UniRec-0.1B: Unified Text and Formula Recognition with 0.1B Parameters

Yongkun Du[1,2], Zhineng Chen[1,2(✉)], Yazhen Xie[1,2], Weikang Bai[1,2], Hao Feng[3], Wei Shi[3], Yuchen Su[1,2], Can Huang[3], and Yu-Gang Jiang[1,2(✉)]

[1] Institute of Trustworthy Embodied AI, Fudan University, China
[2] Shanghai Key Laboratory of Multimodal Embodied AI, Fudan University, China
[3] ByteDance, China

Abstract. Text and formulas constitute the core informational components of many documents. Accurately and efficiently recognizing both is crucial for developing robust and generalizable document parsing systems. Recently, vision-language models (VLMs) have achieved impressive unified recognition of text and formulas. However, they are large-sized and computationally demanding, restricting their usage in many applications. In this paper, we propose UniRec-0.1B, a unified recognition model with only 0.1B parameters. It is capable of performing text and formula recognition at multiple levels, including characters, words, lines, paragraphs, and documents. To implement this task, we first establish UniRec40M, a large-scale dataset comprises 40 million text, formula and mixed samples, enabling the training of a powerful yet lightweight model. Secondly, we identify two challenges when building such a lightweight but unified expert model. They are: structural variability across levels and semantic entanglement between textual and formulaic content. To tackle these, we introduce a hierarchical supervision training that explicitly guides structural comprehension, and a semantic-decoupled tokenizer that separates text and formula representations. Finally, we develop a comprehensive evaluation benchmark covering Chinese and English documents from multiple domains and with multiple levels. Experimental results on this and public benchmarks demonstrate that UniRec-0.1B outperforms both general-purpose VLMs and leading document parsing expert models, while achieving 2-9× speedup, validating its effectiveness and efficiency. Codebase and Dataset: `https://github.com/Topdu/OpenOCR`.

Keywords: Unified Text and Formula Recognition · Lightweight Model · Document Parsing

1 Introduction

Document parsing [81] serves as a key component in real-world applications like document understanding, digital education, and information retrieval, etc. As the

✉ Corresponding authors. E-mail:{zhinchen,ygj}@fudan.edu.cn

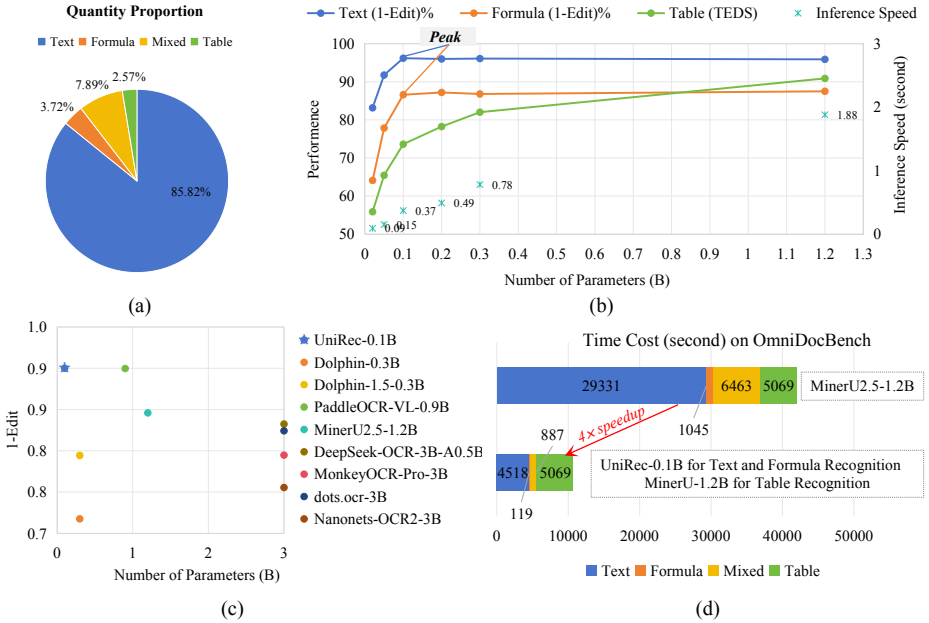

Fig. 1: (a) Status Quo: Text and formulas take up 97.43% of page regions in quantity in OmniDocBench [53]. **(b) Scaling Law Analysis**: As model size increases, text and formula recognition rapidly improve and peak around 0.1B parameters, whereas table performance continues to grow more steadily with diminishing returns. Meanwhile, inference latency increases with the model size, revealing a clear trade-off between performance and efficiency. **(c) Result**: Comparison with advanced document parsing models in our constructed UniRec-Bench. UniRec-0.1B outperforms or on par with existing models in accuracy while only has 3%-33% of their parameters, demonstrating strong efficiency for text and formula recognition tasks. **(d) Advantage**: Employing UniRec-0.1B for text and formula recognition, combined with MinerU2.5-1.2B for table recognition, achieves approximately 4× speedup over MinerU2.5-1.2B for all recognition tasks.

core tasks of document parsing, text and formula recognition have traditionally been addressed as distinct tasks, each with extensive research [3, 16, 18, 19, 21, 37, 46, 56–58, 67, 77, 79] and significant progress over the past decades.

Recent advances in vision-language models (VLMs) [2, 9, 26, 29, 63, 74, 75, 87] have enabled end-to-end document parsing [5, 8, 10, 11, 14, 35, 38, 42, 48, 50, 51, 55, 71, 72], where a single large model unifies text, formula, table, and chart understanding within one framework. While conceptually elegant, such unified paradigms typically rely on billion-scale parameters, resulting in substantial computational cost and inference latency. On the other hand, a closer examination of real-world document distributions reveals a critical imbalance. As shown in Fig. 1(a), text and formulas dominate document content, accounting for 97.43% of page regions in quantity in OmniDocBench [53]. More importantly, Fig. 1(d)

shows that these two modalities consume 36839 seconds in total, accounting for 87.90% of the total parsing time when using MinerU2.5 [51]. This indicates that optimizing text and formula recognition is the key for accelerating document parsing systems.

Furthermore, the scaling law analysis in Fig. 1(b) reveals performance in text and formula recognition improves rapidly with model scaling and saturates at around 0.1B parameters, beyond which additional scaling yields negligible gains. In contrast, table recognition continues to benefit from larger models, in accordance with its substantially higher task complexity. These observations suggest that blindly increasing model capacity to unify all tasks is inefficient: larger models significantly increase inference latency while offering diminishing returns for the dominant modalities. Overall, the results highlight the inherently heterogeneous nature of document parsing and motivate a divide-and-conquer strategy rather than relying on a single large model. As illustrated in Fig. 1(c) and (d), such a collaborative framework achieves approximately a 4× overall speedup while maintaining competitive performance. This paradigm strikes a more favorable balance between efficiency and accuracy, providing a practical and scalable solution for real-world document parsing scenarios.

Motivated by this, we introduce UniRec-0.1B, a lightweight recognition model with only 0.1 billion parameters, specializing in dominant and saturation-prone tasks (text and formulas). We aim to use this model to jointly recognize textual and formula content across multiple levels (e.g., character, word, line, and paragraph), and thus significantly improving the efficiency of document parsing. However, the first challenge encountered is training data. Existing datasets are limited in both scale and diversity, making it difficult for small models to achieve comprehensive learning in multi-task and multi-level settings. To address this, we first construct a large-scale Chinese and English text-formula recognition dataset that includes approximately 40 million samples. The dataset integrates public text recognition data, Wikipedia articles, and LaTeX formulas extracted from arXiv papers. It covers a wide range of real-world scenarios such as digital-born documents, photographs, scanned pages, and handwritten content. This provides a solid data foundation for unified recognition.

