CL-OCR: Fusing Layout Analysis and Adapting Recognition for Document Parsing in the Wild

Haijie Yuan[1,2]* **Yiming Zhang**[1,2]* **Dekang Ran**[1,2] **Yi Chen**[1,3] **Xinming Wang**[1,3]
Jun Cao[1,2] **Wenzhuo Liu**[1,2]† **Xiao-Hui Li**[1,2] **Fei Zhu**[4] **Fei Yin**[1,2] **Cheng-Lin Liu**[1,2]†
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
[2]School of Artificial Intelligence, UCAS [3]Zhongguancun Academy
[4]Centre for Artificial Intelligence and Robotics, HKISI-CAS
{yuanhaijie2024, zhangyiming2025, liuwenzhuo2020}@ia.ac.cn liucl@nlpr.ia.ac.cn

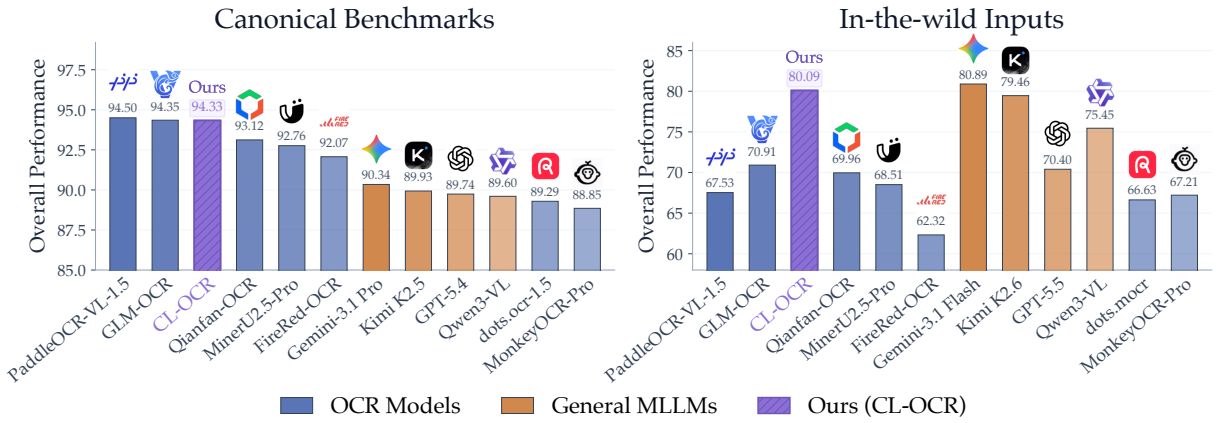


Figure 1: **Performance on canonical benchmarks and challenging in-the-wild inputs.** OCR models perform well on canonical benchmarks but degrade on challenging real-world inputs, where they no longer consistently surpass general MLLMs. CL-OCR improves both settings through layout fusion and recognition adaptation.

Abstract

Recent OCR models achieve strong results on document benchmarks, with several 1–2B OCR-specialized models surpassing much larger general multimodal LLMs (MLLMs). However, complex real-world documents often contain blur, cropping, curved text, handwriting, and unfamiliar structures, which remain challenging for current OCR systems. We trace this gap to a shift in the error source: errors on canonical benchmarks mainly arise from layout analysis, including localization, element classification, and reading order, while errors under these challenging conditions mainly arise from content recognition. To this end, we propose CL-OCR, a two-stage framework that first fuses semantic structure from an MLLM-based detector with geometric candidates from a vision detector to obtain calibrated layouts, then enhances recognition via a shared LoRA adapter gated by a lightweight router. We further construct OCRWildBench and a companion training split to study these conditions. Experiments show that CL-OCR remains competitive on canonical benchmarks and achieves the strongest overall performance among OCR-specialized models on OCRWildBench.

1 Introduction

Optical character recognition (OCR) aims to convert document images into machine-readable text and structured representations. In document intelligence, OCR systems are expected to handle text, tables, formulas, layout regions, and reading order, producing outputs usable by downstream applications. These requirements have motivated recent OCR-specialized models, including MinerU2.5/Pro (Niu et al., 2026; Wang et al., 2026a), MonkeyOCR (Li et al., 2025), Qianfan-OCR (Dong et al., 2026), and PaddleOCR-VL (Cui et al., 2025a). Evaluated on benchmarks such as OmniDocBench (Ouyang et al., 2025), OCRBench v2 (Fu et al., 2025), olmOCR-Bench (Poznanski et al., 2025), and Ocean-OCR Bench (Chen et al., 2025), these models have shown strong recognition,

*Equal contribution.
†Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37416–37437

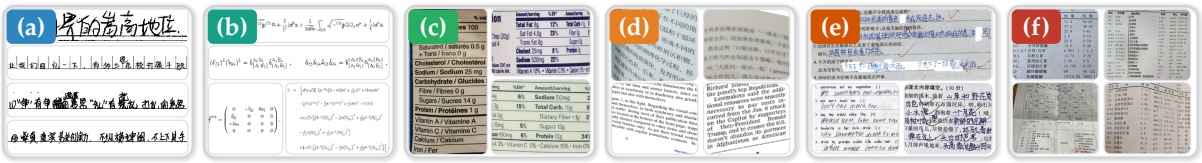


Figure 2: **Representative recognition failures on in-the-wild inputs.** Six challenging scenarios: (a) handwritten text, (b) handwritten formulas, (c) dense product tables, (d) non-canonical captures (blur, skew, curvature), (e) interleaved printed and handwritten text, and (f) medical prescription tables.

layout understanding, structural fidelity, and output consistency. Notably, several 1–2B parameter models have been reported to outperform much larger general MLLMs.

However, strong benchmark performance does not imply that OCR tasks are fully solved in practice. Most benchmarks are built from clean, well-structured documents or augmented samples, whereas OCRWildBench focuses on non-canonical captures, handwriting, formulas, mixed printed–handwritten content, and dense structured tables (Fig. 2). To quantify this gap, we evaluate OCR-specialized models under both canonical benchmarks and challenging in-the-wild settings. Surprisingly, as shown in Fig. 1, these models perform strongly on benchmarks but degrade substantially on in-the-wild inputs, where they no longer consistently outperform general MLLMs. These results indicate that current OCR-specialized models remain insufficiently robust to complex, uncontrolled conditions, and naturally raise the question: ***what systematic errors arise on challenging inputs?***

In this paper, we systematically analyze the errors underlying this benchmark-to-wild gap. Interestingly, we find that the two settings exhibit distinct error profiles, as demonstrated in Fig. 3. **(1)** On canonical benchmarks, errors stem mainly from layout analysis, including inaccurate localization, incorrect element classification, and disrupted reading order. Once the relevant region is correctly identified, recognition results are generally reliable. **(2)** In contrast, in-the-wild inputs result in more recognition failures: Models misread complex formulas and handwritten text; make both content- and structure-level errors in dense structured documents such as ingredient labels and medical prescriptions; and hallucinate or produce garbled outputs when text is truncated, skewed, or curved. We also observe lower recognition performance when handwritten and printed text are interleaved on the same page.

Based on the systematic error analysis, we propose a two-stage framework that improves both layout analysis and content recognition. In the layout stage, we perform ***fused layout analysis***: semantic categories and reading order from an MLLM-based detector serve as a skeleton, which we align with geometric candidates from a vision-only detector to obtain tighter boundaries while preserving the MLLM reading order. In the recognition stage, we introduce ***continual recognition adaptation***. We collect and annotate 11,864 crops, use 8,898 exclusively for adapter training, and reserve 2,966 for held-out evaluation. A lightweight router gates the shared adapter to improve recognition on the challenging inputs represented in OCRWildBench while limiting canonical degradation. Following this continual learning design, we name the framework **CL-OCR**[1]. Our contributions are summarized as follows:

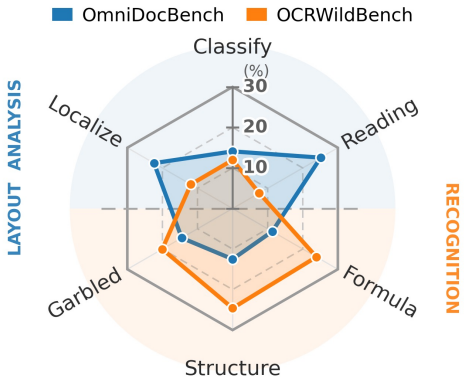


Figure 3: **Dominant failure modes analysis.** Canonical benchmarks mainly expose layout errors, while challenging real-world inputs reveal more recognition failures.

- We systematically analyze OCR models on canonical and in-the-wild settings, showing that errors on benchmarks mainly stem from layout analysis, whereas failures on complex real-world inputs are recognition-related.
- We propose CL-OCR, a two-stage framework composed of fused layout analysis and contin-

[1]Project page: https://github.com/zlysm/CL-OCR

Table 1: **Task-level gap between general MLLMs and OCR models.** We compare standard recognition, wild recognition, and layout-sensitive parsing with the same six representative models. All scores are rounded to two decimals and higher is better. Δ_{best} denotes Best OCR minus Best General.

Task	Metric ↑	General MLLMs				OCR Models				Δ_{best}
		Gemini	GPT	Qwen	Best Gen.	GLM-OCR	MinerU2.5P	Paddle-VL1.5	Best OCR	OCR–Gen.
Recognition	Std. Text	95.30	91.20	95.00	95.30	95.50	96.20	**96.50**	96.50	+1.20
	Std. Formula	88.10	90.74	87.93	90.74	93.65	88.59	**94.21**	94.21	+3.47
	Std. Table	87.64	87.26	85.88	87.64	**93.89**	93.50	92.76	93.89	+6.25
	Wild Text	**82.52**	72.58	77.82	82.52	75.00	71.18	73.58	75.00	-7.52
	Wild Formula	**78.00**	58.20	63.90	78.00	56.90	50.40	65.05	65.05	-12.95
	Wild Table	**79.93**	73.23	77.68	79.93	71.78	73.58	59.73	73.58	-6.35
Layout	Overall AVG	90.34	89.74	89.60	90.34	94.35	92.76	**94.50**	94.50	+4.16
	page@F1	83.47	82.88	83.98	83.98	90.56	90.69	**90.90**	90.90	+6.92
	Order Score	94.40	91.90	94.10	94.40	95.30	**96.00**	95.80	96.00	+1.60

ual recognition adaptation, to improve layout calibration and recognition without compromising canonical performance.

- We build OCRWildBench and a companion training split covering diverse challenging conditions, and show that CL-OCR improves canonical structural parsing while achieving the best OCR-model overall performance on OCRWildBench.

