ViDocParse: A Multi-Level Benchmark for Vietnamese Document Parsing

Anonymous ACL submission

Abstract

Document parsing converts page images into structured representations that preserve text, layout, tables, figures, formulas, and reading order, serving as a key input layer for Retrieval-Augmented Generation and document-centered NLP. Despite recent progress in Vision-Language Models (VLMs), Vietnamese remains underrepresented in document parsing benchmarks. Its tone and vowel diacritics, syllable-based writing system, and dense multi-column layouts pose challenges for current parsers and can degrade downstream evidence extraction. We introduce **ViDocParse**, a multi-level Vietnamese document parsing dataset with 11,439 document-page images across nine categories, annotated with block-level and span-level labels following the OmniDocBench schema. Using ViDocParse, we benchmark seven contemporary OCR/VLM systems under identical conditions and provide fine-grained analyses across language, layout, table structure, image quality, and block type. Our results show that no single model dominates across all document attributes, and that layout-sensitive errors such as column merging, table omission, and inline-formula boundary drift remain major challenges for Vietnamese document parsing. We further fine-tune dots.mocr with low-rank adaptation (LoRA) on three training splits—generated, manually annotated, and mixed—releasing three checkpoints together with the dataset and evaluation code to support future research on Vietnamese document AI.

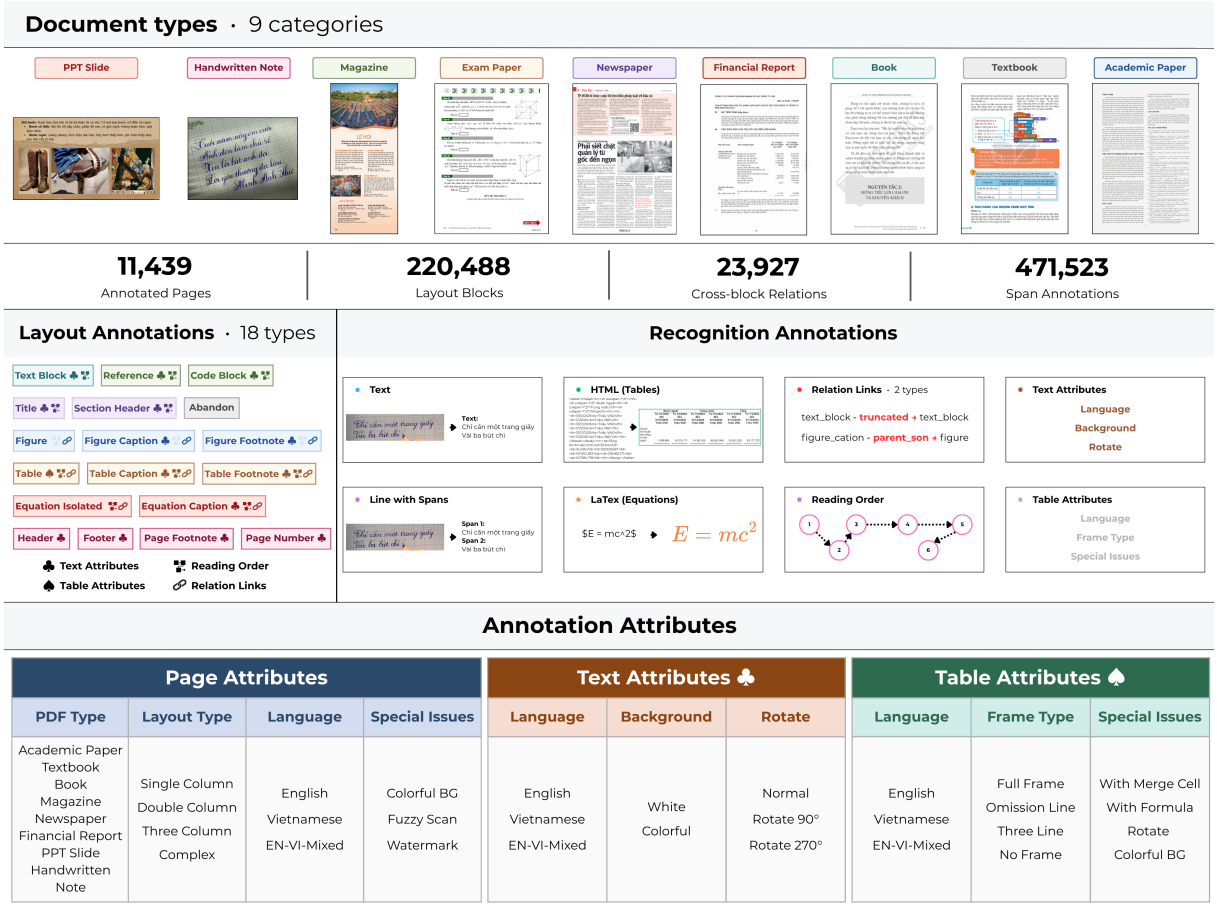
1 Introduction

Document-centered NLP increasingly relies on visually rich sources, including papers, textbooks, financial reports, exam papers, slides, and scanned records. To support Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), question answering, information extraction, or other Large Language Model (LLM) pipelines, these documents need to be converted into structured, machine-readable content. This task, known as Document Content Extraction (DCE), goes beyond traditional Optical Character Recognition (OCR): it requires recognizing text while preserving layout, tables, formulas, captions, and reading order. Errors at this stage can corrupt retrieved evidence and propagate to downstream language generation.

Recent Vision-Language Models (VLMs), such as PaddleOCR-VL-1.5 (Cui et al., 2026) and dots.mocr (Zheng et al., 2026), have achieved strong results on general document parsing benchmarks such as OmniDocBench (Ouyang et al., 2025). However, their performance on Vietnamese remains underexplored. Vietnamese poses distinct linguistic and visual challenges: tone and vowel diacritics encode lexical meaning, while spacing and reading order strongly affect word interpretation. As a result, errors in diacritics, column order, line grouping, or table reconstruction can introduce semantic drift that downstream NLP models may not recover from. These challenges are further amplified by diverse Vietnamese document layouts, including dense newspapers, textbooks, exam papers, financial reports, and handwritten notes.

