DocFusion: A Unified Framework for Document Parsing Tasks

Anonymous ACL submission

Abstract

Document parsing is essential for analyzing complex document structures and extracting fine-grained information, supporting numerous downstream applications. However, existing methods often require integrating multiple independent models to handle various parsing tasks, leading to high complexity and maintenance overhead. To address this, we propose DocFusion, a lightweight generative model with only 0.28B parameters. It unifies task representations and achieves collaborative training through an improved objective function. Experiments reveal and leverage the mutually beneficial interaction among recognition tasks, and integrating recognition data significantly enhances detection performance. The final results demonstrate that DocFusion achieves state-of-the-art (SOTA) performance across four key tasks.

1 Introduction

Document parsing plays a crucial role in extracting structured data from complex documents, serving as a foundational technology for downstream applications. It is particularly important in Retrieval-Augmented Generation (RAG) workflows (Ren et al., 2023; Zhang et al., 2022), where extracting organized and contextually rich information from documents can significantly enhance the performance of large language models (LLMs) (Jiang et al., 2023; Zhao et al., 2024a; Gao et al., 2024). However, information in real-world documents is often embedded in complex structures, such as hierarchical layouts, mathematical expressions, and tables, which makes automatic parsing substantially challenging.

To address these issues, research on document parsing has primarily focused on four key tasks: document layout analysis (**DLA**), mathematical expression recognition (**MER**), table recognition (**TR**), and optical character recognition (**OCR**).

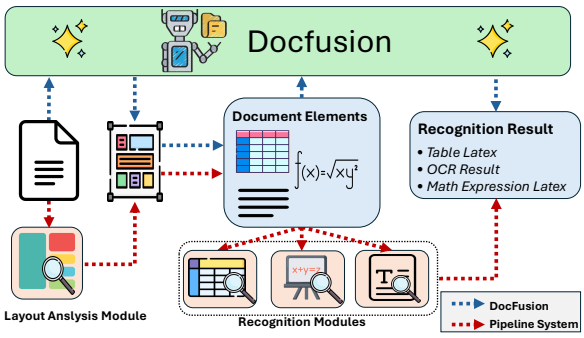

Figure 1: Pipeline systems integrate multiple modules into a Framework. In contrast, DocFusion incorporates multiple functionalities within a single model.

Existing methods can be categorized into two main approaches: multi-module pipeline systems and end-to-end page-level OCR models. As shown in Figure 1, multi-module pipeline systems decompose document parsing tasks into independent modules, allowing each module to adopt the most suitable model. For example, DocLayout-YOLO (Zhao et al., 2024c) has demonstrated excellent performance in DLA, while UniMERNet (Wang et al., 2024a) achieves SOTA results in MER. Although this approach improves performance for specific tasks, integrating multiple models into a single system increases overall complexity. Moreover, these systems fail to fully exploit task-level collaboration, leading to inefficiencies in parameter usage. In contrast, end-to-end page-level OCR models, such as Nougat (Blecher et al., 2023) and GOT (Wei et al., 2024), can seamlessly integrate multiple recognition tasks. While the outputs of these models demonstrate a well-organized logical structure, the models lack the ability of DLA to generate bounding boxes (bboxes) for layout elements. As a result, they cannot preserve the spatial relationships between documents and their corresponding layouts. This limitation is critical in RAG workflows, where preserving spatial relationships is essential for

achieving interpretability. The absence of DLA also affects single-task applications such as MER and TR, which depend on accurate layout analysis for reliable results. These limitations highlight the urgent need for an approach to reduce system complexity and integrate multiple tasks. Therefore, we aim to develop a model capable of simultaneously handling DLA, MER, OCR, and TR tasks.

In this paper, we propose DocFusion, a unified generative multi-task model designed to address four key document parsing tasks. DocFusion leverages multi-task collaboration to achieve comprehensive optimization in document parsing. To handle complex layouts, we introduce Dual Attention (Ding et al., 2022), which combines spatial and channel information interactions. This mechanism enhances DocFusion's ability to process complex tasks with greater accuracy. To address the challenge of loss convergence in detection tasks (DLA) within a generative framework, we design a specialized objective function. The challenge arises from the conflict between the continuous nature of coordinate data and the discrete nature of token generation. Our objective function applies a one-dimensional convolution to smooth the discrete generation distribution. This approach significantly accelerates loss convergence and enables efficient joint training.

Experimental results demonstrate that DocFusion achieves leading performance across all four tasks. Additionally, the recognition tasks mutually enhance each other's performance, leading to overall improvements compared to single-task setups. Notably, OCR improves DLA by providing enriched textual context, enabling more precise layout analysis. Further experiments validate the effectiveness of the improved objective function, demonstrating its key role in enabling task collaboration and performance gains.

