Towards Better Dissemination and Preservation: End-to-End Chinese Historical Document Digitization

Anonymous ACL submission

Abstract

Historical documents serve as the carrier of massive Chinese history and culture. Increasing works try to digitize historical documents by recognizing the context of books with Optical Character Recognition (OCR) for better preservation and propagation. However, previous works are unpractical for digitization since they focused on isolated fundamental tasks, such as single character recognition or line detection, whereas their outputs are low-level components such as isolated characters instead of readable context, can not fulfill the applicable digitization. To this end, we introduce the first end-to-end benchmark for digitizing Chinese historical documents, targeting well-formatted and human-readable outputs. This task is challenging due to the visual variability such as diverse page layouts and the need for deep textual understanding to maintain semantic coherence and consistency. To address these issues, we propose two complementary components: 1) Document Image Augmentation tailored to simulate visual artifacts and layout diversity. 2) Correction-Based Post-Editing that corrects textual errors to enforce semantic coherence. Experiments demonstrate the advantage of our proposed model over cutting-edge baselines, underscoring the necessity of introducing this new setting, thereby facilitating a solid precondition for protecting and propagating the already scarce resources.

1 Introduction

Chinese historical documents have served as the record medium of splendid history and knowledge for thousand years, among which there exist classic heirloom works such as *Yongle Dadian* (永乐大典) and *Siku Quanshu* (四库全书). These masterpieces hold profound significance in shaping traditional Chinese cultural identity (Shi et al., 2023) and nature (Su et al., 2022).

However, although with massive inventory, only nearly 20% (Wang, 2024) of them have been digitized, putting them in a low-resource situation. Current existing resources of historical document are mainly in images, previous efforts can be divided into three manners: 1) Recovering (Borenstein et al., 2023) the images with flaws into complete; 2) Normalization (Robertson and Goldwater, 2018; Lyu et al., 2021) to align historical texts with their modern counterparts. 3) Recognition of the context in the historical document image, mainly with OCR technique (Li et al., 2022; Ren et al., 2015; Dan et al., 2022).

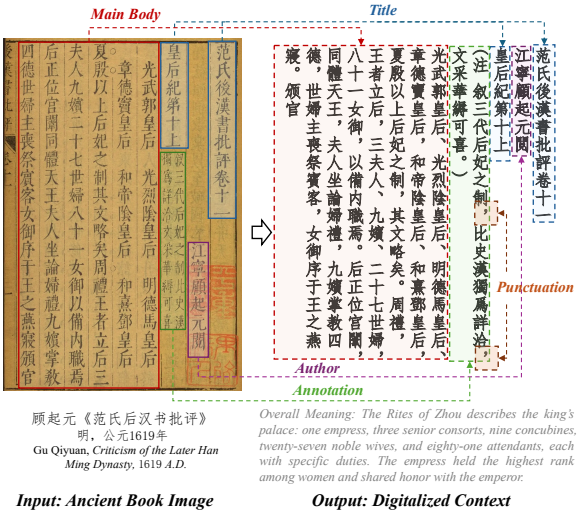


Figure 1: The illustration of our end-to-end Chinese Historical Document Digitization.

Despite their effectiveness, previous works' modelings are impractical to fulfill the one-stop digitization, the former recovery and normalization only focus on a specific pre- or post-recognition stage in the entire workflow while the recognition works somehow stick to low-level fundamental tasks, where their targets are unreadable components, such as isolated characters from character recognition (Shi et al., 2023), line positions from line detection (Shi et al., 2019), writing orders from order prediction (Ma et al., 2020), lack-

ing formats and punctuations, requiring complex combination to form the final readable context, cannot fulfill the effective digitization.

We thus propose a new task: end-to-end digitization of Chinese historical documents, which performs one-stop conversion from document images to structured, readable text as shown Figure 1, which include punctuated, semantically coherent content with explicit layout, including title, author, annotations (注疏), and main body. This task is designed to align with real-world digitization needs for ancient books, enabling holistic, globally optimized solutions rather than piecemeal pipelines.

To effectively benchmark this task, we construct a dataset named Ancient Chinese Book Recognition (ACBR). On the basis of classic document images, we build the dataset by hiring native speakers to collect the context in the image, which includes the 400,748 characters distributed in 1,992 pages from 46 books, with a time span of 10 dynasties from Han (汉) to Ming (明), across 1,500 years. Different from previous datasets that did not have distinctions between characters, our dataset divides them into title, author, annotation (注疏), and main body. Moreover, we also include the necessary components to be readable: punctuation (句读). Our dataset is designed to cover corner cases as many as possible, the perspectives include the variations of different writing styles from neat to scribble, the topics from government documents to religions, formations from preface (序言) to poems, and even with the stamps that could disturb the recognition. Thereby our ACBR can facilitate the exhaustive benchmark of our digitization task.

Yet end-to-end digitization of ancient book pages is challenging. The first set of difficulties arises from the visual modality as shown in Figure 2, encompassing: **1) Layout** that each page could include various text blocks for different information, such as main body, title or annotations (注疏), these blocks have totally different positions and sizes, not to mention the absence of punctuation between and inside these blocks; **2) Writing styles** stemming from individual habits, such as the Kai (正楷), which is famous for its neatness, but Mi (米芾) is scribbled, having characters naturally joined-up and overlapped; **3) Noise** which could include seals (印章), inscriptions (落款), and pagination (页码) that could seriously disturb the recognition. Additionally, previous subtasks such as character recognition only targets isolated characters, the shift from sub-tasks to complete context digitization introduces the second challenge of ensuring **4) Contextual Coherence**: how to eliminate the misrecognition that totally incoherent with the context, a tiny misrecognition can lead to a misinterpretation of the entire sentence, as shown in Figure 2 d).

