Boosting Document Image Translation via Layout-aware Semantic Paragraph Clustering

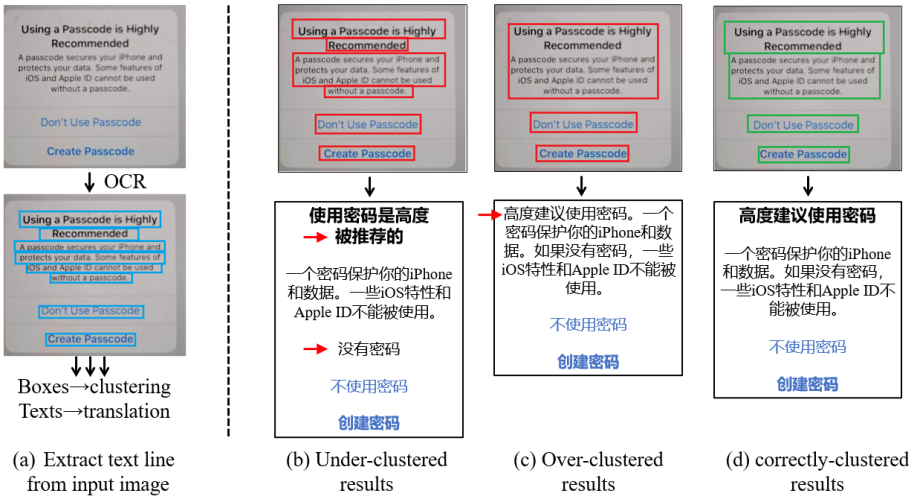
Zhiyuan Chen[1,2], Yaping Zhang[1,2], Zhiyang Zhang[1,2], Yupu Liang[1,2], Yue Xu[3], Yunfei Lu[3], Dandan Tu[3], Chengqing Zong[1,2], and Yu Zhou[1,4(✉)]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{chenzhiyuan2023, zhangzhiyang2020, liangyupu2021}@ia.ac.cn
{yaping.zhang, cqzong, yzhou}@nlpr.ia.ac.cn
[3] Huawei Technologies Co., Ltd., China
{xuyue56, luyunfei6, tudandan}@huawei.com
[4] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China

Abstract. Document Image Translation (DIT) requires accurate preservation of both textual semantics and spatial layout during cross-lingual conversion. Existing approaches predominantly rely on geometric heuristics for post-OCR text clustering, often failing to capture the semantic coherence essential for high-quality translation. To address this, we propose a novel Layout-aware Semantic Paragraph Clustering algorithm designed to reconstruct coherent paragraphs from fragmented OCR results. Our method operates through an iterative framework composed of three synergistic modules: (1) a **Spatial Neighbor Selection** module that identifies spatially proximate OCR fragments based on geometric constraints, (2) a **Semantic Concatenation Model** that evaluates semantic coherence for candidate fragment merging, and (3) a **Completeness Judgment Model** that determines whether aggregated segments constitute semantically independent paragraphs. Through iterative optimization, our framework reconstructs semantically coherent and spatially consistent paragraph structures, significantly enhancing downstream translation quality. Extensive experiments on DIT700K and cross-domain evaluations demonstrate substantial improvements over state-of-the-art methods, with BLEU score improvements of up to 17.37 points on complex layouts. Our framework serves as an effective post-processing component that enhances both traditional cascaded systems and modern multimodal language models.

1 Introduction

Document Image Translation (DIT) seeks to translate text embedded within document images while faithfully preserving the original layout, a task crucial for global information access and cross-lingual communication. Many state-of-the-art DIT systems adopt a cascaded pipeline ([12, 17]), where Optical Character



(a) Extract text line from input image (b) Under-clustered results (c) Over-clustered results (d) correctly-clustered results

Fig. 1. An example in en-zh translation. Red arrows indicate errors in translation and paragraph structure. **(a)** The original image input processed by OCR. **(b)** Under-clustered results: paragraphs split into multiple segments, disrupting the flow and meaning of the translation. **(c)** Over-clustered results: poorly structured translation. **(d)** Correct results: both paragraph positions and translations are correct.

Recognition (OCR) first extracts text lines. Subsequently, these lines must be accurately grouped into semantically coherent paragraphs before being processed by a machine translation engine. The quality of this paragraph clustering step is critical. From a functional perspective, a high-quality clustering result should form independent complete paragraphs, which means paragraphs that a text translation model can accurately translate without relying on additional context. as illustrated in Figure 1, suboptimal clustering, either over-grouping disparate content or excessively splitting cohesive text, can severely degrade the readability and accuracy of the final translation. However, existing paragraph clustering methods predominantly rely on physical layout cues (e.g., geometric rules [2, 19], or neural network based approaches [23, 29]). Such approaches are often fragile, struggling with complex layouts, OCR noise, or image distortions [21], and crucially, they tend to overlook the rich semantic information and logical reading order that humans use to distinguish paragraphs.

