

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing

Junbo Niu^{1,2*}, Zheng Liu^{1,2*}, Zhuangcheng Gu^{1*}, Bin Wang^{1*‡}, Linke Ouyang^{1*}
 Zhiyuan Zhao^{1*}, Tao Chu^{1*}, Tianyao He^{1*}, Fan Wu^{1*}, Qintong Zhang^{1,2*}, Zhenjiang Jin^{1*}
 Guang Liang¹, Rui Zhang¹, Wenzheng Zhang^{1,2}, Yuan Qu¹, Zhifei Ren¹, Yuefeng Sun¹
 Yuanhong Zheng¹, Dongsheng Ma¹, Zirui Tang^{1,3}, Boyu Niu^{1,3}, Ziyang Miao¹, Hejun Dong¹
 Siyi Qian^{1,2}, Junyuan Zhang¹, Jingzhou Chen^{1,2}, Fangdong Wang¹, Xiaomeng Zhao¹, Liqun Wei¹
 Wei Li¹, Shasha Wang¹, Ruiliang Xu¹, Yuanyuan Cao¹, Lu Chen¹, Qianqian Wu¹, Huaiyu Gu¹
 Lindong Lu¹, Keming Wang¹, Dechen Lin¹, Guanlin Shen¹, Xuanhe Zhou^{1,3}, Linfeng Zhang³
 Yuhang Zang¹, Xiaoyi Dong¹, Jiaqi Wang¹, Bo Zhang¹, Lei Bai¹, Pei Chu¹, Weijia Li¹, Jiang Wu¹
 Lijun Wu¹, Zhenxiang Li¹, Guangyu Wang¹, Zhongying Tu¹, Chao Xu¹, Kai Chen¹
 Yu Qiao¹, Dahua Lin¹, Bowen Zhou¹, Wentao Zhang^{1,2}✉, Conghui He¹✉

¹Shanghai Artificial Intelligence Laboratory, ²Peking University, ³Shanghai Jiao Tong University

We introduce MinerU2.5, a 1.2B-parameter document parsing vision-language model that achieves state-of-the-art recognition accuracy while maintaining exceptional computational efficiency. Our approach employs a coarse-to-fine, two-stage parsing strategy that decouples global layout analysis from local content recognition. In the first stage, the model performs efficient layout analysis on downsampled images to identify structural elements, circumventing the computational overhead of processing high-resolution inputs. In the second stage, guided by the global layout, it performs targeted content recognition on native-resolution crops extracted from the original image, preserving fine-grained details in dense text, complex formulas, and tables. To support this strategy, we developed a comprehensive data engine that generates diverse, large-scale training corpora for both pretraining and fine-tuning. Ultimately, MinerU2.5 demonstrates strong document parsing ability, achieving state-of-the-art performance across multiple benchmarks, surpassing both general-purpose and domain-specific models across multiple recognition tasks, while maintaining significantly lower computational overhead.

* Equal contribution ✉ Corresponding author ‡ Project leader

Correspondence: Conghui He, heconghui@pjlab.org.cn

Code: <https://github.com/opendatalab/MinerU>

Model: <https://huggingface.co/opendatalab/MinerU2.5-2509-1.2B>

Date: September 26, 2025

Contents

1	Introduction	4
2	Related Work	5
2.1	Traditional Pipelines	5
2.2	General-Purpose Vision Language Models	6
2.3	Domain-Specific Vision Language Models	6
3	MinerU2.5	6
3.1	Model Architecture	6
3.2	Two-Stage Parsing Strategy	7
3.3	Training Recipe	8
3.3.1	Stage 0-Modality Alignment	8
3.3.2	Stage 1-Document Parsing Pre-training	9
3.3.3	Stage 2-Document Parsing Fine-tuning	9
3.3.4	Data Augmentation Strategies	10
3.4	Model Deployment	10
4	Data Engine	11
4.1	Overall Workflow	11
4.1.1	Data Curation	11
4.1.2	Pre-training Dataset Preparation	12
4.1.3	Fine-tuning Dataset Construction	13
4.2	Task Reformulation and Enhancement	13
4.2.1	Layout Analysis	13
4.2.2	Formula Recognition	15
4.2.3	Table Recognition	16
4.3	Iterative Mining via Inference Consistency	17
5	Evaluation	19
5.1	Full-Document Parsing Task	19
5.1.1	Evaluation Details and Metrics	21
5.1.2	Evaluation Results	21
5.2	Element-Specific Parsing Task	22
5.2.1	Layout Analysis	22
5.2.2	Table Recognition	23
5.2.3	Formula Recognition	24
6	Conclusion	25
A	Qualitative examples	30
A.1	Overview	31
A.1.1	Among PDF types	31
A.1.2	Among Table types	34
A.1.3	Among Formula types	36
A.2	Compare to Previous Versions	38
A.2.1	Table	38
A.2.2	Formula	40
A.2.3	Layout&OCR	42
A.3	Compare with Others	44
A.3.1	Table	44
A.3.2	Formula	50

A.3.3	Layout&OCR	53
B	Prompt Details	56
B.1	Layout Detection	56
B.2	Text Recognition	56
B.3	Formula Recognition	56
B.4	Table Recognition	57

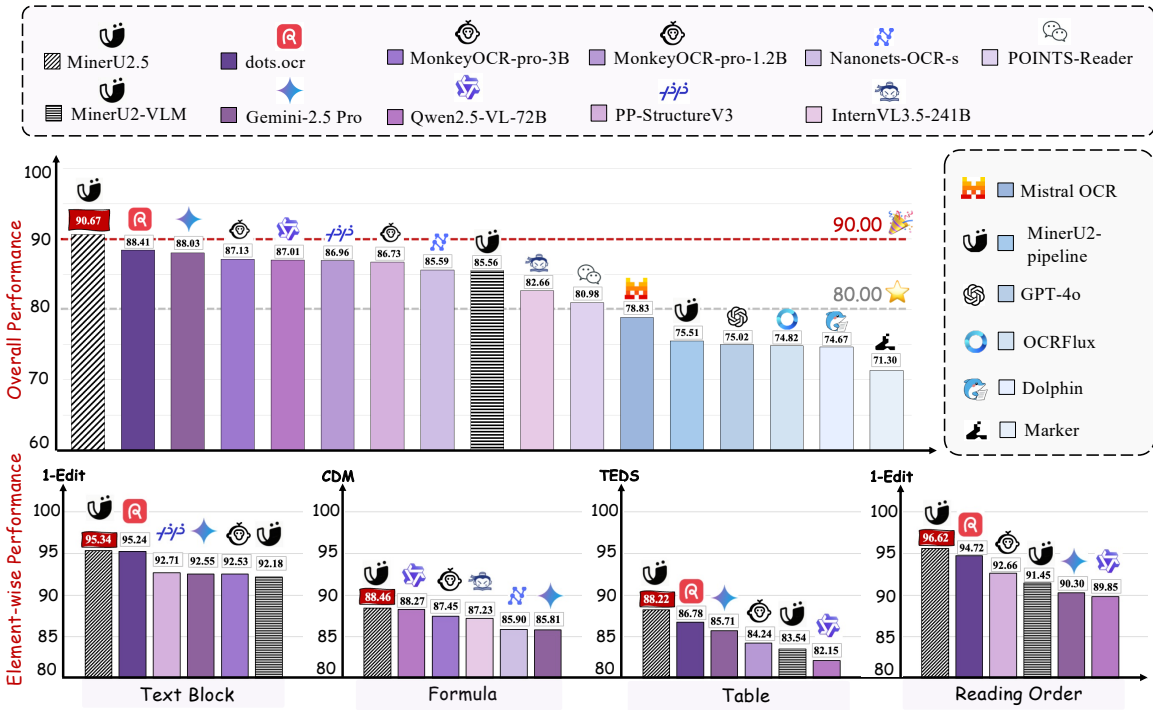


Figure 1: **Performance Highlights of MinerU2.5 on OmniDocBench.** MinerU2.5 consistently outperforms both general-purpose VLMs (e.g., Gemini-2.5 Pro, Qwen2.5-VL-72B, GPT-4o) and domain-specific models (e.g., MonkeyOCR, dots.ocr, PP-StructureV3), establishing new performance records in text recognition, formula recognition, table recognition, and reading order prediction. Detailed results are presented in Table 5.

1 Introduction

Document parsing [56] serves as a fundamental task in multimodal understanding, underpinning a variety of downstream applications such as information extraction [18, 42], Retrieval-Augmented Generation (RAG) [19, 55, 57] and intelligent document analysis [2, 4, 39]. In contrast to natural images, document images are characterized by significantly higher resolutions, denser content, and more complex structural layouts [20, 50, 51]. These inherent properties introduce a unique set of challenges. Firstly, the high resolution and fine-grained layout structures necessitate models capable of processing images at their native resolution. Secondly, the text-dense and often lengthy nature of documents imposes stringent requirements on the parameter efficiency and robustness of the models. Thirdly, the success of OCR is contingent not only on precise text recognition but also heavily on reliable layout analysis and efficient inference.

Contemporary approaches to document parsing can be broadly categorized into two paradigms: pipeline-based approaches [8, 24, 31, 45] and end-to-end approaches based on VLMs [1, 3, 7, 36, 51]. The former employs a modular design, decomposing the task into discrete stages such as layout detection, reading order prediction, and recognition of text lines, formulas, and tables. Each stage is handled by a specialized model. While this approach offers interpretability, it suffers from a cumbersome workflow and the potential for error propagation across modules. The latter paradigm exhibits superior semantic modeling capabilities, yet it is still widely constrained by the hallucination problem in long-document processing and suffers from severe efficiency bottlenecks when dealing with high-resolution inputs. A critical factor limiting the performance and efficiency of VLM-based

parsing is token redundancy, arising from large blank or low-information regions within the document image.

In response to the aforementioned challenges, we introduce a new document parsing framework, **MinerU2.5**. The key innovation is a decoupled architecture that separates *global layout analysis* from *local content recognition* via an efficient coarse-to-fine, two-stage inference mechanism. In the first stage, the model conducts fast and holistic layout analysis on downsampled document images, capturing the global structural organization with minimal computational cost. In the second stage, guided by the detected layout, it crops key regions from the original high-resolution input and performs fine-grained recognition within local windows, thereby preserving native resolution and ensuring high accuracy. This decoupled strategy not only reduces computational cost by an order of magnitude, primarily by avoiding the enormous number of visual tokens with $\mathcal{O}(N^2)$ complexity inherent in end-to-end native-resolution approaches [3, 6, 36], but also brings multiple advantages: it significantly enhances the interpretability of parsing, effectively mitigates the common hallucination problem in VLMs, and allows the two stages to be independently optimized and iterated, resulting in more robust and efficient parsing capabilities. Ultimately, with its lightweight design of only 1.2B parameters, MinerU2.5 exhibits strong adaptability and efficiency in scenarios with long documents and high-density content while ensuring high parsing accuracy. Furthermore, to overcome the challenges of insufficient data diversity, sample imbalance, and inconsistent annotation quality in document parsing, we have developed a closed-loop data engine for complex documents. This engine systematically collects, processes, and generates large-scale, high-quality document corpora. This ensures that our model exhibits precise parsing capabilities and robustness across a wide spectrum of layouts, document types, and complex elements.

MinerU2.5 not only achieves state-of-the-art (SOTA) performance across a wide range of public benchmarks but also represents a qualitative leap in practical application and user experience over the previous MinerU2 version, as demonstrated by the examples in [Appendix A](#). Its key improvements include:

- **Comprehensive and Granular Layout Analysis:** It not only preserves non-body elements like headers, footers, and page numbers to ensure full content integrity, but also employs a refined and standardized labeling schema. This enables a clearer, more structured representation of elements such as lists, references, and code blocks.
- **Breakthroughs in Formula Parsing:** Delivers high-quality parsing of complex, lengthy mathematical formulae and accurately recognizes mixed-language (Chinese-English) equations.
- **Enhanced Robustness in Table Parsing:** Effortlessly handles challenging cases, including rotated tables, borderless tables, and tables with partial borders.

2 Related Work

2.1 Traditional Pipelines

Early OCR systems [8, 24, 31, 45] decompose document parsing into modular pipelines, sequentially executing layout detection [43, 58], text recognition [8], and reading order [49]. For instance, Marker [31] implements a sequential pipeline integrating Surya OCR [32] with layout analysis and reading order prediction modules to process diverse document types. MinerU [45] leverages PDF-Extract-Kit [29] to orchestrate multiple specialized models for layout detection, formula recognition and table extraction. This modular architecture enables specialized optimization of individual components and facilitates targeted refinement of specific subtasks through well-defined module boundaries. However, pipeline-based methods are prone to error propagation across stages and exhibit limited robustness when confronted with complex layouts such as multi-column text or cross-page structures. Moreover,

modular systems often entail multiple interdependencies in practice, rendering usage, maintenance, and updates cumbersome and less efficient.

2.2 General-Purpose Vision Language Models

General-purpose vision language models (VLMs) [1, 3, 7, 62] have emerged as an alternative paradigm for document understanding. Gemini2.5 Pro [7] demonstrates strong OCR capabilities among general VLMs, surpassing traditional pipeline models like MinerU [45] in text parsing and approaching specialized systems like UniMERNet [44] in formula recognition, showcasing the potential of VLMs in OCR applications. Among open-source models, Qwen2.5-VL-72B [3] achieves the best results, using native-resolution vision encoders [10] to adapt to different image sizes, demonstrating the effectiveness of arbitrary-resolution processing in OCR tasks. However, these general models exhibit inherent limitations for document-centric tasks. Proprietary models like Gemini2.5 Pro [7] are expensive and slow in processing, while open-source models require massive parameter scales for optimal performance, limiting practical deployment. Additionally, both types remain susceptible to hallucinations in densely populated text regions, affecting reliability in complex document layouts.

2.3 Domain-Specific Vision Language Models

End-to-End Approaches. Recent domain-specific models [4, 6, 15, 23, 34, 36, 51] adopt end-to-end architectures that unify document parsing within a single model, eliminating the need for cascaded processing stages. GOT [51], as an early representative of end-to-end approaches, pioneered the OCR 2.0 paradigm by establishing both model architecture and data methodology that unified recognition across diverse modalities—text, formulas, tables, and charts—within a single framework. Subsequent models like Ocean-OCR [6], olmOCR [34], and dots.ocr [36] leverage native resolution vision encoders to process documents and construct massive document corpora, further advancing the performance of end-to-end architectures. However, end-to-end designs face scalability challenges: joint optimization of layout and content often reduces accuracy on complex documents, while native-resolution processing introduces prohibitive $\mathcal{O}(N^2)$ complexity. Despite strengths in semantic modeling, these models suffer from hallucinations on long documents and severe inefficiency with high-resolution inputs, where token redundancy from blank or low-information regions becomes a major bottleneck.

Multi-Stage Approaches. Recently, multi-stage methods [11, 17] leveraging VLMs decouple layout analysis from content recognition, combining the efficiency of pipeline approaches with the accuracy of unified models. Dolphin [11] employs a Swin-Transformer VLM that first performs page-level layout, then conducts efficient parallel parsing of identified regions. However, Swin-Transformer’s fixed resolution severely limits crop parsing—sub-regions with extreme aspect ratios suffer from distortion when resized to predetermined dimensions, degrading recognition quality while increasing computational overhead. MonkeyOCR [17] adopts a similar multi-stage strategy but employs a native resolution vision encoder in its second stage, improving both performance and efficiency. However, MonkeyOCR requires multiple specialized models across different stages, increasing system complexity and deployment overhead. A single unified model with native resolution parsing presents a promising direction to address these limitations, which is precisely the goal that MinerU2.5 pursues.

3 MinerU2.5

3.1 Model Architecture

Figure 2 illustrates the overall architecture of MinerU2.5, which is inspired by the classical Qwen2-VL framework [47]. The overall model architecture consists of three major components:

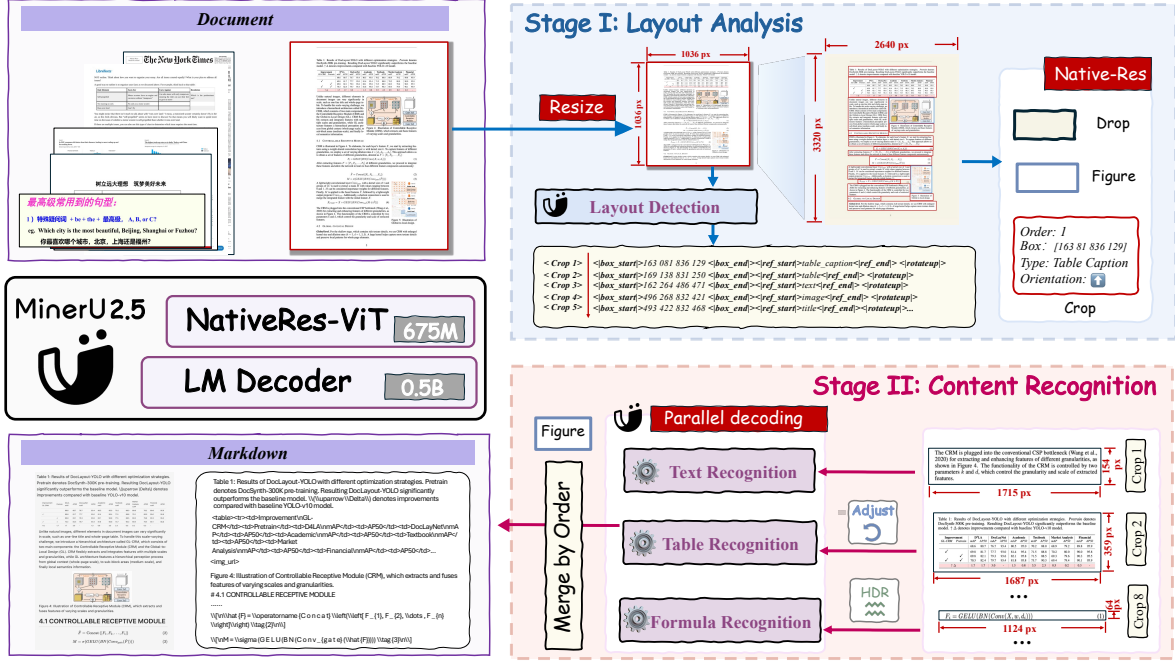


Figure 2: **The framework of MinerU2.5.** In stage I, MinerU2.5 performs rapid, global layout analysis on a downsampled page. In stage II, MinerU2.5 leverages the layout results to crop key regions from the original high-resolution document, performing fine-grained content recognition (e.g., text, table, and formula recognition) within these native-resolution local regions. The detailed prompts used in the inference are illustrated in [Appendix B](#).

Language Model. For the decoder, we employ a 0.5B-parameter Qwen2-Instruct model [41], as document parsing tasks typically exhibit relatively low dependency on large-scale language models. To better accommodate diverse resolutions and aspect ratios in cropped image parsing, we replace the original 1D-RoPE [38] with M-RoPE [47], thus enhancing the model’s generalization ability across varying resolutions.

Vision Encoder. Inspired by Qwen2-VL, MinerU2.5 incorporates a native-resolution encoding mechanism. Although the Qwen2.5-VL series [3] adopts window attention to improve efficiency, this design causes performance degradation in document parsing tasks. Therefore, we employ a 675M-parameter NaViT [10] initialized from Qwen2-VL. This vision encoder supports dynamic image resolutions and employs 2D-RoPE for positional encoding, enabling it to flexibly handle inputs of various resolutions and aspect ratios.

Patch Merger. To balance efficiency and performance, the architecture uses pixel-unshuffle [37] on adjacent 2×2 vision tokens, preprocessing the aggregated vision tokens before passing them into the large language model. This design effectively achieves a trade-off between computational efficiency and task performance.

3.2 Two-Stage Parsing Strategy

In high-resolution document parsing with VLMs, a large proportion of low-information blank regions introduces severe token redundancy, which substantially reduces overall efficiency. Existing end-to-end visual encoding strategies for VLMs face inherent limitations:

- **Crop-based approaches** [51, 62] can partially reduce computational overhead but inevitably sacrifice semantic consistency and layout information.
- **Native-resolution approaches** [3, 13, 36] preserve fine-grained details in high-resolution inputs, yet produce an enormous number of visual tokens with $\mathcal{O}(N^2)$ complexity, rendering them computationally impractical.

To address this dilemma, we propose a **two-stage parsing strategy**. This design decouples layout analysis from local content recognition, thereby improving interpretability, enhancing optimization potential for downstream tasks such as OCR, and effectively reducing the risk of hallucinations. Below, we provide more details of each stage.

Stage I: Layout Analysis. In the first stage, the input image is uniformly resized to a thumbnail of 1036×1036 pixels, enabling global layout analysis while controlling computational cost. The parameter choice is determined through systematic analysis: the thumbnail size must balance global visibility and efficiency—too small leads to detail loss, while too large triggers the quadratic complexity of NaViT. In contrast to native-aspect-ratio thumbnails, adopting a fixed thumbnail size results in more stable bounding-box localization and facilitates more efficient training.

Stage II: Content Recognition. In the second stage, the model leverages the detected layout to crop the native high-resolution image into local regions, which are then parsed at fine granularity. Cropped regions are fed at native resolution with an upper bound of $2048 \times 28 \times 28$ pixels, avoiding detail loss from overly small crops while preventing redundant computation from excessively large ones. This design ensures a robust trade-off between accuracy and efficiency across diverse document parsing scenarios.

3.3 Training Recipe

As described in Section 3.1, MinerU2.5 consists of three core components: vision encoder, patch merger, and language model. Prior to the pre-training phase of MinerU2.5, the vision encoder is initialized from Qwen2-VL-2B-Instruct, while the language model is initialized from Qwen2-Instruct-0.5B. The overall training procedure of MinerU2.5 is divided into three stages, as summarized in Table 1.

3.3.1 Stage 0-Modality Alignment

To ensure that MinerU2.5 acquires the fundamental vision–language alignment ability as well as the OCR recognition capability, we first conduct two-stage modality alignment training on Visual Question Answering (VQA) datasets.

Language-Image Alignment. Only the two-layer MLP within the patch merger is trained, while both the vision encoder and the language model are frozen. We use image-caption pairs¹ for training to effectively project visual features into the LLM embedding space, thus achieving alignment of the modal representation.

Visual Instruction Tuning. All model parameters are unfrozen. The focus is on knowledge accumulation and ability expansion, particularly strengthening visual alignment and OCR capability. The training data² mainly covers image captioning, interleaved text-image pairs, visual alignment, and OCR data. The goal is to enable MinerU2.5 to follow instructions across diverse visual tasks and generate reasonable responses.

¹This dataset is sourced from [LLaVA-Pretrain](#).

²This dataset is sourced from [LLaVA-Instruct](#).

		Stage-0		Stage-1	Stage-2
		a	b		
Vision	Max Resolution	$2048 \times 28 \times 28$	$4096 \times 28 \times 28$	$2048 \times 28 \times 28$	$2048 \times 28 \times 28$
	#Tokens per Image	4 ~ 2048	4 ~ 4096	4 ~ 2048	4 ~ 2048
Data	Dataset	Image Caption	VQA	Layout&OCR	Layout&OCR
	#Samples	558K	665K	6.9M	630K
Model	Trainable	MLP Adaptor	All	All	All
	Sequence Length	4096	4096	8192	16384
	Data Augmentation	No	No	Yes	Yes
Training	Batch Size	128	64	256	256
	LR: ψ_{ViT}	1×10^{-3}	1×10^{-5}	4×10^{-6}	4×10^{-6}
	LR: $\{\theta_{MLP}, \phi_{LM}\}$	1×10^{-3}	1×10^{-5}	4×10^{-5}	4×10^{-5}
	Epoch	1	1	2	3

Table 1: Training setup and hyperparameters in three training stages.

Empirical results demonstrate that MinerU2.5, after VQA-based modality alignment training, exhibits significant improvements in tasks such as layout analysis and content recognition. Conversely, skipping this stage leads to higher losses and a clear drop in overall performance.

3.3.2 Stage 1-Document Parsing Pre-training

The objective of the document parsing pre-training stage is to enable MinerU2.5 to acquire two fundamental capabilities: **layout analysis** and **content recognition**. At this stage, all parameters of the model remain fully trainable.

Training Data. We leveraged a large-scale mixture of model-labeled data and public datasets to ensure both sufficient scale and document diversity. For layout analysis, in consideration of training efficiency, full document images were resized to a fixed resolution with corresponding relative coordinates, and the prompt “Layout Detection:” was used. For content recognition, we employed single-element image samples of text blocks, formula blocks, and table blocks as inputs, with prompts “Text Recognition:”, “Formula Recognition:”, and “Table Recognition:” respectively. More details are shown in the [Appendix B](#).

Training Configuration. The model, initialized from Stage 0, was trained for 2 epochs. Each epoch consisted of a total of 6.9M samples, including 2.3M for layout analysis, 2.4M for text blocks, 1.1M for formula blocks, and 1.1M for table blocks.

Through this document parsing pre-training, the model has acquired strong layout analysis and content recognition capabilities, demonstrating excellent performance across most simple and medium-level scenarios. The resulting model not only serves as a **strong baseline** for downstream fine-tuning, but also functions as an **efficient hard-sample miner** within our data engineering pipeline, facilitating the identification of challenging cases for human annotation and further improving document parsing performance.

3.3.3 Stage 2-Document Parsing Fine-tuning

The objective of the document parsing fine-tuning stage is to further enhance parsing performance in challenging scenarios, while maintaining the detection and parsing capabilities already acquired by

MinerU2.5.