In this study, we first tackle visual challenges with a proposed Document Image Augmentation, which progressively refines training images to closely mimic real ancient book pages via three augmentation strategies. We then ensure contextual coherence through our Correction-Based Post-Editing, leveraging a concise set of editing operations to emulate the human process of fixing incoherent errors, strengthened by our novel Bidirectional Edition Alignment that maximizes correction effectiveness, thereby distinguishing our model from prior purely visual methods.

We finally benchmark our dataset with our method and a set of representative baselines. The empirical experiments highlight the advantage of our method in recognizing historical document images and validate our motivation of proposing this new task for furthering the preservation and propagation of ancient Chinese book.

2 Related Work

2.1 Historical Document Digitization

Previous efforts of digitization generally include: 1) Recovering the images with flaws before the recognition to avoid error propagation (Borenstein et al., 2023). 2) Normalization (Robertson and Goldwater, 2018; Lyu et al., 2021) after the recognition to align historical texts with their modern counterparts (Rijhwani et al., 2020; Dong and Smith, 2018). 3) Recognition of the context in the historical document image, mainly with OCR technique (Li et al., 2022; Ren et al., 2015; Dan et al., 2022). Some work further tried OCR-free approaches such as Donut (Kim et al., 2022a) with image-encoder-text-decoder; Dessurt (Davis et al., 2023) and Pix2Struct (Lee et al., 2023) that were pretrained by masking image regions.

Nevertheless, previous work cannot fulfill the one-stop digitization: the recovery and normalization only focus on a specific pre- or post-recognition stage while the recognitions stick to low-level fundamental tasks, their targets are unreadable components, such as isolated characters from text recognition (Wang et al., 2019b; Long et al., 2020), line positions from line detection (Shi

et al., 2019; Ma et al., 2024; Mechi et al., 2019, 2021), writing order from order prediction (Ma et al., 2020; Wang et al., 2021), requiring complex downstream combination to form the readable context, cannot fulfill effective digitization.

Unlike previous works, our benchmark and model stand out as the first to focus on the practical setting of end-to-end Chinese historical document digitization, thereby guiding the holistic optimization of these real-world challenges.

2.2 Chinese Historical Document Digitization

The digitization of Chinese historical documents remains limited and largely follows the routine of English works, focusing on the recognition step (Peng et al., 2022; Wang et al., 2019a; Liu et al., 2019). However, besides from sticking to low-level sub-tasks, these methods also lack adaptation to the unique characteristics of Chinese documents: Ma et al. (2020) analyzed the document layouts but only detected the bounding box of text blocks; M5hisdoc (Shi et al., 2023) and SCUT-CAB (Cheng et al., 2022) further considered writing orders but missed punctuations; HisDoc1B (Shi et al., 2025) took the punctuation into consideration but their labels are annotated by trained model, can not guarantee the quality of it.

On the other hand, our benchmark and model distinguish themselves from previous studies by covering both reading orders and various layout text blocks, also collecting manually segmented punctuation, thereby facilitating effective end-to-end Chinese historical document digitization.

3 Task and Dataset

In this section, we first introduce the definition and challenges of our task, we then introduce how we build the ACBR dataset and our Document Image Augmentation that augment the train set to address the visual challenges as following.

3.1 End-to-End Ancient Chinese Book Digitization

As shown in Figure 1, we first define the input is the image of ancient book without any text. The output will be the punctuated context of title T, author A, main body M, annotation N in the page. The output should be readable and segmented sentence. Our task is then formulated as extracting a sequence of text elements from image I:

$$Output = [T, A, M, N] \quad (1)$$

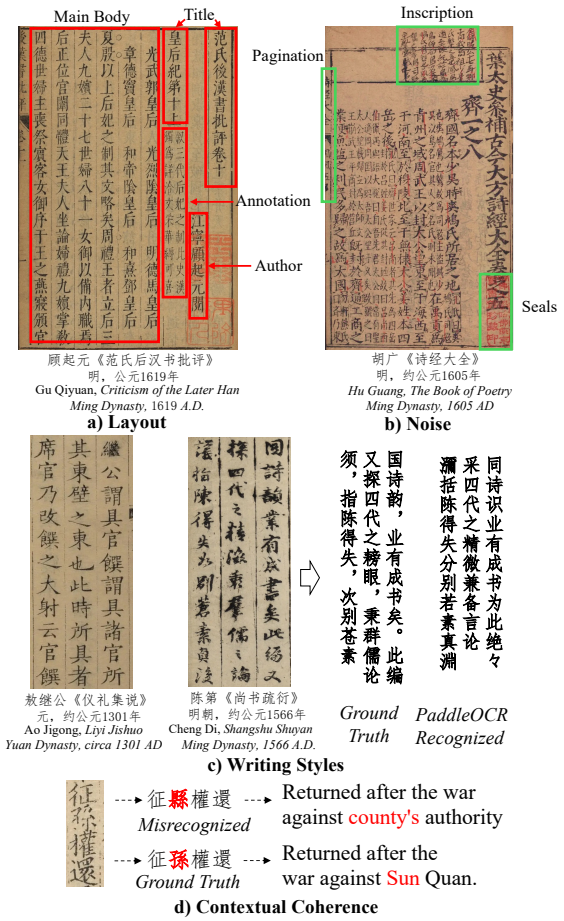


Figure 2: Illustration of challenges.