Despite its importance, Vietnamese lacks a large-scale, multi-level benchmark for document-page parsing. To address this gap, we introduce **ViDocParse**, a Vietnamese document parsing dataset and benchmark for converting document-page images into structured content. The dataset contains 11,439 pages across nine document categories, with block-level and span-level annotations following the OmniDocBench schema. Using this dataset, we evaluate seven contemporary OCR/VLM systems and analyze their behavior across language, layout, table, image-quality, and block-level attributes. Our contributions are threefold:

- We introduce **ViDocParse**, a multi-level annotated Vietnamese document parsing dataset



Annotation Attributes

Page Attributes				Text Attributes ♣			Table Attributes ♠		
PDF Type	Layout Type	Language	Special Issues	Language	Background	Rotate	Language	Frame Type	Special Issues
Academic Paper Textbook Book Magazine Newspaper Financial Report PPT Slide Handwritten Note	Single Column Double Column Three Column Complex	English Vietnamese EN-VI-Mixed	Colorful BG Fuzzy Scan Watermark	English Vietnamese EN-VI-Mixed	White Colorful	Normal Rotate 90° Rotate 270°	English Vietnamese EN-VI-Mixed	Full Frame Omission Line Three Line No Frame	With Merge Cell With Formula Rotate Colorful BG

Figure 1: Overview of ViDocParse data diversity. Representative samples from all nine document categories illustrating the range of layouts, scan qualities, and content types.

with 11,439 document-page images across nine categories.

- We benchmark seven OCR/VLM systems—dots.mocr (Zheng et al., 2026), DeepSeek-OCR-2 (Wei et al., 2026), Chandra OCR 2 (Datalab, 2026), HunyuanOCR (Team et al., 2025), PaddleOCR-VL-1.5 (Cui et al., 2026), LightOn-OCR-2 (Taghadouini et al., 2026), and MinerU 2.5-Pro (Wang et al., 2024a)—under identical conditions.

- We provide adaptation baselines by fine-tuning dots.mocr (Zheng et al., 2026) with LoRA (Hu et al., 2022) on generated, manual, and mixed training splits.

2 Related Work

2.1 Evolution of Document Parsing Methods

Document content extraction has evolved through three paradigm shifts. First-generation systems were rule-based pipelines: Tesseract (Smith, 2007) and its derivatives applied connected-component analysis followed by character-level classification, achieving reasonable accuracy on clean, single-column printed text but failing on complex layouts. Second-generation approaches introduced deep-learning-based modular pipelines, in which separate models handle layout detection (e.g., PubLayNet (Zhong et al., 2019)), table structure recognition (Zhong et al., 2020), and formula conversion (Deng et al., 2017). MinerU (Wang et al., 2024a) exemplifies this paradigm: it chains a layout detector, a text-line recognizer, and a formula-to-LaTeX converter, achieving strong results on structured PDFs but suffering from error propagation across modules. Transformer-based document understanding models such as LayoutLMv3 (Huang et al., 2022) and Donut (Kim et al., 2022) bridged this generation and the next, jointly modeling text and visual features within a single architecture but still relying on task-specific heads or OCR preprocessing.

The third and current generation replaces the

pipeline with a single Vision-Language Model that conditions text generation directly on the document image. Built on the Qwen2-VL backbone (Wang et al., 2024b), recent VLMs — dots.mocr (Zheng et al., 2026) (the successor of dots.ocr (Li et al., 2025)), PaddleOCR-VL-1.5 (Cui et al., 2026), DeepSeek-OCR-2 (Wei et al., 2026), Chandra 2 (Datalab, 2026), HunyuanOCR (Team et al., 2025) and LightOn-OCR-2 (Taghadouini et al., 2026) — jointly perform layout detection, reading-order inference and content generation in one forward pass, with prompt-engineering details differing across models. These VLMs achieve state-of-the-art on English-centric benchmarks, but their performance on Vietnamese remains insufficiently studied.

2.2 Benchmarks for Document Parsing

Early benchmarks for document parsing focused on individual subtasks. PubLayNet (Zhong et al., 2019) provided 360,000 document images annotated for layout detection, while ICDAR table competitions evaluated table structure recognition in isolation. These benchmarks cover only English documents and assess one module at a time, making them insufficient for evaluating modern end-to-end VLMs.

OmniDocBench (Ouyang et al., 2025) addressed this gap by introducing a multi-level evaluation framework across nine document types with 19 layout categories and 15 attribute labels. It supports end-to-end evaluation using metrics such as *Edit Distance* ($\text{Edit}_{\text{dist}}$), *Table Edit-Distance-based Similarity* (TEDS), and *Character Detection Metric* (CDM). OmniDocBench has become a widely used reference benchmark for evaluating document parsing systems.

However, OmniDocBench and other existing benchmarks (e.g., DocPedia (Feng et al., 2024)) are predominantly composed of English and Chinese documents. For non-Latin scripts, KITAB-Bench (Heakl et al., 2025) recently targeted Arabic OCR with 8,809 samples across 36 sub-domains, revealing a 60% CER gap between traditional OCR and VLMs. To our knowledge, there is currently no comparable benchmark or dataset for Vietnamese document parsing, leaving a gap for Vietnamese NLP.

2.3 Challenges Specific to Vietnamese Document Parsing

Tonal and Diacritical Complexity. Vietnamese uses a Latin-derived script with vowel-modifying marks (e.g., $a \rightarrow \breve{a}$, $\hat{a}$) and six tone marks. At 200 DPI, these are often visually indistinguishable. Since tone marks are phonemic, a single error can transform a word into a semantically unrelated one (e.g., *ma, má, mà, mả, mã, mạ*), potentially corrupting retrieved evidence in RAG pipelines.

Word Segmentation. As an isolating language (Vu et al., 2018), Vietnamese morphemes are syllable-based. Reading-order errors do not merely put words out of sequence—they create entirely different, meaningless morpheme combinations that are unrecoverable by downstream language models (Nguyen and Nguyen, 2020).

Layout Distribution Shift. Although contemporary document VLMs often report multilingual capabilities, their training and evaluation settings are still largely shaped by high-resource languages and dominant document styles, especially English and Chinese. Vietnamese documents, however, follow neither the Western borderless-table / inline-caption style nor the Chinese vertical / heavy-grid layout: they typically use fully bordered tables, lateral figure captions, and dense multi-column newspaper layouts. These patterns are underrepresented in common pretraining and evaluation distributions, making robust transfer to Vietnamese document layouts difficult.