2 Related Work

Pipeline-based document parsing. Traditional document parsing decomposes the task into text detection, recognition, layout analysis, and reading-order recovery. Tesseract (Smith, 2007) is a representative OCR engine integrating page segmentation and text recognition. Deep learning based pipelines implement each stage with a dedicated network, such as EAST (Zhou et al., 2017) and DBNet (Liao et al., 2020) for text detection and CRNN (Shi et al., 2017) for sequence-based recognition. Large-scale layout resources such as PubLayNet (Zhong et al., 2019) and DocLayNet (Pfitzmann et al., 2022) further support document layout analysis across scientific and diverse document domains. Recent works extend to multilingual OCR and structured document conversion, represented by PaddleOCR 3.0 (Cui et al., 2025b) and Docling (Livathinos et al., 2025).

Vision-language document parsing. Another paradigm performs document parsing through image-to-sequence generation. Donut (Kim et al., 2022) directly generates structured outputs from document images, while Nougat (Blecher et al., 2024) targets markup generation for scientific documents. OCR-specialized models extend to full-page recognition, layout-aware parsing, and structured output generation, with representative works including GOT-OCR (Wei et al., 2024), MinerU2.5 (Niu et al., 2026), MonkeyOCR (Li et al., 2025), and PaddleOCR-VL (Cui et al., 2025a). Beyond OCR-specialized systems, general MLLMs, such as Qwen2.5-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025), are increasingly used for document understanding and scientific-agent workflows (Wang et al., 2026c), where evaluation must consider response behavior alongside task correctness (Wang et al., 2026b). Related work on large reasoning models further treats factuality alignment as a distinct alignment objective (Wang et al., 2026d).

OCR benchmarks. Recent OCR evaluation has expanded from text transcription to document parsing. The ICDAR Robust Reading series laid the foundation for standardized text detection and recognition evaluation (Karatzas et al., 2015), while recent benchmarks such as OCRBench v2 (Fu et al., 2025) and OmniDocBench (Ouyang et al., 2025) evaluate OCR systems in broader text-centric and document parsing scenarios (Fu et al., 2025; Ouyang et al., 2025). Specialized benchmarks also expose structure-sensitive failure modes in formulas and tables, including image-to-LaTeX generation, PubTabNet, WTW, and TabRecSet (Deng et al., 2017; Zhong et al., 2020; Long et al., 2021; Yang et al., 2023). Studies on camera-captured and handwritten documents (Ma et al., 2018; Marti and Bunke, 2002; Zhu et al., 2019) show that real-world OCR is strongly affected by non-canonical inputs, such as geometric distortion,

visual degradation, and irregular writing styles.

3 Empirical Analysis and Motivation

To characterize the benchmark-to-wild robustness gap, we construct OCRWildBench, a crop-level diagnostic benchmark with a held-out test split of 2,966 samples covering text, formula, and table recognition under challenging real-world conditions (Fig. 2). A disjoint 8,898-sample training split from the same benchmark conditions is used exclusively for recognition adaptation. Following prior handwriting, formula, and table benchmarks (Marti and Bunke, 2002; Zhu et al., 2019; Deng et al., 2017; Zhong et al., 2020; Long et al., 2021), we compare representative OCR models and general MLLMs on OCRWildBench and canonical benchmarks, yielding two findings in Tab. 1:

- OCR models outperform general MLLMs on canonical benchmarks, but lose this advantage on OCRWildBench. Overall, the strongest baseline OCR model trails the strongest general MLLM by 10.0 points (70.9 vs. 80.9; Tab. 4).
- Errors shift from layout analysis in full-page documents to content recognition in cropped regions, where models still misread characters, formulas, handwriting, and table cells.

These findings suggest that robust real-world OCR requires improving both page-level layout analysis and region-level content recognition, without degrading performance on standard regions. Motivated by the above observation and analysis, we propose the CL-OCR framework.

4 Method: CL-OCR

Overview and Notation. CL-OCR is a two-stage framework for robust document parsing, as illustrated in Fig. 4. Given a document image I, the goal is to produce a structured Markdown document $\hat{Y}$. CL-OCR first performs layout parsing to obtain a calibrated page layout, and then performs region-level recognition on the calibrated regions.

A page layout is represented as an ordered sequence $L = (\ell_1, \ldots, \ell_N)$, where each region $\ell_i = (b_i, c_i, a_i)$ consists of a bounding box b_i, an element category c_i, and a content orientation a_i. We use $\hat{L}$ to denote the calibrated layout parsing result, including refined region boundaries, categories, orientations, and reading order. For each calibrated region $\hat{\ell}_i \in \hat{L}$, CL-OCR predicts a region-level transcription $\hat{y}_i$. The final Markdown output $\hat{Y}$ is obtained by serializing these region-level predictions according to the reading order in $\hat{L}$. The detailed layout calibration and recognition procedures are described in Sec. 4.1 and Sec. 4.2, respectively.

4.1 Fused Layout Analysis

Complementary layout sources. Let M_θ denote the frozen MLLM and V denote the visual detector. CL-OCR uses two complementary layout sources. The MLLM is prompted with a layout-analysis prompt π_{lay}, which instructs it to predict layout regions, semantic categories, orientations, and reading order. It produces an ordered layout proposal

$$L^m = (\ell_1^m, \ldots, \ell_{N_m}^m) = M_\theta(I; \pi_{\text{lay}}), \quad (1)$$

where the superscript m indicates that the proposal is generated by the MLLM, and each region is represented as $\ell_i^m = (b_i^m, c_i^m, a_i^m)$. In parallel, the visual detector produces unordered geometric candidates

$$G = V(I) = \{g_j\}_{j=1}^{N_g}, \quad g_j = (b_j^g, c_j^g, s_j^g), \quad (2)$$

where b_j^g, c_j^g, and s_j^g denote the detector-predicted box, category, and confidence score, respectively. The MLLM proposal provides semantic labels, reading order, and orientation, but its boundaries may be coarse. The visual detector provides sharper localization, but its candidates are unordered and may follow a different label space. To consider both sides of the coin, CL-OCR therefore preserves the MLLM semantic order and uses detector evidence to calibrate geometry.

Compatibility based calibration. The calibrated layout is obtained by $\hat{L} = \mathcal{R}(L^m, G)$, where $\mathcal{R}$ denotes the layout-fusion function and $\hat{L}$ is the final layout parsing result. We first map both sources into a shared category space $\mathcal{C}$ by $\bar{c}_i^m = \psi_m(c_i^m)$ and $\bar{c}_j^g = \psi_g(c_j^g)$. For an MLLM region ℓ_i^m and a detector candidate g_j, their compatibility score is

$$\kappa_{ij} = \lambda_{\text{geo}} \operatorname{IoM}(b_i^m, b_j^g) + \lambda_{\text{cls}} \mathbb{I}[\bar{c}_i^m = \bar{c}_j^g] + \lambda_{\text{conf}} s_j^g,$$

where $\operatorname{IoM}(b, b') = |b \cap b'| / \min(|b|, |b'|)$ tolerates granularity mismatch between coarse regions and fine detector boxes. Each detector candidate is assigned to its most compatible MLLM region if the score exceeds τ_{match}. Let $A_i = \{j : i =$

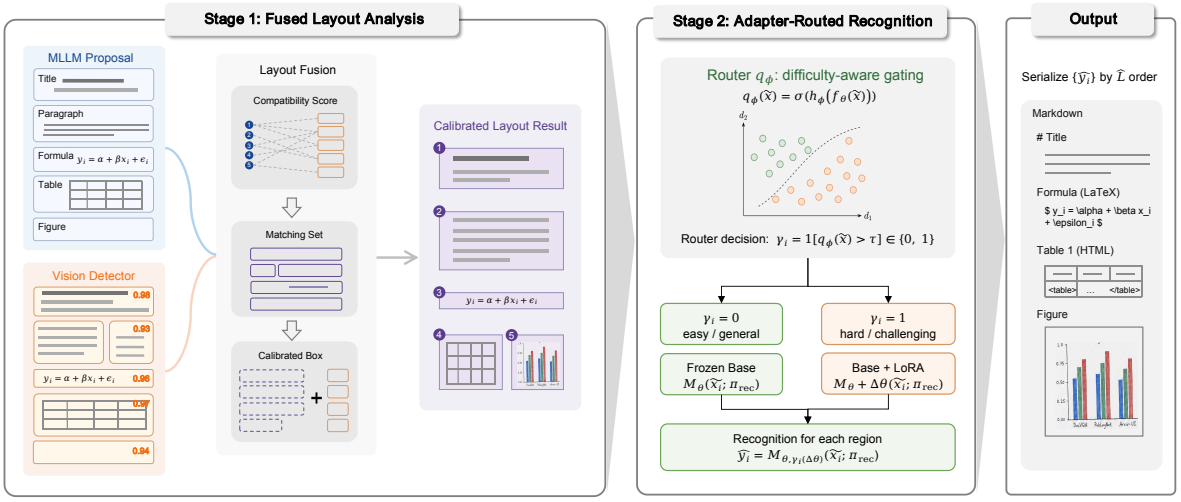


Figure 4: **CL-OCR framework for document parsing in the wild.** CL-OCR constructs a calibrated page layout, selectively adapts recognition for challenging regions, and serializes the recognized outputs into Markdown.

$\arg\max_k \kappa_{kj}, \kappa_{ij} > \tau_{\text{match}}\}$ be the assigned candidates for region ℓ_i^m. The calibrated box is

$$\hat{b}_i = \begin{cases} \text{Union}\{b_j^g : j \in A_i\}, & |A_i| > 0, \\ b_i^m, & |A_i| = 0. \end{cases} \quad (3)$$

The calibrated region is $\hat{\ell}_i = (\hat{b}_i, \bar{c}_i^m, a_i^m)$. Unmatched detector candidates with confidence above τ_{res} are inserted into the nearest local reading position by row-aware geometric sorting. This yields the final calibrated layout $\hat{L} = (\hat{\ell}_1, \ldots, \hat{\ell}_{|\hat{L}|})$, which preserves the MLLM order while improving region boundaries.

4.2 Continual Recognition Adaptation

After layout calibration, CL-OCR recognizes each calibrated region independently. For each region $\hat{\ell}_i \in \hat{L}$, we crop and rectify the corresponding image region $\tilde{x}_i = \mathcal{N}_I(\hat{\ell}_i)$, where $\mathcal{N}_I(\cdot)$ denotes region cropping and orientation normalization. The goal of recognition is to predict the region-level transcription $\hat{y}_i$ from the normalized crop $\tilde{x}_i$.

