MosaicDoc: A Large-Scale Bilingual Benchmark for Visually Rich Document Understanding

Ketong Chen, Yuhao Chen, Yang Xue*

School of Electronic and Information Engineering, South China University of Technology
eeckt@mail.scut.edu.cn, chardmaplesan@gmail.com, yxue@scut.edu.cn

Abstract

Despite the rapid progress of Vision-Language Models (VLMs), their capabilities are inadequately assessed by existing benchmarks, which are predominantly English-centric, feature simplistic layouts, and support limited tasks. Consequently, they fail to evaluate model performance for Visually Rich Document Understanding (VRDU), a critical challenge involving complex layouts and dense text. To address this, we introduce DocWeaver, a novel multi-agent pipeline that leverages Large Language Models to automatically generate a new benchmark. The result is MosaicDoc, a large-scale, bilingual (Chinese and English) resource designed to push the boundaries of VRDU. Sourced from newspapers and magazines, MosaicDoc features diverse and complex layouts (including multi-column and non-Manhattan), rich stylistic variety from 196 publishers, and comprehensive multi-task annotations (OCR, VQA, reading order, and localization). With 72K images and over 600K QA pairs, MosaicDoc serves as a definitive benchmark for the field. Our extensive evaluation of state-of-the-art models on this benchmark reveals their current limitations in handling real-world document complexity and charts a clear path for future research.

Code — https://github.com/DOCLAB-SCUT/MosaicDoc
Extended version — https://arxiv.org/abs/2511.09919

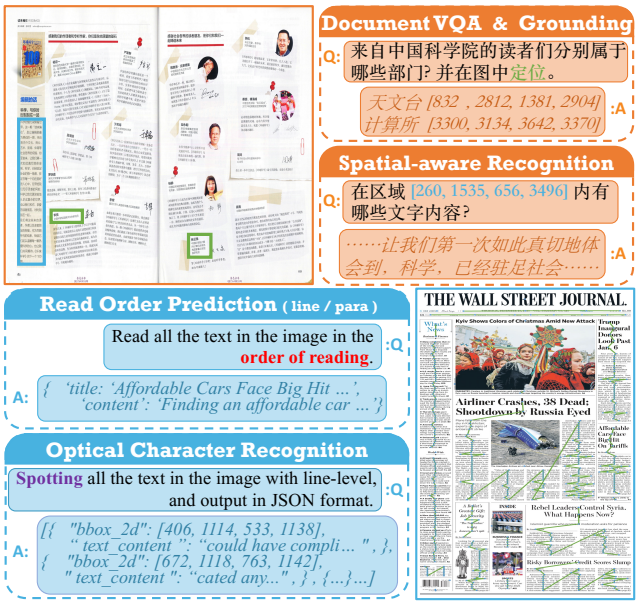

Figure 1: Examples of VRDU Tasks in the MosaicDoc Benchmark.

Introduction

With the rapid advancement of Document AI, the field is converging on unified, end-to-end models. Document Visual Question Answering (DocVQA) has emerged as a key paradigm, capable of unifying diverse tasks into a single, prompt-based framework (Tang et al. 2023; Ye et al. 2023; Feng et al. 2023). However, the development of these powerful models is fundamentally constrained by the benchmarks used for their evaluation. The central challenge lies in Visually Rich Document Understanding (VRDU), which requires models to comprehend documents with complex layouts, dense text, and diverse visual styles. This remains a domain where current datasets fall critically short.

Existing benchmarks, such as VisualMRC (Tanaka, Nishida, and Yoshida 2021), and LLM-generated ones like Docmatix (Laurençon et al. 2024), are overwhelmingly English-centric and rest upon visually simple documents. They lack stylistic variety and layout complexity representative of real-world artifacts like newspapers and magazines. This limitation is particularly acute for non-English documents. Existing Chinese datasets also fail to capture real-world visual complexity. For instance, XFUND (Xu et al. 2022) largely replicates the simplistic layouts of its English counterparts, while DuReadervis (Qi et al. 2022), despite using long webpage, utilizes only the visually simple top portions, neglecting their richer content. Crucially, they lack reading order annotations, forcing models to rely on a flawed top-to-bottom reading assumption that fails on multi-column or non-Manhattan layouts. As our experiments confirm, this leads to significant performance degradation in state-of-the-art (SOTA) models.

To address these challenges, we propose DocWeaver, a novel multi-agent pipeline powered by LLMs for generating high-fidelity, multi-task document annotations. DocWeaver

*Correspond author

employs a modular architecture of specialized agents, where Extractors parse document elements, Generators create diverse questions, and five distinct Hallucination Guardrails rigorously validate each sample for quality and factual consistency. This pipeline explicitly computes and utilizes correct reading order to ensure genuine layout comprehension. Using this method, we introduce MosaicDoc, a large-scale bilingual benchmark for Visually Rich Document Understanding. Sourced from modern newspapers and magazines, MosaicDoc provides the first comprehensive resource for evaluating models on complex, real-world documents in both Chinese and English. It contains over 600K question-answer pairs across 72K images from 196 publishers, with rich annotations supporting a wide array of tasks, including DocVQA, OCR, reading order prediction, and content-aware localization. We conduct a rigorous evaluation of thirteen SOTA models on MosaicDoc, establishing a new performance baseline.

In summary, our contributions are three-fold:

- A Novel Automated Pipeline. We introduce **DocWeaver**, a fully automated multi-agent pipeline that leverages LLMs to generate high-fidelity, multi-task annotations, addressing the criticla data creation bottleneck for visually rich documents.
- A Challenging New Benchmark. We present **MosaicDoc**, a large-scale bilingual benchmark from complex newspapers and magazines. It introduces unprecedented diverse in layout and style to facilitate more rigorous research in VRDU.
- A Rigorous Performance Analysis. We provide a comprehensive benchmark analysis of 13 state-of-the-art models on MosaicDoc. This evaluation reveals systemic weaknesses in current approaches, particularly in handling dense layouts and multi-span reasoning, thereby establishing a new more challenging baseline for the field.