To address these limitations, we propose a novel paragraph aggregation strategy that shifts the focus from purely physical layout to leveraging textual semantics and logical reading order, mimicking human comprehension. Our core contribution is a dual-model collaborative framework designed for robust OCR post-processing in cascaded DIT. This framework iteratively reconstructs paragraph structures by:

1. Employing a Completeness Judgment Model (CJM) to assess if a text segment forms a semantically complete paragraph.

2. Using a Spatial Neighbor Selection (SNS) module to identify spatially proximate candidate segments for merging with incomplete ones.
3. Utilizing an Semantic Concatenation Model (SCM) to determine if an incomplete segment and a selected neighbor can be coherently concatenated.

This iterative process aims to generate paragraphs that are not only structurally sound but also semantically meaningful, thereby significantly enhancing the quality of the final translation.

In summary, our main contributions are:

- A novel, human-inspired paragraph aggregation strategy prioritizing logical layout (semantic coherence, reading order) over physical cues.
- A dual-model collaborative framework for iterative paragraph hierarchy reconstruction, designed as an effective OCR post-processing step for cascaded DIT.
- Extensive experiments demonstrating superior paragraph clustering effects and significantly improved DIT performance over state-of-the-art methods.

2 Related Work

2.1 Document Image Translation

Document Image Translation (DIT) aims to translate document images by leveraging both their visual and textual content. [27] Early work [1] focused only on translating OCR-extracted text, but often lost crucial spatial context. Subsequent research integrated visual layout information [10, 27, 28], either through external parsers or intrinsic layout-oriented encoders, to better handle complex-layout documents beyond simple sentence or paragraph-level inputs.

Despite these advancements in incorporating layout awareness, a significant limitation persists: many methods still operate on sequences of potentially noisy or fragmented OCR results, thereby neglecting the logical layout of the document image, which is vital for high-quality and contextually accurate translation.

2.2 Post-OCR Paragraph Clustering

In cascaded DIT systems, post-OCR processing is required to aggregate fragmented text lines from OCR, which provides both text and bounding box coordinates, into coherent paragraphs for translation.

Early research employed geometric and rule-based approaches. Both categories include algorithms to find column gaps by searching for white space [3] or text alignment [19]. Limitations of these approaches include susceptibility to input noise and false positive column boundaries from monospace font families. Especially when handling scene text with perspective distortions from camera angles, vision-based algorithms can be fragile and inconsistent. [21]

Methods based on deep neural networks like Deep Layer Aggregation [25], Graph Neural Network (GNN) [18], Graph Convolutional Network (GCN) [26]

also exist [21,24]; however, due to their insufficient incorporation of semantic information, these methods may lack the capability to effectively address complex situations encountered in real-world scenarios, such as intricate shapes, rotations, and distortions.

3 Method Description

3.1 Task Formulation

Formally, the goal of clustering is to find an optimal partition $P^* = \{p_1, \ldots, p_M\}$ of input lines $L = \{l_1, \ldots, l_N\}$ maximizing an objective $\mathcal{F}$: $\text{argmax}_{P \in \Pi(L)} \mathcal{F}(P \mid L, \Theta)$ where $\Pi(L)$ is the set of all valid partitions of L; $\mathcal{F}(P \mid L, \Theta)$, parameterized by Θ, evaluates partition P's quality.

3.2 Model Architecture

Overall Workflow. Our model employs a synergistic workflow to reconstruct coherent paragraphs from OCR-derived text lines. The process initiates with the Completeness Judgment Model , which assesses each text line or initial short segment for semantic completeness. Lines or segments identified as incomplete by the CJM are then prioritized for extension. To find suitable candidates for merging, the Spatial Neighbor Selection module identifies physically proximate text lines, excluding impossible candidates. Subsequently, the Semantic Concatenation Model evaluates whether an incomplete segment and a selected spatial neighbor can be meaningfully and coherently concatenated. If a concatenation is validated by the SCM, the segments are merged, and the newly formed segment is re-evaluated by the CJM. This loop terminates when all segments satisfy CJM's completeness criterion or SCM rejects all potential mergers with neighbors.