Low-rank continual learning. To improve recognition on challenging real-world crops without training scenario-specific models, we use a single LoRA adapter shared across all challenging cases. For a LoRA-equipped linear layer with frozen weight W_ℓ, the adapted weight is

$$W'_\ell = W_\ell + \frac{\alpha_\ell}{r} B_\ell A_\ell, \quad (4)$$

where $A_\ell \in \mathbb{R}^{r \times d_{\text{in}}}$, $B_\ell \in \mathbb{R}^{d_{\text{out}} \times r}$, and $r \ll \min(d_{\text{in}}, d_{\text{out}})$. The base MLLM is frozen, and

Algorithm 1 CL-OCR

Input: Document image I; frozen MLLM M_θ; visual detector V; trained adapter $\Delta\theta$ and router q_ϕ
Output: Markdown document $\hat{Y}$
 # Fused layout analysis
$L^m = (\ell_1^m, \ldots, \ell_{N_m}^m) \leftarrow M_\theta(I; \pi_{\text{lay}})$
$G = \{g_j\}_{j=1}^{N_g} \leftarrow V(I)$
$\kappa_{ij} \leftarrow \lambda_{\text{geo}} \text{IoM}(b_i^m, b_j^g) + \lambda_{\text{cls}} \mathbb{I}[\bar{c}_i^m = \bar{c}_j^g] + \lambda_{\text{conf}} s_j^g$
$A_i \leftarrow \{j \mid i = \arg\max_k \kappa_{kj}, \quad \kappa_{ij} > \tau_{\text{match}}\}$
if $|A_i| > 0$ **then**
  $\hat{b}_i \leftarrow \text{Union}\{b_j^g : j \in A_i\}$
else
  $\hat{b}_i \leftarrow b_i^m$
end if
$\hat{L} \leftarrow \{(\hat{b}_i, \bar{c}_i^m, a_i^m)\}_{i=1}^{N_m}$
 # Router-gated recognition
for $\hat{\ell}_i \in \hat{L}$ **do**
  $\tilde{x}_i \leftarrow \mathcal{N}_I(\hat{\ell}_i)$
  $\gamma_i \leftarrow \mathbb{I}[q_\phi(\tilde{x}_i) > \tau]$
  $\hat{y}_i \leftarrow M_{\theta,\gamma_i(\Delta\theta)}(\tilde{x}_i; \pi_{\text{rec}})$
end for
$\hat{Y} \leftarrow \text{Serialize}(\{\hat{y}_i\}, \hat{L})$
return $\hat{Y}$

only the adapter parameters $\Delta\theta = \{A_\ell, B_\ell\}_\ell$ are updated. Given the challenging training set $D_{\text{wild}}^{\text{train}}$, the adapter is optimized by

$$\mathcal{L}_{\text{rec}} = -\sum_{(\tilde{x},y) \in D_{\text{wild}}^{\text{train}}} \log p_{\theta,\Delta\theta}(y \mid \tilde{x}, \pi_{\text{rec}}), \quad (5)$$

where π_{rec} denotes the recognition prompt.

Router-gated activation. The above continual adaptation can effectively improve the recognition performance on in-the-wild inputs. However, an always-on adapter may alter the behavior of the base recognizer on standard documents. CL-OCR

therefore uses a router $q_\phi(\tilde{x}) = \sigma(h_\phi(f_\theta(\tilde{x})))$, where f_θ is a frozen crop feature and h_ϕ is a small MLP. The router is trained with binary cross-entropy loss (BCE), where $z = 1$ for challenging crops and $z = 0$ for standard crops:

$$\mathcal{L}_{\text{route}} = \sum_{(\tilde{x},z)} \text{BCE}\left(q_\phi(\tilde{x}), z\right). \tag{6}$$

At inference time, the router determines whether to activate the shared adapter for each crop: $\gamma_i = \mathbb{I}[q_\phi(\tilde{x}_i) > \tau]$, where $\gamma_i \in \{0, 1\}$ is the router decision. The recognized content of region $\hat{\ell}_i$ is then predicted by

$$\hat{y}_i = M_{\theta,\gamma_i(\Delta\theta)}\left(\tilde{x}_i; \pi_{\text{rec}}\right), \tag{7}$$

where $M_{\theta,\gamma_i(\Delta\theta)}$ denotes the frozen MLLM with the shared adapter conditionally activated. Thus, standard crops follow the frozen base path with $\gamma_i = 0$, while challenging crops activate the adapted path with $\gamma_i = 1$.

4.3 Training and Inference

CL-OCR keeps both the MLLM and the visual detector frozen. The layout stage is a deterministic fusion procedure over the MLLM layout proposal and detector candidates, and introduces no trainable parameters. Training updates only the adapter $\Delta\theta$ and router ϕ with

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mu\mathcal{L}_{\text{route}}. \tag{8}$$

During inference, CL-OCR first applies fused layout analysis to produce the calibrated layout $\hat{L}$. It then crops and normalizes each calibrated region, applies router-gated recognition to obtain $\{\hat{y}_i\}_{i=1}^{|\hat{L}|}$, and serializes the region-level predictions according to the calibrated reading order: $\hat{Y} = \text{Serialize}\left(\{\hat{y}_i\}_{i=1}^{|\hat{L}|}, \hat{L}\right)$.

5 Experiments

Evaluation setup. Full-page parsing is evaluated primarily on OmniDocBench v1.5 (Ouyang et al., 2025), with Ocean-OCR (Chen et al., 2025) and olmOCR-Bench (Poznanski et al., 2025). Recognition is evaluated on OCRWildBench, whose subsets correspond to the challenging conditions in Fig. 2. We compare CL-OCR with recent general MLLMs and OCR models, including the GPT, Gemini, Qwen, MinerU, PaddleOCR-VL, GLM-OCR, MonkeyOCR, Qianfan-OCR, and FireRed-OCR families.

Implementation details. CL-OCR instantiates the MLLM parser with MinerU2.5 Pro (Wang et al., 2026a) and the visual detector with PP-DocLayoutV3. Unless otherwise specified, layout fusion uses $(\lambda_{\text{geo}}, \lambda_{\text{cls}}, \lambda_{\text{conf}}) = (0.80, 0.15, 0.05)$, $\tau_{\text{match}} = 0.7$, and $\tau_{\text{res}} = 0.3$; Table 7 examines their sensitivity. The adapter is a language-model LoRA (Hu et al., 2022) with rank $r = 64$ and scaling factor $\alpha = 128$, while the router is a two-layer MLP. Both modules are trained for 3 epochs on 4 NVIDIA A100 80GB GPUs with AdamW (Loshchilov and Hutter, 2019), a batch size of 64, and gradient clipping at norm 1.0. The learning rate is set to 5×10^{-4} with a 0.03 warmup ratio and cosine decay. Further details are provided in Appendix A.3.

5.1 Main Results

Comparison results on OmniDocBench. Table 2 shows that CL-OCR remains highly competitive in canonical full-page parsing. It achieves an Overall AVG of 94.30, only slightly below the best result, while attaining the best Table TEDS-S score. Its strongest results are concentrated in formula parsing and strict table structure, where boundary accuracy is particularly important. This suggests that layout fusion leverages the visual detector's more accurate geometric boundaries while preserving the MLLM parser's semantic-category and reading-order modeling. On standard documents, CL-OCR benefits mainly from better layout calibration, consistent with the OP-LER ablation results in Section. 5.2, which improve structural parsing without introducing reading-order drift.

Comparison on additional benchmarks. Table 3 further evaluates CL-OCR beyond OmniDocBench. CL-OCR achieves the numerically best Overall score on olmOCR-Bench and performs strongly on structure-heavy subsets, including ArXiv, old scanned math, and tables. These cases depend jointly on accurate region boundaries, formula recognition, and reading order, indicating that layout calibration of CL-OCR improves full-page structural parsing rather than a single local metric. Results on Ocean-OCR show the same trend: CL-OCR obtains lower Edit and higher F1, BLEU, and METEOR on average across Chinese and English documents, showing improved text fidelity and structural consistency beyond OmniDocBench-specific layout metrics.

Table 2: **Document parsing performance comparison on OmniDocBench v1.5.** CL-OCR achieves competitive overall performance while substantially improving strict table structure.

Model Type	Methods	Param	Overall↑	Text^{Edit}↓	Formula^{CDM}↑	Table^{TEDS}↑	Table^{TEDS-S}↑	Read Order^{Edit}↓
General MLLMs	Gemini-3.1-Pro	–	90.34	0.047	88.10	87.64	92.54	0.056
	GPT-5.4	–	89.74	0.088	90.74	87.26	91.87	0.081
	Qwen3-VL	235B	89.60	0.050	87.93	85.88	89.97	0.059
	Ovis2.6-30B-A3B	30B	92.36	0.037	90.32	90.46	94.00	0.046
OCR Models	HunyuanOCR	1B	90.57	0.085	86.01	**94.19**	95.96	0.082
	MinerU2.5	1.2B	90.93	0.045	88.86	88.44	92.42	0.044
	FireRed-OCR	2B	92.07	**0.035**	90.98	88.72	92.38	0.041
	Logics-Parsing-v2	4B	92.56	0.043	91.41	90.54	93.85	0.044
	PaddleOCR-VL	0.9B	92.86	**0.035**	91.22	90.89	94.76	0.043
	MinerU2.5 Pro	1.2B	92.76	0.038	88.59	93.50	96.46	0.040
	Youtu-Parsing	2.5B	93.37	0.042	91.22	93.10	96.47	**0.026**
	GLM-OCR	0.9B	94.35	0.045	93.65	93.89	96.50	0.047
	PaddleOCR-VL-1.5	0.9B	**94.50**	**0.035**	**94.21**	92.76	95.79	0.042
	CL-OCR	1.3B	94.30	**0.035**	92.69	93.70	**97.05**	0.040

Table 3: **Additional document benchmarks. (a)** Selected olmOCR-Bench scenarios. **(b)** Ocean-OCR results averaged over English and Chinese documents.