Training Data. To achieve this goal, it is crucial to construct a compact yet high-quality dataset:

- To preserve the model’s fundamental capabilities, we sampled high-quality and diverse examples from the pre-training dataset via data engineering and incorporated them into Stage 2 training, ensuring broad coverage across different document element types.
- From a large-scale, multi-source PDF corpus, we employed data engineering to identify cases where the model still underperformed. We summarized these difficult scenarios and conducted targeted data collection with manual annotation to obtain high-quality samples representing challenging cases.

Training Configuration. We fine-tuned the pre-trained model for 3 epochs. Each epoch contained a total of 630K samples, consisting of 43K for layout analysis, 300K for text blocks, 147K for formula blocks, and 140K for table blocks.

With this targeted data iteration strategy, Stage 2 fine-tuning enables the model to not only retain its established document parsing abilities but also achieve significant improvements in previously challenging scenarios.

3.3.4 Data Augmentation Strategies

To enhance the model’s robustness in handling diverse documents in an open-world setting, we designed a variety of targeted data augmentation strategies during both Stage 1 and Stage 2. These augmentations simulate common types of document interference, and can be categorized as shown in [Table 2](#).

Augmentation Type	Operations
Spatial Transformations	Scaling, Grid Distortion, Rotation
Background Transformations	Texture, Weather effect, Image background, Watermark, Scanlines, Shadow
Color Transformations	Brightness Contrast, Illumination, RGB Shift
Degradation Transformations	PSF Blur, Vibration Blur, Gaussian Blur, Erosion / Dilation

Table 2: Data augmentation strategies for document parsing.

Note that spatial transformations are not applied to layout analysis samples. For different element types, we carefully design augmentation parameters and probabilities in order to strike a balance between model performance and robustness.

3.4 Model Deployment

We implement an efficient offline inference pipeline for MinerU2.5 based on vLLM [16]. While vLLM provides high-throughput serving for large language models, we introduce two additional optimizations tailored for our two-stage document parsing pipeline to further minimize end-to-end latency. First, we employ an asynchronous backend to handle batching submission of page-level requests, enabling better overlap between CPU and GPU workloads. Second, we decouple Stage I and Stage II into independent inference tasks, allowing downstream processing to begin as soon as individual results become available, rather than waiting for entire batches.

A key challenge during deployment was suppressing degenerate token repetition without penalizing legitimate repetitive structures (e.g., tables, equations, or structured content). To address this, we dynamically adjust sampling parameters like `frequency_penalty` and `presence_penalty` in Stage II based on the layout type detected in Stage I. For instance, higher penalties are applied to text paragraphs, while lower values are used for tabular content.

Furthermore, we carefully tuned key vLLM scheduling parameters, including `max_num_batched_tokens`, `max_num_seqs`, and `cuda_graph_sizes`, to improve batch utilization and kernel launch efficiency.

We evaluate all compared models on OmniDocBench [30], a dataset of 1,355 document pages with an average of over 1,100 tokens per page. All models are tested using their official inference scripts under a consistent batched parallel processing protocol, with vLLM startup overhead excluded for fair comparison. After preliminary optimization, MinerU2.5 achieves an end-to-end throughput of **2.12 pages/s**. The end-to-end generation throughput, measured only on valid output tokens from Stage II, reaches **2337.25 tokens/s**³. As shown in Table 3, MinerU2.5 outperforms MonkeyOCR-Pro-3B by 4× and dots.ocr by 7× in page throughput, demonstrating strong inherent efficiency for large-scale document parsing. Notably, even without any deployment optimizations, MinerU2.5 achieves a baseline throughput of 0.95 pages/s and 1045.14 tokens/s, already surpassing other compared models under default configurations.

Model	Parameters	Backend	Hardware	Tokens/sec	Pages/sec
MinerU2-VLM [45]	0.9B	SGLang [59]	A100 80G	3091.23	2.84
dots.ocr [36]	3.0B	vLLM [16]		311.06	0.28
MonkeyOCR-pro-3B [17]	3.7B			520.16	0.47
MonkeyOCR-pro-1.2B [17]	1.9B			589.76	0.53
Nanonets-OCR-s [26]	3.7B			605.92	0.55
MinerU2.5	1.2B	vLLM	RTX 4090 48G	1875.82	1.70
			A100 80G	2337.25	2.12
			H200 141G	4938.31	4.47

Table 3: Inference performance comparison of specialized VLMs and MinerU2.5 across different backends and GPUs.

4 Data Engine

The state-of-the-art performance of MinerU2.5 is underpinned by a systematic Data Engine designed to generate large-scale, high-quality training data with uniform annotation standards. This engine first establishes a vast and diverse foundation through rigorous data curation and refined automated annotation for pre-training. Building upon this foundation, we introduce our novel Iterative Mining via Inference Consistency (IMIC) strategy, which efficiently identifies complex “hard cases” for targeted human annotation. This multi-stage approach creates a virtuous cycle of improvement, progressively enhancing the model’s capabilities. The entire process is illustrated in Figure 3.

4.1 Overall Workflow

4.1.1 Data Curation

Our process begins with a large-scale internal document pool comprising publicly available web data and commercially procured documents. While diverse, this raw pool suffers from a significant

³The end-to-end generation throughput is calculated based on the number of valid tokens produced by Stage II divided by the total processing time for both stages.

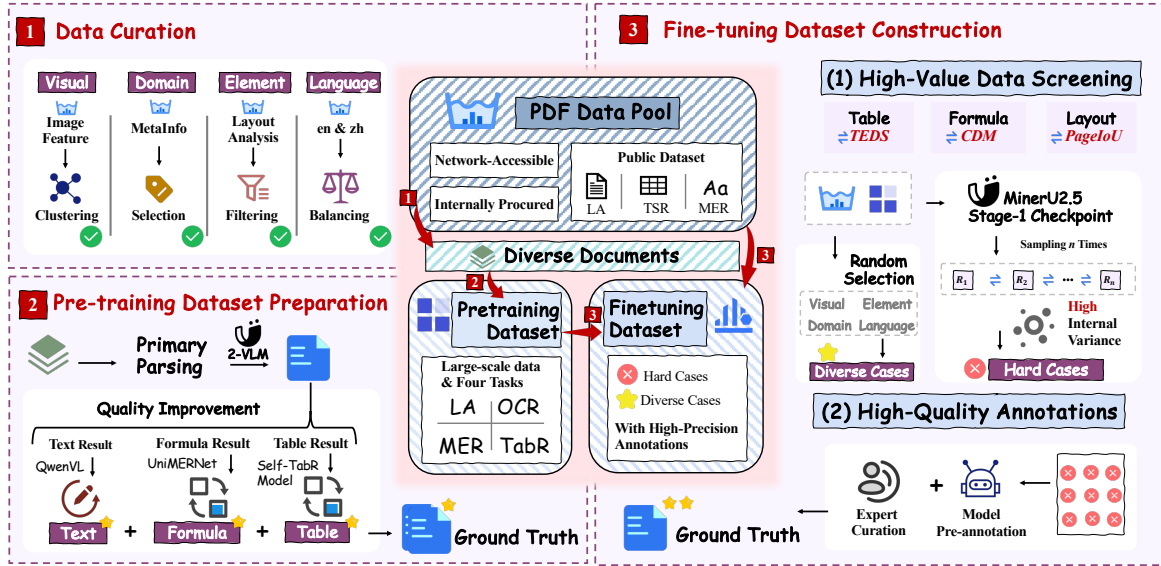


Figure 3: **Overview of the Data Engine.** Our data pipeline consists of three core stages. (1) **Data Curation:** We filter a massive, raw document pool to construct a diverse and balanced dataset based on layout, document type, element balance, and language. (2) **Pre-training Data Preparation:** We generate automated annotations for the curated data and then refine them using specialized, powerful models for text, tables, and formulas to ensure high quality. (3) **Fine-tuning Dataset Construction:** We employ our Iterative Mining via Inference Consistency (IMIC) strategy to automatically discover hard cases, which then undergo meticulous expert curation to create a high-quality SFT dataset.

long-tail distribution. To mitigate this imbalance and enhance training robustness, we implement a rigorous curation process to build a balanced Chinese-English dataset with high diversity across multiple dimensions:

- **Layout Diversity:** We employ page-level image clustering to select exemplars from a wide spectrum of visual layouts and styles.
- **Document Type Diversity:** Using document metadata (e.g., discipline, tags), we perform stratified sampling to ensure a balanced representation of types such as academic papers, textbooks, reports, and presentations.
- **Element Balance:** A preliminary detection model helps ensure a balanced class distribution of key elements like titles, paragraphs, tables, formulas, and figures in the curated set.
- **Language Balance:** We filter the data to maintain a comparable volume of Chinese and English documents.

4.1.2 Pre-training Dataset Preparation

Initial annotations for the curated dataset are generated using our MinerU2-pipeline, establishing a baseline for subsequent refinement. To move beyond this baseline quality, we perform a multi-step refinement process using specialized, expert models for different content types:

- **Textual Content:** We leverage the powerful Qwen2.5-VL-72B-Instruct to verify and correct initial text recognition results on cropped text regions.
- **Formula Content:** Recognized formulas are substituted with higher-fidelity outputs from an

in-house UniMERNet model, which we retrained on our extensive formula dataset to boost its accuracy.

- **Table Content:** All table structures are re-generated using an in-house, high-performance table parsing model.

This refinement workflow yields a high-quality pre-training dataset of image-annotation pairs, covering our four core tasks: layout analysis, text recognition, formula recognition, and table recognition.

4.1.3 Fine-tuning Dataset Construction

While pre-training ensures broad capabilities, the noise inherent in automated annotations creates a ceiling for model performance. To break through this ceiling, our fine-tuning strategy pivots to high-value, difficult examples. We designed an Iterative Mining via Inference Consistency (IMIC) strategy to automatically filter these hard cases from the large-scale data pool. To ensure annotation quality, these select samples are processed through an AI-assisted pipeline: they are first pre-annotated by a foundation model, such as Gemini-2.5-Pro for complex tables, and then meticulously reviewed and corrected by human experts. The final Supervised Fine-Tuning (SFT) dataset combines these high-quality hard cases with a smaller, randomly sampled set of regular examples, equipping MinerU2.5 to excel in complex, real-world parsing scenarios.

4.2 Task Reformulation and Enhancement

To move beyond the limitations of existing document analysis methods, we systematically reformulated the core tasks of layout analysis, formula recognition, and table recognition. This involved defining more robust standards, designing novel task paradigms, and introducing specialized metrics and representations.

4.2.1 Layout Analysis

A Unified Tagging System. A fundamental challenge in layout analysis is the lack of a standardized tagging system. Existing datasets suffer from widespread inconsistencies in element definitions, granularity, and scope. To address this, we engineered a hierarchical and comprehensive tagging system by analyzing a vast corpus of documents. Our system is defined by three key principles:

- **Comprehensive Coverage:** It includes non-body content often ignored by others, such as headers, footers, and page numbers, which is critical for downstream applications like RAG.
- **Fine Granularity:** It decomposes complex elements. For instance, figures are sub-categorized into image, chart, and chemical structure, with distinct tags for their associated captions.
- **Semantic Distinction:** Visually distinct text blocks like code, algorithms, references, and lists are assigned their own categories to preserve crucial semantic information.

Table 4 presents a comparison with mainstream tagging systems, highlighting the superior coverage and granularity of our proposed system.

An Enhanced Multi-Task Paradigm. Traditional methods often treat layout analysis as a standard object detection task, which ignores element rotation and defers reading order prediction to downstream modules. This approach not only impairs the recognition of rotated elements but also increases system coupling. We propose an enhanced paradigm that redefines layout analysis as a multi-task problem. This paradigm simultaneously predicts four key attributes for each document element in a single inference pass: its **Position**, **Class**, **Rotation Angle**, and **Reading Order**. This integrated design effectively resolves the challenge of parsing rotated elements and streamlines the entire document analysis pipeline.

Category	MinerU2-pipeline	PaddleOCR	MinerU2.5
Textual	text	text, toc, abstract	text
	title	title, page_title	title
	×	×	phonetic
	image_caption	common_caption	image_caption
	image_footnote	common_footnote	image_footnote
	table_caption	common_caption	table_caption
	table_footnote	common_footnote	table_footnote
	×	code	code
	×	×	code_caption
	×	×	algorithm
	×	ref_text, ref_block	reference
	×	×	list
Image	image	image, seal, chart, molecular	image
Table	table	table	table
Equation	equation	equation	equation
	×	×	equation_block
Page Margins	×	header	header
	×	footer	footer
	×	aside_text	aside_text
	×	page_number	page_number
	×	page_footnote	page_footnote

Table 4: Comparison of category support across different OCR systems.

PageIoU: A New Metric for Layout Quality. Layout analysis is typically evaluated with object detection metrics like mAP, which rely on a fixed Intersection over Union (IoU) threshold. While effective for well-defined objects, this approach is ill-suited for document layouts where text block boundaries are often ambiguous. This can lead to a discrepancy where quantitative IoU-based scores do not align with qualitative visual assessment.

As illustrated in Figure 4, a prediction that coarsely covers a paragraph (Case 1) can achieve a perfect recall score ($\text{Recall@IoU0.5} = 1.0$), while a more accurate line-by-line prediction (Case 2) is penalized for not matching the paragraph-level ground truth, yielding a lower score ($\text{Recall@IoU0.5} = 0.6$). Visually, however, Case 2 is clearly a better fit.

To better evaluate document layout analysis, we introduce **PageIoU**, a page-level coverage metric that measures the spatial consistency between predicted layouts and ground-truth annotations. Let the predicted layout be

$$P = \{bbox_i \mid i = 1, 2, \dots, n\},$$

and the ground truth be

$$G = \{bbox_j \mid j = 1, 2, \dots, m\},$$

where each $bbox$ denotes a bounding box on the page. We first compute coverage maps for both prediction and ground truth. For example, the ground-truth coverage map is defined as:

$$G_{cover} = \left\{ \sum_{j=1}^m 1_{p \in bbox_j} \mid p \in M \right\},$$

where p is a page pixel and M denotes the non-background region of the page. Similarly, P_{cover} can be

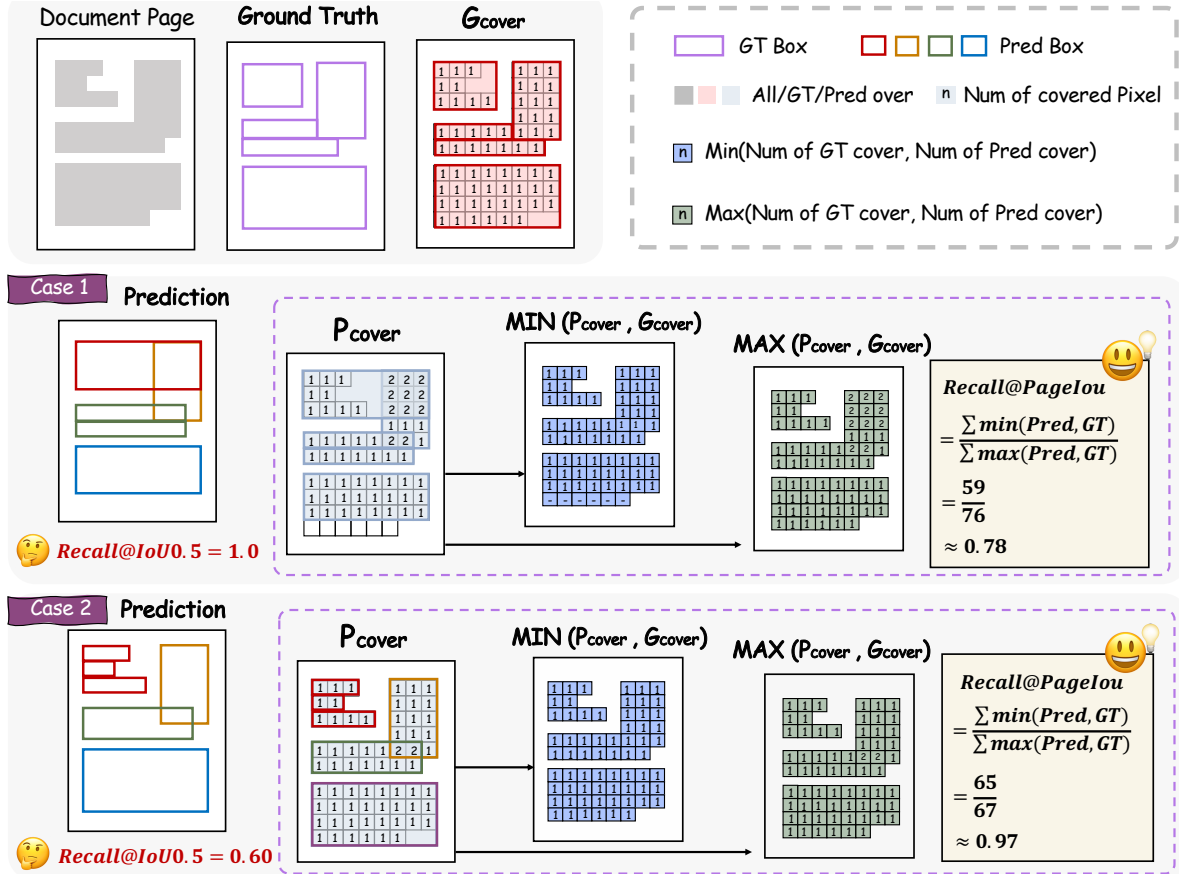


Figure 4: **Illustration of the proposed PageIoU metric.** Case 1 and Case 2 show that IoU-based recall may produce contradictory results compared with visual inspection, whereas PageIoU provides a page-level coverage score that aligns more closely with qualitative observations.

obtained. Based on these, PageIoU is defined as:

$$PageIoU(P, G) = \frac{|P_{cover} \cap G_{cover}|}{|P_{cover} \cup G_{cover}|} = \frac{\sum_{p \in M} \min\{P_{cover}(p), G_{cover}(p)\}}{\sum_{p \in M} \max\{P_{cover}(p), G_{cover}(p)\}}.$$

Here, $|\cdot|$ denotes the summation over all pixel values, while $\cap$ and $\cup$ correspond to the pixel-wise minimum and maximum of coverage counts, respectively. As shown in Figure 4, PageIoU aligns with human perception, scoring the qualitatively poor prediction 0.78 and the superior one 0.97.

4.2.2 Formula Recognition

Decoupling Atomic and Compound Formulas. Existing models struggle with long or multi-line formulas, and VLMs are prone to severe structural hallucinations. We identify the root cause as the tendency to treat all formulas as monolithic entities, failing to account for internal complexity. To this end, MinerU2.5 introduces a “whole-part” decoupling philosophy, classifying formulas into two types based on their structural and semantic integrity:

- **Atomic Formulas:** The smallest, indivisible semantic units with a tight 2D topology (e.g., a single fraction, a matrix).

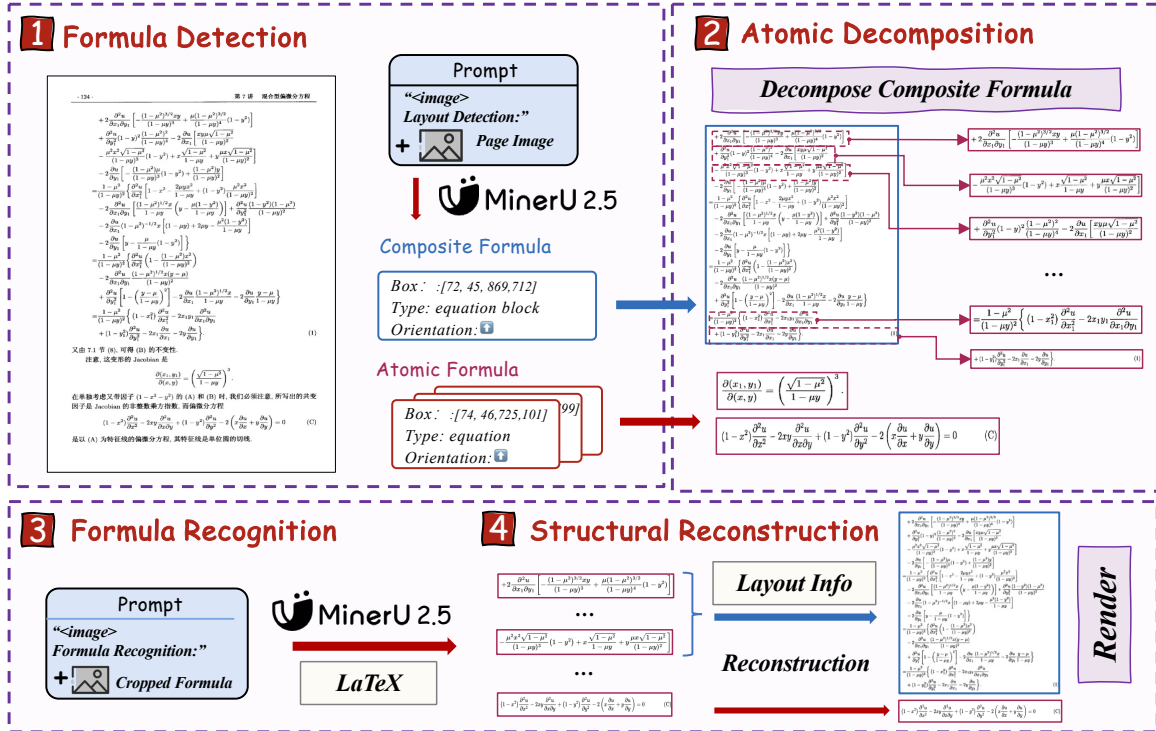


Figure 5: **The proposed ADR framework.** First, a compound formula is decomposed into atomic lines via layout analysis. Next, each line is individually recognized into LaTeX. Finally, the individual results are structurally recombined to produce the complete output.

- **Compound Formulas:** An ordered set of atomic formulas composed vertically with specific alignment relationships (e.g., a multi-line derivation aligned at the equal signs).

The Atomic Decomposition & Recombination (ADR) Framework. To handle the complexity of compound formulas, we propose the ADR framework, which implements a multi-stage “divide and conquer” strategy. As illustrated in [Figure 5](#), the ADR pipeline is powered by our versatile MinerU2.5 model, which acts as both a layout analyzer and a recognition engine, guided by task-specific prompts. The process begins with an initial layout analysis pass, where MinerU2.5, guided by a layout detection prompt, identifies and classifies all formula regions on the page as either atomic or compound. Next, in the decomposition stage, each identified compound formula is segmented into an ordered sequence of its constituent atomic formula lines, which are then cropped as individual images. In the third stage, these simple, semantically independent atomic formula images are fed back into the MinerU2.5 model. This time, using a formula recognition prompt, the model performs high-precision translation of each image into its corresponding LaTeX string. Finally, a lightweight recombination step uses the positional information from the initial layout pass to structurally reassemble the individual LaTeX strings into a single, coherent block, correctly formatting them within environments like align. This approach transforms a single, difficult recognition task into a series of simpler ones, ensuring both high-fidelity recognition of each component and the logical integrity of the overall structure.

4.2.3 Table Recognition

Overcoming Long-Sequence Dependencies. A primary challenge in table recognition is parsing complex, long tables, especially for VLM-based approaches that target HTML. We attribute this

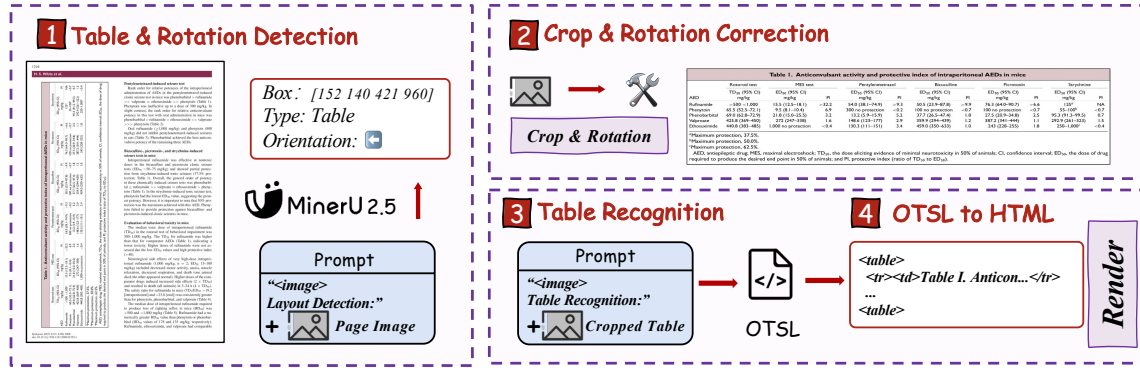


Figure 6: **The Table Recognition Pipeline.** The pipeline first detects a table and its rotation, then corrects its geometry. Next, the rectified image is recognized into the OTSL result, which is finally converted to standard HTML.

difficulty to two inherent weaknesses of the HTML representation: (1) its complex, non-visual syntax must be learned implicitly by the model; and (2) its high token redundancy results in excessively long sequences, degrading performance on large tables. (The issue of rotated tables is effectively handled by our enhanced layout paradigm.)

OTSL: An Optimized Table Structure Language. To robustly handle complex tables, we propose a four-stage recognition pipeline, as depicted in Figure 6. The first two stages handle geometric normalization: the system detects the table’s bounding box and rotation angle, then corrects the image by cropping and rotating it to a canonical orientation. For the crucial third stage, table recognition, we leverage the Optimized Table-Structure Language (OTSL) [25], an intermediate representation developed by IBM [citation, 2023]. We adopted OTSL for its significant advantages over HTML as a target for VLMs. Its minimalist design features a direct structural correspondence to a table’s visual 2D matrix, reducing the number of structural tokens from over 28 to just 5 and shortening the average sequence length by approximately 50%. This makes it a far more effective target for model generation. The final stage is a straightforward conversion from the OTSL output into standard HTML.