Subsequently, achieving unified recognition by using such a small parameter budget raises two technical challenges. First, **structural variability**: Document elements across different levels exhibit significant structural diversity, making it difficult for the model to adapt to multi-granularity representations simultaneously. Second, **semantic entanglement**: Existing approaches typically employ a coupled tokenizer to process both text and formula modalities, where the same token may represent distinct semantics in purely text versus formula contexts. This coupling leads to semantic entanglement between modalities. While large language models (LLMs) can partially absorb such ambiguity through their huge model capacity, small models are far more sensitive to it, resulting in significant performance degradation.

To overcome these challenges, we introduce two novel techniques: Hierarchical Supervision Training (HST) and Semantics-Decoupled Tokenizer (SDT).

HST explicitly inserts hierarchical tokens into the label sequence. They guide the model in recognizing and distinguishing different structural levels, thereby enhancing its ability to model hierarchical structures. SDT constructs independent vocabularies for textual and formula modalities, fundamentally eliminating the cross-modality semantic confusion. Note that despite being simple and straightforward, existing document parsing systems, such as PaddleOCR-VL [11], MinerU2.5 [51], MonkeyOCR [39], Dolphin-1.5 [24], etc, all have missed such designs.

To evaluate the performance of UniRec-0.1B comprehensively, we build UniRec-Bench, a benchmark of Chinese and English documents covering text, formula, text-formula mixed content with multiple levels and domains based on OmniDocBench [53]. Experimental results demonstrate that UniRec-0.1B achieves highly competitive accuracy compared to large-scale VLM-based document parsing models [5, 8, 10, 11, 35, 38, 42, 48, 50, 55, 71, 72] on various text and formula recognition tasks. Furthermore, by replacing the recognition modules of existing document parsing models with UniRec-0.1B, we observe a 2-9× overall parsing speedup. These results highlight the effectiveness of UniRec-0.1B in efficient and unified recognition.

Our main contributions are summarized as follows:

- We construct UniRec40M, a comprehensive dataset containing 40 million multi-level text-formula samples in Chinese and English. It fills the data gap in unified text and formula recognition.
- We propose UniRec-0.1B, a lightweight unified recognition model. It addresses the granularity- and modality-mixed challenges by introducing Hierarchical Supervision Training and Semantics-Decoupled Tokenizer.
- Extensive evaluations show that UniRec-0.1B outperforms or is on par with leading large-scale VLM-based models in accuracy across various text and formula recognition tasks, while delivering 2-9× inference speedup.

2 Related Work

Text and Formula Recognition. Text and formula recognition are traditionally treated as distinct tasks and both mainly focus on line-level recognition, aiming to convert text instances or mathematical formulas into symbolic or character sequences. In text recognition, existing methods are typically categorized into Connectionist Temporal Classification (CTC)-based methods [17, 19, 27, 32, 57] and encoder-decoder-based methods [3,15,22,28,37,47,56,58,69,70,77,78,82–86]. While CTC-based models emphasize fast inference, encoder-decoder-based ones generally have higher accuracy with increased computational cost. Most formula recognition (or mathematical expression recognition, MER) methods utilize the encoder-decoder framework [13,46,67,73,79,80]. Adversarial and structure-aware models [73, 79], and Transformer or VLMs [46, 67] have significantly advanced this field. These approaches progressively improve syntactic parsing and contextual understanding of formulas. Despite these advancements, both text and formula recognition remain largely constrained to word- or line-level inputs, with

paragraph- or multi-line recognition still relying on additional text and formula detection models.

Document Parsing with VLMs. With the rise of VLMs [2, 9, 26, 29, 63, 74, 75, 87], document parsing is increasingly moving toward unified models [5, 8, 10, 11, 24, 35, 38, 42, 48, 50, 55, 71, 72]. Pioneering efforts such as Nougat [5], GOT [71], Dolphin [23, 24] introduce end-to-end frameworks that jointly extract and recognize multiple document elements, including text, formulas, tables, and charts, within a single model. These approaches demonstrate the feasibility of replacing traditional Optical Character Recognition (OCR) pipelines [12, 61, 68] with unified VLM-based solutions. Building on this paradigm, recent methods such as Ocean-OCR [8], olmOCR [55], dots.ocr [38], DeepSeek-OCR [72], and HunyuanOCR [62] employ multi-modal large language models [2, 40, 64] trained on extensive in-house document corpora to enhance end-to-end document parsing performance. While these unified models simplify system design and improve holistic reasoning, they often require large model sizes and incur substantial computational cost. An alternative line of work adopts multi-stage or hybrid designs that decouple layout analysis from content recognition, aiming to balance efficiency and recognition accuracy [11, 24, 39, 51]. Dolphin [24] utilizes a Swin-Transformer-based VLM for page-level layout parsing followed by parallel region recognition. MinerU2.5 [51] adopts a similar two-stage framework with a single vision-language model. MonkeyOCR [39] and PaddleOCR-VL [11] also follow a multi-stage strategy, integrating expert models for layout and reading-order analysis while employing VLMs for unified recognition of text, formulas, and tables. Although these methods unify multiple recognition tasks to reduce dependency on specialized models, they often increase parameter counts and reduce efficiency. In contrast, UniRec-0.1B focuses on lightweight unification of text and formula recognition, the two most dominant elements in documents. It maintains competitive accuracy while significantly improving parsing speed.

Training Data for Document Parsing. Existing text and formula recognition datasets, such as RealData [3], Union14M-L [34], and UniMERNet-1M [67], mainly contain line-level samples. Multi-Level and mixed text-formula datasets remain scarce. Document parsing models, such as GOT [71], Dolphin [24], dots.ocr [38], MinerU2.5 [51], PaddleOCR-VL [11], and DeepSeek-OCR [72] have built their own in-house datasets, which unfortunately have not been released publicly. olmOCR [55] releases 270,000 full-page documents consisting of plain text, but with limited coverage of formulas and without multi-level annotations. Recently, DocHumming [36] demonstrates the potential of end-to-end document parsing through a carefully designed data pipeline and training strategy. Its systematic design pushes the upper bound of end-to-end parsing performance in complex document understanding scenarios. UniRec40M fills this gap by offering a large-scale, multi-level dataset that covers both text and formula instances, enabling comprehensive document parsing research. As shown in Tab. 1, we present a detailed comparison between UniRec40M and existing datasets. This comparison highlights the unique characteristics of UniRec40M, particularly its large scale,

Dataset	Task		Level					Language		Size
	Text	Formula	Character	Word	Line	Paragraph	Multi-Paragraph	English	Chinese	
ST [30]	✓		✓	✓				✓		7M
MJ [33]	✓		✓	✓				✓		9M
Real [3]	✓		✓	✓				✓		3.5M
Union14M-L [34]	✓		✓	✓				✓		3.2M
BCTR [7]	✓		✓	✓	✓			✓	✓	1.1M
IM2LATEX [13]		✓			✓			✓		75.3K
Pix2tex [4]		✓			✓			✓		233.8K
UniMER-1M [67]		✓			✓			✓		1.1M
olmOCR [55]	✓	✓					✓	✓		270K
Infinity-Doc [66]	✓	✓					✓	✓	✓	400K
UniRec40M	✓	✓	✓	✓	✓	✓	✓	✓	✓	40M

Table 1: Comparison of Public Datasets and the Proposed UniRec40M Dataset.

TeX Generated						Scene Text			Handwritten			Domain Data			Sum
EN			CH												
Text	Formula	Mixed	Text	Formula	Mixed	LSVT	MTWI	HierText	HWDB	TAL	Note	IR Report	News-paper	K-12	
9.05	12.85	7.91	6.84	0.05	0.24	0.26	0.15	1.05	0.38	0.02	0.08	0.35	0.13	0.25	39.60
1.68	2.57	0.64	1.86	0.05	0.24	1.30	1.03	0.32	1.13	0.20	0.41	0.70	0.25	0.25	12.63

Table 2: Data composition of UniRec40M (in million). The last row presents the sampling counts of each epoch during training.

bilingual support, and unified coverage of both text and formula recognition at multiple levels.