(a) olmOCR-Bench

Method	ArXiv	Old math	Tables	Old scans	Overall
Mistral OCR	77.2	67.5	60.6	29.3	72.0 ± 1.1
PaddleOCR-VL*	85.7	71.0	84.1	37.8	80.0 ± 1.0
Infinity-Parser*	84.4	83.8	85.0	47.9	82.5 ± 0.8
Chandra OCR*	82.2	80.3	88.0	**50.4**	83.1 ± 0.9
olmOCR	83.0	82.3	84.9	47.7	82.4 ± 1.1
CL-OCR	**88.9**	**84.9**	**89.6**	38.3	**83.4** ± 0.9

(b) Ocean-OCR document subset

Method	Edit↓	F1↑	BLEU↑	METEOR↑
Mathpix	0.144	0.925	0.747	0.846
Qwen2.5-VL	0.117	0.932	0.839	0.900
Ocean-OCR	0.060	0.950	0.909	0.931
MinerU2.5	0.058	0.955	0.863	0.919
CL-OCR	**0.030**	**0.960**	**0.915**	**0.957**

Comparison results on OCRWildBench. Table 4 compares region-level recognition on OCRWildBench. Since inputs are cropped and normalized, layout and reading-order errors are largely excluded. CL-OCR achieves the best Overall score among OCR models and approaches the strongest general MLLM, showing that recognition adaptation improves robustness under the benchmark conditions. The largest improvements appear on dense wild tables and formula serialization, with additional gains on non-canonical text; handwriting remains only partially improved and still lags behind the strongest general MLLMs. CL-OCR is not best on all subsets, especially mixed exams and professional tables, which remain challenging.

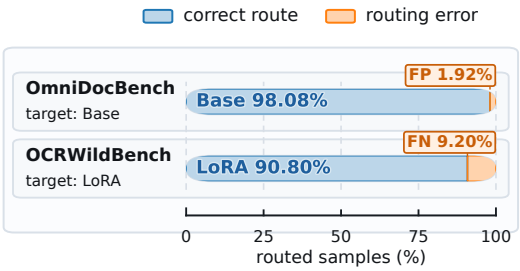


Figure 5: **Routing accuracy.** The router assigns pages to the base recognizer and wild crops to the adapter.

These results indicate that the dominant bottleneck shifts from page layout to region-level recognition under such conditions, making selective adaptation critical.

5.2 Ablation Analysis

Effectiveness of router-gated adaptation. Tables 6 and 5 show that recognition adaptation must be conditionally activated. Always-on LoRA (A-LoRA) yields the largest OCRWildBench gain but drops canonical Overall AVG from 94.33 to 93.24, as wild-tuned parameters shift the output distribution on standard regions. Router-gated R-LoRA cuts this drop to 0.03 points while retaining 95.7% of the wild-recognition gain, confirming that LoRA works best as a conditional expert rather than a global replacement.

Effectiveness of layout fusion. Table 6 compares different uses of detector evidence. Replacing the MLLM layout with the detector improves localization-sensitive metrics (Formula CDM 92.13) but degrades Text and Read Order Edit (0.059, 0.079), as a vision-only detector cap-

Table 4: **Recognition performance comparison on OCRWildBench.** Scores are grouped by subset for text, formula, and table recognition. The best and second-best results are highlighted within each model group separately.

Method	NSDS		SCUT-EPT		LatexOCR		Mixed Exam		WTW		Prof. Tables		Overall↑
	cNED	B4	cNED	B4	CDM	B4	cNED	B4	TEDS	Cell	TEDS	Cell	
General MLLMs													
Gemini-3.1-Flash	**89.3**	**85.9**	71.5	56.1	**87.8**	**68.2**	97.1	**95.2**	**68.2**	**87.8**	**75.3**	**88.4**	**80.9**
GPT-5.5	81.4	76.0	59.7	37.9	72.7	43.7	92.1	88.4	60.5	82.1	68.9	81.4	70.4
GLM-5V-Turbo	80.2	76.5	61.9	41.2	79.8	47.6	93.2	88.0	62.0	78.3	74.5	86.2	72.5
Kimi K2.6	89.0	84.5	**82.9**	66.4	86.2	47.2	**97.7**	93.7	64.1	82.2	73.1	86.7	79.5
Qwen3-VL	85.6	81.3	81.6	**69.3**	82.2	45.6	78.0	71.1	62.7	86.8	74.1	87.1	75.5
OCR Models													
dots.mocr	77.0	73.3	69.4	50.3	78.4	48.2	90.4	83.0	40.8	55.6	60.9	72.3	66.6
FireRed-OCR	70.2	63.2	54.9	42.2	70.5	43.7	85.5	75.5	44.2	51.7	66.1	80.2	62.3
GLM-OCR	81.7	75.8	65.5	48.2	72.9	40.9	91.3	87.5	54.8	70.0	**75.2**	**87.1**	70.9
MinerU2.5 Pro	71.4	66.8	66.9	46.1	63.1	37.7	91.1	84.8	63.4	78.4	71.2	81.3	68.5
MonkeyOCR-Pro	82.0	76.2	66.9	48.8	67.7	31.9	83.7	77.9	55.0	71.5	65.5	79.5	67.2
PaddleOCR-VL-1.5	76.7	70.3	73.0	**56.4**	82.3	47.8	84.0	81.1	40.6	74.5	57.2	66.6	67.5
Qianfan-OCR	81.1	76.2	49.3	37.0	80.3	48.3	**94.3**	**91.6**	55.3	67.7	72.5	85.9	70.0
CL-OCR	**88.8**	**84.8**	**74.1**	55.5	**84.1**	**62.6**	92.6	87.9	**82.2**	**89.4**	74.7	84.4	**80.1**

Table 5: **Recognition routing trade-off.** The table separates canonical-page degradation from retained wild-crop improvement.

Methods	Router	Canonical page		Wild crop		
		OmniDoc AVG ↑	Drop ↓	Overall ↑	Gain ↑	Gain retained ↑
Base	–	**94.33**	–	68.51	–	–
A-LoRA	×	93.24	1.09	**80.61**	**+12.10**	**100.0%**
R-LoRA	✓	94.30	**0.03**	80.09	+11.58	95.7%

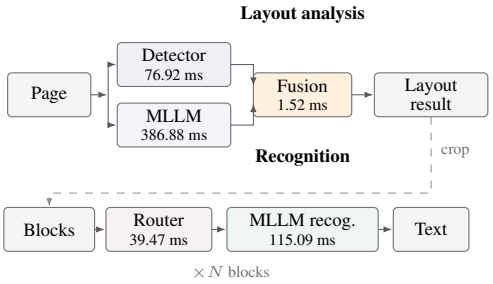


Figure 6: **Latency timeline of CL-OCR.** Layout analysis runs in parallel before routed block-level recognition.

tures local boxes but not semantic categories or page-level order. OP-LER instead keeps the semantic and ordering predictions of MLLM and uses the detector only to calibrate boundaries, raising Overall AVG from 92.76 to 94.33 and Formula CDM by 4.15 points while leaving Read Order Edit unchanged. The gain of layout fusion thus comes from correcting boundary errors, rather than from reconstructing page structure with the detector.

Sensitivity to fusion parameters. Table 7 shows that the fusion procedure is stable over practical settings. Across the tested thresholds, Read Order Edit ranges from 0.037 to 0.039 and Overall from 94.337 to 94.388, with a maximum absolute deviation of 0.051 from the default Overall. Weight configurations that retain substantial geometry show similarly stable Read Order Edit (0.037–0.039) and Overall (94.337–94.513). In contrast, small or zero geometry weights increase Read Order Edit to 0.070–0.324 even when Overall remains high, indicating that Overall alone is insufficient for parameter selection. We therefore retain the default configuration.

Routing accuracy and cost. Figures 5 and 6 explain why R-LoRA preserves canonical stability with minimal overhead. The router sends 98.08% of OmniDocBench regions to the base recognizer and 90.80% of OCRWildBench samples to the LoRA path, avoiding over-correction while still adapting hard inputs. The added cost is limited, layout fusion takes only 1.52 ms, the detector runs in parallel with the MLLM layout call, and the router is faster than a single region-level recognition step. Overhead of CL-OCR thus stems from the lightweight router rather than layout fusion.

Table 6: **Component ablation on OmniDocBench v1.5.** Each row isolates one component: layout fusion (OP-LER) or recognition adaptation (A-/R-LoRA), with oracle GT layout as the upper bound.

Variant	Layout construction			Recognition	Overall	Text	Formula	Table		Read Order
	MLLM proposal	Detector candidate	Order preserved		AVG ↑	Edit ↓	CDM ↑	TEDS ↑	TEDS-S ↑	Edit ↓
Oracle GT layout†	–	–	–	Base MLLM	95.39	0.020	93.73	94.45	97.59	0.005
Baseline	✓	–	✓	Base MLLM	92.76	0.038	88.59	93.50	96.46	**0.040**
Detector	–	✓	–	Base MLLM	93.14	0.059	92.13	93.22	96.60	0.079
OP-LER only	✓	✓	✓	Base MLLM	**94.33**	**0.035**	**92.74**	**93.75**	**97.12**	**0.040**
OP-LER + A-LoRA	✓	✓	✓	A-LoRA	93.24	0.038	90.37	93.15	96.35	**0.040**
CL-OCR	✓	✓	✓	R-LoRA	94.30	**0.035**	92.69	93.70	97.05	**0.040**
Δ OP-LER vs. baseline (pos. means better)					+1.57	+0.003	+4.15	+0.25	+0.66	0.000
Δ A-LoRA vs. OP-LER (neg. means worse)					-1.09	-0.003	-2.37	-0.60	-0.77	0.000

Table 7: **Layout-fusion sensitivity on OmniDocBench v1.5.** Parameters are varied without retraining.

(a) Thresholds

Setting	τ_{match}	τ_{res}	RO Edit ↓	Overall ↑
Default	**0.7**	**0.3**	**0.038**	**94.337**
Match low	0.6	0.3	0.039	94.354
Match high	0.8	0.3	0.038	94.388
Residual low	0.7	0.0	0.037	94.383
Residual high	0.7	0.6	0.038	94.365

(b) Compatibility weights

Setting	λ_g	λ_c	λ_s	Mix	RO Edit ↓	Overall ↑
Default	**0.80**	**0.15**	**0.05**		**0.038**	**94.337**
Class-heavy	0.65	0.30	0.05		0.039	94.474
Conf.-heavy	0.65	0.10	0.25		0.037	94.410
Geo-only	1.00	0.00	0.00		0.037	94.367
Class-only	0.00	1.00	0.00		0.315	94.410
Conf.-only	0.00	0.00	1.00		0.324	94.870
Equal weights	0.33	0.33	0.33		0.039	94.513
Low-geo	0.05	0.48	0.48		0.070	94.038
No-geo	0.00	0.50	0.50		0.316	94.446

Note. Panel (a) fixes weights and panel (b) thresholds at default. The bars encode $g/c/s$ left-to-right (blue/purple/gold), with gray denoting the remainder to 1; values are rounded to two decimals. Overall excludes RO Edit; the maximum region-count change is 0.45%.