3 ViDocParse: Dataset Construction and Annotation

ViDocParse is a multi-level annotated dataset for Vietnamese document parsing[1]. It contains 11,439 document-page images spanning nine categories that span the breadth of Vietnamese printed and handwritten content (Table 1 and Figure 1). Each image is paired with a JSON annotation file containing both block-level and span-level labels following the OmniDocBench schema.

3.1 Annotation Schema

Each page carries six page-level attributes (*data_source*, *language*, *layout*, *watermark*, *col-*

[1]Dataset and evaluation code are available anonymously for review. Code: https://anonymous.4open.science/r/ViDocParse-E6B1. Dataset: https://huggingface.co/datasets/emnlp2026anon/ViDocParse/ (Access token: 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).

Category	Pages	Source	Lang.
Academic Paper	3,204	Gen.+Manual	EN/VI/Mix
Book	544	Gen.+Manual	VI
Exam Paper	768	Manual	VI
Financial Rep.	1,962	Gen.+Manual	VI/Mix
Handwritten	1,088	Gen.+Manual	VI
Magazine	605	Manual	VI
Newspaper	1,233	Manual	VI
PPT Slide	861	Gen.+Manual	EN/VI/Mix
Textbook	1,174	Gen.+Manual	VI
Total	**11,439**		

Table 1: ViDocParse dataset composition by document category. "Gen." = AI-generated; "Manual" = scanned/photographed.

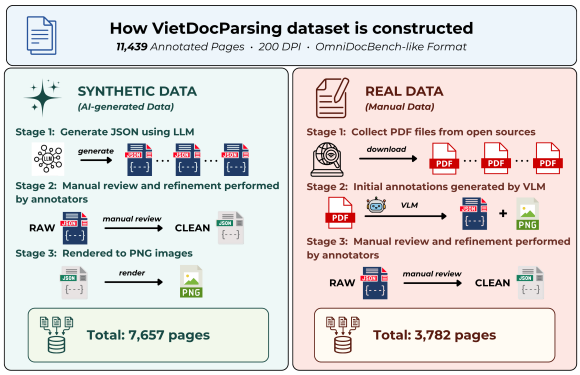


Figure 2: Dataset construction pipeline. Left: generated pipeline (JSON generated by Gemini 3.0 Flash, reviewed, then rendered to PNG). Right: manual pipeline with semi-automatic annotation and quality control.

orful_background, *fuzzy_scan*) and a list of block-level detections. Every block records a `category_type` drawn from an 18-class taxonomy adapted from OmniDocBench (Ouyang et al., 2025), a polygon bounding box (`poly`), and category-specific content. Text blocks carry a UTF-8 Markdown `text` field plus three Vietnamese-aware attributes (*text_language*, *text_background*, *text_rotate*); table blocks store a minified `html` string plus a *frame_type* attribute $\in$ {`full_line`, `less_line`, `fewer_line`, `wireless_line`}; and the optional `line_with_spans` field decomposes a text block into `text_span`, `equation_inline` and `footnote_mark` sub-elements. The full taxonomy and attribute lists are given in Appendix B.

3.2 Data Collection and Quality Control

The dataset construction proceeds through two parallel pipelines, as illustrated in Figure 2. The **generated pipeline** (synthetic data) utilizes Gemini 3.0 Flash to generate structured document layouts in the OmniDocBench JSON format. These generated representations then undergo manual review and refinement before being rendered into 200 DPI PNG images, yielding pixel-perfect alignment. The **manual pipeline** scans or photographs real-world documents (books, exam papers, newspapers, magazines, etc.) and annotates them in three steps: (i) a layout detection model proposes initial bounding boxes; (ii) dots.ocr (Li et al., 2025) performs initial text recognition; (iii) human annotators correct every block. All Vietnamese strings are validated against a 74,000-token Vietnamese vocabulary dictionary to catch residual tone-mark and vowel errors.

Consent and privacy. Manually collected documents were obtained from public, institutional, or author-controlled sources and were reviewed before inclusion. Pages containing sensitive personal identifiers were excluded or anonymized. The released dataset is intended for research on Vietnamese document parsing and does not include documents requiring private individual consent.

Quality control. Annotations pass through automated validation scripts (enforcing language- and background-attribute consistency, inline-formula delimiter normalisation, table HTML minification, frame-type classification, and text-rotation detection via Tesseract OSD (Smith, 2007)) followed by two rounds of manual review. The full procedure is detailed in Appendix C.

3.3 Dataset Splits

The dataset is partitioned into four splits via metadata manifests: `test` (1,108 pages, held out across all nine categories), `train_generated` (7,247 pages, AI-generated only), `train_manual` (3,084 pages, human-annotated only), and `train_mixed` (10,331 pages, the union of the two training splits). The three training splits let us isolate the contribution of synthetic versus real-world supervision in our fine-tuning experiments (§6). All benchmark numbers in this paper use the held-out `test` split exclusively; the full per-split-per-category breakdown is given in Appendix A.

4 Benchmarking Methodology

We evaluate all seven models on the held-out `test` split (1,108 pages, see §3.3) under identical conditions using the OmniDocBench (Ouyang et al., 2025) evaluation toolkit. For each model we feed in the raw document image together with a task-

appropriate prompt drawn from its official HuggingFace card (verbatim prompts are listed in Appendix D), and the produced Markdown is matched block-by-block against ground truth.

Primary metrics. We report four metrics, each instantiated as a normalised character- or node-level Levenshtein-style distance:

- **Text Edit Distance** ($\text{Text}^{\text{Edit}}$, ↓): normalised character-level edit distance (Marzal and Vidal, 1993) between predicted and ground-truth text blocks.
- **Formula Edit Distance** ($\text{Formula}^{\text{Edit}}$, ↓): normalised edit distance between predicted and ground-truth LaTeX strings of isolated formulas.
- **Table TEDS** (↑): Tree-Edit-Distance-based Similarity (Zhong et al., 2020) between predicted and ground-truth HTML DOM trees. We additionally report TEDS-S, the structure-only variant that ignores cell text.
- **Reading-Order Edit Distance** (RO^{Edit}, ↓): edit distance over the sequence of block reading-order indices.