3.2 Challenges

The challenges of our task lie in the two modalities towards recognition. The first is the ***visual variations***, which include three aspects:

- **Layout**: The layout of historical documents include two levels: 1) Block-level, where text blocks are distributed on the page delivering different information such as annotation and title, as shown in Figure 2 a), distinguishing between them is needed for digitization; 2) Char-level, where the characters are also dispersed without any punctuation, making sentence segmentation necessary for digitization.

- **Writing styles**: As shown in Figure 2 c), the left style features neat, clear typography, while the right is wild, with varying word sizes and overlapping characters. We show that the widely used PaddleOCR (Du et al., 2021) struggles with this task on the right part of Figure 2 c).

- **Noise**: Besides being blurred due to poor storage, there are noises deliberately added by the

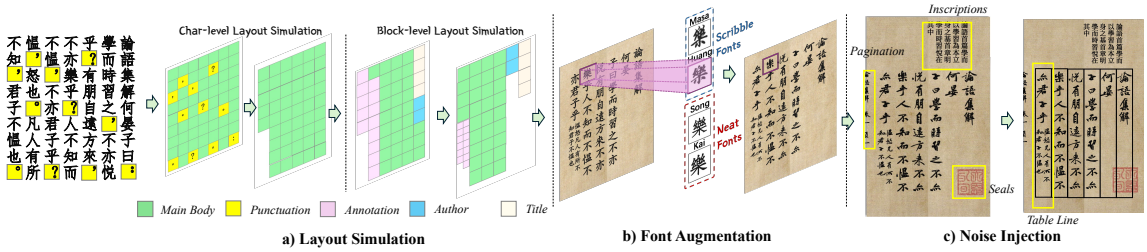


Figure 3: The illustration of our Document Image Augmentation.

depositories or authors, such as seals, inscriptions, and pagination as shown in Figure 2 b).

The second challenge arises due to the shift from isolated sub-tasks to complete digitization. Previous sub-tasks such as character recognition only needs to consider isolated characters, single misrecognition will not cause serious error in these sub-tasks, but in our digitization, such an incoherence will lead to the misunderstanding of the entire sentence semantics as shown in Figure 2 d). The challenging point here can be summarized as:

- **Contextual Coherence**: A tiny misrecognition can lead to a misinterpretation of the entire sentence, how to eliminate the misrecognition that totally incoherent with the surrounding context to ensure the contextual coherence of the recognized context is crucial to digitization.

3.3 Dataset Collection and Annotation

We construct a new dataset called **A**ncient **C**hinese **B**ook **R**ecognition **(ACBR)** for benchmarking. ACBR focuses on one of the most challenging and practical cases of end-to-end historical document recognition, thereby facilitating the solid benchmarking for downstream evaluation.

For the train and dev set, we collect 300,548 characters from the context of 1,492 classic Chinese literary works and sample 1,200 for the train set, the remained 292 works for the dev set. We build the input images by printing each sample into an image of 1024 $\times$ 1024 with the font of Song(宋), composing of the characters' pixel maps that are concatenated with a common classic Chinese writing order: from up to bottom and starting a new line on the left of current one.

For the testset, we collect 100,200 characters from 500 pages by hiring native speakers to collect historical document samples from *Ancient Classics* (识典古籍)[1]. We applied the following standards to simulate practical scenarios: 1) Each sample must contain a minimum of 20 and a maximum of 300 characters. 2) Irregular images such as covers and table of contents are not included. 3) Only complete, single-image pieces are accepted, partial or cropped images are excluded. 4) Only documents from Han (汉) to Ming (明) are included. Works from earlier or later periods are excluded due to being either too ancient or modern.

[1]https://www.shidianguji.com

3.4 Document Image Augmentation

After building the vanilla dataset, we further propose Document Image Augmentation to address the visual challenges. As shown in Figure 3, we propose three strategies to augment the input image in the train set in a pipeline manner to come close to the real document image step by step.

Layout Simulation

We first address the complex layout, which involves two levels: block-level and character-level layouts. As shown in Figure 3 a), we follow these steps when rendering our training images: 1) For the character-level layout, we render images by removing all punctuation while preserving the original word order; 2) For the block-level layout, title and author blocks are placed on a new line starting from the left edge, while annotation blocks are inserted into the main body block, positioned close to the annotated word. By adopting this approach, we emphasize that different blocks convey different types of information, making it necessary for the model to distinguish them.

Font Augmentation

We then address the challenge of diverse writing styles. Current pretrained vision-language models are unable to handle the wide variety of writing styles, as they are typically exposed only to standard fonts such as Song (宋体); while these fonts clearly represent character structures, they lack generalization to more cursive or scribbled

	Train	Dev	Test
#Chars	244,308	56,240	100,200
#Pages	1,200	292	500
#Chars / Pages	203.65	192.60	201.20
#Punctuations / Pages	32.60	30.78	33.23
#Books	34	9	13
#Samples / Books	35.29	32.44	38.46
#Dynasties	11	7	10
#Samples / Dynasties	109.09	41.71	50.00

Table 1: Statistics of our ACBR dataset.