Completeness Judgment Model. The CJM is designed to assess whether a given sequence of text lines constitutes a semantically complete paragraph. Given a candidate paragraph $p_c = (l_i, l_{i+1}, \ldots, l_j)$, which is a sequence of text lines, the CJM outputs a probability score $S_{comp}(p_c)$ indicating the likelihood that p_c is a complete paragraph:

$$S_{comp}(p_c) = P(\text{is\_complete} \mid p_c, \Theta_{CJM})$$

where Θ_{CJM} are the parameters of this model. This score is crucial for determining appropriate paragraph break points, ensuring that generated paragraphs are self-contained units of meaning.

Spatial Neighbor Selection. The SNS module leverages the geometric layout of text lines to identify potential candidates for grouping. For each text line

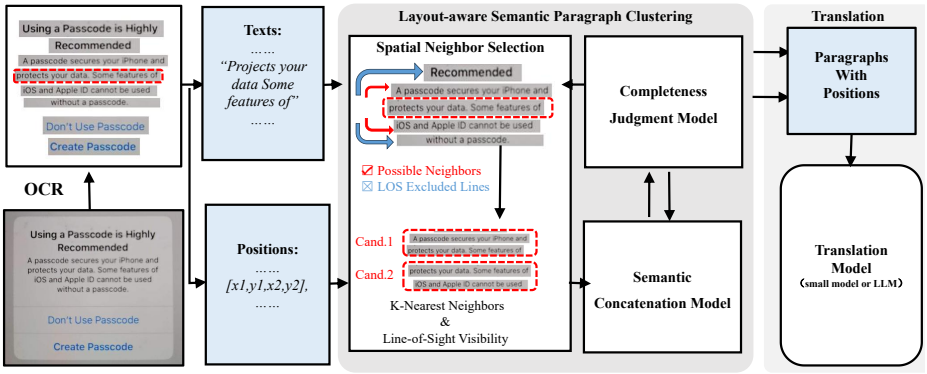


Fig. 2. Model architecture: Overview of our paragraph aggregation framework. It processes OCR-extracted text lines using a Completeness Judgment Model, Semantic Concatenation Model, and Spatial Neighbor Selection module to reconstruct coherent paragraphs. Output paragraphs are sent to a translation model. The solid arrows indicate the direction of data flow.

$l_k \in L$, defined by its vertices $V_k = \{v_{k,1}, \ldots, v_{k,m}\}$, its centroid $c_k = (x_k, y_k)$ is computed as: $c_k = \frac{1}{m}\sum_{s=1}^{m} v_{k,s}$

Given a line l_i, the SNS module determines its set of spatial neighbors, $N_S(l_i) \subset L \setminus \{l_i\}$. This set is constructed using a combination of the following criteria:

- **Distance Thresholding:** Lines l_j whose centroids c_j are within a specified Euclidean distance $\delta_{spatial}$ from c_i.

$$N_S(l_i) = \{l_j \in L \setminus \{l_i\} \mid \|c_i - c_j\|_2 \leq \delta_{spatial}\}$$

- **K-Nearest Neighbors (KNN):** The K lines $l_j \in L \setminus \{l_i\}$ whose centroids c_j are closest to c_i based on Euclidean distance.
- **Line-of-Sight (LoS) Visibility:** Lines $l_j \in L \setminus \{l_i\}$ such that the line segment connecting c_i and c_j, denoted $\overline{c_i c_j}$, does not intersect the bounding box of any other text line l_k (where $k \neq i, j$).

$$N_S(l_i) = \{l_j \in L \setminus \{l_i\} \mid \forall l_k \in L \setminus \{l_i, l_j\}, \overline{c_i c_j} \cap \text{Bbox}(l_k) = \emptyset\}$$

where $\text{Bbox}(l_k)$ represents the bounding box of line l_k.

The output $N_S(l_i)$ provides a spatially constrained set of lines, pruning the search space for paragraph construction by focusing on physically proximate candidates.

Semantic Concatenation Model. The SCM evaluates whether a text sequence Seq_B should logically and semantically follow another sequence Seq_A to form an extended, coherent text block. Given two sequences of text lines,

$Seq_A = (l_{a_1}, \ldots, l_{a_n})$ and $Seq_B = (l_{b_1}, \ldots, l_{b_m})$, the SCM outputs a score $S_{concat}(Seq_A, Seq_B)$. This score represents the probability of the event should _concatenate, signifying that Seq_B indeed forms such a logical and semantic continuation of Seq_A:

$$S_{concat}(Seq_A, Seq_B) = P(\text{should\_concatenate} \mid Seq_A, Seq_B, \Theta_{SCM})$$

where Θ_{SCM} are the model parameters. This model is trained using the Text-line Order Prediction (TOP) task, as detailed in the following section. The SCM is crucial for deciding whether to merge adjacent lines or segments into a single paragraph, considering both semantic relatedness and logical flow.