6 Conclusion

This paper presents CL-OCR, a two-stage framework for document parsing in the wild. We find that canonical benchmarks primarily expose layout errors, whereas challenging real-world inputs more often expose recognition failures. CL-OCR therefore fuses semantic order from an MLLM with geometric boundaries from a vision detector and selectively activates a recognition adapter. Experiments demonstrate competitive canonical parsing and strong recognition performance on OCRWildBench. Together, these results show that selective adaptation can improve recognition under challenging conditions while largely preserving canonical performance. Future work will examine broader generalization and multi-stage continual learning.

Limitations

CL-OCR improves robustness by targeting failure modes that are common but not exhaustive. Its gains depend on the coverage of OCRWildBench and on the reliability of the upstream layout proposal, detector candidates, and router decisions; when a document falls outside these assumptions, such as heavily degraded scans, rare scripts, unfamiliar handwriting styles, or domain-specific table conventions, errors may still propagate through the pipeline. More broadly, our results show that selective layout calibration and recognition adaptation can close part of the benchmark-to-wild gap, but they should not be interpreted as evidence of fully open-domain OCR generalization.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant U23B2029.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Lukas Blecher. 2025. Latex-ocr: Optical character recognition for latex formulas. https://github.com/lukas-blecher/LaTeX-OCR. GitHub repository.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2024. Nougat: Neural optical understanding for academic documents. In *International Conference on Learning Representations*.

Song Chen, Xinyu Guo, Yadong Li, Tao Zhang, Mingan Lin, Dongdong Kuang, Youwei Zhang, Lingfeng Ming, Fengyu Zhang, Yuran Wang, Jianhua Xu, Zenan Zhou, and Weipeng Chen. 2025. Ocean-ocr: Towards general ocr application via a vision-language model. *Preprint*, arXiv:2501.15558.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. 2023. M^{6}Doc: 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 (CVPR)*, pages 15138–15147.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025a. PaddleOCR-VL: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *Preprint*, arXiv:2510.14528.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025b. PaddleOCR 3.0 technical report. *Preprint*, arXiv:2507.05595.

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*, volume 70 of *Proceedings of Machine Learning Research*, pages 980–989. PMLR.

Daxiang Dong, Mingming Zheng, Dong Xu, Chunhua Luo, Bairong Zhuang, Yuxuan Li, Ruoyun He, Haoran Wang, Wenyu Zhang, Wenbo Wang, Yicheng Wang, Xue Xiong, Ayong Zheng, Xiaoying Zuo, Ziwei Ou, Jingnan Gu, Quanhao Guo, Jianmin Wu, Dawei Yin, and Dou Shen. 2026. Qianfan-OCR: A unified end-to-end model for document intelligence. *Preprint*, arXiv:2603.13398.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2025. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *Advances in Neural Information Processing Systems 38*, pages 107930–107981.

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*.

Liufeng Huang, Bangdong Chen, Chongyu Liu, Dezhi Peng, Weiying Zhou, Yaqiang Wu, Hui Li, Hao Ni, and Lianwen Jin. 2023. Ensexam: A dataset for handwritten text erasure on examination papers. In *Document Analysis and Recognition – ICDAR 2023*, pages 470–485. Springer.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. 2015. ICDAR 2015 competition on robust reading. In *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR 2015)*, pages 1156–1160, Nancy, France.

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 *Computer Vision – ECCV 2022*, pages 498–517. Springer.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Biao Yang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. 2025. MonkeyOCR: Document parsing with a structure-recognition-relation triplet paradigm. *Preprint*, arXiv:2506.05218.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11474–11481.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valéry Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2025. Docling: An efficient open-source toolkit for ai-driven document conversion. *CoRR*, abs/2501.17887.

Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. 2021. Parsing table structures in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 924–932.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. 2018. Docunet: Document image unwarping via a stacked u-net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4709.

Urs-Viktor Marti and Horst Bunke. 2002. The iam-database: An english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Zirui Tang, Boyu Niu, and 40 others. 2026. MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 13–42, San Diego, California, USA. Association for Computational Linguistics.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2025. OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24838–24848.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter W. J. Staar. 2022. DocLayNet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3743–3751.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. olmOCR: Unlocking trillions of tokens in PDFs with vision language models. Poster presented at the CODEML: Championing Open-source DEvelopment in Machine Learning Workshop at ICML 2025.

Baoguang Shi, Xiang Bai, and Cong Yao. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633.

Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, Bangrui Xu, Junbo Niu, Mengzhang Cai, Jiantao Qiu, Qintong Zhang, Dongsheng Ma, Yuefeng Sun, Hejun Dong, Wenzheng Zhang, and 24 others. 2026a. Mineru2.5-pro: Pushing the limits of data-centric document parsing at scale. *Preprint*, arXiv:2604.04771.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Xinming Wang, Weinong Wang, Hongming Yang, Yansong Lin, Zheng Ruan, Shangpin Peng, Qiming Peng, Nan Qiao, Fengyuan Lu, Guoqing Ma, Marito Li, Songyang Zhang, Saiyong Yang, Han Hu, Yonglong Tian, and Xu-Yao Zhang. 2026b. Beyond correctness: Benchmarking and aligning response behaviors in hybrid-thinking mllms. *arXiv preprint arXiv:2608.12781*.

Xinming Wang, Jian Xu, Sheng Lian, Yi Chen, Haiyang Guo, Fei Zhu, Yuanqi Shao, Minsi Ren, Hongzhu Yi, Aslan H. Feng, Hongming Yang, Tailin Wu, Han Hu, Shiming Xiang, Xu-Yao Zhang, and Cheng-Lin Liu. 2026c. The hitchhiker's guide to autonomous research: A survey of scientific agents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20. Early Access.

Xinming Wang, Jian Xu, Bin Yu, Sheng Lian, Yi Chen, Boran Wang, Yingjian Zhu, Hongzhu Yi, Hong-Ming Yang, Han Hu, Cheng-Lin Liu, and Xu-Yao Zhang. 2026d. MR-ALIGN: Meta-reasoning informed factuality alignment for large reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 4191–4214, San Diego, California, United States. Association for Computational Linguistics.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. 2024. General OCR theory: Towards OCR-2.0 via a unified end-to-end model. *Preprint*, arXiv:2409.01704.

Fan Yang, Lei Hu, Xinwu Liu, Shuangping Huang, and Zhenghui Gu. 2023. A large-scale dataset for end-to-end table recognition in the wild. *Scientific Data*, 10(1):110.

Hesuo Zhang, Lingyu Liang, and Lianwen Jin. 2020. Scut-hccdoc: A new benchmark dataset of handwritten chinese text in unconstrained camera-captured documents. *Pattern Recognition*, 108:107559.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: Data, model, and evaluation. In *Computer Vision – ECCV 2020*, pages 564–580. Springer.

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

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. EAST: An efficient and accurate scene text detector.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2642–2651.

Yuanzhi Zhu, Zecheng Xie, Lianwen Jin, Xiaoxue Chen, Yaoxiong Huang, and Ming Zhang. 2019. SCUT-EPT: New dataset and benchmark for offline chinese text recognition in examination paper. *IEEE Access*, 7:370–382.

Table 8: **Full results on olmOCR-Bench.** CL-OCR improves the overall score and performs best on structure-sensitive scenarios including ArXiv, old scanned math, and tables.

Method	Scenario score ↑								Overall ↑
	ArXiv	Old scans math	Tables	Old scans	Headers & footers	Multi column	Long tiny text	Base	
Mistral OCR API	77.2	67.5	60.6	29.3	93.6	71.3	77.1	99.4	72.0±1.1
Marker 1.10.1	83.8	66.8	72.9	33.5	86.6	80.0	85.7	99.3	76.1±1.1
MinerU 2.5.4*	76.6	54.6	84.9	33.7	96.6	78.2	83.5	93.7	75.2±1.1
DeepSeek-OCR	77.2	73.6	80.2	33.3	96.1	66.4	79.4	99.8	75.7±1.0
Nanonets-OCR2-3B	75.4	46.1	86.8	40.9	32.1	81.9	**93.0**	99.6	69.5±1.1
PaddleOCR-VL*	85.7	71.0	84.1	37.8	**97.0**	79.9	85.7	98.5	80.0±1.0
Infinity-Parser 7B*	84.4	83.8	85.0	47.9	88.7	**84.2**	86.4	99.8	82.5±0.8
Chandra OCR 0.1.0*	82.2	80.3	88.0	**50.4**	90.8	81.2	92.3	**99.9**	83.1±0.9
olmOCR v0.4.0	83.0	82.3	84.9	47.7	96.1	83.7	81.9	99.7	82.4±1.1
CL-OCR	**88.9**	**84.9**	**89.6**	38.3	96.4	82.2	87.1	99.7	**83.4±0.9**

A Additional Results and Experimental Details

A.1 Additional Benchmark Results

Comparison results on olmOCR-Bench. Table 8 reports the full olmOCR-Bench results. CL-OCR achieves the best Overall score and ranks first on ArXiv, old scanned math, and tables. These results show that layout fusion improves structure-sensitive full-page parsing across different document scenarios. CL-OCR is weaker on old scans, suggesting that heavily degraded pages still require stronger degradation-aware recognition.

Comparison results on Ocean-OCR. Table 9 reports English and Chinese results separately. CL-OCR obtains the lowest Edit distance in both languages and the best F1, BLEU, and METEOR in most settings. The gain is especially clear on Chinese documents, where Edit distance is reduced from 0.062 to 0.038. These results show that CL-OCR improves text fidelity beyond OmniDocBench-specific layout metrics.

A.2 Evaluation Metrics

We use task-specific metrics for layout, text, formula, and table outputs. Unless otherwise stated, higher scores indicate better performance. For OmniDocBench, Overall is the arithmetic mean of $(1 - \text{Text Edit}) \times 100$, Formula CDM, and Table TEDS. Read Order Edit is reported separately and is not included in Overall.