3 UniRec40M Dataset

3.1 Data Composition

As shown in Fig. 2, UniRec40M is built from three data sources to ensure diversity across domains and modalities.

(1) Online TeX source [20]: We first collect arXiv TeX source files and Wikipedia HTML pages, which are also converted into TeX format. This results in approximately 2 million TeX files. For each TeX source, every valid text or formula token is assigned with a unique color by inserting the corresponding LaTeX color commands. The TeX files are then rendered into PDFs, producing documents with distinct colors for different textual and formulaic content. By performing color-based alignment between the TeX sources and the rendered PDFs, we can get labels at word and line levels, while the paragraph level can be obtained by further parsing LaTeX grammars. This pipeline enables the automatic generation of large-scale, multi-level data covering text, formula, and mixed text-formula content.

(2) Digital-born PDF documents: To broaden the distribution of document types, we collect digital-born PDFs such as industry research (IR) reports and newspapers. Then, we extract samples consisting of textual blocks and their cor-

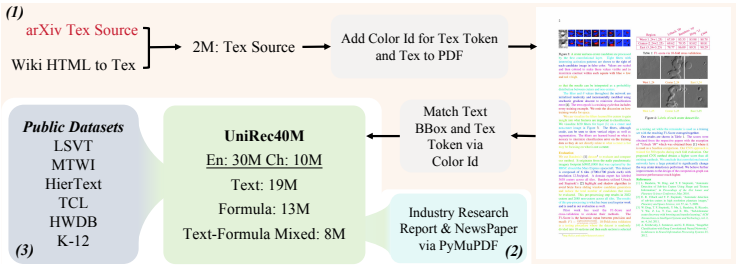

Fig. 2: The construction and composition of UniRec40M. Detailed descriptions are provided in Sec. 3.

responding image regions by PyMuPDF. This component primarily supplements the dataset with multi-level, multi-domain text recognition data.

(3) Public datasets: We further integrate a range of public datasets to enhance coverage across languages and domains. It includes: Chinese scene text datasets: LSVT [59], MTWI [31]. English scene text dataset: HierText [43]. Handwritten datasets: CASIA-HWDB [41], TAL [60]. Examination-style dataset: K-12 [60]. Additionally, we include handwritten notes (Note) labeled with Qwen3VL-235B-A30B [74] and manually refined, serving as high-quality handwritten supplements.

Tab. 2 summarizes the composition of UniRec40M, which has nearly 30M English and 10M Chinese samples. Specifically, the dataset includes 19M plain text, 13M formula-only, and 8M text-formula mixed instances. This large-scale, multilingual collection provides rich annotations for a broad coverage of real-world document formats, and thus can serve as the foundation for training our UniRec-0.1B.

3.2 Sampling Strategy

To ensure balanced learning across data modalities, we adopt a proportion-balanced sampling strategy. Each data type is either sub-sampled or re-sampled per training epoch, maintaining stable ratios among text, formula, and mixed samples. The last row of Tab. 2 reports the sampling counts used in each epoch during training.

4 UniRec-0.1B Model

The architecture of UniRec-0.1B follows a standard encoder-decoder framework, as illustrated in Fig. 3. The image encoder processes the input image using a native-resolution strategy, where the input image maintains its original aspect ratio, with the maximum width and height capped at 960 and 1408 pixels, respectively.

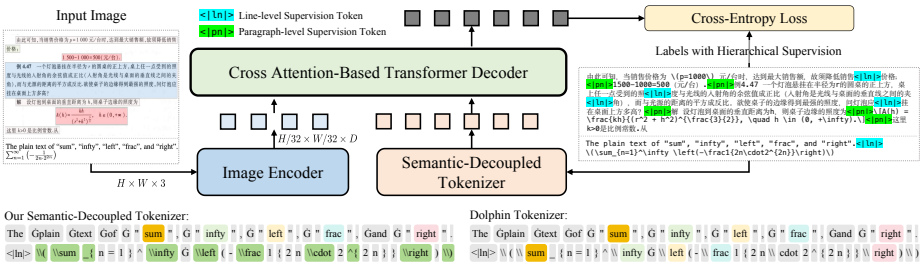

Fig. 3: The architecture overview of UniRec-0.1B. 'Ġ' represents a space. Detailed descriptions are provided in Sec. 4.

Formally, the input image is defined as $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$. The image encoder $\mathbf{Encoder}(\cdot)$ is implemented with FocalNet [76], producing a dense visual feature map:

$$\mathbf{F}_{map} = \mathbf{Encoder}(\mathbf{I}) \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times D}$$

where $D = 768$ denotes the feature dimensionality.

To obtain a sequence representation, the spatial dimensions of $\mathbf{F}_{map}$ are flattened into a set of visual tokens:

$$\mathbf{F} = \text{Flatten}(\mathbf{F}_{map}) \in \mathbb{R}^{N \times D}, \quad N = \frac{H}{32} \times \frac{W}{32}.$$

This tokenized feature sequence serves as the input to the subsequent multimodal decoder. The text and formula supervision `Label` is defined with hierarchical supervision tokens. A Semantic-Decoupled Tokenizer (SDT) is employed to decouple textual and formulaic elements, yielding a discrete token sequence:

$$\mathbf{Y} = \text{SDT}(\texttt{Label}) = \{\texttt{<BOS>}, y_0, y_1, \ldots, y_{l-1}, y_l\},$$

where `<BOS>` and y_l = `<EOS>` represent the beginning and end of sequence, respectively, and l denotes the token length. Each token is mapped to a continuous embedding via a text embedding layer $\mathbf{T} = \mathbf{E}_{text}(\mathbf{Y}) \in \mathbb{R}^{(l+2) \times D}$, where the embedding matrix is defined as $\mathbf{E}_{text} \in \mathbb{R}^{|V| \times D}$ and $|V|$ is the vocabulary size.

The decoder $\mathbf{Decoder}(\cdot)$ consists of six Transformer layers with cross-attention modules. Each layer uses a hidden size D and $D/64$ attention heads. Given the textual embeddings $\mathbf{T}$ and the visual features $\mathbf{F}$, the decoder performs autoregressive generation under a causal mask $\mathbf{M}_{causal}$:

$$\tilde{\mathbf{Y}} = \mathbf{Decoder}(\mathbf{T}, \mathbf{F}, \mathbf{M}_{causal}) = \{\tilde{y}_0, \tilde{y}_1, \ldots, \tilde{y}_l\}.$$

At each time step t, the model predicts the next token probability $P(\tilde{y}_t | \tilde{y}_{0:t-1}, \mathbf{F}) = \text{Softmax}(h_t W_o)$, where h_t is the decoder hidden state and $W_o \in \mathbb{R}^{D \times |V|}$ is the output projection matrix. The model is trained with a cross-entropy loss over all token positions:

$$\mathcal{L}_{CE} = -\sum_{t=0}^{l} \log P(y_t | y_{0:t-1}, \mathbf{F}).$$

4.1 Hierarchical Supervision Training

Text and formulas exhibit a natural hierarchical structure, typically organized into lines and paragraphs. Most recognition models treat the content as a flat sequence, ignoring the hierarchical spatial relationships between lines and paragraphs. This simplification limits the model's ability to learn the spatial layout representation.