3.3 Training Strategy

The CJM is trained as a binary classifier to distinguish between complete and incomplete paragraphs. Training data is directly extracted from ground truth paragraph annotations.

For the SCM, we introduce the **Text-line Order Prediction (TOP)** task. Inspired by Next Sentence Prediction (NSP) from BERT [4] and Sentence Order Prediction (SOP) from ALBERT [9], TOP is specifically tailored for sequences of OCR-derived text lines. Unlike NSP, which primarily assesses thematic relatedness, and SOP, which focuses on local coherence (often by detecting reversed sentence pairs), TOP uniquely integrates spatial layout information from the source document or scene image into its sample construction.

The TOP task involves classifying pairs of text sequences, $[Seq_A, Seq_B]$, as either positive or negative. Samples are constructed as follows:

- **Positive Samples** ($[Seq_A, Seq_B]$): Consist of two text sequences where Seq_A immediately precedes Seq_B within the same original paragraph. These include:
  (1) *Consecutive lines*: e.g., $[line_i, line_{i+1}]$.
  (2) *Contiguous segments*: A paragraph randomly partitioned into two multi-line segments (Seq_A, Seq_B), simulating erroneous paragraph breaks.
- **Negative Samples**: Designed to challenge the model's understanding of sequence order and coherence.
  (1) *Swapped order*: Reversing a positive pair (e.g., $[Seq_B, Seq_A]$).
  (2) *Non-contiguous*: Pairing non-adjacent sequences from the same paragraph (e.g., $[Seq_A, Seq_C]$ from a paragraph Seq_A, Seq_B, Seq_C).
  (3) *Spatially successive but unrelated*: Pairing sequences (single or multi-line) that are spatially close in the document/image but originate from different paragraphs (i.e., different semantic contexts).

By simulating complex real-world scenarios, this strategy enables the SCM to develop a robust understanding of semantic coherence and logical flow, thereby enhancing robustness in handling diverse challenges commonly encountered in real-world OCR applications.

Both models are initialized using pre-trained weights from ALBERT [9], while the classifier layer employs Kaiming Uniform initialization. Then, we fine-tune the models on the DIT700K dataset [28].

4 Experiments

4.1 Experiment Settings

Datasets and Evaluation Metrics To evaluate the proposed method, we conduct extensive experiments on DIT700K [28], currently the largest publicly available benchmark for Document Image Translation (DIT). This comprehensive dataset comprises 718,000 high-quality document images with bilingual (English-Chinese) content, featuring three distinctive characteristics: comprehensive multi-level, fine-grained annotations, including word text and bounding boxes, sentence prefixes, order, and translation, and full document translation.

For performance evaluation, we employ the following standard metrics:

- BLEU [14]: n-gram precision, focusing on translation accuracy.
- chrF++ [15]: Combines character and word n-gram F-score, balancing precision and recall.

Setups Following the methodology of [28] for constructing the DIT700K test set, we constructed two distinct test sets: one featuring complex layouts and another with simple layouts. Each set comprises 1024 examples and was designed to evaluate the performance of both baseline methods and our proposed approach across varying degrees of layout complexity. To eliminate potential OCR noise that could confound translation quality assessment, we utilized ground truth text content and bounding boxes directly from these test sets for all experiments.

The translation model is based on the Qwen2.5-7B architecture [16]. It was fine-tuned for 2000 steps on the DIT700K training set using Low-Rank Adaptation (LoRA) [6]. The classifier heads in both the SCM and CJM modules consist of a fully-connected layer with a dropout rate of 0.1. Within the SNS module, the spatial threshold, denoted as $\delta_{spatial}$, was set to half the image width, and K, the number of nearest neighbors, was configured to 3 for the K-nearest neighbors selection. Model parameters were initialized using the albert-xlarge-v2 checkpoint from the ALBERT model [9].

Baselines We compare our method against several representative baselines:

- **EasyOCR [8]:** A rule-based method for paragraph aggregation.
- **Docxchain [24]:** A CNN-based layout parser.
- **DIMTDA [10]:** An end-to-end translation model based purely on visual information.
- **LayoutXLM-Dec [22], LiLT-XLM-Dec [20],LayoutLMv3-Dec [7], BROS-Dec [5], and ZoomDIT [28]:** These models integrate both visual and textual layout information for document understanding and translation.
- **GPT-4o (via Azure API) [13]:** A multimodal Large Language Model (LLM). For this baseline, the model receives only the raw image and a translation instruction as input, without any pre-extracted textual content.