Layout metrics. For category c, let $G_c(\mathbf{x})$ and $P_c(\mathbf{x})$ denote the rasterized ground-truth and predicted coverage maps, and let $M(\mathbf{x})$ be the foreground mask. The evaluated foreground support is

$$\Omega_c = \big\{\mathbf{x} \mid \big(G_c(\mathbf{x}) > 0 \vee P_c(\mathbf{x}) > 0\big) \wedge M(\mathbf{x}) = 1\big\}. \tag{9}$$

The matched, predicted, and ground-truth masses are

$$\begin{aligned} I_c &= \sum_{\mathbf{x}\in\Omega_c} \min(G_c(\mathbf{x}), P_c(\mathbf{x}))\,, \\ A_c^P &= \sum_{\mathbf{x}\in\Omega_c} P_c(\mathbf{x}), \\ A_c^G &= \sum_{\mathbf{x}\in\Omega_c} G_c(\mathbf{x}). \end{aligned} \tag{10}$$

The page-level precision, recall, and F1 are

$$\begin{aligned} \text{page@P}_c &= \frac{I_c}{A_c^P}, \\ \text{page@R}_c &= \frac{I_c}{A_c^G}, \\ \text{page@F1}_c &= \frac{2\,\text{page@P}_c\,\text{page@R}_c}{\text{page@P}_c + \text{page@R}_c}. \end{aligned} \tag{11}$$

BLEU-4. Given the normalized reference sequence r and prediction sequence h, BLEU-4 is computed as

$$\text{BLEU}_4 = \text{BP} \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log p_n\right), \tag{12}$$

where p_n is the modified n-gram precision and BP is the brevity penalty. Sentence-level smoothing is applied.

Table 9: **Bilingual document results on Ocean-OCR.** English and Chinese results are reported separately to evaluate cross-language text fidelity.

Model	Edit ↓		F1 ↑		Precision ↑		Recall ↑		BLEU ↑		METEOR ↑	
	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
Document OCR systems												
Mathpix	0.064	0.223	0.930	0.919	0.950	0.952	0.911	0.889	0.901	0.593	0.924	0.768
PP-StructureV3	0.068	0.210	0.871	0.929	0.856	0.924	0.892	0.935	0.796	0.570	0.902	0.802
MinerU2-pipeline	0.099	0.225	0.663	0.919	0.635	0.908	0.703	0.934	0.504	0.571	0.670	0.810
PaddleOCR	0.323	0.649	0.707	0.864	0.690	0.912	0.730	0.821	0.517	0.537	0.674	0.699
General MLLMs												
Gemini-2.5 Pro	0.080	0.204	0.922	0.927	0.940	0.959	0.906	0.898	0.877	0.690	0.921	0.862
GPT-4o	0.085	0.450	0.919	0.686	0.929	0.694	0.910	0.703	0.870	0.354	0.922	0.495
Qwen2.5-VL-72B	0.093	0.140	0.923	0.940	0.936	0.956	0.912	0.926	0.879	0.798	0.924	0.876
InternVL3-76B	0.125	0.282	0.828	0.871	0.842	0.889	0.817	0.856	0.728	0.527	0.829	0.759
Qwen2-VL-7B	0.165	0.270	0.849	0.883	0.834	0.847	0.873	0.942	0.795	0.578	0.859	0.763
MiniCPM-V2.6-8B	0.244	0.437	0.804	0.778	0.793	0.721	0.837	0.875	0.695	0.431	0.640	0.642
Specialized OCR/VLMs												
MinerU2-VLM	0.048	0.182	0.936	0.941	0.926	0.927	0.947	0.958	0.893	0.611	0.950	0.837
Ocean-OCR	0.057	0.062	0.937	0.962	0.932	0.956	**0.956**	**0.974**	0.906	**0.912**	0.945	0.916
MonkeyOCR Pro	0.064	0.190	0.929	0.934	0.918	0.925	0.944	0.948	0.884	0.699	0.941	0.850
SmolDocling	0.080	0.878	0.899	0.157	0.895	0.140	0.912	0.268	0.839	0.048	0.907	0.151
dots.ocr	0.083	0.179	0.904	0.931	0.920	0.951	0.890	0.913	0.849	0.639	0.911	0.842
GOT	0.084	0.117	0.895	0.928	0.891	0.934	0.906	0.929	0.835	0.805	0.874	0.848
MinerU2.5	0.033	0.082	0.945	0.965	0.948	**0.966**	0.942	0.964	0.909	0.817	0.950	0.887
CL-OCR	**0.021**	**0.038**	**0.951**	**0.968**	**0.951**	**0.966**	0.955	0.970	**0.918**	**0.912**	**0.963**	**0.950**

charNED. Let $\phi(\cdot)$ denote text normalization, including Unicode NFKC normalization and filtering to Chinese characters, English letters, digits, and common punctuation. Character-level normalized edit-distance similarity is

$$\mathrm{charNED}(y,\hat{y}) = 1 - \frac{\mathrm{ED}\big(\phi(y),\phi(\hat{y})\big)}{\max\big(1,|\phi(y)|,|\phi(\hat{y})|\big)}, \tag{13}$$

where $\mathrm{ED}(\cdot,\cdot)$ is the Levenshtein edit distance.

CDM. For formula recognition, the reference and predicted LaTeX strings are normalized and rendered into images. Let $\mathcal{G}$ and $\mathcal{P}$ denote the rendered character sets of the reference and prediction, and let $\mathcal{M}$ denote the matched character pairs. CDM is defined as

$$\begin{aligned} P_{\mathrm{CDM}} &= \frac{|\mathcal{M}|}{|\mathcal{P}|}, \\ R_{\mathrm{CDM}} &= \frac{|\mathcal{M}|}{|\mathcal{G}|}, \\ \mathrm{CDM} &= \frac{2P_{\mathrm{CDM}}R_{\mathrm{CDM}}}{P_{\mathrm{CDM}}+R_{\mathrm{CDM}}}. \end{aligned} \tag{14}$$

Predictions that cannot be rendered are assigned a score of 0, and samples with unrenderable references are excluded.

TEDS. For table recognition, the prediction and reference are normalized as HTML and parsed into ordered trees $T_{\hat{y}}$ and T_y. Tree-Edit-Distance-based Similarity is

$$\mathrm{TEDS}(y,\hat{y}) = 1 - \frac{\mathrm{TED}(T_y,T_{\hat{y}})}{\max\big(1,|T_y|,|T_{\hat{y}}|\big)}, \tag{15}$$

where $\mathrm{TED}(\cdot,\cdot)$ denotes tree edit distance.

CellNED. CellNED evaluates table cell text independently of the HTML tree. After table normalization, cell texts are extracted in row-major order:

$$C_y = (c_1,\ldots,c_m), \qquad C_{\hat{y}} = (\hat{c}_1,\ldots,\hat{c}_n). \tag{16}$$

The normalized concatenated cell strings are

$$\begin{aligned} \bar{C}_y &= \phi(c_1\|\cdots\|c_m), \\ \bar{C}_{\hat{y}} &= \phi(\hat{c}_1\|\cdots\|\hat{c}_n). \end{aligned} \tag{17}$$

CellNED is defined as

$$\mathrm{CellNED}(y,\hat{y}) = 1 - \frac{\mathrm{ED}(\bar{C}_y,\bar{C}_{\hat{y}})}{\max\big(1,|\bar{C}_y|,|\bar{C}_{\hat{y}}|\big)}. \tag{18}$$

A.3 Implementation Details

CL-OCR keeps the base MLLM parser and the visual detector frozen during both training and inference. Only the shared LoRA adapter and the

Table 10: **Training hyperparameters of CL-OCR.**

Hyperparameter	Adapter	Router
batch size	64	
learning rate	5×10^{-4}	
lr schedule	cosine decay	
lr warmup ratio	0.03	
weight decay	0.05	
gradient clipping	1.0	
epoch	3	
optimizer	AdamW	
DeepSpeed stage	zero2	
LoRA r	64	–
LoRA α	128	–

Table 11: **OCRWildBench split statistics.** OCRWildBench contains 8,898 training samples and 2,966 test samples across six subsets.

Subset	Main challenge	Train	Test	Total
WTW	Wild and dense table recognition	1,344	448	1,792
NSDS	Non-canonical scene/document text	1,629	543	2,172
Mixed Exam	Interleaved printed and handwritten content	1,386	462	1,848
Prof. Tables	Professional tables and structured forms	1,560	520	2,080
LatexOCR	Formula and LaTeX expression recognition	1,500	500	2,000
SCUT-EPT	Handwritten text recognition	1,479	493	1,972
Overall	–	**8,898**	**2,966**	**11,864**

lightweight router are optimized on the training split. The adapter is used for recognition adaptation, while the router decides whether each crop should follow the base recognizer or the adapted path. Table 10 reports the training hyperparameters used for both modules.

B Details of OCRWildBench

Dataset overview. OCRWildBench evaluates crop-level recognition after removing page-level layout and reading-order factors. The held-out test split contains 2,966 orientation-normalized crops from six subsets. We also construct an 8,898-crop training split using a 3:1 train/test ratio within each subset, which is used exclusively for recognition adaptation.

Subset construction. The six subsets instantiate the failure modes analyzed in the main paper. NSDS contains degraded or irregular text captures; SCUT-EPT contains edited handwriting; LatexOCR contains formula crops; Mixed Exam contains interleaved printed and handwritten content; WTW and Prof. Tables contain dense tables that require both cell transcription and structure preservation.

Annotation and split. Each crop is annotated in the output format used for evaluation: plain text for text regions, normalized LaTeX for formulas, and HTML-style structure-preserving markup for tables. We use Gemini 3.1 Pro only to generate an initial draft annotation; all final labels are manually corrected against the original crop image. During correction, annotators fix character errors, formula syntax, table cell boundaries, row/column alignment, and hallucinated content. Samples with unresolved visual ambiguity are removed or sent for an additional review pass. To avoid near-duplicate leakage, crops from the same source image or visually repeated table/text region are kept in the same split.

C Third-party Data Statement

Annotation and redistribution policy. For all third-party sources, OCRWildBench annotations are newly produced or corrected by us. Upstream labels are not used directly as final OCR targets unless they satisfy our annotation protocol. For restricted datasets, especially those with non-commercial, no-derivatives, or application-based access terms, we follow the corresponding license restrictions and do not redistribute raw images beyond what is permitted by the original data owners. If OCRWildBench is released publicly, the release will provide derived annotations, evaluation scripts, and source-specific acquisition instructions where necessary.

De-identification policy. OCRWildBench is designed for OCR robustness evaluation rather than identity recognition, so we avoid retaining personally identifiable information. During data construction and human correction, annotators remove or mask visible names, student IDs, phone numbers, email addresses, addresses, account numbers, signatures, and other sensitive fields whenever they appear. For examination-paper sources, we additionally screen school names, student information, scores, and teacher comments. For web-crawled professional tables, we exclude pages containing

Table 12: Third-party sources used in OCRWildBench. We use the listed images or regions as visual inputs and re-annotate them with crop-level OCR labels following our protocol.