4.3 Iterative Mining via Inference Consistency

To enable continuous model improvement and the efficient expansion of our high-quality training dataset, we introduce the IMIC (Iterative Mining via Inference Consistency) strategy. IMIC automatically identifies the most challenging samples—or “hard cases”—for the current model from a large corpus of unlabeled data. This allows us to direct limited human annotation efforts toward the data that offers the maximum value for model improvement.

The core principle of IMIC leverages the stochasticity inherent in model inference. For a given sample, if the model has learned its features robustly, multiple inference passes with stochastic sampling enabled should yield highly consistent outputs. Conversely, significant divergence across outputs suggests the sample lies near the model’s decision boundary—a ‘hard case’ where its predictions are uncertain. Such samples are the most valuable candidates for manual annotation, as they directly target the model’s specific weaknesses.

As illustrated in Figure 7, the implementation is tailored to each recognition task:

- **Layout analysis:** For full document pages, we perform multiple inference runs and measure consistency by calculating the pairwise PageIoU between the resulting layouts. Samples falling below a predefined similarity threshold are flagged as hard cases for precise manual annotation.

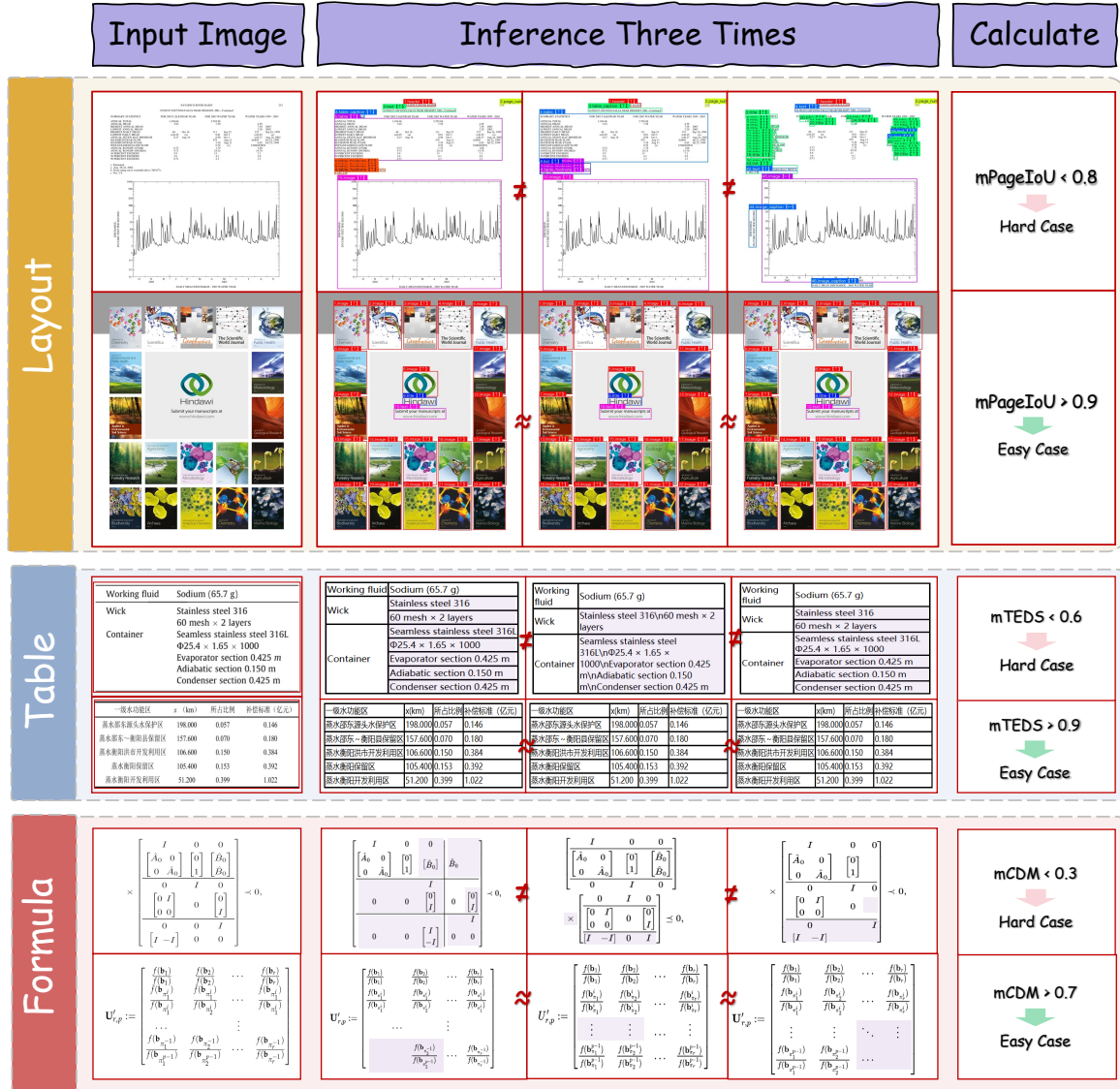


Figure 7: Illustration of the proposed **IMIC (Iterative Mining via Inference Consistency)** strategy. From top to bottom: (a) Layout analysis, (b) Table recognition, and (c) Formula recognition. For each task, the model performs multiple stochastic inference runs, and the pairwise consistency between outputs is calculated with task-specific metrics (PageIoU, TEDS, CDM). Samples with low consistency are automatically identified as hard cases and prioritized for manual annotation.

- **Formula Recognition:** For cropped formula images, consistency is assessed using the pairwise CDM [46] across multiple outputs. Samples with low consistency are prioritized for manual correction.
- **Table Recognition:** For cropped table images, we use the TEDS (Tree-Edit-Distance-based Similarity) score to evaluate consistency across multiple recognized structures. Low-consistency samples are routed to the manual annotation workflow.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing

Model Type	Methods	Parameters	Overall $\uparrow$	Text $\text{Edit}\downarrow$	Formula $\text{CDM}\uparrow$	Table $\text{TEDS}\uparrow$	Table $\text{TEDS-S}\uparrow$	Read Order $\text{Edit}\downarrow$
Pipeline Tools	Marker-1.8.2 [31]	-	71.30	0.206	76.66	57.88	71.17	0.250
	MinerU2-pipeline [45]	-	75.51	0.209	76.55	70.90	79.11	0.225
	PP-StructureV3 [8]	-	86.73	0.073	85.79	81.68	89.48	0.073
General VLMs	GPT-4o [1]	-	75.02	0.217	79.70	67.07	76.09	0.148
	InternVL3-76B [62]	76B	80.33	0.131	83.42	70.64	77.74	0.113
	InternVL3.5-241B [48]	241B	82.67	0.142	87.23	75.00	81.28	0.125
	Qwen2.5-VL-72B [3]	72B	87.02	0.094	<u>88.27</u>	82.15	86.22	0.102
	Gemini-2.5 Pro [7]	-	88.03	0.075	85.82	85.71	90.29	0.097
Specialized VLMs	Dolphin [11]	322M	74.67	0.125	67.85	68.70	77.77	0.124
	OCRFlux [5]	3B	74.82	0.193	68.03	75.75	80.23	0.202
	Mistral-OCR [40]	-	78.83	0.164	82.84	70.03	78.04	0.144
	POINTS-Reader [22]	3B	80.98	0.134	79.20	77.13	81.66	0.145
	olmOCR-7B [34]	7B	81.79	0.096	86.04	68.92	74.77	0.121
	MinerU2-VLM[45]	0.9B	85.56	0.078	80.95	83.54	87.66	0.086
	Nanonets-OCR-s [26]	3.7B	85.59	0.093	85.90	80.14	85.57	0.108
	MonkeyOCR-pro-1.2B [17]	1.9B	86.96	0.084	85.02	84.24	89.02	0.130
	MonkeyOCR-pro-3B [17]	3.7B	87.13	0.075	87.45	81.39	85.92	0.129
	dots.ocr [36]	3B	<u>88.41</u>	<u>0.048</u>	83.22	<u>86.78</u>	<u>90.62</u>	<u>0.053</u>
	MinerU2.5	1.2B	90.67	0.047	88.46	88.22	92.38	0.044

Table 5: Performance comparison of document parsing methods on OmniDocBench across text, formula, table, and reading order extraction tasks.

Model Type	Models	Slides	Academic Papers	Book	Textbook	Exam Papers	Magazine	Newspaper	Notes	Financial Report
Pipeline Tools	Marker-1.8.2 [31]	0.1796	0.0412	0.1010	0.2908	0.2958	0.1111	0.2717	0.4656	0.0341
	MinerU2-pipeline [45]	0.4244	0.0230	0.2628	0.1224	0.0822	0.395	0.0736	0.2603	0.0411
	PP-StructureV3 [8]	0.0794	0.0236	0.0415	0.1107	0.0945	0.0722	<u>0.0617</u>	0.1236	0.0181
General VLMs	GPT-4o [1]	0.1019	0.1203	0.1288	0.1599	0.1939	0.142	0.6254	0.2611	0.3343
	InternVL3-76B [62]	0.0349	0.1052	0.0629	0.0827	0.1007	0.0406	0.5826	<u>0.0924</u>	0.0665
	InternVL3.5-241B [48]	0.0475	0.0857	0.0237	0.1061	0.0933	0.0577	0.6403	0.1357	0.1117
	Qwen2.5-VL-72B [3]	0.0422	0.0801	0.0586	0.1146	<u>0.0681</u>	0.0964	0.238	0.1232	0.0264
	Gemini-2.5 Pro [7]	0.0326	<u>0.0182</u>	0.0694	0.1618	0.0937	0.0161	0.1347	0.1169	0.0169
Specialized VLMs	Dolphin [11]	0.0957	0.0453	0.0616	0.1333	0.1684	0.0702	0.2388	0.2561	0.0186
	OCRFlux [5]	0.0870	0.0867	0.0818	0.1843	0.2072	0.1048	0.7304	0.1567	0.0193
	Mistral-OCR [40]	0.0917	0.0531	0.0610	0.1349	0.1341	0.0581	0.5643	0.3097	0.0523
	POINTS-Reader [22]	0.0334	0.0779	0.0671	0.1372	0.1901	0.1343	0.3789	0.0937	0.0951
	olmOCR-7B [34]	0.0497	0.0365	0.0539	0.1204	0.0728	0.0697	0.2916	0.122	0.0459
	MinerU2-VLM[45]	0.0745	0.0104	0.0357	0.1276	0.0698	0.0652	0.1831	0.0803	0.0236
	Nanonets-OCR-s [26]	0.0551	0.0578	0.0606	0.0931	0.0834	0.0917	0.1965	0.1606	0.0395
	MonkeyOCR-pro-1.2B [17]	0.0961	0.0354	0.053	0.111	0.0887	0.0494	0.0995	0.1686	0.0198
	MonkeyOCR-pro-3B [17]	0.0904	0.0362	0.0489	0.1072	0.0745	0.0475	0.0962	0.1165	0.0196
	dots.ocr [36]	0.0290	0.0231	0.0433	<u>0.0788</u>	0.0467	<u>0.0221</u>	0.0667	0.1116	0.0076
	MinerU2.5	<u>0.0294</u>	0.0235	<u>0.0332</u>	0.0499	<u>0.0681</u>	0.0316	0.054	0.1161	<u>0.0104</u>

Table 6: Document Parsing Performance in Text Edit Distance on OmniDocBench: evaluation using edit distance across 9 PDF page types.

Evaluation

In this section, we present a comprehensive quantitative evaluation of MinerU2.5 to demonstrate its effectiveness in document parsing tasks. Specifically, we compare MinerU2.5 against leading general-purpose VLMs including GPT-4o [1], Gemini-2.5 Pro [7], and Qwen2.5-VL [3], as well as state-of-the-art domain-specific VLMs such as dots.ocr [36], MonkeyOCR [17], and olmOCR [34]. The evaluation is organized into two parts: Section 5.1 presents full-document parsing results across multiple benchmarks, while Section 5.2 focuses on element-specific capabilities including layout analysis, formula recognition, and table recognition.

5.1 Full-Document Parsing Task

We evaluate MinerU2.5’s full document parsing performance on three prominent benchmarks: OmniDocBench [30], Ocean-OCR [6] benchmarks, and olmOCR-bench [34]. These benchmarks provide comprehensive evaluation from different dimensions, covering diverse document types, various

Model	Edit Distance ↓		F1-score ↑		Precision↑		Recall↑		BLEU↑		METEOR↑	
	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
Mathpix [27]	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 [8]	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 [45]	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 [8]	0.323	0.649	0.707	0.864	0.690	0.912	0.730	0.821	0.517	0.537	0.674	0.699
Gemini-2.5 Pro [7]	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 [1]	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 [3]	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 [62]	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 [47]	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 [53]	0.244	0.437	0.804	0.778	0.793	0.721	0.837	0.875	0.695	0.431	0.640	0.642
MinerU2-VLM [45]	<u>0.048</u>	0.182	0.936	0.941	0.926	0.927	<u>0.947</u>	0.958	0.893	0.611	0.950	0.837
Ocean-OCR [6]	0.057	0.062	<u>0.937</u>	<u>0.962</u>	0.932	<u>0.956</u>	0.956	0.974	<u>0.906</u>	0.912	<u>0.945</u>	0.916
MonkeyOCR-pro-1.2B [17]	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 [28]	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 [36]	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[51]	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	<u>0.082</u>	0.945	0.965	<u>0.948</u>	0.966	0.942	<u>0.964</u>	0.909	<u>0.817</u>	0.950	<u>0.887</u>

Table 7: Evaluation results on Ocean-OCR bench on dense English (en) and Chinese (zh) OCR for document-level pages. Some model results are sourced from the OceanOCR official reports.

Model	Overall	AR	OSM	TA	OS	HF	MC	LTT	Base
MinerU2-pipeline[45]	55.6	61.8	13.5	60.9	17.3	96.6	59.0	39.1	96.6
Nanonets-OCR-s[26]	60.7	63.9	41.0	77.7	39.5	40.7	69.9	53.4	<u>99.3</u>
GPT-4o[1]	63.2	44.1	37.6	69.1	40.9	94.2	68.9	54.1	96.7
MonkeyOCR-pro-1.2B[17]	64.3	65.4	26.9	60.3	31.2	93.3	66.2	<u>81.7</u>	89.5
Qwen2.5-VL-72B[3]	64.8	<u>72.2</u>	<u>51.1</u>	67.3	38.6	73.6	68.3	49.1	98.3
MonkeyOCR-pro-3B[17]	68.8	67.7	28.4	74.6	36.1	91.2	76.6	80.1	95.3
olmOCR[34]	71.8	63.9	41.0	72.9	43.9	<u>95.1</u>	77.3	81.2	98.9
dots.ocr[36]	<u>73.6</u>	66.3	35.8	88.3	<u>40.9</u>	94.1	82.4	81.2	99.5
MinerU2.5	75.2	76.6	54.6	<u>84.9</u>	33.7	96.6	<u>78.2</u>	83.5	93.7

Table 8: Evaluation results on olmOCR-bench grouped by document types, including arXiv Math(AR), Old Scans Math (OSM), Tables (TA), Old Scans (OS), Headers Footers (HF), Multi Column (MC) and Long Tiny Text (LTT). Results on AR and OSM are replaced with ExpRate, and other results are sourced from the official reports of olmOCR-bench and dots.ocr. The Overall Score (Overall) represents the average across all document types.

quality conditions, and different parsing challenges to thoroughly assess the model’s robustness and generalization capabilities.

- **OmniDocBench** [30]: This evaluation dataset is designed for diverse document parsing in real-world scenarios, encompassing nine document types, four layout types, and three language types. It offers a comprehensive assessment of parsing scores for text, formulas, tables, and reading order in full-document parsing, as well as for element-specific parsing tasks.
- **olmOCR-bench** [34]: This evaluation dataset comprises 1,402 PDF documents sourced from various repositories, organized into seven subsets. Certain test patterns are applicable across all document types (e.g., presence, absence, reading order), while others are specifically targeted at challenging yet crucial content extraction objectives (e.g., tables, mathematical formulas).
- **Ocean-OCR benchmark** [6]: This evaluation dataset consists of 100 images from English papers and 100 images from Chinese papers. It primarily evaluates the ability of text parsing and

employs several text OCR-related evaluation metrics, such as Normalized Edit Distance, F1 Score, Precision, Recall, BLEU, and METEOR.

5.1.1 Evaluation Details and Metrics

For OmniDocBench [30], we evaluate on the latest version with three key improvements:

- Enhanced resolution for Notes and Newspapers from 72 to 200 DPI, enabling more accurate evaluation of fine-grained text and handwritten content.
- An addition of 374 pages to balance Chinese-English content distribution and enrich mathematical formula coverage. Currently, it contains a total of 1,355 pages.
- Evaluation methodology updated to hybrid matching algorithm.

The Overall score combines three core metrics:

$$\text{Overall} = \frac{(1 - \text{Text}^{\text{Edit}}) \times 100 + \text{Table}^{\text{TEDS}} + \text{Formula}^{\text{CDM}}}{3}$$

For olmOCR-bench [34], we replace the formula scores of Arxiv Math (AR) and Old Scans Math (OSM) with the more reliable ExpRate of CDM [46]. The original evaluation compares LaTeX formulas by parsing them into abstract syntax trees and matching Unicode tokens, which is overly sensitive to syntax variations (e.g., `\cdots` vs. `\dotsb`) that render identically but are scored as different. To avoid this bias, we adopt ExpRate, which directly compares rendered outputs, assigning 1 for exact matches and 0 otherwise.

5.1.2 Evaluation Results

MinerU2.5 demonstrates exceptional performance across all benchmarks, achieving state-of-the-art results in most metrics (Tables 5 to 8).

As shown in Table 5, MinerU2.5 achieves an overall score of 90.67 on OmniDocBench, outperforming the second-best model dots.ocr [36] by 2.26 points and Gemini-2.5 Pro [7] by 2.64 points. In text recognition tasks, MinerU2.5 achieves the lowest edit distance of 0.047, marginally better than dots.ocr at 0.048 and significantly outperforming Gemini-2.5 Pro, which scores 0.075. For formula recognition, MinerU2.5 leads with a CDM score of 88.46, exceeding both dots.ocr at 86.78 and Gemini-2.5 Pro at 85.71. In table recognition tasks, MinerU2.5 achieves the highest TEDS score of 88.22 and TEDS-S score of 92.38. For reading order evaluation, it maintains the best edit distance of 0.044. The document-type specific results presented in Table 6 demonstrate that MinerU2.5 achieves best or second-best performance in 6 out of 9 categories. For textbooks, it delivers the best performance with an edit distance of 0.0499, substantially outperforming dots.ocr’s 0.0788. For newspapers, MinerU2.5 leads with a score of 0.054, surpassing all competing models. In both financial reports and slides categories, MinerU2.5 achieves second-best performance with scores of 0.0104 and 0.0294 respectively.

For the results of the Ocean-OCR benchmark presented in Table 7, MinerU2.5 demonstrates exceptional performance in dense OCR tasks. On English documents, it achieves the lowest edit distance of 0.033 and the highest F1-score of 0.945, accompanied by best-in-class BLEU and METEOR scores of 0.909 and 0.950 respectively. For Chinese documents, MinerU2.5 achieves the highest F1-score of 0.965 and Precision of 0.966, while maintaining strong BLEU and METEOR scores of 0.817 and 0.887 respectively.

The results of olmOCR-bench are shown in Table 8, where MinerU2.5 achieves an overall score of 75.2, surpassing dots.ocr’s 73.6 by 1.6 points. In the arXiv Math category, it leads with a score of 76.6, outperforming Qwen2.5-VL-72B [3]’s 72.2 by 4.4 points. For Old Scans Math, MinerU2.5 dominates with a score of 54.6, exceeding all other evaluated models. In the Long Tiny Text category, it achieves 83.5, surpassing MonkeyOCR-pro-1.2B [17] which scores 81.7.

Method	Textual			Image			Table			Equation			Page Margins			Full Page		
	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑
OmniDocBench [30]																		
LayoutLMv3 [14]	90.4	48.2	58.1	72.1	51.2	57.2	72.6	55.1	61.0	-	36.9	-	-	-	-	-	-	-
MinerU2-VLM [45]	90.3	95.6	91.9	87.2	91.0	90.9	96.0	97.1	97.8	87.4	95.8	90.5	-	-	-	-	-	-
DocLayout-YOLO [58]	95.4	98.3	96.5	<u>87.6</u>	96.7	<u>94.7</u>	94.9	<u>98.1</u>	<u>98.4</u>	<u>95.3</u>	90.6	93.8	-	98.7	-	92.3	97.7	94.1
PP-StructureV3 [8]	96.8	96.7	96.6	86.4	92.1	92.9	96.6	97.4	98.2	96.5	97.6	96.7	92.9	86.2	<u>88.1</u>	<u>94.8</u>	96.2	<u>94.6</u>
MinerU2.5	97.2	<u>98.0</u>	97.5	89.6	<u>94.3</u>	95.0	<u>96.0</u>	98.1	98.4	92.4	99.6	<u>94.7</u>	<u>89.9</u>	<u>95.4</u>	91.4	95.8	<u>97.0</u>	95.9
D ⁴ LA [9]																		
LayoutLMv3 [14]	86.9	41.2	52.4	59.3	32.0	31.4	59.3	41.8	43.3	-	50.5	-	-	-	-	-	-	-
MinerU2-VLM [45]	88.3	88.9	87.9	<u>56.7</u>	35.0	38.1	<u>89.1</u>	<u>84.1</u>	<u>90.6</u>	<u>38.3</u>	<u>99.4</u>	79.1	-	-	-	-	-	-
DocLayout-YOLO [58]	86.3	97.8	90.8	41.5	92.9	62.6	87.6	89.0	89.8	31.9	80.2	91.1	-	<u>95.0</u>	-	82.6	95.4	<u>87.3</u>
PP-StructureV3 [8]	<u>88.5</u>	93.5	90.0	50.1	82.3	<u>67.9</u>	87.1	81.1	89.7	24.6	85.9	<u>92.1</u>	76.8	84.2	<u>79.1</u>	<u>85.7</u>	91.0	86.0
MinerU2.5	91.8	98.3	94.6	53.8	94.3	72.8	91.9	78.9	91.4	46.0	100.0	91.0	<u>75.9</u>	97.6	84.2	90.4	<u>92.5</u>	90.2
DocLayNet [33]																		
LayoutLMv3 [14]	88.8	59.3	67.9	79.0	50.3	61.9	75.2	54.9	61.8	-	31.9	-	-	-	-	-	-	-
MinerU2-VLM [45]	88.1	96.1	91.7	85.5	78.1	91.3	94.9	<u>94.4</u>	95.6	83.9	<u>97.0</u>	90.0	-	-	-	-	-	-
DocLayout-YOLO [58]	86.9	96.8	91.2	85.8	<u>96.2</u>	91.3	92.0	95.7	94.8	80.5	86.9	82.8	-	<u>97.7</u>	-	88.0	<u>96.3</u>	90.9
PP-StructureV3 [8]	90.9	97.3	93.8	91.7	90.4	<u>94.2</u>	96.4	93.7	96.7	88.8	96.0	<u>92.1</u>	76.8	79.3	<u>77.4</u>	<u>92.4</u>	95.7	<u>93.0</u>
MinerU2.5	<u>90.2</u>	99.6	94.8	92.5	96.3	95.9	<u>96.3</u>	93.5	97.1	88.9	98.6	93.5	<u>76.3</u>	98.9	86.3	92.8	97.7	94.6

Table 9: Comparison of layout analysis performance (Precision@PageIoU, Recall@PageIoU, F1-score@PageIoU) across different methods and content types on multiple layout analysis benchmarks.

5.2 Element-Specific Parsing Task

5.2.1 Layout Analysis

We validate the effectiveness of our layout analysis by performing a fair, zero-shot comparison with leading methods on three publicly available datasets:

- **OmniDocBench** [30]: A recent benchmark for document parsing that includes detailed layout annotations.
- **D⁴LA** [9]: Contains 11,092 noisy document images annotated with 27 categories, split into 8,868 training and 2,224 test images. We use its test set with annotations for evaluation.
- **DocLayNet** [33]: A large-scale dataset of 80,863 pages from 7 document types, manually annotated with 11 categories. We use its validation set with annotations for evaluation.

We compare our MinerU2.5 with several recent methods, including LayoutLMv3 [14], MinerU2-VLM [45], DocLayout-YOLO [58] and PP-StructureV3 [8]. For a equitable assessment, we evaluate all models without dataset-specific training. To account for differences in detection granularity and category definitions, we unified the evaluation by mapping all labels to five broad categories and using the PageIoU metric, which assesses the spatial overlap without considering category labels for the “Full Page” score.

The results in Table 9 show that MinerU2.5 significantly outperforms other models, achieving the top Full Page F1-score@PageIoU across all benchmarks. It also secures leading F1-scores@PageIoU for the majority of individual element types. This consistent superiority confirms that the PageIoU metric provides a robust basis for comparison, effectively capturing model performances independent of annotation inconsistencies.