Figure 4: Statistics of neatness levels in our testset.

characters. To overcome this limitation, we propose a Font Augmentation method utilizing two distinct font sets. As shown in Figure 3 b), the first set consists of neat fonts, like Song (宋体), representing standard characters. The second set includes scribble calligraphic fonts, such as Mifu (米芾), which capture writing styles beyond neat fonts. Each training image is re-rendered with one font from each set, enhancing the model's ability to generalize across varied writing styles.

Noise Injection

We finally move to the challenge of noise artifacts. Different from the common recognition (Liao et al., 2022) where all the text in the image are the target, there are texts in the ancient books are considered as noise such as seals and page numbers. We thus inject the noise into the image in our train set to enhance its robustness towards the noise. Particularly, three types of noise are injected: 1) Seals inserted with Seal Script (篆体) that randomly appear at any position in the image; 2) Inscriptions are generated by LLM with a prompt instructed to introduce the book, with a deliberate different font to distinguish from the author's scripts; 3) Pagination is generated similar to previous one, but printed in the same font. 4) Table lines are added to mimic the horizontal or vertical rules often found in ancient books.

3.5 Dataset Statistic and Analysis

We show detailed statistics of our ACBR data in Table 1. We can tell that there are average around 32 punctuations per sample, which are missed in the historical document image and post a hard challenge for the recognition to recover the punctuations properly. Besides, we also ensure the diversity of writing styles by extending the books and dynasties pool as large as possible, especially in the testset where only around 38 pages per book.

To quantify recognition difficulty, we partition the test set into three neatness levels using the average minimum edit distance per character between PaddleOCR (Du et al., 2021) outputs and ground truth. Samples with score ≥ 0.90 are labeled scribbled, those in $[0.85, 0.90)$ as medium, and those ≤ 0.85 as neat. As shown in Figure 4, 191 samples are scribbled, 209 are medium, and 98 are neat.

4 Correction-Based Post-Editing

Previous pure visual models may misrecognize characters that are totally incoherent with the surrounding context. Motivated by this limitation, we explore an alternative approach that leverages contextual information to correct these errors: we design a Correction-Based Post-Editing framework that emulate how a human editor would address these errors. We further adopt an LLM to produce the correct edition sequence and subsequently apply these editions with our Bidirectional Edition Alignment mechanism as shown in Figure 5.

4.1 Correct Editions

We design three correct editions to process the recognized sentence follow the editions in Levenshtein Distance, matching each character with an edition, specifically: $\mathcal{E} = \{\text{Insertion}(A), \text{Deletion}(A), \text{Equal}(A)\}$. Here, $\text{Insertion}(A)$ inserts the parameter character A into the sentence, $\text{Deletion}(A)$ removes current character A from the sentence, and $\text{Equal}(A)$ accepts (matches) the current character A as correct.

We omit the traditional $\text{Substitution}(A, B)$ operation and realize it as a consecutive pair $\text{Deletion}(A)$ followed by $\text{Insertion}(B)$, constraining the edition to a unified set of atomic operations, thereby reducing the search space while preserving functionality as shown in Figure 5 a).

We then fine-tune an LLM to generate edition sequences for recognized sentences. We adopt the VLM's output with possible errors in training as input. The output is organized based on minimal edit actions between the recognized sentence and the correct label in a dynamic programming approach.

4.2 Bidirectional Edition Alignment

After generating the edit sequence, we need a robust alignment between edits and characters. Even

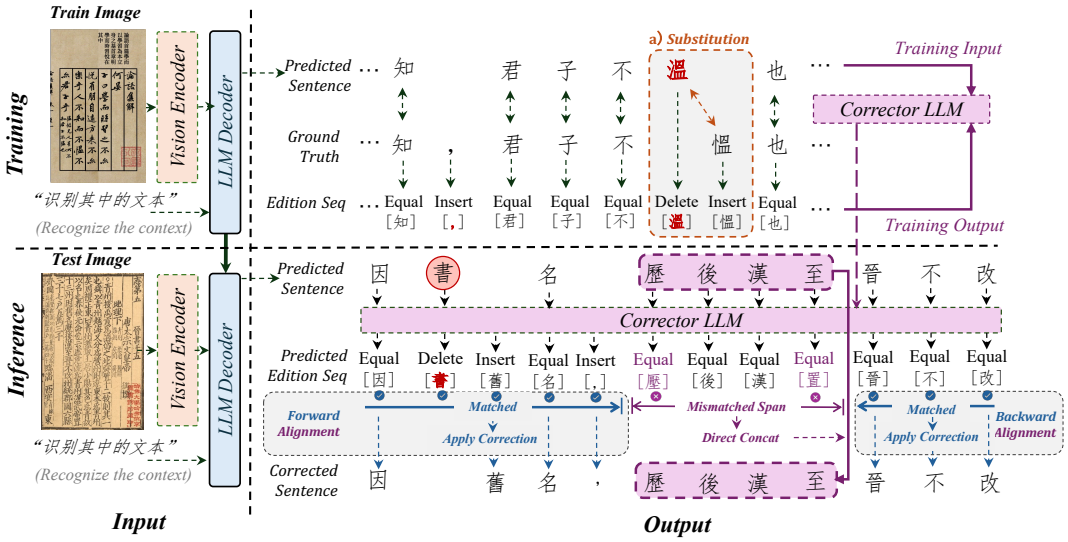


Figure 5: The illustration of our Correction-Based Post-Editing.

small mismatches can render the final sentence unreadable. As shown in Figure 5, we propose a bidirectional alignment strategy to maximize valid editions, it includes two conservative passes over the edit sequence $E = [e_1, e_2, \ldots, e_n]$ and the text sequence $T = [c_1, c_2, \ldots, c_m]$:

Forward Alignment Starting from the left boundary to the right as shown in Figure 5, we match pairs $[e_1, c_1], [e_2, c_2], \ldots$ and iteratively update the left corrected prefix T'_f by applying e_i to c_j until $i > n, j > m$, or an invalid edition occurs. Let the stopping indices be (i_f, j_f). The update rule is

$$\text{T'}_f \leftarrow T'_f + \text{apply}(e_i, c_j) \quad \text{if } e_i \text{ is valid} \tag{2}$$

where validity follows the criterion:

$$\text{Invalid}(\text{e}_i, c_j) \quad \text{if } (e_i \in \zeta) \wedge (e_i.p \neq c_j) \tag{3}$$

where $\zeta = \{\text{Equal}, \text{Deletion}\}$, $e_i.p$ represent the parameter of edition e_i, otherwise deemed valid.

Backward Alignment Symmetrically, starting from the right boundary to the left, we match pairs $[e_n, c_m], [e_{n-1}, c_{m-1}], \ldots$ and iteratively update the right corrected suffix T'_b until $i < 1, j < 1$, or an invalid edition occurs. Let the stopping indices be (i_b, j_b). The update mirrors the forward:

$$\text{T'}_b \leftarrow \text{apply}(e_i, c_j) + T'_b \quad \text{if } e_i \text{ is valid.} \tag{4}$$

Composition Given the two maximal valid regions, we compose the final corrected text T' by concatenation. If a non-overlapping mismatch span exists ($j_f < j_b$), we preserve the middle as:

$$\text{T'} \leftarrow T'_f + T[(j_f + 1) : (j_b - 1)] + T'_b \tag{5}$$

Otherwise if the two valid aligned regions meet, we resolve by concatenating the two passes result.

Our bidirectional alignment maximizes valid editions by halting early at the first inconsistency at both ends, and directly concatenates the unmatched middle, thereby preventing index shift and errors while maintaining readability.

5 Experiment

5.1 Dataset and Experiment Setting

We evaluate our method and other baselines on the proposed datasets. We employ a LoRA fine-tuned InternVL2.5-8B (Chen et al., 2024) for our Vision-Language Model, and a LoRA fine-tuned Qwen2.5-7B for the corrector, the implementing details can be found in Appendix A.

Following prior works (Wei et al., 2024a; Yousef and Bishop, 2020), our evaluation relies Character Error Rate (CER), F1, and BLEU. CER quantifies normalized edit distance at the character level. BLEU rewards correct ordering and n-gram consistency. F1 is computed purely over individual character recognitions without considering sentence structure. These measures can jointly assess character and sequence level quality.

5.2 Main Result

In Table 2, we present a comprehensive comparison with cutting edge baselines, include: common OCRs: 1) PaddleOCR (Du et al., 2021), 2) EasyOCR (JaidedAI, 2024); Document recognitions: 1) Donut (Kim et al., 2022b), 2) EffOCR (Carlson et al., 2024); VLM baselines, include off-the-

Method	↑ P.	↑ R.	↑ F1.	↓ CER	↑ BLEU
Common OCR Baselines					
PaddleOCR(off-the-shelf) (Du et al., 2021)	0.6348	0.6119	0.6348	0.8256	0.1335
EasyOCR(off-the-shelf) (JaidedAI, 2024)	0.6712	0.6328	0.6514	0.7924	0.1577
Document Recognition Baselines					
Donut(off-the-shelf) (Kim et al., 2022b)	0.4011	0.2111	0.2766	0.8767	0.0012
EffOCR (Carlson et al., 2024)	0.7153	0.7419	0.7283	0.7335	0.1634
VLM Baselines					
GPT-4o(off-the-shelf) (OpenAI, 2024)	0.7553	0.7358	0.7454	0.5973	0.2031
Deepseek-VL2(off-the-shelf) (Wu et al., 2024)	0.7712	0.7473	0.7590	0.6548	0.2384
GOT-OCR2.0 (Wei et al., 2024b)	0.6137	0.6892	0.6492	0.7041	0.1213
Vary (Wei et al., 2023)	0.6762	0.6691	0.6726	0.7283	0.1033
Qwen2-VL-7B (Wang et al., 2024)	0.8387	0.8093	0.8237	0.4526	0.3527
Qwen2.5-VL-7B (Yang et al., 2024)	0.8687	0.8399	0.8540	0.4241	0.3590
InternLM-XComposer (Dong et al., 2024)	0.7734	0.8147	0.7935	0.5267	0.3185
InternVL2.5-8B (Chen et al., 2024)	0.8473	0.8267	0.8368	0.3939	0.3784
Ours	**0.9089**	**0.8392**	**0.8726**	**0.3205**	**0.4085**

Table 2: Comparison with baselines.

Layout	InternVL2.5-8B		Ours	
	↑ F1.	↓ CER	↑ F1.	↓ CER
Main Body	0.8413	0.3804	0.8631 (+2.2%)	0.3221 (-5.8%)
Title	0.8823	0.3414	0.9045 (+2.2%)	0.2939 (-4.8%)
Annotation	0.7531	0.4721	**0.8078 (+5.5%)**	**0.3945 (-7.8%)**
Author	0.8522	0.3765	0.8749 (+2.3%)	0.3191 (-5.7%)
Punctuation	0.6334	-	**0.6983 (+6.5%)**	-

Table 3: Performance on different layouts.

