UniOne: A Document Parsing Dataset for Cross Task Association Modeling (Supplementary Materials)

Anonymous Authors

Anonymous Institute

1 Introduction

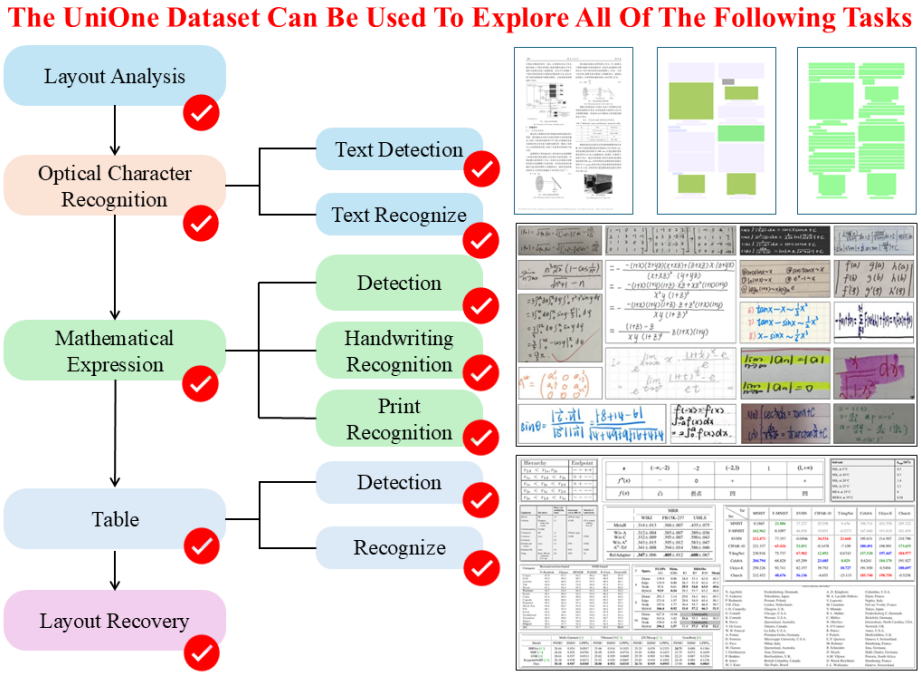

Fig. 1. The scope that the UniDoc dataset can cover in the field of document parsing

Currently, document parsing technology is facing the problem of "task silos" caused by fragmented datasets. With the widespread application of deep learning in document understanding, the construction of a unified and shared dataset for collaborative development of upstream and downstream tasks has become an inevitable trend. We have built the first UniOne document dataset that supports the parsing of upstream and downstream tasks. By systematically

integrating tasks such as layout analysis, text line detection and recognition, and table recognition, we have innovatively established a cross-task annotation dataset. This dataset: (1) at the layout analysis level, includes 236,790 paragraph-level annotations across 14,481 pages, covering 11 semantic categories; (2) at the text line detection level, based on the layout analysis data, further adds fine-grained annotations for 340,890 lines in 198,901 text paragraphs; (3) for complex scenarios, it introduces 8,000 challenging handwritten mathematical expressions, 18,717 printed mathematical formulas, 26,849 formula texts with unified recognition annotations, and 1,169 tables extracted from papers to fully support document content parsing. To our knowledge, this dataset is the first to achieve cross-task joint modeling from macro layouts to micro elements, breaking through the limitations of traditional single-task datasets and providing essential infrastructure for building the next generation of intelligent document parsing systems. The UniOne dataset can be accessed here: https://github.com/MaxTEX310/UniOne.