Our contributions are summarized as follows:

- We propose DocFusion, a unified generative multi-task model that standardizes task formulations and achieves SOTA performance across four key document parsing tasks: DLA, MER, TR, and OCR.
- Experimental results demonstrate that incorporating multi-task data significantly outperforms single-task setups, providing insights into the benefits of multi-task learning.
- We propose an improved objective function to directly address the conflict between the continuous nature of coordinate data and the discrete nature of token generation in detection tasks within the generative framework.
- We constructed a large-scale dataset containing 1.5M LaTeX-annotated math expressions and 100K tables, standardized for consistency, providing a valuable resource for advancing document parsing research.

Tool	Size	Type	DLA	MER	TR	OCR
UniMER (2024a)	325M	M		✓		
DocLayout(2024c)	20M	M	✓			
StructTable (2024)	938M	M			✓	
ViTLP (2024)	253M	M	✓		✓	✓
Nougat (2023)	350M	M		✓	✓	✓
GOT (2024)	580M	M		✓	✓	✓
open-parse (2024)	-	S	✓		✓	✓
LlamaParse (2024)	-	S	✓	✓	✓	✓
DeepDoc (2024)	-	S	✓		✓	✓
MinerU (2024)	-	S	✓	✓	✓	✓
DocFusion	289M	M	✓	✓	✓	✓

Table 1: Capabilities of document parsing tools. **Type**: **M** represents a model, while **S** denotes a system. **DLA**: Document Layout Analysis. **MER**: Math Expression Recognition. **TR**: Table Recognition. **OCR**: Optical Character Recognition.

2 Related Work

Document Parsing Models. Document parsing models have seen remarkable progress across various tasks. DLA has evolved from vision-based methods (Wick and Puppe, 2018; Bao et al., 2021) to multimodal approaches integrating textual features (Xu et al., 2022; Huang et al., 2022). OCR has transitioned from template matching (Smith, 2007) to deep learning-based solutions (Bušta et al., 2017; Chen et al., 2021; Mosbah et al., 2024). MER progressed from symbol segmentation (Miller and Viola, 1998) to CNN-RNN hybrids (Le et al., 2019) and Transformer-based models (Wang et al., 2024a). Similarly, TR now employs methods like grid segmentation and image-to-sequence techniques to reconstruct structured data (Qasim et al., 2019; Huang et al., 2023; Xia et al., 2024). Page-level end-to-end OCR models like Nougat (Blecher et al., 2023) and GOT (Wei et al., 2024) simplify workflows by integrating multi recognition tasks.

Modular Pipeline Systems. The advancements in task-specific models have driven the development of modular pipeline systems, which process

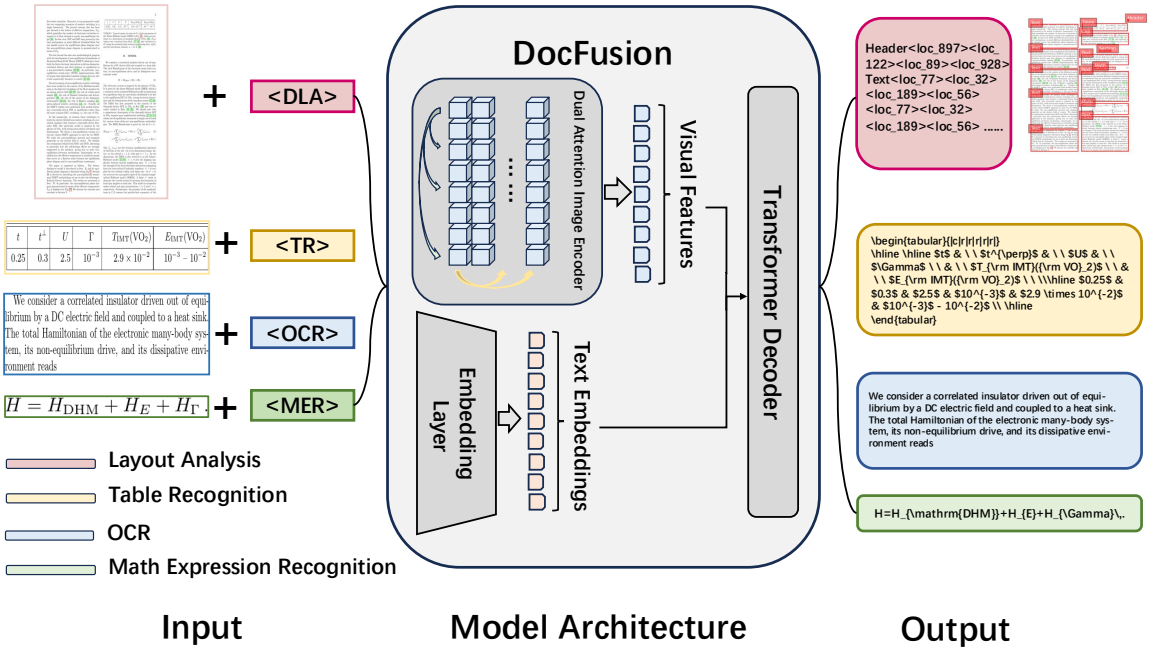

Figure 2: The model comprises three key components: a visual encoder, a text embedding layer and a Transformer decoder. The image features extracted by the visual encoder and the instruction embeddings are combined and then passed to the Transformer decoder, which produces the final output sequence.