Overall score. Following OmniDocBench v1.6, we collapse the three content metrics into a single 0–100 *Overall* score by averaging each metric's 0–100 form. Writing $T = \text{Text}^{\text{Edit}}$, $F = \text{Formula}^{\text{Edit}}$ and $\tau = \text{Table}^{\text{TEDS}}$,

$$\text{Overall} = \tfrac{1}{3}\big[(1-T)\cdot 100 + (1-F)\cdot 100 + \tau\big]. \quad (1)$$

Hardware and inference. All experiments run end-to-end on an NVIDIA H100 NVL 94 GB GPU; each model is loaded, evaluated over the full test set, and unloaded before the next to avoid memory interference. For VLM systems we use greedy decoding with a maximum generation length of 8,192 tokens; MinerU 2.5-Pro uses its default two-step pipeline (layout analysis followed by per-crop recognition).

5 Results and Analysis

5.1 Overall Benchmark Results

Table 2 reports the headline numbers on the `test` split. End-to-end VLMs occupy the entire top half of the ranking, and the gap between the leader (dots.mocr, Overall = 86.91) and the only pipeline OCR system (MinerU 2.5-Pro, 60.08) is driven almost exclusively by Text Edit Distance (0.115 vs. 0.538): a five-fold character-level error. Chandra 2 ranks second overall and posts the best Text Edit Distance (0.105) and the best structure-only Table TEDS-S (94.22), suggesting a strong text-and-table head despite a comparatively weak formula module. The remaining VLMs cluster between 72.9 and 81.8 Overall and trade strengths along different axes, which motivates the per-attribute analysis that follows.

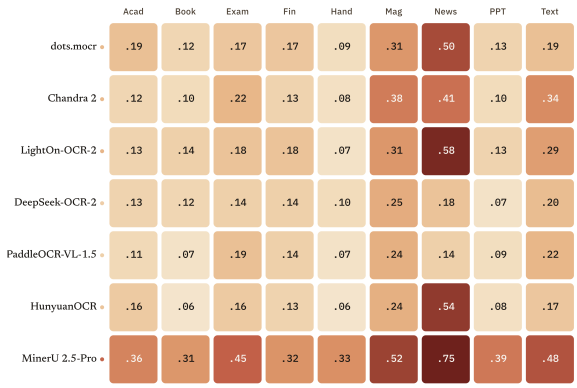


Figure 3: Text Edit Distance (↓) broken down by document category. Rows are sorted by Overall score (Table 2); columns: *Acad* = academic paper, *Fin* = financial report, *Hand* = handwritten note, *Mag* = magazine, *News* = newspaper, *Text* = textbook.

5.2 Analysis by Attribute

Our annotation schema records both page-level and block-level attributes (§3.1), which we use to slice the test set and re-compute Text Edit Distance within each slice. Figure 3 visualises performance by document category, Figure 4 reports the Vietnamese–English gap, Table 3 groups slices by background, layout, and image artefacts, and Table 4 reports an HTML-structure proxy of Table TEDS-S broken down by table frame type.

Per-category behaviour. PaddleOCR-VL-1.5 dominates on academic papers (0.113) and newspapers (0.137), confirming its strength on dense columnar Latin layouts. HunyuanOCR is the most consistent on Vietnamese-heavy categories (best on *book*, *handwritten*, *magazine* and *textbook*, all $\leq$ 0.24). The widest gap appears on *newspaper*, where Vietnamese Gothic-style fonts push dots.mocr, HunyuanOCR and LightOn-OCR-2 above 0.49 while PaddleOCR-VL-1.5 stays at 0.137. MinerU 2.5-Pro is the worst on every category — the bottom row of Figure 3.

Model	Overall ↑	Text^{Edit} ↓	Formula^{Edit} ↓	Table^{TEDS} ↑	Table$^{\text{TEDS-S}}$ ↑	RO^{Edit} ↓
dots.mocr	**86.91**	0.115	**0.182**	**90.42**	92.78	**0.186**
Chandra 2	82.67	**0.105**	0.279	86.39	**94.22**	0.205
LightOn-OCR-2	81.81	0.160	0.274	88.80	93.96	0.292
DeepSeek-OCR-2	76.92	0.167	0.316	79.06	83.85	0.243
PaddleOCR-VL-1.5	73.61	0.116	0.437	76.21	82.49	0.228
HunyuanOCR	72.90	0.135	0.439	76.07	84.04	0.263
MinerU 2.5-Pro	60.08	0.538	0.433	77.39	86.26	0.331

Table 2: Overall benchmark results on the ViDocParse test split, sorted by Overall score. **Bold** = best, underlined = second-best. ↑: higher is better; ↓: lower is better. RO^{Edit} denotes Reading-Order Edit Distance.

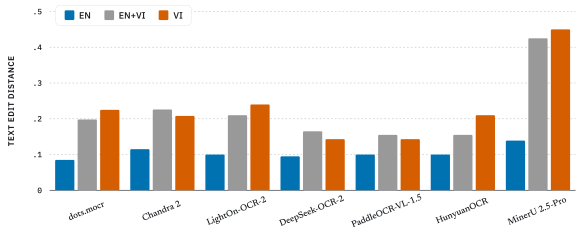


Figure 4: Text Edit Distance (↓) by page language. Every model is at least 1.4× worse on Vietnamese than on English; the EN+VI bucket lies in between.

Model	White	Color	1-col	M-col	WM	Fuzzy
dots.mocr	0.215	**0.206**	0.153	0.294	0.185	0.088
DeepSeek-OCR-2	0.140	0.231	0.126	0.177	0.199	0.074
Chandra 2	0.194	0.373	0.161	0.275	0.330	0.081
HunyuanOCR	0.188	0.220	**0.113**	0.293	**0.172**	**0.050**
PaddleOCR-VL-1.5	**0.134**	0.259	0.133	**0.161**	0.207	0.062
LightOn-OCR-2	0.217	0.345	0.164	0.314	0.292	0.054
MinerU 2.5-Pro	0.430	0.502	0.377	0.514	0.474	0.313

Table 3: Text Edit Distance by page background, layout, and image artifacts. Lower is better. **Bold** indicates the best result, and underlining indicates the second-best result. Color = colorful background; 1-col = single-column layout; M-col = multi-column layout; WM = watermark.

Model	full	less	fewer	wireless
dots.mocr	0.696	0.000	0.530	0.659
DeepSeek-OCR-2	0.661	0.715	0.718	0.700
Chandra 2	**0.814**	**0.864**	0.889	**0.894**
HunyuanOCR	0.609	0.710	0.715	0.710
PaddleOCR-VL-1.5	0.650	0.715	0.658	0.635
LightOn-OCR-2	0.790	**0.864**	**0.900**	0.890
MinerU 2.5-Pro	0.667	0.743	0.740	0.718

Table 4: Approximate Table TEDS-S by table frame type (*full_line*, *less_line*, *fewer_line*, *wireless_line*). (TEDS-S ↑; **bold** = best, underlined = second-best.)