In our experiments, our proposed method, alongside EasyOCR and Docxchain, is introduced as post-processing modules for text information in cascaded models. The training configurations and hyperparameter settings for all relevant models generally adhere to those specified in their respective original publications or follow the setup detailed in [28] for consistency, unless otherwise noted.

Table 1. Results on DIT700K EN (EN-ZH) and DIT700K ZH (ZH-EN) datasets under simple and complex layouts, evaluated by BLEU and chrF++ scores. The values in **bold** and underlined represent the best and second-best results respectively.

DIT700K EN (en-zh)						
methods	**simple**		**complex**		**Average**	
	BLEU	**chrF++**	**BLEU**	**chrF++**	**BLEU**	**chrF++**
Multimodal LLM						
GPT-4o	44.02	47.30	34.72	39.61	39.37	43.46
End2End DIT model						
DIMTDA	34.65	45.93	25.49	35.11	30.07	40.52
LayoutXLM-Dec	41.97	51.46	32.07	43.06	37.02	47.26
LayoutLMv3-Dec	32.58	43.37	28.90	39.87	30.74	41.62
BROS-Dec	41.36	51.68	33.31	44.35	37.34	48.02
ZoomDIT	44.45	**54.52**	37.07	47.42	40.76	**50.97**
Cascaded DIT System						
Rule-based	37.20	44.35	30.32	33.67	33.76	39.01
CNN-based	48.04	43.51	48.84	43.47	48.44	43.49
Ours	**59.02**	51.15	**57.24**	**49.31**	**58.13**	50.23
DIT700K ZH (zh-en)						
Multimodal LLM						
GPT-4o	41.03	57.78	33.65	48.4	37.34	53.09
End2End DIT model						
LayoutXLM-Dec	42.83	67.23	31.53	55.17	37.18	61.20
LiLT-XLM-Dec	37.05	61.74	29.04	51.58	33.05	56.66
ZoomDIT	44.45	**67.25**	**39.86**	**62.59**	**42.16**	**64.92**
Cascaded DIT system						
Rule-based	23.77	34.51	22.61	30.75	23.19	32.63
CNN-based	26.39	47.41	28.85	54.11	27.62	50.76
Ours	**45.96**	65.22	37.57	61.66	41.77	63.44

4.2 Comparison with Prior State-of-the-Art

Table 1 presents comprehensive performance comparisons across all methods on both DIT700K EN→ZH and ZH→EN translation benchmarks, demonstrating the effectiveness of our semantic-spatial paragraph reconstruction framework.

English-to-Chinese Translation Performance. Our framework achieves exceptional performance on the EN→ZH translation task, establishing new state-of-the-art results across both layout categories. For documents with simple layouts, our method attains the highest BLEU score of 59.02, representing a substantial 14.57-point improvement over the strongest baseline (ZoomDIT: 44.45). This improvement demonstrates superior translation accuracy through effective semantic paragraph reconstruction. While ZoomDIT leads in chrF++ (54.52)—a metric that balances character and word-level precision and recall—our model's chrF++ score of 51.15 remains highly competitive, showing only a 3.37-point difference while significantly outperforming all other methods.

The performance advantage becomes even more pronounced for complex layouts, where our approach achieves the top scores for both BLEU (57.24) and chrF++ (49.31). Compared to ZoomDIT, we observe improvements of 20.17 and 1.89 points respectively, validating our framework's robustness in handling irregular document structures where purely geometric clustering approaches typically struggle. These substantial improvements across complex layouts underscore the critical importance of semantic understanding in document structure reconstruction.

Chinese-to-English Translation Performance. For the ZH→EN translation direction, our method maintains strong competitive performance while demonstrating consistent improvements over traditional approaches. On simple layouts, our framework achieves the leading BLEU score of 45.96, outperforming ZoomDIT (44.45) by 1.51 points. Although ZoomDIT secures the highest chrF++ score (67.25), our model's chrF++ of 65.22 represents only a 2.03-point difference while substantially surpassing all other competing methods by margins exceeding 4 points.

For complex layouts, ZoomDIT achieves the best overall performance with BLEU (39.86) and chrF++ (62.59) scores. However, our method exhibits highly competitive results with BLEU of 37.57 and chrF++ of 61.66, maintaining performance gaps of only 2.29 and 0.93 points respectively. Importantly, our approach significantly outperforms all other baseline methods, with the second-best performer (LayoutXLM-Dec) achieving only 31.53 BLEU on complex layouts—a 6.04-point deficit compared to our method.

Table 2. Ablation study on DIT700K EN→ZH translation showing individual component contributions. Performance degradation compared to the full model is shown in red. Evaluation setup identical to Table 1.