complex document structures through specialized modules. For instance, Open-Parse(Filimonov, 2024) performs well in incrementally parsing complex layouts but lacks support for MER. Other systems, such as DeepDoc(Yu, 2024) and Llama-Parse(Liu, 2024), extend the scope of modular pipelines to handle more diverse tasks. In particular, MinerU(Zhao et al., 2024b) stands out by supporting advanced features such as complex layout parsing and Markdown conversion. However, despite their flexibility, modular systems face significant challenges in practical deployment. The variability in environmental dependencies between modules increases the complexity of maintenance. Furthermore, tasks that could be efficiently handled by a single module are often divided among multiple modules, leading to unnecessary system overhead. These limitations highlight the need for more unified and efficient frameworks to address the growing demands of document parsing.

3 DocFusion

We introduce the model architecture (3.1) and explain how detection tasks are integrated into the generative framework. Then, we discuss the challenges (3.2) of detection tasks within this framework. Next, we explain the improved objective function (3.3)

3.1 Architecture

As shown in Figure 2, the architecture of DocFusion consists of three main components: a vision encoder, a text embedding layer, and a Transformer decoder. Since the task instructions are limited and predefined, no Transformer encoder is included; instead, task-specific prompts are directly embedded, simplifying the architecture. To unify the representation of object detection and text recognition tasks, we adopt a coordinate quantization representation (Xiao et al., 2023). Specifically, images are quantized into a fixed resolution (e.g., 1000×1000), and coordinates are represented as discrete tokens (e.g., <loc_1>, <loc_2>, ..., <loc_1000>). This approach enables the use of a unified regression framework for detection tasks, simplifying multi-task integration. To address the challenges posed by densely structured content, the vision encoder incorporates a Dual Attention mechanism (Ding et al., 2022), which captures interactions across channel and spatial dimensions, enhancing feature extraction for intricate document layouts. Additionally, the traditional feed-forward network (FFN) is removed, reducing both parameter count and computational cost, further improving model efficiency. The vision encoder processes input images $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ into visual features, flattened as token

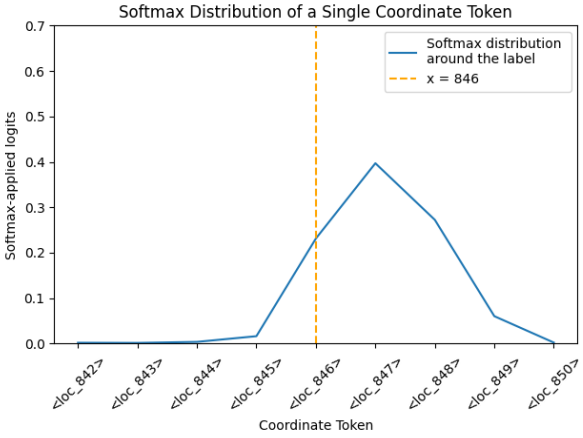

Figure 3: The Softmax distribution of logits for a target token and its neighboring tokens after the loss has stabilized.

embeddings $\mathbf{V} \in \mathbb{R}^{N_v \times D_v}$. These embeddings are transformed for compatibility with task-specific prompt embeddings $\mathbf{T}\text{prompt} \in \mathbb{R}^{N_t \times D}$. The combined input $\mathbf{X} = [\mathbf{V}'; \mathbf{T}\text{prompt}]$ is then passed to the Transformer decoder to generate predictions. By integrating Dual Attention, coordinate quantization, and optimizing its architecture, DocFusion efficiently handles complex document parsing tasks with high precision and computational efficiency.

3.2 Challenges and Motivations

While representing object detection as text regression enables joint training of layout analysis and page element recognition under a unified cross-entropy-based framework, it inherently forces continuous coordinates into discrete token spaces. This mismatch creates several challenges, especially in fine-tuning small coordinate adjustments, where the model struggles to produce accurate gradients, reducing training stability. As shown in Figure 3, small unavoidable deviations in coordinate labels smooth out the softmax distribution, preventing the target token's probability from forming a sharp peak. This makes it harder for the model to escape local optima and limits its learning capacity. Additionally, traditional cross-entropy loss, which is designed for discrete classification tasks, does not handle continuous changes well, further increasing inaccuracies during training.