To explicitly model this hierarchy, we introduce line-level and paragraph-level supervision tokens, `<|ln|>` and `<|pn|>`, during dataset construction, as shown in Fig. 3. The token `<|ln|>` denotes a line break within a paragraph, while `<|pn|>` indicates the end of a paragraph. These supervision signals encourage the model to learn hierarchical spatial dependencies and improve layout awareness during training. During inference, predicted `<|ln|>` tokens are removed, and `<|pn|>` tokens are replaced with two newline characters to reconstruct the paragraph structure in the generated output.

4.2 Semantic-Decoupled Tokenizer

Existing tokenizers are usually trained on a joint corpus that mixes plain text and formula. This approach often leads to semantic coupling between textual and formula tokens. For example, in the Dolphin Tokenizer (see Fig. 3), tokens such as *sum*, *infty*, *left*, *frac*, and *right* are assigned shared embeddings despite representing semantically distinct concepts in plain text and formula modalities. While LLMs can leverage contextual cues to disambiguate such cases, this coupling introduces unnecessary learning complexity for smaller models like UniRec-0.1B.

To mitigate this issue, we propose a semantic-decoupled tokenizer. Specifically, we train two independent tokenizers: one on plain text and another on mathematical formulas. The tokens from the formula tokenizer are then added to the text tokenizer as special tokens, excluding those already existing in the text tokenizer. As shown in Fig. 3, this method ensures that tokens such as *sum*, *infty*, *left*, *frac*, and *right* maintain distinct embeddings across modalities, thereby achieving effective semantic decoupling and reducing representational ambiguity during model training.

5 Experiments

5.1 Benchmark and Implementation Details

UniRec-Bench: Recent document parsing benchmarks, such as FoxPage [71], olmOCR-Bench [55], and OmniDocBench [53], primarily focus on evaluating page-level parsing capabilities. They lack fine-grained evaluation at the granularity of text block. To bridge this gap, we extend OmniDocBench [53] by extracting text, formula, and mixed text-formula blocks from full-page documents. Then, each text block is further categorized into five hierarchical levels and two languages. Consistent with OmniDocBench, the dataset spans nine document

Methods	Size	Modality				Level					Language		
		Avg	Text	Formula	Mix	Character	Word	Line	Paragraph	Multi-Paragraph	CH	EN	Mix
			14301	620	1314	167	2525	3549	7334	726	9448	3885	968
PP-OCRv5 [12]	42M	-	0.125	-	-	0.597	0.208	0.100	0.097	0.138	0.097	0.197	0.109
PP-Recv5 [12]	20M	-	-	-	-	0.033	0.056	0.094	-	-	-	-	-
OpenOCR-Rec [19]	30M	-	-	-	-	0.054	0.061	0.079	-	-	-	-	-
Mathpix [49]	-	-	-	0.322	-	-	-	-	-	-	-	-	-
Pix2Tex [46]	23M	-	-	0.337	-	-	-	-	-	-	-	-	-
UniMERNet-B [67]	0.3B	-	-	0.238	-	-	-	-	-	-	-	-	-
Dolphin [24]	0.3B	0.283	0.118	0.502	0.227	0.072	0.070	0.065	0.153	0.204	0.141	0.047	0.183
Dolphin-1.5 [24]	0.3B	0.206	0.050	0.365	0.202	0.061	0.059	0.056	0.037	0.115	0.049	0.038	0.110
MonkeyOCR-Pro [39]	3B	0.205	0.297	0.177	0.142	0.043	0.055	0.111	0.490	0.157	0.408	0.065	0.147
dots.ocr [38]	3B	0.176	0.113	0.221	0.194	0.129	0.148	0.192	0.065	**0.092**	0.121	0.096	0.107
MinerU2.5 [51]	1.2B	0.154	0.167	0.140	0.155	0.033	0.051	0.068	0.260	0.153	0.217	0.055	0.138
Nanonets-OCR2 [48]	3B	0.245	0.312	0.193	0.229	0.174	0.077	0.099	0.504	0.263	0.392	0.139	0.221
DeepSeek-OCR [72]	3B-A0.5B	0.168	0.103	0.238	0.162	0.794	0.219	0.061	0.067	0.107	0.103	0.106	0.091
PaddleOCR-VL [11]	0.9B	**0.100**	0.041	**0.125**	0.135	0.037	0.042	0.046	**0.031**	0.107	0.044	**0.022**	0.087
UniRec-0.1B	0.1B	**0.100**	**0.038**	0.134	**0.128**	**0.025**	**0.032**	**0.043**	0.032	0.099	**0.041**	0.023	**0.071**
w/o HST *w* SDT	0.1B	0.113	0.050	0.144	0.143	0.030	0.040	0.047	0.049	0.111	0.053	0.034	0.083
w/o HST *w/o* SDT	0.1B	0.159	0.062	0.255	0.161	0.054	0.066	0.066	0.053	0.114	0.070	0.035	0.092

Table 3: Comparison with text recognition expert models, formula recognition expert models, and document parsing expert models on UniRec-Bench across modalities, levels, and languages. The last two rows present ablation experiments for UniRec-0.1B.

domains. This process yields UniRec-Bench (see Tab. 3 and Tab. 4), a comprehensive benchmark that emphasizes multi-type, level, lingual, and domain text block evaluation. Benefiting from its fine-grained categorization, UniRec-Bench can support the inspection of different parsing models at the block level, which is largely missed in existing studies.

Training: We use AdamW optimizer [45] with a weight decay of 0.01 for training. The learning rate (LR) is set to 1×10^{-4} and the global batchsize is set to 64. One cycle LR scheduler [44] with 0.5 epochs linear warm-up is used throughout the 10 training epochs. Data augmentation like rotation, distortion, motion blur and gaussian noise, are randomly performed. The maximum token length is set to 1024. The vocabulary size $|V|$ is 56371. Training takes approximately 80 hours on 8 A800 40GB GPUs. Note that UniRec-0.1B is trained from scratch without loading pre-trained models. The accuracy is computed based on the edit distance between the model prediction and the ground-truth.

5.2 Ablation Study

We carry out ablation studies to assess the effectiveness of the proposed HST and SDT on UniRec-0.1B. The results are summarized in Tab. 3 and Tab. 4.

Effectiveness of HST. The second-to-last row in Tab. 3 and Tab. 4 reports the performance without HST. Incorporating HST improves the edit distance by 1.2% on text, 1.0% on formulas, and 1.5% on their mix, indicating consistent benefits across all scenarios. For multi-level text, the five levels get 0.5%, 0.8%, 0.4%, 1.7%, and 1.2% improvements, respectively. Gains are particularly notable at the paragraph and multi-paragraph levels, demonstrating that hierarchical supervision effectively captures structural dependencies across lines and paragraphs. Meanwhile, benefits are also observed across languages. For multiple document domains, the most pronounced improvements are observed on

Methods	Size	Domain								
		Book	PPT2PDF	Research Report	Textbook	Exam Paper	Magazine	Literature	Note	Newspaper
		970	488	782	994	1579	1040	1157	812	6479
PP-OCRv5 [12]	42M	0.100	0.101	0.051	0.114	0.155	0.081	0.088	0.250	0.132
Dolphin [24]	0.3B	0.033	0.060	0.040	0.069	0.119	0.043	0.021	0.238	0.167
Dolphin-1.5 [24]	0.3B	0.024	**0.045**	0.027	**0.039**	0.075	0.029	0.020	0.184	0.045
MonkeyOCR-Pro [39]	3B	0.025	0.075	0.030	0.044	0.082	0.023	**0.010**	0.200	0.585
dots.ocr [38]	3B	0.027	0.064	0.034	0.056	0.085	0.028	0.021	0.126	0.184
MinerU2.5 [51]	1.2B	0.024	0.107	0.031	0.044	0.086	**0.019**	0.015	0.151	0.302
Nanonets-OCR2 [48]	3B	0.067	0.124	0.051	0.097	0.133	0.061	0.092	0.265	0.556
DeepSeek-OCR [72]	3B-A0.5B	0.071	0.136	0.053	0.111	0.137	0.061	0.037	0.224	0.105
PaddleOCR-VL [11]	0.9B	**0.021**	0.060	0.023	0.042	0.085	**0.019**	0.013	0.067	0.039
UniRec-0.1B	0.1B	0.025	0.063	**0.018**	0.054	**0.058**	0.021	0.012	**0.055**	**0.038**
w/o HST *w* SDT	0.1B	0.026	0.091	0.024	0.071	0.087	0.026	0.015	0.073	0.050
w/o HST *w/o* SDT	0.1B	0.025	0.120	0.030	0.065	0.090	0.033	0.019	0.075	0.071

Table 4: Comparison with existing models and ablations across domains on UniRec-Bench.