Related Work

We review key datasets in Document Visual Understanding (DVU), focusing on two core tasks where current benchmarks show significant limitations: Reading Order Prediction (ROP) and Document Visual Question Answering (DocVQA).

Datasets for Reading Order Prediction

Reading Order Prediction, which aims to infer coherent reading sequences from 2D layouts, is a cornerstone of document comprehension. While methods like LayoutReader (Wang et al. 2021) and Dolphin (Feng et al. 2025) have drawn attention, progress is constrained by the lack of robust and diverse datasets. Early examples like ReadingBank (Wang et al. 2021) provides word-level text sequences for 500K documents but lacks block-level annotations for structural evaluation, which is critical for complex layouts like multi-column Manhattan (Liu, Li, and Wei 2025). Other efforts, such as ROOR (Zhang et al. 2024), augments existing datasets like FUNSD (Jaume, Ekenel, and Thiran 2019) with reading order but remains limited in scale and layout diversity.

More recently, large-scale ROP datasets have emerged, including DocGenome (Xia et al. 2024) on scientific papers and olmOCR-mix-0225 (Poznanski et al. 2025) on web-scraped PDFs. However, a critical limitation persists across all these benchmarks: they are dominated by documents with simple, homogeneous layouts(e.g., single- or two-column academic papers). This fails to capture the complexity of real-world documents with non-Manhattan or multi-column layouts, thereby providing an inadequate challenge for evaluating robust reading order prediction. Furthermore, they are often single-task and mono-lingual.

Datasets for Document Visual Question Answering

DocVQA benchmark similar lack real-world complexity. Many datasets target individual pages or components, with multi-page understanding being a nascent and flawed area. For instance, MP-DocVQA (Tito, Karatzas, and Valveny 2023) extends SP-DocVQA (Mathew, Karatzas, and Jawahar 2021) but lacks questions that require reasoning across pages, while DUDE (Van Landeghem et al. 2023) suffers from short contextual spans and inconsistent quality due to crowdsourcing. Other datasets are domain-specific, such as TAT-DQA (Zhu et al. 2024) for financial reports emphasizing structured tables and numerical reasoning within a narrow domain, SlideVQA (Tanaka et al. 2023) for low-density presentation slides, and InfographicsVQA (Mathew et al. 2022), which prioritizes visual elements like charts and icons but underrepresents dense text and complex layouts. A crucial deficiency across nearly all these datasets is the scarcity of questions requiring multi-span reasoning.

Recent LLM-based generation methods, such as RefChartQA (Vogel et al. 2025) for charts and TVG (Liu et al. 2024a) for tables, primarily automate annotation for well-structured elements rather than introducing new, complex visual domains. Even TRIG (Li et al. 2025) provides visual grounding, ultimately falls back on human validation, failing to establish a fully automated, scalable pipeline.

Our work directly addresses these identified gaps. The DocWeaver pipeline introduces a fully automated generation and validation process, moving beyond the manual verification required by prior work. Most critically, the resulting MosaicDoc benchmark introduces visually complex and stylistically diverse newspaper and magazine documents in both English and Chinese. By providing rich, multi-task annotations for this challenging new domain, MosaicDoc fills a crucial void in the DVU landscape, offering a more realistic and rigorous testbed for future research.

The DocWeaver Pipeline

To automate the creation of a large-scale, multi-task benchmark from visually rich document, we introduce DocWeaver, a multi-agent collaborative pipeline. As illustrated in Figure 2, DocWeaver is designed to systematically decompose complex documents, generate high-quality questions, and rigorously validate the outputs without manual intervention. The pipeline operates in three phases: (1) Document Decomposition and Structuring, (2) High-Fidelity QA Generation, and (3) Automated Quality Assurance.

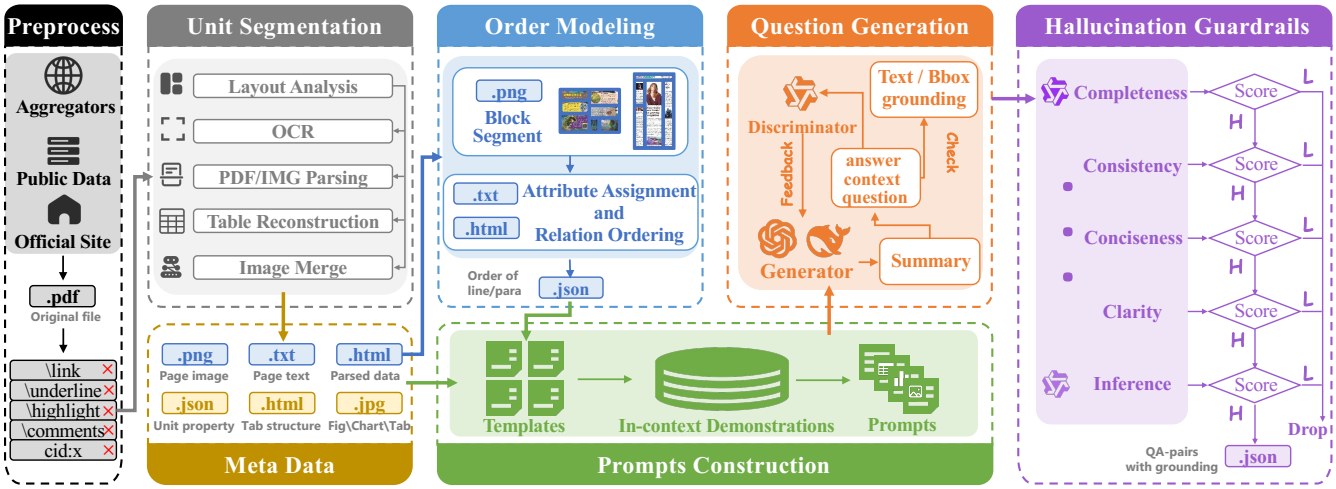

Figure 2: Overview of DocWeaver, a Multi-Agent Pipeline for Generating the MosaicDoc Benchmark.