In multi-task settings, these issues become even more challenging. The conflict between discrete loss functions and continuous coordinate optimization can skew gradients, causing one task to dominate at the cost of others. This imbalance reduces performance in other tasks and harms the model's ability to predict coordinates accurately, limiting its overall effectiveness in complex document parsing tasks. Solving these problems is critical to improving both localization accuracy and training stability across tasks.

3.3 Objective function

To address these challenges, we propose an improved objective function that applies a one-dimensional convolution over the probability distribution, refining the model's sensitivity to small coordinate changes while preserving the discrete treatment of cross-entropy. This approach helps alleviate the mismatch between discrete tokens and continuous coordinates, improves gradient quality, and prevents the coordinate prediction task from dominating the optimization process. In doing so, it enhances localization accuracy, supports stable multi-task training, and achieves better alignment with the desired properties identified in the motivating considerations.

Let the model's output logits be denoted as $\mathbf{Z} \in \mathbb{R}^{B \times L \times V}$, where B is the batch size, L is the sequence length, and V is the vocabulary size. The target labels are denoted as $\mathbf{T} \in \mathbb{N}^{B \times L}$. The range of indices corresponding to coordinate tokens is defined as $[s, e]$, representing their positions in the vocabulary.

The standard softmax probability distribution is first computed as:

$$\mathbf{P} = \text{softmax}(\mathbf{Z}) \tag{1}$$

A mask is then applied to zero out probabilities outside the range $[s, e]$, creating a modified probability tensor $\mathbf{P}'$:

$$\mathbf{P}'_{ijk} = \begin{cases} \mathbf{P}_{ijk}, & \text{if } k \in [s, e] \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Next, a one-dimensional convolution kernel $\mathbf{K} \in \mathbb{R}^{1 \times 1 \times k}$ is constructed based on a Gaussian distribution, where k is the kernel size (an odd integer greater than 1), and σ is the standard deviation. The kernel weights are computed as:

$$\mathbf{K}_i = \exp\left(-\frac{(i - \frac{k-1}{2})^2}{2\sigma^2}\right) \tag{3}$$

The kernel is then applied to $\mathbf{P}'$ via one-dimensional convolution along the vocabulary dimension:

$$\mathbf{C} = \text{conv1d}(\mathbf{P}', \mathbf{K}, \text{padding} = \frac{k-1}{2}) \tag{4}$$

This convolution preserves the size of the input and output tensors.The convolution result $\mathbf{C}$ is integrated back into the original probability distribution $\mathbf{P}$ within the index range $[s, e]$, while retaining the original probabilities outside this range:

$$\mathbf{P}''_{ijk} = \begin{cases} \mathbf{C}_{ijk}, & \text{if } k \in [s, e] \\ \mathbf{P}_{ijk}, & \text{otherwise} \end{cases} \tag{5}$$

The final objective function is computed as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{B} \sum_{j=1}^{L} \mathbf{M}_{ij} \log \mathbf{P}''_{ij\mathbf{T}_{ij}} \tag{6}$$

4 Dataset Construction and Refinement

In this section, we briefly describe the reconstruction of the DLA dataset and the collection of the MER, TR, and OCR datasets, with more detailed information provided in the appendix.

DLA Dataset. DocLaynet (Pfitzmann et al., 2022) was chosen for its comprehensive layout annotations, but its formula annotations, where content and numbering share the same bounding box, introduce noise for MER tasks. To address this, we re-extracted formulas from arXiv LaTeX files, trained a lightweight model to re-annotate the pages with manual verification.

MER Dataset. The UniMER-1M (Wang et al., 2024a) has significantly advanced MER research but contains many redundant spaces in LaTeX code. Although some spaces are syntactically necessary, most are unnecessary, increasing output length and computational overhead. To address this, we constructed a new dataset by extracting content from LaTeX files, normalizing style variations and verifying accuracy through re-rendering. Models trained on our dataset produce LaTeX code that is approximately 34.2% shorter for complex expressions and 37.5% shorter for simple expressions on the UniMER-1M test set, demonstrating improved efficiency and performance.

TR Dataset. In the TR task of DocFusion, we adopted LaTeX as the output format for two main reasons: (1) to ensure consistency with the MER task's output format, enabling better multi-task collaboration; and (2) because LaTeX facilitates both the extraction of cell content and the restoration of the original table layout. Existing LaTeX-based TR datasets either lack sufficient scale or fail to separate tables from captions, conflicting with our DLA task. To overcome these limitations, we constructed a high-quality TR dataset with 100K samples by following a similar approach to the MER dataset.

OCR Dataset. The dataset also sourced from DocLaynet (Pfitzmann et al., 2022), provides detailed layout and character annotations. We extracted cropped images for each layout element and paired them with corresponding character-level text annotations.