Per-language behaviour. Figure 4 shows the Vietnamese–English gap as the dominant axis of error: every model is at least 1.4× worse on Vietnamese, reaching 2.6× for dots.mocr (0.225 vs. 0.086) and 3.2× for MinerU 2.5-Pro (0.451 vs. 0.139). The EN+VI bucket lies between the two monolingual buckets, suggesting the slowdown is driven by Vietnamese tokens, not code-switching.

Layout, background and artifacts. Table 3 shows that VLMs handle single-column white pages well (Text^{Edit} $\leq$ 0.16) but degrade sharply on multi-column layouts; PaddleOCR-VL-1.5 is the most robust multi-column extractor (0.161), DeepSeek-OCR-2 has the lowest watermark and fuzzy-scan errors, and colourful backgrounds remain hardest for Chandra 2 and LightOn-OCR-2 ($\geq$ 0.34).

Table frame type. Table 4 reports TEDS-S split by frame type. Chandra 2 and LightOn-OCR-2 dominate across all four conditions ($\geq$ 0.79), while VLMs without an explicit table head (dots.mocr, HunyuanOCR, DeepSeek-OCR-2) cluster around 0.66–0.72. The dots.mocr score of 0.000 on *less_line* is a total table-detection failure on this small subset, not a metric zero (§5.3).

5.3 Error Analysis

We isolate three orthogonal failure dimensions: per-block-category accuracy, the share of error attributable to Vietnamese diacritics, and qualitative structural failure modes that recur in manual inspection.

Per-block-category accuracy. Table 5 reports Edit Distance at the granularity of individual block categories, using best-window matching against each model's full-page output. PaddleOCR-VL-1.5 is the strongest text-block extractor (0.038), more than three times better than dots.mocr (0.134). For tables and isolated equations, however, Chandra 2 leads (0.133 and 0.125 respectively), echoing its strong table behaviour on Table 4. References (cita-

Model	text	title	section	ref.	table	equation	caption
dots.mocr	0.134	0.131	0.086	0.055	0.186	0.147	0.147
DeepSeek-OCR-2	0.064	0.095	0.050	0.015	0.136	0.218	0.078
Chandra 2	0.072	**0.062**	**0.031**	0.015	**0.133**	**0.125**	0.055
HunyuanOCR	0.126	0.082	0.065	0.009	0.157	0.253	0.152
PaddleOCR-VL-1.5	**0.038**	0.076	0.039	**0.007**	0.144	0.230	**0.034**
LightOn-OCR-2	0.121	0.073	0.054	0.011	0.134	0.148	0.136
MinerU 2.5-Pro	0.273	0.218	0.227	0.043	0.216	0.314	0.233

Table 5: Edit Distance by block category (best-window match). Columns are abbreviated: *text* = `text_block`, *section* = `section_header`, *ref.* = `reference`, *equation* = `equation_isolated`, *caption* = `figure_caption`. ($\text{Edit}_{\text{dist}} \downarrow$; **bold** = best, underlined = second-best.)

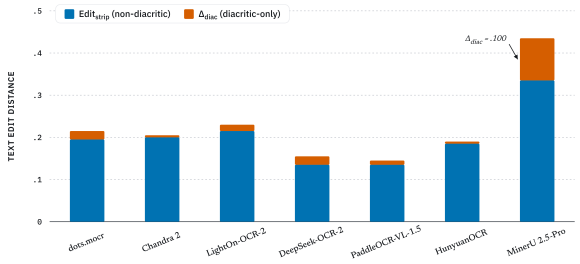


Figure 5: Decomposition of Text Edit Distance into a non-diacritic residual ($\text{Edit}_{\text{strip}}$, blue) and a diacritic-only delta (Δ_{diac}, orange). For modern VLMs the orange tier is barely visible; only MinerU 2.5-Pro loses substantial mass to diacritics.

Model	Col. merge	Tbl. miss	Row merge	Drift
dots.mocr	35.1	22.9	10.9	66.6
DeepSeek-OCR-2	**14.0**	3.0	11.9	**56.8**
Chandra 2	29.6	1.0	**9.6**	71.6
HunyuanOCR	34.7	7.0	10.6	74.8
PaddleOCR-VL-1.5	14.4	1.2	9.9	70.9
LightOn-OCR-2	37.8	**0.6**	**9.6**	75.1
MinerU 2.5-Pro	74.3	1.8	11.4	99.5

Table 6: Frequency (%) of four structural failure modes: multi-column reading-order merging, table omission, table- row merging, and inline-formula boundary drift. (Rate (%) $\downarrow$; **bold** = best, underlined = second-best.)

tion lines) are the easiest category for every VLM (≤ 0.06); MinerU 2.5-Pro is the only system that visibly degrades even there (0.043).

Diacritic-only contribution. Figure 5 decomposes total Text Edit Distance into two stacked components: the residual after stripping Vietnamese tone and vowel diacritics ($\text{Edit}_{\text{strip}}$, blue) and the diacritic-only delta $\Delta_{\text{diac}} = \text{Edit}_{\text{full}} - \text{Edit}_{\text{strip}} \geq 0$ (orange). For every VLM, Δ_{diac} stays in the 0.4–1.9% range, i.e. tone marks contribute very little once the underlying word has been recognised correctly. The picture is dramatically different for MinerU 2.5-Pro ($\Delta_{\text{diac}} = 0.100$), which loses ten edit-distance points to diacritics alone — consistent with its connectionist OCR backbone producing tone-stripped Latin output.

Qualitative failure modes. Table 6 enumerates four structural failure modes that recur under manual inspection: column-order merging on multi-column pages, missing or row-merged tables, and inline-formula boundary drift. DeepSeek-OCR-2 and PaddleOCR-VL-1.5 keep column merging below 15% of multi-column pages, while LightOn-OCR-2 and dots.mocr exceed 35%. Inline-formula boundary drift is the most universal failure: every model exceeds 56%, and MinerU 2.5-Pro effectively never produces a well-bounded inline formula (99.5%). We attribute the universality of this failure to the inline-equation–vs–surrounding-text ambiguity that VLMs only see in small quantities during pre-training.