Document Decomposition and Structuring

This initial phase transforms raw PDF documents into a structured, machine-readable format.

Data Preprocessing The pipeline parses digital PDFs using PyMuPDF[1] and pdfminer.six[2] to extract raw text and bounding boxes. Semantically irrelevant elements such as annotations, hyperlinks, underlines, and highlights are discarded. To handle encoding errors in PDFs(e.g., malformed symbols like <?> or cid:x), we employ chardet[3] to detect encoding, ensuring text recovery. Each page is rendered into two raster images: one at high resolution for precise layout analysis, and another at randomized DPI to simulate real-world rendering variance and increase dataset diversity.

Unit Segmentation We employ a suite of specialized agents to decompose each document page into its constituent semantic units. The process begins with a Layout Detection Agent, PP-DocLayout (Sun et al. 2025) fine-tuned on the M^{6}Doc magazine dataset (Cheng et al. 2023), which identifies 13 distinct layout element types (e.g., titles, paragraphs, and subhead). To recover text missed by PDF parsers, a Multi-Engine OCR Agent then processes each detected layout element with three different OCR engines[4]. The final text is determined via a weighted-voting scheme that considers both engine reliability and confidence scores.

$$\hat{s} = \arg\max_{t \in \mathcal{T}} \sum_{i=1}^{N} w_i p_i \mathbf{1}[s_i = t] \quad (1)$$

where $\mathcal{T} = \{s_1, s_2, ...s_N\}$ denotes all candidate strings from each OCR engine. $\mathbf{1}[\cdot]$ is the indicator function, w_i is the weight reflecting the expected reliability of the i-th engine, and p_i is the confidence score. Following recognition, global char-level positions are recovered from PaddleOCR's CTC decoder.

Subsequently, specialized agents handle structural and relational information. A Table Structure Agent, using the TableStructureRec toolbox[5], converts the detected tables to a structured HTML format. Concurrently, a Reading Order Agent powered by MinerU (Wang et al. 2024) predicts an initial reading order for documents with standard layouts. To handle content spanning multiple pages, a Cross-Page Linking Agent leverages DeepSeek[6] to compute semantic similarity between adjacent pages. Pages with a similarity score exceeding 0.8 are then automatically merged into a single logical document, up to a four pages limit. This structured decomposition provides a high-fidenlity foundation for all subsequent annotation tasks.

Complex Reading Order Modeling A key challenge in VRDU is modeling the non-linear and diverse reading orders often found in newspapers and magazines. Our approach addresses this through two tailored strategies (Please refer to Appendix A for more details).

For Documents with Structured Data (e.g., magazines and Chinese newspapers) where official HTML versions are available, we use the HTML structure as a reference template. Text spans extracted from the PDF are mapped to the template using fuzzy matching with edit distance, constrained by bounding box and contextual text proximity. The final correct reading order is then derived from the validated sequence within each layout block.

For Documents without Structured Data (e.g., English newspapers), we employ a hybrid approach combining structural analysis and semantic clustering. Page images are first binarized to detect layout lines, which are used to partition the page into rectangular content blocks, each representing a distinct article. Then, within each block, line-level reading order is inferred using PDF metadata (e.g., font features, positional cues) and semantic coherence.

[1]https://pymupdf.readthedocs.io
[2]https://pdfminersix.readthedocs.io/
[3]https://pypi.org/project/chardet/
[4]https://github.com/PaddlePaddle/PaddleOCR
https://pypi.org/project/pytesseract/
https://github.com/JaidedAI/EasyOCR

[5]https://github.com/RapidAI/TableStructureRec
[6]https://www.deepseek.com/

Dataset	Sources	#Images	Tokens *(avg±std)*	#Tasks	Lang	#Ques	Unique *(%)*
Document Visual Question Answering (DocVQA)							
MP-DocVQA	Industry docs	12.7K	$489.6_{\pm 411.2}$	2	en	41.4K	84.2
TAT-DQA	Finance reports	2.8K	$850.9_{\pm 252.0}$	2	en	16.6K	88.0
InfographicsVQA	Infographics	5.5K	$382.7_{\pm 231.7}$	3	en	27.1K	98.5
DuReader_{vis}	Web	12.6K	$3,186.8_{\pm 1,902.3}$	2	zh	14.1K	99.7
DUDE	Multi	27.9K	$2,944.0_{\pm 3,937.1}$	4	en	30.1K	86.9
Reading Order Prediction (ROP)							
ROOR	Scan forms	0.2K	$323.6_{\pm 149.3}$	3	en	-	-
ReadingBank	Ebooks	500K	$314.6_{\pm 144.6}$	2	en	-	-
olmOCR-mix-0225	Web, Ebooks	266K	$661.0_{\pm 598.7}$	1	en	-	-
MosaicDoc (Ours)	Magzines	42.7K	$1,075.1_{\pm 860.6}$	6	en, zh	304.6K	99.7
	Newspapers	29.6K	$3,557.9_{\pm 2051.2}$	6	en, zh	318.5K	99.8

Table 1: Comparison of MosaicDoc with existing benchmarks for Document Visual Understanding. MosaicDoc is the first large-scale benchmark to combine visually rich sources (magazines, newspapers), bilingual support (English and Chinese), and comprehensive multi-task annotations. This directly addresses the key limitations of prior work, which are often restricted to simpler layouts, a single language, or fewer tasks.