5 Experiments

5.1 Implementation Details

We conducted our experiments using the PyTorch framework on eight NVIDIA H100 GPUs, with an initial learning rate of 1e-5, a per-GPU batch size of 12, and employing a cosine learning rate scheduler to progressively adjust the model parameters.

5.2 Evaluation Metrics

5.2.1 Evaluation for Recognition

We employ traditional metrics such as BLEU (Papineni et al., 2001) and Edit Distance (Levenshtein, 1966) to evaluate generated sequences. Additionally, we introduce task-specific metrics like CDM (Wang et al., 2024b) and CSR to better assess the quality and usability of LaTeX-based outputs.

BLEU: The BLEU score is used for evaluating machine-generated text, measuring n-gram overlap with reference texts while incorporating a brevity penalty to ensure balanced outputs.

Edit distance: Also known as Levenshtein distance, measures the minimum number of operations insertions, deletions, or substitutions required to transform one string into another.

CSR: This score refers to the percentage of generated LaTeX outputs that can be successfully compiled into PDF. It reflects the correctness of the model's predictions and practical usability.

ExpRate: The ExpRate (Li et al., 2022) measures the proportion of samples where the predicted text matches the reference text without any errors.

CDM: The CDM evaluates MER by comparing image-rendered expression at the character level with spatial localization, ensuring fairness and accuracy over text-based metrics like BLEU.

5.2.2 Evaluation for Detection

Since the DLA task in DocFusion does not use confidence scores, we did not use the widely

Model	size	OCR		MER			TR	
		BLEU↑	EditDis↓	CDM↑	ExpRate↑	CSR↑	F1↑	CSR↑
Pix2tex (2022)	-	-	-	76.5	41.7	95.9	-	-
Texify (2023)	312M	-	-	88.6	71.7	97.8	-	-
UniMERNet (2024a)	325M	-	-	**99.0**	89.5	99.7	-	-
Qwen-VL-PLUS (2023)	-	85.3	0.120	-	-	-	-	-
Qwen-VL-OCR (2023)	-	94.9	0.055	-	-	-	-	-
StructEqTable (2024)	938M	-	-	-	-	-	90.6	89.3
GOT (2024)	580M	86.7	0.115	87.7	67.3	97.8	86.9	81.6
DocFusion	289M	**99.1**	**0.007**	98.7	**94.2**	**99.8**	**92.1**	**92.5**

Table 2: Comparison of DocFusion with other models on three recognition tasks. *Note:* Due to differences in training styles across models, line break were consistently removed when calculating BLEU and Edit Distance.

Model	Size	DocLayNet			DocLayNet-Scientific			FPS↑
		Precision↑	Recall↑	F1↑	Precision↑	Recall↑	F1↑	
DETR (2020)	41M	87.1	**91.6**	**89.3**	95.9	**96.2**	96.0	3.7
DocLayout-YOLO (2024c)	20M	86.7	91.1	88.9	94.4	95.5	95.0	**85.2**
DocFusion	289M	**88.0**	88.4	88.2	**96.8**	**96.2**	**96.4**	7.5

Table 3: The performance of the models on DLA, where DocLayNet-Scientific refers to the scientific document subset of the Doclaynet. *Note:* DETR and DocLayout-YOLO are limited to object detection tasks only.

adopted Average Precision (AP) metric from the object detection field. Instead, we focus on the following metrics:

Precision: Precision measures the proportion of correctly identified positive instances among all predicted positives.

Recall: Recall measures the proportion of correctly identified positive instances among all actual positives.

F1-score: The F1-score balances precision and recall, serving as their harmonic mean. This metric is particularly useful for evaluating the trade-off between precision and recall in the DLA task.

FPS: FPS measures the number of frames processed by the model per second, providing an indication of the model's inference speed and efficiency.

5.3 Main Results

We use UnimerNet (Wang et al., 2024a) for MER, StructEqTable (Xia et al., 2024) for TR, DocLayout-YOLO (Zhao et al., 2024c) for DLA, and Qwen-VL-OCR (Bai et al., 2023) for OCR as baselines, as well as other widely used models for comparison. These baselines were selected for their strong performance and task relevance. The results show that DocFusion demonstrates competitive performance against other SOTA methods.

5.3.1 MER performance

We evaluated our model using the test subset of the UniMER-1M (Wang et al., 2024a), with a focus on the Simple Printed Expression (SPE) and Complex Printed Expression (CPE) subsets, as DocFusion is specifically designed for processing printed documents. As shown in Table 2, DocFusion performs exceptionally well across multiple evaluation metrics, particularly in CSR and ExpRate. Notably, its ExpRate exceeds that of the second-ranked UniMERNet by 5.2%, demonstrating its superior reliability in real-world document parsing. The results presented here combine the performance of both SPE and CPE, with detailed separate results provided in the appendix.