PPT2PDF and *Exam Paper* subsets, with increases of 2.8% and 2.9%, respectively. This is mainly because PPT and exam documents often have complex layouts compared to other domains, and HST better models such structural complexity, leading to enhanced recognition.

Effectiveness of SDT. The last row in Tab. 3 and Tab. 4 shows the results further without SDT. Using SDT yields 1.2% improvement on text, 11.1% on formulas, and 1.8% on their mix. The results confirm the benefit of SDT. Specifically, the substantial gain in formula recognition convincingly suggests that decoupling formula and text tokens alleviates semantic entanglement between modalities. In addition, SDT slightly decreases performance in certain domains, such as *Textbook*, with a 0.6% drop. Nevertheless, when SDT is combined with HST, all domains benefit, leading to an average gain of 2.4%.

5.3 Comparison with State-of-the-arts

We compare UniRec-0.1B with state-of-the-art text recognition expert models, formula recognition expert models, document parsing models, and general VLMs on UniRec-Bench and OmniDocBench [53]. UniRec-Bench focuses on text and formula block recognition, while OmniDocBench [53] targets full-page parsing.

Results on UniRec-Bench Comparison with text recognition expert models. For text recognition, we benchmark against advanced open-source models, including PP-OCRv5 [12], PP-Recv5 [12], and OpenOCR-Rec [19]. PP-OCRv5 combines a text-line detector (PP-Detv5) and a recognizer (PP-Recv5) to perform paragraph-level text recognition. Since PP-Recv5 and OpenOCR-Rec operate only text instances rather than paragraphs, they are evaluated at the character-, word-, and line-levels. Although the two small expert models are known for their strong performance in line-level recognition, results in Tab. 3 show that UniRec-0.1B consistently outperforms them across multiple text levels in Chinese, English, and mixed-language text. In addition, UniRec-0.1B largely surpasses the pipeline-based PP-OCRv5 across all domains, as shown in Tab. 4.

Method Type	Methods	Overall$^{\text{Edit}}\downarrow$	Overall$^{\text{Edit}}\downarrow$		Text$^{\text{Edit}}\downarrow$		Formula$^{\text{Edit}}\downarrow$		Table$^{\text{TEDS}}\uparrow$		Table$^{\text{Edit}}\downarrow$		Reading Order$^{\text{Edit}}\downarrow$	
			EN	CH	EN	CH	EN	CH	EN	CH	EN	CH	EN	CH
Pipeline Models	Docling-2.14.0 [61]	0.749	0.589	0.909	0.416	0.987	0.999	1	61.3	25.0	0.627	0.810	0.313	0.837
	OpenParse-0.7.0 [25]	0.730	0.646	0.814	0.681	0.974	0.996	1	64.8	27.5	0.284	0.639	0.595	0.641
	Unstructured-0.17.2 [65]	0.651	0.586	0.716	0.198	0.481	0.999	1	0	0.1	1	0.998	0.145	0.387
	Pix2Text-1.1.2.3 [6]	0.424	0.320	0.528	0.138	0.356	0.276	0.611	73.6	66.2	0.584	0.645	0.281	0.499
	Marker-1.7.1 [54]	0.397	0.296	0.497	0.085	0.293	0.374	0.688	67.6	54.0	0.609	0.678	0.116	0.329
	Mathpix [49]	0.278	0.191	0.364	0.105	0.381	0.306	0.454	77.0	67.1	0.243	0.320	0.108	0.304
	MinerU-pipeline [68]	0.203	0.162	0.244	0.072	0.111	0.313	0.581	77.4	79.5	0.166	0.150	0.097	0.136
	PP-StructureV3 [12]	0.176	0.145	0.206	0.058	0.088	0.295	0.535	77.2	83.9	0.159	0.109	0.069	0.091
General VLMs	InternVL2-76B [9]	0.442	0.440	0.443	0.353	0.290	0.543	0.701	63.0	60.2	0.547	0.555	0.317	0.228
	GPT-4o [1]	0.316	0.233	0.399	0.144	0.409	0.425	0.606	72.0	62.9	0.234	0.329	0.128	0.251
	InternVL3-78B [87]	0.257	0.218	0.296	0.117	0.210	0.380	0.533	69.0	73.9	0.279	0.282	0.095	0.161
	Qwen2.5-VL-72B [2]	0.238	0.214	0.261	0.092	0.180	0.315	0.434	81.4	83.0	0.341	0.262	0.106	0.168
	Gemini2.5-Pro [26]	0.180	0.148	0.212	0.055	0.168	0.356	0.439	85.8	86.4	0.130	0.119	0.049	0.121
Specialized VLMs	Nougat [5]	0.713	0.452	0.973	0.365	0.998	0.488	0.941	39.9	0.0	0.572	1	0.382	0.954
	SmolDocling-256M [50]	0.655	0.493	0.816	0.262	0.838	0.753	0.997	44.9	16.5	0.729	0.907	0.227	0.522
	olmOCR-7B [55]	0.398	0.326	0.469	0.097	0.293	0.455	0.655	68.1	61.3	0.608	0.652	0.145	0.277
	GOT [71]	0.349	0.287	0.411	0.189	0.315	0.360	0.528	53.2	47.2	0.459	0.520	0.141	0.280
	OCRFlux-3B [10]	0.294	0.238	0.349	0.112	0.256	0.447	0.716	69.0	80.0	0.269	0.162	0.126	0.263
	Nanonets-OCR-s [48]	0.289	0.283	0.295	0.134	0.231	0.518	0.546	76.8	79.4	0.343	0.201	0.135	0.200
	Dolphin [24]	0.259	0.205	0.313	0.092	0.204	0.447	0.606	76.1	66.9	0.193	0.282	0.088	0.160
	MinerU2-VLM [52]	0.186	0.133	0.238	0.045	0.115	0.273	0.506	82.1	83.4	0.150	0.209	0.066	0.122
	MonkeyOCR-1.2B [39]	0.184	0.146	0.221	0.068	0.118	0.272	0.452	81.3	85.5	0.149	0.134	0.093	0.179
	MonkeyOCR-3B [39]	0.172	0.138	0.206	0.067	0.107	0.246	0.421	81.5	87.5	0.139	0.111	0.100	0.185
	dots.ocr [38]	0.143	0.125	0.160	**0.032**	0.066	0.329	0.416	**88.6**	89.0	0.099	0.092	**0.040**	0.067
	DeepSeek-OCR [72]	0.158	0.134	0.181	0.046	0.097	0.285	0.433	82.6	89.0	0.138	0.088	0.067	0.105
	olmOCR2 [55]	0.214	0.161	0.267	0.048	0.185	0.392	0.543	83.7	78.5	0.123	0.165	0.081	0.174
	MinerU2.5 [51]	0.143	0.111	0.174	0.050	0.074	0.258	0.473	88.3	89.2	**0.089**	0.083	0.045	0.068
	w UniRec-0.1B	0.120	0.107	0.133	0.044	0.068	0.247	0.314	88.3	89.2	**0.089**	0.083	0.047	0.067
	PaddleOCR-VL [11]	0.115	0.105	0.126	0.041	**0.062**	0.241	0.316	88.0	**92.1**	0.093	**0.062**	0.045	**0.063**
	w UniRec-0.1B	**0.113**	**0.102**	**0.123**	0.040	0.065	**0.234**	**0.298**	88.0	**92.1**	0.093	**0.062**	0.042	0.065

Table 5: Evaluation of document parsing on OmniDocBench [53]. *w* UniRec-0.1B means replacing the text and formula recognition modules of MinerU2.5 [51] or PaddleOCR-VL [11] by UniRec-0.1B, while keep their layout analysis and table recognition remain unchanged.