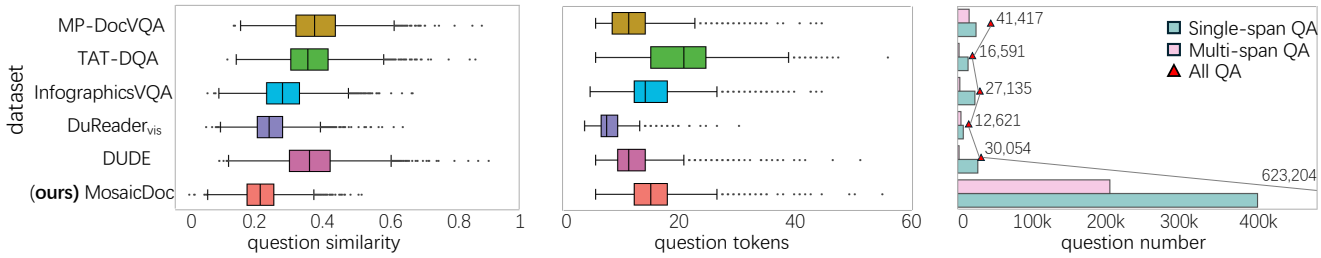

Figure 3: Left panels show the distributions of question similarity and token length, while right panel compares number of multi-span and single-span question-answer pairs instances

High-Fidelity QA Generation

This phase leverages the structured document representation to generate diverse and challenging question-answer pairs.

Prompts Construction To elicit a wide range of question types, we developed specialized prompt templates for text, tables, and charts (see Appendix B for details). These templates guide the model to generate a diverse question set. Additionally, we build an in-context learning repository containing examples for single-span, multi-span, and table- and chart-based QA. We source examples from datasets such as MultiSpanQA(Li et al. 2022), InfographicsVQA(Mathew et al. 2022) and TableBench(Wu et al. 2025), supplemented with 400 manually crafted examples for domain-specific coverage. During generation, we randomly sample four well-designed and four dataset-sourced examples to serve as few-shot demonstrations within the prompt.

Question Generation We employ powerful LLMs, GPT-4o[7] and DeepSeek-R1, as QA generators. To ensure high quality and factual grounding, the generation process follows a structured, multi-step workflow. First, the model generates a summary of the input content to identify key information. Based on the summary, a question is formulated. Then, the model is explicitly instructed to locate the corresponding answer-bearing context within the original content and to extract the answer directly, avoiding reliance on prior knowledge. Furthermore, we introduce QWQ-32B (Team 2025) as an auxiliary discriminator that assesses whether the retrieved context supports cross-sentence reasoning. If so, the context along with a prompt is returned to the generator with explicit instructions to formulate multi-span questions with detailed reasoning chains and supporting evidence. The final answer and its evidence are then matched back to precisely locate their positions in the document image.

Automated Quality Assurance

Hallucination Guardrails Despite rigorous prompting, minor hallucinations may still occur. Inspired by the G-EVAL(Liu et al. 2023) framework, we implement a final filtering stage using QWQ as an LLM evaluator serving as an automated "hallucination guardrail". Each generated QA pair is scored from 1 to 5 across five distinct criteria, and

[7]https://openai.com/es-ES/api/

only those passing a high threshold on all criteria are retained (see Appendix B for details). The five evaluation criteria are:

- **Completeness:** Assess if the answer contains all essential elements without omitting any critical information.
- **Consistency:** Verify the answer factually aligns with source document.
- **Conciseness:** Evaluate if the answer avoids redundancy information.
- **Clarity:** Determine if the answer is free from ambiguous references (e.g., "this year", "here").
- **inference:** Validates the precision of any calculations or logical reasoning.

The MosaicDoc Benchmark

This section details the construction, quality assurance process, and key statistical properties of the MosaicDoc benchmark.

Data Sources

To construct a benchmark that reflects real-world visual diversity, we carefully curated a collection of documents from official public websites, content-rich aggregation platforms, and open-source PDF repositories, ensuring compliance with all data usage and licensing requirements (details in Appendix D). The resulting MosaicDoc benchmark is composed of four subsets: Chinese magazines, Chinese newspapers, English magazines, and English newspapers. The data is sourced from 196 distinct publishing institutions across 24 diverse domains, including science, finance, culture, and the arts. The full distribution of sources is visualized in Figure 4.

Quality Assurance

Automated Filtering As described in the DocWeaver section, every generated question-answer pair is passed through our five-criteria "Hallucination Guardrail" system, with only high-scoring pairs being retained (results in Table 2). Additionally, we apply a similar automated check for reading order. If a document page contains more than five detected reading order errors judged by distance-based heuristics (see Appendix C), the entire page is discarded.

Human Validation To establish a gold-standard test set for evaluation, we conducted a meticulous manual validation. We randomly sampled 200 document images and their corresponding annotations from across all four subsets. For this sample, human annotators manually corrected any remaining errors in the QA pairs and reading order sequences, following the same criteria as our automated guardrails. This high-fidelity, human-validated subset is used for all experimental evaluations reported in this paper, ensuring our results are grounded in the most accurate data possible.

Dataset Properties and Statistics

MosaicDoc was created to fill a critical resource gap in Document AI research. Its properties demonstrate its significant advantages over existing datasets in scale, complexity, and task diversity.