5.3.2 TR performance

We constructed a benchmark dataset consisting of 2,500 table images extracted from LaTeX documents, including both simple and relatively complex tables, all of which were manually verified. To accommodate the model parameters and maximum sequence length, the LaTeX ground truth for the test set was limited to a maximum of 1,024 tokens. Using LatexNodes2Text, we extracted the content of each table cell to computed F1 scores (The detailed extraction method is presented in the appendix). As shown in Table 2,

Train Dataset	OCR		MER		TR		DLA
	BLEU↑	EditDis↓	CDM↑	CSR_{MER}↑	F1↑	CSR_{TR}↑	F1↑
Task-Specific	98.8	0.010	98.5	99.8	91.2	92.7	87.8
OCR+DLA	98.5	0.010	-	-	-	-	**88.9**
OCR+MER+TR	**99.1**	**0.008**	**98.9**	**99.9**	**92.3**	**94.6**	-

Table 4: Ablation experiments on task collaboration, comparison of task performance when using **Task-specific** training, where each task is trained independently, and other joint multi-task strategies.

DocFusion performs excellently on this benchmark, with both F1 and CSR scores exceeding those of the second-ranked model by 1.6%, demonstrating its superior ability to handle both simple and complex table structures.

5.3.3 OCR performance

We separated 3,000 English image samples from the originally constructed dataset as the test set. As shown in Table 2, DocFusion demonstrates outstanding performance in both BLEU and EditDis, achieving more precise recognition of layout elements. This performance improvement is primarily attributed to DocFusion's joint training on layout analysis and text recognition tasks, which enhances the model's efficiency and effectiveness in handling complex document structures. These results further validate the effectiveness of the proposed training strategy, especially for document parsing tasks involving both text content and layout element recognition.

5.3.4 DLA performance

DocFusion generates layout element labels and coordinates by sequentially predicting tokens without relying on confidence scores. Since the commonly used Average Precision (AP) metric in object detection depends on confidence scores, it cannot be directly applied in this evaluation. To ensure a fair comparison with confidence-based models, we adopt an alternative evaluation method. For these models, we compute Precision, Recall, and F1-score at different thresholds and select the maximum F1-score across all thresholds as the final evaluation metric. As shown in Table 3, DocFusion may not achieve outstanding performance on the entire DocLaynet test set but performs well in the domain of scientific document detection. This could be attributed to its ability to generate bounding boxes with clean edges. In terms of processing speed, although DocFusion has more parameters than DETR, another Transformer-based model, it achieves faster processing due to the use of Flash-Attention. Compared to YOLO, DocFusion is slightly slower but does not require threshold tuning to achieve optimal performance, offering high performance without additional adjustments.

5.4 Ablation Study

5.4.1 OCR-Driven Enhancement of DLA

This section explores the impact of OCR on DLA performance. As shown in Table 4, the results in the DLA column from the first and second rows indicate that adding the OCR task improves DLA performance, with an F1 increase of up to 1.3%. This result demonstrates the effectiveness of using textual information in joint training. Compared to independent training that relies only on visual features, OCR significantly enhances the model's robustness and generalization. For example, tables and mathematical expressions are layout elements with clear visual features, which the model can often recognize effectively. In contrast, text or titles have less distinctive visual features, making it challenging to predict their labels based on visual information alone. Textual cues play a crucial role in these cases. These findings confirm that OCR is essential for improving DLA performance. By providing complementary textual information, OCR strengthens the collaboration between visual and semantic features, resulting in better overall performance.

5.4.2 Collaboration of Recognition Tasks

In this section, we explore the collaboration among the recognition tasks OCR, TR, and MER. As shown in Table 4, the experimental results from the first and third rows demonstrate that joint training yields better performance compared to training each task individually. Specifically, OCR achieves a 0.3% improvement in BLEU score, MER sees

Objective Function	OCR		MER		TR		DLA
	BLEU↑	EditDis↓	CDM↑	CSR_{MER}↑	F1↑	CSR_{TR}↑	F1↑
CE	97.3	0.009	97.8	96.5	90.2	89.1	87.9
CE*	**99.1**	**0.007**	**98.7**	**99.8**	**92.1**	**92.5**	**88.2**

Table 5: Ablation analysis of the improved objective function was conducted on the same dataset across four tasks: OCR, MER, TR, and DLA. **CE** represents training with the standard cross-entropy loss, while **CE*** denotes training with the improved objective function.