Method	PubTabNet		FinTabNet		CC-OCR		OCRBench v2		In-house TR Benchmark	
	TEDS↑	TEDS-S↑	TEDS↑	TEDS-S↑	TEDS↑	TEDS-S↑	TEDS↑	TEDS-S↑	TEDS↑	TEDS-S↑
RapidTable [35]	86.57	96.43	73.77	84.84	50.93	65.84	65.55	77.73	51.96	71.94
MiniCPM-V 4.5 [54]	80.30	87.67	<u>85.41</u>	<u>89.18</u>	68.49	77.55	80.28	85.65	55.47	69.61
InternVL3.5-241B [48]	83.75	88.76	84.74	87.92	62.87	69.52	79.5	85.81	56.32	69.3
Qwen2.5-VL-7B [3]	81.60	86.78	82.58	87.46	78.29	84.26	77.44	84.71	57.34	73.17
Qwen2.5-VL-72B [3]	84.39	87.91	82.90	87.13	<u>81.22</u>	<u>86.48</u>	81.33	86.58	62.79	76.91
GPT-4o [1]	76.53	86.16	83.94	87.00	66.98	79.04	70.51	79.55	46.99	70.29
Gemini-2.5 Pro [7]	-	-	-	-	85.56	90.07	88.94	<u>89.47</u>	<u>69.72</u>	<u>81.29</u>
dots.ocr [36]	90.65	93.76	84.12	87.86	75.42	81.65	82.04	86.27	66.91	79.27
Nanonets-OCR-s [26]	63.58	75.68	68.06	73.6	66.15	71.33	69.66	76.28	54.35	66.12
MinerU2-VLM [45]	88.11	90.85	78.49	83.03	64.61	71.8	73.22	78.24	63.54	76.66
MinerU2.5	<u>89.07</u>	93.11	95.97	97.61	79.76	85.16	<u>87.13</u>	90.62	71.48	82.83

Table 10: Table Recognition Performance. MinerU2.5 achieves SOTA performance on most benchmarks among TEDS and TEDS-S metrics, and the remaining ones are also generally competitive with the SOTA. (CCOCR and OCRBench v2 are OCR evaluation benchmarks, we only select the subsets that contain tables. PubTabNet and FinTabNet have a large number of images, so we have not evaluate Gemini-2.5 Pro on them.).

5.2.2 Table Recognition

We evaluate representative methods, covering traditional table recognition methods, general multi-modal large models and document parsing models, on five table recognition benchmarks as shown in Table 10. Below is an introduction to each benchmark:

- **PubTabNet** [61] is the first large-scale table recognition dataset that provides annotations (in HTML format) of table images, captured from scientific articles. PubTabNet contains 9k tables in its test set.
- **FinTabNet** [60] is a dataset containing tables from the annual reports of 500 companies. The major challenge of this benchmark is that financial tables largely differ from scientific and government document tables in that the former has fewer graphical lines, larger gaps within each table, and more color variations. FinTabNet contains 10k tables in its test set.
- **CC-OCR** [52] and **OCRBench v2** [12] are both designed to evaluate the OCR capabilities of multimodal large models and contain several OCR tasks. We only retain the data related to document recognition and those images that include tables. After filtering, CC-OCR remains 300 images and OCRBench v2 remains 700 images.
- **In-house TR Benchmark.** To better evaluate the table recognition accuracy of different methods, we considering various table attributes such as the number of table rows and columns, the number of merged cells, the length of the table, the length of the cell content, the type of cell content, the line style of the table, and construct a very diverse evaluation set, which contains approximately 500 tables.

MinerU2.5 achieves SOTA performance on most benchmarks, and shows competitive results with the SOTA on the remaining ones. Specifically, for PubTabNet, Rapidtable [35] achieves the best performance in the TEDS-S metric, while dots.ocr [36] excel in the TEDS metric. Meanwhile, despite using only 20% of the PubTabNet training set, MinerU2.5 still demonstrate comparable results, coming second and third in TEDS and TEDS-S, respectively. For FinTabNet, MinerU2.5 achieves the best result and outperform other methods by a significant margin, this could be mainly credited to the large-scale high-quality table data we extracted from financial reports for training. On CC-OCR benchmark, MinerU2.5 came third after Gemini-2.5 Pro and Qwen2.5-VL-72B. On OCRBench v2 benchmark, MinerU2.5’s performance is competitive to that of Gemini-2.5 Pro, and it significantly outperform other

Method	Public Dataset					In-house Dataset		
	CPE	HWE	SCE	SPE	LaTeX-80M ^M	Chinese	Fuzzy Math	Complex
UniMERNet* [44]	98.2	96.5	95.4	99.2	83.9	84.0	84.3	67.9
PP-Formula_plus-L [21]	98.2	94.7	95.7	99.2	85.9	84.0	86.5	76.5
Gemini-2.5-flash [7]	89.2	90.0	85.1	97.5	78.7	88.1	89.4	80.1
Qwen2.5-VL-72B [3]	88.9	91.8	95.5	96.2	83.4	90.8	86.7	81.4
GPT-4o [1]	82.7	85.9	87.8	96.7	73.4	88.3	85.0	78.6
InternVL3.5-241B [48]	91.7	93.2	95.1	97.8	<u>86.9</u>	82.7	<u>90.3</u>	<u>82.0</u>
dots.ocr [36]	86.8	90.5	94.7	97.5	81.8	74.4	86.2	77.4
MinerU2.5	96.6	94.4	96.4	98.4	90.6	<u>90.7</u>	92.6	82.2

Table 11: Formula Recognition Performance (CDM metric used for evaluation). MinerU2.5 achieves 4 SOTA results and one second-best result across 7 benchmarks. LaTeX-80M^M denotes the matrix benchmark of LaTeX-80M dataset. * indicates that the UniMERNet results are based on an improved version compared to the publicly available open-source implementation.

methods. On the diverse In-house TR Benchmark, MinerU2.5 and Gemini-2.5 Pro both significantly outperform other methods, with MinerU2.5 achieving a slight advantage over Gemini-2.5 Pro.

5.2.3 Formula Recognition

For formula recognition, comparison models include various approaches, covering specialized formula recognition models, document parsing models, and general vision-language models. The evaluation datasets consist of the following:

- **UniMER-Test** [44] is a comprehensive evaluation dataset for general formula recognition. Targeted at real-world formula recognition across various scenarios, UniMER-Test includes four subsets: CPE (complex printed equations), HWE (handwritten equations), SPE (screen printed equations), and SCE (simple printed equations).
- **LaTeX-80M^M** is a matrix subset of LaTeX-80M⁴, featuring intricate mathematical structures encompassing matrices, conditional expressions, and nested combinations.
- **In-house dataset** consists of the following subsets: (1) Chinese, targeted at evaluation on real-world document equations which contain Chinese characters. (2) Fuzzy math, which focuses on authentic mathematics textbooks and exam documents characterized by compromised visual quality due to factors like blur, degeneration, watermarks, and so on. (3) Complex, an extremely difficult dataset aimed at assessing the ability of converting the most complex mathematical formulas to LaTeX codes.

Results are shown in Table 11 and the CDM [46] metric is used for evaluation. Across all seven evaluation datasets, MinerU2.5 achieves the best results in four datasets and one second-best result, demonstrating SOTA formula recognition capabilities. Specifically, on public datasets, MinerU2.5 achieves best CDM results of 96.4 on SCE and 90.6 on LaTeX-80M^M, showcasing leading performance in scenarios involving blurred screenshots and complex matrices. Besides, on CPE, HWE, and SPE, while being slightly outperformed by specialized formula recognition models, MinerU2.5 still deliver comparable performance. On in-house evaluation datasets, MinerU2.5’s performance in Chinese text recognition is on par with Qwen2.5-VL-72B, leading to a second-place result of 90.6. Meanwhile, MinerU2.5 achieves the best results on both the real-world mathematic documents (Fuzzy Math) and extremely hard formula recognition (Complex).

⁴<https://github.com/Oleehy0/TexTeller>

6 Conclusion

In this paper, we present MinerU2.5, a 1.2B-parameter vision-language model that achieves a new state-of-the-art in efficient document parsing through its innovative decoupled, coarse-to-fine strategy. By separating global layout analysis from local recognition, it delivers unprecedented accuracy in a lightweight model, effectively resolving the trade-off between performance and cost. Beyond its standalone capabilities, the primary significance of MinerU2.5 lies in its role as a foundational tool for the LLM era. Its ability to rapidly convert vast, unstructured document collections into clean, structured data is invaluable for curating high-quality pre-training corpora. Furthermore, by preserving the semantic integrity of tables, formulas, and layouts, it is poised to significantly enhance the quality and reliability of Retrieval-Augmented Generation (RAG) systems, unlocking the vast knowledge contained within complex documents for next-generation AI applications.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Haoli Bai, Zhiguang Liu, Xiaojun Meng, Wentao Li, Shuang Liu, Nian Xie, Rongfu Zheng, Liangwei Wang, Lu Hou, Jiansheng Wei, et al. Wukong-reader: Multi-modal pre-training for fine-grained visual document understanding. *arXiv preprint arXiv:2212.09621*, 2022.
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [4] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.
- [5] chatdoc.com. Ocrflux. <https://github.com/chatdoc-com/OCRFlux>, 2025. Accessed:2025-09-25.
- [6] Song Chen, Xinyu Guo, Yadong Li, Tao Zhang, Mingan Lin, Dongdong Kuang, Youwei Zhang, Lingfeng Ming, Fengyu Zhang, Yuran Wang, et al. Ocean-ocr: Towards general ocr application via a vision-language model. *arXiv preprint arXiv:2501.15558*, 2025.
- [7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- [8] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.
- [9] Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. Vision grid transformer for document layout analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19462–19472, 2023.
- [10] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n’pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36: 2252–2274, 2023.
- [11] Hao Feng, Shu Wei, Xiang Fei, Wei Shi, Yingdong Han, Lei Liao, Jinghui Lu, Binghong Wu, Qi Liu, Chunhui Lin, et al. Dolphin: Document image parsing via heterogeneous anchor prompting. *arXiv preprint arXiv:2505.14059*, 2025.
- [12] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024.
- [13] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.
- [14] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091, 2022.
- [15] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.
- [16] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- [17] Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. *arXiv preprint arXiv:2506.05218*, 2025.

- [18] Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R Manmatha, and Vijay Mahadevan. Doctr: Document transformer for structured information extraction in documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19584–19594, 2023.
- [19] Demiao Lin. Revolutionizing retrieval-augmented generation with enhanced pdf structure recognition. *arXiv preprint arXiv:2401.12599*, 2024.
- [20] Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. Hrvda: High-resolution visual document assistant. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15534–15545, 2024.
- [21] Hongen Liu, Cheng Cui, Yuning Du, Yi Liu, and Gang Pan. Pp-formulanet: Bridging accuracy and efficiency in advanced formula recognition. *arXiv preprint arXiv:2503.18382*, 2025.
- [22] Yuan Liu, Zhongyin Zhao, Le Tian, Haicheng Wang, Xubing Ye, Yangxiu You, Zilin Yu, Chuhan Wu, Xiao Zhou, Yang Yu, et al. Points-reader: Distillation-free adaptation of vision-language models for document conversion. *arXiv preprint arXiv:2509.01215*, 2025.
- [23] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.
- [24] Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, et al. Docling: An efficient open-source toolkit for ai-driven document conversion. *arXiv preprint arXiv:2501.17887*, 2025.
- [25] Maksym Lysak, Ahmed Nassar, Nikolaos Livathinos, Christoph Auer, and Peter Staar. Optimized table tokenization for table structure recognition. In *International Conference on Document Analysis and Recognition*, pages 37–50. Springer, 2023.
- [26] Souvik Mandalm. Nanonets-ocr-s. <https://nanonets.com/research/nanonets-ocr-s/>, 2025. Accessed:2025-09-25.
- [27] Mathpix. Mathpix. <https://mathpix.com/>, 2025. Accessed:2025-09-25.
- [28] Ahmed Nassar, Andres Marafioti, Matteo Omenetti, Maksym Lysak, Nikolaos Livathinos, Christoph Auer, Lucas Morin, Rafael Teixeira de Lima, Yusik Kim, A Said Gurbuz, et al. Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion. *arXiv preprint arXiv:2503.11576*, 2025.
- [29] OpenDataLab. Pdf-extract-kit. <https://github.com/opendatalab/PDF-Extract-Kit>, 2025. Accessed:2025-09-25.
- [30] Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24838–24848, 2025.
- [31] Vik Paruchuri. Marker. <https://github.com/datalab-to/marker>, 2025. Accessed:2025-09-25.
- [32] Vikas Paruchuri and Datalab Team. Surya: A lightweight document ocr and analysis toolkit. <https://github.com/VikParuchuri/surya>, 2025. Accessed:2025-09-25.
- [33] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. 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, 2022.
- [34] Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *arXiv preprint arXiv:2502.18443*, 2025.
- [35] RapidAI. Rapid table. <https://github.com/RapidAI/RapidTable>, 2024. Accessed: 2025-9-25.
- [36] rednote. dots.ocr: Multilingual document layout parsing in a single vision-language model. <https://github.com/rednote-hilab/dots.ocr>, 2025. Accessed:2025-09-25.

- [37] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.
- [38] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [39] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19254–19264, 2023.
- [40] Mistral AI Team. Mistral-ocr. https://mistral.ai/news/mistral-ocr?utm_source=ai-bot.cn, 2025. Accessed:2025-09-25.
- [41] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- [42] Jianqiang Wan, Sibao Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. Omniparser: A unified framework for text spotting key information extraction and table recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15641–15653, 2024.
- [43] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024.
- [44] Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. Unimernet: A universal network for real-world mathematical expression recognition. *arXiv preprint arXiv:2404.15254*, 2024.
- [45] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.
- [46] Bin Wang, Fan Wu, Linke Ouyang, Zhuangcheng Gu, Rui Zhang, Renqiu Xia, Botian Shi, Bo Zhang, and Conghui He. Image over text: Transforming formula recognition evaluation with character detection matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19681–19690, 2025.
- [47] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [48] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.
- [49] Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. Layoutreader: Pre-training of text and layout for reading order detection. *arXiv preprint arXiv:2108.11591*, 2021.
- [50] Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. Vrdu: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5184–5193, 2023.
- [51] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024.
- [52] Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, et al. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. *arXiv preprint arXiv:2412.02210*, 2024.
- [53] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- [54] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu

- Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.
- [55] Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. *arXiv preprint arXiv:2412.02592*, 2024.
- [56] Qintong Zhang, Bin Wang, Victor Shea-Jay Huang, Junyuan Zhang, Zhengren Wang, Hao Liang, Conghui He, and Wentao Zhang. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*, 2024.
- [57] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.
- [58] Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*, 2024.
- [59] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024.
- [60] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 697–706, 2021.
- [61] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer, 2020.
- [62] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

A Qualitative examples

This section presents qualitative examples illustrating the capabilities of the MinerU2.5 through document parsing outputs generated for various pages. This section is structured as follows: [Section A.1](#) illustrates the MinerU2.5’s performance on Document Parsing, Table Recognition and Formula Recognition among all types of documents. [Section A.2](#) showcases specific attribute pages with improved performance. [Section A.3](#) demonstrates MinerU2.5’s performance on some complex pages compared to other models.

Examples demonstrating the Document Parsing performance among PDF types are provided in [Figures 8 to 10](#), including Academic literature, Books, Textbooks, Research Report, Financial Report, Slides, Exam Paper, Note, Newspaper and Magazine.

Table Recognition performance among various types of tables are demonstrated in [Figures 11 and 12](#), including table with images, table with colorful background, table with formula, table with empty cells, hand-write table, large table, rotated table, no-line table, three-line table and full-line table.

The performance of Formula Recognition among types of formula are demonstrated in [Figures 13 and 14](#), including formula with background, formula with Chinese, formula with matrix, formula with condition and nested condition, hand-write formula, blurred formula, multi-column formula, degradation formula.

[Figures 15 to 18](#) demonstrate that MinerU2.5’s document parsing ability improved when encounter rotated tables, table with merged cells, formula with Chinese and multi-line and complex formula, comparing with previous version (MinerU2-VLM, MinerU2-pipeline). Moreover, MinerU2.5 achieve finer bounding bbox in layout detection and performs better in watermark pages than previous version, as illustrated in [Figures 19 and 20](#).

MinerU2.5 achieves outstanding performance in scenarios involving PDF pages with complex elements, and its performance is relatively better compared to existing state-of-the-art models.

[Figures 21 to 26](#) showcase the scenarios with complex tables in the page, including full-page table, content dense table, colorful table with amounts of empty cells, a tightly-arranged multiple table, table with irregular merged cells, a table without lines. MinerU2.5 can achieve better parsing outputs on these pages, while other models encounter errors such as table structure error, table structure lost, table content lost and table split error.

[Figures 27 to 29](#) illustrates the performance of MinerU2.5 in the page with nested conditional expressions, complex matrix and nested matrix compared to other SOTA models, MinerU2.5 can correctly parse the complex formula while others might generate wrong outputs.

[Figures 30 to 32](#) shows MinerU2.5’s outstanding performance in pages with complex layout, e.g., alternating texts and images, with very-few frame tables, and pages with watermark compared with others.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing

A.1 Overview

A.1.1 Among PDF types

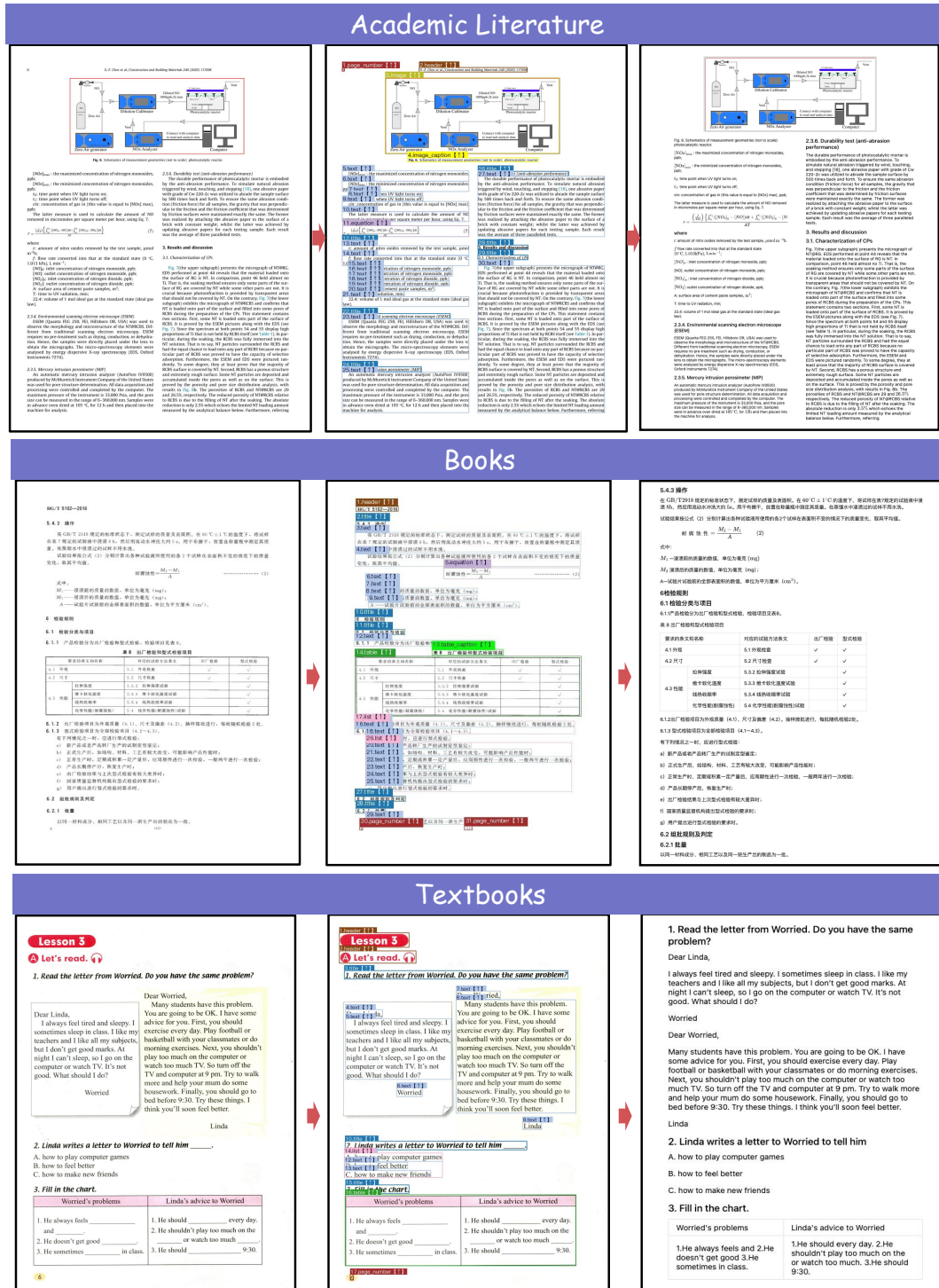


Figure 8: The Layout and rendered markdown output for Academic literature, Books, Textbooks.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing

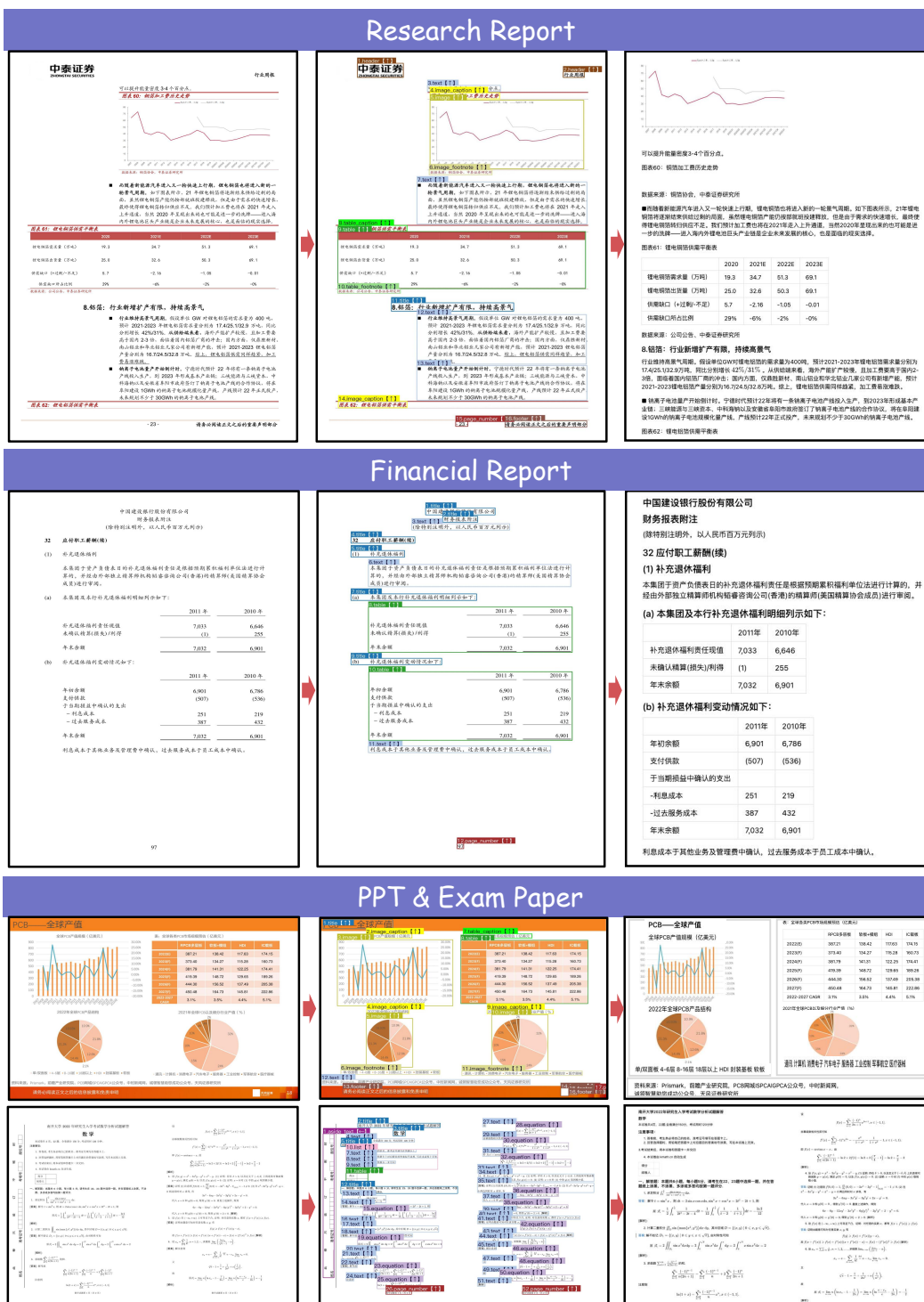


Figure 9: The Layout and rendered markdown output for Research Report, Financial Report, Slides and Exam Paper.