6 Fine-tuning

6.1 Motivation and Setup

Guided by the benchmark results in §5, we select dots.mocr (3.1B) (Zheng et al., 2026) for domain-specific fine-tuning on ViDocParse. With training splits ranging from ~3,000 to ~10,000 samples, updating all 3.1B parameters risks overfitting; we therefore adopt LoRA (Hu et al., 2022), which provides a parameter-efficient adaptation strategy and has been shown to reduce forgetting during fine-tuning (Biderman et al., 2024). Low-rank adapters ($r = 64$, $\alpha = 128$, dropout 0.05) are injected into the attention projections (`q`, `k`, `v`, `o`) and the MLP projections (`gate`, `up`, `down`) of every transformer block in the language model, yielding $\sim$ 2.4% trainable parameters ($\sim$ 74M of 3.1B). The vision encoder is kept frozen to reduce memory cost and preserve pretrained visual representations.

LoRA is applied to each of the three training splits (`train_generated`, `train_manual`,

Setup	Overall ↑	Text ↓	Formula ↓	TEDS ↑
Zero-shot baseline	86.91	0.115	0.182	90.42
LoRA ($r=64$, $\alpha=128$)				
`train_gen.`	85.07	0.156	0.226	**93.41**
`train_man.`	**88.79**	**0.112**	**0.127**	90.27
`train_mixed`	<u>87.22</u>	<u>0.114</u>	<u>0.180</u>	<u>91.05</u>

Table 7: Fine-tuning results on the ViDocParse test split. Zero-shot baseline is the off-the-shelf dots.mocr checkpoint (row 1 of Table 2). **Bold** = best, <u>underlined</u> = second-best.

`train_mixed`), yielding three fine-tuned checkpoints. All runs use the AdamW optimiser with 3% linear warm-up and cosine decay, learning rate 1×10^{-4}, weight decay 0.01, gradient clipping at 1.0, per-device batch size 2 with 8 gradient-accumulation steps (effective batch size 16), and a maximum sequence length of 16,384 tokens. We train for 3 epochs on a single NVIDIA H100 NVL 94 GB GPU with gradient checkpointing and bfloat16 mixed precision. Full hyperparameters are listed in Appendix E.

6.2 Fine-tuning Results

Table 7 compares each fine-tuned checkpoint against the zero-shot dots.mocr baseline along the three content metrics that constitute the Overall score (Eq. 1).

Effect of training split. The manual-only split (`train_manual`, 3,084 pages) yields the largest Overall gain (+1.88, from 86.91 to 88.79), with improvements on both Text Edit Distance (0.115 → 0.112) and, most notably, Formula Edit Distance (0.182 → 0.127, a 30% relative reduction). The mixed split (`train_mixed`) provides a smaller Overall gain (+0.31) and a modest Table TEDS improvement (90.42 → 91.05), but its formula gain is marginal (0.182 → 0.180). The generated-only split (`train_generated`) produces the strongest table reconstruction (93.41 TEDS, +2.99 over the baseline) but degrades text (0.115 → 0.156) and formula recognition (0.182 → 0.226), pulling its Overall below the zero-shot baseline (85.07, −1.84).

These results suggest that manually annotated pages provide more effective supervision under real-world Vietnamese document conditions. In contrast, generated pages appear particularly useful for learning table structure, as reflected by their higher Table TEDS score, but they may introduce a distribution mismatch for text and formula recognition. The mixed split yields intermediate performance, indicating that simply combining generated and manual data does not necessarily preserve the strengths of both sources.

7 Conclusion

We presented **ViDocParse**, a multi-level annotated dataset and benchmark for Vietnamese document parsing, comprising 11,439 document-page images across 9 categories. In addition to the dataset, we provide a systematic benchmark of seven OCR/VLM systems under identical conditions and three LoRA-adapted dots.mocr checkpoints trained on generated, manual, and mixed training splits.

Our results highlight three main findings. First, end-to-end VLM-based parsers substantially outperform the pipeline-based OCR system, with dots.mocr achieving the best Overall score (86.91) compared with MinerU 2.5-Pro (60.08). Second, no single model dominates across all attribute slices, suggesting that Vietnamese document parsing remains highly dependent on layout, document type, and content structure. Third, while Vietnamese diacritics still affect recognition quality, the major remaining challenges lie in layout-sensitive phenomena such as reading-order errors, table reconstruction, column merging, and inline-formula segmentation.

Beyond OCR accuracy, ViDocParse provides a reusable resource for evaluating document-to-text conversion as a prerequisite for Vietnamese retrieval, question answering, information extraction, and document-grounded generation. We hope this benchmark supports future research on Vietnamese document AI and encourages similar resources for other underrepresented languages.

8 Limitations

Domain and language coverage. ViDocParse's nine categories cover academic, journalistic and educational Vietnamese material but under-represent bureaucratic / legal documents (forms, identity papers, court rulings) and medical records (prescriptions, lab reports); performance on those domains should not be extrapolated from our benchmark. Although the schema and tooling generalise to other tonal, diacritic-heavy languages (Thai, Khmer, Lao, Burmese), the empirical findings target Vietnamese only and should not be read as a stand-in for those languages.

Annotation methodology. The manual pipeline is semi-automatic: a layout detector and dots.ocr (Li et al., 2025) produce initial proposals that humans then correct. Initial-model biases (e.g. rare diacritic combinations, extremely faded scans) may therefore be under-represented in the labels. Span-level annotations for inline equations and footnote marks reuse the bounding box of the parent text line as a coarse approximation, because pixel-precise span-level labelling is prohibitively expensive at this scale.

Statistical and metric scope. All benchmark numbers come from a *single* inference run per model (greedy decoding, fixed prompt) on an NVIDIA H100 NVL 94 GB GPU; rankings could shift under different inference budgets or model-specific tricks (e.g. chain-of-thought, structured-output decoding). Fine-tuned checkpoints use a single random seed, so the gains in §6.2 carry no variance interval. On the metric side, Text Edit Distance does not penalise semantically equivalent rewrites (e.g. paragraph reordering) and Table TEDS inherits known sensitivities to HTML normalisation (Zhong et al., 2020); we mitigate but cannot eliminate these biases through multi-metric reporting.