2 UniOne

2.1 Layout Analysis Data Section

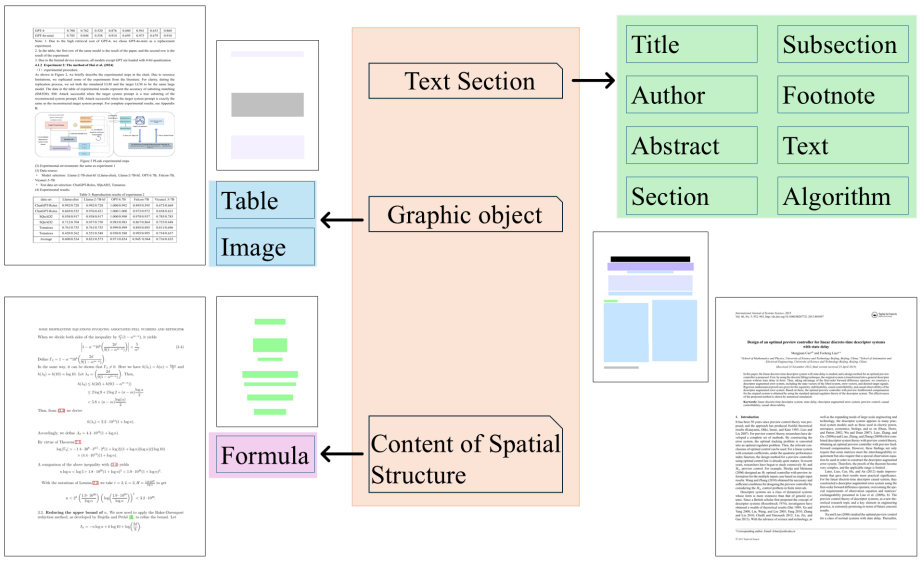

Fig. 2. The element division of document layout that we advocate

According to the semantic function and typesetting characteristics, the document element system is divided into: text element, graphic object, spatial struc-

ture content element (the typical representative is the mathematical expression that needs to be arranged independently).

Layout Analysis Dataset Description The documents processed come from a variety of channels, including arXiv Sci-Hub Textbooks, test papers, etc. The sources and composition of different subsets are shown in the table 1 . It is worth noting that most of the current mainstream layout analysis data sets are built based on English documents, but considering the systematic differences between Chinese and Latin languages in character structure, typesetting rules, etc., this study focuses on the Chinese document scene, and is committed to filling the long-standing shortage of Chinese data sets in this field.

Fig. 3. Example of layout analysis data, where different colors represent different categories of layout elements

Table 1. Statistics of Categories and Corresponding Labels

Category	Label	Total Pages
English Science and Technology Textbook	ET	1268
English Academic Paper	EA	762
Chinese Science and Technology Textbook	CT	5311
Chinese General Document	CN	4648
Chinese Academic Paper	CA	221
Chinese-English Exam Paper	TP	2271

Table 2. Statistics of Subset Classification Categories

Subset	Title	Author	Abstract	Section	Subsection	Table	Image	Formula	Text	Footnote	Algorithm
ET	18	16	1	229	90	106	1070	2898	14974	27	0
EA	43	44	42	518	355	167	322	661	6566	17	26
CT	50	8	1	2936	181	155	2672	21406	60421	24	0
CN	96	2	0	7325	1	545	2267	112	58135	0	0
CA	20	17	17	335	2	79	114	231	2146	19	8
TP	58	10	0	1767	1	267	957	231	22193	1	0
Total	285	97	61	13110	630	1319	7402	25539	164435	88	34

A total of 236,790 paragraph boxes were annotated on 14,481 pages, and the spatial coordinates and semantic categories of the text blocks were accurately captured using bounding boxes. The data was saved in YOLO format; At the same time, explicitly record the logical reading order of the layout, and the annotation order in YOLO format is the reading order.

2.2 Text Detection Dataset Section

We propose a holistic recognition paradigm: a breakthrough approach that treats embedded formulas and their surrounding regular text as a unified semantic unit, and integrates the three independent stages of segmentation, recognition, and association in traditional pipeline systems into an end-to-end process.At the level of data construction, the holistic recognition paradigm proposed in this study needs to reconstruct the data annotation system to support new task requirements. **On the basis of our proposed layout analysis dataset**, we further extended fine-grained annotation for 8 types of text blocks, adding 340,890 row level annotation units and establishing bidirectional hierarchical associations between paragraphs and rows. This hierarchical annotation architecture not only fully preserves the global spatial distribution characteristics of layout elements, but also provides an incremental parsing path for OCR engines from macro layout partitioning to micro text line granularity.