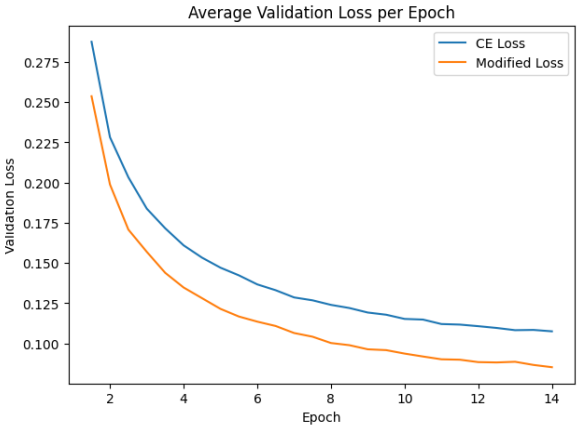

Figure 4: Validation loss curves under identical hyperparameter settings, where the only variation is the choice of the objective function.

increases of 0.4% in CDM and 0.1% in CSR, and TR benefits most significantly, with a 2.1% improvement in F1 score for cell parsing and a 2.0% increase in CSR. This collaboration enables the model to leverage shared information across tasks, enhancing individual task performance and improving overall document parsing capabilities. These results demonstrate that multi-task collaboration effectively enhances performance by leveraging shared information.

5.4.3 Results of improved objective function

In this section, we compared the original cross-entropy and the improved objective function in recognition and detection tasks. As shown in Table 5, the results demonstrate that the improved objective function led to significant performance gains across both task categories. In recognition tasks, the BLEU score in the OCR task saw an improvement of 1.8%. Additionally, the CDM metric in the MER task increased by 0.9%, while the F1 score in the TR task rose by 2.1%. Notably, for the CSR metric, which assesses LaTeX compilation success, the MER and TR tasks achieved gains of 3.3% and 3.8%, respectively, highlighting enhanced usability and correctness of the LaTeX outputs. For the detection task, the F1 score of the DLA task increased by 0.34%. This improvement can be attributed to the improved objective function, which alleviates the issue of coordinate token errors dominating the gradient. By addressing this imbalance, the objective function not only enhances the performance of recognition tasks but also improves the accuracy of predicting layout element categories in the detection task itself. These results collectively show that the improved objective function effectively addresses key challenges in loss minimization, ensuring that tasks such as DLA can operate within a generative framework. It avoids gradient dominance issues while achieving better task balance in a multi-task learning setup, demonstrating its robustness and versatility.

6 Conclusion

In this work, we introduced DocFusion, the first approach to integrate the four modules of a document parsing pipeline into a unified model by designing a objective function tailored to handle diverse data types across tasks. Our method achieved SOTA performance on multiple benchmarks. To enable downstream applications, we re-annotated the widely used DocLayNet dataset and constructed a large-scale formula-to-LaTeX dataset, applying a unified standardization process. Through detailed analysis, we observed that DocFusion, as a lightweight model with fewer parameters, effectively integrates multiple tasks into a single framework, demonstrating both efficiency and versatility in handling complex document parsing challenges. In the future, we aim to extend DocFusion to larger models and further improve dataset standardization to enhance its performance and applicability across broader tasks and domains.

Limitations

In this section, we discuss the limitations of the proposed model, DocFusion. While the model has demonstrated strong performance across multiple document layout analysis subtasks on specific datasets, its design is constrained by a parameter size of 289M and a maximum output length of 1024 tokens. These constraints may impact its ability to handle highly complex layouts or extremely long sequences, requiring further optimization for specific use cases. Additionally, DocFusion's reliance on PDF screenshots in the LaTeX recognition task limits its generalization to handwritten or other non-standard formats. For the detection task, although the model achieves competitive accuracy, its processing speed poses challenges for real-time or high-throughput applications, indicating a need for further improvements in computational efficiency to meet broader application demands.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *Cornell University - arXiv,Cornell University - arXiv*.

Lukas Blecher. 2022. pix2tex - latex ocr. Accessed: 2024-02-29, cited in pages 1, 2, 3, 7, 10, 11.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *Preprint*, arXiv:2308.13418.

Michal Bušta, Lukàš Neumann, and Jirí Matas. 2017. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2231.

Jingye Chen, Bin Li, and Xiangyang Xue. 2021. Scene text telescope: Text-focused scene image super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. 2022. Davit: Dual attention vision transformers. *Preprint*, arXiv:2204.03645.

Sergey Filimonov. 2024. Openparse.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Yongshuai Huang, Ning Lu, Dapeng Chen, Yibo Li, Zecheng Xie, Shenggao Zhu, Liangcai Gao, and Wei Peng. 2023. Improving table structure recognition with visual-alignment sequential coordinate modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11134–11143.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking (2022). *arXiv preprint arXiv:2204.08387*.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. *Preprint*, arXiv:2305.09645.

Anh Duc Le, Bipin Indurkhya, and Masaki Nakagawa. 2019. Pattern generation strategies for improving recognition of handwritten mathematical expressions. *Pattern Recognition Letters*, 128:255–262.

V.I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Proceedings of the USSR Academy of Sciences,Proceedings of the USSR Academy of Sciences*.