References

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Paul Greengard, Connor Stolber, Luke Edelman, Leo Gao, Maximilian Franke, Timothy Havens, and Peter Savarese. 2024. LoRA learns less and forgets less. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, and 1 others. 2026. PaddleOCR-VL-1.5: Towards a multi-task 0.9B VLM for robust in-the-wild document parsing. *arXiv preprint arXiv:2601.21957*.

Datalab. 2026. Chandra OCR 2: A 4b-parameter document understanding model. `https://huggingface.co/datalab-to/chandra-ocr-2`. HuggingFace model card.

Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M. Rush. 2017. Image-to-markup generation with coarse-to-fine attention. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 980–989.

Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. 2024. DocPedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences*, 67(12):220106.

Ahmed Heakl, Muhammad Abdullah Sohail, Mukul Ranjan, Rania Elbadry, Ghazi Shazan Ahmad, Mohamed El-Geish, Omar Maher, Zhiqiang Shen, Fahad Shahbaz Khan, and Salman Khan. 2025. Kitab-bench: A comprehensive multi-domain benchmark for arabic ocr and document understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22006–22024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, pages 4083–4091.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, volume 13688 of *Lecture Notes in Computer Science*, pages 498–517. Springer.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Yumeng Li, Guang Yang, Hao Liu, Bowen Wang, and Colin Zhang. 2025. dots.ocr: Multilingual document layout parsing in a single vision-language model. *arXiv preprint arXiv:2512.02498*.

Andrés Marzal and Enrique Vidal. 1993. Computation of normalized edit distance and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):926–932.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, and 1 others. 2025. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24838–24848.

Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 629–633. IEEE.

Said Taghadouini, Adrien Cavaillès, and Baptiste Aubertin. 2026. LightOnOCR: A 1B end-to-end multilingual vision-language model for state-of-the-art OCR. *arXiv preprint arXiv:2601.14251*.

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, and 1 others. 2025. HunyuanOCR technical report. *arXiv preprint arXiv:2511.19575*.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and 1 others. 2024a. MinerU: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2026. DeepSeek-OCR 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*.

Handong Zheng, Yumeng Li, Kaile Zhang, Liang Xin, Guangwei Zhao, Hao Liu, Jiayu Chen, Jie Lou, Qi Fu, Rui Yang, and 1 others. 2026. Multimodal ocr: Parse anything from documents. *arXiv preprint arXiv:2603.13032*.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: Data, model, and evaluation. In *European Conference on Computer Vision (ECCV)*, volume 12366 of *Lecture Notes in Computer Science*, pages 564–580. Springer.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. PubLayNet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.

Appendix

A Dataset Splits and Statistics

Table 8 summarises the four metadata-manifested splits of ViDocParse. The `test` split is fully held out and used for every benchmark number in §5–§6. The three training splits are designed to isolate the contribution of synthetic versus real-world supervision: `train_generated` contains only AI-generated pages rendered from Gemini-produced OmniDocBench JSON layouts at 200 DPI, with pixel-perfect labels, `train_manual` contains only human-annotated scans and photographs, and `train_mixed` is their union.

Split	Pages	Description
`test`	1,108	Held out (all 9 categories)
`train_generated`	7,247	AI-generated docs only
`train_manual`	3,084	Human-annotated docs only
`train_mixed`	10,331	Union of `train_generated` and `train_manual`

Table 8: ViDocParse dataset splits.

B Annotation Schema Details

The annotation taxonomy follows OmniDocBench (Ouyang et al., 2025) with two Vietnamese-specific extensions, namely the *text_language* attribute (with the mixed-language value `text_en_vi_mixed`) and the *text_rotate* attribute (used to flag rotated columns in Vietnamese magazine and newspaper pages).

Block categories (18 total). We define 14 text-bearing categories — `text_block`, `title`, `section_header`, `reference`, `footnote`, `header`, `footer`, `page_number`, `caption_table`, `caption_figure`, `caption_chart`, `caption_equation`, `list_item`, and `code_block` — and 4 non-text categories: `table`, `figure`, `equation_isolated`, and `chart`. Every block carries a polygon bounding box (`poly`), a category-specific content field (`text` for text-bearing blocks, `html` for tables, `latex` for isolated equations), and a set of attribute sub-fields described below.

Page-level attributes. Every page is tagged with six attributes: (i) `data_source` ∈ {`generated`, `manual`}; (ii) `language` ∈ {`vi`, `en`, `en_vi_mixed`}; (iii) `layout` ∈ {`single_column`, `multi_column`}; (iv) `watermark` ∈ {0, 1}; (v) `colorful_background` ∈ {0, 1}; (vi) `fuzzy_scan` ∈ {0, 1}.

Span-level annotations. The optional `line_with_spans` field decomposes a text block into ordered sub-elements of three types: `text_span` (free Markdown text), `equation_inline` (LaTeX content delimited by `$...$`), and `footnote_mark` (raised reference markers). Each span carries its own polygon and content string. For inline equations and footnote marks the polygon is the bounding rectangle of the parent text line, used as a coarse approximation because pixel-precise span-level labelling is prohibitively expensive at the 11,439-page scale; we discuss this limitation in the main Limitations section.

C Quality Control Procedure

Annotations pass through two automated stages followed by two rounds of manual review.

Automated validation. A suite of fix scripts is run repeatedly until convergence on the full corpus. The scripts enforce six properties: (a) the page-level `language` attribute is consistent with the distribution of block-level `text_language` values; (b) the page-level `colorful_background` attribute is consistent with block-level `text_background` values; (c) inline formulas use a single delimiter style (`$...$`); (d) table `html` fields are minified to a single line with no inter-tag whitespace; (e) the table `frame_type` attribute is classified as `full_line`, `less_line`, `fewer_line` or `wireless_line` via rule-based DOM heuristics; (f) the `text_rotate` attribute is verified against Tesseract OSD orientation predictions (Smith, 2007).

Vocabulary check. Every Vietnamese string is segmented into syllables, and each syllable is checked against a curated 74,000-token Vietnamese vocabulary dictionary. Out-of-vocabulary syllables surface a flag for human review and are the most common source of post-hoc tone-mark corrections.

Manual review. To establish a rigorous baseline, we first evaluated inter-annotator agreement on a pilot set of 108 double-covered pages. All annotation and review work was conducted by the authors, who served as annotators without monetary compensation; no external crowdworkers or paid participants were recruited. We measured Cohen's Kappa (κ) between two independent annotators and observed substantial to near-perfect agreement: block category type ($\kappa = 0.726$), text background

($\kappa = 0.791$), text rotation ($\kappa = 0.970$), table frame type ($\kappa = 0.857$), and table layout ($\kappa = 1.000$).