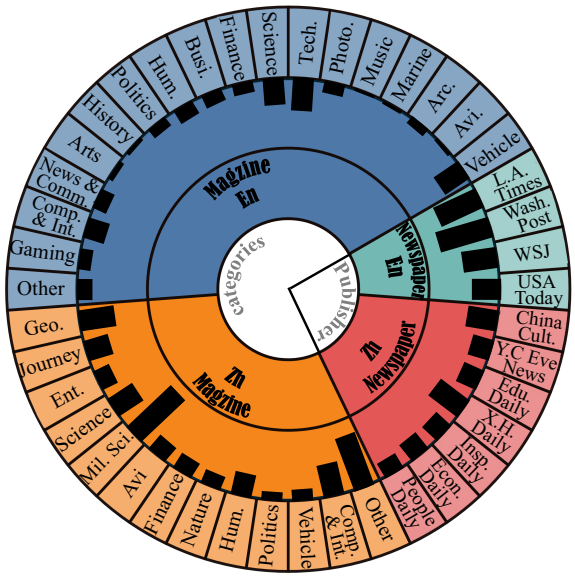

Figure 4: Source category distribution of MosaicDoc dataset (abbreviated; full details in Appendix C).

	Previous	After	Ratio *(%)*
en-Magzine	178.2k	160.1k	89.8
zh-Magzine	164.7k	144.5k	87.7
en-Newspaper	184.7k	163.6k	88.6
zh-Newspaper	176.1k	154.9k	87.9

Table 2: The counts of QA pairs after guardrails filtering

Scale and Composition The full benchmark comprises over 72.3K document images from newspapers and magazines, annotated with fine-grained OCR and reading order labels. Layered on top are over 620K question-answer pairs that query content from text, tables and charts.

Textual and Layout Complexity MosaicDoc is defined by its information density. As shown in Table 1, the average number of tokens per document is significantly higher than in most other VQA datasets. While DuReader-vis also features high token counts, its web-page screenshots often contain sparse layouts where answers are confined to small regions. In contrast, MosaicDoc's sources feature dense, structured text across the entire page. The complexity of these layouts is further quantified by average sentence-level BLEU metrics (see Figure 5), confirming that MosaicDoc presents a more challenging structural reasoning task.

Question Quality and Diversity The questions generated by DocWeaver are not only numerous but also sophisticated. Using a suite of Qwen models[8], we sample and analyze 5,000 questions per dataset, computing semantic similarity scores and token lengths. We found that questions in MosaicDoc have a higher average token length and signifi-

[8]https://huggingface.co/Qwen/Qwen3-Embedding-0.6B

Sou.	Type	Expert Models		Expert VLMs			General VLMs				
		Donut	ViTLP	Vary	TextMonkey	mPlug-DocOWL2	CogVLM2	InternVL3	Qwen2.5-VL	GPT-4o	Gemini-2.5
Parameters		253M	259M	7B	7B	7B	19B	9B	7B	API	API
Mag.	S	8.13 / –	6.91 / –	1.78 / –	5.43 / –	13.75 / 3.11	31.86 / 18.79	54.11 / 48.68	57.86 / 58.59	47.64 / 18.76	65.78 / 59.27
	M	1.26 / –	1.28 / –	0.49 / –	4.65 / –	5.38 / 5.73	23.77 / 18.38	40.78 / 38.68	43.81 / 40.39	42.08 / 18.03	55.63 / 63.67
	T	13.11 / –	10.04 / –	8.4 / –	3.02 / –	18.54 / 8.27	30.42 / 19.04	53.53 / 45.27	54.62 / 54.84	46.19 / 28.06	68.37 / 70.64
	C	3.38 / –	5.78 / –	2.18 / –	1.38 / –	22.72 / 14.95	30.46 / 15.71	33.42 / 33.32	41.42 / 41.88	32.80 / 32.69	42.65 / 50.39
	All	5.87 / –	5.14 / –	1.81 / –	4.78 / –	11.70 / 3.97	28.93 / 18.41	48.43 / 43.98	51.98 / 51.33	48.86 / 20.75	61.27 / 68.82
News.	S	4.54 / –	3.80 / –	3.87 / –	2.16 / –	9.80 / 1.62	18.43 / 7.98	51.81 / 38.44	61.19 / 51.63	32.34 / 10.70	68.47 / 59.05
	M	0.50 / –	0.68 / –	0.62 / –	2.86 / –	1.45 / 0.04	11.50 / 2.10	29.26 / 26.54	36.42 / 35.71	25.14 / 6.55	55.14 / 53.97
	T	2.62 / –	0.79 / –	6.15 / –	0.00 / –	6.12 / 2.73	8.82 / 10.90	39.05 / 33.59	41.37 / 50.02	21.42 / 13.74	48.55 / 75.16
	C	5.72 / –	3.52 / –	7.69 / –	5.43 / –	27.10 / 16.53	22.09 / 28.82	39.09 / 34.74	42.35 / 43.67	29.57 / 29.80	48.22 / 62.91
	All	3.63 / –	7.69 / –	3.70 / –	2.32 / –	8.97 / 1.98	16.37 / 6.65	45.80 / 31.77	52.99 / 44.79	29.65 / 10.17	62.47 / 59.76

Table 3: The ANLSL score of Baseline performance on the MosaicDoc dataset. The figures before the "/" denote English subset, while those after it denote Chinese subset. The type of questions are abbreviated as (S)ingle-span, (M)ulti-span, (T)able and (C)hart. Sou. means source, Mag. means magzine and News. means newspaper.