These results indicate that unified multi-level text recognition not only eliminates the need for a separate text detection module but also improves the overall accuracy, validating the effectiveness of UniRec-0.1B.

Comparison with formula recognition expert models. For formula recognition, we compare ours with commercial Mathpix [49], open-source Pix2Tex [46] and UniMERNet-B [67]. As reported in Tab. 3, UniRec-0.1B outperforms them by 18.8%, 20.3%, and 10.4%, respectively. The performance gain mainly stems from the rich and diverse formula recognition data in UniRec40M that leads to sufficient model training, and the decoupled tokenizer design that well separates textual and formula semantics.

Comparison with document parsing models. UniRec-0.1B also exhibits clear advantages when compared with VLM-based document parsing models. Against Dolphin-1.5 [24], a model with 0.3B parameters, UniRec-0.1B achieves improvements of 1.2% on text, 23.1% on formulas, and 7.4% on their mix. The substantial gain in formula recognition is largely attributed to UniRec-0.1B's text-formula SDT design, whereas Dolphin-1.5 uses a semantically coupled tokenizer, making it difficult to distinguish text from formula elements at small parameter scales. Moreover, as reported in Tab. 6, UniRec-0.1B is more than 2× faster than Dolphin-1.5. Compared to PaddleOCR-VL [11], UniRec-0.1B achieves comparable performance using only 11% of its parameters, consistently

Methods	Size	Character	Word	Line	Paragraph	Multi-Paragraph	Block_{avg}	Page_{avg}
MonkeyOCR [39]	3B	0.64	0.76	1.25	4.43	5.60	3.62	58.39
DeepSeek-OCR [72]	3B-A0.5B	0.18	0.94	0.93	4.33	6.92	3.47	58.95
MinerU2.5 [51]	1.2B	0.08	0.24	0.66	3.16	7.92	2.54	42.72
PaddleOCR-VL [11]	0.9B	0.60	0.66	0.9	2.29	3.32	1.88	31.92
Dolphin-1.5 [24]	0.3B	0.10	0.14	0.34	0.95	1.14	0.78	13.16
UniRec-0.1B	0.1B	**0.03**	**0.05**	**0.10**	**0.54**	**0.62**	**0.37**	**6.20**

Table 6: Inference speed (in second) comparison across models. The speed is measured on a single A800 40GB GPU. For fairness, all models are run without inference acceleration, using Torch or PaddlePaddle in dynamic graph mode with KV Cache and a batch size of 1.

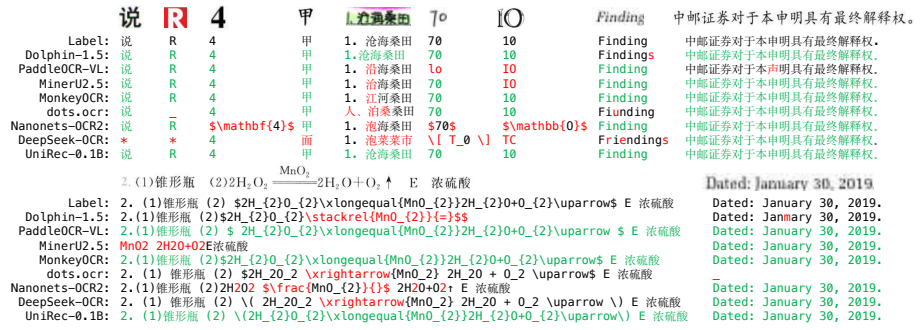

Fig. 4: Recognition result visualizations. Red characters indicate recognition errors, _ indicate missed characters, and * indicate recognition results that are meaningless hallucination. Green indicates correct recognition. More cases are presented in the supplementary.

across multiple levels, languages, and domains. Notably, in the *Exam Paper* domain, UniRec-0.1B outperforms PaddleOCR-VL by 2.7%. This prominent gain is likely due to the inclusion of exam-related samples in UniRec40M, highlighting the importance of incorporating multi-domain data during dataset construction. Regarding inference speed, compared to PaddleOCR-VL, UniRec-0.1B reduces the inference time from 1.88 s to 0.37 s on the block level, and from 31.92 s to 6.2 s on the page level, both exceeding 5× speedup.

Furthermore, UniRec-Bench reveals that end-to-end full-page parsing models such as dots.ocr [38], Nanonets-OCR2 [48], and DeepSeek-OCR [72] perform worse at the character, word, and line levels. The first eight cases in Fig. 4 also confirm this observation. In contrast, multi-stage methods such as Dolphin [24], MonkeyOCR [39], MinerU2.5 [51], and PaddleOCR-VL [11] excel in fine-grained character-, word-, and line-level recognition. End-to-end methods more emphasize holistic perception, while multi-stage methods integrate multiple expert models that are good at specialized fine-grained recognition. Nevertheless, our UniRec-0.1B absorbs advantages from both sides. It not only achieves

SOTA accuracy in character, word, and line levels, but also attains highly competitive accuracy in paragraph and multi-paragraph levels. These results again validate the effectiveness of UniRec-0.1B.

Note that in *Note* domain, UniRec-0.1B gets the best accuracy, surpassing PaddleOCR-VL by 1.2%, while other models lag significantly behind. This suggests that most methods are optimized for machine-written documents but struggle with handwritten ones. Similarity, in *Newspaper* domain, UniRec-0.1B again achieves the best result, with PaddleOCR-VL and Dolphin-1.5 performing nearby, and the rest fall far behind. This is explained as *Newspaper* pages often contain blurred or degraded text. The results indicate that most existing methods are biased toward digital-born documents and tend to fail on visually noisy inputs.

Results on OmniDocBench To evaluate the document parsing performance of UniRec-0.1B, we conduct experiments summarized in Tab. 5. Specifically, we integrate UniRec-0.1B into two leading two-stage document parsing methods, MinerU2.5 and PaddleOCR-VL, by replacing their text and formula recognition modules with UniRec-0.1B. The two methods perform layout analysis in the first stage, followed by region-level recognition in the second stage, where detected document components (e.g., text or formula regions) are cropped and recognized individually. In our setup, the second-stage recognition of text and formula regions is implemented by UniRec-0.1B.

As shown in Tab. 5, when combined with MinerU2.5, the full-page edit distance improves from 0.143 to 0.120, achieving a 2.3% performance gain. Moreover, as reported in Tab. 6, UniRec-0.1B reduces the average page parsing time from 42.72 s to 6.2 s, resulting in a nearly 7× speedup. A similar trend is observed with PaddleOCR-VL, where the full-page edit distance improves further by 0.2%, reaching the new SOTA, and the parsing time accelerated significantly. These results clearly demonstrate the practicality and effectiveness of UniRec-0.1B in document parsing systems. Moreover, they also verify the necessity and benefit of unified recognition of both text and formulas with a lightweight model.