Following this validation, the workload was divided among the annotators. One annotator was responsible for running the AI-generated data pipeline, while another manually annotated the pages based on initial proposals from dots.ocr (Li et al., 2025). The annotators then swapped roles to cross-check each other's work. Any unresolved disagreements were escalated to a senior annotator to ensure high consensus. To quantify the impact of this quality-control pipeline, we compared raw pre-reconciliation checkpoints against the final gold-standard annotations on the stratified held-out test split of 1,108 pages. The reconciliation process introduced 29,729 corrections (avg. 26.83 per page), spanning spelling normalisation (2,354 instances), attribute standardisation (10,412), and relation alignments (16,963), confirming that cross-checking and senior-annotator arbitration yield non-trivial gains over the initial pipeline output.

D Model Prompts and Inference Configuration

Table 9 lists the exact prompts and Python entry points used to invoke each evaluated model, following the official HuggingFace model cards as of December 2025. All VLM systems use greedy decoding (temperature $= 0$, top-$p = 1$) with a maximum generation length of 8,192 tokens; MinerU 2.5-Pro uses its default two-step pipeline (layout analysis $\rightarrow$ per-block recognition).

Model	Prompt / configuration
dots.mocr	`prompt_ocr`: "Convert the document to markdown. No header/footer."
DeepSeek-OCR-2	`<image>\n<|grounding|>Convert the document to markdown.`
Chandra 2	`prompt_type="ocr_layout"` $\rightarrow$ `parse_markdown()`
HunyuanOCR	Full document parsing prompt (formula = LaTeX, table = HTML)
PaddleOCR-VL-1.5	`PaddleOCRVL()` full document parser pipeline
LightOn-OCR-2	Prompt-free; image-only conversation input
MinerU 2.5-Pro	`two_step_extract()`: layout $\rightarrow$ crop recognition

Table 9: Prompts and entry points used for each evaluated model.

E Fine-tuning Setup

Table 10 summarises the hyperparameters used for each of the three fine-tuning strategies applied to dots.mocr (3B). All runs share the AdamW optimiser with linear warm-up over the first 3% of training steps followed by cosine decay, weight decay 0.01, and gradient clipping at 1.0. LoRA adapters ($r = 64$, $\alpha = 128$, dropout 0.05) are injected into the attention projections (`q`, `k`, `v`, `o`) and MLP projections (`gate`, `up`, `down`) of every transformer block in the language model, yielding $\sim$2.4% trainable parameters. Fine-tuning data is read from a shared Parquet/Arrow index that encodes the image path and the OmniDocBench-style JSON path for each sample, allowing all three runs to reuse the same on-disk cache.

Hyperparameter	LoRA
Trainable params (%)	$\sim$2.4
Base precision	bf16
Adapter rank (r)	64
Adapter α	128
Adapter dropout	0.05
Learning rate	$1 \cdot 10^{-4}$
Batch size (per GPU)	2
Grad. accumulation	8
Epochs	3
Max seq. length	16,384
LR warmup	3% linear

Table 10: Fine-tuning hyperparameters for LoRA (Hu et al., 2022) applied to dots.mocr on three training splits. All runs use an NVIDIA H100 NVL 94 GB GPU.

F Evaluation Metric Definitions

For completeness, we restate the four primary metrics used in the benchmark. Let y and $\hat{y}$ denote the ground-truth and predicted token sequences of length n and $\hat{n}$, respectively.

Edit Distance. The normalised Levenshtein edit distance is

$$\mathrm{Edit}(y, \hat{y}) = \frac{\mathrm{Lev}(y, \hat{y})}{\max(n, \hat{n})} \in [0, 1],$$

where Lev counts character-level insertions, deletions and substitutions. $\mathrm{Text}^{\mathrm{Edit}}$, $\mathrm{Formula}^{\mathrm{Edit}}$ and $\mathrm{RO}^{\mathrm{Edit}}$ instantiate this definition on raw text, isolated LaTeX formula strings, and the integer reading-order index sequence, respectively.

TEDS and TEDS-S. Tree-Edit-Distance-based Similarity (Zhong et al., 2020) compares the HTML DOM trees of two tables. TEDS uses the full

tree (structure + cell text); TEDS-S compares only structural nodes and is therefore robust to cell-level OCR errors. Both lie in $[0, 100]$ with higher being better.

Overall. The Overall score collapses Text, Formula and Table into a single 0–100 number using the OmniDocBench v1.6 convention given in Eq. 1.

G Computational Budget and Release Plan

Compute. A single end-to-end benchmark run over the held-out test split takes 2.5–4 GPU-hours per model on an NVIDIA H100 NVL 94 GB GPU; the full seven-model benchmark therefore costs ~ 20 GPU-hours. LoRA fine-tuning costs ~ 8 GPU-hours per training split (measured on the `train_mixed` split), so the three fine-tuned LoRA checkpoints amount to ~ 24 additional GPU-hours. The full experimental envelope of the paper is approximately 50–60 H100 GPU-hours. Disk usage is dominated by the raw dataset (~ 60 GB), which we preprocess down to ~ 22 GB of JPEG cache prior to training; a working installation requires at least 200 GB of disk and 32 GB of host RAM.

Release plan. The dataset, the evaluation orchestrator, the per-block metric scripts, and the three fine-tuned dots.mocr checkpoints (LoRA $\times$ `train_generated`, `train_manual`, `train_mixed`) are available anonymously for review. The evaluation code is hosted at https://anonymous.4open.science/r/ViDocParse-E6B1. The dataset can be accessed at https://huggingface.co/datasets/emnlp2026anon/ViDocParse/ using the read token `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`. All checkpoints, prompts, evaluation seeds, and hardware identifiers are pinned via per-model YAML configs and a per-split Parquet index to guarantee bit-for-bit reproducibility.

Planned extensions. We expect a follow-up release to extend the dataset along two axes identified in the Limitations section: (i) bureaucratic and legal documents, where layout regularity (forms, tables of contents, official seals) creates a qualitatively different benchmark; and (ii) medical records, where domain-specific abbreviations and handwritten annotations push current VLMs further out of distribution.