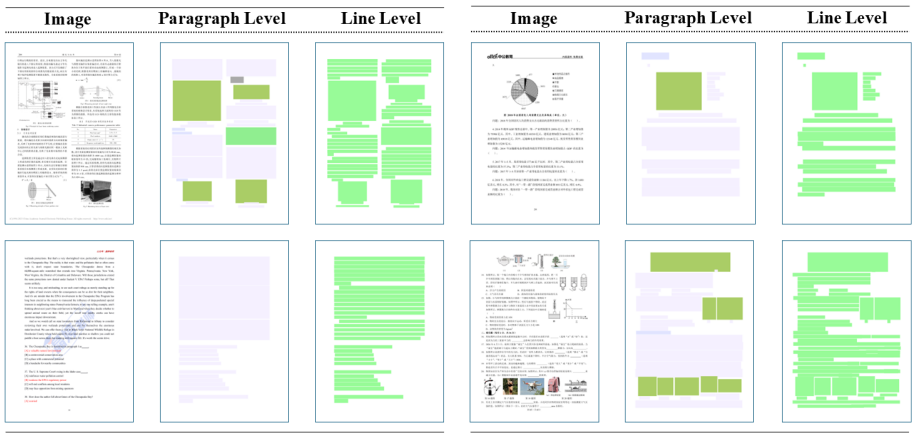

Fig. 4. Example of Paragraph-Line two level annotated data

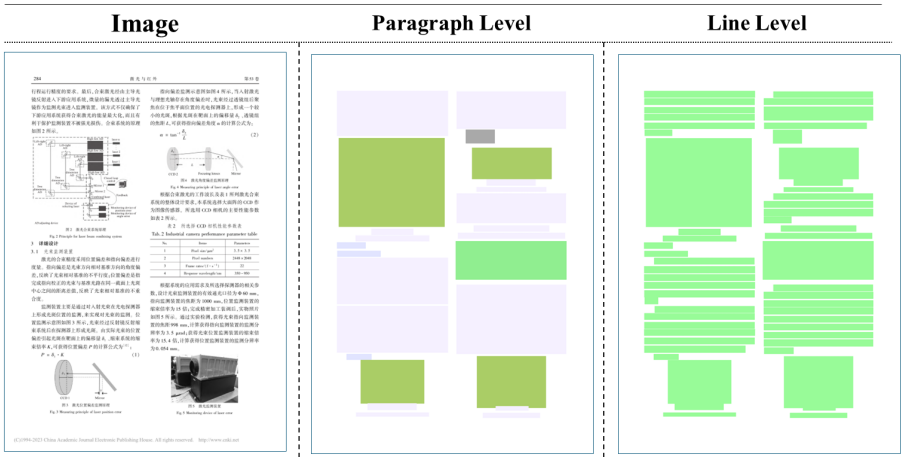

Fig. 5. Enlarge paragraph line two-level annotation data

2.3 Formula-Text Dataset Section

We extracted 399 and 810 pages of Chinese science and engineering academic textbooks and Chinese English test papers based on the text line detection dataset, and completed the unified recognition and labeling of text formulas for 26,849 line paragraphs.

2.4 Table Dataset Section

For the tables, we extracted 1,196 tables from academic papers and annotated them finely using LaTeX. As a mature typesetting system.