Method	↑ F1.	↓ CER
Basic	0.8368	0.3939
Document Image Augmentation		
+Font Augmentation	0.8423	0.3775
+Layout Simulation	0.8486	0.3537
+Noise Injection	0.8429	0.3792
+*All*	**0.8526**	**0.3447**
Correction-Based Post-Editing		
+Correct Editions	0.8517	0.3635
+Correct Editions, Edition Alignment	**0.8623**	**0.3471**
Ours	**0.8726**	**0.3205**

Table 4: The result of ablation study.

shelf 1)Deepseek-VL2 (Wu et al., 2024) 2) GPT-4o (OpenAI, 2024); and LoRA finetuned 1) Qwen-2-VL (Wang et al., 2024); 2) GOT-OCR2.0 (Wei et al., 2024b); 3) Vary (Wei et al., 2023); 4) InternLM-XComposer (Dong et al., 2024); 5) InternVL2.5-8B (Chen et al., 2024).

From Table 2, we observe that all baselines exhibit noticeably low performance, highlighting the difficulty of our digitization. Among these baselines, VLM-driven baselines outperform traditional methods, emphasizing the effectiveness of unified generation models that leverage rich label semantics by encoding natural language labels into the output. Furthermore, our proposed model surpasses all previous studies. This result underscores the effectiveness of our end-to-end framework when applied to document images, validating our motivation to address inherent challenges by integrating our augmentation and correction.

5.3 Performance on Different Layouts

We first investigate our method's performance on different layouts to check if our model can handle the layout challenges. We apply the common sentence segmentation metrics on punctuations as previous works did (Srinivasan and Dyer, 2021).

As shown in Table 3, our model outperforms the baseline across all types. Notably, our model shows significant advantage over the annotations and punctuations, the former are appear smaller than the common main body as shown in Figure 2, while the latter do not exist in the image, are considered to be more difficult when compared with other layouts. These results indicate that our model effectively addresses the visual challenge posed by complex layouts in digitization.

5.4 Ablation Study

We then investigate the contribution of our Document Image Augmentation and Correction-Based Post-Editing. We use "Basic" to refer to the removing of two components, relying solely on the raw image input from the vanilla dataset.

As depicted in Table 4, using only raw images results in notably low performance, which is expected since the VLM is not pre-trained on historical document images. A significant improvement is observed when Document Image Augmentation is introduced, we attribute this to its ability to rein-

Method	Perplexity
Ground Truth	23.56
EffOCR	35.83
InternVL2.5-8B	32.78
Ours (w/o Post-Correction)	29.86
Ours	**26.13**

Table 5: Results of contextual coherence.

		InternVL2.5-8B	Ours
↑ **F1**	Neat	0.8832	0.9152 (+3.20%)
	Medium	0.8292	0.8729 (+4.37%)
	Scribble	0.7183	0.8184 (**+10.01%**)
↓ **CER**	Neat	0.3243	0.2604 (-6.39%)
	Medium	0.3914	0.3212 (-7.02%)
	Scribble	0.5465	0.4513 (**-9.52%**)

Table 6: Results for different neatness levels.

force robustness and generalization towards document images. Furthermore, our Correction-Based Post-Editing, which instead of sticking to pure-visual solution, incorporates contextual semantics to redeem semantically incoherent errors, further enhancing the performance.

6 Analysis and Discussion

6.1 Analysis of Contextual Coherence

We first investigate whether our method can better ensure the contextual coherence. We adopt the perplexity metrics (Wan et al., 2024) of Qwen3 (Yang et al., 2025) to evaluate the contextual coherence of each method, a lower perplexity represents that the LLM is less perplexed by the recognized texts, showing that the texts are more fluid and coherent.

As shown in Table 5, our method achieves the best coherence performance close to the Ground Truth among all methods. We attribute this improvement to our end-to-end design and our Post-Editing strategy that rely on surrounding semantics to correct the incoherent misrecognition.

6.2 Impact of Document Neatness

We further investigate the effects of our proposed method across different levels of annotated neatness. Specifically, we compare our method' s performance with the strongest baseline across the three neatness levels. As shown in Table 6, performance decreases as documents become more scribbled, since these irregularities hinder recognition. Notably, the more scribbled the document is, the larger advantage our model has, which we attribute to our font augmentation strategy that robust our model in challenging cases.

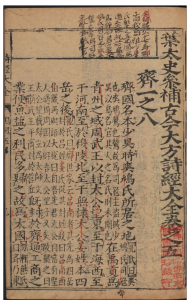
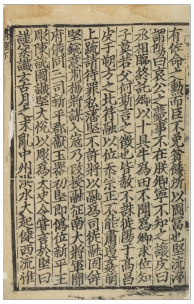
Input Image	Output (partial)
	**InternVL2.5-8B** 叶太史参补古今大方诗经 大全佛燕京珍藏齐一之八 全转服工
	Ours 叶太史参补古今大方诗经 大全卷之五 齐一之八
	Ground Truth 叶太史参补古今大方诗经 大全卷之五 齐一之八
	**InternVL2.5-8B** 符验以位秩， 示正不能，庸愚察顽。 上疏请待罪牧藩区，
	Ours 符融以位忝宗正， 不能肃遏奸萌， 上疏请待罪私藩。
	Ground Truth 符融以位忝宗正， 不能肃遏奸萌， 上疏请待罪私藩。

Table 7: Cases studies.

Besides, we also analyze the impact of fonts and our design of correct editions in Appendix B & C.