Figure 10: The Layout and rendered markdown output for Note, Newspaper and Magazine.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing

A.1.2 Among Table types

Table with Image					
g/L		B	1	2	3
NaOH			2	2	2
保险粉			2	2	2
SUNMORL WX-245			1	1	
NEOCRYSAL BC-650				1	
染料		Xg/L	Xg/L	Xg/L	Xg/L
AC-EK1g/L	初始				
UN-SE蓝1g/L	初始				
XF翠蓝1g/L	初始				

g/L		B	1	2	3
NaOH		2	2	2	2
保险粉			2	2	2
SUNMORL WX-245				1	1
NEOCRYSAL BC-650					1
染料		Xg/L	Xg/L	Xg/L	Xg/L
AC-EK1g/L	初始				
UN-SE蓝1g/L	初始				
XF翠蓝1g/L	初始				

Handwritten Table									
山西省	晋	太原	山西省	晋	太原				
内蒙古自治区	内蒙古	呼和浩特	内蒙古自治区	内蒙古	呼和浩特				
辽宁省	辽	沈阳	辽宁省	辽	沈阳				
吉林省	吉	长春	吉林省	吉	长春				
黑龙江省	黑	哈尔滨	黑龙江省	黑	哈尔滨				
上海市	沪	上海	上海市	沪	上海				
江苏省	苏	南京	江苏省	苏	南京				
浙江省	浙	杭州	浙江省	浙	杭州				
安徽省	皖	合肥	安徽省	皖	合肥				
广东省	粤	广州	广东省	粤	广州				
广西壮族自治区	桂	南宁	广西壮族自治区	桂	南宁				
海南省	琼	海口	海南省	琼	海口				
重庆市	渝	重庆	重庆市	渝	重庆				
四川省	川	成都	四川省	川	成都				
贵州省	贵	贵阳	贵州省	贵	贵阳				
云南省	云	昆明	云南省	云	昆明				
西藏自治区	藏	拉萨	西藏自治区	藏	拉萨				
陕西省	陕	西安	陕西省	陕	西安				
甘肃省	甘	兰州	甘肃省	甘	兰州				
青海省	青	西宁	青海省	青	西宁				
新疆维吾尔自治区	新	乌鲁木齐	新疆维吾尔自治区	新	乌鲁木齐				
山东省	鲁	济南	山东省	鲁	济南				
河南省	豫	郑州	河南省	豫	郑州				
Colorful Table									
日期/季度/年份	指标名称	T15	T10	T15					
2015-06-12	保费	724.67	805.37	1950.11					
2015-09-15	保费	391.78	419.74	891.21					
2015-12-22	保费	27.46	31.48	29.89					
2016-01-28	保费	265.92	142.04	99.04					
2017-11-13	保费	222.65	191.97	254.29					
2018-10-18	保费	568.83	1530.78	2009.05					
2019-04-10	保费	662.43	1465.46	1517.01					
2020-07-13	保费	8064.47	7902.07	6790.54					
2021-03-18	保费	2009.72	2852.76	2849.22					
2021-12-13	保费	5341.16	3579.57	3065.12					
2022-04-26	保费	2197.51	2264.42	3399.59					
2022-07-04	保费	2340.62	3688.16	4627.22					
2022-10-31	保费	6161.69	6618.07	5339.34					
2015-06-12	保费	724.67	805.37	1950.11					
2015-09-15	保费	391.78	419.74	891.21					
2015-12-22	保费	27.46	31.48	29.89					
2016-01-28	保费	265.92	142.04	99.04					
2017-11-13	保费	222.65	191.97	254.29					
2018-10-18	保费	568.83	1530.78	2009.05					
2019-04-10	保费	662.43	1465.46	1517.01					
2020-07-13	保费	8064.47	7902.07	6790.54					
2021-03-18	保费	2009.72	2852.76	2849.22					
2021-12-13	保费	5341.16	3579.57	3065.12					
2022-04-26	保费	2197.51	2264.42	3399.59					
2022-07-04	保费	2340.62	3688.16	4627.22					
2022-10-31	保费	6161.69	6618.07	5339.34					
Large Table									
014354.OF	东方财富九个月持有 A	偏债混合型	2022-05-09	2022-07-08	014354.OF	东方财富九个月持有 A	偏债混合型	2022-05-09	2022-07-08
014065.OF	长信永利收益一年持有 A	偏债混合型	2022-04-13	2022-07-12	014065.OF	长信永利收益一年持有 A	偏债混合型	2022-04-13	2022-07-12
014067.OF	国泰稳健收益一年持有	混合型 FOF	2022-04-18	2022-07-15	014067.OF	国泰稳健收益一年持有	混合型 FOF	2022-04-18	2022-07-15
014070.OF	汇添富均衡成长三个月持有 A	混合型 FOF	2022-04-18	2022-07-15	014070.OF	汇添富均衡成长三个月持有 A	混合型 FOF	2022-04-18	2022-07-15
011210.OF	新疆前海开源中证 1-3 年国开行 A	被动指数型	2022-04-20	2022-07-19	011210.OF	新疆前海开源中证 1-3 年国开行 A	被动指数型	2022-04-20	2022-07-19
011629.OF	银河核心优势	偏股混合型	2022-04-20	2022-07-19	011629.OF	银河核心优势	偏股混合型	2022-04-20	2022-07-19
015699.OF	平安均衡成长 2 年持有 A	偏股混合型	2022-05-06	2022-07-20	015699.OF	平安均衡成长 2 年持有 A	偏股混合型	2022-05-06	2022-07-20
014072.OF	汇安裕同 A	中长期纯债型	2022-04-21	2022-07-20	014072.OF	汇安裕同 A	中长期纯债型	2022-04-21	2022-07-20
014993.OF	西部利得聚优一年持有	偏债混合型	2022-06-02	2022-07-22	014993.OF	西部利得聚优一年持有	偏债混合型	2022-06-02	2022-07-22
014290.OF	南方顺利一年定开	中长期纯债型	2022-04-25	2022-07-22	014290.OF	南方顺利一年定开	中长期纯债型	2022-04-25	2022-07-22
561153.OF	富国中证 500ESG 基准 ETF	被动指数型	2022-04-26	2022-07-25	561153.OF	富国中证 500ESG 基准 ETF	被动指数型	2022-04-26	2022-07-25
014462.OF	光大保德信汇信 A	偏股混合型	2022-05-25	2022-07-26	014462.OF	光大保德信汇信 A	偏股混合型	2022-05-25	2022-07-26
014313.OF	新华创新增长一年持有 A	偏股混合型	2022-04-27	2022-07-26	014313.OF	新华创新增长一年持有 A	偏股混合型	2022-04-27	2022-07-26
513223.OF	招商中证全球中国互联网 ETF	国际 (QDII) 股票型	2022-05-27	2022-07-27	513223.OF	招商中证全球中国互联网 ETF	国际 (QDII) 股票型	2022-05-27	2022-07-27
014408.OF	创金合信优选产业国际一年封闭 A	偏股混合型	2022-04-29	2022-07-28	014408.OF	创金合信优选产业国际一年封闭 A	偏股混合型	2022-04-29	2022-07-28
011637.OF	前海开源金壹顺 A	偏股混合型	2022-04-29	2022-07-28	011637.OF	前海开源金壹顺 A	偏股混合型	2022-04-29	2022-07-28
011964.OF	国泰中证 500 联接 A	被动指数型	2022-04-29	2022-07-28	011964.OF	国泰中证 500 联接 A	被动指数型	2022-04-29	2022-07-28
012328.OF	天弘中证新能源指数增强 A	增强指数型	2022-04-29	2022-07-28	012328.OF	天弘中证新能源指数增强 A	增强指数型	2022-04-29	2022-07-28
014288.OF	淳厚添丰 A	中长期纯债型	2022-04-29	2022-07-28	014288.OF	淳厚添丰 A	中长期纯债型	2022-04-29	2022-07-28
014343.OF	泰康丰盛纯债一年定开	中长期纯债型	2022-04-29	2022-07-28	014343.OF	泰康丰盛纯债一年定开	中长期纯债型	2022-04-29	2022-07-28
501220.OF	国泰行业轮动一年封闭 A	股票型 FOF	2022-04-29	2022-07-28	501220.OF	国泰行业轮动一年封闭 A	股票型 FOF	2022-04-29	2022-07-28
513293.OF	汇添富纳斯达克生物科技 ETF	国际 (QDII) 股票型	2022-05-27	2022-07-29	513293.OF	汇添富纳斯达克生物科技 ETF	国际 (QDII) 股票型	2022-05-27	2022-07-29
014311.OF	大成优质精选 A	偏股混合型	2022-05-06	2022-08-05	014311.OF	大成优质精选 A	偏股混合型	2022-05-06	2022-08-05
159621.OF	博时中证成渝地区双城经济圈成份 ETF	被动指数型	2022-05-10	2022-08-09	159621.OF	博时中证成渝地区双城经济圈成份 ETF	被动指数型	2022-05-10	2022-08-09
015578.OF	南方宝晋 A	偏债混合型	2022-05-11	2022-08-10	015578.OF	南方宝晋 A	偏债混合型	2022-05-11	2022-08-10
014466.OF	汇添富添富一年定开 A	中长期纯债型	2022-05-12	2022-08-11	014466.OF	汇添富添富一年定开 A	中长期纯债型	2022-05-12	2022-08-11
014613.OF	南方月利 6 个月持有 A	中长期纯债型	2022-05-16	2022-08-12	014613.OF	南方月利 6 个月持有 A	中长期纯债型	2022-05-16	2022-08-12
014384.OF	国投瑞银稳健一年定开	中长期纯债型	2022-05-13	2022-08-12	014384.OF	国投瑞银稳健一年定开	中长期纯债型	2022-05-13	2022-08-12
014443.OF	汇丰晋信丰盈 A	中长期纯债型	2022-05-13	2022-08-12	014443.OF	汇丰晋信丰盈 A	中长期纯债型	2022-05-13	2022-08-12
014374.OF	景顺长城稳健平衡养老目标三年持有	混合型 FOF	2022-05-13	2022-08-12	014374.OF	景顺长城稳健平衡养老目标三年持有	混合型 FOF	2022-05-13	2022-08-12
014391.OF	华安添禧	中长期纯债型	2022-05-12	2022-08-12	014391.OF	华安添禧	中长期纯债型	2022-05-12	2022-08-12
014387.OF	光大保德信永利一年定开	中长期纯债型	2022-05-16	2022-08-15	014387.OF	光大保德信永利一年定开	中长期纯债型	2022-05-16	2022-08-15
014388.OF	渤海汇金汇宸一年定开	中长期纯债型	2022-05-16	2022-08-15	014388.OF	渤海汇金汇宸一年定开	中长期纯债型	2022-05-16	2022-08-15
014482.OF	华夏盛颐可持续一年持有 A	偏股混合型	2022-05-18	2022-08-17	014482.OF	华夏盛颐可持续一年持有 A	偏股混合型	2022-05-18	2022-08-17
014710.OF	平安惠利纯债 A	中长期纯债型	2022-05-24	2022-08-19	014710.OF	平安惠利纯债 A	中长期纯债型	2022-05-24	2022-08-19
014433.OF	国泰智慧科技 1 个月滚动 A	偏股混合型	2022-05-20	2022-08-19	014433.OF	国泰智慧科技 1 个月滚动 A	偏股混合型	2022-05-20	2022-08-19
014474.OF	中欧安恒一年定开	中长期纯债型	2022-05-20	2022-08-19	014474.OF	中欧安恒一年定开	中长期纯债型	2022-05-20	2022-08-19
014492.OF	天弘惠享一年定开	混合型 FOF	2022-05-24	2022-08-23	014492.OF	天弘惠享一年定开	混合型 FOF	2022-05-24	2022-08-23
014510.OF	国信永丰聚兴一年定开	中长期纯债型	2022-05-24	2022-08-23	014510.OF	国信永丰聚兴一年定开	中长期纯债型	2022-05-24	2022-08-23
159621.OF	国泰 CSI 中国 A 股 ESG 通用 ETF	被动指数型	2022-05-24	2022-08-23	159621.OF	国泰 CSI 中国 A 股 ESG 通用 ETF	被动指数型	2022-05-24	2022-08-23
014948.OF	中保投资景润 3 个月定开 A	中长期纯债型	2022-05-24	2022-08-23	014948.OF	中保投资景润 3 个月定开 A	中长期纯债型	2022-05-24	2022-08-23

Figure 11: The rendered outputs for various types of Tables.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing

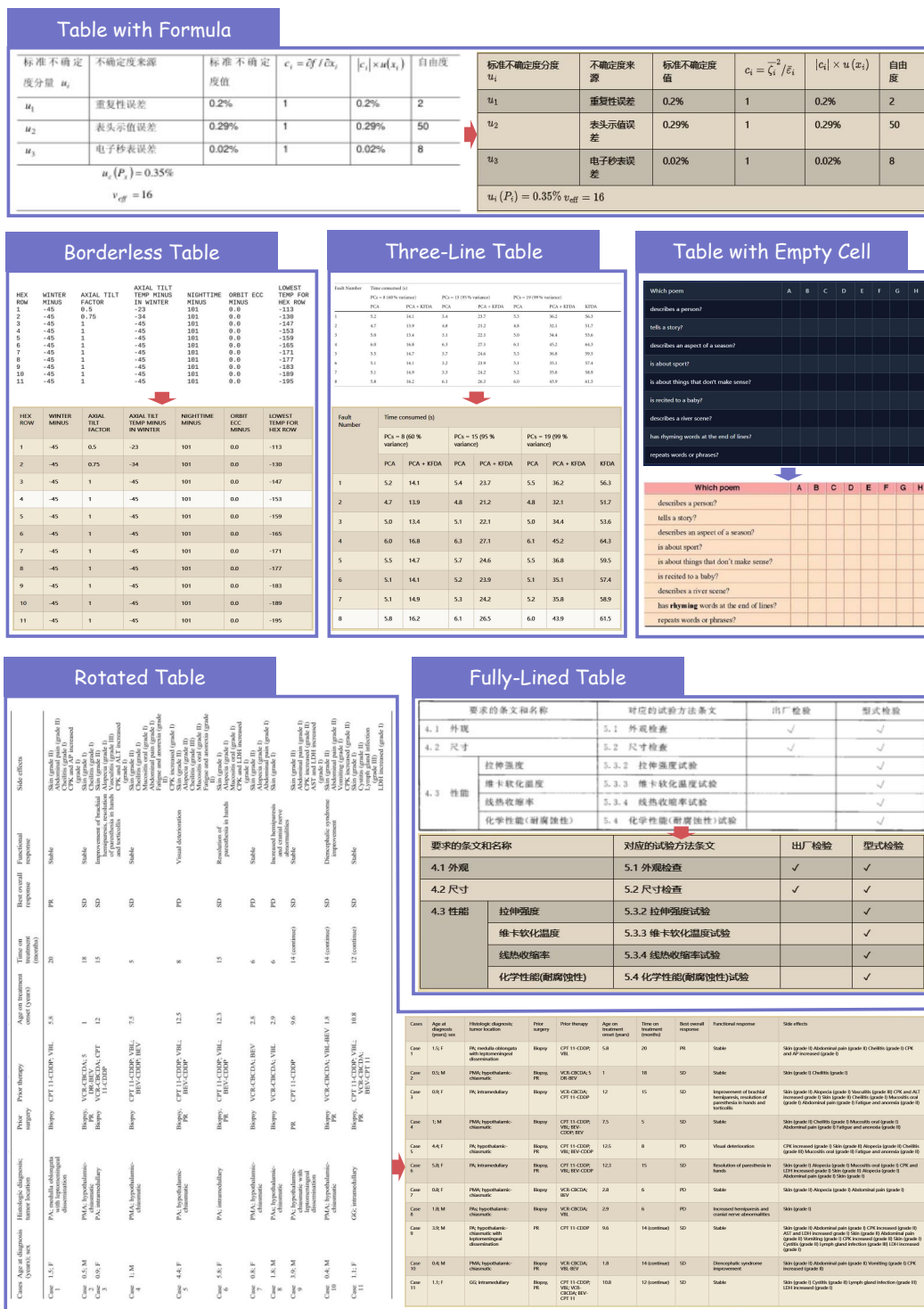
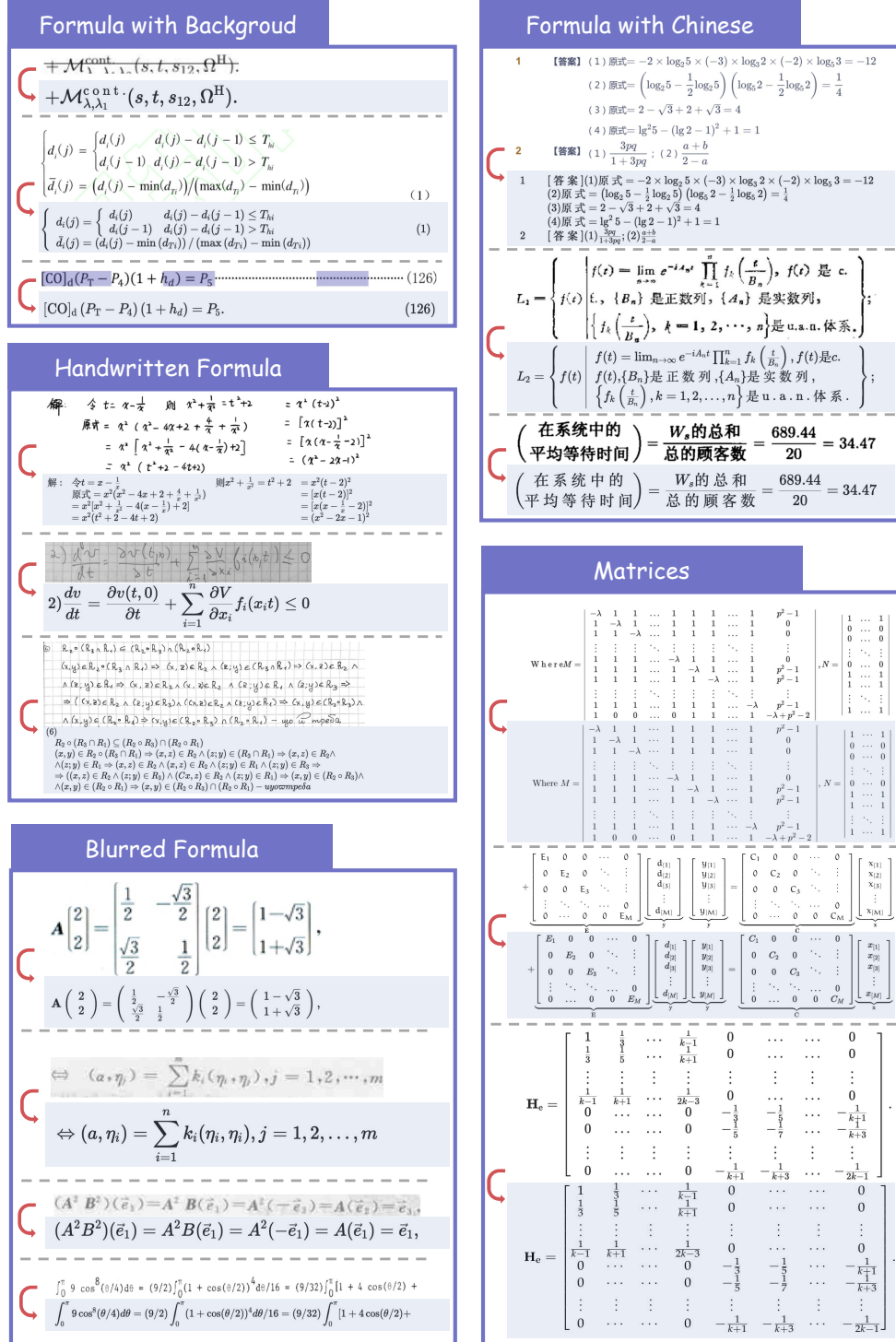


Figure 12: The rendered outputs for various types of Tables.

A.1.3 Among Formula types



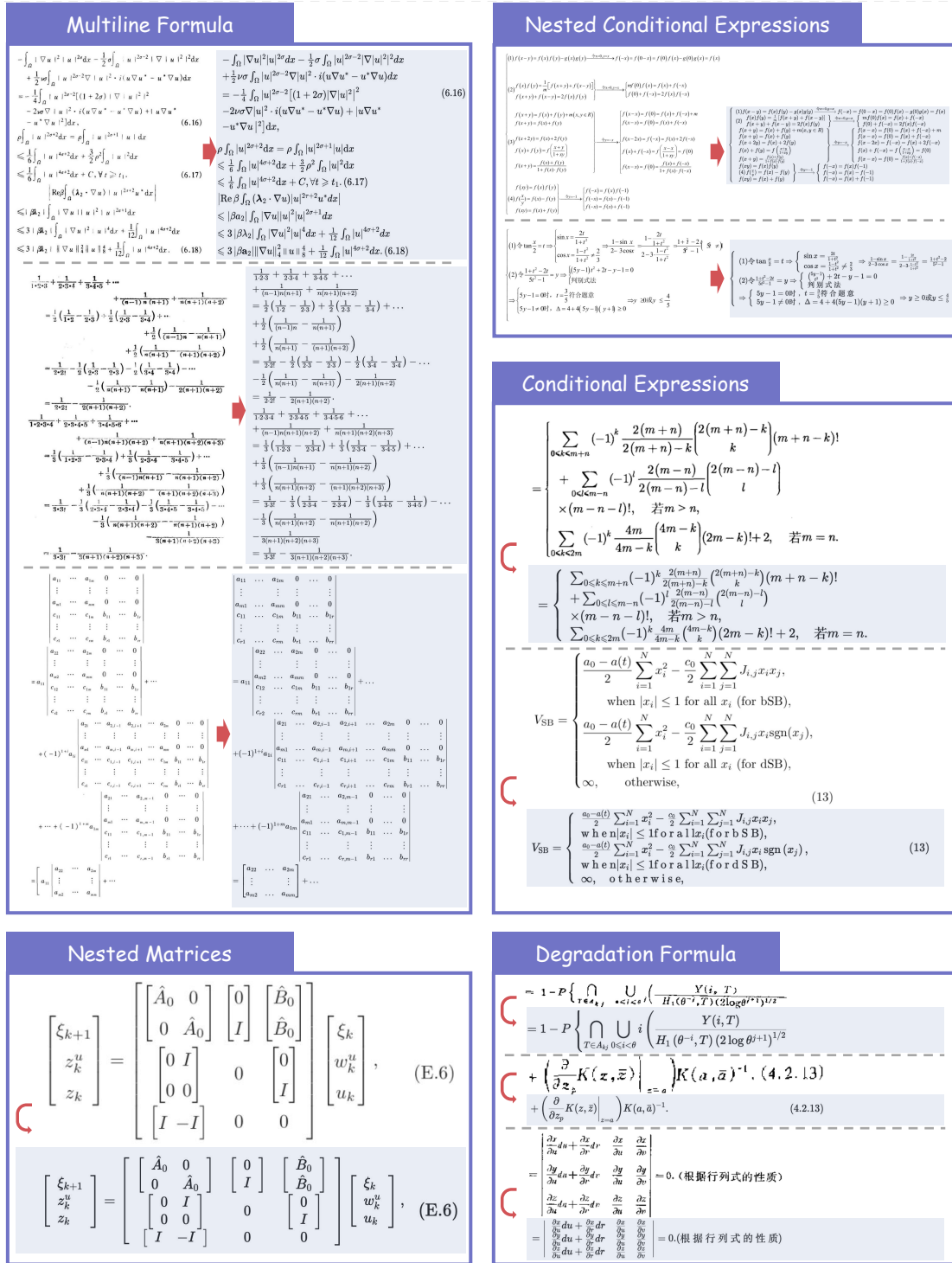


Figure 14: The rendered outputs for various types of Formulas.

A.2 Compare to Previous Versions

A.2.1 Table

Image

1216

H. S. White et al.

AED	MES test		Pentylenetetrazol		Bicuculline		Picrotoxin		Strychnine	
	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI
Rufamide	>1000	>1000	15.5 (12.5-18.1)	>3.2	54.0 (38.7-74.9)	>4.5	78.3 (64.8-96.7)	>4.5	95.3 (71.5-125)	NA
Phenoin	65.5 (52.5-79.5)	9.5 (8.1-10.4)	6.9	300 no protection	<0.2	100 no protection	<0.7	100 no protection	<0.7	55-100b
Phenobarbital	68.0 (52.5-72.5)	21.8 (16.0-29.0)	3.2	13.2 (8.9-19.0)	5.2	37.7 (26.5-47.4)	1.8	27.5 (18.8-38.8)	2.5	90.3 (61.5-125)
Valproate	428.8 (289-650)	272 (247-298)	1.6	148.6 (123-177)	2.9	358.9 (284-438)	1.2	387.2 (311-464)	1.1	282.9 (231-333)
Ethosuximide	448.8 (289-650)	1,000 no protection	<0.4	130.3 (111-151)	3.4	459.9 (387-531)	1.0	243 (228-258)	1.8	250.1 (201-300)

Rank order for relative potencies of the intraperitoneal administration of AEDs in the pentylenetetrazol-induced tonic seizure test in mice was phenobarbital > ethosuximide > phenoin > strychnine > valproate > pentylenetetrazol (Table 1). Phenoin was ineffective up to a dose of 300 mg/kg. In slight contrast, the rank order for relative anticonvulsant potency in this test with oral administration in mice was phenobarbital > ethosuximide > valproate > pentylenetetrazol > phenoin (Table 2).

Oral rufamide (>1,000 mg/kg) and phenoin (800 mg/kg) did not inhibit pentylenetetrazol-induced seizures in rats (Table 3). Phenobarbital achieved the best anticonvulsant potency of the remaining three AEDs.

Bicuculline-, picrotoxin-, and strychnine-induced seizure tests in mice

Intraperitoneal rufamide was effective at nontoxic doses in the bicuculline and picrotoxin clonic seizure tests (ED₅₀ < 50-75 mg/kg) and showed partial protection from strychnine-induced tonic seizures (37.5% protection; Table 1). Overall, the general order of potency in these chemically induced tonic seizure tests was phenobarbital > ethosuximide > valproate > pentylenetetrazol > phenoin (Table 1). In the strychnine-induced tonic seizure test, phenoin had the lowest ED₅₀ value, suggesting the greatest potency. However, it is important to note that 50% protection was the maximum achieved with this AED. Phenoin failed to provide protection against bicuculline- and picrotoxin-induced tonic seizures in mice.

Evaluation of behavioral toxicity in mice

The median toxic dose of intraperitoneal rufamide (TD₅₀) in the rotarod test of behavioral impairment was 500-1,000 mg/kg. The TD₅₀ for rufamide was higher than that for comparator AEDs (Table 1), indicating a lower toxicity. Higher doses of rufamide were not assessed due to the low ED₅₀ values and high protective index (>40).