6 Conclusion

In this paper, we have presented UniRec-0.1B, a 0.1B ultra-lightweight model that is capable of jointly recognizing text and formulas across multiple hierarchical levels effective and efficient. To build this model, we first construct UniRec40M, a comprehensively labeled dataset comprising approximately 40 million text, formula, and their mix samples in both Chinese and English. It fills the data gap in unified text-formula recognition. Furthermore, we have made two key innovations to unleash the recognition capability of this small model: Hierarchical Supervision Training and Semantics-Decoupled Tokenizer. The former explicitly models the hierarchical relationships between lines and paragraphs, while the latter decouples token representations for text and formulas. UniRec-0.1B, the resulted model, has been extensively validated on the

popular OmniDocBench and the newly constructed UniRec-Bench, which evaluates text and formula at document blocks of multiple levels, languages, and domains. The results demonstrate that UniRec-0.1B achieves leading accuracy compared to existing models, meanwhile it achieves 2-9× speedup in inference speed when parsing documents. This clearly validates its effectiveness, and the practicality that we use an ultra-lightweight model for unified text and formula recognition. Nevertheless, the results on UniRec-Bench reveal some limitations of existing models, e.g., excelling at paragraph-level recognition but struggling with fine-grained recognition. These constitute future research directions. Overall, we hope that our UniRec-0.1B, UniRec40M, and UniRec-Bench will continue to advance the field of document parsing.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 625B2057, and in part by the Science and Technology Commission of Shanghai Municipality under Grant 25511104000

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: ECCV. pp. 178–196. Springer (2022)
4. Blecher, L.: pix2tex: LaTeX-OCR – using a ViT to convert images of equations into LaTeX code. https://github.com/lukas-blecher/LaTeX-OCR (2021)
5. Blecher, L., Cucurull, G., Scialom, T., Stojnic, R.: Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418 (2023)
6. breezedeus: Pix2text. `https://github.com/breezedeus/Pix2Text` (2022), accessed: 2025-06-23
7. Chen, J., Yu, H., Ma, J., Guan, M., Xu, X., Wang, X., Qu, S., Li, B., Xue, X.: Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. CoRR **abs/2112.15093** (2021)
8. Chen, S., Guo, X., Li, Y., Zhang, T., Lin, M., Kuang, D., Zhang, Y., Ming, L., Zhang, F., Wang, Y., et al.: Ocean-ocr: Towards general ocr application via a vision-language model. arXiv preprint arXiv:2501.15558 (2025)
9. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR. pp. 24185–24198 (2024)
10. chatdoc com: Ocrflux. `https://github.com/chatdoc-com/OCRFlux` (2025), accessed:2025-09-25
11. Cui, C., Sun, T., Liang, S., Gao, T., Zhang, Z., Liu, J., Wang, X., Zhou, C., Liu, H., Lin, M., et al.: Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. arXiv preprint arXiv:2510.14528 (2025)

12. Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., et al.: Paddleocr 3.0 technical report. arXiv preprint arXiv:2507.05595 (2025)
13. Deng, Y., Kanervisto, A., Ling, J., Rush, A.M.: Image-to-markup generation with coarse-to-fine attention. In: ICML. pp. 980–989 (2017)
14. Du, Y., Chen, P., Ying, X., Chen, Z.: Docptbench: Benchmarking end-to-end photographed document parsing and translation. arXiv preprint arXiv:2511.18434 (2025)
15. Du, Y., Chen, Z., Jia, C., Gao, X., Jiang, Y.G.: Out of length text recognition with sub-string matching. In: AAAI. pp. 2798–2806 (2025)
16. Du, Y., Chen, Z., Jia, C., Yin, X., Li, C., Du, Y., Jiang, Y.G.: Context perception parallel decoder for scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **47**(6), 4668–4683 (2025). https://doi.org/10.1109/TPAMI.2025.3545453
17. Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., Jiang, Y.G.: SVTR: Scene text recognition with a single visual model. In: IJCAI. pp. 884–890 (2022)
18. Du, Y., Chen, Z., Su, Y., Jia, C., Jiang, Y.G.: Instruction-guided scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **47**(4), 2723–2738 (2025). https://doi.org/10.1109/TPAMI.2025.3525526
19. Du, Y., Chen, Z., Xie, H., Jia, C., Jiang, Y.G.: SVTRv2: Ctc beats encoder-decoder models in scene text recognition. In: ICCV. pp. 20147–20156 (2025)
20. Duan, C., Tan, Z., Bartsch, S.: LaTeX rainbow: Universal LaTeX to PDF document semantic & layout annotation framework. In: Proceedings of the Second Workshop on Information Extraction from Scientific Publications. pp. 56–67 (2023)
21. Fang, S., Mao, Z., Xie, H., Wang, Y., Yan, C., Zhang, Y.: ABINet++: Autonomous, bidirectional and iterative language modeling for scene text spotting. IEEE Trans. Pattern Anal. Mach. Intell. **45**(6), 7123–7141 (2023)
22. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: CVPR. pp. 7098–7107 (2021)
23. Feng, H., Shi, W., Zhang, K., Fei, X., Liao, L., Yang, D., Du, Y., Wu, X., Tang, J., Liu, Y., Chen, H., Huang, C.: Dolphin-v2: Universal document parsing via scalable anchor prompting. arXiv preprint arXiv:2602.05384 (2026)
24. Feng, H., Wei, S., Fei, X., Shi, W., Han, Y., Liao, L., Lu, J., Wu, B., Liu, Q., Lin, C., et al.: Dolphin: Document image parsing via heterogeneous anchor prompting. arXiv preprint arXiv:2505.14059 (2025)
25. Filimoa: open-parse. https://github.com/Filimoa/open-parse (2024), accessed: 2025-06-23
26. Google DeepMind: Gemini 2.5. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/ (2025)
27. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: ICML. pp. 369–376 (2006)
28. Guan, T., Shen, W., Yang, X., Feng, Q., Jiang, Z., Yang, X.: Self-Supervised Character-to-Character distillation for text recognition. In: ICCV. pp. 19473–19484 (2023)
29. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
30. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: CVPR. pp. 2315–2324 (2016)

31. He, M., Liu, Y., Yang, Z., Zhang, S., Luo, C., Gao, F., Zheng, Q., Wang, Y., Zhang, X., Jin, L.: ICPR2018 contest on robust reading for multi-type web images. In: ICPR. pp. 7–12 (2018)
32. Hu, W., Cai, X., Hou, J., Yi, S., Lin, Z.: Gtc: Guided training of ctc towards efficient and accurate scene text recognition. In: AAAI. vol. 34, pp. 11005–11012 (2020)
33. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. CoRR **abs/1406.2227** (2014)
34. Jiang, Q., Wang, J., Peng, D., Liu, C., Jin, L.: Revisiting scene text recognition: A data perspective. In: ICCV. pp. 20486–20497 (2023)
35. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: ECCV. pp. 498–517. Springer (2022)
36. Li, G., Lyu, P., Zhang, C., Shen, H., Wu, L., Wan, X., Zeng, G., Hu, H., Ma, C., Zhou, Y.: Towards real-world document parsing via realistic scene synthesis and document-aware training. arXiv preprint arXiv:2603.23885 (2026)
37. Li, H., Wang, P., Shen, C., Zhang, G.: Show, attend and read: A simple and strong baseline for irregular text recognition. In: AAAI. pp. 8610–8617 (2019)
38. Li, Y., Yang, G., Liu, H., Wang, B., Zhang, C.: dots.ocr: Multilingual document layout parsing in a single vision-language model (2025)
39. Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Guo, Z., Zhang, J., Wang, X., Bai, X.: Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. arXiv preprint arXiv:2506.05218 (2025)
40. Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D., et al.: Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434 (2024)
41. Liu, C.L., Yin, F., Wang, D.H., Wang, Q.F.: Casia online and offline chinese handwriting databases. In: ICDAR. pp. 37–41 (2011)
42. Liu, Y., Yang, B., Liu, Q., Li, Z., Ma, Z., Zhang, S., Bai, X.: Textmonkey: An ocr-free large multimodal model for understanding document. arXiv preprint arXiv:2403.04473 (2024)
43. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Towards end-to-end unified scene text detection and layout analysis. In: CVPR. pp. 1049–1059 (2022)
44. Ilya Loshchilov, Hutter, F.: SGDR: stochastic gradient descent with warm restarts. In: ICLR (2017)
45. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
46. Lukas Blecher: pix2tex - latex ocr. `https://github.com/lukas-blecher/LaTeX-OCR` (2022), accessed: 2025-06-23
47. Luo, C., Jin, L., Sun, Z.: MORAN: A multi-object rectified attention network for scene text recognition. Pattern Recognit. **90**, 109–118 (2019)
48. Mandal, S., Talewar, A., Ahuja, P., Juvatkar, P.: Nanonets-ocr-s: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging (2025)
49. Mathpix: Mathpix snip: Convert images and pdfs to latex, docx, and more. `https://mathpix.com/` (2025)
50. Nassar, A., Marafioti, A., Omenetti, M., Lysak, M., Livathinos, N., Auer, C., Morin, L., de Lima, R.T., Kim, Y., Gurbuz, A.S., et al.: Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion. arXiv preprint arXiv:2503.11576 (2025)