Configuration	Simple Layout		Complex Layout		Average	
	BLEU	**chrF++**	**BLEU**	**chrF++**	**BLEU**	**chrF++**
Ours Model	59.02	51.15	57.24	49.31	58.13	50.23
w/o SCM	33.37 (-25.65)	38.59 (-12.56)	27.56 (-29.68)	33.72 (-15.59)	30.47 (-27.66)	36.16 (-14.07)
w/o CJM	47.06 (-11.96)	49.50 (-1.65)	26.18 (-31.06)	29.30 (-20.01)	36.62 (-21.51)	39.40 (-10.83)
w/o SNS	58.30 (-0.72)	50.07 (-1.08)	47.65 (-9.59)	44.17 (-5.14)	52.98 (-5.15)	47.12 (-3.11)

4.3 Ablation Study

We systematically ablate key components to evaluate their individual contributions on the DIT700K EN→ZH benchmark (Table 2). The analysis reveals the critical importance of each module, with performance degradations clearly quantified relative to the full model.

SCM Criticality: Removing the Semantic Concatenation Model causes the most severe performance degradation across all metrics, with average BLEU dropping by 27.66 points and chrF++ by 14.07 points. The impact is particularly pronounced on complex layouts (29.68 BLEU point decrease), confirming that semantic coherence assessment is essential for handling irregular document structures where spatial cues alone prove insufficient.

CJM Significance: The Completeness Assessment Model removal shows asymmetric impact across layout types. While simple layouts experience moderate degradation (11.96 BLEU points), complex layouts suffer severe performance loss (31.06 BLEU points), highlighting CJM's critical role in determining paragraph boundaries when geometric heuristics fail.

SNS Contribution: Although the Spatial Neighbor Selection module shows the smallest individual impact, it provides crucial spatial constraints that prevent semantically plausible but spatially implausible merging decisions. The 5.15 average BLEU point degradation, with more pronounced impact on complex layouts (9.59 points), demonstrates its importance as a necessary filtering mechanism.

4.4 Enhancing LLM translation via paragraph clustering.

To demonstrate the value of our method as a pre-processing step for modern translation systems, we evaluate its impact on Large Language Models (LLMs) by comparing their performance on raw OCR lines versus lines structured by our clustering approach and existing approach. We evaluate its impact by comparing three distinct input configurations: (1) raw OCR text lines, (2) text grouped by a rule-based method, and (3) text structured by our proposed paragraph clustering method. Evaluation methods are the same as the main experiment on DIT700K en-zh dataset. As shown in Table 3, our framework is a crucial component that significantly enhances cascaded translation systems. By providing semantically coherent context, our method dramatically improves translation quality, elevating the BLEU score on the GPT-4o model from 35.62 (raw text) to 53.43, and on DeepseekV3 from 35.58 to 50.02. This substantial improvement demonstrates the framework's ability to unlock the full potential of downstream translation models, pushing their performance to a state-of-the-art level.

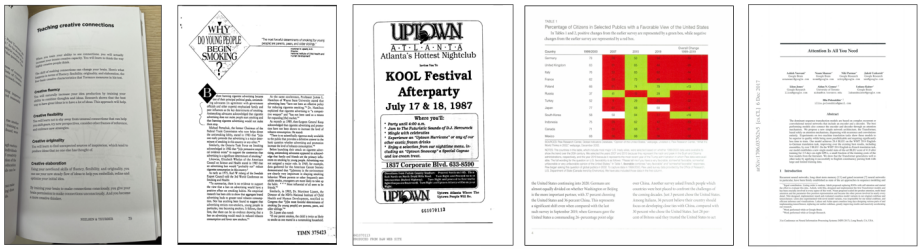
4.5 Cross-Domain Generalization and LLM Comparison

Cross-Domain Generalization Analysis. To assess the transferability of our approach across diverse document types, we evaluate our models using the DITrans dataset [27], which comprises meticulously annotated document paragraphs from diverse domains, exhibiting significant variations in textual content

Table 3. LLM Translation Enhancement on Complex Layouts

Clustering methods	DeepSeek-V3		GPT-4o	
	BLEU	chrF++	BLEU	chrF++
Raw OCR lines	35.58	40.23	35.62	42.25
Rule-based clustering	38.70	42.37	41.62	44.14
Our clustering	**50.02**	**44.52**	**53.43**	**47.31**

and layout structure (Figure 3). Both models were fine-tuned exclusively on the 'book' domain of DITrans and subsequently evaluated on four distinct domains: 'News', 'Political reports', 'Advertisement', and 'Scientific articles'. Both CJM and SCM were trained exclusively on the 'book' domain and evaluated on four target domains: News, Political reports, Advertisement, and Scientific articles. Table 4 shows CJM's superior cross-domain generalization (85.33 average F1) compared to SCM (78.95), with both models performing best on Political and News domains that structurally resemble the training domain. Performance degradation on Advertisement and Scientific domains reflects their distinct layout characteristics and specialized vocabularies.