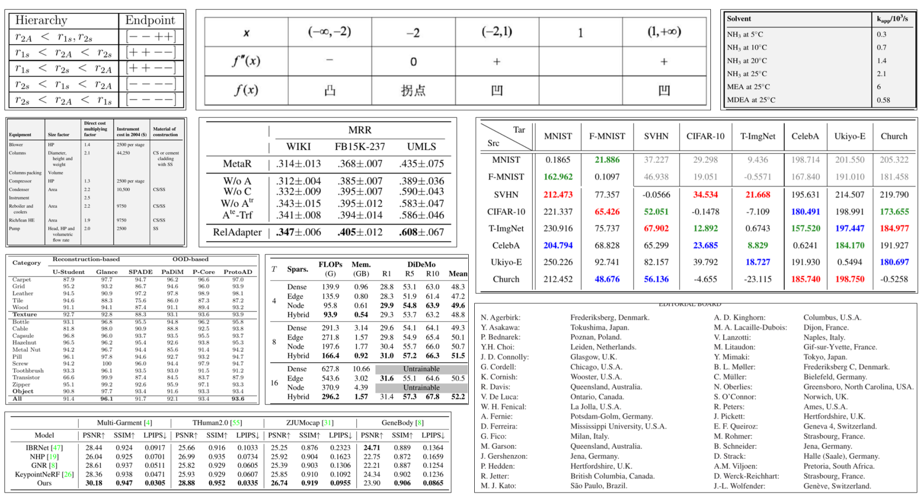

Fig. 6. Example of table recognition data

2.5 Mathematical Formula Dataset (Handwritten/Printed) Section

We collected 8,000 handwritten mathematical formulas and 18,717 printed mathematical formulas with sequence lengths ranging from 64 to 512, mainly from courses such as mathematical analysis and advanced algebra for college students. For the handwritten dataset, we also provided printed images paired with handwriting to support subsequent research on cross modal alignment between handwriting and printing.

As shown in the table, compared to traditional datasets such as CROHME and HME100K, its advantages mainly lie in three dimensions: symbol coverage breadth, complex structure modeling, and long sequence processing capability, making it a challenging dataset.

Table 3. Comparison of Token Types and Sequence Length Across Datasets

Dataset	Number of Samples	Token Space	Average	Median	Maximum
CROHME	9,821	115	14.579	12	203
HME100K	99,109	250	16.664	14	183
MLHME38k	38,000	207	24.203	22	183
IM2LATEX-100K	94,002	516	65.023	55	151
Print-type (Ours)	**18,717**	**308**	**118.824**	**97**	**506**
Hand-type (Ours)	**8,000**	**308**	**62.712**	**48**	**488**