Neurological side effects of very high-dose intraperitoneal rufamide (1,000 mg/kg, n = 2; ED₅₀ 15-100 mg/kg) included decreased motor activity, ataxia, muscle relaxation, decreased respiration, and death (one animal died; the other appeared normal). Higher doses of the comparator drugs induced increased side effects (> 2 x TD₅₀) and resulted in death (all animals) in 3-24 h (4 x TD₅₀). The safety ratio for rufamide in mice (TD₅₀/ED₅₀ > 19.2 [interpretation] and > 23.8 [oral]) was considerably greater than for phenoin, phenobarbital, and valproate (Table 4).

The median dose of intraperitoneal rufamide required to produce loss of righting reflex in mice (ED₅₀) was >500 and <1,000 mg/kg (Table 5). Rufamide had a numerically greater ED₅₀ value than phenoin or phenobarbital (ED₅₀ values of 178 and 135 mg/kg, respectively). Rufamide, ethosuximide, and valproate had comparable

Epilepsia, 49(7):1215-1230, 2008
doi:10.1111/j.1528-1482.2008.01852.x

MinerU2.5

1216

H. S. White et al.

AED	MES test		Pentylenetetrazol		Bicuculline		Picrotoxin		Strychnine	
	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI
Rufamide	>1000	>1000	15.5 (12.5-18.1)	>3.2	54.0 (38.7-74.9)	>4.5	78.3 (64.8-96.7)	>4.5	95.3 (71.5-125)	NA
Phenoin	65.5 (52.5-79.5)	9.5 (8.1-10.4)	6.9	300 no protection	<0.2	100 no protection	<0.7	100 no protection	<0.7	55-100b
Phenobarbital	68.0 (52.5-72.5)	21.8 (16.0-29.0)	3.2	13.2 (8.9-19.0)	5.2	37.7 (26.5-47.4)	1.8	27.5 (18.8-38.8)	2.5	90.3 (61.5-125)
Valproate	428.8 (289-650)	272 (247-298)	1.6	148.6 (123-177)	2.9	358.9 (284-438)	1.2	387.2 (311-464)	1.1	282.9 (231-333)
Ethosuximide	448.8 (289-650)	1,000 no protection	<0.4	130.3 (111-151)	3.4	459.9 (387-531)	1.0	243 (228-258)	1.8	250.1 (201-300)

Rank order for relative potencies of the intraperitoneal administration of AEDs in the pentylenetetrazol-induced tonic seizure test in mice was phenobarbital > ethosuximide > phenoin > strychnine > valproate > pentylenetetrazol (Table 1). Phenoin was ineffective up to a dose of 300 mg/kg. In slight contrast, the rank order for relative anticonvulsant potency in this test with oral administration in mice was phenobarbital > ethosuximide > valproate > pentylenetetrazol > phenoin (Table 2).

Oral rufamide (> 1,000 mg/kg) and phenoin (800 mg/kg) did not inhibit pentylenetetrazol-induced seizures in rats (Table 3). Phenobarbital achieved the best anticonvulsant potency of the remaining three AEDs.

Bicuculline-, picrotoxin-, and strychnine-induced seizure tests in mice

Intraperitoneal rufamide was effective at nontoxic doses in the bicuculline and picrotoxin clonic seizure tests (ED₅₀ < 50-75 mg/kg) and showed partial protection from strychnine-induced tonic seizures (37.5% protection; Table 1). Overall, the general order of potency in these chemically induced seizure tests was phenobarbital > ethosuximide > valproate > pentylenetetrazol > phenoin (Table 1). In the strychnine-induced tonic seizure test, phenoin had the lowest ED₅₀ value, suggesting the greatest potency. However, it is important to note that 50% protection was the maximum achieved with this AED. Phenoin failed to provide protection against bicuculline- and picrotoxin-induced tonic seizures in mice.

Evaluation of behavioral toxicity in mice

The median toxic dose of intraperitoneal rufamide (TD₅₀) in the rotarod test of behavioral impairment was 500-1,000 mg/kg. The TD₅₀ for rufamide was higher than that for comparator AEDs (Table 1), indicating a lower toxicity. Higher doses of rufamide were not assessed due to the low ED₅₀ values and high protective index (>40).

Neurological side effects of very high-dose intraperitoneal rufamide (1,000 mg/kg, n = 2; ED₅₀ 15-100 mg/kg) included decreased motor activity, ataxia, muscle relaxation, decreased respiration, and death (one animal died; the other appeared normal). Higher doses of the comparator drugs induced increased side effects (> 2 x TD₅₀) and resulted in death (all animals) in 3-24 h (4 x TD₅₀). The safety ratio for rufamide in mice (TD₅₀/ED₅₀ > 19.2 [interpretation] and > 23.8 [oral]) was considerably greater than for phenoin, phenobarbital, and valproate (Table 4).

The median dose of intraperitoneal rufamide required to produce loss of righting reflex in mice (ED₅₀) was >500 and <1,000 mg/kg (Table 5). Rufamide had a numerically greater ED₅₀ value than phenoin or phenobarbital (ED₅₀ values of 178 and 135 mg/kg, respectively). Rufamide, ethosuximide, and valproate had comparable

Epilepsia, 49(7):1215-1230, 2008
doi:10.1111/j.1528-1482.2008.01852.x

MinerU2-pipeline

Table Structure Error

AED	MES test		Pentylenetetrazol		Bicuculline		Picrotoxin		Strychnine	
	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI
Rufamide	>1000	>1000	15.5 (12.5-18.1)	>3.2	54.0 (38.7-74.9)	>4.5	78.3 (64.8-96.7)	>4.5	95.3 (71.5-125)	NA
Phenoin	65.5 (52.5-79.5)	9.5 (8.1-10.4)	6.9	300 no protection	<0.2	100 no protection	<0.7	100 no protection	<0.7	55-100b
Phenobarbital	68.0 (52.5-72.5)	21.8 (16.0-29.0)	3.2	13.2 (8.9-19.0)	5.2	37.7 (26.5-47.4)	1.8	27.5 (18.8-38.8)	2.5	90.3 (61.5-125)
Valproate	428.8 (289-650)	272 (247-298)	1.6	148.6 (123-177)	2.9	358.9 (284-438)	1.2	387.2 (311-464)	1.1	282.9 (231-333)
Ethosuximide	448.8 (289-650)	1,000 no protection	<0.4	130.3 (111-151)	3.4	459.9 (387-531)	1.0	243 (228-258)	1.8	250.1 (201-300)

Rank order for relative potencies of the intraperitoneal administration of AEDs in the pentylenetetrazol-induced tonic seizure test in mice was phenobarbital > ethosuximide > phenoin > strychnine > valproate > pentylenetetrazol (Table 1). Phenoin was ineffective up to a dose of 300 mg/kg. In slight contrast, the rank order for relative anticonvulsant potency in this test with oral administration in mice was phenobarbital > ethosuximide > valproate > pentylenetetrazol > phenoin (Table 2).

Oral rufamide (> 1,000 mg/kg) and phenoin (800 mg/kg) did not inhibit pentylenetetrazol-induced seizures in rats (Table 3). Phenobarbital achieved the best anticonvulsant potency of the remaining three AEDs.

Bicuculline-, picrotoxin-, and strychnine-induced seizure tests in mice

Intraperitoneal rufamide was effective at nontoxic doses in the bicuculline and picrotoxin clonic seizure tests (ED₅₀ < 50-75 mg/kg) and showed partial protection from strychnine-induced tonic seizures (37.5% protection; Table 1). Overall, the general order of potency in these chemically induced tonic seizure tests was phenobarbital > ethosuximide > valproate > pentylenetetrazol > phenoin (Table 1). In the strychnine-induced tonic seizure test, phenoin had the lowest ED₅₀ value, suggesting the greatest potency. However, it is important to note that 50% protection was the maximum achieved with this AED. Phenoin failed to provide protection against bicuculline- and picrotoxin-induced tonic seizures in mice.

Evaluation of behavioral toxicity in mice

The median toxic dose of intraperitoneal rufamide (TD₅₀) in the rotarod test of behavioral impairment was 500-1,000 mg/kg. The TD₅₀ for rufamide was higher than that for comparator AEDs (Table 1), indicating a lower toxicity. Higher doses of rufamide were not assessed due to the low ED₅₀ values and high protective index (>40).

Neurological side effects of very high-dose intraperitoneal rufamide (1,000 mg/kg, n = 2; ED₅₀ 15-100 mg/kg) included decreased motor activity, ataxia, muscle relaxation, decreased respiration, and death (one animal died; the other appeared normal). Higher doses of the comparator drugs induced increased side effects (> 2 x TD₅₀) and resulted in death (all animals) in 3-24 h (4 x TD₅₀). The safety ratio for rufamide in mice (TD₅₀/ED₅₀ > 19.2 [interpretation] and > 23.8 [oral]) was considerably greater than for phenoin, phenobarbital, and valproate (Table 4).

The median dose of intraperitoneal rufamide required to produce loss of righting reflex in mice (ED₅₀) was >500 and <1,000 mg/kg (Table 5). Rufamide had a numerically greater ED₅₀ value than phenoin or phenobarbital (ED₅₀ values of 178 and 135 mg/kg, respectively). Rufamide, ethosuximide, and valproate had comparable

MinerU2-VLM

Table Repeat

AED	MES test		Pentylenetetrazol		Bicuculline		Picrotoxin		Strychnine	
	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI	ED ₅₀ (95% CI)	PI
Rufamide	>1000	>1000	15.5 (12.5-18.1)	>3.2	54.0 (38.7-74.9)	>4.5	78.3 (64.8-96.7)	>4.5	95.3 (71.5-125)	NA
Phenoin	65.5 (52.5-79.5)	9.5 (8.1-10.4)	6.9	300 no protection	<0.2	100 no protection	<0.7	100 no protection	<0.7	55-100b
Phenobarbital	68.0 (52.5-72.5)	21.8 (16.0-29.0)	3.2	13.2 (8.9-19.0)	5.2	37.7 (26.5-47.4)	1.8	27.5 (18.8-38.8)	2.5	90.3 (61.5-125)
Valproate	428.8 (289-650)	272 (247-298)	1.6	148.6 (123-177)	2.9	358.9 (284-438)	1.2	387.2 (311-464)	1.1	282.9 (231-333)
Ethosuximide	448.8 (289-650)	1,000 no protection	<0.4	130.3 (111-151)	3.4	459.9 (387-531)	1.0	243 (228-258)	1.8	250.1 (201-300)

Rank order for relative potencies of the intraperitoneal administration of AEDs in the pentylenetetrazol-induced tonic seizure test in mice was phenobarbital > ethosuximide > phenoin > strychnine > valproate > pentylenetetrazol (Table 1). Phenoin was ineffective up to a dose of 300 mg/kg. In slight contrast, the rank order for relative anticonvulsant potency in this test with oral administration in mice was phenobarbital > ethosuximide > valproate > pentylenetetrazol > phenoin (Table 2).

Oral rufamide (> 1,000 mg/kg) and phenoin (800 mg/kg) did not inhibit pentylenetetrazol-induced seizures in rats (Table 3). Phenobarbital achieved the best anticonvulsant potency of the remaining three AEDs.

Bicuculline-, picrotoxin-, and strychnine-induced seizure tests in mice

Intraperitoneal rufamide was effective at nontoxic doses in the bicuculline and picrotoxin clonic seizure tests (ED₅₀ < 50-75 mg/kg) and showed partial protection from strychnine-induced tonic seizures (37.5% protection; Table 1). Overall, the general order of potency in these chemically induced tonic seizure tests was phenobarbital > ethosuximide > valproate > pentylenetetrazol > phenoin (Table 1). In the strychnine-induced tonic seizure test, phenoin had the lowest ED₅₀ value, suggesting the greatest potency. However, it is important to note that 50% protection was the maximum achieved with this AED. Phenoin failed to provide protection against bicuculline- and picrotoxin-induced tonic seizures in mice.

Evaluation of behavioral toxicity in mice

The median toxic dose of intraperitoneal rufamide (TD₅₀) in the rotarod test of behavioral impairment was 500-1,000 mg/kg. The TD₅₀ for rufamide was higher than that for comparator AEDs (Table 1), indicating a lower toxicity. Higher doses of rufamide were not assessed due to the low ED₅₀ values and high protective index (>40).

Neurological side effects of very high-dose intraperitoneal rufamide (1,000 mg/kg, n = 2; ED₅₀ 15-100 mg/kg) included decreased motor activity, ataxia, muscle relaxation, decreased respiration, and death (one animal died; the other appeared normal). Higher doses of the comparator drugs induced increased side effects (> 2 x TD₅₀) and resulted in death (all animals) in 3-24 h (4 x TD₅₀). The safety ratio for rufamide in mice (TD₅₀/ED₅₀ > 19.2 [interpretation] and > 23.8 [oral]) was considerably greater than for phenoin, phenobarbital, and valproate (Table 4).

The median dose of intraperitoneal rufamide required to produce loss of righting reflex in mice (ED₅₀) was >500 and <1,000 mg/kg (Table 5). Rufamide had a numerically greater ED₅₀ value than phenoin or phenobarbital (ED₅₀ values of 178 and 135 mg/kg, respectively). Rufamide, ethosuximide, and valproate had comparable

Figure 15: Compare with Previous Version, MinerU2.5 performs better in rotated tables.

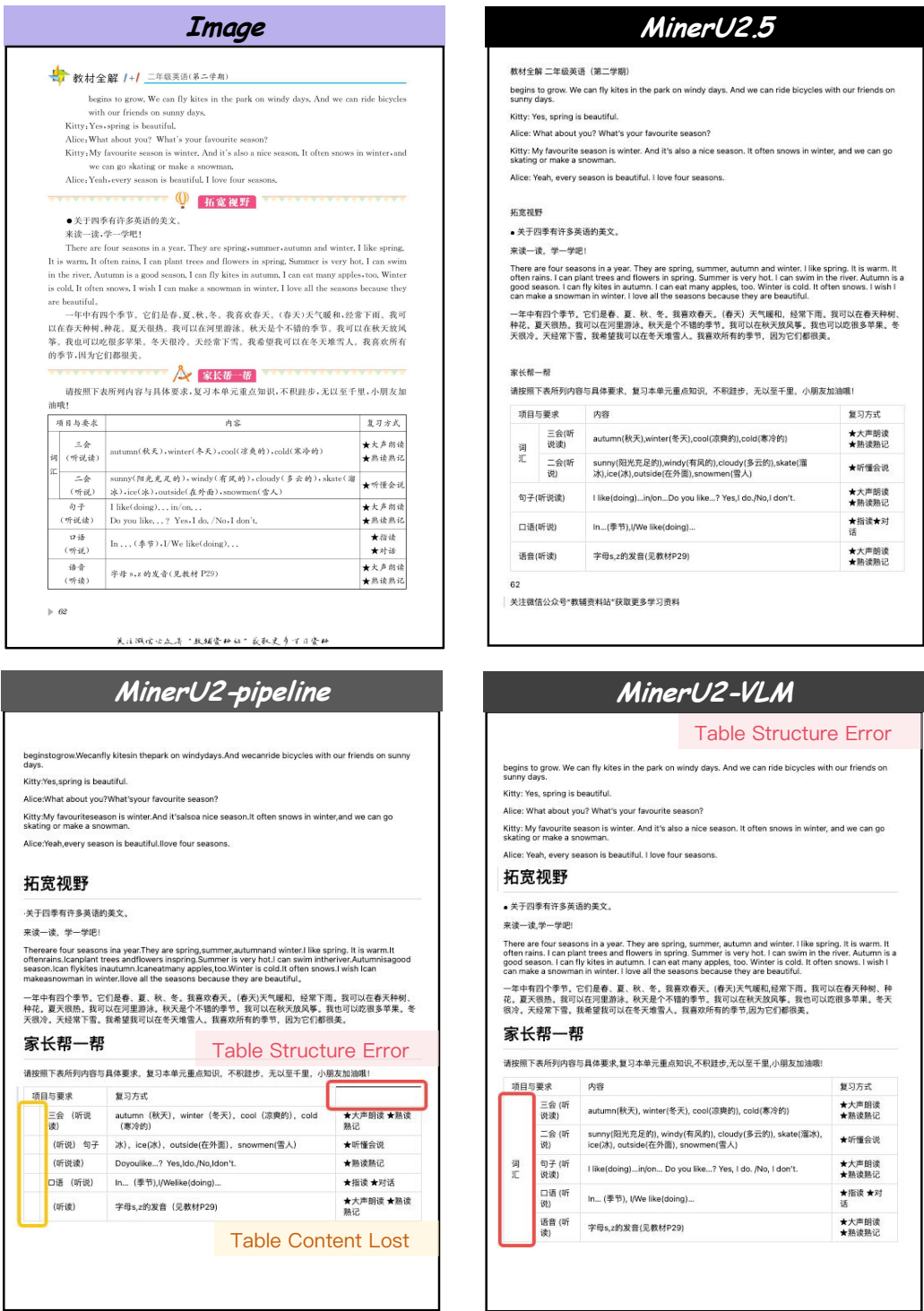


Figure 16: Compare with Previous Version, MinerU2.5 performs better in tables with merged cells.

A.2.2 Formula

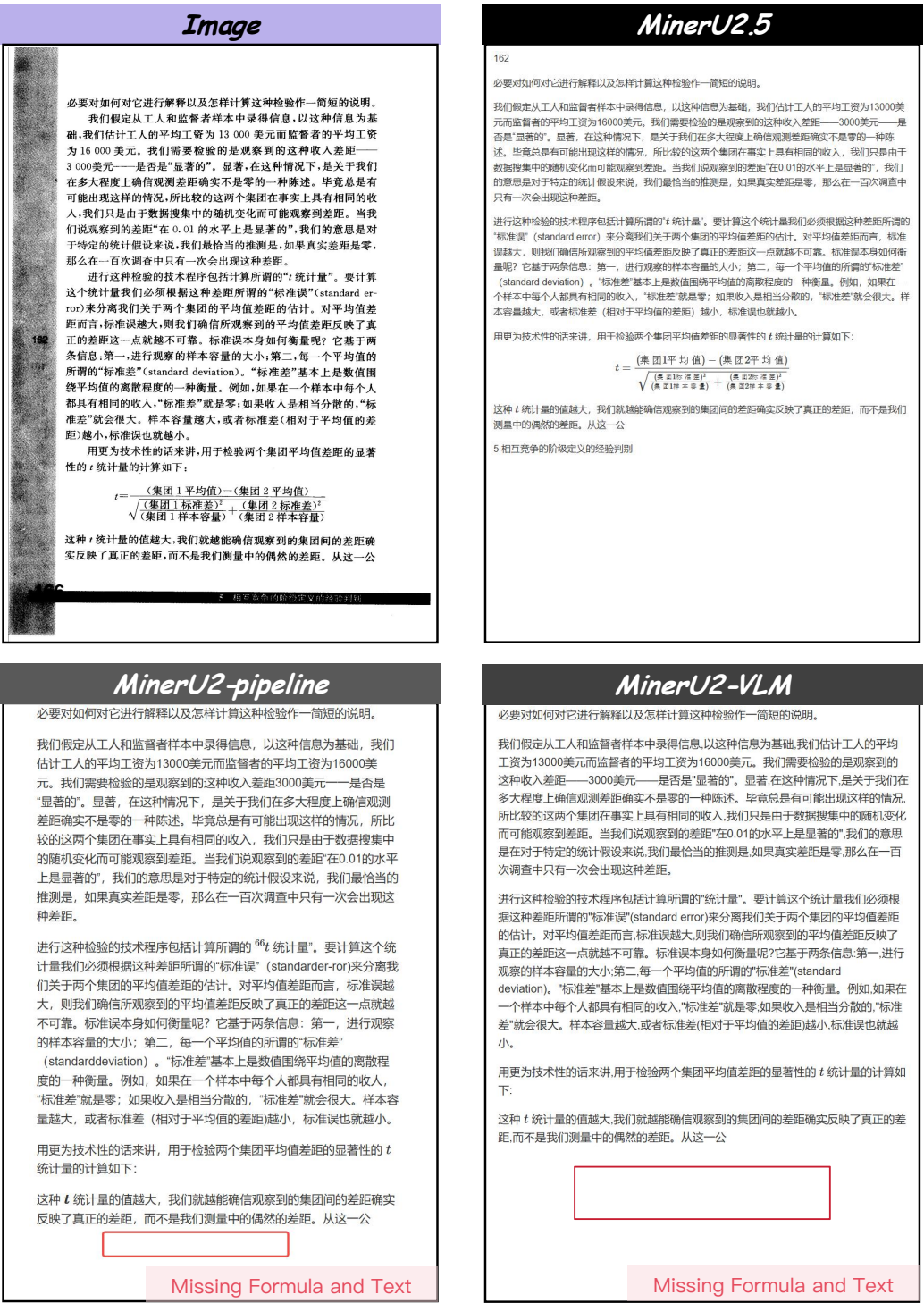


Figure 17: Compare with Previous Version, MinerU2.5 performs better in Formula with Chinese.

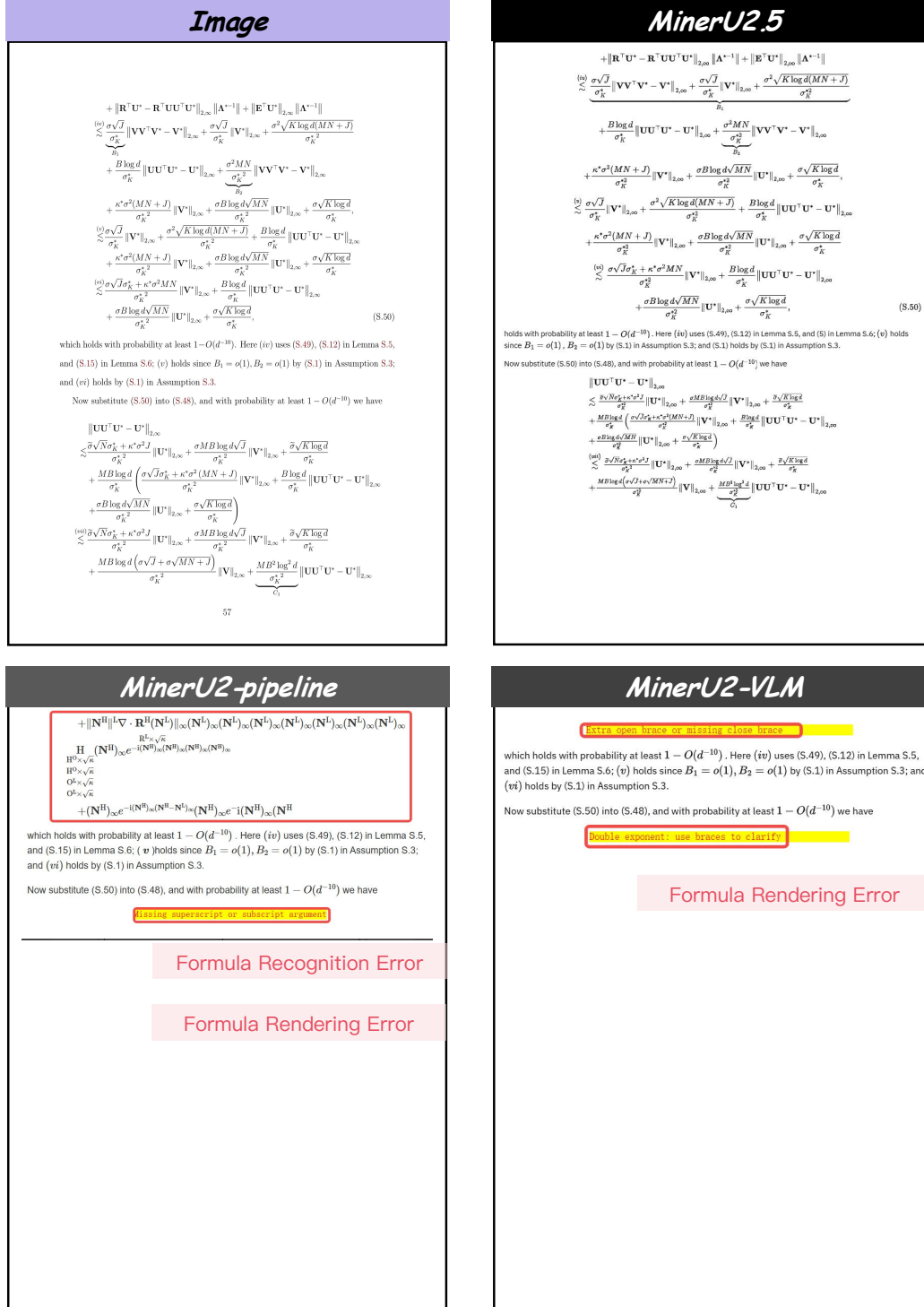


Figure 18: Compare with Previous Version, MinerU2.5 performs better in multi-lines and complex Formula.