(a) Paper Books (b) News (c) Advertisements (d)Political Reports (e) Scientific Article

Fig. 3. Five domains from DITrans dataset.

Table 4. F1 Scores for CJM and SCM Models Across Different Datasets

Cross-Domain Evaluation					
models	**News**	**Political reports**	**Advertisement**	**Scientific articles**	**Avg.**
CJM	91.3	93.7	79.8	76.5	85.33
SCM	86.2	83.3	73.6	72.7	78.95

Comparison with Large Language Models. We compare our specialized models against state-of-the-art LLMs (DeepSeek-V3 [11] and GPT-4o) on completeness assessment and text-line order prediction tasks using both zero-shot

and few-shot (2 examples) prompting strategies. Our models significantly outperform LLMs on both tasks (Table 5). CJM achieves 92.6% F1 on completeness assessment (10.1 points above GPT-4o few-shot), while SCM reaches 86.7% on order prediction (2.2 points above GPT-4o few-shot). Error analysis reveals systematic biases limiting LLM effectiveness: (1) **Length bias**—disproportionately associating longer segments with completeness regardless of semantic coherence; (2) **Coherence detection limitations**—struggling with inter-line continuity assessment in multi-line layouts. These biases, particularly evident on the most challenging 10% of test examples, demonstrate that specialized architectures remain superior for fine-grained document layout tasks despite LLMs' general capabilities.

Table 5. F1 Score Comparison with DeepSeek-V3 and GPT-4o

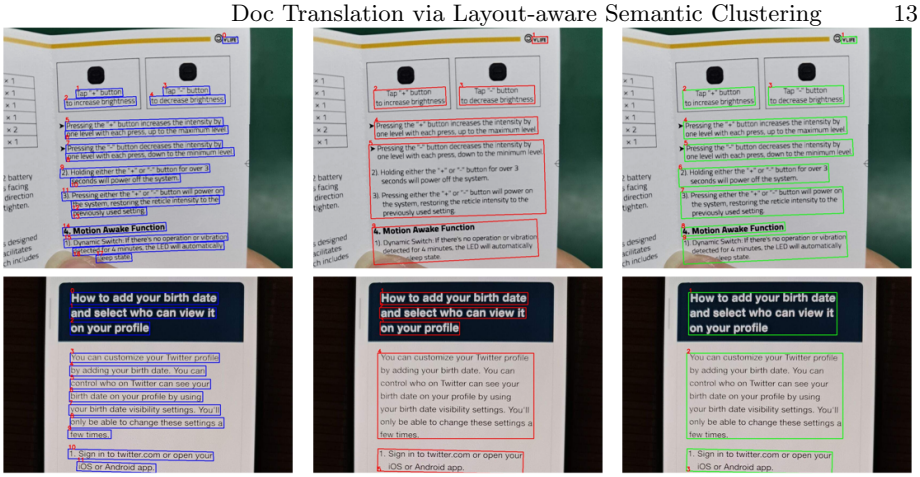
F1 Score Comparison with SOTA LLM		
models	**Completeness**	**Text-line Order**
DeepSeek-V3 (zero-shot)	75.4%	78.4%
DeepSeek-V3 (few-shot)	67.3%	68.9%
GPT-4o (zero-shot)	80.3%	81.2%
GPT-4o (few-shot)	82.5%	84.5%
Ours	**92.6%**	**86.7%**

4.6 Visualization

As demonstrated in Figure 4, in contrast to the baseline's over-clustering or under-clustering (b), which erroneously merges section headings with subsequent descriptive text or wrongly separate paragraphs (e.g., the 'Motion Awake Function' heading incorrectly merged with its three-line description; the title is separated into three lines), our approach (c) accurately distinguishes hierarchical structures and preserves logical semantic boundaries.

5 Conclusion

This paper proposes a novel text line clustering framework, which uniquely prioritizes textual semantics and logical reading order, leveraging a Completeness Judgment Model (CJM), an Semantic Concatenation Model (SCM), and a Spatial Neighbor Selection (SNS) module. This iterative process effectively reconstructs semantically coherent paragraphs, serving as an advanced OCR post-processing step. Experimental results on the DIT700K benchmark demonstrate that our framework significantly outperforms existing methods in paragraph clustering and substantially enhances downstream DIT performance, pushing cascaded DIT systems to a higher performance level.