Bohan Li, Ye Yuan, Dingkang Liang, Xiao Liu, Zhilong Ji, Jinfeng Bai, Wenyu Liu, and Xiang Bai. 2022. When counting meets hmer: Counting-aware network for handwritten mathematical expression recognition.

Jerry Liu. 2024. Llamaparse.

Zhiming Mao, Haoli Bai, Lu Hou, Jiansheng Wei, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Visually guided generative text-layout pre-training for document intelligence. *arXiv preprint arXiv:2403.16516*.

Erik G Miller and Paul A Viola. 1998. Ambiguity and constraint in mathematical expression recognition. In *AAAI/IAAI*, pages 784–791.

Lamia Mosbah, Ikram Moalla, Tarek M. Hamdani, Bilel Neji, Taha Beyrouthy, and Adel M. Alimi. 2024. Adocrnet: A deep learning ocr for arabic documents recognition. *IEEE Access*, 12:55620–55631.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*.

Vik Paruchuri. 2023. Texify. Accessed: 2024-02-29, cited in pages 1, 2, 4, 6, 7.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter W J Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. page 3743–3751.

Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. 2019. Rethinking table recognition using graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 142–147. IEEE.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2023. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. *Preprint*, arXiv:2110.07367.

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

Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. 2024a. Unimernet: A universal network for real-world mathematical expression recognition. *Preprint*, arXiv:2404.15254.

Bin Wang, Fan Wu, Linke Ouyang, Zhuangcheng Gu, Rui Zhang, Renqiu Xia, Bo Zhang, and Conghui He. 2024b. Cdm: A reliable metric for fair and accurate formula recognition evaluation. *Preprint*, arXiv:2409.03643.

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

Christoph Wick and Frank Puppe. 2018. Fully convolutional neural networks for page segmentation of historical document images. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. 2024. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2023. Florence-2: Advancing a unified representation for a variety of vision tasks (2023). *URL https://arxiv. org/abs/2311.06242*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2022. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *Preprint*, arXiv:2012.14740.

Zhichang Yu. 2024. Deepdoc.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2022. Adversarial retriever-ranker for dense text retrieval. *Preprint*, arXiv:2110.03611.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024a. A survey of large language models. *Preprint*, arXiv:2303.18223.

Xiaomeng Zhao, Kaiwen Liu, and Bin Wang. 2024b. Deepdoc.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024c. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *Preprint*, arXiv:2410.12628.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint*.

A Appendix

A.1 DLA Dataset Reconstruction

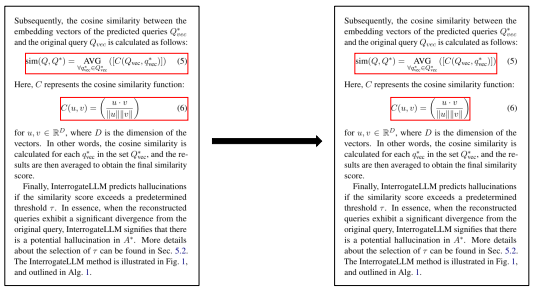

Figure 5: The corresponding numbers were removed from the annotated data for mathematical expression detection.

In DocLaynet and other similar datasets, the annotation of mathematical formulas has certain limitations, as show in figure 5, the content of math expression and numbering are typically annotated within the same bounding box. This annotation approach introduces noise in subsequent Mathematical Expression Recognition (MER) tasks. To address this issue, we extracted formulas from arXiv LaTeX source files using regular expressions and assigned unique colors and bounding boxes to each element. Then, we employed a fuzzy matching algorithm to ensure annotation accuracy and eliminate overlaps. Finally, we trained a

lightweight detection model and, combined with manual verification, re-annotated pages containing formulas. These improvements significantly enhance the dataset's applicability to subsequent MER tasks.

A.2 MER and TR data standardization

Issue	Original	Standardized
Bracket	`\{`	`\lbrace`
Subsup	`a^1_2`	`a_2^1`
Prime	a′	`a^{\prime}`
Fraction	`\over`	`\frac`
Space	`\tabular{l c}`	`\tabular{lc}`

Table 6: Examples of LaTeX standardization for various symbols and expressions.

We chose to standardize the output format as LaTeX for two recognition tasks involving non-plain-text elements. For MER, converting to LaTeX was essential as it provides a precise representation of mathematical formulas. For TR, in addition to ensuring format consistency, converting to LaTeX also allows for the restoration of the original content through compilation, and enables the extraction of cell elements using tools such as LatexNodes2Text, thus enhancing processing flexibility. We used regular expressions to extract relevant content from the LaTeX source files of research papers. However, due to variations in author writing styles, the same formula or table may appear in multiple forms, increasing the complexity of training. As show in table 6 , we analyzed these different representations, standardized them to eliminate ambiguities and ensured consistency. To verify the accuracy of the standardized LaTeX code, we re-rendered it into images, creating a high-quality dataset that aligns with the actual input-output content.