A.2.3 Layout&OCR

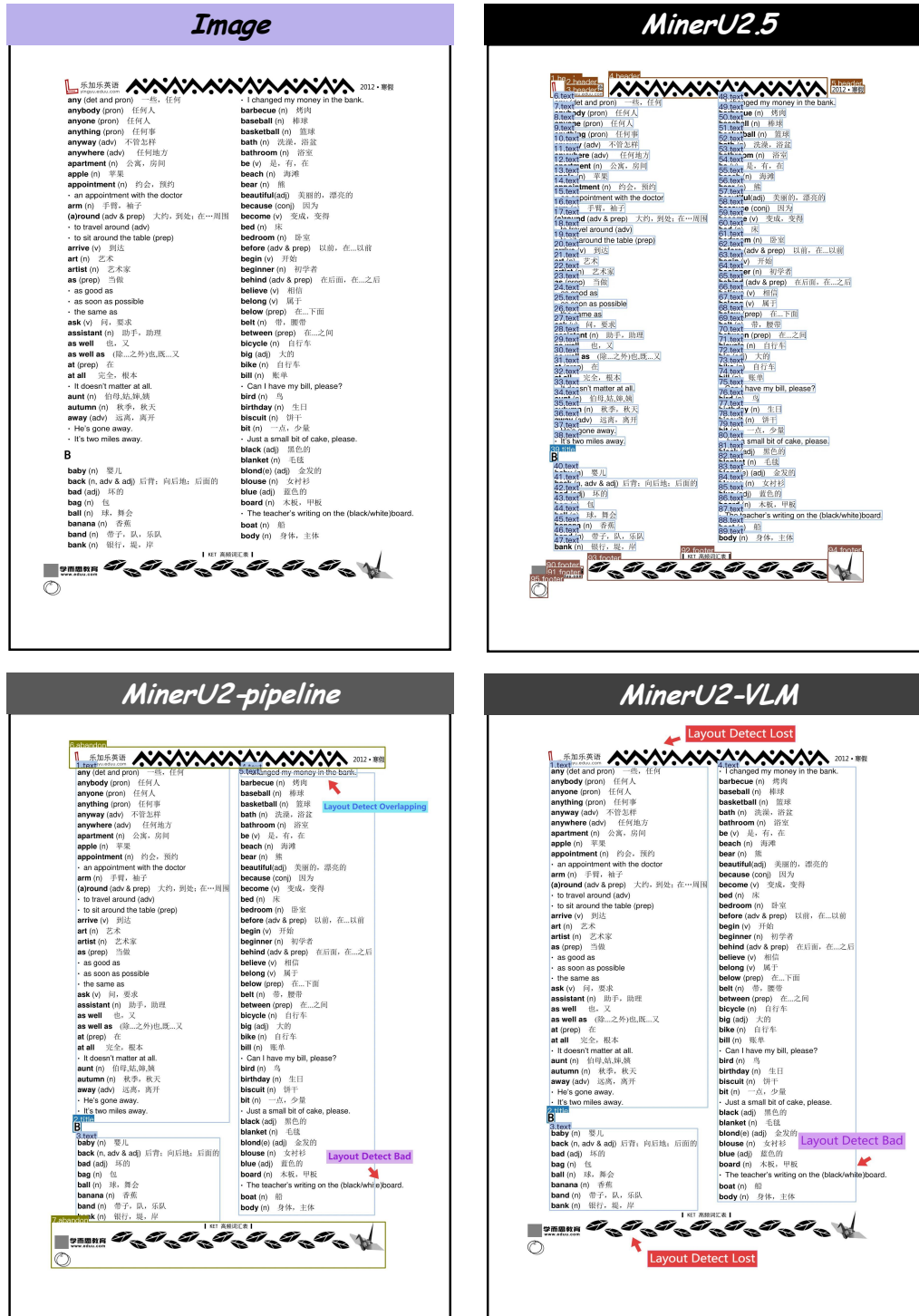


Figure 19: Compare with Previous Version, MinerU2.5 achieve finer layout detection.

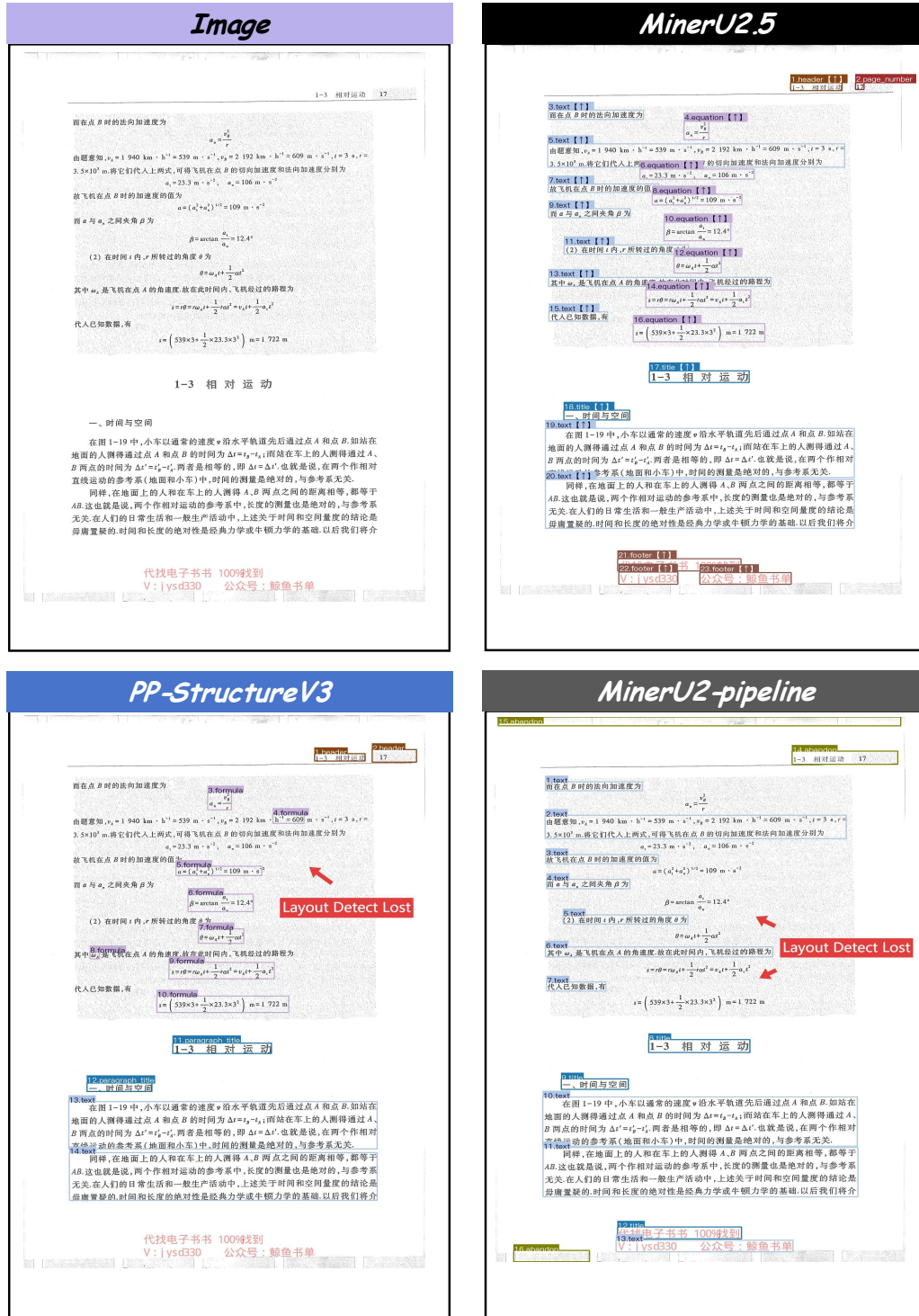


Figure 20: Compare with Previous Version, MinerU2.5 achieve fewer detection omissions in watermark page.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing

Image

JOURNAL OF ECV CLASS FILES VOL. 14, NO. 8, AUGUST 2015

TABLE 2
Quantitative evaluation of state-of-the-art LFSR algorithms. The reported average PSNR and SSIM for Spectral, 2 $\times$, 3 $\times$, 4 $\times$ and Angular 2 $\times$, 3 $\times$, 4 $\times$ and $\times 8$ are highlighted in red and blue, respectively. The best and second best performance, respectively.

Algorithm	Scale	PSNR (dB)					SSIM				
		Original	Reflective	EPIC	Starling	EPIC	Original	Reflective	EPIC	Starling	EPIC
Spectral											
Reada		25.50	25.50	25.49	25.33	25.49	0.857	0.857	0.857	0.857	0.857
Wet [14]		28.86	14.12	28.41	31.24	28.99	0.924	0.858	0.826	0.851	0.798
Wet [14]	2 $\times$	31.45	10.07	31.45	34.07	31.45	0.847	0.793	0.793	0.772	0.772
Wet [14]	3 $\times$	33.97	13.85	29.95	32.98	32.13	0.924	0.861	0.826	0.851	0.824
Wet [14]	4 $\times$	34.24	13.84	32.42	34.24	32.42	0.924	0.861	0.826	0.851	0.824
Wet [14]	8 $\times$	38.08	12.43	31.93	33.13	33.63	0.927	0.891	0.852	0.864	0.845
EPIC [38]		31.46	10.06	31.46	34.07	31.46	0.847	0.793	0.793	0.772	0.772
EPIC [38]	2 $\times$	32.73	33.98	31.46	31.42	33.98	0.851	0.890	0.890	0.864	0.826
EPIC [38]	3 $\times$	32.29	34.34	30.02	32.92	34.34	0.862	0.893	0.893	0.866	0.890
EPIC [38]	4 $\times$	36.34	32.64	30.34	32.34	37.00	0.933	0.894	0.912	0.912	0.897
EPIC [38]	8 $\times$	41.78	34.63	32.64	34.63	41.78	0.940	0.912	0.912	0.912	0.897
Angular											
Reada		26.95	26.95	26.95	26.95	26.94	0.746	0.746	0.746	0.746	0.746
Wet [14]		27.33	19.78	27.33	29.88	27.33	0.746	0.746	0.746	0.746	0.746
Wet [14]	2 $\times$	27.96	10.03	27.96	30.34	28.00	0.791	0.817	0.817	0.786	0.784
Wet [14]	3 $\times$	30.11	10.03	30.11	32.98	30.11	0.847	0.847	0.847	0.847	0.847
Wet [14]	4 $\times$	31.76	26.77	26.77	26.77	26.77	0.924	0.841	0.721	0.803	0.779
Wet [14]	8 $\times$	32.22	32.22	32.22	32.22	32.22	0.924	0.841	0.721	0.803	0.779
EPIC [38]		29.14	10.02	26.95	26.94	26.94	0.746	0.746	0.746	0.746	0.746
EPIC [38]	2 $\times$	30.10	11.47	30.10	32.98	30.10	0.847	0.847	0.847	0.847	0.847
EPIC [38]	3 $\times$	30.62	33.39	29.91	31.76	33.39	0.858	0.896	0.896	0.896	0.896
EPIC [38]	4 $\times$	31.18	33.39	29.91	31.76	33.39	0.858	0.896	0.896	0.896	0.896
EPIC [38]	8 $\times$	31.18	33.37	29.91	31.76	33.39	0.858	0.896	0.896	0.896	0.896
2D											
Reada		25.58	25.58	25.58	25.58	25.58	0.746	0.746	0.746	0.746	0.746
Wet [14]		26.28	19.74	26.28	29.88	26.28	0.746	0.746	0.746	0.746	0.746
Wet [14]	2 $\times$	27.41	10.03	27.41	30.34	28.00	0.791	0.817	0.817	0.786	0.784
Wet [14]	3 $\times$	29.10	10.03	29.10	32.98	29.10	0.847	0.847	0.847	0.847	0.847
Wet [14]	4 $\times$	30.81	26.77	26.77	26.77	26.77	0.924	0.841	0.721	0.803	0.779
Wet [14]	8 $\times$	32.22	32.22	32.22	32.22	32.22	0.924	0.841	0.721	0.803	0.779
EPIC [38]		27.14	10.02	26.95	26.94	26.94	0.746	0.746	0.746	0.746	0.746
EPIC [38]	2 $\times$	27.96	10.03	27.96	30.34	28.00	0.791	0.817	0.817	0.786	0.784
EPIC [38]	3 $\times$	28.40	30.37	27.96	28.40	28.40	0.897	0.796	0.796	0.796	0.891
EPIC [38]	4 $\times$	30.31	30.37	27.96	28.40	28.40	0.897	0.796	0.796	0.796	0.891
EPIC [38]	8 $\times$	30.31	30.37	27.96	28.40	28.40	0.897	0.796	0.796	0.796	0.891
3D											
Reada		15.48	15.37	15.37	15.34	15.34	0.321	0.319	0.319	0.319	0.319
Wet [14]		16.47	15.37	15.37	15.34	15.34	0.321	0.319	0.319	0.319	0.319
Wet [14]	2 $\times$	17.76	10.36	16.32	16.34	16.35	0.392	0.392	0.392	0.392	0.392
Wet [14]	3 $\times$	17.76	11.22	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	4 $\times$	18.60	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8 $\times$	18.70	11.39	17.76	17.76	17.76	0.392	0.392	0.392	0.392	0.392
Wet [14]	8<										

inner25

Journal of LATEX CLASS FILES, VOL. 16, NO. 8, AUGUST 2015

Table 2: Evaluation results of the dot-ocr algorithm, using the average F1 and EER for Table 2a, 2b, and 2c, and also include the best and second best algorithms

Algorithms	Size	Execution Time (s)	Reference and Recall (%)	F1-Score (%)				EER				Area Under ROC Curve	Missed Characters	Error Rate (%)
				dot-ocr	dot-ocr	dot-ocr	dot-ocr	dot-ocr	dot-ocr	dot-ocr	dot-ocr			
Baseline	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2a	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2b	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2c	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2d	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2e	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2f	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2g	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2h	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2i	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2j	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2k	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2l	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2m	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2n	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2o	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2p	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2q	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2r	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2s	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2t	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2u	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2v	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2w	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2x	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2y	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2z	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2aa	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ab	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ac	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ad	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ae	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2af	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ag	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ah	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ai	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2aj	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ak	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2al	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2am	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2an	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ao	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ap	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2aq	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ar	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2as	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2at	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2au	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2av	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2aw	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ax	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ay	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2az	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ba	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bb	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bc	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bd	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2be	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bf	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bg	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bh	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bi	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bj	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bk	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bl	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bm	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bn	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bo	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bp	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bq	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2br	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bs	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bt	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bu	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bv	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bw	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bx	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2by	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2bz	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ca	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2cb	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2cc	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2cd	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2ce	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2cf	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Table 2cg	100	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.			

Figure 22: Compare with others in content dense table.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing

Image

全球科研实力对比						
一级主题词	二级主题词	三级主题词	四级主题词	五级主题词	六级主题词	七级主题词
	计算机视觉	图像处理	机器翻译			
			社会媒体处理			
			图像处理			
			特征提取			
			图像分割			
			三维重建			
			图像增强			
			图像识别			
			图像检索			
			目标检测			
		语音技术	运动行为分析			
			图像标注			
			视频标注			
			虚拟现实			
			增强现实			
	机器学习	集成学习	语音合成			
			语音识别			
			语音增强			
			语音翻译			
			语音系统			
		数据降维	监督学习			
			半监督学习			
			强化学习			
			Boosting			
			Bagging			

32

MinerU2.5

全球科研实力对比						
一级主题词	二级主题词	三级主题词	四级主题词	五级主题词	六级主题词	七级主题词
	计算机视觉	图像处理	机器翻译			
			社会媒体处理			
			图像处理			
			特征提取			
			图像分割			
			三维重建			
			图像增强			
			图像识别			
			图像检索			
			目标检测			
		语音技术	运动行为分析			
			图像标注			
			视频标注			
			虚拟现实			
			增强现实			
	机器学习	集成学习	语音合成			
			语音识别			
			语音增强			
			语音翻译			
			语音系统			
		数据降维	监督学习			
			半监督学习			
			强化学习			
			Boosting			
			Bagging			

32

Gemini-2.5-Pro

全球科研实力对比						
一级主题词	二级主题词	三级主题词	四级主题词	五级主题词	六级主题词	七级主题词
	计算机视觉	图像处理	机器翻译			
			社会媒体处理			
			图像处理			
			特征提取			
			图像分割			
			三维重建			
			图像增强			
			图像识别			
			图像检索			
			目标检测			
		语音技术	运动行为分析			
			图像标注			
			视频标注			
			虚拟现实			
			增强现实			
	机器学习	集成学习	语音合成			
			语音识别			
			语音增强			
			语音翻译			
			语音系统			
		数据降维	监督学习			
			半监督学习			
			强化学习			
			Boosting			
			Bagging			

32

dots.ocr

一级主题词	二级主题词	三级主题词	四级主题词	五级主题词	六级主题词	七级主题词
	计算机视觉	图像处理	机器翻译			
			社会媒体处理			
			图像处理			
			特征提取			
			图像分割			
			三维重建			
			图像增强			
			图像识别			
			图像检索			
			目标检测			
		语音技术	运动行为分析			
			图像标注			
			视频标注			
			虚拟现实			
			增强现实			
	机器学习	集成学习	语音合成			
			语音识别			
			语音增强			
			语音翻译			
			语音系统			
		数据降维	监督学习			
			半监督学习			
			强化学习			
			Boosting			
			Bagging			

Table Structure Error

Table Content Lost

Figure 23: Compare with others in Colored table with many empty cells.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing

Image											
中国银河证券 研究院											
CHINA GALAXY SECURITIES											
公司点评											
附录:											
(一) 公司财务报表摘要											
资产负债表(百万元)											
	2022A	2022E	2024E	2025E	附释表(百万元)	2022A	2022E	2024E	2025E		
流动资产	10511.08	12286.84	14612.28	16677.48	货币资金	8776.53	10059.40	13121.98	14972.29		
货币资金	1594.99	1271.84	1806.96	2383.76	货币成本	6706.94	8488.58	10931.11	13062.99		
应收票据	1752.30	1603.56	1905.54	2148.95	其他应收及应付款	93.50	58.84	43.74	83.16		
应收账款	236.61	489.01	765.85	636.29	其他资产	121.70	148.89	164.31	200.67		
预付账款	626.84	848.86	964.23	1252.29	管理资产	431.14	511.76	565.36	676.89		
存货	207.96	445.85	463.77	513.59	财务资产	338.66	401.45	431.61	441.35		
其他	3302.68	3719.81	4428.87	5081.21	资产减值损失	-15.91	0.00	0.00	0.00		
非流动资产	1340.05	1215.04	1246.49	1284.36	公允价值变动损益	-17.91	0.00	0.00	0.00		
长期股权投资	851.84	751.84	951.84	1151.84	投资净收益	71.41	101.62	125.19	156.62		
固定资产	1961.87	1962.00	2138.80	2272.54	公允价值变动	85.40	108.27	144.21	188.24		
无形资产	758.09	760.09	692.09	676.29	营业外收入	4.09	0.00	0.00	0.00		
其他	199.24	235.21	2462.19	2653.18	营业外支出	16.26	0.00	0.00	0.00		
资产总计	22661.3	24436.88	26077.21	29462.24	利润总额	826.95	1082.27	1446.21	1881.24		
负债和负债	7990.76	8261.34	8422.72	9079.24	所得税	84.37	97.83	129.26	172.74		
应付票据	3661.17	3461.17	3661.17	3961.17	公允价值变动	742.56	984.23	1316.95	1706.50		
应付账款	2575.93	2610.53	2546.02	2705.14	少数股东损益	51.88	66.82	104.87	127.05		
其他	1564.70	1861.68	2027.57	2324.88	归属母公司净利润	695.48	918.53	1212.89	1581.46		
所有者权益	4862.19	5168.29	6162.09	6762.19	EBITDA	1611.39	1839.10	2205.74	2706.54		
少数股东权益	4661.62	5261.62	5863.62	6465.62	EPS (元)	0.45	0.60	0.79	1.03		
其他	516.57	516.57	516.57	516.57							
负债合计	12972.95	13843.53	14804.81	15861.43	主要财务比率	2022A	2023E	2024E	2025E		
资产负债率	80.06	87.68	98.12	110.87%	营业收入	12.11%	24.10%	26.76%	29.09%		
少数股东权益占比	8873.13	9771.68	10930.56	12312.29	营业利润率	-4.05%	29.55%	33.63%	36.08%		
少数股东权益占比	22661.3	24436.88	26077.21	29462.24	毛利率	51.49%	6.24%	7.23%	8.14%		
					净利率	7.89%	8.47%	9.22%	9.32%		
					ROE	7.80%	8.42%	11.50%	12.84%		
					ROIC	5.49%	6.24%	7.23%	8.14%		
					总资产周转率	57.52%	56.65%	55.08%	53.80%		
					净资产收益率	133.92%	130.66%	122.60%	116.48%		
					总资产周转率	1.28	1.49	1.67	1.81		
					净资产收益率	0.49	0.44	0.53	0.61		
					总资产周转率	0.39	0.46	0.50	0.58		
					净资产收益率	5.01	6.85	6.80	7.90		
					总资产周转率	3.40	4.17	5.16	6.27		
					净资产收益率	0.45	0.60	0.79	1.03		
					净资产收益率	-0.29	-0.23	0.67	0.77		
					净资产收益率	5.77	6.32	7.11	8.14		
					净资产收益率	17.85	13.50	16.19	7.82		
					净资产收益率	1.39	1.27	1.13	0.99		
					净资产收益率	13.88	11.22	9.25	7.66		
					净资产收益率	1.40	1.13	0.93	0.72		
					净资产收益率	692.48	915.83	1212.89	1581.46		
					净资产收益率	1811.39	1839.10	2205.74	2706.54		
					净资产收益率	0.45	0.60	0.79	1.03		

资料来源: 银河证券, 中国银河证券研究院

请务必阅读正文最后的中国银河证券股份有限公司免责声明。

1

MinerU2.5											
中国银河证券 研究院											
CHINA GALAXY SECURITIES											
公司点评											
附录:											
(一) 公司财务报表摘要											
资产负债表(百万元)											
	2022A	2023E	2024E	2025E							
流动资产	10511.08	12286.84	14612.28	16677.48	货币资金	8776.53	10059.40	13121.98	14972.29		
货币资金	1594.99	1271.84	1806.96	2383.76	货币成本	6706.94	8488.58	10931.11	13062.99		
应收票据	1752.30	1603.56	1905.54	2148.95	其他应收及应付款	93.50	58.84	43.74	83.16		
应收账款	236.61	489.01	765.85	636.29	其他资产	121.70	148.89	164.31	200.67		
预付账款	626.84	848.86	964.23	1252.29	管理资产	431.14	511.76	565.36	676.89		
存货	207.96	445.85	463.77	513.59	财务资产	338.66	401.45	431.61	441.35		
其他	3302.68	3719.81	4428.87	5081.21	资产减值损失	-15.91	0.00	0.00	0.00		
非流动资产	1340.05	1215.04	1246.49	1284.36	公允价值变动损益	-17.91	0.00	0.00	0.00		
长期股权投资	851.84	751.84	951.84	1151.84	投资净收益	71.41	101.62	125.19	156.62		
固定资产	1961.87	1962.00	2138.80	2272.54	公允价值变动	85.40	108.27	144.21	188.24		
无形资产	758.09	760.09	692.09	676.29	营业外收入	4.09	0.00	0.00	0.00		
其他	199.24	235.21	2462.19	2653.18	营业外支出	16.26	0.00	0.00	0.00		
资产总计	22661.3	24436.88	26077.21	29462.24	利润总额	826.95	1082.27	1446.21	1881.24		
负债和负债	7990.76	8261.34	8422.72	9079.24	所得税	84.37	97.83	129.26	172.74		
应付票据	3661.17	3461.17	3661.17	3961.17	公允价值变动	742.56	984.23	1316.95	1706.50		
应付账款	2575.93	2610.53	2546.02	2705.14	少数股东损益	51.88	66.82	104.87	127.05		
其他	1564.70	1861.68	2027.57	2324.88	归属母公司净利润	695.48	918.53	1212.89	1581.46		
所有者权益	4862.19	5168.29	6162.09	6762.19	EBITDA	1611.39	1839.10	2205.74	2706.54		
少数股东权益	4661.62	5261.62	5863.62	6465.62	EPS (元)	0.45	0.60	0.79	1.03		
其他	516.57	516.57	516.57	516.57							
负债合计	12972.95	13843.53	14804.81	15861.43	主要财务比率	2022A	2023E	2024E	2025E		
资产负债率	80.06	87.68	98.12	110.87%	营业收入	12.11%	24.10%	26.76%	29.09%		
少数股东权益占比	8873.13	9771.68	10930.56	12312.29	营业利润率	-4.05%	29.55%	33.63%	36.08%		
少数股东权益占比	22661.3	24436.88	26077.21	29462.24	毛利率	51.49%	6.24%	7.23%	8.14%		
					净利率	7.89%	8.47%	9.22%	9.32%		
					ROE	7.80%	8.42%	11.50%	12.84%		
					ROIC	5.49%	6.24%	7.23%	8.14%		
					总资产周转率	57.52%	56.65%	55.08%	53.80%		
					净资产收益率	133.92%	130.66%	122.60%	116.48%		
					总资产周转率	1.28	1.49	1.67	1.81		
					净资产收益率	0.49	0.44	0.53	0.61		
					总资产周转率	0.39	0.46	0.50	0.58		
					净资产收益率	5.01	6.85	6.80	7.90		
					总资产周转率	3.40	4.17	5.16	6.27		
					净资产收益率	0.45	0.60	0.79	1.03		
					净资产收益率	-0.29	-0.23	0.67	0.77		
					净资产收益率	5.77	6.32	7.11	8.14		
					净资产收益率	17.85	13.50	16.19	7.82		
					净资产收益率	1.39	1.27	1.13	0.99		
					净资产收益率	13.88	11.22	9.25	7.66		
					净资产收益率	1.40	1.13	0.93	0.72		
					净资产收益率	692.48	915.83	1212.89	1581.46		
					净资产收益率	1811.39	1839.10	2205.74	2706.54		
					净资产收益率	0.45	0.60	0.79	1.03		