51. Niu, J., Liu, Z., Gu, Z., Wang, B., Ouyang, L., Zhao, Z., Chu, T., He, T., Wu, F., Zhang, Q., et al.: Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. arXiv preprint arXiv:2509.22186 (2025)
52. opendatalab: Mineru2.0-2505-0.9b. https://huggingface.co/opendatalab/MinerU2.0-2505-0.9B (2025)
53. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., et al.: Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In: CVPR. pp. 24838–24848 (2025)
54. Paruchuri, V.: Marker. https://github.com/datalab-to/marker (2025), accessed: 2025-09-25
55. Poznanski, J., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Rangapur, A., Wilhelm, C., Lo, K., Soldaini, L.: olmocr: Unlocking trillions of tokens in pdfs with vision language models. arXiv preprint arXiv:2502.18443 (2025)
56. Sheng, F., Chen, Z., Xu, B.: NRTR: A no-recurrence sequence-to-sequence model for scene text recognition. In: ICDAR. pp. 781–786 (2019)
57. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2017). https://doi.org/10.1109/TPAMI.2016.2646371
58. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: ASTER: An attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell. **41**(9), 2035–2048 (2019)
59. Sun, Y., Ni, Z., Chng, C.K., Liu, Y., Luo, C., Ng, C.C., Han, J., Ding, E., Liu, J., Karatzas, D., Chan, C.S., Jin, L.: ICDAR 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In: ICDAR. pp. 1557–1562 (2019)
60. TAL: Tal open dataset. https://ai.100tal.com/dataset (2023)
61. Team, D.: Docling. https://github.com/docling-project/docling (2024), accessed: 2025-06-23
62. Team, H.V., Lyu, P., Wan, X., Li, G., Peng, S., Wang, W., Wu, L., Shen, H., Zhou, Y., Tang, C., Yang, Q., Peng, Q., Luo, B., Yang, H., Zhang, X., Zhang, J., Peng, H., Yang, H., Xie, S., Zhou, L., Pei, G., Wu, B., Yan, R., Wu, K., Yang, J., Wang, B., Liu, K., Zhu, J., Jiang, J., Linus, Hu, H., Zhang, C.: Hunyuanocr technical report (2025)
63. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
64. Tencent: Hunyuan-0.5b. https://github.com/Tencent-Hunyuan/Hunyuan-0.5B (2025)
65. Unstructured-IO: unstructured. https://github.com/Unstructured-IO/unstructured (2022), accessed: 2025-06-23
66. Wang, B., Wu, B., Li, W., Fang, M., Huang, Z., Huang, J., Wang, H., Liang, Y., Chen, L., Chu, W., Qi, Y.: Infinity parser: Layout aware reinforcement learning for scanned document parsing (2025)
67. Wang, B., Gu, Z., Liang, G., Xu, C., Zhang, B., Shi, B., He, C.: UniMERNet: A universal network for real-world mathematical expression recognition. arXiv preprint arXiv:2404.15254 (2024)
68. Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., et al.: Mineru: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839 (2024)
69. Wang, P., Da, C., Yao, C.: Multi-Granularity Prediction for scene text recognition. In: ECCV. pp. 339–355 (2022)

70. Wang, Y., Xie, H., Fang, S., Xing, M., Wang, J., Zhu, S., Zhang, Y.: PETR: Rethinking the capability of transformer-based language model in scene text recognition. IEEE Trans. Image Process. **31**, 5585–5598 (2022)
71. Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al.: General ocr theory: Towards ocr-2.0 via a unified end-to-end model. arXiv preprint arXiv:2409.01704 (2024)
72. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)
73. Wu, J.W., Yin, F., Zhang, Y.M., Zhang, X.Y., Liu, C.L.: Handwritten mathematical expression recognition via paired adversarial learning. IJCV **128**, 2386–2401 (2020)
74. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
75. Yang, B., Wen, B., Ding, B., Liu, C., Chu, C., Song, C., Rao, C., Yi, C., Li, D., Zang, D., et al.: Kwai keye-vl 1.5 technical report. arXiv preprint arXiv:2509.01563 (2025)
76. Yang, J., Li, C., Dai, X., Gao, J.: Focal modulation networks. In: NeurIPS. pp. 4203–4217 (2022)
77. Yu, D., Li, X., Zhang, C., Han, J., Liu, J., Ding, E.: Towards accurate scene text recognition with semantic reasoning networks. In: CVPR. pp. 12113–12122 (2020)
78. Yue, X., Kuang, Z., Lin, C., Sun, H., Zhang, W.: RobustScanner: Dynamically enhancing positional clues for robust text recognition. In: ECCV. pp. 135–151 (2020)
79. Zhang, J., Du, J., Yang, Y., Song, Y.Z., Wei, S., Dai, L.: A tree-structured decoder for image-to-markup generation. In: ICML. pp. 11076–11085. PMLR (2020)
80. Zhang, J., Du, J., Zhang, S., Liu, D., Hu, Y., Hu, J., Wei, S., Dai, L.: Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. PR **71**, 196–206 (2017)
81. Zhang, Q., Wang, B., Huang, V.S.J., Zhang, J., Wang, Z., Liang, H., He, C., Zhang, W.: Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. arXiv preprint arXiv:2410.21169 (2024)
82. Zhao, S., Du, Y., Chen, Z., Jiang, Y.G.: Decoder pre-training with only text for scene text recognition. In: ACM MM. pp. 5191–5200 (2024)
83. Zhao, S., Quan, R., Zhu, L., Yang, Y.: CLIP4STR: A simple baseline for scene text recognition with pre-trained vision-language model. IEEE Trans. Image Process. **33**, 6893–6904 (2024)
84. Zhao, Z., Tang, J., Lin, C., Wu, B., Huang, C., Liu, H., Tan, X., Zhang, Z., Xie, Y.: Multi-modal in-context learning makes an ego-evolving scene text recognizer. In: CVPR. pp. 15567–15576 (2024)
85. Zheng, T., Chen, Z., Fang, S., Xie, H., Jiang, Y.G.: CDistNet: Perceiving multi-domain character distance for robust text recognition. Int. J. Comput. Vis. **132**(2), 300–318 (2024)
86. Zhong, H., Yang, Z., Li, Z., Wang, P., Tang, J., Cheng, W., Yao, C.: Vl-reader: Vision and language reconstructor is an effective scene text recognizer. In: ACM MM. pp. 4207–4216 (2024)
87. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)