7 Case Study

We launch case studies to make a more intuitive comparison between our method and the strongest baseline InternVL2.5-8B in Table 7. We show that our method can effectively handle the layout recognition in the first example, where the baseline mistakenly take the seals as the title while our method successfully recognize it. In the second example, we illustrate that our method also performs better in recognizing the common main body: the baseline get errors in both the segmentation and characters, whereas ours successfully avoids the problems above and helps the final recognition.

8 Conclusion

In this study, we highlight previous Chinese historic document recognitions are inapplicable to real-world situation and propose a novel task: end-to-end Chinese historic document digitization that aims to recognize formatted and segmented sentence at one stop. We further propose ABCR dataset to fulfill the evaluation. With our Document Image Augmentation and Correction-Based Post-Editing, our method builds a strong benchmark and effectively promote the preservation and dissemination of Chinese historic document.

Limitations

The limitations of our work can be considered from two complementary perspectives. First, the sources of historical documents are relatively restricted, we did not incorporate documents that contain additional elements such as graphs, equations, or images, nor did we systematically analyze documents with more complex visual or typographic structures. A more comprehensive investigation can provide a deeper understanding of how different layouts influence of our approach.

Second, our study primarily focuses on a single language. Although we obtained promising results in this language, the extent to which these findings generalize to other linguistic contexts remains uncertain. Future work should include cross-lingual evaluations, domain adaptation or transfer learning experiments, and larger-scale benchmarking to determine the stability of the approach across languages, genres, and historical periods.

Ethical Statement

For building the dataset, we hired 10 annotators, each collects 50 samples with an average hourly wage above 75 CNY, which exceeds the local minimum of 19 CNY per hour.

References

Nadav Borenstein, Phillip Rust, Desmond Elliott, and Isabelle Augenstein. 2023. PHD: Pixel-based language modeling of historical documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 87–107, Singapore. Association for Computational Linguistics.

Jacob Carlson, Tom Bryan, and Melissa Dell. 2024. Efficient OCR for building a diverse digital history. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8105–8115, Bangkok, Thailand. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Hiuyi Cheng, Cheng Jian, Sihang Wu, and Lianwen Jin. 2022. Scut-cab: A new benchmark dataset ofăancient chinese books withăcomplex layouts forădocument layout analysis. In *Frontiers in Handwriting Recognition*, pages 436–451, Cham. Springer International Publishing.

Yongping Dan, Zongnan Zhu, Weishou Jin, Zhuo Li, and Mario Versaci. 2022. Pf-vit: Parallel and fast vision transformer for offline handwritten chinese character recognition. *Intell. Neuroscience*, 2022.

Brian Davis, Bryan Morse, Brian Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. 2023. End-to-end document recognition andăunderstanding withădessurt. In *Computer Vision – ECCV 2022 Workshops*, pages 280–296, Cham. Springer Nature Switzerland.

Rui Dong and David Smith. 2018. Multi-input attention for unsupervised OCR correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2363–2372, Melbourne, Australia. Association for Computational Linguistics.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, and 4 others. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *Preprint*, arXiv:2401.16420.

Yuning Du, Chenxia Li, Ruoyu Guo, Cheng Cui, Weiwei Liu, Jun Zhou, Bin Lu, Yehua Yang, Qiwen Liu, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. 2021. Pp-ocrv2: Bag of tricks for ultra lightweight ocr system. *Preprint*, arXiv:2109.03144.

JaidedAI. 2024. Easyocr.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022a. Ocr-free document understanding transformer. In *Computer Vision – ECCV 2022*, pages 498–517, Cham. Springer Nature Switzerland.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022b. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. *Preprint*, arXiv:2210.03347.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2022. Trocr: Transformer-based optical character recognition with pre-trained models. *Preprint*, arXiv:2109.10282.

Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. 2022. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *Preprint*, arXiv:2202.10304.

Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. 2019. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recogn.*, 90(C):337–345.

Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. 2020. Textsnake: A flexible representation for detecting text of arbitrary shapes. *Preprint*, arXiv:1807.01544.

Lijun Lyu, Maria Koutraki, Martin Krickl, and Besnik Fetahu. 2021. Neural ocr post-hoc correction of historical corpora. *Transactions of the Association for Computational Linguistics*, 9:479–493.

Hsing-Yuan Ma, Hen-Hsen Huang, and Chao-Lin Liu. 2024. Reading between the lines: Image-based order detection in ocr for chinese historical documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23808–23810.

Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. 2020. Joint layout analysis, character detection and recognition for historical document digitization. *Preprint*, arXiv:2007.06890.

Olfa Mechi, Maroua Mehri, Rolf Ingold, and Najoua Essoukri Ben Amara. 2019. Text line segmentation in historical document images using an adaptive u-net architecture. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 369–374.

Olfa Mechi, Maroua Mehri, Rolf Ingold, and Najoua Essoukri Ben Amara. 2021. A two-step framework for text line segmentation in historical arabic and latin document images. *Int. J. Doc. Anal. Recognit.*, 24(3):197–218.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Dezhi Peng, Lianwen Jin, Yuliang Liu, Canjie Luo, and Songxuan Lai. 2022. Pagenet: Towards end-to-end weakly supervised page-level handwritten chinese text recognition. *Preprint*, arXiv:2207.14807.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Alexander Robertson and Sharon Goldwater. 2018. Evaluating historical text normalization systems: How well do they generalize? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 720–725, New Orleans, Louisiana. Association for Computational Linguistics.

Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2019. Aster: An attentional scene text recognizer with flexible rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2035–2048.

Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, and Lianwen Jin. 2023. M5hisdoc: A large-scale multi-style chinese historical document analysis benchmark. In *Advances in Neural Information Processing Systems*, volume 36, pages 78483–78495. Curran Associates, Inc.

Yongxin Shi, Dezhi Peng, Yuyi Zhang, Jiahuan Cao, and Lianwen Jin. 2025. A large-scale dataset for chinese historical document recognition and analysis. *Scientific Data*, 12(1):169.

Srivatsan Srinivasan and Chris Dyer. 2021. Better Chinese sentence segmentation with reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 293–302, Online. Association for Computational Linguistics.

Benpeng Su, Xuxing Liu, Weize Gao, Ye Yang, and Shanxiong Chen. 2022. A restoration method using dual generate adversarial networks for chinese ancient characters. *Visual Informatics*, 6(1):26–34.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge fusion of large language models. In *The Twelfth International Conference on Learning Representations*.

Mingguang Wang. 2024. Accelerating the digitization of ancient books 加快古籍数字化进程. *People's Daily* 人民日报.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. 2019a. Decoupled attention network for text recognition. *Preprint*, arXiv:1912.10205.

Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. 2019b. Shape robust

text detection with progressive scale expansion network. *Preprint*, arXiv:1903.12473.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. LayoutReader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4744, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2023. Vary: Scaling up the vision vocabulary for large vision-language models. *Preprint*, arXiv:2312.06109.

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. 2024a. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *Preprint*, arXiv:2409.01704.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, and 1 others. 2024b. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Mohamed Yousef and Tom E. Bishop. 2020. Origaminet: Weakly-supervised, segmentation-free, one-step, full page text recognition by learning to unfold. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Implementing details

For the VLM, we fine-tuned the InternVL2.5-8B with four Nvidia A6000 GPUs in DeepSpeed ZeRO-2. Training was conducted for 30 epochs with a global effective batch size of 4 (per-device train/eval batch size = 1, gradient accumulation steps = 4). Inputs were truncated to a maximum length of 4096 tokens, and images were resized to 490 pixels on the longer side. Mixed-precision was enabled with bfloat16. The ViT backbone and the sampler were frozen throughout training. We applied fine-tuning via LoRA (the LoRA alpha is 128 and the LoRA rank is 64.), while other components of the language model were frozen. Optimization used AdamW with learning rate 5e-5, weight decay 0.1, $_2$ = 0.95, and a cosine learning-rate schedule with 1% warm-up.

For the corrector, we LoRA fine-tuned a Qwen2.5-7B with a 2,048-token cutoff. The LoRA alpha is 128 and the LoRA rank is 64. Optimization used AdamW with an initial learning rate of 1e-4, cosine decay, and a 10% warm-up ratio. We trained for 30 epochs with per-device batch size 1 and gradient accumulation of 8 steps (effective batch size 8) under bfloat16 precision.

B Impact of Fonts

We further investigate which type of font in our font augmentation can benefit the recognition more. We train our VLM with two sets of fonts for the train set: 1) Neat fonts such as Song (宋). 2) Scribbled fonts such as Huang (黄庭坚).

As shown in Table 8, performances within each group are similar. Between the two groups, neat fonts significantly outperform scribbled ones due to their clarity to convey character structure. Scribbled fonts like Mifu (米芾) help with specific styles but offer limited generalization. However, combining Song (宋体) with scribbled fonts further improves performance, indicating that scribbled fonts help address corner cases within the broad generalization of neat fonts.

C Impact of Correct Editions

As we state in Section 4.1, we abandon the substitution edition in the traditional Levenshtein Distance design of 4 corrections: *Insertion, Deletion, Substitution, Equal* and replace it with a pair of continuous *Deletion, Insertion* to narrow the searching space for the corrector LLM. We thus investigate the impact of our design by compare our 3-correction design with the traditional 4-correction design.

We use “Basic” to refer to the removing of

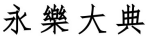
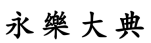
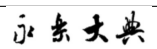
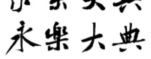
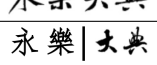
Font	Type	Illustration	↑ F1.	↓ CER
Song (宋体)	Neat		0.8614	0.3402
Kai (楷体)			0.8581	0.3423
Mi (米芾)	Scribbled		0.8534	0.3561
Huang (黄庭坚)			0.8589	0.3583
Song + Mi	Mixed		**0.8683**	**0.3282**
Song + Huang			0.8663	0.3293

Table 8: Result of different fonts.

Method	↑ F1.	↓ CER	↓ Corrector Inference Time (100 samples)
Basic	0.8368	0.3939	
Correction-Based Post-Editing			
+ 4-Correction	0.8461	0.3713	**23.18s**
+ 3-Correction	**0.8609**	**0.3486**	26.42s
Ours	**0.8726**	**0.3205**	27.26s

Table 9: The result of different correct editions.

two components, relying solely on the raw image from vanilla dataset. The "Correction-Based Post-Editing" refers to the adding of corrector post-editing, while "Ours" refers to the adding of both post-editing and document image augmentation.

As shown in the Table 9, by dropping substitution, our 3-correction narrows the search space and reduces ambiguity, it outperforms the 4-correction design in both F1 and CER, at the cost of a slight increase in inference time since the 3-correction design uses two corrections to replace the traditional substitution, leading to a longer target sequence length.