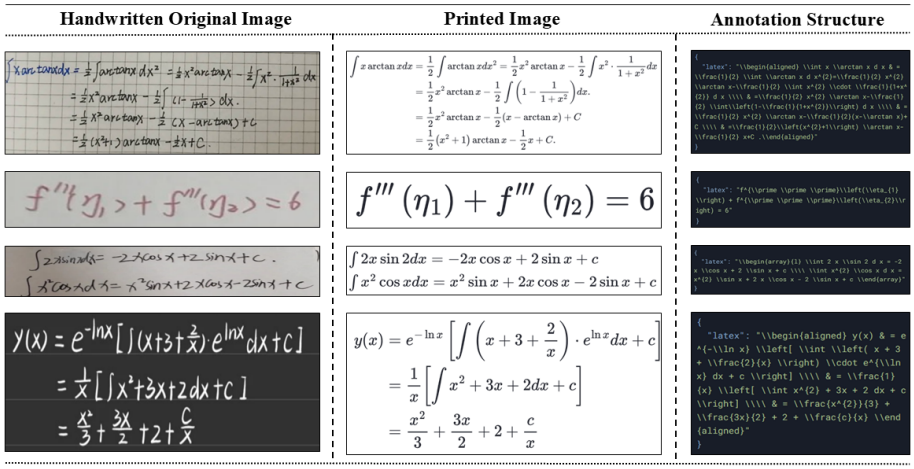

Fig. 7. Handwritten and paired printed mathematical formula annotation

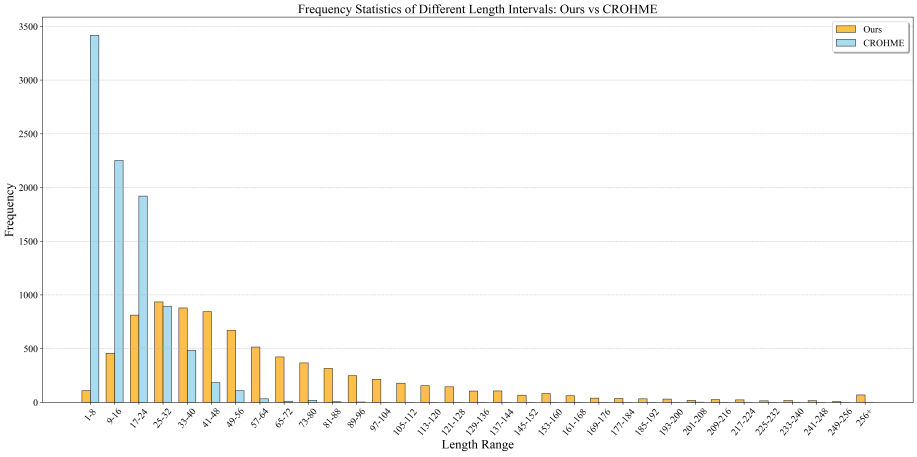

Fig. 8. Comparison of sequence length distribution between the Handwritten Dataset we constructed and the CROHME Dataset

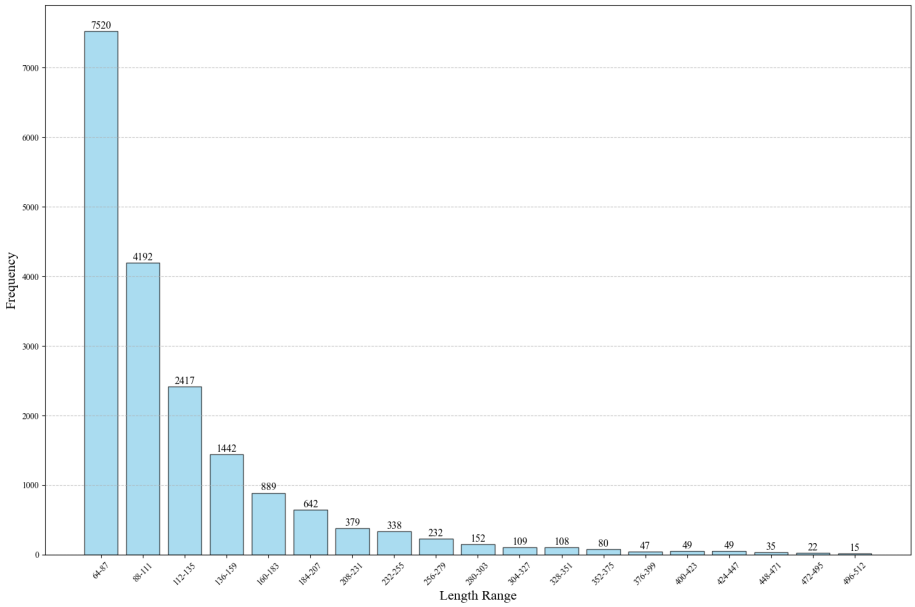

Fig. 9. The length distribution of the 18,717 Printed Mathematical Formulas we constructed

3 Expectation

In the long-term perspective of technological development, when computing power resources no longer constitute a technological bottleneck, the document parsing pipeline technology paradigm advocated in this article will gradually evolve into a solution dominated by multimodal large models. This technological leap marks an important milestone in the development of artificial intelligence. We foresee and look forward to the arrival of this technological turning point, when the universal cognitive ability of basic models will break through the limitations of current engineering architectures and achieve the beautiful vision of technology benefiting all mankind.