Model	Backbone	Params	Book	Slides	Research Report	Textbook	Exam Paper	Magazine	Academic Literature	Notes	Newspaper	Average
DiT-L [24]	ViT-L	361.6M	43.44	13.72	45.85	15.45	3.40	29.23	**66.13**	0.21	23.65	26.90
LayoutLMv3 [17]	RoBERTa-B	138.4M	42.12	13.63	43.22	21.00	5.48	31.81	64.66	0.80	30.84	28.84
DocLayout-YOLO [53]	v10m	19.6M	**43.71**	**48.71**	**72.83**	**42.67**	**35.40**	51.44	64.64	9.54	**57.54**	**47.38**
SwinDocSegmenter [4]	Swin-L	223M	42.91	28.20	47.29	32.44	20.81	**52.35**	48.54	**12.38**	38.06	35.89
GraphKD [5]	R101	44.5M	39.03	16.18	39.92	22.82	14.31	37.61	44.43	5.71	23.86	27.10
DOCX-Chain [50]	-	-	30.86	11.71	39.62	19.23	10.67	23.00	41.60	1.80	16.96	21.27

Table 6. Component-level layout detection evaluation on OmniDocBench layout subset: mAP results by PDF page type.

Model Type	Model	Language			Table Frame Type				Special Situation				Overall
		EN	*ZH*	*Mixed*	*Full*	*Omission*	*Three*	*Zero*	*Merge Cell*(+/-)	*Formula*(+/-)	*Colorful*(+/-)	*Rotate*(+/-)	
OCR-based Models	PaddleOCR[23]	76.8	71.8	80.1	67.9	74.3	81.1	74.5	70.6/75.2	71.3/74.1	72.7/74.0	23.3/74.6	73.6
	RapidTable[37]	**80.0**	**83.2**	**91.2**	**83.0**	**79.7**	**83.4**	78.4	**77.1/85.4**	**76.7/83.9**	**77.6/84.9**	25.2/**83.7**	**82.5**
Expert VLMs	StructEqTable[55]	72.8	75.9	83.4	72.9	76.2	76.9	**88.0**	64.5/81.0	69.2/76.6	72.8/76.4	**30.5**/76.2	75.8
	GOT-OCR [45]	72.2	75.5	85.4	73.1	72.7	78.2	75.7	65.0/80.2	64.3/77.3	70.8/76.9	8.5/76.3	74.9
General VLMs	Qwen2-VL-7B [44]	70.2	70.7	82.4	70.2	62.8	74.5	80.3	60.8/76.5	63.8/72.6	71.4/70.8	20.0/72.1	71.0
	InternVL2-8B [8]	70.9	71.5	77.4	69.5	69.2	74.8	75.8	58.7/78.4	62.4/73.6	68.2/73.1	20.4/72.6	71.5

Table 7. Component-level Table Recognition evaluation on OmniDocBench table subset. *(+/-)* means *with/without* special situation.

ponents, with tables, images, and ignored components excluded from the final reading order calculation.

5. Benchmarks

Based on the distinct characteristics of these algorithms, we categorize document content extraction methods into three main classes:

- **Pipeline Tools:** These methods integrate layout detection and various content recognition tasks (such as OCR, table recognition, and formula recognition) into a document parsing pipeline for content extraction. Prominent examples include MinerU [42] (v0.9.3), Marker [34] (v1.2.3), and Mathpix[4].
- **Expert VLMs:** These are large multimodal models specifically trained for document parsing tasks. Representative models include GOT-OCR2.0 [45] and Nougat [7].
- **General VLMs:** These are general-purpose large multimodal models inherently capable of document parsing. Leading models in this category include GPT-4o [2], Qwen2-VL-72B [44], and InternVL2-76B [8].

5.1. End-to-End Evaluation Results

Overall Evaluation Results. As illustrated in Table 2, pipeline tools such as MinerU and Mathpix, demonstrate superior performance across sub-tasks like text recognition, formula recognition, and table recognition. Moreover, the general Vision Language Models (VLMs), Qwen2-VL, and GPT4o, also exhibit competitive performance. Almost all algorithms score higher on English than on Chinese pages.

Performance Across Diverse Page Types. To gain deeper insights into model performance on diverse document types, we evaluated text recognition tasks across different page types. Intriguingly, as shown in Table 3, pipeline tools perform well for commonly used data, such as academic papers and financial reports. Meanwhile, for more specialized data, such as slides and handwritten notes, general VLMs demonstrate stronger generalization. Notably, most VLMs fail to recognize when dealing with the Newspapers, while pipeline tools achieve significantly better performance.

Performance on Pages with Visual Degradations. In Table 4, we further analyze performance on pages containing common document-specific challenges, including fuzzy scans, watermarks, and colorful backgrounds. VLMs like InternVL2 and Qwen2-VL exhibit higher robustness in these scenarios despite visual noise. Among pipeline tools, MinerU remains competitive due to its strong layout segmentation and preprocessing capabilities.

Performance on Different Layout Types. Page layout is a critical factor in document understanding, especially for tasks involving reading order. OmniDocBench annotates layout attributes such as single-column, multi-column, and complex custom formats. Across all models, we observe a clear drop in accuracy on multi-column and complex layouts. MinerU shows the most consistent reading order prediction, though its performance dips on handwritten single-column pages due to recognition noise.

Discussion on End-to-End Results. 1) While general VLMs often lag behind specialized pipelines and expert models on standard documents (e.g., academic papers), they generalize better to unconventional formats (e.g., notes) and perform more robustly under degraded conditions (e.g., fuzzy scans). This is largely due to their broader training data, enabling better handling of long-tail scenarios compared to models trained on narrow domains. 2) VLMs, how-