(a) Detected text lines (Input) (b) Clustered paragraphs by baseline method (c) Clustered paragraphs by proposed method

Fig. 4. Illustration of clustering on real-world photographed documents. (a) Text lines as inputs (b) baseline (Docxchain) results and (c) results of proposed method. The red boxes in (b) indicate **over-clustering or under-clustering**.

Acknowledgments The research work has been supported by the Natural Science Foundation of China under Grant No. 62476275 and No.62476271. This work is also supported by Young Scientists Fund of The State Key Laboratory of Multimodal Artificial Intelligence Systems (ES2P100120, MAIS2024316).

References

1. Afli, H., Way, A.: Integrating optical character recognition and machine translation of historical documents. In: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). pp. 109–116 (2016)
2. Binmakhashen, G.M., Mahmoud, S.A.: Document Layout Analysis: A Comprehensive Survey. ACM Comput. Surv. **52**(6), 109:1–109:36 (Oct 2019), `https://doi.org/10.1145/3355610`
3. Breuel, T.M.: An algorithm for finding maximal whitespace rectangles at arbitrary orientations for document layout analysis. In: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. pp. 66–70. IEEE (2003)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
5. Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., Park, S.: Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents (2022), `https://arxiv.org/abs/2108.04539`
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)

7. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM international conference on multimedia. pp. 4083–4091 (2022)
8. JaidedAI: Easyocr. Available online: https://github.com/JaidedAI/EasyOCR (2023), accessed: 2025-06-01
9. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
10. Liang, Y., Zhang, Y., Ma, C., Zhang, Z., Zhao, Y., Xiang, L., Zong, C., Zhou, Y.: Document image machine translation with dynamic multi-pre-trained models assembling. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 7084–7095 (2024)
11. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
12. Ma, C., Zhang, Y., Tu, M., Han, X., Wu, L., Zhao, Y., Zhou, Y.: Improving end-to-end text image translation from the auxiliary text translation task. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 1664–1670. IEEE (2022)
13. OpenAI: Gpt-4o system card (2024), https://arxiv.org/abs/2410.21276
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040/
15. Popović, M.: chrF++: words helping character n-grams. In: Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Kreutzer, J. (eds.) Proceedings of the Second Conference on Machine Translation. pp. 612–618. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/W17-4770, https://aclanthology.org/W17-4770/
16. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), https://arxiv.org/abs/2412.15115
17. Sable, N.P., Shelke, P., Deogaonkar, N., Joshi, N., Kabadi, R., Joshi, T.: Dochandler: Document scanner, manipulator, and translator based on image and natural language processing. In: 2023 International Conference on Emerging Smart Computing and Informatics (ESCI). pp. 1–6. IEEE (2023)
18. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE transactions on neural networks **20**(1), 61–80 (2008)
19. Smith, R.W.: Hybrid Page Layout Analysis via Tab-Stop Detection. In: 2009 10th International Conference on Document Analysis and Recognition. pp. 241–245 (Jul 2009). https://doi.org/10.1109/ICDAR.2009.257, https://ieeexplore.ieee.org/document/5277715, iSSN: 2379-2140

20. Wang, J., Jin, L., Ding, K.: Lilt: A simple yet effective language-independent layout transformer for structured document understanding (2022), https://arxiv.org/abs/2202.13669
21. Wang, R., Fujii, Y., Popat, A.C.: Post-ocr paragraph recognition by graph convolutional networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 493–502 (2022)
22. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.: Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. arXiv preprint arXiv:2104.08836 (2021)
23. Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D., Lee Giles, C.: Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5315–5324 (2017)
24. Yao, C.: Docxchain: A powerful open-source toolchain for document parsing and beyond. arXiv preprint arXiv:2310.12430 (2023)
25. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2403–2412 (2018)
26. Zhang, S., Tong, H., Xu, J., Maciejewski, R.: Graph convolutional networks: a comprehensive review. Computational Social Networks **6**(1), 1–23 (2019)
27. Zhang, Z., Zhang, Y., Liang, Y., Xiang, L., Zhao, Y., Zhou, Y., Zong, C.: Layoutdit: Layout-aware end-to-end document image translation with multi-step conductive decoder. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 10043–10053 (2023)
28. Zhang, Z., Zhang, Y., Liang, Y., Xiang, L., Zhao, Y., Zhou, Y., Zong, C.: From chaotic ocr words to coherent document: A fine-to-coarse zoom-out network for complex-layout document image translation. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 10877–10890 (2025)
29. Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International conference on document analysis and recognition (ICDAR). pp. 1015–1022. IEEE (2019)