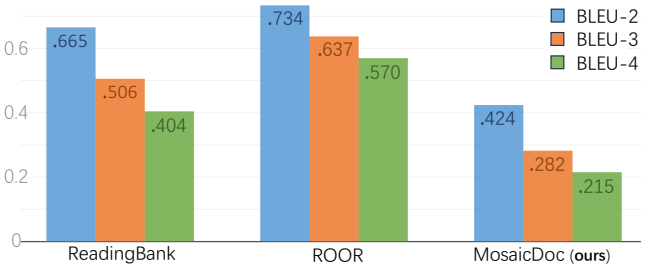

Figure 5: The BLEU scores are calculated for the left-to-right and top-to-bottom order to measure the layout complexity of MosaicDoc.

cantly lower semantic similarity to each other compared to other datasets, indicating greater diversity and less repetition. With a uniqueness score of over 99.7%, nearly every question is distinct. Crucially, MosaicDoc contains a substantial volume of multi-span QA pairs, addressing a well-known gap in existing DocVQA datasets and enabling the evaluation of complex answer extraction scenarios that require aggregating information from multiple spans within a document.

Comprehensive Multi-Task Support Unlike single-task benchmarks like ReadingBank or domain-specific ones like DuReader-vis, MosaicDoc provides a unified platform for a broad range of VRDU tasks. Its rich annotations support: (1) Document VQA, (2) Word- and line-level OCR, (3) Block- and line-level Reading Order Prediction, and (4) Content-aware Localization. By providing a single, comprehensive resource that bridges these tasks, MosaicDoc serves as a more holistic and challenging benchmark for advancing Document AI.

Evaluation

To demonstrate the challenges posed by MosaicDoc and to establish a new performance baseline for Visually Rich Document Understanding (VRDU), we conduct a comprehensive evaluation of 13 SOTA models.

Experimental Setting

Baselines We categorize the 13 recent SOTA models into three groups:

- **Expert Models** include pre-LLM architectures like LayoutReader(Wang et al. 2021), Donut(Kim et al. 2022), and ViTLP(Mao et al. 2024). LayoutReader uses line-level text bounding boxes as input, whereas Donut and ViTLP consume only the image and question. None of these models incorporate large language models (LLMs), and all have smaller parameter counts (<0.3B).
- **Expert VLMs** comprise models like Vary-7B(Wei et al. 2024a), mPLUG-DocOwl2-7B(Hu et al. 2025), TextMonkey-7B(Liu et al. 2024b), GOT-OCR-0.5B(Wei et al. 2024b), and olmOCR-8B(Poznanski et al. 2025). These are built on LLM backbones and are specifically pretrained on large document datasets for tasks like DocVQA and ROP.
- **General VLMs** include powerful, general-purpose vision-language models like CogVLM2-19B(Hong et al. 2024), InternVL3-9B(Zhu et al. 2025), and Qwen2.5-VL-7B(Bai et al. 2025). In addition, we accessed GPT-4o and Gemini-2.5[9](Comanici et al. 2025) through their APIs. While not specialized for documents, they possess state-of-the-art capabilities across a wide range of vision-language tasks.

Implementation Details All models are evaluated in a zero-shot setting using their official default configurations. Images are provided at the maximum resolution supported by each model. All VLM evaluations were conducted within the VLLM framework on NVIDIA A100 GPUs with 80GB memory. Further implementation details and qualitative results are provided in Appendix E.

Results and Analysis

We evaluate model performance on three core VRDU tasks: Document VQA, Page-Level OCR, and Reading Order Prediction.

[9]https://gemini.google.com/

Document VQA We evaluate DocVQA performance using the Average Normalized Levenshtein Similarity for List (ANLSL) metric ((Tito, Karatzas, and Valveny 2021)). The comprehensive results, broken down by question type, are presented in Table 3.

Our analysis reveals several key findings. First, General VLMs significantly outperform both Expert Models and a majority of the specialized Expert VLMs. This suggests that the massive, diverse pre-training data and superior architectural scaling of general models provide a more robust foundation for VRDU than domain-specific pre-training on simpler documents.

Second, Expert VLMs struggle unexpectedly on MosaicDoc's sources. We primarily attribute this to token reduction strategies employed by models like TextMonkey and mPLUG-DocOwl2. While effective for simple documents, these methods, which merge or discard visual tokens, likely cause critical semantic loss when applied to the information-dense layouts of magazines and newspapers.

Third, all models exhibit a dramatic and consistent performance drop on multi-span questions. Even the top-performing local model Qwen2.5-VL and the remote model Gemini-2.5 see a sharp decline in ANLS score compared to their single-span performance. This highlights a critical and universal weakness in current models' ability to perform multi-span extraction that requires synthesizing information across complex layouts. Interestingly, performance on table and chart questions is comparatively higher, likely because these tasks rely more on locating structured objects than on fine-grained reading of dense paragraphs.

Page-level OCR To assess raw text recognition capability, we evaluate page-level Character Recognition Rate (OCR) using CRR and Output-based Character Recognition Rate (OCRR). CRR measures accuracy against the ground truth, while OCRR normalizes by the length of the model's output, measuring precision within the generated text.

As shown in Table 4, all models perform significantly worse on the newspaper subset, particularly for Chinese newspapers, underscoring the extreme difficulty of their dense and complex layouts. We also observe a widespread and common failure mode where VLMs, after generating a certain amount of text, begin producing repetitive sequences until reaching their token limit, resulting in lower OCRR. This severely penalizes the CRR score and indicates a breakdown in contextual understanding when processing long, dense visual inputs, even when the initial recognized text is accurate. This failure to read the full page effectively limits a model's ability to answer content-related questions.

Reading Order Prediction We evaluate ROP using the Micro-F1 score on text line sequences. Since VLMs produce plain text, we determine the correctness of the predicted order by matching text blocks to the ground truth sequence.