Subset	Source	License / terms	Use in OCRWildBench
NSDS	In-house data; M^{6}Doc (Cheng et al., 2023)	In-house research use; M^{6}Doc: CC BY-NC-ND 4.0	Non-canonical scene/document text crops with new OCR annotations.
SCUT-EPT	SCUT-EPT (Zhu et al., 2019)	Non-commercial research use; commercial use requires permission	Exam-paper crops re-annotated for erasures, corrections, overwritten strokes, insertions, and local reading-order ambiguity.
LatexOCR	LatexOCR (Blecher, 2025; Deng et al., 2017)	Official LatexOCR license and usage terms	Formula samples corrected and normalized into LaTeX targets.
WTW	Wired Table in the Wild (Long et al., 2021)	CC BY-NC 4.0	Table regions re-annotated for OCR-oriented cell content and structure recognition.
Prof. Tables	Publicly accessible web tables	Website-specific terms; non-commercial research use only	Public table regions crawled, cropped, and annotated for table content and structure.
Mixed Exam	SCUT-HCCDoc (Zhang et al., 2020); SCUT-EnsExam (Huang et al., 2023)	SCUT-HCCDoc: non-commercial research/education use; SCUT-EnsExam: CC BY-NC-ND 4.0	Mixed printed-handwritten exam crops re-annotated for text, handwriting, formulas, insertions, corrections, and local reading order.

private individuals' information or sensitive business records. Samples that cannot be safely de-identified are removed from the benchmark.

Use of in-house data. The in-house portion of NSDS consists of data collected and curated by our group for internal research use. Before annotation, these data are screened to remove private, identifying, or sensitive content. The in-house raw images will not be redistributed unless the corresponding release permission is obtained. If OCRWildBench is released publicly, this portion will be provided only in forms permitted by our internal data policy, such as derived annotations, metadata, or evaluation protocols.

D Large Language Models Usage

Large language models were used for two limited purposes: (1) manuscript writing assistance, including sentence rephrasing, grammar checking, and readability improvement; and (2) preliminary OCR draft generation during OCRWildBench annotation. All draft annotations were manually corrected against the original crop images, and final labels were produced by human annotators. LLMs were not used for research ideation, methodology design, experimental design, result interpretation, or scientific conclusion generation. All research concepts, analyses, and conclusions were developed by the authors. The authors take full responsibility for the manuscript and confirm that the use of large language models complies with ethical standards and does not constitute plagiarism or scientific misconduct.

E Qualitative Inference Comparison

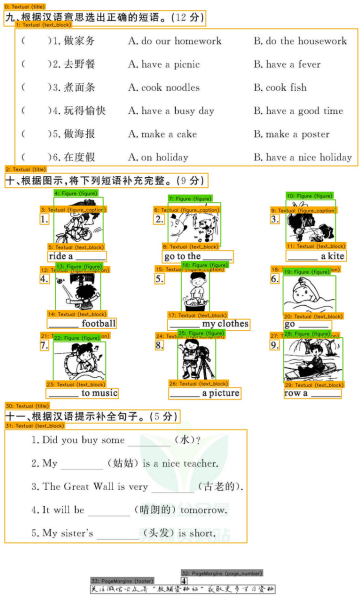
Comparison results on OmniDocBench. Figures 7, 8, and 9 provide page-level parsing and recognition comparisons on representative OmniDocBench v1.5 samples. Each composite figure compares the ground-truth layout with MinerU2.5 Pro and CL-OCR in the first row, and further compares the corresponding recognition outputs in the second row. The examples cover handwritten bilingual notes, mixed text-figure exam pages, and illustration-heavy instructional pages. These cases show the same pattern as Table 2: on canonical full-page documents, CL-OCR improves layout category assignment and region boundary calibration while preserving reading-order consistency, and the recognition outputs remain better aligned with the page structure.

Comparison results on OCRWildBench. Figures 10, 11, and 12 summarize crop-level recognition comparisons on OCRWildBench. The cases are grouped by text, table, and formula recognition to highlight the main input types covered by the benchmark.

(a) Ground Truth (b) MinerU2.5 Pro (c) CL-OCR

(d) MinerU2.5 Pro Recognition (e) CL-OCR Recognition

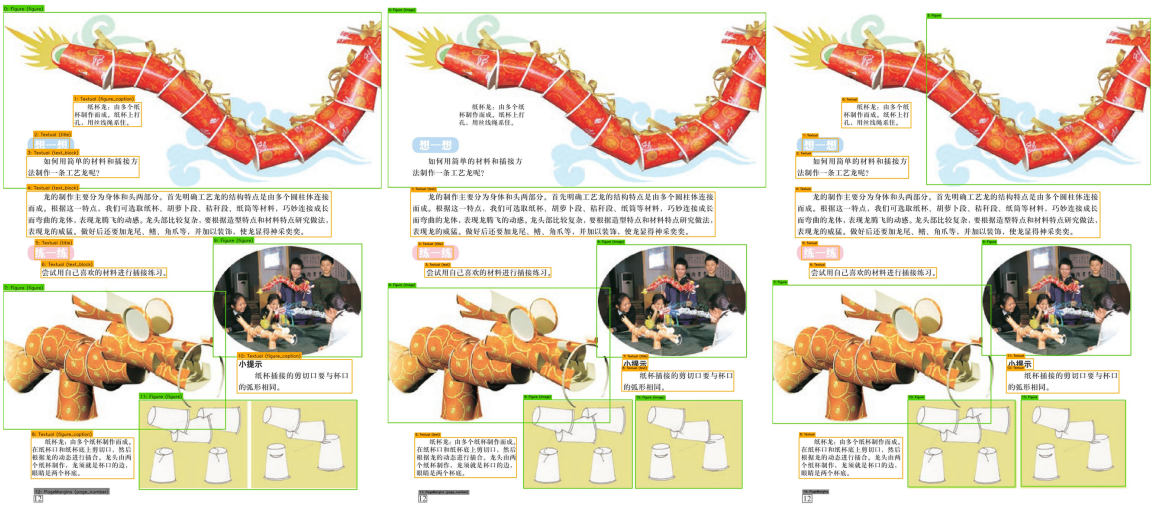
Figure 7: **OmniDocBench v1.5 handwritten-note parsing and recognition.**



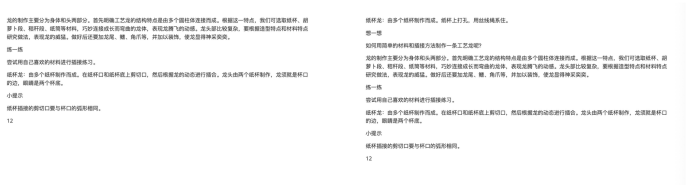
(a) Ground Truth (b) MinerU2.5 Pro (c) CL-OCR

(d) MinerU2.5 Pro Recognition (e) CL-OCR Recognition

Figure 8: **OmniDocBench v1.5 mixed text-figure parsing and recognition.**



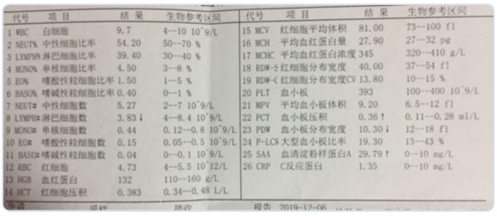
(a) Ground Truth (b) MinerU2.5 Pro (c) CL-OCR



(d) MinerU2.5 Pro Recognition (e) CL-OCR Recognition

Figure 9: **OmniDocBench v1.5 illustration-heavy page parsing and recognition.**

Figure 10: **OCRWildBench text recognition comparison.**



Ground Truth

代号	项目	结果	生物参考区间	代号	项目	结果	生物参考区间
1 WBC	白细胞	9.7	4—10 10^9/L	15 MCV	红细胞平均体积	81.00	73—100 fl
2 NEUT%	中性细胞比率	54.20	50—70 %	16 MCH	平均血红蛋白量	27.90	27—32 pg
3 LYMPH%	淋巴细胞比率	39.40	30—40 %	17 MCHC	平均血红蛋白浓度	345	320—410 g/L
4 MONO%	单核细胞比率	4.50	3—8 %	18 RDW-SD	红细胞分布宽度	40.00	37—54 fl
5 EO%	嗜酸性粒细胞比率	1.50	1—5 %	19 RDW-CV	红细胞分布宽度CV	13.80	10—15 %
6 BASO%	嗜碱性粒细胞比率	0.40	0—1 %	20 PLT	血小板	393	100—400 10^9/L
7 NEUT#	中性细胞数	5.27	2—7 10^9/L	21 MPV	平均血小板体积	9.20	6.5—12 fl
8 LYMPH#	淋巴细胞数	3.83↓	4—8.4 10^9/L	22 PCT	血小板压积	0.36↑	0.11—0.28 ml/L
9 MONO#	单核细胞数	0.44	0.12—0.8 10^9/L	23 PDW	血小板分布宽度	10.30↓	12—18 fl
10 EO#	嗜酸性粒细胞数	0.15	0.05—0.5 10^9/L	24 P-LCR	大型血小板比率	19.30	13—43 %
11 BASO#	嗜碱性粒细胞数	0.04	0—0.1 10^9/L	25 SAA	血清淀粉样蛋白A	29.79↑	0—10 mg/L
12 RBC	红细胞	4.73	4—5.5 10^12/L	26 CRP	C反应蛋白	1.35	0—10 mg/L
13 HGB	血红蛋白	132	110—160 g/L				
14 HCT	红细胞压积	0.383	0.34—0.48 L/L				

MinerU2.5 Pro

代号	项目	结果	生物参考区间	代号	项目	结果	生物参考区间
1	RBC白细胞	9.7	4-10 10^9/L	15	MCY红细胞平均体积	81.00	73-100 fl
2	NEUT%中性细胞比率	54.20	50-70 %	16	MCH平均血红蛋白量	27.90	27-32 pg
3	LYMPH-淋巴细胞比率	39.40	30-40 %	17	MYCN平均血红蛋白浓度	345	320-410 g/L
4	MONON单核细胞比率	4.50	3-8 %	18	RDW红细胞分布宽度	40.00	37-54 fl
5	EOW嗜酸性粒细胞比率	1.50	1-5 %	19	RDW红细胞分布宽度CV	13.80	10-15 %
6	BAS26嗜碱性粒细胞比率	0.40	0-1 %	20	PLT血小板	393	100-400 10^9/L
7	NEUT%中性细胞数	5.27	2-7 10^9/L	21	MPV平均血小板体积	9.20	6.5-12 n1
8	LYMPH-淋巴细胞数	3.83↓	4-8.4 10^9/L	22	PCT血小板压积	0.36↑	0.11-0.28 m/L
9	MONOD单核细胞数	0.44	0.12-0.8 10^9/L	23	PDW血小板分布宽度	10.30↓	12-18 f1
10	EOG嗜酸性粒细胞数	0.15	0.05-0.5 10^9/L	24	P-LCS大型血小板压积	19.30	13-43 %
11	BAS26嗜碱性粒细胞数	0.04	0-0.1 10^9/L	25	SAA血清淀粉样蛋白A	29.79↑	0-10 mg/L
12	RBC红细胞	4.73	4-5.5 10^12/L	26	CRP C反应蛋白	1.35	0-10 mg/L
13	HGB血红蛋白	132	110-160 g/L				
14	HCT红细胞压积	0.383	0.34-0.48 L/L				

