on. Similarly, MinerU [42] first utilizes a layout detection model to segment the document page into different regions, then applies task-specific models for corresponding regions. Finally, it outputs the complete content in Markdown format with a reading order algorithm. By leveraging lightweight models and parallelized operations, pipeline-based methods can achieve efficient parsing speeds.

2.2. VLM-based Document Content Extraction

Document understanding and optical character recognition (OCR) are crucial tasks for evaluating the perception capabilities of vision-language models (VLMs). By incorporating extensive OCR corpus into the pretraining stage, VLMs like GPT4o [2] and Qwen2-VL [3] have demonstrated comparable performance in document content extraction tasks. Unlike pipeline-based methods, VLMs perform document parsing in an end-to-end manner. Furthermore, without requiring specialized data fine-tuning, these models are able to deal with diverse and even unseen document types for their generalization capabilities.

To integrate the efficiency of lightweight models and the generalizability of VLMs, many works [7, 14, 29, 32, 45, 46] have focus on training specialized end-to-end expert models for document parsing. These VLM-driven models excel at comprehending both visual layouts and textual contents, balancing a trade-off between accuracy and efficiency.

2.3. Benchmarks for Document Content Extraction

Document content extraction requires the ability to understand document layouts and recognize various types of content. However, current benchmarks fall short of a comprehensive page-level evaluation, as they focus solely on evaluating the model's performance on module-level recognition. PubLayNet [49] and concurrent benchmarks [9, 26, 35] specialize in evaluating a model's ability to detect document page layouts. OCRBench [31] proposes five OCR-related tasks with a greater emphasis on evaluating the model's visual understanding and reasoning capabilities. Only line-level assessments are provided for text recognition and handwritten mathematical expression recognition (HMER). Similarly, single-module benchmarks [20, 26, 40, 54] disentangle the task into different dimensions and focus narrowly on specific parts. Such paradigm overlooks the importance of structural and semantic information like the reading order and fails to evaluate the model's overall ability when processing the full-page documents as a whole.

Page-level benchmarks have been proposed alongside some recent VLM-driven expert models [7, 29, 45]. However, the robustness of these benchmarks is compromised by limitations in data size, language, document type, and annotation. For example, Nougat [7] evaluates models using only printed English documents collected from arXiv while the page-level benchmark introduced by GOT-OCR [45] consists of only 90 pages of Chinese and English documents in total. Commonly-seen document types like handwritten notes, newspapers, and exam papers are further neglected. Lacking detailed annotations, the benchmarks can only conduct naive evaluation between the full-page results of Ground Truths and predictions without special handling for different output formats and specialized metrics for different content types. The evaluation of the model performance can be severely biased due to limited document domains, unaligned output format and mismatched metrics. *Therefore, there is an urgent need for a more finely annotated, diverse, and reasonable page-level document content extraction benchmark.*