The results in Table 5 reveal a highly consistent trend across most models: high precision paired with low recall. This indicates that while the text fragments models do recognize are often in the correct local sequence (e.g., within a single column), they fail to capture the entire content of the page. This problem is exacerbated in newspapers, where horizontally adjacent blocks are often spatially closer than vertically sequential lines within an article, challenging simple top-down heuristics. Despite strong performance across the three subsets, Gemini can only effectively order text within the same column and fails to establish the correct sequential relationship between paragraph-level texts in multi-column or non-Manhattan layouts—a challenge that lies at the core of reading order recovery in such complex document layouts (see results and visualizations in Appendix F.4). This failure to comprehend the global layout contributes to the repetitive outputs seen in the OCR task and fundamentally undermines the model's ability to understand the document as a whole.

Model	**Magzine** (en / zh)		**Newspaper** (en / zh)	
	CRR↑	**OCRR↑**	**CRR↑**	**OCRR↑**
GOT-OCR	22.27 / 31.46	45.62 / 41.15	3.09 / 7.15	1.13 / 5.17
olmOCR	73.76 / 53.85	86.33 / 74.20	32.58 / 1.19	52.00 / 12.45
InternVL3	66.06 / 59.29	74.65 / 61.38	56.03 / 31.77	58.59 / 44.89
Qwen2.5-VL	55.24 / 57.29	23.60 / 28.40	34.46 / 39.62	11.17 / 31.40
GPT-4o	69.81 / 30.33	75.88 / 39.15	37.55 / 2.69	53.17 / 15.34
Gemini-2.5	89.42 / 87.34	90.32 / 80.31	87.90 / 66.64	91.04 / 78.18

Table 4: Page-level OCR recognition results on MosaicDoc

Model	**Magzine** (en / zh)			**Newspaper** (en / zh)		
	P	**R**	**F1**	**P**	**R**	**F1**
LayoutReader	5.93/10.6	– / –	– / –	5.40/3.19	– / –	– / –
GOT-OCR	2.55/7.05	0.61/2.72	0.99/3.93	14.2/13.5	0.55/0.17	1.07/0.34
olmOCR	92.7/91.1	73.3/58.7	81.9/71.4	60.1/22.6	23.5/0.36	33.8/0.71
InternVL3	86.1/87.6	62.1/70.7	72.2/76.6	82.2/74.8	52.2/28.4	63.9/41.2
Qwen2.5-VL	92.6/93.5	50.6/63.6	65.4/75.7	90.0/92.0	35.5/41.5	50.9/57.2
GPT-4o	85.9/74.6	60.1/33.2	70.7/45.9	74.6/60.0	33.2/17.3	45.1/26.9
Gemini-2.5	91.9/93.8	84.2/85.4	87.9/89.4	95.7/94.2	81.3/40.2	87.9/56.3

Table 5: Text Line reading order prediction results

Conclusion

In this work, we introduced DocWeaver, a novel multi-agent pipeline for automated data generation, and used it to construct MosaicDoc, a large-scale bilingual benchmark for VRDU. By focusing on complex newspaper and magazine documents (a domain largely neglected by previous research), MosaicDoc provides the community with a more realistic and challenging testbed. Its diversity in language, layout, and task support addresses critical limitations of existing datasets. Our extensive evaluation of state-of-the-art models on MosaicDoc reveals significant and previously unmeasured systemic weaknesses, particularly in handling dense layouts and performing multi-span reasoning, thereby charting a clear path for future research. While DocWeaver has proven effective, we view this as a foundational step. Future work will focus on extending the pipeline to diverse new domains, such as historical and handwritten documents, to further push the boundaries of robust document intelligence.

Acknowledgments

This research is supported in part by the China National Key Research and Development Program (2022YFC3301702) and National Natural Science Foundation of China (grant no.61771199).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cheng, H.; Zhang, P.; Wu, S.; Zhang, J.; Zhu, Q.; Xie, Z.; Li, J.; Ding, K.; and Jin, L. 2023. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15138–15147.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Feng, H.; Wang, Z.; Tang, J.; Lu, J.; Zhou, W.; Li, H.; and Huang, C. 2023. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*.

Feng, H.; Wei, S.; Fei, X.; Shi, W.; Han, Y.; Liao, L.; Lu, J.; Wu, B.; Liu, Q.; Lin, C.; et al. 2025. Dolphin: Document Image Parsing via Heterogeneous Anchor Prompting. *arXiv preprint arXiv:2505.14059*.

Hong, W.; Wang, W.; Ding, M.; Yu, W.; Lv, Q.; Wang, Y.; Cheng, Y.; Huang, S.; Ji, J.; Xue, Z.; et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Hu, A.; Xu, H.; Zhang, L.; Ye, J.; Yan, M.; Zhang, J.; Jin, Q.; Huang, F.; and Zhou, J. 2025. mPLUG-DocOwl2: High-resolution Compressing for OCR-free Multi-page Document Understanding. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 5817–5834. Association for Computational Linguistics.

Jaume, G.; Ekenel, H. K.; and Thiran, J.-P. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, 1–6. IEEE.

Kim, G.; Hong, T.; Yim, M.; Nam, J.; Park, J.; Yim, J.; Hwang, W.; Yun, S.; Han, D.; and Park, S. 2022. OCR-Free Document Understanding Transformer. In *European Conference on Computer Vision (ECCV)*.

Laurençon, H.; Marafioti, A.; Sanh, V.; and Tronchon, L. 2024. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.

Li, H.; Tomko, M.; Vasardani, M.; and Baldwin, T. 2022. MultiSpanQA: A dataset for multi-span question answering. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1250–1260.