资料来源: 银河证券, 中国银河证券研究院

请务必阅读正文最后的中国银河证券股份有限公司免责声明。

2

Gemini-2.5-Pro											
中国银河证券 研究院											
CHINA GALAXY SECURITIES											
公司点评											
附录:											
(一) 公司财务报表摘要											
资产负债表(百万元)											
	2022A	2023E	2024E	2025E	附释表(百万元)	2022A	2023E	2024E	2025E		
流动资产	10311.08	12286.84	14612.28	16677.48	货币资金	8776.53	10059.40	13121.98	14972.29		
货币资金	1594.99	1271.84	1806.96	2383.76	货币成本	6706.94	8488.58	10931.11	13062.99		
应收票据	1752.30	1603.56	1905.54	2148.95	其他应收及应付款	93.50	58.84	43.74	83.16		
应收账款	236.61	489.01	765.85	636.29	其他资产	121.70	148.89	164.31	200.67		
预付账款	626.84	848.86	964.23	1252.29	管理资产	431.14	511.76	565.36	676.89		
存货	207.96	445.85	463.77	513.59	财务资产	338.66	401.45	431.61	441.35		
其他	3302.68	3719.81	4428.87	5081.21	资产减值损失	31.60	0.00	0.00	0.00		
非流动资产	12479.05	12724.12	12444.30	12394.36	公允价值变动损益	17.91	0.00	0.00	0.00		
固定资产	801.64	712.94	604.93	507.34	净营业利润	7.41	104.22	134.23	156.42		
无形资产	1945.47	1965.49	2193.48	2227.24	其他收益	64.65	108.27	127.17	149.18		
长期资产	7065.09	7066.09	6900.09	6762.09	营业外收入	4.09	0.00	0.00	0.00		
其他	1992.24	2274.21	2442.19	2394.19	营业外支出	16.56	0.00	0.00	0.00		
资产总计	22890.13	25010.96	27056.58	29071.84	净利润	838.99	198.27	144.25	171.58		
负债	7202.19	8200.14	8422.12	8579.34	所得税	84.87	40.45	43.24	44.13		
短期借款	3649.13	3649.13	3649.13	3649.13	净利润	744.86	158.44	114.91	130.52		
应付账款	2576.93	2615.33	2644.02	2734.14	少数股东损益	31.88	66.82	134.37	127.06		
应付债券	1654.76	1691.48	1707.57	1734.76	归属于母公司净利润	804.00	157.82	111.91	127.48		
其他	4262.19	4576.19	4562.19	4576.19	归属于少数股东净利润	44.11	93.93	127.39	137.54		
长期借款	4666.62	5265.62	5865.62	6465.62	EPS (元)	0.45	0.00	0.00	0.79	13.75	
其他	316.27	316.27	316.27	316.27							
负债合计	12271.96	13943.53	14644.36	15088.43							
所有者权益	6262.64	8777.99	8860.19	11678.79	主要股东名称	2022A	2023E	2024E	2025E		
少数股东权益	100.00	100.00	100.00	100.00	银河证券股份有限公司	12.11%	24.10%	20.76%	28.90%		
归属于母公司所有者权益	6678.13	8777.68	8860.19	11678.79	资本公积	8.03%	23.91%	23.48%	30.36%		
其他	22890.13	25010.96	27056.58	29071.84	其他综合收益	4.09%	32.20%	33.67%	33.00%		
					公允价值变动损益	21.66%	21.66%	21.66%	21.66%		
					营业外收入	7.99%	8.41%	8.27%	9.32%		
					净利润	7.80%	8.42%	10.11%	12.48%		
					归属于母公司净利润	8.49%	8.24%	7.23%	8.14%		
					归属于少数股东净利润	4.12%	16.65%	16.16%	16.65%		
					净利润	13.93%	13.93%	13.93%	14.91%		
					归属于母公司净利润	1.28	1.49	1.67	1.83		
					归属于少数股东净利润	0.45	0.44	0.40	0.40		
					净利润	0.30	0.45	0.50	0.55		
					归属于母公司净利润	0.1	0.60	0.80	0.78		
					归属于少数股东净利润	0.40	4.17	5.16	6.27		
					净利润	0.45	0.60	0.76	0.78		
					归属于母公司净利润	0.29	0.23	0.67	0.77		
					归属于少数股东净利润	0.77	0.32	7.11	8.14		
					净利润	13.88	13.50	15.19	15.72		
					归属于母公司净利润	1.59	1.27	1.59	1.72		
					归属于少数股东净利润	1.34	12.23	12.35	7.66		
					净利润	1.43	1.11	0.99	0.72		

数据来源：公司年报、Wind数据库整理

谨此声明：本报告及其中所有数据均来自公开渠道，不构成任何投资建议，也不构成任何要约或要约邀请。

1

Table Split Error

Image		MinerU2.5	
附 全书大归纳		全书大归纳 附 全书大归纳 核心知识归纳 一元一次方程	
核心知识归纳			
一元一次方程			
知识点	内容		
方程的概念	使方程左右两边的值相等的未知数的值就是该方程的解。提示:将方程的解代入方程,等式成立。		
等式的基本性质	(1)等式两边都加上(或都减去)同一个数或同一个整式,所得结果仍是等式。如果 $a=b$, 那么 $a+c=b+c$, $a-c=b-c$;(2)等式两边都乘以(或都除以)同一个数(除数不能为0),所得结果仍是等式。如果 $a=b$, 那么 $ac=bc$, $c \neq 0$ 。		
方程的变形规则	(1)方程两边都加上(或都减去)同一个数或同一个整式,方程的解不变;(2)方程两边都乘以(或都除以)同一个不等于0的数,方程的解不变。		
移项	将方程中的某些项改变符号后,从方程的一边移到另一边的变形。注意:移项后要改变符号。		
一元一次方程的定义	只含有一个未知数,并且含有未知数的式子都是整式,未知数的次数都是1,像这样的方程叫做一元一次方程。提示:一般式为 $ax=b$ (a, b 为常数,且 $a \neq 0$)。		
解一元一次方程的步骤	(1)去分母;(2)去括号;(3)移项;(4)合并同类项;(5)将未知数的系数化为1。提示:应根据方程特点,灵活选择解题步骤。		
列一元一次方程解应用题的几种常见题型及特点	类型	题中涉及的数量关系及公式	等量关系
	和、差、倍、分问题		明确“倍数”关系及“多”、“少”关系等
	等积变形问题	长方体体积=长×宽×高; 圆柱体体积= $\pi r^2 h$ (其中 h 为高, r 为底面半径)	变形前后体积相等
	相遇问题		快行的距离+慢行的距离=总距离(相向而行)
	行程问题	路程=速度×时间; 时间=路程÷速度; 速度=路程÷时间	快行的距离-慢行的距离=距离差(同向而行)
	追及问题		静水速度+水流速度=顺水速度 静水速度-水流速度=逆水速度(顺水路程=逆水路程)
	水流问题		
	调配问题	从调配后的数量关系中找等量关系。注意:调配对象流动的方向和数量。	
	比例分配问题	全部数量=各份的数量之和。提示:可把一份数量设为 x 。	
	其他问题		
284		关注微信公众号“初高教辅站”获取更多初中教辅资料	

Gemini-2.5-Pro		dots.ocr	
附 全书大归纳		Table Structure Error	
核心知识归纳			
一元一次方程			
知识点	内容		
方程的概念	使方程左右两边的值相等的未知数的值就是该方程的解。提示:将方程的解代入方程,等式成立。		
等式的基本性质	(1)等式两边都加上(或都减去)同一个数或同一个整式,所得结果仍是等式。如果 $a=b$, 那么 $a+c=b+c$, $a-c=b-c$;(2)等式两边都乘以(或都除以)同一个数(除数不能为0),所得结果仍是等式。如果 $a=b$, 那么 $ac=bc$, $c \neq 0$ 。		
方程的变形规则	(1)方程两边都加上(或都减去)同一个数或同一个整式,方程的解不变;(2)方程两边都乘以(或都除以)同一个不等于0的数,方程的解不变。		
移项	将方程中的某些项改变符号后,从方程的一边移到另一边的变形。注意:移项后要改变符号。		
一元一次方程的定义	只含有一个未知数,并且含有未知数的式子都是整式,未知数的次数都是1,像这样的方程叫做一元一次方程。提示:一般式为 $ax=b$ (a, b 为常数,且 $a \neq 0$)。		
解一元一次方程的步骤	(1)去分母;(2)去括号;(3)移项;(4)合并同类项;(5)将未知数的系数化为1。提示:应根据方程特点,灵活选择解题步骤。		
列一元一次方程解应用题的几种常见题型及特点	类型	题中涉及的数量关系及公式	等量关系
	和、差、倍、分问题		明确“倍数”关系及“多”、“少”关系等
	等积变形问题	长方体体积=长×宽×高; 圆柱体体积= $\pi r^2 h$ (其中 h 为高, r 为底面半径)	变形前后体积相等
	相遇问题		快行的距离+慢行的距离=总距离(相向而行)
	行程问题	路程=速度×时间; 时间=路程÷速度; 速度=路程÷时间	快行的距离-慢行的距离=距离差(同向而行)
	追及问题		静水速度+水流速度=顺水速度 静水速度-水流速度=逆水速度(顺水路程=逆水路程)
	水流问题		
	调配问题	从调配后的数量关系中找等量关系。注意:调配对象流动的方向和数量。	
	比例分配问题	全部数量=各份的数量之和。提示:可把一份数量设为 x 。	
	其他问题		
284		关注微信公众号“初高教辅站”获取更多初中教辅资料	

Figure 25: Compare with others in Table with irregular merged cells.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing



Figure 26: Compare with others in Tables with No Frame.

A.3.2 Formula

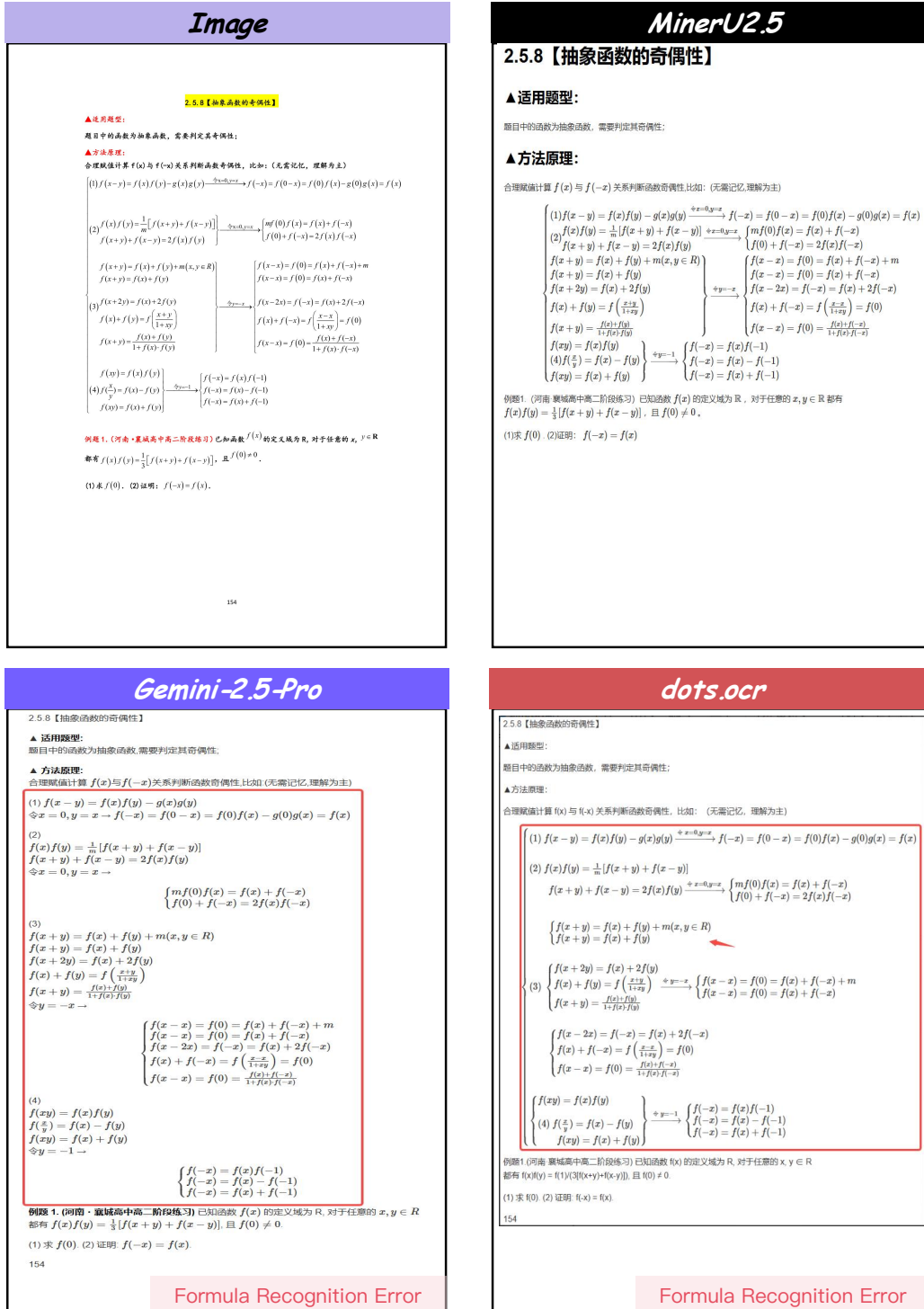


Figure 27: Compare with others in Nested conditional expressions.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing

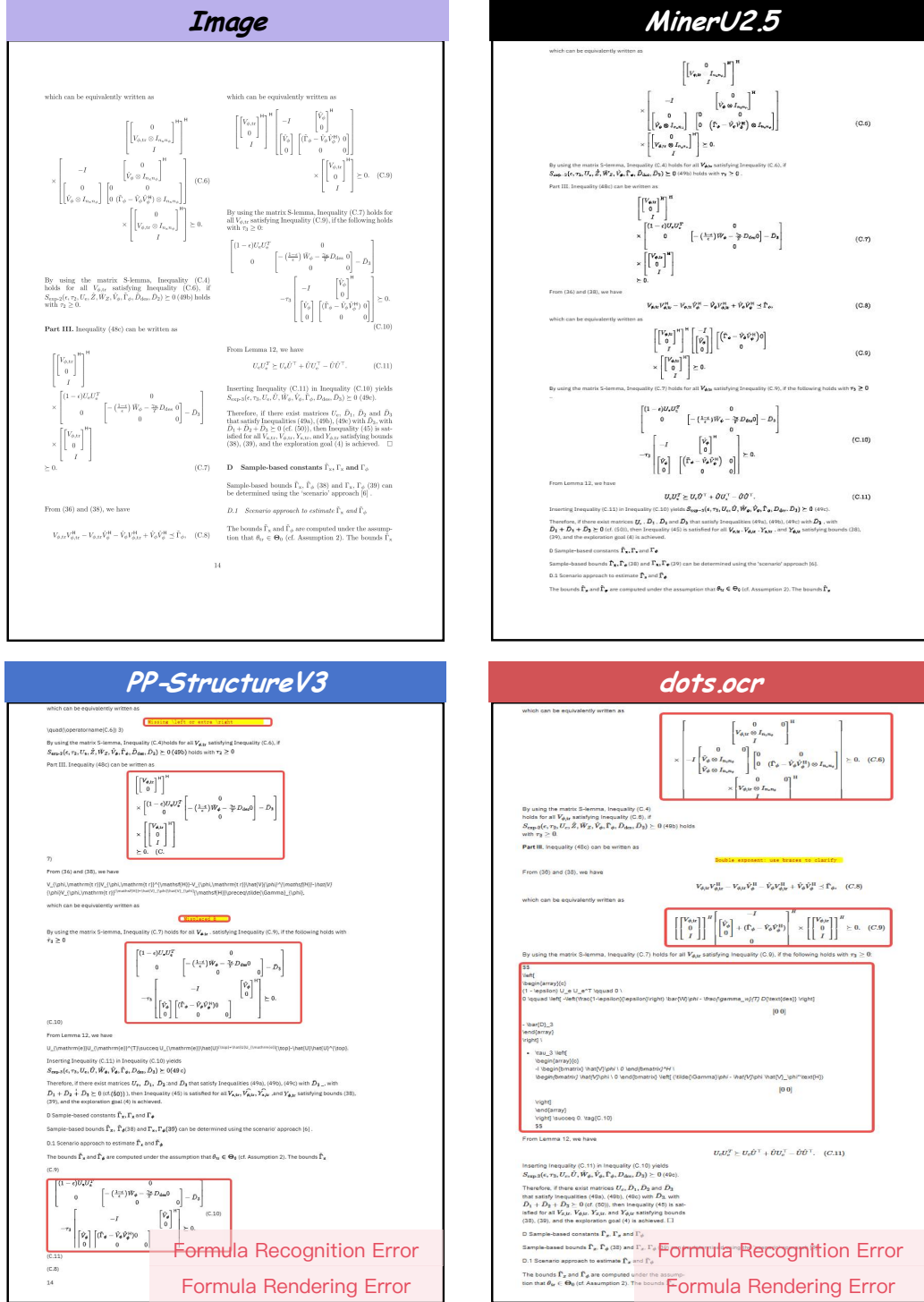


Figure 29: Compare with others in Nested matrix.

A.3.3 Layout&OCR



Figure 30: Compare with others in Academic literature with alternating text and images.

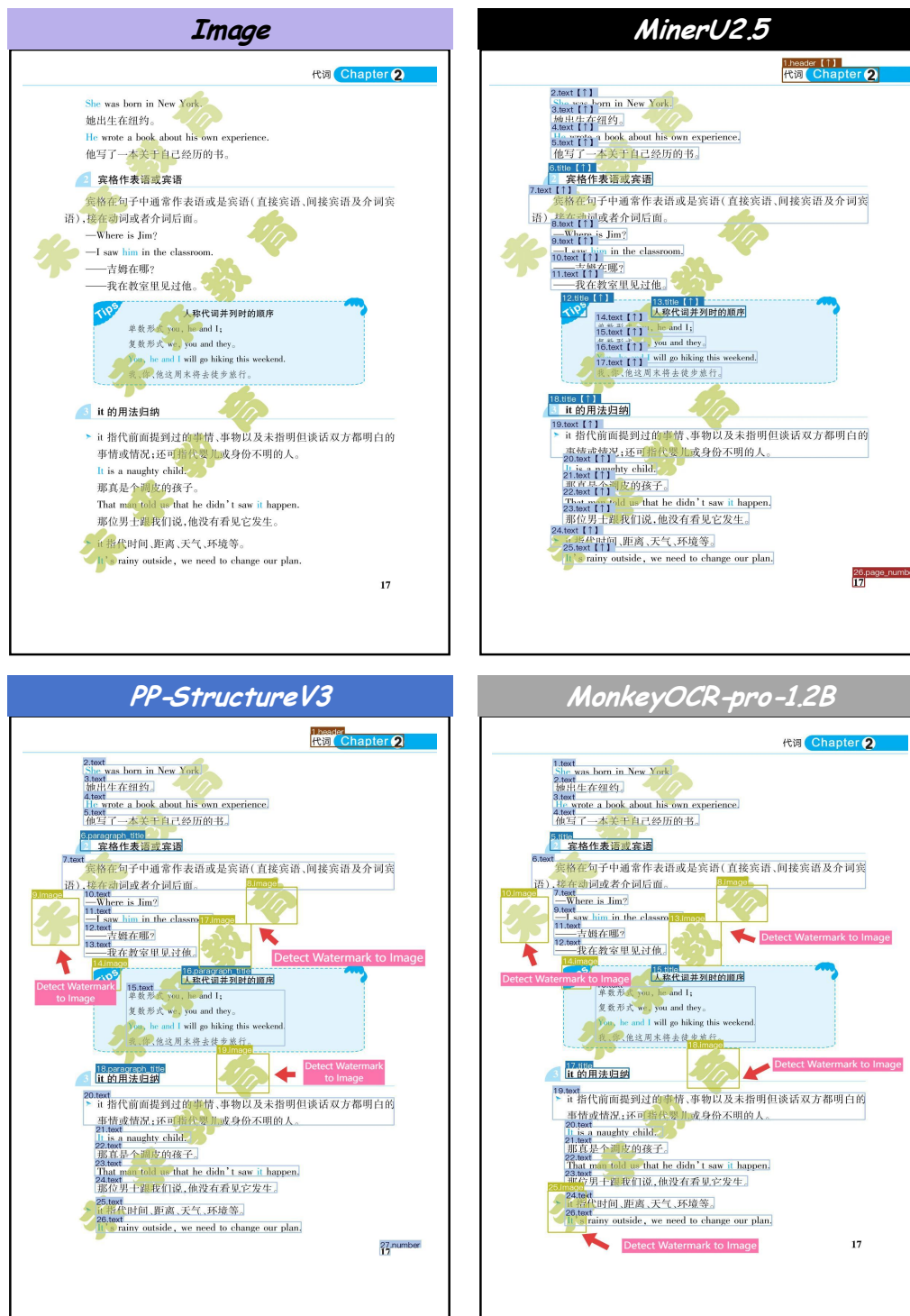


Figure 32: Compare with others in Textbooks with watermarks.

B Prompt Details

Here, we provide a detailed description of the different prompts used during the two-stage inference of MinerU2.5, along with their corresponding output formats.

B.1 Layout Detection

The layout detection output will include the relative coordinates, category, and rotation direction of each element in the document. Each element will be output in sequence, ensuring traceability for all layout data. The input image will be resized to a resolution of 1036×1036 .

Output format:

- **Box Coordinates:** x_1, y_1, x_2, y_2
- **Document Element Category:** title, text, image, etc.
- **Rotation Direction:** up, down, left, right

Example:

```
<|box_start|>100 200 300 400<|box_end|><|ref_start|>title<|ref_end|><|rotate_up|>
<|box_start|>400 500 600 700<|box_end|><|ref_start|>text<|ref_end|><|rotate_up|>
```

B.2 Text Recognition

The output will contain the recognized text results. The input image will retain its native resolution; however, the number of image tokens will be limited to the range of 4 to 2048. If this limit is exceeded, the image will be scaled accordingly.

Output format:

- **OCR Results:** The raw OCR output

Example:

```
The results of the analyses of the uncertainty of the field data and related assumptions are
shown in Figs 13 and 14.
```

B.3 Formula Recognition

Any formulas found in the image will be extracted and converted into LaTeX format. The input image will retain its native resolution; however, the number of image tokens will be limited to the range of 4 to 2048. If this limit is exceeded, the image will be scaled accordingly.

Output format:

- **LaTeX Format:** The LaTeX representation of the formula

Example:

```
\[
\hat{F} = \operatorname{Concat}\left(\left[ F_{\{1\}}, F_{\{2\}}, \dots, F_{\{n\}} \right]\right) \tag{2}
\]

\[\begin{aligned} M = & \sigma \bigl( \mathrm{GELU}(\mathrm{BN}(\mathrm{Conv}_{\mathrm{gate}}(\hat{F}))) \bigr) \tag{3} \end{aligned}
\]
```

B.4 Table Recognition

The output will include the recognized tables, structured in an OTSL (Open Table Structure Language) format for easy data processing. The input image will retain its native resolution; however, the number of image tokens will be limited to the range of 4 to 2048. If this limit is exceeded, the image will be scaled accordingly.

Output format:

- **OTSL Format:** The table represented in OTSL format

Example:

```
<fcel>Site<fcel>Cl<fcel>N03<fcel>S04<fcel>Na<fcel>Ca<fcel>K<fcel>Mg<fcel>NH4<fcel>References
<nl>
<fcel>Cl dominance sites<lcel><lcel><lcel><lcel><lcel><lcel><lcel><lcel><lcel><lcel><nl>
<fcel>Comba<fcel>109.8<fcel>12.1<fcel>23.3<fcel>86.8<fcel>43.4<fcel>4.8<fcel>15.1<fcel>13.2<
fcel>Present study<nl>
<fcel>Alibagh<fcel>236<fcel>9<fcel>36<fcel>220<fcel>46<fcel>5<fcel>64<fcel>8<fcel>Naik et al
. (2002)<nl>
<fcel>Goa<fcel>113.4<fcel>5.5<fcel>27.4<fcel>97.2<fcel>41.5<fcel>2.5<fcel>24.5<fcel>5.5<fcel>
Parashar et al. (2001)<nl>
<fcel>Bombay<fcel>138<fcel>-<fcel>10<fcel>115<fcel>36<fcel>3.6<fcel>24<fcel>-<fcel>Sequeira
(1976)<nl>
<fcel>Na dominance sites<lcel><lcel><lcel><lcel><lcel><lcel><lcel><lcel><lcel><nl>
<fcel>Colaba<fcel>171<fcel>34<fcel>52<fcel>179<fcel>133<fcel>6<fcel>59<fcel>12<fcel>Naik et
al. (2002)<nl>
<fcel>Silent Valley<fcel>43.0<fcel>21.0<fcel>20.0<fcel>46.0<fcel>43.0<fcel>4.0<fcel>14.0<
fcel>3.0<fcel>Rao et al. (1995)<nl>
<fcel>Chembur<fcel>164.5<fcel>29.5<fcel>70.4<fcel>168.2<fcel>89.5<fcel>6.9<fcel>36.5<fcel>
41.1<fcel>Khemani et al. (1994)<nl>
<fcel>Bhubaneswar<fcel>18<fcel>10<fcel>19.1<fcel>15<fcel>20.2<fcel>1.8<fcel>5.2<fcel>18.7<
fcel>Das et al. (2005)<nl>
```