CL-OCR

代号	项目	结果	生物参考区间	代号	项目	结果	生物参考区间
1 RBC	白细胞	9.7	4—10 10^9/L	15 MCY	红细胞平均体积	81.00	73—100 fL
2 NEUT%	中性细胞比率	54.20	50—70 %	16 MCH	平均血红蛋白量	27.90	27—32 pg
3 LYMPH%	淋巴细胞比率	39.40	30—40 %	17 MCHC	平均血红蛋白浓度	345	320—410 g/L
4 MON%	单核细胞比率	4.50	3—8 %	18 RDW	红细胞分布宽度	40.00	37—54 fL
5 EON	嗜酸性粒细胞比率	1.50	1—5 %	19 RDW+	红细胞分布宽度CV	13.80	10—15 %
6 BASO%	嗜碱性粒细胞比率	0.40	0—1 %	20 PLT	血小板	393	100—400 10^9/L
7 NEUT#	中性细胞数	5.27	2—7 10^9/L	21 MPV	平均血小板体积	9.20	6.5—12 fL
8 LYMPH#	淋巴细胞数	3.83↓	4—8.4 10^9/L	22 PCT	血小板压积	0.36↑	0.11—0.28 ml/L
9 MON#	单核细胞数	0.44	0.12—0.8 10^9/L	23 PDW	血小板分布宽度	10.30↓	12—18 fL
10 EOG	嗜酸性粒细胞数	0.15	0.05—0.5 10^9/L	24 L-CLS	大型血小板比率	19.30	13—43 %
11 BASO#	嗜碱性粒细胞数	0.04	0—0.1 10^9/L	25 SAA	血清淀粉样蛋白A	29.79↑	0—10 mg/L
12 RBC	红细胞	4.73	4—5.5 10^12/L	26 CRP	C反应蛋白	1.35	0—10 mg/L
13 HGB	血红蛋白	132	110—160 g/L				
14 HCT	红细胞压积	0.383	0.34—0.48 L/L				

Nutrition Facts
Serving Size 8 fl oz (237ml)
Servings Per Container 2
Amount Per Serving
Calories 20 Calories from Fat 0
% Daily Value*
Total Fat 0g 0%
Saturated Fat 0g 0%
Trans Fat 0g
Cholesterol 0mg 0%
Sodium 0mg 0%
Total Carbohydrate 4g 1%
Dietary Fiber 0g 0%
Sugars 3g
Protein 1g
Vitamin A 0% • Vitamin C 0%
Calcium 2% • Iron 0%
*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs:
Calories: 2,000 2,500
Total Fat Less than 65g 80g
Saturated Fat Less than 20g 25g
Cholesterol Less than 300mg 300mg
Sodium Less than 2,400mg 2,400mg
Total Carbohydrate 300g 375g
Dietary Fiber 25g 30g

GT (raw html)

Nutrition Facts		
Serving Size 8 fl oz (237ml)		
Servings Per Container 2		
Amount Per Serving		
Calories 20	Calories from Fat 0	
	% Daily Value*	
Total Fat 0g	0%	
Saturated Fat 0g	0%	
Trans Fat 0g		
Cholesterol 0mg	0%	
Sodium 0mg	0%	
Total Carbohydrate 4g	1%	
Dietary Fiber 0g	0%	
Sugars 3g		
Protein 1g		
Vitamin A 0%	Vitamin C 0%	
Calcium 2%	Iron 0%	
*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs:		
	Calories:	2,000 2,500
Total Fat	Less than	65g 80g
Saturated Fat	Less than	20g 25g
Cholesterol	Less than	300mg 300mg
Sodium	Less than	2,400mg 2,400mg
Total Carbohydrate		300g 375g
Dietary Fiber		25g 30g

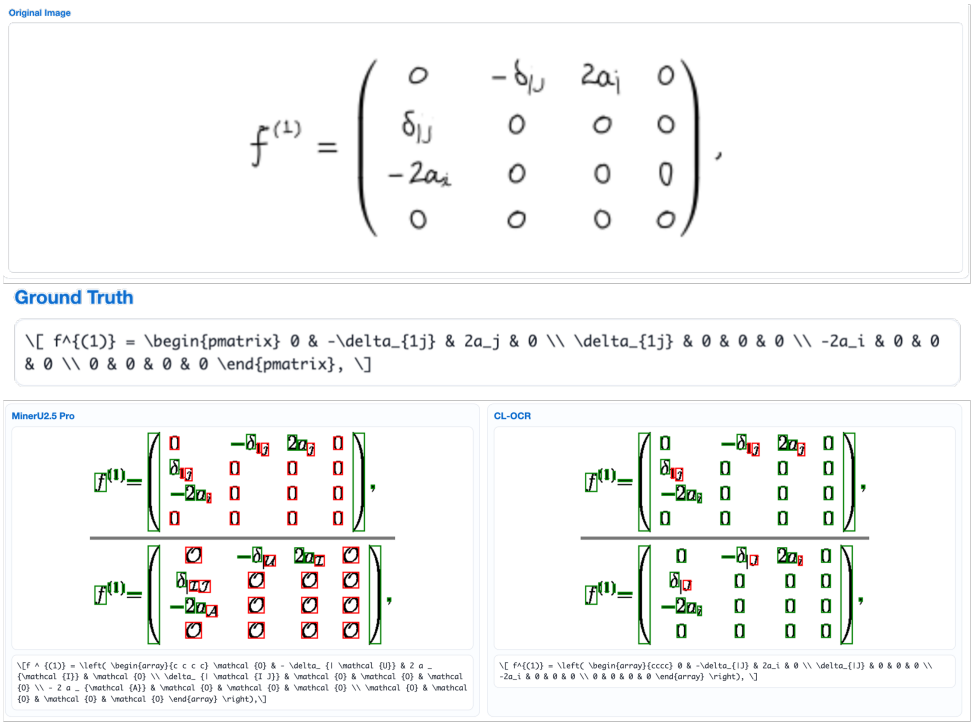
MinerU2.5 Pro

Nutrition Facts		
Serving Size 8 oz (23mT)		
Servings Per Container 2		
Amount Per Serving		
Calories 20	Calories from Fat 0	% Daily Value*
Total Fat 0g	0	0%
Saturated Fat 0g	0	0%
Trans Fat 0g	0	0%
Cholesterol 0mg	0	0%
Sodium 0mg	0	0%
Total Carbohydrate 4g	1	1%
Dietary Fiber 0g	0	0%
Sugars 3g	0	0%
Protein 1g	0	0%
Vitamin A 0%	+ Vitamin C 0%	
Calcium 2%	+ Iron 0%	
*Percent Daily Values are based on a 2,000 calorie diet. Your daily values maybe be higher or lower depending on your calorie needs.		
Calcium 2%	+ Iron 0%	
*Percent Daily Values are based on a 2,000 calorie diet. Your daily values maybe be higher or lower depending on your calorie needs.		
Total Fat	Less than 65g	85g
Total Fat	Less than 30g	30g
Cholesterol	Less than 300mg	300mg
Cholesterol	Less than 2,400mg	2,400mg
Total Carbohydrate	300g	375g
Dietary Fiber	30g	30g

CL-OCR

Nutrition Facts		
Serving Size 8 fl oz (23mT)		
Servings Per Container 2		
Amount Per Serving		
Calories 20	Calories from Fat 0	
		% Daily Value*
Total Fat 0g		0%
Saturated Fat 0g		0%
Trans Fat 0g		
Cholesterol 0mg		0%
Sodium 0mg		0%
Total Carbohydrate 4g		1%
Dietary Fiber 0g		0%
Sugars 3g		
Protein 1g		
Vitamin A 0%	Vitamin C 0%	
Calcium 2%	Iron 0%	
*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs:		
	Calories:	2,000 2,500
Total Fat	Less than	65g 80g
Sat Fat	Less than	20g 25g
Cholesterol	Less than	300mg 300mg
Sodium	Less than	2,400mg 2,400mg
Total Carbohydrate		300g 375g
Dietary Fiber		25g 30g

Figure 11: **OCRWildBench table recognition comparison.**



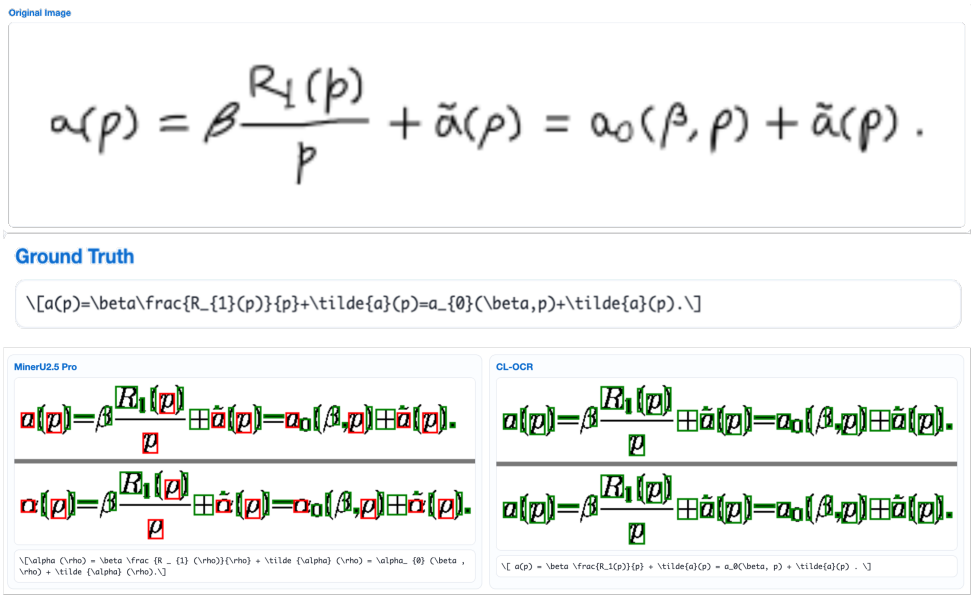


Figure 12: **OCRWildBench formula recognition comparison.**