Li, M.; Zhang, R.; Chen, J.; Gu, J.; Zhou, Y.; Dernoncourt, F.; Zhu, W.; Zhou, T.; and Sun, T. 2025. Towards visual text grounding of multimodal large language model. *arXiv preprint arXiv:2504.04974*.

Liu, S.; Li, Y.; and Wei, J. 2025. XY-Cut++: Advanced Layout Ordering via Hierarchical Mask Mechanism on a Novel Benchmark. *arXiv preprint arXiv:2504.10258*.

Liu, S.; Zhang, Z.; Hu, P.; Ma, J.; Du, J.; Wang, Q.; Zhang, J.; and Liu, C. 2024a. See then Tell: Enhancing Key Information Extraction with Vision Grounding. *arXiv preprint arXiv:2409.19573*.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Liu, Y.; Yang, B.; Liu, Q.; Li, Z.; Ma, Z.; Zhang, S.; and Bai, X. 2024b. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

Mao, Z.; Bai, H.; Hou, L.; Shang, L.; Jiang, X.; Liu, Q.; and Wong, K. 2024. Visually Guided Generative Text-Layout Pre-training for Document Intelligence. In Duh, K.; Gómez-Adorno, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, 4713–4730. Association for Computational Linguistics.

Mathew, M.; Bagal, V.; Tito, R.; Karatzas, D.; Valveny, E.; and Jawahar, C. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1697–1706.

Mathew, M.; Karatzas, D.; and Jawahar, C. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2200–2209.

Poznanski, J.; Rangapur, A.; Borchardt, J.; Dunkelberger, J.; Huff, R.; Lin, D.; Wilhelm, C.; Lo, K.; and Soldaini, L. 2025. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *arXiv preprint arXiv:2502.18443*.

Sun, T.; Cui, C.; Du, Y.; and Liu, Y. 2025. PP-DocLayout: A Unified Document Layout Detection Model to Accelerate Large-Scale Data Construction. *arXiv preprint arXiv:2503.17213*.

Tanaka, R.; Nishida, K.; and Yoshida, S. 2021. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13878–13888.

Tang, Z.; Yang, Z.; Wang, G.; Fang, Y.; Liu, Y.; Zhu, C.; Zeng, M.; Zhang, C.; and Bansal, M. 2023. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19254–19264.

Team, Q. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Tito, R.; Karatzas, D.; and Valveny, E. 2021. Document collection visual question answering. In *International Conference on Document Analysis and Recognition*, 778–792. Springer.

Tito, R.; Karatzas, D.; and Valveny, E. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144: 109834.

Van Landeghem, J.; Tito, R.; Borchmann, Ł.; Pietruszka, M.; Joziak, P.; Powalski, R.; Jurkiewicz, D.; Coustaty, M.; Anckaert, B.; Valveny, E.; et al. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19528–19540.

Vogel, A.; Moured, O.; Chen, Y.; Zhang, J.; and Stiefelhagen, R. 2025. Refchartqa: Grounding visual answer on chart images through instruction tuning. *arXiv preprint arXiv:2503.23131*.

Wang, B.; Xu, C.; Zhao, X.; Ouyang, L.; Wu, F.; Zhao, Z.; Xu, R.; Liu, K.; Qu, Y.; Shang, F.; et al. 2024. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Wang, Z.; Xu, Y.; Cui, L.; Shang, J.; and Wei, F. 2021. Layoutreader: Pre-training of text and layout for reading order detection. *arXiv preprint arXiv:2108.11591*.

Wei, H.; Kong, L.; Chen, J.; Zhao, L.; Ge, Z.; Yang, J.; Sun, J.; Han, C.; and Zhang, X. 2024a. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, 408–424. Springer.

Wei, H.; Liu, C.; Chen, J.; Wang, J.; Kong, L.; Xu, Y.; Ge, Z.; Zhao, L.; Sun, J.; Peng, Y.; et al. 2024b. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Wu, X.; Yang, J.; Chai, L.; Zhang, G.; Liu, J.; Du, X.; Liang, D.; Shu, D.; Cheng, X.; Sun, T.; et al. 2025. Tablebench: A comprehensive and complex benchmark for table question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25497–25506.

Xia, R.; Mao, S.; Yan, X.; Zhou, H.; Zhang, B.; Peng, H.; Pi, J.; Fu, D.; Wu, W.; Ye, H.; et al. 2024. DocGenome: An Open Large-scale Scientific Document Benchmark for Training and Testing Multi-modal Large Language Models. *arXiv preprint arXiv:2406.11633*.

Xu, Y.; Lv, T.; Cui, L.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; and Wei, F. 2022. XFUND: A Benchmark Dataset for Multilingual Visually Rich Form Understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, 3214–3224. Dublin, Ireland: Association for Computational Linguistics.

Ye, J.; Hu, A.; Xu, H.; Ye, Q.; Yan, M.; Xu, G.; Li, C.; Tian, J.; Qian, Q.; Zhang, J.; Jin, Q.; He, L.; Lin, X.; and Huang, F. 2023. UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2841–2858. Singapore: Association for Computational Linguistics.

Zhang, C.; Tu, Y.; Zhao, Y.; Yuan, C.; Chen, H.; Zhang, Y.; Chai, M.; Guo, Y.; Zhu, H.; Zhang, Q.; et al. 2024. Modeling Layout Reading Order as Ordering Relations for Visually-rich Document Understanding. *arXiv preprint arXiv:2409.19672*.

Zhu, F.; Liu, Z.; Feng, F.; Wang, C.; Li, M.; and Chua, T. S. 2024. Tat-llm: A specialized language model for discrete reasoning over financial tabular and textual data. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 310–318.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.